phgfoundation
making science work for health

Pathogen Genomics Into Practice

July 2015

Authors

Leila Luheshi, Sobia Raza, Sowmiya Moorthie, Alison Hall, Laura Blackburn, Chris Rands, Gurdeep Sagoo, Susmita Chowdhury, Mark Kroese and Hilary Burton

Acknowledgements

The PHG Foundation is grateful for the expert advice and guidance provided by the workshop participants, consultees, and the Project Steering Group for their support and direction. Full acknowledgments are listed on p.228.

We would like to thank the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre for funding this work.

This report can be downloaded from our website:
www.phgfoundation.org

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK

Tel: +44 (0) 1223 761 900

Correspondence to:
sobia.raza@phgfoundation.org
leila.luheshi@phgfoundation.org

How to reference this report:
***Pathogen Genomics Into Practice* PHG Foundation (2015)**

ISBN 978-1-907198-18-2

The PHG Foundation is an independent, not for profit think-tank (registered in England and Wales, charity no. 1118664, company no. 5823194), working to achieve better health through the responsible and evidence based application of biomedical science.

Contents

Introduction

Part I

Part II

Part III

Part IV

Executive summary

Pathogen genomics in this report is defined as the application of genome sequencing technologies to the characterisation and analysis of pathogens for the purpose of informing clinical and public health investigations of infectious disease. In principle, this technology has the power to transform the management of infectious disease in England.

Introduction

Pathogen genomic methods offer two key advantages over existing microbiological methods for investigating infectious disease:

- Whole genome sequencing can be used to discriminate between pathogens with greater sensitivity and often specificity than current methods, enabling outbreaks to be resolved or ruled out with greater speed, accuracy and confidence

- Genomic sequencing is an immensely powerful technology in that it can provide a description of a wide range of clinically and epidemiologically relevant characteristics of a pathogen, including identity, virulence determinants, drug resistance and relatedness to other pathogens. The 'generic' nature of genomes (which are all constructed from the same types of molecule) also means that the same technologies used for genomic analysis of one organism can, in principle, be applied to any other

Given sufficient understanding of the clinical and epidemiological significance of pathogen genome variation, this technology could in the future be used as a frontline tool in the analysis and management of most (if not all) of the pathogens that represent a threat to human health.

Can pathogen genomics improve patient and population health?

There is now a substantial body of peer-reviewed literature demonstrating how in principle pathogen genomics can be used to improve the management of infectious disease through improved diagnosis, detection and tracking of antimicrobial resistance and outbreak control. However, due to current limitations of genomic technology and of our understanding of the clinical and epidemiological significance of genomic variation for many important pathogens, utility of pathogen genomics, with the exception of tuberculosis and HIV, is currently limited almost entirely to use in outbreak detection and

If the rate-limiting step in the performance of existing infectious disease management systems arises from care pathway factors outside of current microbiological practice ... implementation of pathogen genomics may have little or no impact on health outcomes.

control. The vast majority of current diagnostic microbiology practice seems likely to continue in its current form at least until genomic technology improves to a point where it can deliver clinically useful results that compete with existing traditional microbiology methods on both turnaround time and cost.

Most evidence supporting the utility of pathogen genomics in microbiology practice centres around demonstrations of its ability to enhance the sensitivity and specificity of outbreak investigations, particularly for healthcare-associated infections (HCAIs) such as MRSA but also for community-acquired infections such as tuberculosis. However, from a health policy perspective it is vital to note that these investigations have mostly been retrospective and were unable to measure objectively their impact on patient and population health. This limits their value in determining whether use of pathogen genomics in real world clinical and public health settings would have resulted in significantly improved health outcomes for either individuals or populations. To do this would require prospective trials designed specifically to test the effectiveness of healthcare and public health service models that incorporate genomic information compared to existing practices.

As yet no such trials have been published, although several small scale, pathogen-specific pilot studies are now underway in England. Analysis of the results of such evaluations will be essential to determine whether infectious disease management services, informed by pathogen genomics, can be realised within our current health service operating frameworks. They will also help to determine whether the costs involved in establishing such services and tailoring clinical and public healthcare pathways to exploit them outweigh the benefits.

With this significant gap in the evidence base of effectiveness and cost effectiveness to support implementation, the case for the use of pathogen genomics relies on the assumption that the demonstrable improvements in analytical performance of this technology – compared to existing methods used for outbreak detection and investigation – are almost certain to deliver significant improvements in health outcomes. However, if the rate-limiting step in the performance of existing infectious disease management systems arises from care pathway factors outside of current microbiological practice, or if the major needs arise in diseases where testing is not currently amenable to improvement through genomic analysis, implementation of pathogen genomics may have little or no impact on health outcomes.

Developing a successful pathogen genomics informed infectious disease management system

Our project has reviewed the current state of science and clinical practice in pathogen genomics and gathered evidence from individuals working across the health economy, including health policy makers, clinical and public health practitioners and academic researchers developing genomic analysis tools. We have leveraged their expertise, in combination with the in-house expertise of the PHG Foundation in genomics and public health, to identify two objectives that the health system should seek to achieve in order to realise the potential benefits of genomics in the field of infectious disease management.

- **Ensure effective genomics service implementation and delivery, where this is justifiable on the basis of evidence, in the short term**

- **Drive innovation and expansion in the range of genomics informed / enabled services that can be developed and delivered in the long term**

What is key to achieving these objectives? The results of our research and analysis show very clearly that the effectiveness of any efforts to implement pathogen genomics will depend on the implementation of a nationally coordinated system of service development and delivery. We have identified two key features of this system that must be delivered if benefits of genomics technology to patients and populations are to be realised:

Data integration

Individual pathogen genomes cannot be usefully analysed in isolation. All clinically and epidemiologically meaningful information derived from pathogen genomes depends on our ability to compare them to other genomic data *e.g.* a pathogen genome isolated from a patient with the same disease. Equally it depends on our ability to combine the genomic data about a pathogen with relevant epidemiological information – for example when and where it was isolated and associated clinical information.

The timely collation, integration and sharing of genomic and clinical / epidemiological metadata across all parts of the health system involved in the delivery of pathogen genomics informed infectious disease services is therefore essential. This is particularly the case where we seek to realise benefits from this technology to deliver improvements in outbreak detection and resolution, where failure to share and integrate genomic and clinical data across different NHS and PHE laboratories will fundamentally undermine these efforts.

Not only will effective data integration serve to maximise the effectiveness of services that can be delivered now, such as outbreak investigation for certain HCAIs, it is also essential for driving innovation and the expansion of services that need to be developed for future use. Accordingly, it will be vital that as much of the data generated by clinical and public health services as possible is made available to the research and development community. Access to such databases will enable them to increase understanding of the significance of pathogen genome variation and to develop the tools to analyse and interpret this variation that will become vital parts of future clinical and public health services. Although the development and refinement of analytic tools and methods can be accelerated by widening access to an integrated data resource, the widespread clinical deployment of pathogen genomics will be contingent on the availability of robust computational software for data analysis and adequate computing infrastructure. The development of accessible and automated computing software, underpinned by scalable and sustainable computational infrastructure, should therefore be prioritised by pathogen genomics service providers in order to support national use.

The results of our research and analysis show very clearly that the effectiveness of any efforts to implement pathogen genomics will depend on the implementation of a nationally coordinated system of service development and delivery.

Strategic coordination and leadership

Management of infectious disease and its impact on human health in England requires the input of a wide range of organisations including those with responsibility for public health (PHE and local authorities), delivering healthcare (NHSE), managing food safety (FSA) and animal health (APHA). It also depends on the input of a wide range of professional groups, ranging from infectious disease physicians, medical microbiologists and infection control nurses, to clinical laboratory scientists and academic researchers. Each of these organisations and professional groups has a stake in realising the effective development and implementation of pathogen genomics services. Consequently these efforts can only succeed where there are clear mechanisms to achieve strategic coordination of policy at an organisational level and where there are mechanisms to ensure that professional groups are supported to work together to share and develop the knowledge, expertise and best practice that will enable them to deliver the highest quality care to their patients and to protect the health of our population.

The roadmap

Within our report we detail – and provide evidence for – over 30 recommendations to support the achievement of the objectives set out above. We have presented these recommendations within the framework of a roadmap, which has two parallel routes to achieving patient and population benefit from pathogen genomics:

- Steps needed to achieve implementation of the pathogen genomics informed services for which we currently have sufficient evidence of utility, and the ability to deliver accurate and meaningful analysis to clinicians and public health practitioners

- Steps needed to enable the research and development that will broaden the range of services that can be improved by the introduction of pathogen genomics in the longer term

We recognise that organisations and groups within the UK are already progressing along these paths. However, in their current form – focused on discrete service pilots for a selected number of pathogens, being deployed across a small number of locations, and involving only a limited proportion of relevant stakeholder organisations and professional groups – we believe they are unlikely to achieve the scale and depth of benefit that full-scale implementation of pathogen genomics could eventually bring.

The catalyst – enhancing service effectiveness now and accelerating innovation and service development into the future

As we highlight above, systems for effective data integration (that can be used for all pathogens and for all potential applications of genomics) and strategic coordination of policy and practice will be required to accelerate the rate at which genomics services are developed and implemented and to enhance the effectiveness with which they are delivered.

It is our view that these objectives can only be achieved through the development of a new, catalytic 'core' within this roadmap. We define the catalyst as: *A set of real or virtual structures that amplifies and integrates the current activities in pathogen genomics to accelerate and increase the effectiveness of their impact on patient and population health.*

Our proposed catalyst performs four functions:

1. Infrastructure to provide a repository for data, knowledge and samples necessary to fulfil the data integration demands of the system

2. A focus for collaboration within and between the health services, academia and industry

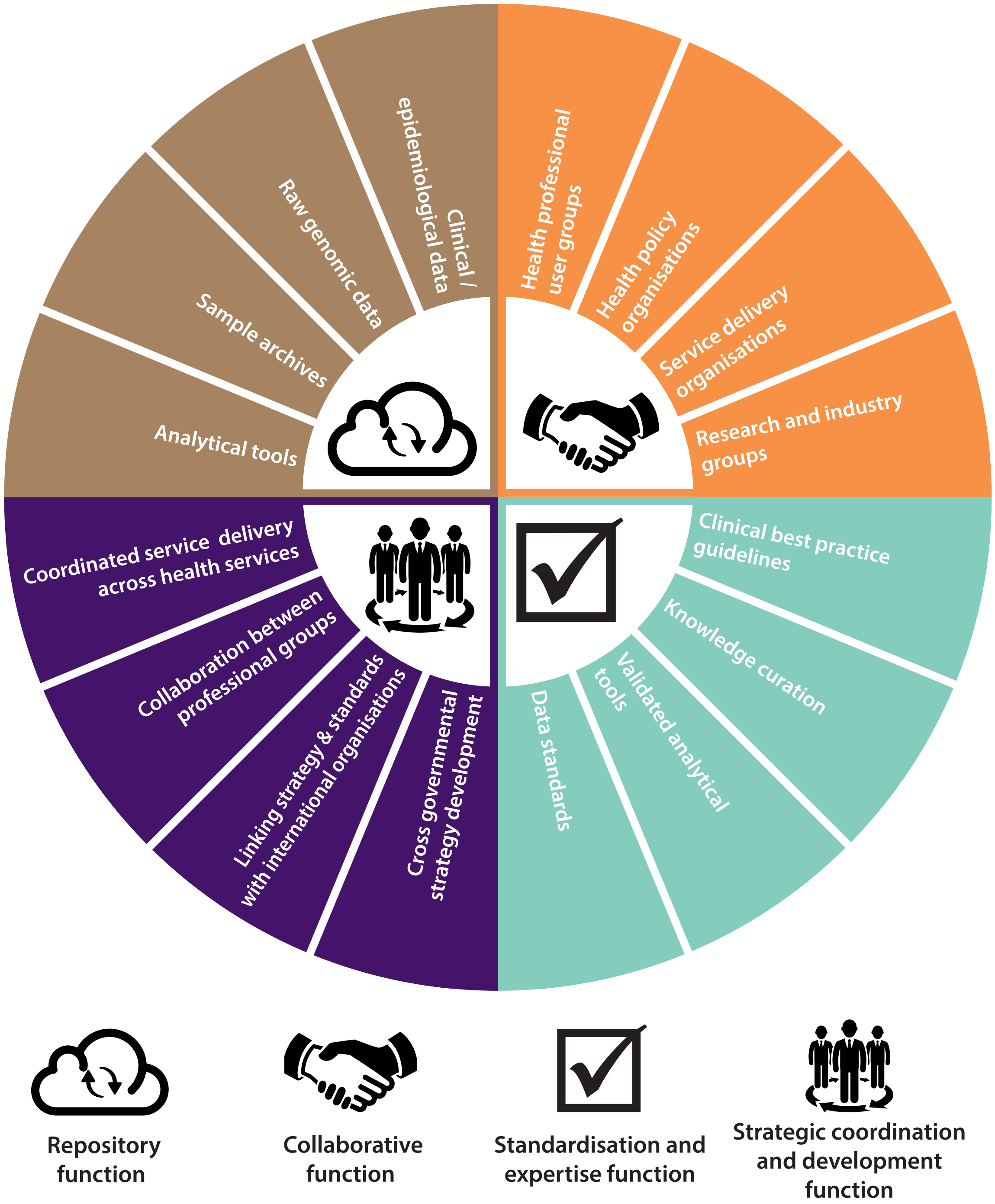
3. A mechanism to facilitate development and diffusion of standards and sharing of expertise

4. Establishment of a leadership group that can oversee and drive forward the strategic coordination and development of policies and practices for the use of pathogen genomics across all relevant stakeholder organisations in England

It is the conclusion of our report that without the establishment of these functions many of the proposed benefits of pathogen genomics for patient and population health are unlikely to be achieved.

Enhancing service effectiveness and accelerating the rate of innovation and service development can only be achieved through the development of a new, catalytic 'core' within this roadmap.

Figure 1 Catalyst


Clinical / epidemiological data
Raw genomic data
Sample archives
Analytical tools
Health professional user groups
Health policy organisations
Service delivery organisations
Research and industry groups
Clinical best practice guidelines
Knowledge curation
Validated analytical tools
Data standards
Coordinated service delivery across health services
Collaboration between professional groups
Linking strategy & standards with international organisations
Cross governmental strategy development
Repository function
Collaborative function
Standardisation and expertise function
Strategic coordination and development function

Conclusions – an implementation dilemma?

Adoption and delivery of the roadmap – and in particular the catalyst proposed within this report – would require significant investment of resources by both policy and delivery organisations within the health system in England. Any decision to commit such resources will require sufficiently strong evidence to support the proposition that these investments would provide the anticipated returns in terms of health and economic benefits.

As noted earlier, there is currently a lack of direct evidence demonstrating that when implemented as part of real world pathways of infectious disease management and patient care, pathogen genomics can deliver on its promise. Furthermore, current implementation pilots, targeted at individual pathogens and developed in the absence of an integrated system-wide approach to data and knowledge integration and service delivery, are low risk but of limited reward as they are restricted in their capacity to generate this evidence. This limitation stems from the value and impact of pathogen genomic information being directly correlated with the amount of information available, the effectiveness with which it is integrated with other sources of information, its accessibility to innovators and the degree of coordination of the systems required to deliver services that rely upon it. The current absence of the systems necessary to meet these requirements significantly reduces the likelihood that the pilots will be successful in demonstrating positive health outcomes, and even where they do, their generalisability and wider adoption and diffusion across the health service will be severely hampered.

Continuing with the current gradual and fragmentary approach to implementation therefore poses a risk that must be acknowledged and addressed: it is less likely to generate the desired impact in terms of positive outcomes for patients and the population in England and is more likely to lead to less efficient use of resources within the health system than adopting the type of system-wide and integrated approach embodied by the catalyst we propose.

Continuing with the current gradual and fragmentary approach to implementation therefore poses a risk that must be acknowledged and addressed...it is more likely to lead to the less efficient use of resources within the health system than adopting the type of system-wide and integrated approach embodied by the catalyst we propose.

Conversely, we must also acknowledge that any decision to invest in building and operating the catalyst would entail taking a calculated risk, requiring its establishment prior to the availability of sufficient evidence to support all aspects of development. Nevertheless, it is the conclusion of our analysis that unless this risk is taken, the opportunity to realise the benefits of pathogen genomics for our population may well be lost.

Furthermore, if through its investment in genomics, England aspires to lead the world in precision medicine, then it must recognise that pathogen genomics, if implemented effectively, represents an opportunity to prove that genomics can truly 'transform' health services. UK scientists and clinicians have laid the foundations for this transformation, but the real challenge begins now with the need for health services leaders to direct and invest to establish the necessary systems and infrastructure to make pathogen genomics part of routine and effective clinical and public health practice. If they can achieve this, then they will truly lead the world.

1 The convergence of science and policy: why now is the right time to bring pathogen genomics into practice

The Pathogen Genomics Into Practice project, and this report in particular, is the PHG Foundation's response to the convergence of demand for improvements in the management of infectious disease.

1.1 Infectious diseases: a persistent threat to the health of the nation

The introduction of vaccination and antibiotic therapy during the 20th century has contributed significantly to dramatic reductions in the prevalence of infectious diseases in the UK. As a result, non communicable diseases have overtaken infectious diseases as the principal causes of morbidity and mortality in developed nations, including England. However the threat from infectious diseases remains and must be addressed: the 2011 Annual Report of the Chief Medical Officer (CMO) for England stated that '*In 2010 infectious disease accounted for 7% of all deaths*' and that the economic burden from infectious disease was estimated to be £30 billion a year.

Healthcare-associated infections, pandemics and the rising spectre of antimicrobial resistance pose particularly significant challenges to the effectiveness of existing systems for the management of infectious disease.

Healthcare-associated infections, pandemics and the rising spectre of antimicrobial resistance pose particularly significant challenges to the management of infectious disease and demand urgent responses from health policy makers and health practitioners alike. In addition to these particularly 'high profile' threats, there also remains a significant burden placed on the health of the nation by the occurrence of more common infections such as those causing gastrointestinal illness and respiratory infections. These infections, which disproportionately affect the health of the very young, old and the immunocompromised, remain prevalent despite advances in vaccination and antibiotic therapy. They constitute the majority of the burden of infectious disease on the health of the population and the economy in England.

1.2 Improving the effectiveness of infectious disease management in England: the limitations of existing microbiological practice

These persistent threats arising from infectious diseases pose continuing challenges to management systems and, in particular to the effectiveness of microbiological methods for investigating pathogens. Microbiology services play a central part in the management of infectious disease by:

- Identifying the pathogens causing infections
- Determining the most appropriate drugs with which to treat them
- Investigating sources and routes of transmission
- Undertaking surveillance to detect new emerging pathogens and development of antimicrobial resistance in existing pathogens

They are used by a range of practitioners across the health system in the context of individual patient care and disease prevention at population level. It is of particular concern, therefore, that existing microbiological methods, many of which were developed over 100 years ago, are limited in the information they can provide. These limitations may be overcome by new technologies such as pathogen genome sequencing.

1.3 The transformative power of pathogen genomic technology and knowledge

The Chief Medical Officer stated in her 2011 report on infectious disease that:

"…the exciting potential opportunities are from the impact of developing technologies, particularly those that the advances in genomic medicine make possible.

Whole genome sequencing of infectious agents gives the ultimate in resolution between two related pathogens. Rapid technological advances in DNA sequencing have led to the availability of benchtop sequencers that are drastically reduced in cost and likely to become cheaper. These can sequence multiple bacterial or viral genomes in less than a day. The use of these methods will almost certainly become the standard diagnostic approach and have the potential to be the impetus for a step change in the effectiveness of surveillance. Specified pathogens isolated in diagnostic laboratories can be sequenced and this information fed into current surveillance systems to track disease trends. Such a system could also be used to monitor the emergence and spread of clinically important bacterial drug resistance."

This sets out very clearly that pathogen genomic science has, in principle, the ability to transform the microbiology led investigation and management of infectious disease.

The transformative power attributed to pathogen genomics arises from:

- **Universality of the genomic code –** The blueprint for the construction and function of every pathogen is 'written' in the same universal language (DNA or RNA), and thus only a single technology (genome sequencing) is required to read and decode the blueprints of a highly diverse range of organisms.

- **Multi-functionality of genomic information –** Once decoded, the genome of a pathogen reveals a host of clinically relevant information including:

  - Identity

  - Resistance to different drugs or vaccines

  - Relatedness to similar pathogens isolated from other patients or the environment

  - Ability to cause illness

- **High resolution of genomic information –** the genome sequence of a pathogen consists of millions of potentially discriminatory pieces of information. Comparison of these high resolution pictures of pathogens allows their relatedness to be determined with an accuracy that is orders of magnitude greater than that achievable with current methods.

- **Recent advances in genomic technology and knowledge –** genome sequencing technology is now sufficiently affordable, rapid, stable and reliable for use in clinical applications. These advances have been complemented by significant advances in our ability to analyse and understand pathogen genomes, and together they are enabling genomics-informed healthcare to become a reality.

Well-targeted funding of translational research projects has also played a role in ensuring that the necessary research and development is now being undertaken to place England at the forefront of efforts to capitalise on these powerful features of genomic technology, particularly in the area of infectious disease.

1.4 The Pathogen Genomics Into Practice project

1.4.1 Rationale

The Pathogen Genomics Into Practice project, and this report in particular, is the PHG Foundation's response to the convergence of: demand for improvements in the management of infectious disease, the scientific and technological capacity to deliver pathogen genomic analysis to the clinic, and the political will to make this a reality.

We have undertaken a programme of research, analysis and extensive stakeholder engagement to produce a roadmap of the policies and practices necessary to realise our aim to supporting the development and delivery of genomics informed infectious disease services that are evidence based, high quality, available population wide, and on an equitable basis.

1.4.2 The report

The purpose of this report is to place our policy roadmap into context, to provide the evidence base that underpins our conclusions, and to inform the diverse range of organisations involved in delivering infectious disease management in England both of the potential of pathogen genomics, and the challenges each will face to realise the benefits.

This report is divided into four parts:

- **Part I** introduces pathogen genomes and explains for a non-expert audience how these are in many ways more diverse and complex, but also more accessible for analysis, than human genomes. We introduce pathogen genome structure and function and how genomes can be sequenced and analysed to derive information with clinical and public health utility.

- **Part II** describes what the application of pathogen genomics to microbiological investigations can achieve in principle, given the current state of technology and knowledge. We present our synthesis of the evidence for the utility of pathogen genomic information in different aspects of infectious disease management. We also identify areas of microbiological practice that are both close to and further from being able to harness the benefits of genomics.

- **Part III** constitutes the core of the analysis arising from our research and stakeholder engagement activities. Embedded within this analysis are recommendations for the policies and practices required to ensure the successful establishment of pathogen genomics in mainstream infectious disease management. These recommendations arise from our understanding of:

    - The current configuration of services for the delivery of microbiological investigations

    - The scientific, clinical and economic evidence base required to support implementation

    - Principles underpinning how genomics strategies and services will need to be developed and configured, locally and nationally, to realise the benefits of genomics as part of a national infectious disease management system

    - The challenges and opportunities in managing, and exploiting for current and future population health benefit, the vast quantity of genomic data that will be generated as pathogen genomics is deployed

> *Well-targeted funding of translational research projects have played a role in ensuring that the necessary research and development is now being undertaken to place England at the forefront of efforts to capitalise on these powerful features of genomic technology, particularly in the area of infectious disease.*

- Part IV sets out the conclusions of our analysis and maps out how our recommendations can be taken forward by the relevant stakeholder groups to ensure:

    - Acceleration of implementation and maximisation of the effectiveness of pathogen genomics informed infectious disease management services that can be developed and delivered now

    - Timely and appropriate expansion of the range of pathogen genomics informed services, and the technology and knowledge required to deliver these

1.5 Methodology

This report and its findings are the synthesis of eighteen months of research, analysis and stakeholder consultation undertaken by the PHG Foundation. In addition to our own research into peer reviewed academic literature, and the other public sources of information on activity within the English health services, we have relied heavily on external expertise to inform our analysis and our conclusions. This expertise has been gathered through two workshops (whose participants are listed in the acknowledgements at the end of this report) and also through extensive one-to-one or small group engagement with academic researchers, clinical and public health service practitioners, and policy makers. Many of these experts have contributed significantly to the writing of this report, through their insights, unpublished information about their research or service development, and through review of the factual accuracy of this report. Their contributions are acknowledged on **p.228.**

We are extremely grateful to all of the participants in this project, who have given their time and expertise without compensation to support our work. Whilst this work has benefitted greatly from their input, final responsibility for the content, analysis and any errors within this report lies entirely with the PHG Foundation authors (listed at the front of this document). The views expressed in this report solely represent those of the PHG Foundation, and are not necessarily those of individuals or organisations who have contributed to its development.

Part I

In the following three chapters we introduce the key scientific and technical concepts and knowledge that are relevant to understanding how pathogen genome information can be obtained and used to improve the management of infectious disease. We focus on the three principal elements:

- Pathogen genomes – we describe their structure, diversity and adaptability and consider how these may help or hinder the utility of pathogen genomic information in the context of infectious disease management

- Sequencing pathogen genomes – we consider the current genome sequencing technologies that have been applied to pathogen genomes and the benefits and limitations associated with their use in a clinical or public health context. We also highlight the potential of emerging, but as yet unproven, sequencing technologies to impact on the practice of pathogen genomics

- Analysing pathogen genomes – we describe the principles underlying the computational analysis of pathogen genomes and how different types of information can be derived from them. We also consider the benefits and limitations of current pathogen genome analysis approaches in the context of their ability to deliver clinically useful information

2 An introduction to pathogen genomes

This chapter introduces pathogen genomes and describes some features of bacterial and viral genomes that determine the utility or limitations of sequencing them as a tool for informing the management of infectious disease.

2.1 Background

As is the case for human genomes, the genome sequence of a bacterium or virus is effectively a blueprint describing the potential characteristics or traits of that organism. The genome also provides a record of ancestry, revealing genetic relationships with other members of the same species and also more distantly related ones. Sequencing pathogen genomes can therefore, similarly to sequencing human genomes, be used to characterise identity, predict activity, and understand genetic relationships. These are the broad aims of most microbiological investigations undertaken in clinical and public health laboratories, and the high precision and detail with which deep (whole) genome sequencing enables them to be achieved has the potential to substantially reduce the burden of infectious diseases.

At the fundamental molecular level the genomes of bacteria, viruses and humans share much in common. They are constructed from the same types of nucleic acids (DNA and RNA) and use the same genetic code consisting of the nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T) for DNA or uracil (U) for RNA. This means pathogens are broadly amenable to the same sequencing technologies applied to the genomes of humans and other organisms. Nevertheless, the ways in which these DNA or RNA molecules are structured, inherited and subject to change over time differ greatly between humans and the immense diversity of pathogenic organisms that can infect them. In some cases microbiologists rely on these differences to extract clinical and public health utility from the analysis of pathogen genomes, and in other cases these differences pose significant challenges to analysis and failure to understand them may undermine the utility of the analysis.

2.2 The bacterial genome – a multi-part, dynamic storage and transfer medium

2.2.1 The bacterial chromosome

The core bacterial genome usually consists of a single, circular, double stranded DNA molecule (*i.e.* one chromosome), although some bacterial genomes are comprised of multiple chromosomes. The genomes range in size from 15,000 to 13,000,000 base pairs (bp), but are most commonly a few million bp that collectively encode approximately 3000 protein coding genes. By contrast a human genome is roughly 1000 times larger than this at over three billion bp, contains approximately 19,000 protein coding genes and typically consists of 23 chromosomes pairs of chromosomes in each cell. It is striking that while the human genome is 1000 times larger than the average bacterial genome it contains only around six times the number of protein coding genes. This greater gene density in bacteria reflects the fact that bacterial genes have fewer introns (noncoding sequences between protein-coding exon sequences), and neighbouring genes have much smaller intervening intergenic regions of noncoding, potentially regulatory, DNA.

Table 2.1 Approximate haploid genome size, gene content, and gene density for human, bacterial and viral genomes

Genome	Genome size (bp)	Number of protein coding genes	Gene density
Human	3,400,000,000	19000	0.0006
Bacterium (*M. tuberculosis*)	4,000,000	4000	0.1
Bacteria (MRSA)	2,900,000	2700	0.09
Virus (HIV)	10,000	9	0.09

While a small, compact genome may be advantageous for the bacterial cell, in that it can be replicated and transmitted to 'daughter cells' following division in a relatively rapid and energy efficient manner, the lack of a second copy of its chromosome (except in a few rare cases) may mean that if mutations occur in its genes that cause them to malfunction, the bacterium is likely to be adversely affected, as it has no 'back up' working copy of the gene on another copy of the chromosome.

The relatively simple construction of the bacterial chromosome and small size are advantageous in that they enable it to be sequenced rapidly (in a matter of hours) and at low cost, for a price of less than £100 (see chapter 3). Moreover, the small number of functional elements harboured within pathogen genomes makes the resulting sequence much easier to interpret than for the human genome, which is not only large but contains many 'junk' sequences that may not be important for human health[1]. Additionally, functional information on the genome may not even be necessary to provide medically useful information from pathogen sequences. Simply tracing the evolution of pathogen lineages through genome sequencing may provide useful enough epidemiological information to help contain an outbreak, without the need for functional annotation of the genomes.

However, pathogens are highly diverse and evolve very rapidly compared to human genomes. This means that inference of their ancestry and tracing their movement can be challenging. Such identification generally relies on comparing the sequences of different pathogen genomes, and such comparisons are harder for more divergent species.

The situation is further complicated for bacteria by the common presence of extra-chromosomal elements such as plasmids, mobile DNA elements called transposons and even DNA from other organisms such as bacteriophage viruses. These extra-chromosomal genetic elements can contain genes of significance to the infectious nature of the organism (its pathogenicity) so it is desirable that they are also analysed during bacterial genome sequencing to ensure a complete picture of the organism can be constructed.

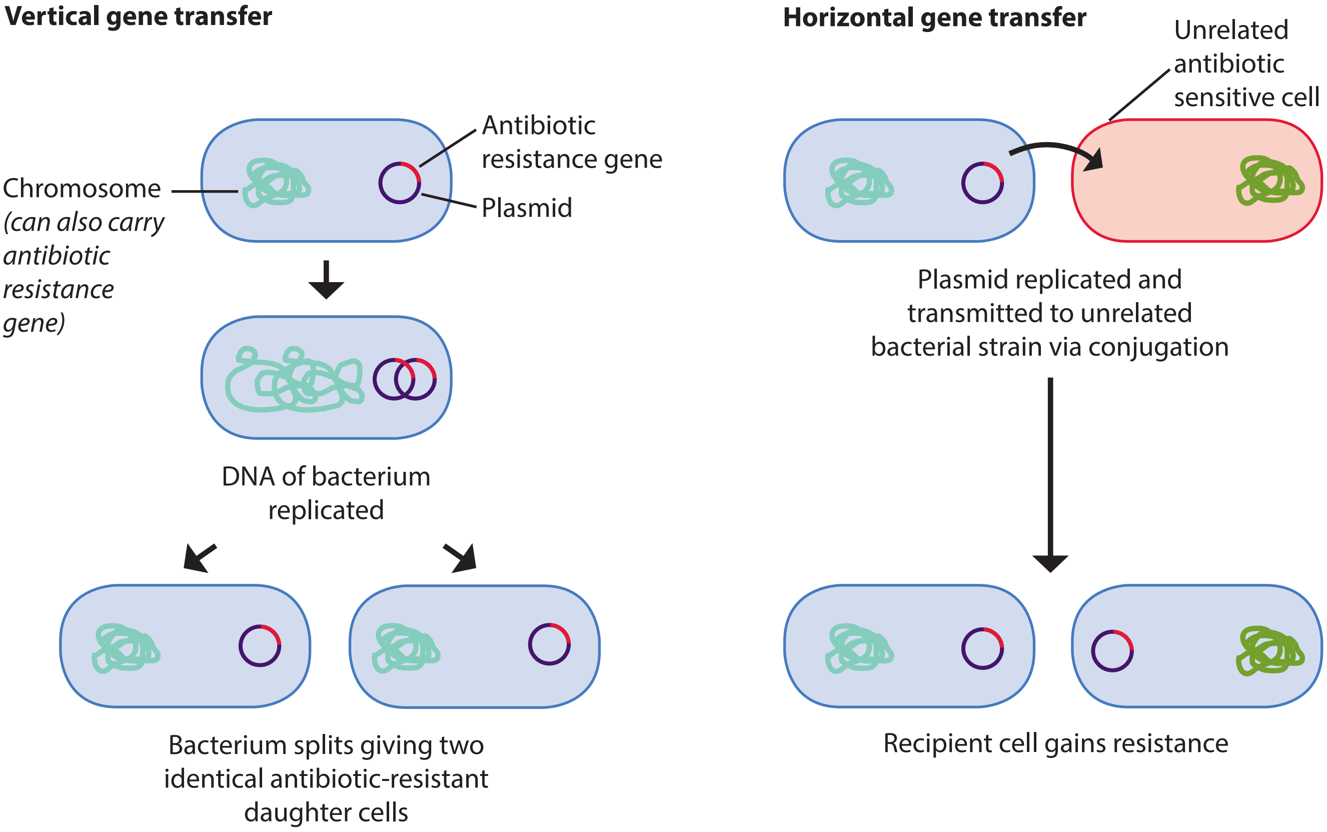
2.2.2 Plasmids – vectors of virulence and antibiotic resistance

Plasmids are small, typically closed circular DNA molecules that do not carry genes essential for the survival of the bacterium. Instead they carry genes that potentially confer adaptive advantages to the bacteria under specific conditions such as environmental stress or the application of antibiotics. Thus, sequencing the plasmids as well as the chromosome of a bacterium is an important part of understanding its potential to cause disease or evade treatment or our immune system.

Both plasmids and bacterial chromosomes can be replicated during cell division and inherited by daughter cells, so-called vertical transmission. Crucially for our understanding of both infectious disease and its resistance to treatment, plasmids can also replicate and be transmitted 'horizontally' from one bacterium to another without cell division occurring, using a process known as conjugation. This behavior represents an important mechanism by which antibiotic resistance and virulence can be transmitted not only within generations of the same bacterial strain, but also between unrelated bacterial strains and species. Thus plasmids enable bacterial strains to switch rapidly from being harmless to being pathogenic, and can in a single step transform a pathogenic bacterium from being antibiotic-susceptible to antibiotic-resistant. It should be noted however, that not all bacterial species can donate or receive plasmids and so plasmid-mediated horizontal gene transfer is not universal.

In species where plasmid transmission does occur, sequencing of plasmid DNA may be crucial to understanding and following the transmission of plasmid mediated antibiotic resistance. For example, this is of particular significance for monitoring drug-resistance in Gram-negative bacteria since there are few new antibiotics or strategies in development for this category of bacteria, and also for understanding the spread of resistance to carbapenams, an antibiotic class of 'last resort' for many complicated bacterial infections[2].

Figure 2.1 Transfer of antibiotic resistance: vertical *versus* horizontal gene transfer


Vertical gene transfer
Chromosome
(can also carry antibiotic resistance gene)
Antibiotic resistance gene
Plasmid
DNA of bacterium replicated
Bacterium splits giving two identical antibiotic-resistant daughter cells
Horizontal gene transfer
Unrelated antibiotic sensitive cell
Plasmid replicated and transmitted to unrelated bacterial strain via conjugation
Recipient cell gains resistance

Box 2.1 What do genes on plasmids do?

- **Resist antibiotics –** gram-negative bacteria carry plasmids whose genes confer resistance to antibiotics. Spread of these is a major threat to public health
- **Allow bacteria to compete with one another** – genes on some plasmids inhibit the growth of similar or closely related bacterial strains, potentially resulting in the replacement of benign bacteria that colonise humans with pathogenic strains that cause disease
- **Produce toxins** – enterotoxins, produced by some *E. coli* and neurotoxins, such as those produced by *Clostridium tetani*, are examples of bacterial toxins that cause disease in humans
- **Resist the toxic effects of heavy metals** – genes that allow bacteria to live in the presence of heavy metals enable them to survive in extremely harsh environments
- **Invade human cells** – genes produced on some plasmids enable bacteria to attach to and enter human gut cells damagmaging them and causing disease in enteroinvasive *E. coli* infections

2.2.3 Bacteriophages – viral vectors for the transfer of resistance and virulence genes

The bacterial genome can also contain genes acquired from infecting viruses called bacteriophages. Bacteriophages are able to insert their genetic material into bacterial cells, where the viral genes can become integrated into the bacteria's own genome, a process known as transduction. Therefore, as with plasmids, bacteriophage genes can also confer pathogenic properties on the bacteria that the viruses infect. For example, when bacteriophages insert their DNA into the chromosome of the V*ibrio cholerae* bacterium, the bacterial genome acquires several of the viral genes that encode toxins responsible for causing diarrhoea in cholera. Similarly to plasmids, these bacteriophages can then be transmitted to other cells in two ways; they can either be passed to daughter cells during chromosome replication (as they are now an integral part of the bacterium's own genome, or released from the bacterial chromosome, assemble as free virus particles and be released during bacterial lysis to infect other cells. This latter process can be significant in clinical and in public health terms as it potentially enables the one-step creation of new pathogenic strains of bacteria through the bacteriophage-mediated transfer of pathogenic genes between bacteria cohabiting the same host or environment.

2.2.4 Transposons – non autonomous vectors for mobilising resistance and virulence

Transposons (or transposable genetic elements), also known as 'jumping genes', are DNA sequences that can move from one location on the genome to another. Transfer of transposons occurs between plasmids, or between plasmid and chromosome within a bacterial cell. Unlike plasmids, transposons cannot replicate autonomously and so must be embedded within either plasmids or chromosomes in order to be transmitted horizontally or vertically between bacteria. Their ability to transfer antibiotic resistance genes between plasmids and to the bacterial chromosome means that they are significant mediators of the spread of antibiotic resistance.

2.3 Genomics of important bacteria for human health in the UK

Bacteria and viruses share many biological features as microorganisms but also have important differences. Critically from a medical perspective, bacterial infections can be stopped by antibiotics, while antivirals inhibit various stages of the viral life cycle but tend not to destroy it completely. Pathogenic bacteria are involved in a large number of different important human diseases and illnesses. Arguably the most important bacteria for health in the UK that genomics technologies have the potential to reduce the health burden from are *Mycobacterium tuberculosis,* MRSA, and *Clostridium difficile*, each of which is associated with hundreds of deaths deaths in the UK annually in addition to other social and economic costs.

2.3.1 *Mycobacterium tuberculosis*

Mycobacterium tuberculosis (*M. tuberculosis)* causes the lung infection tuberculosis (TB) that can often be fatal if left untreated. Tuberculosis is very widespread globally with an estimated 9 million new cases and 1.5 million deaths in 2013[3], and 280 deaths in the UK[4]. *M. tuberculosis* can be treated with antibiotics, but there are problems with antibiotic resistance with the emergence of mutidrug-resistant (MDR) and extensively drug resistant (XDR) TB strains.

M. tuberculosis genome is around 4 million base pairs in length and contains nearly 4,000 protein coding genes[5]. This is a relatively large genome for a pathogen, and the lifecycle is also relatively slow, with a generation time of approximately one day for each complete replication cycle. The mutation rate varies between different *M. tuberculosis* lineages, but is orders of magnitude lower than the rate for RNA viruses. An improved understanding of functional differences between the genome sequences of different *M. tuberculosis* strains should help improve the targeted use of antibiotics.

2.3.2 Methicillin-resistant *Staphylococcus aureus*

Methicillin-resistant *Staphylococcus aureus* (MRSA) is responsible for nearly 300 deaths in England and Wales in 2012[6] and is also prevalent globally. MRSA is, by definition, resistant to at least some antibiotics and thus can be difficult to treat. MRSA is problematic in a hospital setting as patients with already weakened immune systems are particularly vulnerable to infection and the bug can spread across hospitals.

The MRSA genome consists of a circular chromosome almost 3 million base pairs in length, accompanied by a smaller plasmid genome, and contains approximately 2,500 protein coding genes. A number of different putative biomarkers of antibiotic resistance and virulence have been identified within the genome. Different strains of MRSA have genomes that differ by approximately 6% from each other[7], so the genome sequences are relatively homogeneous compared to viruses. Nonetheless, MRSA genome sequences diverge rapidly enough to provide potentially useful information for infection control measures[8].

2.3.3 *Clostridium difficile*

Clostridium difficile (*C. difficile)* is a spore-forming bacterium that causes a form of diarrhoea associated with around 1,500 deaths per year in England and Wales[9] and is also found across the world. *C. difficile* infections can be caused by antibiotics that interfere with the balance of bacteria in the gut and thus the intestinal microbiome.

The C. difficile genome consists of a circular chromosome of around 4 million base pairs and a circular plasmid of nearly 8,000 base pairs. The genome contains over 3,500 protein coding genes and a relatively large proportion for a bacterium (11% of the genome) of mobile genetic elements, some of which are associated with the pathogen's antibiotic resistance susceptibility and virulence[10]. *C. difficile* is both multidrug-resistant and highly contagious; systematic genome sequencing of the pathogen could improve infection control. For example a longitudinal study using whole genome sequencing and analysis to better define the epidemiology of *C. difficile* infection, found that patients with symptomatic *C. difficile* infection are not the only source of transmission[11].

2.4 The virus genome – a compact and diverse set of genetic instructions

2.4.1 Composition of viral genomes

Viral genomes vary considerably in size from approximately 2,000 to 1,200,000 base pairs (bps)[12], but they are generally near the smaller end of that range. The genomes are typically smaller than bacterial genomes, and contain often only a handful of protein coding genes compared to the thousands found in most bacterial genomes.

2.4.2 Classes of viral genomes

Viral genomes are highly diverse, not only in their underlying genetic sequence but also the basic mechanisms by which they evolve and replicate. This diversity coupled with their rapid evolution can make pathogenic viruses difficult for the host's immune system and antiviral drugs to target effectively. Genomic technologies provide a high resolution picture of these complex genomes and thus can help detect and trace the spread of viruses and antiviral resistance. This epidemiological information can improve containment measures during outbreaks and facilitate more directed use of antiviral drugs (chapter 6 and chapter 7). In the latter case, sequencing can be used pre-treatment to target therapies appropriately and post-treatment to detect drug resistance mutations if the initial treatment fails.

While eukaryotes (a large group of organisms that includes all animals, plants and fungi) have linear DNA genomes, virus genomes are much more variable., even more so than bacterial genomes. Five main different structural features of viral genomes are as follows:

1. Virus genomes can be linear sequences or circular loops

2. The genomes may consist of either double-stranded or single-stranded sequences

3. They can contain either DNA, or more commonly RNA nucleic acids

4. The genomes can be positive-sense or negative-sense depending on the direction in which the sequence is 'read' by the enzymatic cellular machinery

5. Some viruses (retroviruses) have genomes that are reverse transcribed, where their genomes are converted from RNA to DNA when they are incorporated into the host genome

Based on these five criteria, seven different classes of virus genome have been identified[12]. Each of these different types of viruses is associated with important human diseases and illnesses (table 2.2). Pathogenic retroviruses can be particularly difficult for our immune system to deal with because they 'hide' within the host cell's own genome, and RNA viruses can be problematic as they evolve particularly rapidly.

Table 2.2 Examples of viruses in different virus classes and associated medical illnesses

Virus class	Class symbol	Virus family example	Disease example
Double stranded DNA	dsDNA	Adenoviruses	Respiratory infections
Single-stranded DNA	ssDNA	Parvoviruses	Fifth disease
Double-stranded RNA	dsRNA	Reoviruses	Diarrhoea
Single-stranded positive-sense RNA	+ssRNA	Flaviviridae	Hepatitis C
Single-stranded negative-sense RNA	-ssRNA	Orthomyxoviruses	Influenza
Single-stranded RNA reverse transcribing	ssRNA retro	Orthoretroviruses	AIDS
Double-stranded DNA reverse transcribing	dsRNA retro	Hepadnaviruses	Hepatitis B

2.5 Genomics of important viruses for human health in the UK

There are a wide variety of important human diseases and illnesses caused by viruses (table 2.2). The main viral disease burden in the UK that genomic technologies have the potential to help alleviate, and thus the primary focus of this report, arises from infection with either human immunodeficiency virus (HIV) or hepatitis C virus (HCV).

HIV infection can lead to acquired immunodeficiency syndrome (AIDS), a slow progressive disease that is one of the most significant global health burdens, resulting in over a million deaths worldwide every year[13], and around 500 deaths annually in the UK[14]. As of 2012, it was estimated that almost 100,000 people in the UK are infected with HIV, approximately 22% of whom are unaware of their infection[15].

2.5.1 HIV

HIV is a retrovirus with each viron (individual virus particle) consisting of a viral envelope and a protein shell that contains an RNA genome. The RNA genome of HIV is almost 10,000 nucleotides in length and although different HIV strains are highly variable they share several common genomic features. HIV genomes consist of three important structural proteins encoded by the *gag, pol*, and *env* genes. Additionally, HIV genomes contain several other elements that produce proteins including the essential Tat and Rev proteins, and a variety of other accessory regulatory proteins. The replication of the virus requires a reverse transcriptase enzyme to convert the RNA genome into a DNA copy that is 'pasted' into the human host cell's genome. The DNA is later transcribed by the host cell to form an RNA copy of the virus genome. There are two types of HIV, named HIV-1 and HIV-2, which are genetically divergent from each other. HIV-1 is much more prevalent, virulent and infectious compared to HIV-2, and thus of more importance to public health.

2.5.2 HCV

Infection with HCV can cause hepatitis C, a disease that affects the liver. At least 150 million people are currently affected by hepatitis C, which equates to approximately 3% of the global population and 0.4% of the UK population [16,17]. An estimated 350,000-500,000 people die annually worldwide from Hepatitis C related liver diseases[16].

The HCV has an RNA genome of similar size to the HIV genome at 9,600 nucleotides long, which consists of a large 'gene' encoding a polyprotein between two small regions that contain regulatory elements. HCV replicates while bound to host cells, normally hepatocytes in the liver. HCV takes control of some of the host cell's replication machinery to translate the polyprotein part of the genome into a single protein of approximately 3,000 amino acids. The polyprotein is then processed by enzymes into at least 11 distinct structural and nonstructural protein components.

2.5.3 Influenza A

The risk of a pandemic flu outbreak due to a mutated strain of influenza A has been rated as the largest emergency threat to the UK[18]. The most recent flu pandemic was in 2009, the so called 'swine flu' pandemic, which led to the deaths of an estimated 284,000 people globally and over a thousand within the UK[19]. The outbreak known as 'bird flu' caused an estimated 608 cases and 359 deaths across multiple Asian and African countries between 2003 and 2012[20].

The influenza A genome is approximately 13,500 bp in length and consists of eight RNA segments that produce 11 different proteins. Two important components are segment 4 that encodes the hemagglutinin (H) protein and segment 6 that encodes the neuraminidase (N) protein, which are used to name the different influenza strains. Reassortments of these eight segments between different viral strains is the major way by which the virus is able to evolve so rapidly and create novel, potentially dangerous, strains.

2.6 Mutation, evolution and diversity – the pathogen genome as a moving target

2.6.1 Mutation

Mutation can be defined as any change in the structure or sequence of the genome of an organism. The effect of any genomic mutation on the functions and behaviour of an organism depends on factors including the location, size and type of the mutation. In some cases mutation disrupts the sequence of a gene and prevents its correct expression leading to a 'loss of function'. In others mutations cause 'gain of function' either by the introduction of new genes into the genome or by changing the behaviour of the protein products of existing genes. Perhaps the most important instances of genomic mutation in pathogens, in public health terms, are those that result in acquisition of resistance to antibiotics or the ability to evade our immune system.

One of the main reasons that pathogens can be effective and persistent disease causative agents is that their genomes tend to be small and evolve very rapidly as their genetic sequence changes over time.

Mutations normally accumulate vertically over the generations at the point of genome replication. Mutations most commonly take the form of individual nucleotide changes (point mutations), but larger-scale changes where a number of bases are added (inserted), removed (deleted) or 'copied and pasted' (transposed) also occur. The rates of point mutations per year are typically high compared to that for human genomes as pathogens are able to reproduce rapidly and thus have a very short generation time.

The principal mechanisms through which mutations can occur are:

- **Errors in replication** – where the wrong base(s) get incorporated during the process of genome replication prior to cell division
- **Chemical or radiological mutagenesis** – where an external chemical or radiological agent damages the DNA and causes changes to the genome sequence
- **Exchange of genetic material through** transformation **or** transduction – where exchange of DNA fragments through recombination (or for bacterial genomes plasmids, transposons or bacteriophage DNA) alters the composition of the genome

Whilst the chances of function altering mutation occurring in any individual pathogen cell are extremely low, pathogens colonise humans in their billions, and divide / replicate extremely rapidly (often in less than an hour) **–** during an infection or colonisation there are always a proportion of cells present that undergo mutation. The clinical significance of these mutant cells, at both the individual and population level, depends on:

- Whether they confer a selective advantage *e.g.* resistance to antibiotics, or evasion of immune surveillance
- Whether the population is subjected to appropriate selective pressure *e.g.* antibiotic treatment or immune activation
- The rate at which they are able to divide and outcompete the other cells that do not carry the advantageous mutation

One of the main reasons that pathogens can be effective and persistent disease causative agents is that their genomes tend to be small and evolve very rapidly as their genetic sequence changes over time.

Point mutation rates in viral genomes are higher than those in bacterial genomes, but mutation rates are still variable across the different types of virus RNA viruses tend to have higher mutation rates than DNA viruses because RNA viruses lack DNA polymerases (proofreading enzymes), and therefore are more 'error', and thus mutation prone. The rapid evolution of viruses can lead to new emerging threats, not only due to novel antiviral resistance genes developing but also genetic changes that can enable zoonotic transmission from other animal species across to humans[21] (chapter 8).

In addition to the process by which viruses gradually accumulate mutations over time which are are passed vertically between generations during replication, viral evolution can occasionally proceed through horizontal changes, where dramatic change in genome composition can occur in a single step. This occurs when segments of DNA are transferred between different viruses, or more drastically when different strains of virus combine together to form a completely new virus. The latter in particular leads to large scale changes that can facilitate the transmission of viruses between different host species. This can be medically very important, for example, the combining of different influenza strains has been implicated in the emergence of several different human flu epidemics that originated in animals (chapter 8).

A graph showing the average rates of spontaneous mutation in viruses, can be found at: www.nature.com[22].

2.6.2 Evolution and diversity

The rapid acquisition of mutations that confer selective advantage in different environments allows bacteria to adapt to survive under a wide range of conditions. This adaptive evolution, and the consequent diversity within and between pathogen species, has significant adverse consequences for human health. It enables them to evade our natural immunity, the immune protection afforded by vaccinations and our limited repertoire of antibiotics. This can lead to an ongoing 'evolutionary arms race' where clinicians attempt to find new drugs or vaccines that inhibit infection by bacteria and viruses which are simultaneously under selection to evolve resistance to those same vaccines and drugs. This is a particular challenge for rapidly evolving pathogens, such as RNA viruses *e.g.* influenza, where new vaccine formulations are required annually to keep pace with its evolution.

2.6.3 Drugs and vaccines: drivers of pathogen evolution

It is important to note that the rate at which pathogen genomes evolve depends both on the intrinsic rates of the processes that affect the genome sequence itself *i.e.* mutation, plasmid-mediated horizontal gene transfer and recombination, and also the extent to which these genomic changes confers a relative 'fitness' benefit to the individual bacteria that have acquired them. Thus whilst an individual bacterium might easily acquire a plasmid carrying antibiotic resistance genes, or the ability to evade immune recognition, this offers no selective survival advantage to the bacterium over other cells not carrying these genes, unless they are exposed to the the relevant antibiotic or a suitably primed immune response. Similarly, it is the continuous exposure of HIV to antiretroviral treatment that drives the selection of mutant copies of the HIV virus present within an individual, favouring those that are able to withstand the effects of the drugs. The same mutations may occur in untreated patients, but in the absence of a drug that biases towards their survival they are less likely to dominate their population.

Such examples demonstrate clearly the extent to which human interventions can drive the evolution of pathogen genomes. The utility of genome sequencing for determining the extent to which both drugs and vaccines are changing the identity and behaviour of pathogens that infect us will be explored further in chapter 7.

The different rates at which bacterial genomes evolve has a significant impact on our understanding and management of bacterial infections. Species with slow evolving genomes, whether due to low replication rates or low intrinsic susceptibilities to undergo mutation or gene transfer, are slower to acquire new functions, such as resistance or immune evasion properties, over time and are thus potentially easier to manage and treat. Species subject to rapid genomic evolution may, on the other hand, more rapidly develop treatment resistance, and are more likely to adapt even when treatment strategies change.

While the small size and generic molecular constituents of pathogens mean they can be sequenced quickly and cheaply, their diversity and complexity means that their analysis and interpretation is likely to be challenging.

2.6.4 Capitalising on the evolution of pathogen genomes: a source of epidemiological information

Bacterial genome evolution occurs on a timescale that can provide vital clues to epidemiologists wishing to detect, monitor and intervene to curtail infectious disease outbreaks. As the rate of change of the genome sequence is relatively stable for many, but not all, bacterial species, the extent of the differences between the genomes of two isolates of the same species can be used to calculate how closely or distantly related they are. Using genetic differences to determine 'family trees' of bacterial isolates is known as phylogenetics. chapter 7 will explore how this method could be used in clinical and public health settings to identify and characterise transmission events during infectious disease outbreaks.

2.7 Conclusions

Pathogen genomes are as diverse as the organisms for which they provide the operational blueprints. While their small size and generic molecular constituents mean they can be sequenced quickly and cheaply, their diversity and complexity means that their analysis and interpretation is likely to be challenging. Thus, whilst it should be possible to apply adaptations of a single genome sequencing method to 'read' the genomes of a diverse range of bacterial and viral pathogens, the analysis of the meaning and clinical significance of what is read in each case will most likely require distinct solutions for each different pathogen and each different use to which genomic information can be put in managing infectious disease.

In the next two chapters we summarise existing methods for pathogen genome sequencing and analysis and consider how ready they are to be adapted for use by clinicians and public health practitioners in infectious disease management.

3 An introduction to pathogen genome sequencing

In this section we outline the different technologies currently in use for the sequencing of pathogen genomes, the practical steps involved in using them to obtain raw genome sequence data and the advantages and disadvantages of their use in different sequencing scenarios relevant to public health and clinical microbiology.

3.1 Introduction

The first complete genome sequence of a bacterium, *Haemophilus influenzae*, was published in 1995 and since then a large number of microbial and viral genome sequences, including many that cause disease in humans, have been determined[†]. During this time, the technology platforms used for the sequencing of genomes have undergone a radical transformation from large, expensive, slow and relatively inaccessible machines that provided extremely accurate sequence information to smaller, cheaper and faster machines that are more prone to errors and whose outputs require a significant degree of computational analysis prior to interpretation.

3.2 From Sanger sequencing to next generation sequencing

Genome sequencing has been possible since the 1970s when the Sanger method was developed and first applied to sequence the genome of a virus[23]. This technique has undergone various modifications and developments in the intervening years to facilitate automation and increase throughput. Nevertheless, the Sanger technique is now considered too laborious and expensive for routinely sequencing whole genomes. More recently, a number of sequencing technologies have been developed (or are in development) that are radically reducing both the cost and time required for sequencing[24,25]. These are collectively described as next generation sequencing (NGS) technologies. While they have been developed with the primary aim of increasing the speed and reducing the cost with which whole genomes can be sequenced they can also be employed for the analysis of specific genes and other genetic elements including RNA.

[†] GOLD: Genomes Online Database, is a resource for pathogen sequencing projects genomesonline.org/cgi-bin/GOLD/index.cgi

NGS platforms allow many millions of target DNA molecules to be sequenced in parallel, resulting in substantial reductions in cost and in the time taken to produce a whole genome sequence. Whilst these advances are significant, there is one important disadvantage of NGS platforms compared to traditional Sanger sequencing: they generate single uninterrupted sequence reads that are only one tenth of the length of their predecessor technology. The data produced by these platforms therefore requires software capable of reconstructing whole genome sequences from far greater numbers of smaller fragments of DNA sequence than is the case with Sanger sequencing. A description of Sanger sequencing can be found in our report (*Next Steps in the Sequence*) but is not included here, as this technology is not capable of delivering pathogen whole genome sequencing (WGS) for clinical and public health use. Strategies for computational analysis of data produced by NGS platforms are discussed in the next chapter.

3.3 Reversible termination sequencing by synthesis

This method is closely based on the original Sanger sequencing-by-synthesis method, but uses special fluorescently labelled terminator nucleotides in which the chemical modification can be removed, rendering the chain termination process reversible[26]. It is currently the predominant NGS method used in both research and clinical settings for genome sequencing of microbes and other organisms.

Following DNA fragmentation, specific sequences (adapters) are added to their ends; this process is referred to as adapter ligation. Using this template, DNA molecules are immobilised onto a glass surface at high density, upon which both amplification and sequencing take place. The tethered fragments are subjected to clonal bridge amplification using surface polymerase chain reaction (PCR) (amplification of DNA)to create dense clusters of identical DNA templates across the plate. The sequencing reaction then begins with the addition of a universal primer, DNA polymerase and four reversible nucleotide terminators labelled with different coloured dyes. Incorporation of a complementary nucleotide into the first position results in termination of polymerisation. At this point, unincorporated nucleotides are washed off and the first base on the template strand is identified by colour imaging. The dye and the terminating group are then cleaved chemically, and the process is repeated, allowing further extension of the DNA fragment. Repetition of this cycle allows identification of specific bases along a template DNA strand as they are incorporated, which can be built into a sequence read.

This chemistry has been commercialised by Illumina® through a number of systems aimed at different applications. The HiSeq systems have been developed for sequencing centres carrying out high throughput sequencing studies. Such systems may be particularly useful in microbiology laboratories where high throughput whole genome sequencing is being undertaken for the purposes of national or regional epidemiological surveillance studies or outbreak detection programmes. The physically smaller 'benchtop' systems available, such as the MiSeq and NextSeq500, deliver significantly lower throughput but are significantly cheaper to buy and so may be more attractive where resources are limited.

Given the sensitivity of microbiology investigations to turnaround time, sequencing platforms with the shortest run time to generate a whole genome sequence are likely to be favoured for use in a clinical setting.

Given the importance of short turnaround times in microbiological investigations, sequencing platforms with the shortest run time to generate a whole genome sequence are also likely to be favoured for use in a clinical setting. While benchtop instruments were designed to deliver shorter run times than their larger, higher throughput counterparts, recent developments in the performance of HiSeq instruments means that the length of a single sequencing run on these instruments is effectively the same as that of its benchtop comparator the MiSeq. Nevertheless, it has been proposed that these benchtop sequencers are better adapted to the needs of diagnostic microbiology laboratories as they require lower throughput to maximise their utilisation. They may, therefore, be more suited to 'random access' use where time sensitive diagnostic applications are prioritised over use for longitudinal, but less time sensitive, surveillance or outbreak investigations. Where the latter is prioritised *e.g.* in specialist public health laboratories, higher throughput is more likely to be required and batching of samples to maximise utilisation is more feasible, favouring use of larger HiSeq type instruments.

Notably, these platforms have been developed for research purposes, and are not compliant with existing *in vitro* diagnostic device regulations. An exception to this is the MiSeqDx, a MiSeq manufactured to meet clinical diagnostic standards, which is the first NGS platform to attain FDA approval as an *in vitro* diagnostic device.

An illustration of the sequencing chemistry employed by Illumina platforms is available at: www.nature.com[27].

3.4 Semi-conductor sequencing

Both Sanger and reversible termination sequencing rely on the direct imaging of the addition of fluorescent nucleotides to determine the order of bases in each DNA fragment they sequence. This imaging step requires the use of modified nucleotides and sophisticated optics that add to the cost and complexity of the sequencing process. This potential limitation has driven the development of alternative non optical methods to detect the addition of nucleotides during sequencing-by-synthesis reactions. The most successful of these methods has been semi-conductor sequencing, which takes advantage of existing low cost mass production semi-conductor technology to detect addition of nucleotides during a sequencing-by-synthesis reaction[28, 29].

Two semi-conductor sequencing platforms are currently available, the Ion Proton and Ion Personal Genome Analyser (Ion PGM), both developed by Life Technologies. Unlike other platforms, semi-conductor sequencing is based on monitoring the release of hydrogen ions (H+), which are another by-product of DNA synthesis. DNA templates are held in specialised wells which are designed as ion sensors, and nucleotides are added sequentially to each well. If a particular nucleotide is incorporated into a growing strand by DNA polymerase the result will be a release of H+ into solution and a concomitant change in acidity (pH). The change in pH is detected as a voltage shift by sensors and can be related to the number of molecules of a particular base incorporated. Currently this system does not detect single molecules and amplification is required prior to sequencing, but the synthesis reaction is detected in real-time.

These machines have not been widely used in clinical and public health microbiology settings, however, a number of proof-of-principle studies demonstrating the utility of the Ion Torrent PGM, a small benchtop instrument, for pathogen sequencing have been published[30,31,32]. The high error rate (1.71% compared to 0.80% for Illumina MiSeq) and lower throughput of this platform mean that is has less utility in high throughput sequencing centres but the fast turnaround time, reduced cost and ability to process smaller numbers of samples, compared to for example the MiSeq, are considered advantages in a diagnostic setting (table 3.1). Moreover in the context of investigating an outbreak in a hospital setting, a comparison of sequencing platforms found both MiSeq and Ion Torrent produced similar clinically actionable data despite their different read metrics and error profiles[33].

An illustration of semi-conductor sequencing as used in Ion Torrent machines is available at: www.genomics.cn[34].

3.5 Single molecule sequencing

One drawback of the sequencing-by-synthesis technologies described above is that they rely on the fragmentation of DNA into small lengths suitable for sequencing, and clonal amplification of these fragments to ensure that there is sufficient DNA present to detect the addition of a new nucleotide. Significant computational costs and complexities arise from the need to reconstitute the original intact genome from the millions of sequenced short fragments, and the process of amplification may introduce systematic errors or artefacts in the sequencing process. These limitations have driven the development of sequencing platforms that directly read the sequence of single DNA molecules. This approach avoids the need for amplification, and the error associated with this, and also enables far longer DNA molecules to be sequenced.

Some single molecule methods are based on the principle of sequencing-by-synthesis, though there are also several novel methodologies, such as monitoring the passage of DNA through nanopores. The only single molecule sequencing system currently available is the PacBio RS II system developed by Pacific Biosciences. This platform is based on the standard sequencing-by-synthesis method that monitors the addition of four different dye-labelled nucleotides in specialised wells containing an immobilised DNA polymerase[35]. The template DNA is added to the well and a single DNA molecule forms a complex with a single polymerase enzyme. The identity of the nucleotides incorporated into the growing strand is monitored by laser excitation of the fluorophore on the nucleotide and detection in the well by a sensitive CCD camera.

The difference between this method and others that use fluorophores is that the dye is attached to the phosphate of the nucleotide rather than the base itself. Thus it is cleaved and released as a natural part of DNA synthesis, resulting in release of the dye without interruption to the sequencing process. The sequencing is therefore both single molecule and real-time. Template DNA is prepared using a proprietary kit, which attaches special hairpin loop adapters to the end of double stranded DNA molecules. The result is circular template DNA which is sequenced without any amplification steps. A particular limitation of this approach is that it requires a larger quantity of highly purified DNA to produce good quality results, and these are not always available when undertaking analysis from clinical samples.

The use of these instruments in genome sequencing projects is relatively rare compared to other NGS platforms, mainly due to their prohibitive cost and low throughput (yield per run is 100Mb). Relatively low levels of parallelisation are currently possible with a PacBio RSII, and so throughput is <1% of that achieved by benchtop platforms such as the MiSeq and Ion Torrent PGM. Whilst this makes this instrument unsuitable for use in diagnostic settings, its ability to read DNA molecules up to 100 times longer than those achievable on a benchtop sequencer may have some advantages in a reference microbiology context. Specifically, by enabling assembly of accurate and complete whole bacterial and viral reference genomes against which sequences from clinical samples can be compared. Indeed this is how the PacBio RS II is currently used in microbial genomics research, predominantly for *de novo* genome assembly and for filling in the gaps missing in draft genome assemblies as these projects require relatively low throughput but benefit from accurate long sequence reads.

Examples of these projects include the 100K Foodborne Pathogen genome project in the USA and a collaboration between the Wellcome Trust Sanger Centre and the Public Health England reference microbiology service to sequence the complete genomes of their reference collection of bacterial strains. It has also been used to determine the genomic architecture of the resistance genes underlying the spread of antibiotic resistant infections in a hospital in the USA[2].

An illustration of single molecule sequencing on PacBio platform is available at: www.nature.com[36].

3.6 Future technologies

The majority of platforms under development for single molecule DNA sequencing using electronic detection are not based on the sequencing-by-synthesis method, but on an entirely new method using either biological or solid state nanopores. These technologies monitor changes in electrical current following the passage of DNA strands or individual bases through a nanopore. The nanopores themselves can either be simply small holes in an inorganic membrane (solid-state nanopores), such as silicon nitride[37] or graphene[38], or specific channels made from modified natural pore-forming proteins[39] embedded in a lipid bilayer or synthetic membrane. Nanopore sequencing technologies are based on one of two approaches – either strand sequencing as the DNA strand itself passes through the nanopore, or base sequencing as DNA bases are cut off from the end of the strand then individually and sequentially fed into the nanopore. A voltage is placed across the membrane that drives the translocation of negatively charged DNA molecules across the pore as the nanopore is the only point at which current can flow across the insulating membrane. As DNA bases pass through the pore, each base blocks the current by a different amount that can be monitored and related back to the nucleotide composition of the strand of DNA.

Numerous companies are developing nanopore-based DNA sequencing platforms, including Oxford Nanopore Technologies, NABSys and base4innovation. Although this technique is extremely promising, there are still some challenges to be overcome prior to the launch of a platform, such as regulating the speed at which bases travel though the nanopore so they can be accurately detected and increased parallelisation to enable a large number of molecules to be sequenced simultaneously[40, 41]. Perhaps the most advanced to date are the platforms under development by Oxford Nanopore Technologies, which is developing the MinION, PromethION and the GridION systems. These are different scales of device operating on the same underlying nanopore technology. They both consist of chips containing a stable lipid bilayer into which nanopores are inserted. Having multiple nanopores per chip, each processing a single molecule at a time, parallelises the method.

The MinION, which is currently being trialled through an early access user programme open to research, public health and clinical laboratories, is a USB drive sized sequencer that plugs directly into to a laptop and can be used in low throughput sequencing. Recently the first reports describing results using the MinION platform have emerged[42, 43, 44, 45]. Notably they have predominantly involved the description of its use for sequencing bacterial and viral genomes. Such a portable, low throughput sequencing device may, in principle, be useful for microbiological investigations, particularly in the field or in community settings where rapid near-patient analysis of samples is required. This would enable more immediate and effective diagnostic and treatment management decisions to be taken. It must, however, be noted that at the time of writing this report this technology is still in the earliest stages of evaluation, requires a level of sample preprocessing that necessitates access to specialist laboratory equipment, and significant computational and 'wet laboratory' expertise to operate it successfully.

An illustration of nanopore sequencing is available at: www2.technologyreview.com[46].

3.7 Performance metrics

Performance of different sequencing platforms is an important factor in the decision on which might be best suited for use in a particular clinical or public health setting. Measuring performance is, however, a challenging task, as these technologies are developing at a rate that renders most performance comparisons almost immediately obsolete, and examples of unbiased and independent comparisons are few and far between[33,47, 48]. Nevertheless, it is important to understand the factors that affect performance and influence the utility of different sequencing technologies. It is also vital to bear in mind that decisions on which machines are optimal for different applications are not only based on the performance metrics outlined below, but also on practical considerations such as sample preparation, ease of use, turnaround time from sample to result, and cost. In addition to errors added during the sequencing process, it is important to note that there other potential sources of error such contamination of the sample upstream, and bioinformatic errors downstream of the sequencing (chapter 4).

3.7.1 Analytical accuracy, systematic errors and quality of base calls

In addition to amplification errors (which can be eliminated in principle by amplification-free single molecule sequencing), all sequencing methodologies suffer from both random and systematic errors. The raw accuracy of the sequencing process is a critically important factor when applied to medical diagnostics. To estimate the likelihood of a correct DNA base call quality scores are usually assigned to each base by software integral to the sequencing platform. Different sequencing technologies are prone to different systematic errors and use different quality scoring algorithms, both of which can influence their utility for different applications and the cross compatibility of data generated by different platforms[33,47, 48].

3.7.2 Read depth, genome coverage and uniformity

Genomic fragments are sequenced multiple times inside a sequencer to provide overlapping sequence 'reads'. A greater number of reads at a given point (read depth) and across the entire genome (average genome coverage) provides greater statistical confidence that a given DNA base in a sequence has been inferred correctly, and can therefore reduce the time and substantial cost needed to perform secondary validations in downstream variant identification. The required read depth varies depending upon the specific application and level of certainty required for the result. The maximum average read depth achievable for a pathogen genome on a particular sequencing machine is limited by the total amount of sequence the machine can generate (table 3.1) per run, but also depends on the size and number of genomes being sequenced.

Coverage across the genome can also be non-uniform, either because some regions are difficult to amplify as they have either very high or very low levels of the GC base pair, or because they are highly repetitive in sequence and so determining their location in genome assembly unambiguously is challenging. Different pathogen genomes vary widely in their GC base pair content, with both extreme high and low GC content genomes being particularly difficult to sequence evenly *e.g. Plasmodium falciparum*. Whilst increased read depth can partially overcome low coverage problems, and different enzymes can be used that are better adapted to low / high GC content genomes, some regions of genomes remain refractory to sequencing by synthesis methods.

3.7.3 Read length (number of bases per read)

Read length is an important factor in certain applications, such as sequencing through repetitive regions or facilitating *de novo* assembly (chapter 4). In addition, it makes alignment to the reference sequence a substantially easier task by reducing the number of possible sequence matches throughout the genome. Thus, depending on the intended application an instrument with a longer read length may be more desirable than one with higher throughput but shorter read lengths. Current NGS platforms offer up to 300 bases per read, with the Pac-Bio RS offering up to 20Kb (Table 3.1). It is anticipated that the single molecule platforms will have substantially longer read lengths. Scientists working with the Oxford Nanopore MinION have informally reported reads between 50-100Kb in length, but these have yet to be published in peer-reviewed journals.

3.7.4 Throughput, capacity and run time

The number of bases of DNA that can be sequenced per run, the number of different samples that can be sequenced simultaneously, and the length of the run itself all have a major impact on the suitability of different sequencing platforms for different laboratories. These factors vary substantially between machines and applications, ranging from 1- 4 days per run with a throughput of 1-25 Gb per day, depending upon read length and sequencing protocol.

Sample multiplexing is also crucial for cost-effective use of NGS, as although NGS has substantially reduced the per-base cost of sequencing, this saving is only realised if the capacity of the instrument is used effectively. Sequencing a single bacterial or viral genome will only require a small fraction of the capacity of even a relatively low throughput benchtop sequencer; consequently, methods for analysing multiple samples in a single run are important. Although DNA sequencing machines are unable to differentiate between matching target DNA isolated from different samples, there are a number of methods that allow multiple samples to be sequenced simultaneously. In addition, many sequencing platforms allow physical separation between samples, by having multiple separate channels. Finally, DNA 'barcode tags' have been developed, these are added to the ends of DNA fragments during library preparation, and provide a unique DNA signature to mark and track individual samples.

Table 3.1 Summary and specifications of currently available NGS platforms (data from company websites, December 2013

Platform	Chemistry	Capability	Uses
High output machines			
HiSeq 1500/2500 (Illumina)	Reversible termination Sequencing by synthesis	Read length: 2 x 250 bp Run time: 2-11 days, <2 days in rapid run mode Output (Max.) 600Gb	Illumina platforms are currently the instruments of choice for the vast majority of clinical sequencing applications. Workhorse machine for large research labs and core facilities doing WGS where turnaround time is not a major issue but flexibility of application and high throughput are more important. Also offers a rapid run mode for lower throughput that can perform exome sequencing of 100x coverage or whole genome sequencing at 30x (of a human genome) in 27 hours
Pac-Bio RS (Pacific Biosciences)	Real-time single molecule Sequencing by synthesis	Read length: up to 40kb averaging at 14Kb (with good starting library) Run time: 2 hours typically, minimum 30min for some applications. Output (max.) 250Mb	These instruments are relatively rare compared to other platforms. Published literature indicates they are used mostly for *de novo* genome assembly for microbes, and for 'finishing' draft genome assemblies as these projects require relatively low throughput. It is being employed in a research setting at UC Davis on the 100k Foodborne Pathogen genome project. It is also being used by the Wellcome Trust Sanger Institute and PHE Colindale to sequence complete genomes of their reference collection of bacterial strains
Ion Proton (Life Technologies)	Single molecule Proton detection	Read length: 200 bases (averaging at 200 bases) Run time: 2-4 hours Output (max.) 10 Gb	Proton instruments are being installed in research centres offering core genomics facilities or undertaking in-house large-scale WGS projects. Still used by only small minority of researchers / clinicians

Table 3.1 cont. Summary and specifications of currently available NGS platforms (data from company websites, December 2013)

Platform	Chemistry	Capability	Uses
Benchtop instruments			
Ion Personal Genome Machine (Life Technologies)	Single molecule Proton detection	Read length: 35-400 bases (averaging at 200 bases) Run time: 2-8 hours Output (max.) 2 Gb	Ion PGM instruments are being used for targeted resequencing, with potential diagnostic applications in cancer, and also for pathogen genome sequencing. These instruments are beginning to be used in clinical genetics and public health settings
MiSeq (Illumina)	Sequencing by synthesis	for longest run and 1 x 36bp for shortest Run time: 4-55 hours Output (max.) 15 Gb	MiSeq instruments are being used for targeted resequencing, with potential diagnostic applications in cancer and also for pathogen genome sequencing. These instruments are beginning to be used in clinical genetics and public health settings. The MiSeqDx is the first next-generation sequencer approved for clinical use by the US FDA (in 2013)
Minion and GridIon (Oxford Nanopore)	Nanopore base single molecule sequencing	From their website *"Oxford Nanpore does not provide traditional / comparative throughput or accuracy specifications for DNA sequencing".* However, their read lengths can be many Kb	MinION is the smaller portable USB device, PromethION is a high throughput tablet sized instrument, and GridION is the larger instrument. The technologies are not being used in a clinical setting as the per base error rate is too high currently, but the technology is improving rapidly as it is being widely adopted by researchers

3.7.5 Reagent and instrument cost

Although reagent cost for sequencing has plummeted over the last decade to less than $0.50/Mb for reagents on the newest NGS platforms this is not the case for the sequencing machines themselves. These are often fairly expensive (ranging from US$100,000-1 million each), and in addition, upstream and downstream costs of sample preparation and data analysis may add significantly to the budget required to run a sequencing service.

3.8 The sequencing pipeline

The sequencing of pathogen genomes is a process that is relatively generic across different species of pathogen, using the same technology platforms and producing data in similar forms. By contrast, the upstream sample preparation required to produce 'sequencer-ready DNA' and the downstream analytical processes required to turn raw sequence data into interpretable genomic information can vary significantly between pathogens. Collectively, sample preparation, sequencing and data analysis can be described as a sequencing pipeline. The basic steps in this pipeline are illustrated in figure 3.1 and the upstream sample preparation processes are described in more detail in the sections below. The analysis of data from pathogen genome sequencing is discussed more fully in the next chapter of this report.

3.9 Clinical sample to purified DNA

Although the exact details of the upstream stages of sample preparation vary with different pathogens, they involve broadly similar steps. Initially, the specific pathogen of interest must be isolated from a clinical sample (*e.g.* blood, faeces, body fluids) that may contain many different microbes. The methods for achieving this differ between bacteriology and virology.

3.9.1 Bacteria

Bacteria are isolated from clinical samples through culture on bacterial growth promoting media. Selective media that promotes growth of particular organisms are available for a number of bacteria of clinical importance (*e.g.* GN Broth is used for enrichment of *Salmonella* and *Shigella* species, whereas *Enterococcus* agar is use for selective culture of *Enterococci,* and MRSA selective media are also commonly used). Hence where there is strong *a priori* knowledge of the pathogen or the media readily available, this step can be straight forward. However, it can be a time consuming process when the pathogen is novel or has special growth requirements. In such cases, it requires appropriate selection of media (which may have to be determined empirically) in order to ensure growth of the bacteria suspected to be causing an infection, while excluding the growth of non-pathogenic commensal species that may also be present in the sample. While selective culture is effective for many pathogens, there are many that cannot be isolated using this procedure[49].

Following the initial isolation of a bacterial pathogen on selective media, a second culture step may be required to amplify the quantity of bacterial cells available to ensure sufficient genomic DNA is available for whole genome sequencing. For some bacterial species this second culture step can be circumvented by exploiting sequencing sample preparation methods (*e.g.* Nextera XT developed by Illumina) that enable smaller quantities of DNA extracted from a single bacterial colony from the primary culture plate to be used to prepare sequencing libraries[50]. However, for many pathogens with long incubation periods and those that are difficult to culture in the laboratory, obtaining sufficient DNA for genome sequencing remains a challenge. The development of culture-free methods for obtaining genome sequences in such cases are discussed in chapter 6.

Once sufficient quantities of the pathogen are available (which may require as little as 24 hours or as much as two weeks in culture) genomic DNA can be extracted and purified. Many commercial kits are available for purification of genomic DNA from bacteria. Pathogen genomes are still too large to be sequenced or analysed intact by most NGS platforms, and so genomic DNA must be broken into smaller fragments prior to sequencing. There are two principal methods of DNA fragmentation: mechanical force (including nebulisation and ultrasound) or enzyme digestion. Manufacturers of second generation sequencing technologies may recommend different methods of fragmentation, as different fragmentation methods influence the fragment size distribution; this in turn has consequences for the amount of starting material required which vary between 50-100 nanograms depending on application.

3.9.2 Viruses

Isolation of viral nucleic acid differs from that of bacteria in that it does not typically involve selective culture and amplification of the virus. Instead viral DNA or RNA are extracted directly from the sample taken from the patient, and is therefore likely to be contaminated with human nucleic acids. The method for separating viral and human nucleic acids for sequencing may vary depending on the starting sample. For example, clinical samples such as cerebrospinal fluid or serum will contain less contaminating human nucleic acids than swabs or faeces. Simple procedures such as centrifugation may be used to separate viral particles (containing their genomes) from human material. Enzymes such as RNAse and DNase can also be used to selectively remove contaminating nucleic material such as RNA and DNA respectively. The exact sample preparation method used will depend on how pure the sample needs to be compared to the time and cost of purification. In the case of RNA viruses, often the isolate RNA is converted in vitro to DNA so that it can be sequenced. This is followed by an amplification step (PCR) to ensure enough sample is available for sequencing.

3.10 Preparation of DNA for sequencing

Following extraction of genomic DNA, it is fragmented (by enzymatic digestion or PCR amplification) into smaller DNA strands suitable for NGS sequence. This collection of genomic DNA fragments is known as a DNA library. The current library preparation methods for use with NGS platforms are usually carried out *in vitro* and involve a number of steps. First, because the fragmentation process can produce DNA molecules whose ends could either be damaged or incompatible with downstream processes, the ends of the DNA strands must be repaired either by filling in or removing protruding ends (blunt ending). This is followed by linking short, synthesised DNA molecules (adapters) to the ends of the genomic DNA fragments by ligation. Adapter molecules with characteristic 'barcodes' *i.e.* nucleic acids specific to particular samples can also be employed to enable the mixing of multiple patient samples (each indexed with a unique DNA barcode) in a single sequencing run. Adapters have the dual purpose of acting as primers to initiate subsequent reactions, and tethers to a solid surface to which the DNA template fragment(s) will be subsequently attached in all the current NGS platforms. Notably, commercial kits such as the Illumina Nextera XT now allow this library preparation process to be simplified, reducing the hands-on time taken to achieve these critical steps.

Finally, size selection is carried out to enrich for the correct template DNA and remove free adapters (*i.e.* not attached to template DNA) and adapter dimers. This is followed by quality control and amplification steps to ensure that correct amounts of the right template DNA are obtained. This step is particularly important if samples containing genomes of different sizes are being mixed in a single sequencing run, as obtaining even depth of coverage across different size genomes requires the sample concentrations to be normalised so large genomes are not lacking in read depth while smaller ones are sequenced more deeply than required.

Prior to sequencing on NGS platforms, the template DNA must be amplified in order to produce a large number of identical DNA template fragments to ensure a high quality sequence data. The Illumina platforms employ a method known as clonal amplification *in situ* in order to increase the number of copies of template DNA and the Life Technologies platforms use emulsion PCR, where multiple copies of a single template DNA molecule are generated within isolated lipid vesicles.

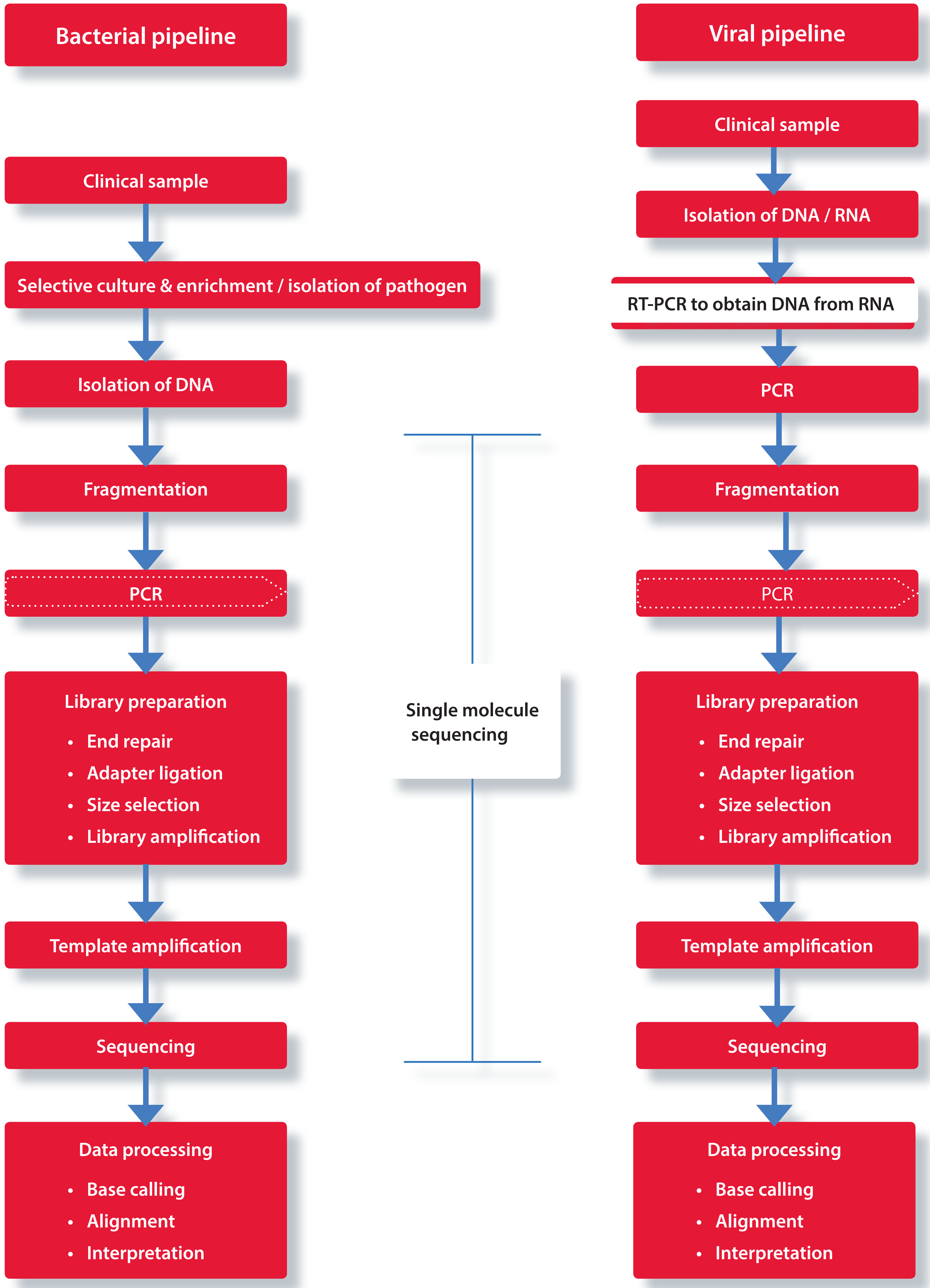
Following amplification, the sample is sequenced to determine the order of bases on each template fragment, using one of a number of different sequencing chemistries. Although sequencing a single read to determine the order of the bases is usually only performed in one direction, it is also possible to sequence both ends of a fragment of DNA by modifying the library preparation protocol slightly, thus artificially increasing read length. This method is known as 'paired-end' sequencing (or 'mate-pair' sequencing for large fragments). As both reads contain positional information, alignment of the reads becomes easier. This is of value for accurate assembly and mapping of the sequence.

3.11 Conclusions

The rapid development of NGS and sample preparation methods has dramatically driven down the cost and time taken to sequence a pathogen genome. It is now possible to produce a draft sequence of a bacterial or viral genome at a cost of between £100-200 in just over a day. The most significant contribution to the overall time taken to determine the genome sequence of a pathogen is the speed with which microbiologists are able to isolate, amplify, extract and prepare the genomic DNA libraries for sequencing. These processes still require a minimum of 24 hours and in some cases where bacterial growth is slow may take weeks. The ability to extract and sequence microbial DNA directly from clinical specimens (for example, through use of metagenomic approaches, discussed in chapter 5) could help reduce turnaround times. Whilst single molecule platforms are beginning to demonstrate limited utility as tools to create high quality, complete reference genomes and to unravel complex genomic architecture, their use in routine clinical or public health microbiology will continue to be limited until these technologies mature further, reducing costs and increasing throughput.

When considering genome sequencing as a whole, however, the challenges and complexities of sample preparation and sequencing pale in comparison to those of data analysis – the topic of the next chapter.

Figure 3.1 Viral and bacterial sequencing pipeline



4 An introduction to bioinformatic analysis of pathogen genomes

Massively parallel sequencing generates an enormous volume of data, the analysis of which requires substantial computational power, purpose-built bioinformatics tools and accurate, comprehensive databases of genomic variation to aid interpretation.

4.1 Introduction

As with human genome analysis, an increasingly significant proportion of the effort associated with analysing pathogen genomes is expended in computational processes. Computational tools are required to assemble whole pathogen genome sequences from the raw fragments of genomic sequence generated by NGS platforms, to interpret variation between sequences of different organisms and to manage the unprecedented volume of pathogen sequence data now being generated around the world. The exact computational approach taken to analysing the genome sequence of a pathogen differs depending on a number of factors, including but not limited to:

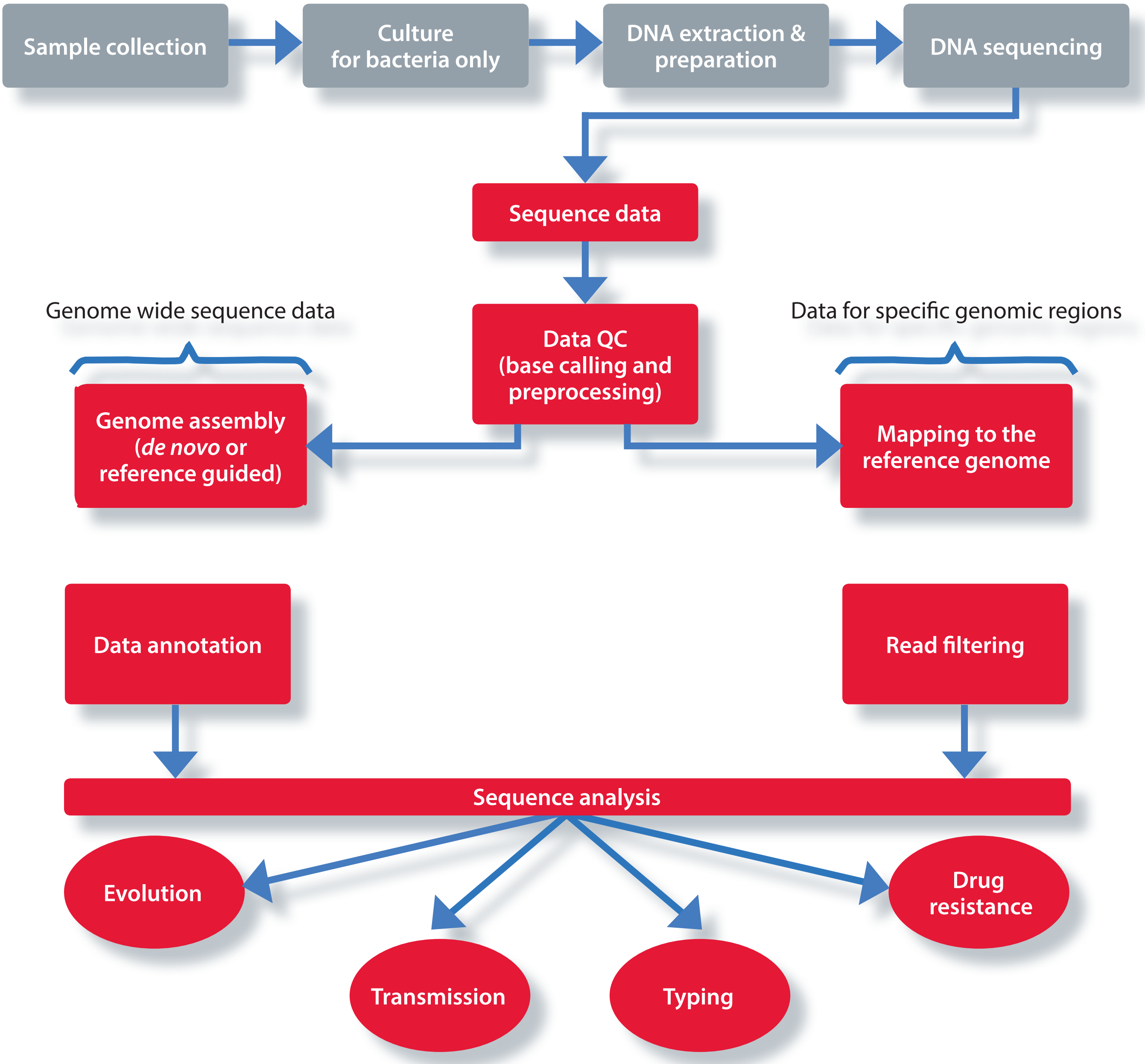
- The type of pathogen being sequenced *e.g.* bacterial *versus* viral
- The upstream processing of pathogen samples *e.g.* sequencing from pure cultures *versus* mixed populations
- The desired downstream use and application of the data *e.g.* decision making in diagnostics, antimicrobial selection or epidemiological investigation

Determination of pathogen genome sequences and interpretation of the significance of their variation requires a combination of analytic disciplines including bioinformatics, population genetics and statistics. In this chapter we describe the data processing steps required to transform raw and fragmentary genomic data from a sequencing platform into a single representation 'pathogen genome', and consider what it will take to reliably detect and interpret variation between these genomes for clinical and public health purposes.

4.2 Pathogen sequences – the data processing steps

There are a number of analytical challenges to be overcome in converting raw sequence data into a pathogen genome sequence. The general process by which this is achieved is illustrated in figure 4.1 and particularly challenging steps are discussed in the subsequent sections.

Figure 4.1 Pathogen sequence data processing steps


Sample collection
Culture for bacteria only
DNA extraction & preparation
DNA sequencing
Sequence data
Data QC (base calling and preprocessing)
Genome wide sequence data
Data for specific genomic regions
Genome assembly (*de novo* or reference guided)
Mapping to the reference genome
Data annotation
Read filtering
Sequence analysis
Evolution
Transmission
Typing
Drug resistance

Adapted from Willson, 2012[51]. Upstream laboratory procedures shown in grey boxes, informatics steps shown in red boxes. Once sequence data has been generated by a sequencing machine it is subject to quality control procedures. Data of an acceptable quality is then used for 'genome assembly', that is reconstruction of the pathogens genome sequence from the sequenced read data, or only data for specific genomic regions of interest will be 'mapped' (matched to an existing known reference sequence), prior to annotation (prescribing functional or structural meaning to the mapped / assembled sequences) and filtered, before its application for sequence analysis.

4.3 Quality control of base calling

Accurate determination of genome sequences is fundamental to effective, high quality clinical and public health applications of genomics. The initial quality control of the data that are used for 'assembling' pathogen genomes is therefore vital. Sequencing machines generate data in the form of 'raw' images, fluorescence read outs, or electrical signals. During the process of base calling the detected 'raw' signals are converted into reads for individual nucleotides (identifying whether the next base in the sequence is an A, T, C or G) within genomic fragments by software that is integrated into the sequencing machine. If a sequencer is unable to determine a base with sufficient confidence then the call is normally denoted with an 'N'. In other cases a base may be assigned incorrectly by mistake.

Ultimately 'miscalled' bases can lead to errors in the subsequent data processing steps, including genome assembly and identification of true genetic variants, and must therefore be detected by quality control (QC) processes. To allow for evaluation of potential errors in base calling, the quality of the sequence data is assessed using either the software tools accompanying the sequencer or standalone applications (*e.g.* FastQC). These tools assign quality scores to individual bases by measuring the probability that a base has been called correctly; Phred-like quality scores (Q scores) being one of the most widely used methods (table 4.1). Generally a score of at least Q25 is the minimum accepted quality score for bases to be included in subsequent data processing steps; however the accepted threshold can vary depending on the application.

Table 4.1 Probabilities of incorrect base calls and accuracy associated with Phred Quality Scores

Phred Quality (Q) Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Following the quality scoring of individual bases, the overall sequence data may then be subjected to further manual and / or automated assessment of the read quality. There are a number of approaches to data QC adopted at this stage, with optimal strategies determined on a case by case basis and through experience. These approaches include various methods for filtering and pre-processing of sequence reads prior to genome assembly, for example:

- Removing reads with low average quality of bases
- Removing reads with 'N' base calls
- Trimming the ends of reads with low quality bases, especially as quality scores may degrade over the course of a read

The exact approach to data QC employed depends on the features of sequence data (*e.g.* quality, quantity and fragment length) and the subsequent tool and method used for constructing genomes from the (short) fragment reads. To date, and especially in the academic research setting, the preprocessing of data generally is not standardised and instead involves an iterative process of testing, evaluation and optimisation to suit the downstream assembly and analysis. Additionally different types of sequencing machines produce different types of sequencing errors and biases unique to each technology, which need to be accounted for specifically in each case. The current absence of national standards or standard operating procedures for defining DNA sequence quality and performing QC assessment on sequence data will pose challenges for the sharing and accurate comparison of data generated and processed in different locations for clinical and public health applications.

4.4 Genome assembly

Whole genomes are assembled from short read fragments generated by NGS sequencing platforms. The assembly process works on the assumption that highly similar DNA fragments originate from the same position in the genome and their similarity can therefore be used to piece together the fragments into longer contiguous sequences (contigs)[52]. In reality this assumption is challenged and the assembly process complicated, by the presence of repetitive sequences within individual genomes. The number, type, and configuration of these varies greatly across species[53]. Some repetitive elements originate from different places in the genome but share the same repeat sequence and since the assembly process searches for overlapping nucleotides, these repeats may be assembled erroneously. Bacteria, for instance, have multiple copies of ribosomal RNA operons (groups of genes that form a single functional unit). Operons tend to be located in different parts of the bacterial genome, individually span several kilobases, and are of highly identical sequences within most species[54]. Other problematic features include shorter repetitive elements and sequence repeats, including insertion sequences whose copy numbers can vary considerably[55,56].

A key challenge to the effective analysis of pathogen genomes is the lack of high quality and complete reference genomes available to guide sequence assembly...In the case of new or emerging infectious diseases the reference-guided approach may limit identification of pathogens not previously sequenced.

4.5 *De novo* versus reference guided assembly

There are two approaches for assembling genomes; *de novo* and reference guided. The latter employs the use of a predetermined 'reference' genome of another (preferably closely genetically related) member of the same species as a template onto which sequenced reads can be placed. *De novo* assembly on the other hand, is the process of merging overlapping sequence reads into a contiguous sequence (contigs), without the use of preliminary information, *i.e.* without a reference genome as a guide. A key challenge to the effective analysis of pathogen genomes is the lack of high quality and complete reference genomes available to guide sequence assembly. Hence reference guided assembly is restricted to those species with an available closely related genome. In the case of new or emerging infectious diseases the reference-guided approach may limit identification of pathogens not previously sequenced. Strategies for efficient and accurate *de novo* assembly are therefore numerous and under continual development and debate as to their utility[57].

Three major algorithmic strategies are used by sequence assemblers for *de novo* assembly of short sequence reads[52] (table 4.2), each with their own advantages and disadvantages. These algorithms have been implemented in a wide range of software tools for the assembly of pathogen genomes (table 4.3). Choosing which software tool to use for genome assembly depends on a number of factors including choice of assembly algorithm, quality of data available and the sequencing platform on which the data was generated.

Table 4.2 Algorithmic strategies for *de novo* assembly*

Assembly Paradigm	Principle	Trade-offs
Greedy	To join together reads that are most similar to each other *i.e.* overlap best, but without contradicting what has already been assembled	'Greedy' principle means only local information is considered at each step and global relationships between reads is overlooked. Repetitive sequences can therefore confuse the assembly process
Overlap-layout-consensus	To construct a graph of the overlap relationships between reads (where the nodes in the graph represent reads, and edges between any pair of reads represent sequence overlap), and apply graph theory to determine 'layout' or relative placement of the reads	Overhead and complexity of the computation, especially when dealing with the large number of very short reads commonly generated when smaller genomes are sequenced to a high depth of coverage
De Bruijn graph	Reads are converted into sets of overlapping k-mers, strings of nucleotides at a set length. The graph structure is then based on k-mers, not reads, and graph nodes represent a series of overlapping k-mers, and reads are then mapped as paths through the graph	Optimal k-mer length for building graphs can vary depending on the variable fragment length and quality produced by different sequencing platforms. Can be stymied by sequencing errors

* Most assembly methods make use of graphs as a means of representing the possible associations between sequenced reads

Table 4.3 Selection of common *de novo* assembly tools used for microbial genomes, information source Dark, 2013[25]. *Tool also performs reference guided assembly

Assembly Tool	Licence	Data Source Compatibility
Velvet: sequence assembler for very short reads	Open-source link	Illumina, SOLiD, pyrosequencing, and Sanger reads
ABySS (Assembly By Short Sequences)	Free for non commercial / academic link	Illumina, SOLiD, pyrosequencing and Sanger reads
Celera Assembler: CABOG (Celera Assembler with Best Overlap Graph)	Open-source link	Pyrosequencing, Illumina, and PacBio reads
Edena (Exact DE Novo Assembler)	Open-source link	Specifically for bacterial genome assemblies and Illumina-based sequences
MaSuRCA (Maryland Super Read Cabog Assembler)	Open-source link	Illumina, SOLiD, and pyrosequencing reads
MIRA* (Mimicking Intelligent Read Assembly)	Open-source link	Sanger, pyrosequencing, Illumina, PacBio reads
SOAP suite	Open-source link	Illumina reads
SOPRA (Statistical Optimisation of Paired Read Assembly)	Open-source link	Pyrosequencing, Illumina, and SOLiD reads

As with *de novo* assembly there are also a wide variety of algorithmic approaches and software for performing reference guided assembly[58]. Software known as 'sequence-alignment' tools are used to map individual short sequenced read fragments against a longer reference sequence in order to reconstruct the original genome sequence of the species. As sequencing technologies have developed, many new alignment tools have surfaced to exploit or perform more optimally with the characteristics of the sequence data specific to each technology, for example the varying lengths of sequenced fragments, or differences in expected in read quality.

4.6 Draft *versus* finished genomes

'Complete' reconstruction of pathogen genomes from NGS technologies is inherently difficult due to both the short read outputs of the sequencing instruments and the complex features of pathogen genomes, such as repeated sequences and areas prone to recombination. Consequently many assembled pathogen genomes contain gaps in regions intractable to current sequencing and assembly methods. These 'incomplete' genomes are referred to as 'draft' assemblies. A 'finished' complete genome requires substantial resources to close the gaps in draft assemblies and therefore many have been left unfinished, with the number of draft genomes drastically outstripping finished assemblies. The continuing lack of finished genomes is a major barrier to the use of reference-guided assembly as this process is most successful when a high quality, finished reference genome is available. The disparity between draft and finished genomes numbers also underscores the need to distinguish between 'good' and 'poor' draft assemblies[59].

Having 'finished' microbial genomes is considered a worthwhile objective since it allows for more detailed genomic analysis[60]. Accordingly, advances in experimental strategies to generate data that can be effectively used by assembly programmes to accurately reconstruct original DNA sequences, as well as improvements *in silico* genome 'finishing' are underway[61,62] .Third-generation, single-molecule sequencing, is expected to simplify genome assembly by generating longer and less error prone sequence reads. In essence, longer read lengths can allow sequence repeats to be placed more accurately, assuming these repeats become shorter than the overall read length, thereby reducing the complexity of assembly[42].

Evaluation of a genome assembly is considered worthwhile, since *de novo* short-read assemblers are known to be error prone[63] and imperfections in assembly can misinform subsequent genome annotation or analysis. Importantly, quality metrics assigned to a genome assembly can alert downstream users as to its accuracy and therefore its utility.

A graph showing numbers of draft and finished microbial genomes is available at www.sciencemag.org[64].

4.7 Genome sequence mapping and annotation

It is not uncommon to map and (re)order microbial contigs against a suitable reference genome, even when prior *de novo* assembly has been performed[63]. Reference genome mapping is performed, where possible, against the closest related species with a 'finished' genome. Sequence alignment against existing reference genomes can aid discovery, functional and structural comparisons, and provide an assessment of the error rate of an assembly.

Having constructed a draft genome from the sequenced reads, the next stage in the analysis pipeline is to 'annotate' the assembly. Annotation in this context is the process of ascribing biological information to the genomic sequences[65]. The genome is scanned to identify elements that represent genes or other (structural) features of interest (noncoding RNAs, operons). Functional details of identified elements (*e.g.* their biological role) are also specified, where this information is known or predicted, for example the biological function of a gene or whether it is associated with antimicrobial resistance. A variety of web based and command-line tools exist to perform automated annotation (table 4.4), as do tools for viewing and sharing annotated genomes. The quality of a genome annotation largely hinges on the underlying accuracy and completeness of the gene database used by the annotation tool.

Table 4.4 Examples of microbial genome annotation tools

Annotation Tool	Access	Annotates
AGeS - Annotation / Analysis of Genome Sequences[66]	Stand alone link	Bacterial genomes
BaSYS - Bacterial Annotation System[67]	Web submission link	Bacterial genomes
BG7 - Bacterial genome annotation system[68]	Stand alone link	Bacterial genomes
GeneMark gene prediction programs[69]	Web submission link	Prokaryotic, eukaryotic and viral sequences
IGS (Institute for Genome Sciences) Annotation Engine	Web submission link	Prokaryotic sequences
Integrated Microbial Genomes (and Metagenomes) Systems[70, 71, 72]	Web submission link	Archaea, bacteria, eukarya, viruses and plasmids
MAKER Web Annotation Service[73,74]	Web submission link	Prokaryotic (and eukaryotic) sequences
NCBI Prokaryotic Genome Annotation Pipeline	Web submission link	Prokaryotic sequences
Prokka: rapid prokaryotic genome annotation[75]	Stand alone link	Prokaryotic sequences
RAST - Rapid Annotation using Subsystem Technology[76]	Web submission link	Bacterial and archaeal genomes
VIGOR - Viral Genome ORF Reader[77,78]	Web submission link	Viral genomes
xBASE Bacterial Genome Annotation Service[79,80]	Web submission link	Bacterial genomes

In addition to the general microbial genome annotators, there is a range of specialised annotation tools for mining genomics sequences when focused gene / element discovery and / or details on functional significance are required. Examples of specialised annotators include antimicrobial resistance gene identifiers[81], other drug resistance predictors[82], pathogen virulence determination tools[83,84,85,86], bacterial pathogenic potential predictors[87], bacterial species predictors[88]. Most microbial annotation tools were developed to facilitate academic research, and the quality and breadth of their content may not be sufficient to robustly support clinical outcomes. One shortcoming of existing annotation tools is the dearth of knowledge and correlation between genotype and phenotype relationships, for example the extent to which the presence of an antimicrobial resistance gene is correlated to the phenotypic resistance profile of the bacterium in which it is found.

4.8 Sequence databases and quality standards

The interpretation of pathogen genome sequences, be it for pathogen identification or other purposes, is entirely dependent on the ability to compare them to existing identified and annotated sequences. Therefore sharing of, and access to existing genome data and clinical and epidemiological metadata is essential as this informs most analytical procedures such as reference guided assembly, species identification, strain typing, or drug resistance analysis[89]. The sequence data that are shared vary widely between databases and may include raw sequence data files, preprocessed sequenced reads, assembled genomes, and annotated assemblies. These types of data may be shared in the public domain or subject to varying degrees of access restriction *e.g.* the requirement to demonstrate accreditation as a researcher or health practitioner prior to gaining access.

The quality and completeness of these databases are fundamental determinants of the accuracy of most forms of pathogen genome analysis[89,90] as these typically include a comparison with existing genetic or genomic data stored in these databases. Analyses performed to inform healthcare and clinical outcomes require more stringent data quality assurances than are routinely accepted for research purposes, as misinterpretation arising from poor quality data may have a direct impact on patient health. As most current sources of pathogen genomic data are populated with 'research grade' data, their use in clinical and public health settings may pose significant challenges.

To assist clinical-grade analysis, some specialised databases apply manual or automated quality control measures to their content prior to making it available to users. One such resource, the Stanford Database HIV drug resistance database, (designed to store and analyse data relating HIV drug resistance[91]), applies a manual data curation process to ensure the accuracy of annotations and sequence data. This enables clinical decisions to be made confidently on the basis of information contained within this resource, and indeed this database is used routinely in HIV genotyping by virologists in the UK. Whilst small scale databases that curate only small numbers of genes for individual organisms can operate using manual curation to ensure sufficient quality control, this is not likely to be a realistic model for larger, multispecies databases, for which automated curation processes will be required.

4.9 Sequence data analysis – identifying and interpreting variation

As a general principle the first step of most genomic analyses is to compare genes or genomes of interest with one another and to look for the extent to which they differ, a process described as 'variant calling'. Whilst this process is based on the same computational approaches for both human and pathogen genomes, the relatively high rates of mutation in pathogen genomes and the consequently vast diversity within and between species means that in practice it is more difficult to accurately detect and interpret variation. The accuracy of this first step is, however, essential to the effective downstream use of any variation identified in answering specific clinical or public health questions *e.g.* to which antibiotics is the pathogen resistant, or are the bacteria infecting patient 'A' likely to have been passed to them from patient 'B'?

The relatively high rates of mutation in pathogen genomes and the consequently vast within and between species diversity make it more difficult to accurately detect and interpret variation than in the human genome.

The first crucial step in variant calling and interpretation is to identify to which genes or genomes your pathogen sequence of interest ought to be compared. This decision is dictated by your pre-existing knowledge of the pathogen under investigation, the intended downstream application of the variant data that will be obtained, the availability of comparator datasets for the pathogen of interest and the type of variation being investigated.

Where the aim is to look for variation between the pathogen sequence being tested and an available single reference genome sequence of the same species, the two sequences will be aligned with one another computationally and the mismatches between them highlighted for further analysis. This approach is typically used when trying to determine how closely related two bacterial or viral isolates are to one another in order to identify or rule out transmission between hosts. It is also of great utility in determining whether pathogen genomes carry variation that may render the organism resistant to certain drug treatments. The lack of high quality finished reference genome sequences and the challenges inherent in determining what constitutes a 'reference sequence' for highly variable pathogen genomes significantly hampers the success of this type of analysis.

Where there are existing databases containing large numbers of variants of gene or genome sequences for a particular pathogen it is also possible to compare the single whole genome sequences (or specified genes from within that sequence) under investigation against every sequence in the variant database to find a 'best match'. This approach is commonly used in strain typing schemes such as Multi Locus Sequence Typing (MLST)[92] where the aim is to classify which species subtype the bacterial isolate being sequenced belongs to for the purposes of disease surveillance or outbreak management.

4.9.1 Analysing sequence variation – determining relationships

An important component of microbiological investigations is the estimation of relatedness of isolates of the same organism from different sources; *i.e.* their evolutionary relationships. This information can be used to identify or exclude disease outbreaks, track transmission routes, determine sources of infection and more broadly to understand the changes occurring in a given pathogen population over space and time. The application of genomic data for surveillance and infection control is discussed in more detail in chapter 7. From an analytic perspective the use of genomic data for inferring evolutionary relationships involves the comparison of sequence variations across all isolates in an investigation. The numbers of isolates to be compared will vary depending on the timescales and geographic breadth of an investigation, as well as the number of isolates implicated in the investigation.

The extent of sequence variation between isolates is used to construct a family tree of relationships between the samples, known as phylogenetic trees or networks (section 7.6.2). Data generated by either *de novo* assembly or reference mapping can be used for this analyses and there are a number of methods and algorithmic approaches for constructing phylogenetic trees. Individual genes, genomic regions or whole genome sequences can be compared by aligning homologous sequence regions across all isolates under investigation, in a process known as multiple sequence alignment (MSA). A range of algorithms and MSA computer programmes have been developed over the years each with different performance capabilities[93]. MLST profiles can also be used for phylogenetic analysis. The advantage of using whole genome sequence data over MLST is that the latter only uses data from seven housekeeping genes and is therefore much less discriminatory than the former. However the most appropriate choice of phylogenetic method will vary depending on the characteristics of the pathogen population under investigations, as well as the desired level of discrimination for the intended purpose – *e.g.* longitudinal analysis of isolates collected over several years and broad geographic space – *versus* smaller number of isolates collected as part of an outbreak investigation in a confined space such as a hospital ward. However before relationships between isolates can be inferred, it is important to factor in rates of mutation and the effect of genetic recombination (chapter 2) for each organism. For example genomic regions which undergo high rates of recombination, such as horizontal gene transfer can obscure phylogenetic signals[94]. Therefore sequence data on each organism requires careful calibration before it can be applied for outbreak analysis.

Another notable factor to consider in a clinical context is the computing cost of executing different analytic methods as this will influence the turnaround times for results. It is noted that *de novo* assembly and read mapping can generally be competed in under an hour; analysis of MLST or gene sequences ranges from seconds or up to an hour[94]. Similarly the time taken to perform phylogenetic analysis can range from minutes to several hours and is highly dependent on the choice of phylogenetic method, algorithm used, and the number of isolates to be compared. Moreover error rates in determining relationships also vary by choice of algorithm and the algorithms suitability to the sequence data being analysed. Therefore the most appropriate analysis strategy requires a careful consideration to balance computing costs, error rates and desired turnaround times, and remains a subject of debate[94].

4.10 Conclusions

Whilst in principle the process of transforming raw sequence data into an interpretable base-by-base description of a pathogen genome is relatively straightforward, in practice the use of bioinformatics techniques and tools for clinical and public health applications requires careful evaluation at a number of levels.

As the above descriptions of basic quality control, assembly, annotation and variant calling demonstrate, the process of pathogen genome analysis is far from being standardised and automated. For each step multiple methods may be applied, giving varying results, the utility of which depends strongly on their intended application. As the preceding chapters on the structure of pathogen genomes and the methods used for generating raw sequence data also demonstrate, different analysis solutions will be required depending on both the pathogen being analysed and the sequencing method used to acquire the data. These variations in methodology, and limitations in both analytical quality and the quality of the underlying knowledge on which interpretation is based pose particular challenges to the delivery of 'clinical grade' genome interpretation. Furthermore, the utility of these techniques in national and international surveillance functions where quality control and consistency of analytical methods will be important, will also depend on these limitations being addressed.

Part II of this report will provide examples of how the bioinformatic steps described here can be applied to analyse infectious diseases in ways that have both clinical and public health utility. Part III explores the challenges of adapting these tools and approaches into forms that meet the needs of diagnostic and public health microbiology laboratories.

Part II

The microbiological investigations typically undertaken as part of the management of infectious disease in both individuals and populations can be broadly characterised as addressing at least one of the following questions:

- What organism is causing the infection?
- What drugs can be used to treat it?
- How is it related to other similar infections?

Information obtained from whole genome sequencing of pathogens could, in principle, contribute to answering all of these questions. The pertinent question for policy makers and practitioners is, however, whether genomic technology and knowledge can currently deliver these answers in a way that is sufficiently superior to existing microbiological methods (as measured by increased sensitivity and specificity, reduced turnaround time and reduced cost) to warrant its introduction into mainstream clinical and public health microbiological practice.

This part of the report aims to answer this crucial question by outlining the current methods used to answer each of these questions and comparing their effectiveness to that of the available genomic methods described in previous chapters.

In the final chapter we also consider some of the potential future applications of genomics to the management of infectious disease, beyond addressing these key microbiological questions. These are even further from being 'implementation' ready, but nevertheless are important to consider given the pace of change in genomic technologies.

5 Identifying pathogens causing infectious disease

Whilst the utility of WGS for the identification of pathogens is not currently likely to add much value to existing microbiological practices, developments in metagenomics has the potential to transform diagnostic microbiology.

5.1 Current methods for identifying pathogens

Identifying the pathogen suspected to be causing an infection is the principle and often only microbiological investigation required for the management and control of infectious disease. Standard methods for identifying bacterial pathogens from patient or environmental samples were established in the 19th century and are based on culturing the presumptive pathogenic organisms and suppressing the growth of any nonpathogenic commensal organisms on bacterial growth media.

These most commonly used methods for bacterial pathogen identification are described as phenotypic, as they rely on detecting 'expressed characteristics' of the organism, such as physical appearance or growth that is restricted to certain conditions. These phenotypic methods may include methods such as mass spectrometry, which can detect unique patterns of molecules present in the different species of bacteria, and serology, where antibodies are used to detect the presence of particular proteins, characteristic of specific bacterial species on the surface of the organisms. The use of genotypic methods, where selected parts of the genome unique to particular species are amplified and detected, is not common in first line testing for bacterial pathogen identification. Exceptions to this include detection of *Mycobacterium tuberculosis* complex from sputum samples and faecal analysis for some gastrointestinal pathogens, which can be undertaken by PCR based methods.

By contrast, detection of viral pathogens, has for the most part transitioned from phenotypic to genotypic methods, where the identifying characteristic of the virus used in detection is a part of its genome rather than any of the expressed products of the genes in that genome. This is usually achieved by PCR amplification of fragments of DNA or RNA unique to the genome of the virus suspected to be infecting the patient. The principal advantages of the genotypic approach to identifying viral pathogens are that:

- Viral DNA / RNA can be isolated from patient samples without the need for culture to select and amplify the pathogen for phenotypic examination

- Eliminating the need for culture reduces significantly the turnaround time to the test result from 18-24 hours (for a typical bacterial culture) to four hours for a viral PCR

- Genotypic tests also typically allow for greater resolution between what might appear phenotypically similar strains or species of pathogen that have significantly different clinical manifestations

- PCR based tests are typically cheaper than the serological (antibody-based) tests they have replaced

5.2 Impact of whole genome sequencing on bacterial and viral pathogen identification

Knowing the identity of the pathogen causing any infection allows clinicians to identify appropriate treatment (*e.g.* to give antibiotics for a bacterial respiratory infection but avoid giving them for a viral infection) and to determine any infection control measures that may be required to prevent its spread. The utility of this information is to a large extent dependent on its timeliness. Diagnostic microbiology practice aims to identify the majority of pathogens within 24 hours, and often much sooner. The absence of timely information about pathogen identity requires clinical and infection control responses to be undertaken empirically, potentially using unnecessarily broad spectrum antibiotics.

Knowing the identity of the pathogen causing any infection allows clinicians to identify appropriate treatment...The utility of this information is to a large extent dependent on its timeliness. Diagnostic microbiology practice aims to identify the majority of pathogens within 24 hours, and often much sooner.

Whilst whole genome sequencing produces high resolution identification of a bacterial pathogen, even with optimised laboratory processes this information is unlikely to be available within a time frame, or at a cost, that can compete with the phenotypic methods currently used. This is largely because in order to obtain genomic DNA to sequence, the bacterial pathogen must first be cultured (a process taking up to 24 hours). If identification can be made direct from the culture using phenotypic methods, there is little obvious additional clinical utility in spending the additional 24 hours required to prepare, sequence and analyse the genome of the causative agent. The utility of whole genome sequencing as a primary means to identify the causative pathogen in bacterial infections seen in acute clinical settings is therefore extremely limited, given the current state of knowledge and technology.

Rapid molecular detection of viral pathogens using PCR-based methods is already standard practice in UK virology laboratories. Whole genome sequencing of viral genomes is also unlikely, therefore, to be of significant additional clinical utility when simply trying to determine the identity of the viral pathogen as this can already be achieved rapidly and cost-effectively.The exception may be where the existing repertoire of tests is exhausted without success, typically because the causative pathogen is novel or extremely rare, in which circumstances unbiased whole genome sequencing direct from the patient's sample may be the only way to detect the cause of the infection. Such unbiased, culture free whole genome sequencing, and its application to both bacterial and viral pathogens is described in more detail below.

5.3 Diagnostic metagenomics – a culture free future?

The requirement to isolate and amplify a pure culture of any suspected bacterial pathogen prior to conducting WGS places significant limitations on its utility in diagnostic microbiology, not only because of the delays these procedures introduce, but because for a significant proportion of samples taken from patients with suspected infections no organism is ever successfully cultured or detected.

The most promising route to overcoming this limitation, and in the process making genome-based testing suitable as a routine first line test to identify a pathogen is an approach known as metagenomics. First developed to sample the genetic diversity of micro-organisms present in different ecosystems, metagenomics involves the (relatively) unbiased extraction of the entire genomic contents of a sample and the direct sequencing of the resulting mixture of multiple genomes without any intervening culture step. Crucially, this approach allows the genomes of organisms that cannot be cultured in the laboratory to be detected, enumerated, and to a limited extent characterised. Metagenomics has been demonstrated to be capable of detecting novel viruses causing infection in humans directly from blood samples, and detecting the presence of an unusual strain of *E. coli* directly from the stool samples of patients involved in a Europe wide outbreak of severe gastrointestinal disease[95,96,97]. More recently there have also been two reports of metagenomics being used in a 'real time' diagnostic context to identify the cause of severe brain and lung infections that could not be determined by existing culture-based microbiological methods[98,99].

Whilst there is great optimism that metagenomic approaches will in the future contribute significantly to the practice of diagnostic microbiology, there remain several significant technical limitations that mean it is not yet ready for use in mainstream microbiology. These include:

- **Complexity of analysis –** when a single bacterial or viral species is sequenced using massively parallel sequencing following purifying and amplifying culture the assumption can be made that the vast array of small fragments that have been sequenced should all 'fit together' to make up the genome of that single organism. When undertaking metagenomic analysis, however, the number and diversity of different species whose genomes are represented in the sequence reads obtained is unknown and so disentangling which pieces of sequence belong to which genome is a huge analytical challenge requiring significant computing power and time. This problem is analogous to being given a single 100 piece jigsaw to complete, with a picture of the complete jigsaw to guide you versus being given millions of jigsaw pieces all mixed together with only a catalogue of all of the known jigsaw pictures in the world for guidance. While there are several computational approaches to overcoming these challenges, they remain too time consuming and costly to be implemented in routine clinical settings.

- **Low depth of sampling –** the ability to sequence and assemble the genome of an organism using massively parallel sequencing depends strongly on the quantity of DNA available. Whilst new preparation methods (chapter 4) mean that sufficiently high depth coverage of a genome by NGS to assemble it *de novo* can be achieved from only a few nanograms of DNA, this can still be difficult to achieve given the small size (in base pairs) of bacterial and viral genomes, and the low abundance of each of the many individual species that may be present within a sample. This limitation will result in incomplete and uneven coverage of each pathogen genome, making subsequent assembly and mapping particularly challenging.

- **Separating microbial DNA from human DNA** – the samples used in 'clinical' metagenomic analysis, most commonly stool or sputum, must be processed mechanically and chemically to remove as much contaminating human DNA as possible. Given the relative sizes of and abundances of human and pathogen genomes, the continued presence of significant quantities of human DNA in a metagenomic sequencing run will result in a large proportion of the sequence data covering the human rather than the bacterial and viral genomes. Whilst human genome sequences 'contaminating' a metagenomic analysis can be removed from the data computationally, it is preferable to remove as much as possible during sample preparation to maximise the depth of coverage of the organisms of interest and improve cost efficiency.

- **Difficulties in clinical interpretation –** whilst metagenomics enables the sensitive detection of pathogen DNA within samples taken from patients, it is particularly challenging to determine whether the pathogens detected by these methods are the causative agents underlying the patient's disease and are therefore the appropriate target against which to direct therapeutics or surveillance operations. If the pathogen cannot be cultured, it cannot be tested to determine if it is truly capable of causing the infectious disease in which it has been implicated. Instead it will be necessary to look for correlations between metagenomic traits *i.e.* the presence of particular pathogen DNA and the presence of disease across large numbers of patients in order to try and infer causality.

5.4 Conclusions

Whilst WGS for the identification of pathogens (bacterial or viral) is not likely to add much value to existing microbiological practices at present, that does not mean this method will not become routinely used in the future. Developments in metagenomics will ultimately enable rapid and routine use of culture free sequencing to detect pathogens and has the potential to transform diagnostic microbiology in the process by providing a single unified workflow that can be applied to all diagnostic specimens requiring genomic analysis.

Furthermore, genomics is likely to be introduced sooner for routine use in addressing other microbiological questions. Thus it may still find utility as a second line additional confirmatory test that serves to clarify inaccurate or ambiguous first line phenotypic test results *e.g.* by distinguishing between the different species that constitute *Mycobacterium tuberculosis* complex, and may therefore contribute to improved patient management.

Developments in metagenomics will ultimately enable rapid and routine use of culture free sequencing to detect pathogens and has the potential to transform diagnostic microbiology.

6 Determining antimicrobial resistance in pathogens

The ability to determine antimicrobial resistance of different pathogens using whole genome sequencing holds great promise. However, without accurate, comprehensive and validated databases of clinically relevant genotype-phenotype correlations, genomic assays are unlikely to outperform existing phenotypic tests.

6.1 Current methods for antimicrobial susceptibility testing

A principal function of clinical microbiological investigation is to determine the susceptibility of pathogens to the variety of antibiotic or antiviral agents that may be used to treat the infections they cause in the patient.

For bacteria this is typically achieved by quantifying the extent to which an organism is able to grow in the presence of varying concentrations of the desired drug. This process is now mostly achieved using the automated bacterial culture systems found within clinical microbiology laboratories in England. National and international standards have been established that provide protocols for antimicrobial susceptibility testing (AST) and defined cut-off values for bacterial growth rates indicative of either susceptibility or resistance of each organism to a given antibiotic. These standards are crucial in ensuring best practice and consistent interpretation of what may often appear to be ambiguous or intermediate results.

Drug susceptibility testing is not standard practice for most viral infections, with a few notable exceptions. These include chronic infections such as HIV infection where drug resistance must be assessed to determine initial treatment, and is subsequently monitored since it may evolve within the population of viruses in the individual patient. Resistance to antiretroviral therapy (ART) can be predicted by sequencing the HIV genes whose gene products are the targets of the ART. Algorithms can then be used to assess whether mutations within those genes are likely to cause resistance, and therapy can be tailored appropriately. Similarly response of Hepatitis C virus (HCV) infections to antiviral therapy is dependent on the sequence of the viral genome, and thus genotyping is used to determine the appropriate therapeutic regime for each HCV patient. Viral genotyping has, until recently, been mainly undertaken by Sanger sequencing in a small number of virology laboratories with access to capillary sequencing technology. A shift is now underway towards use of the same NGS based platforms for whole genome sequencing as tools for ascertaining drug resistance in these viruses, and clinically accredited services utilising this technology for HIV drug susceptibility testing are now available through Public Health England laboratories.

6.2 Impact of WGS on drug susceptibility testing

Acquisition of resistance to antibiotics by bacteria is mediated by the acquisition of new genes, or variation within existing parts of the core or accessory genome that enable the organism to avoid the toxic effects of the drug. A small number of studies have now demonstrated that, in principle, it is possible to use WGS to detect genes and genomic variants that are known to cause antibiotic resistance. In one study, focused on resistance testing for *Staphylococcus aureus*, the authors were able to predict the antibiotic susceptibility determined by standard phenotypic methods in the laboratory with 97% accuracy by comparing the WGS of each bacterial isolate to a database of known genes and gene variants that cause significant resistance to a range of different antibiotics in *S. aureus*. A retrospective WGS study on *E. coli* and *K. pneumoniae* isolates demonstrated that WGS was as sensitive and as specific as currently used phenotypic methods at predicting antimicrobial sensitivity[100] . While such studies show that antibiotic susceptibility testing is possible, they have not as yet established its clinical utility in real world laboratory settings. The principal disadvantage of using WGS is that resistance attributable to the presence of novel resistance genes and alleles not already present in the databases used for predicting susceptibility will not be detected by a WGS based test. In this case, a resistant organism might be predicted by genomics (or indeed any genotypic method) to be susceptible, leading to inappropriate treatment of the patient.

Over time, as WGS is increasingly used for other applications, it is reasonable to expect that these gaps in the existing knowledge of genotype-phenotype relationships for antibiotic resistance mechanisms will be closed. The determining factors then in the application of WGS for this purpose in routine bacterial microbiology are likely to become timeliness and cost relative to existing methods, which currently remain in favour of current phenotypic methods.

6.3 WGS based drug susceptibility testing for *M. tuberculosis*

While genomics based testing cannot currently outperform phenotypic drug susceptibility testing on accuracy, cost or timeliness for almost all bacterial pathogens, there is one group of organisms to which this does not apply. Mycobacteria, in particular *M. tuberculosis*, grow sufficiently slowly that alternative methods such as WGS that do not rely on measuring the effect of antibiotics on growth rate have the potential to outperform standard resistance testing.

Culture growth based determination of drug susceptibility for *M. tuberculosis* can take up to eight weeks for first line treatments, and if resistance is detected, a further extended period of time elapses while susceptibility to second line treatments are determined. The presence of *M. tuberculosis* in a patient sample can, however, be detected within days using culture, sputum smear or PCR-based methods and usually prompts initation of antibiotic therapy even in the absence of susceptibility information to guide the selection of the most appropriate regime. This can have significant negative consequences where the patient is infected with an MDR or XDR strain of *M. tuberculosis.* In this case the patient may receive ineffective therapy for the many weeks taken to detect resistance, prolonging the course of their own disease and their ability to infect others. Furthermore current phenotypic methods do not accurately detect mixed infections with *M. tuberculosis* which may have different antibiotic resistance profiles.

Whilst whole genome sequencing of *M. tuberculosis* requires the organism to be cultured to generate sufficient DNA for sequencing, enough *M. tuberculosis* bacteria for this purpose can be obtained in a few days (but may take up to six weeks). Analysis of the sequenced genome can then predict resistance weeks and perhaps even months sooner than current conventional methods. It remains to be determined whether the sensitivity and specificity of these genomics based predictions will be sufficient to match, if not outperform current phenotypic methods of AST for *M. tuberculosis.* Nevertheless, the dramatic time advantage for genomics in this case has driven the rapid development of methods for undertaking genome-based AST for *M. tuberculosis,* and these are currently being piloted in Public Health England specialist microbiology laboratories.

6.4 Conclusions

The ability to determine antimicrobial resistance of different pathogens using whole genome sequencing holds great promise. A unified genomic approach, in which a single assay (whole genome sequencing) could be applied to all isolates, with only the data analysis varying according to the species being analysed, would offer the benefit of streamlining the multiple parallel phenotypic assays currently undertaken to test for susceptibility to a wide range of antibiotics. Arguably, genomics could provide more accurate results than phenotypic assays that are subject to significant intrinsic variation caused by the large number of external factors that can affect bacterial growth, and the intrinsic biological variation in the organisms themselves. Despite this optimism, it remains the case that without accurate, comprehensive and validated databases of clinically relevant genotype-phenotype correlations, genomic assays are unlikely to outperform existing phenotypic tests. Furthermore, the extended turnaround time associated with culture-dependent genome sequencing also undermines the utility of the information it provides when compared against current phenotypic methodologies for most organisms.

One way to partially circumvent the turnaround time problem may be to develop simplified point of care devices that are able to detect particular subsets of genes or mutations associated with antibiotic resistance within minutes rather than hours to days. Such machines typically use DNA based approaches, but selectively amplify and detect only predefined candidate regions of the genome and so are less comprehensive than WGS, meaning that a negative result would not necessarily mean no antibiotic resistance. Nevertheless it is anticipated that such devices will have a role in the future and may be particularly appropriate in low resource settings where access to clinical microbiology facilities and complex WGS workflows may be limited.

7 Infection control of common pathogens

Genomic epidemiology can be used for the detection, investigation and resolution of infectious disease outbreaks as demonstrated in the case studies below.

7.1 Introduction

The focus of clinical microbiology is principally on investigations to determine the identity and susceptibility of the pathogen causing an infection for the purposes of managing the infected patient. There are, however, many further microbiological investigations routinely undertaken for the wider purpose of monitoring and controlling infectious disease at a population level. Such investigations inform and direct health protection activities such as immunisation, infection control and outbreak management. Examples include:

- Surveillance of pathogens of public health importance, *e.g.* surveillance of influenza informs the choice of strains included in the global vaccine

- Comparison of cases of infections in a cluster to establish that they are linked, identification of sources of infection, or detection of clusters of linked infections within wider groups of cases

- Longitudinal studies of the incidence of different infections with the aim of determining the effectiveness of public health interventions such as vaccination campaigns

What unites these different contributions to health protection is that they rely on the ability of microbiologists to characterise and discriminate between pathogens at a higher resolution than speciation and drug susceptibility.

7.2 Current methods for characterising and discriminating between pathogens

The methods currently used to characterise pathogens under surveillance in the UK vary according to the identity of the pathogen and the purpose of the surveillance. The aim of these investigations is to distinguish one strain from other strains within the species, and to determine characteristics relevant to its treatment and control, such as similarity to vaccine strains or antimicrobial susceptilibity.

The utility of each method is dependent on:

- The resolution with which it can discriminate between strains of the same species
- The magnitude of the clinically and epidemiologically relevant variation between strains
- The rate at which the different strains change their characteristics
- The diversity of strains circulating in the population

Different methods also vary in their utility depending on the situation in which they are being applied; outbreak investigations will typically require higher resolution typing methods than longitudinal surveys of disease prevalence, where the aim is to understand long-term trends in the population rather than to discriminate the sources of individual cases of an infection.

Methods for determining the characteristics and relatedness of bacterial strains can be broadly classified into two groups: phenotypic and genotypic and are generically referred to as 'typing' methods. Examples of each are given in the boxes below:

Box 7.1 Phenotypic typing methods

- **Serotyping** – relies on the observation that closely related bacteria can often be distinguished from one another by looking at differences in the molecules that coat their outer surface. Antibodies are produced in the laboratory that recognise surface molecules specific to different bacterial strains and these are then used in assays that can determine which strain is present in a given sample from a patient. This method is commonly used for typing *Streptococcus pneumoniae* and *Salmonella* subspecies

- **Phage typing** – distinguishes between different bacterial strains on the basis of whether or not they are susceptible to infection by different types of bacteriophage. This method is used for typing *Salmonella*

- **Antibiograms** – distinguishes bacterial strains on the basis of their susceptibility to different antibiotic drugs. The bacteria are cultured in the presence of a range of antibiotics and the extent to which they can or cannot grow in the presence of these drugs is quantified and reported

Box 7.2 Genotypic typing methods

- **Pulsed Field Gel Electrophoresis (PFGE)** – uses enzymes that cut bacterial genomic DNA into large fragments whose size depends on the strain of bacteria from which the genomic DNA was extracted. These fragments are then separated according to their size, and each strain is distinguished by their characteristic fragment size pattern

- **Multi locus Variable Number Tandem Repeat Analysis (MLVA)** – analyses regions of repetitive DNA sequence within the genome that vary in size between different bacterial strains. PCR is used to selectively amplify these repetitive regions and the number of repeats at each genomic location studied is determined. Each different strain can then be classified and distinguished according to the pattern of different repeat lengths across its genome

- **Single locus sequence typing (SLST)** – selectively amplifies and analyses a single part of the genome, either a single gene or a single intergenic region, and uses variation at that locus to distinguish and classify strains of bacteria

- **Multi-locus sequence typing (MLST)** – selectively amplifies and determines the sequence of fragments of several (typically seven) genes. Each unique sequence determined for each gene is given an identifier and the different combinations of these identifiers across the seven genes can be used to discriminate between bacterial strains

7.3 Advantages and limitations of current typing methods

Characteristics of an optimal typing method include:

- Discriminatory power sufficient to address the clinical or epidemiological question

- Reproducibility over time and across different laboratories

- Short assay time to minimise delays in producing actionable results in outbreak situations

- Low cost to maximise affordability and universality of use

- Easy to perform assays that do not require highly specialised training and equipment

- Easy to interpret results that are quantitative and unambiguous

- Standardised classification nomenclature enabling portability of results and reports between laboratories

None of the different typing methods described in the boxes above satisfies all of the above criteria. Phenotypic methods often have good discriminatory power, for example the ability to distinguish between the >2000 serotypes of *Salmonella enterica*, but often also rely on expensive and scarce reagents such as antisera, require highly trained scientists to perform the assays, have long turnaround times (often weeks) and are difficult to reproduce across different laboratories. These methods have, however, been the mainstay of bacterial microbiology for many years and so are often associated with well-defined and standardised classification schemes. These enhance their utility when considering their use in national and international surveillance efforts that encompass multiple laboratories. Nevertheless, their other limitations have driven the adoption of many of the genotypic methods that are described in box 7.2.

The principal advantages of most genotypic typing schemes over phenotypic ones is that they offer enhanced discriminatory power, are more reproducible and therefore more easily comparable between laboratories. This enhances their utility in both continuous epidemiological surveillance and in outbreak investigations. Despite these significant advantages, these methods remain relatively technically complex, can be laborious and time consuming and in some cases remain costly. Furthermore, while genotypic typing may offer enhanced discriminatory power compared to phenotypic testing it is often still insufficiently discriminatory to distinguish between closely related isolates, which can be crucial for performance in outbreak investigations. This is exemplified by the *spa* gene typing methodology used to distinguish the infections caused by different subtypes of methicillin resistant *Staphylococcus aureus* (MRSA). Understanding the transmission pattern of an MRSA outbreak requires epidemiologists to be able to distinguish small differences between the bacteria present in each patient and use the size of these differences to infer whether two cases are likely to be linked by a transmission event or are unrelated. Eighty percent of MRSA infections in the UK belong to the same *spa* type and so are, by this method at least, indistinguishable.

7.4 Whole genome sequencing - a simplified approach to characterisation and discrimination in microbiology?

The above discussion highlights several limitations of existing methods for typing pathogens that have significant impact on their utility in public health and clinical microbiology. These can be conceptualised as relating either to complexity or precision (box 7.3).

The complexity of current microbiology practice arises to a large extent from the vast diversity of organisms being investigated, and the gradual evolution of a correspondingly diverse repertoire of tests that assay characteristics distinct to each organism *e.g.* pneumococcal surface antigens in *S. pneumoniae* and the *spa* gene in *S. aureus*. The one unifying feature of all the organisms investigated in a microbiology laboratory is, of course, their possession of a genome. Virus, bacteria, parasite or fungus, the characteristic and distinguishing features and behaviours of all pathogens (including their interactions with human host and wider environment) are ultimately determined by the sequence of their genome.

Sequencing the whole genome of a pathogen could therefore enable microbiologists not only to predict and classify the complete repertoire of structural and functional characteristics of an organism, but also to determine with the highest possible resolution the relatedness of isolates of the same organism from different sources. Thus, whole genome sequencing (WGS) has the potential to simplify the currently complex landscape of microbiological methods by replacing many (if not all) of them with a single assay, albeit one whose analysis must be adapted to fulfil the requirements of each different pathogen and epidemiological or clinical application. In the following section we describe evidence from the published literature that supports the potential utility of using whole genome sequencing for the surveillance and control of infection. We then identify and discuss the significant gaps that remain to be bridged between these examples of the utility of WGS in principle and the realisation of its effectiveness in practice.

Box 7.3 Challenges of current microbiological typing methods

Number of different assays

Public health microbiology laboratories currently need to maintain the equipment, reagents and scientific expertise to perform and analyse a vast array of different phenotypic and genotypic assays in order to undertake surveillance of the diverse range of organisms that can cause infections. This has significant negative implications for cost, resilience and utility of public health microbiology services.

Insufficiently discriminatory

Even the most advanced genotypic methods such as MLVA, PFGE and MLST remain, on some occasions, insufficiently discriminatory to enable effective epidemiological investigation and curtailment of infectious disease outbreaks of public health significance. This is mainly due to the fact that they assay only a small proportion of the total variation between the genomes of bacterial strains, and thus if epidemiologically or clinically significant differences in bacterial genomes exist outside the genomic regions covered by these methods, they will not be detected by these assays.

7.5 Case study – pneumococcal disease vaccine and drug susceptibility surveillance

Pneumococcal disease – the facts

- Invasive pneumococcal disease is caused by infection with the bacterium *Streptococcus pneumoniae.* (*S. pneumoniae*) The illnesses caused by *S. pneumoniae* range from mild sinusitis or otitis media to invasive bloodstream infections, pneumonia and meningitis

- There are over 90 strains of *S. pneumoniae*, distinguished by their serotype, which vary widely in their prevalence and their propensity to cause invasive pneumococcal infections

- Whilst a significant proportion of the population (approximately 25%) are carriers of this bacteria, only a small percentage of these carriers ever become infected and unwell

- Pneumococcal infection is one of the most common causes of invasive bacterial infection in children, and is a significant cause of morbidity and mortality in very young, elderly and immunocompromised populations

- Ongoing surveillance indicates that there are between 5000-6000 cases of invasive pneumococcal disease reported in the UK each year

- Pneumococcal disease can be treated successfully with antibiotics, but there is an increased incidence of pneumococcal infections that are resistant to at least one commonly used antibiotic

- The current pneumococcal vaccination programme targets only 13 of the >90 strains of *S. pneumoniae*

Vaccine effectiveness surveillance – pneumococcal vaccine escape

The UK national pneumococcal surveillance scheme monitors the effectiveness of the current vaccination programme by cataloguing all cases of invasive pneumococcal disease and serotyping them to determine which strain of *S. pneumoniae* was responsible for the infection. Monitoring the prevalence of infections caused by different strains enables epidemiologists to determine the effectiveness of the vaccine by comparing the incidence of invasive infections caused by vaccine targeted serotypes before and after the introduction of the vaccine. Such surveillance has revealed, however, that whilst pneumococcal vaccines have been highly effective in reducing the incidence of infections by the strains that they target, the overall rate of incidence of invasive pneumococcal disease has not reduced significantly as other non vaccine-targeted strains have expanded in numbers to 'fill the gap' and vaccine-targeted strains appear to be exchanging genes with non-targeted strains that allow them to escape the immunity provided by the vaccine.

Genomic analysis of the effect of vaccines on pneumococcal disease

Low resolution information derived from serotype analysis limits the ability of public health microbiologists and epidemiologists to monitor the causes and dynamics of these changes in the population of *S. pneumoniae* strains, to predict their longer term impact on the effectiveness of current vaccines and to guide the development of new vaccines. This is further complicated by the characteristic ability of different *S. pneumoniae* strains to undergo horizontal gene transfer, enabling the rapid transfer of genes (including those that determine the serotype) from one strain to another and facilitating vaccine escape.

Whole genome sequencing offers a way to overcome the limitations by providing a complete picture of the genomic processes that underlie the escape of strains from vaccine coverage, and the proliferation of non vaccine targeted strains. Several research groups have recently undertaken large scale sequencing projects to compare the genomes of multiple *S. pneumoniae* isolates and study the effects of vaccination on the genomic diversity of *S. pneumoniae* strains[101,102,103,104]. Their results demonstrate that comparative whole genome analysis can effectively detect and dissect the multiple genomic events underlying the apparent escape of some *S. pneumoniae* strains from the effects of the vaccination programme, something which has not been possible with current serotyping methods.

Utility and future prospects

Whilst research studies convincingly demonstrate that whole genome sequencing offers higher resolution analysis of *S. pneumoniae* population dynamics than current serotyping methods, its utility in either real-time or retrospective real world public health surveillance scenarios remains to be determined. The utility of this methodology could arise from:

- Generation of information that provides predictive insights that can be used to guide the development of the pneumococcal vaccination policy in the UK

- Replacement of costly and time consuming serotyping with a cheaper and quicker assay that enables a more informative, higher resolution classification of strain diversity

- Improved monitoring and understanding of the spread of antibiotic resistance between strains

7.6 Genomic epidemiology for outbreak investigation and control

The mutation of bacterial and viral genomes is perhaps their most powerful weapon in the ongoing battle to adapt to their changing environments, evade our immune systems and resist our antibiotics. The detection of mutations by whole genome sequencing promises, however, to enable microbiologists and epidemiologists to turn this weapon against the pathogen, using genomic epidemiology to investigate infectious disease outbreaks and, ultimately, intervene to halt their progress.

7.6.1 What is an outbreak and how are they investigated

The WHO defines disease outbreaks as:

- The occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season. An outbreak may occur in a restricted geographical area, or may extend over several countries. It may last for a few days or weeks, or for several years

- A single case of a communicable disease long absent from a population, or caused by an agent (*e.g.* bacterium or virus) not previously recognised in that community or area, or the emergence of a previously unknown disease, may also constitute an outbreak and should be reported and investigated

One of the principal methods by which microbiologists and epidemiologists involved in outbreak investigation determine whether there is indeed an outbreak, and if so how the infection is being transmitted, is to compare the characteristics of the pathogens causing the infection in each individual to determine how closely related they are to one another. This is typically done using the phenotypic or genotypic typing methods described above.

Where the infections causing apparently similar illnesses linked in time or place are found to be caused by distinct species or strains of pathogen, this is evidence that the cases are merely coincident and not linked by person to person transmission of a single pathogen or acquisition of that pathogen from a common source. If, however, microbiological investigation shows that the cluster of patients are infected with the same species and strain of pathogen, this is strongly suggestive that their infections are indeed linked by chains of person to person transmission or a shared common source. Combining this microbiological data with behavioural and geographic data can then be used to attempt to elucidate sources of the infection and identify individual transmission events within the cluster.

There are two key limitations to this approach:

- For some infections there is a single dominant strain of pathogen that is highly prevalent within the population *e.g.* MRSA Clone x. Thus, most cases of infection are indistinguishable from one another and so determining whether an increase in the incidence of an infection is random or linked to a particular source that can be controlled is not possible

- By definition, cases related by person to person transmission or a common source of infection are likely to be caused by pathogens that are extremely similar. Whilst detecting this similarity is useful for delineating the boundaries of an outbreak, the inability of current methods to resolve small differences between the pathogens within this highly similar cluster of infections means that identifying the source and pattern of transmission is often impossible, as all cases appear identical

7.6.2 What is pathogen genomic epidemiology?

The aim of pathogen genomic epidemiology is to use comparative analysis of the genomes of pathogens isolated from patients suspected to be part of an outbreak, in combination with other epidemiological data, to attempt to determine whether patients are indeed part of an outbreak and if so to establish its source/s and the chain of transmission between patients and any other environmental reservoirs of the infection.

The method used to achieve this is to sequence the whole genomes of pathogens taken from different patients and different places, potentially at different times, and use the number of differences identified between the genomes to construct 'family trees'. These family trees are constructed on the principle that:

'*The extent of sequence variation between the genomes of pathogens isolated from different people or locations in the environment is proportional to how closely related the pathogens are* i.e. *how recently they share a common ancestor.*'

Thus, isolates of the pathogen that have identical or near identical genomes will be placed close together on these trees, and it can be inferred that these infected individuals are likely to have been exposed to the same source of the infection *e.g.* a common foodstuff, or to be a transmission pair *i.e.* one individual has infected the other. Where isolates of the pathogen have genomes that differ widely in their sequence they will be placed further apart on the family tree and epidemiologists can infer that it is unlikely that these infections were directly transmitted between these individuals and that they are also unlikely to share a common source (such as a third person or location in the environment).

Pathogen genomic epidemiology uses comparative analysis of the genomes of pathogens isolated from patients suspected to be part of an outbreak... to attempt to determine whether patients are part of an outbreak and if so, the source and chain of transmission of the outbreak.

7.6.3 What are the advantages over existing epidemiological approaches to outbreak investigation and control?

Current molecular epidemiological investigation of outbreaks depends on the genotypic strain typing approaches described in previous sections to cluster patients based on similarity of the patients causing their infections, and also to delineate the boundaries between outbreaks and sporadic 'background' cases of a disease. As with longitudinal surveillance, the advantage of genomic outbreak epidemiology is that it provides significantly greater resolution of the characteristics of pathogens causing infections and hence higher resolution discrimination of the differences between similar pathogens infecting different individuals. Benefits of this additional resolution include:

- Enabling epidemiologists to determine how closely or distantly related infections belonging to a suspected outbreak cluster are, even where those cases appear identical by existing typing methods, and thus to determine the most likely sources and paths of transmission between patients

- Deduction of information about the emergence of the outbreak strain, for example where the unusual recombination of genetic features from several different strains has occurred

- Identifying genomic sequences unique to the outbreak strain that can be used to develop more rapid molecular assays to detect subsequent outbreak related cases

- The discovery of genomic determinants of virulence and drug resistance that can guide clinical management of individual cases

The majority of studies investigating the feasibility of using whole genome sequencing to delineate and investigate outbreaks of infectious disease have been retrospective in nature. Examples include the high resolution characterisation of tuberculosis outbreaks in Canada and Germany and the UK, retrospective investigation of an outbreak of MRSA in a neonatal intensive care unit in the UK, an investigation of the diverse sources of *C. difficile* infection in the UK and the investigation of an outbreak of drug resistant *Klebsiella pneumoniae* in a hospital in the USA[105,106,107, 108,109]. Whilst these studies, and many others, have now demonstrated the utility of genomic epidemiology in principle, they do not provide evidence to demonstrate whether use of information derived from these investigations, had it been available in 'real time' during the outbreak could have led to significantly improved outcomes for patients.

Below we summarise two studies that have used genomic epidemiology in prospective context to guide infection control and curtail outbreaks, and which suggest that this approach is feasible and useful, at least in well-resourced healthcare facilities.

7.7 Case study - MRSA outbreak in a special care baby unit (SCBU)

Background

Routine surveillance at a special care baby unit within a Cambridge (UK) hospital revealed a suspected ongoing outbreak of methicillin resistant *Staphylococcus aureus* comprising up to 17 cases over a six month period. Standard infection control investigations, using phenotypic methods to determine how similar the bacteria isolated from each patient were, suggested that at least 15 of these cases were linked as they had identical antibiotic susceptibility profiles. Nevertheless, infection control specialists remained puzzled as to how these 15 cases could be linked, as they occurred in three temporally distinct clusters, with no overlap in the time spent by the patients on the ward between these clusters. Antibiotic susceptibility is a very low resolution typing method and so it remained possible that whilst the isolates from these 15 patients appeared identical, suggesting they were part of an outbreak, that they were in fact distinct and unrelated strains that happened to have the same antibiotic susceptibility.

Given this limited information, a deep clean of the ward was undertaken to attempt to eliminate any potential environmental source of infection and surveillance was continued.

Confirming and elucidating an outbreak by whole genome sequencing

The infection control team also asked a research group to undertake whole genome sequencing to provide a higher resolution analysis of the putative outbreak. This WGS-based investigation made several important discoveries. They found that the fifteen apparently related isolates were indeed almost identical, differing by at most 25 single nucleotides out of a genome of several million, confirming the suspicion of the clinical infection control specialists that they were part of an outbreak. They also found an MRSA infected patient that had been ruled out of the outbreak, on the basis of non-concordant antibiotic susceptibility testing that on re-evaluation was found to be incorrect, had in fact also been infected by the same strain as the other outbreak cases. Crucially, they identified ten other cases of MRSA from outside the SCBU that were infected with the same outbreak strain and that could be plausibly linked back to the original SCBU outbreak cases. Thus WGS was able both to confirm the occurrence of an outbreak, but also to more accurately delineate the extent of the outbreak than had been possible with standard typing and epidemiological approaches.

Outbreak resolution

Whilst the confirmation and delineation of the outbreak was in itself useful, it remained to be established how transmission continued to occur even where there was no overlap between infected patients in the ward. Furthermore, the analysis of the relatedness of the different infections that were studied by whole genome sequencing did not support the hypothesis that the infections were being transmitted step wise from one patient to the next, as there was no clear correlation between the time of infection and how closely related the isolates were.

Solving this puzzle became more pressing as further cases, also infected with the outbreak strain, were detected on the SCBU even after the deep clean had been completed and, in one case, 64 days after the previous MRSA positive patient was present on the ward. This led the researchers and infection control team to suspect a healthcare worker might have been responsible for transmitting the infection, as patient-to-patient and environment-to-patient modes of transmission appeared to have been ruled out. Screening of healthcare workers on the SCBU identified one MRSA positive member of staff. Subsequent whole genome sequencing of the MRSA isolated from the staff member revealed that they were carrying the same strain that was responsible for the outbreak, and that in particular their infection was genetically almost identical to that of the patients infected either side of the 64 day gap following the deep clean. This lead the infection control and research teams to conclude that this staff member may have been responsible for the repeated reintroduction of MRSA into the SCBU, and indeed, following their decolonisation, no further cases of infection with the outbreak strain were detected.

7.8 Case study - *Pseudomonas aeruginosa* outbreak in a burns unit.

Background

In this section we describe the use of whole genome sequencing (WGS) to determine the source of a *Pseudomonas aeruginosa (P. aeruginosa*) outbreak in the burns unit of a recently opened hospital. *P. aeruginosa* is a gram negative bacterium that is a common cause of hospital acquired infections. It thrives in moist conditions and commonly causes infections on mucosal surfaces in the body such as airways, the urinary tract and skin. *P. aeruginosa* infection can cause inflammation and sepsis, such as: bloodstream infections in premature neonates; respiratory tract infections in those with cystic fibrosis and patients who have been mechanically ventilated; urinary tract infections in patients with catheters; and skin infections in burns patients.

The bacterium can be found in biofilms on wet surfaces, commonly in the water system, and outbreaks of infection in hospitals have been associated with plumbing components and associated water sources such as taps, sinks, showers and drains. Infection can also be spread on people's hands and on contaminated medical equipment, such as endoscopes. *P. aeruginosa* is a challenging pathogen to deal with as it is adaptable to a wide range of environments meaning that extensive measures, such as deep cleaning and disinfection, are required to remove it from environmental sources. The bacterium is also naturally resistant to many antibiotics, possessing a range of efflux pumps and antibiotic inactivating enzymes, which limits options for treatment.

A high profile outbreak of *P. aeruginosa* occurred in 2012 in a neonatal unit in Belfast, in which three babies died. The source of the infection was determined to be sink taps within the unit[110], and as a result national guidelines were introduced by the Department of Health outlining the procedures for enhanced water sampling on units that care for patients vulnerable to *P. aeruginosa* infection, and also details of deep cleaning procedures and replacement of high-risk plumbing parts[111].

If *P. aeruginosa* infection is suspected, samples are taken from wounds, urine or stool and tested in the microbiology laboratory. Samples are cultured overnight on agar plates or in broth, before genotyping and antibiotic sensitivity assays are carried out. Current genotyping methods such as variable number tandem repeat analysis or multi-locus sequencing typing only sample a limited number of sites in the genome and can result in unrelated cases being clustered together due to the limited resolution of the methods. Antibiotic sensitivity assays are carried out using automated systems with results available in a few hours.

If a patient is found to be infected with *P. aeruginosa*, additional infection control measures are put in place, assessed according to the patient's needs and treatment regime. This may include deep cleaning of the environment or sections of plumbing, or assessments of water use in the ward.

While these measures can be extremely effective in reducing infections in hospitals, conventional typing methods may have not provided enough information to enable the health system to pinpoint and deal with the source of the infection. Whole genome sequencing can benefit infection control by providing much higher resolution information and is able to distinguish between two isolates of a bacteria down to single nucleotide resolution. This means that health providers can track transmission, determining whether patients were infected by the environment, each other, or an external source.

Using whole genome sequencing for outbreak surveillance

One key question for the health system is to assess the effectiveness of whole genome sequencing to investigate outbreaks in a hospital setting, and to find the source of any outbreaks. Such an observational study was carried out in the burns unit of a recently opened hospital[112].

Burns patients are particularly vulnerable to infection with *P. aeruginosa*, given that hydrotherapy is a mainstay of burns treatment and infection risk is high if the pathogen is present in the water supply or plumbing. It has been observed that up to one-third of burns patients can become infected with *P. aeruginosa*.

Patients recruited to the study were screened on admission for carriage of *P. aeruginosa*, and samples were regularly taken from them and their environment for standard microbiology tests during their stay. If a patient became colonised with *P. aeruginosa* during their stay, they were subjected to more enhanced monitoring where patient and water / environment samples were taken more regularly. These samples were assessed using standard microbiological techniques, and whole genome sequencing.

P. aeruginosa was detected in five patients, of these three were infected with *P. aeruginosa* that had the same genotype as that found in the water and plumbing in the rooms they had been nursed in. In the other two patients, *P. aeruginosa* was only found in their rooms during their stays, but not before or after their time in the hospital, nor were their isolates detected in the water supply. This suggested that they had carried the bacteria with them into the hospital, and had not transmitted it during their stay.

The whole genome sequence information, in conjunction with epidemiological data, allowed the research team to determine that the hydrotherapy showers were responsible for three of the infections, and could also demonstrate that transmission was unidirectional, from the showers to the patients.

Response to the outbreak

In response to these findings, the hospital started enhanced infection control measures, in line with the guidelines from the Department of Health. The measures included additional cleaning of the ward environment and hydrotherapy showers which were infected, including the installation of new filtered water systems in the highest risk water outlets on the ward.

Following these interventions, no further *P. aeruginosa* infections were detected during the study period, however a direct causal link cannot be made, partly because of interventions that had been put in place due to an unrelated infection with a different pathogen, of different patients in another part of the ward.

Case study - *Pseudomonas aeruginosa* outbreak in a burns unit.

This study demonstrates how using WGS for surveillance in a frontline hospital setting supports real improvements in infection control. The use of WGS in this case demonstrated unidirectional transmission from the water supply, via plumbing, to patients. The resolution of the method allowed the clinical team to determine which sections of plumbing infected which patient, and also allowed them to demonstrate when transmission did not take place, in the form of two patients who brought *P. aeruginosa* with them into the unit but did not spread it any further. Targeted infection control efforts such as deep cleaning and replacement of colonised plumbing probably played a role in preventing further transmission of infection.

Utility and future prospects

The case studies above, along with many other published retrospective studies, have clearly demonstrated the contribution that genomic epidemiology can make, in principle, to the detection, investigation and even resolution of infectious disease outbreaks. The scientific principles on which this technique is founded have been rigorously tested, and repeatedly shown to yield useful results. What remains to be established is whether in practice genomic epidemiology will have a significant impact on the management of infectious disease when implemented widely as part of standard practice in the UK. This will depend on a host of factors explored in depth in Part III of this report.

8 Surveillance and control of emerging infectious diseases

Currently under-utilised in the context of emerging infectious diseases, whole genome sequencing has several potential uses in the surveillance of many globally significant diseases. Below we highlight examples of where whole genome sequencing has made a difference in the prevention and control of outbreaks.

8.1 Introduction

Emerging infectious diseases (EIDs) are of importance to UK and global health, and have been responsible for many hundreds of thousands of deaths within the UK alone over the past century. An EID can be defined as an infectious disease that shows a substantial and recent increase in prevalence and / or virulence (box 8.1). This epidemiological definition effectively equates EIDs with pandemic and large-scale outbreaks, rather than regular seasonal and sporadic incidences of infectious diseases. The critical distinction between these EIDs and other important infectious diseases, such as those described in the previous chapters, is that EIDs (as defined here) have a relatively low probability of arising and spreading within a given population but a high health cost if they do. There is an emphasis, therefore, on prevention and preparedness for EIDs rather than curing the diseases.

Box 8.1 Definition of EIDs

An emerging infectious disease (EID) is defined in the context of this report as a disease that shows substantial and recent increase in prevalence and / or virulence. Although the words 'substantial' and 'recent' are ambiguous and leave a grey area, it would be unsuitable to impose arbitrary cut-off thresholds to rigorously define these parameters and unnecessary to introduce a more sophisticated model to classify infectious diseases. It is an oversimplification to put infectious diseases into these binary categories of emerging or non-emerging as in reality these diseases fall on a multi-dimensional spectrum depending on many factors. Whilst acknowledging these caveats, this simple classification is sufficient to distinguish major new infectious disease threats from more established and stable ones that are discussed elsewhere (chapter 7).

The emergence of an EID is likely to be caused by large genetic and / or environmental shifts that lead to the increases in the disease's prevalence and / or virulence. An estimated 75% of EIDs initially involve the transfer of a pathogen to humans from another animal[113], and such transitions can be facilitated by underlying sequence changes in the pathogen's genomes.

Given the interconnectedness of the modern world, the risk of importation and transmission of EIDs is raised and the threat they pose to human health is significant. Past EID pandemics have led to many fatalities. The most notable example being the 1918 influenza pandemic, sometimes dubbed 'Spanish flu', which led to the death of at least 20 million and possibly nearer 100 million people[18,114], similar to the number of people who died during World War II. The major types of pathogens that are current emerging diseases threats for the UK are, numbered based on their importance: pandemic influenza caused by strains of influenza A; the Ebola virus underlying the recent Ebola outbreak; and coronoaviruses such as Middle East respiratory syndrome coronovirus (MERS) and severe acute respiratory syndrome (SARS).

8.2 The roles of genomic technologies in tackling EIDs

8.2.1 Genome sequencing as a tool to trace the spread of EIDs and help contain outbreaks

As discussed in chapter 7, whole genome sequencing can be a valuable tool to help reduce the risk of outbreaks occurring and to reduce the spread of disease during outbreaks. Genome sequencing has several different potential utilities in the context of EIDs, which can be classified into four distinct categories:

1. To trace the source of an outbreak, which may be human, or more distantly an animal reservoir of infection or the source of contamination in food products. Such information can be used to minimise the risk of future outbreaks and prevent repeated reintroductions from the same source

2. To identify a novel emerging human pathogen, using an unbiased metagenomic approach to sequencing, where existing diagnostic tests provide no result to explain the occurrence of an apparently infectious disease

3. To identify patterns of drug (antiviral or antibiotic) resistance that can inform drug choice in managing cases within the outbreak , particularly for multi-drug resistant organisms where conventional tests may be insufficient

4. To prospectively identify emerging strains of virus that may pose threats to human health, enabling anticipatory development and stockpiling of appropriate vaccines and suitable preparedness planning by public health authorities

The same generic approaches to genome sequencing and analysis described in chapter 3 and chapter 4 can be applied to EIDs as to more common pathogens. However, the analysis of emerging pathogens may be complicated compared to more common ones as there is likely to be a relative paucity of genomic information, such as high quality reference genomes against which to map outbreak strains, available for the former.

8.3 Case studies of genome sequencing used in an emerging infectious disease outbreak context

Swine flu

It is notable that all the major EID threats to the UK identified above are caused by viruses, rather than bacteria or other microorganisms. Furthermore, all the viruses are single-stranded RNA viruses. RNA viruses may be particularly likely to be involved in outbreaks because they evolve especially rapidly due to the absence of the DNA polymerase proofreading enzymes that leads to more replication errors (chapter 2). Influenza A is the type of influenza associated with epidemic and pandemic outbreaks, rather than types B or C. The influenza A genome is approximately 13,500 nucleotides (nt) in length and consists of eight RNA segments that produce eleven different proteins. Two critical segments are segment 4 that encodes the hemagglutinin (H) protein and segment 6 that encodes the neuraminidase (N) protein. Reassortments of these eight segments between different virus strains is the major way by which the virus is able to evolve so rapidly and create novel strains, and the naming of the strains is based on the particular H and N proteins that are encoded. During the 2009 'swine flu' influenza A pandemic samples were collected from humans and other animals. Whole genome sequencing of these viruses allowed the reconstruction of events that led to the emergence of the H1N1 strain that caused the pandemic[115]. While these results were only available retrospectively, meaning they could not be used to inform the public health management of the pandemic itself, by suggesting that potentially pandemic strains of the virus were circulating in pig populations for many years before the 2009 outbreak they provide crucial intelligence on how public health practitioners and policy makers could minimise risk of such an event occurring again. In particular, they provide evidence to support the pre-emptive monitoring of the swine populations for emerging influenza A strains that have the potential to be human pathogens. Furthermore, should such a potential threat be identified prior to an outbreak, measures such as regulations on livestock movement could be put in place to help prevent transmission of the virus to humans, and thus reduce the risk of future pandemics occurring.

Ebola

There are several different strains of Ebola virus, but all their genomes are approximately 19,000nt long and contain seven different protein coding genes enclosed within the untranslated regions. During the ongoing Ebola outbreak a consortium of scientists has sequenced 99 Ebola virus genomes from 78 patients in Sierra Leone[116]. Analyses of these sequences identify distinct genetic changes to this Ebola virus outbreak and suggest that the virus spread over the last decade via animal hosts (probably bats) from Central Africa. Furthermore, they found that the outbreak in Sierra Leone was triggered by two different viral strains. It is important to note that despite being undertaken in 'real time' *i.e.* during the ongoing outbreak, this study provides information that will mostly be of use to help prevent future outbreaks, perhaps through enhanced surveillance and support the future development of Ebola treatments and vaccines. Lack of understanding of the epidemiological origins and patterns of spread are not at this stage limiting factors in the management of this outbreak, rather it is resource limitations on providing adequate infection control measures and clinical care that are the major barriers to effectively resolving the outbreak.

Coronaviruses

Both the SARS and MERS coronavirus genomes are both approximately 29,000nt, but the MERS virus genome contains slightly fewer genes. Genome sequencing of the MERS coronavirus from a human patient who died from MERS, and sequencing of viral samples from the patient's camels. showed that the full genome sequence of viruses were identical between the patient and the camel samples[117]. These results, combined with more recent evidence that the human virus can cause MERS in camels[118], suggest that MERS is transmittable between camels and humans. The concentrations of the virus in the different patient and camel samples further implies that transmission occurred from camel to human. Therefore, monitoring of camels and their viruses could help identify MERS-like viruses that are potentially dangerous to humans. In cases where potential human threats are found in camel populations, limiting contact between camels and humans might be beneficial.

8.4 Conclusions

The case studies described above demonstrate the potential utility of genomics both for the prospective surveillance of EIDs, aimed at early detection and outbreak prevention, and also for the responsive investigation of ongoing outbreaks of EIDs. Given that so many of these infections originate in animal populations before being transferred to humans, there is a case to be made for undertaking prospective genomic surveillance of these infections not only in humans, but also in the animals in which they are most likely to arise, particularly those used in the food industry.

Surveillance and management of EIDs is predominantly the responsibility of national and international public health authorities rather than local health services. Given the potentially severe impact they are known to have on population health - despite their rarity - there is a strong case for ensuring that public health authorities in England have the capability to use genomics for both surveillance and outbreak investigation of these potentially devastating pathogens.

9 Wider roles of microbial and host genomics in the management of infectious disease

In this section we briefly capture some of the emerging and more indirect applications of pathogen genomics to infectious disease management.

9.1 Introduction

Whilst this report focuses on the direct application of genomics to the management of bacterial and viral infections, it is important to acknowledge that the impact of genomics on the management of infectious disease will, in the future, extend far beyond these pathogens and beyond the direct applications to clinical and public health microbiology described above.

9.2 Genomics of fungal and parasitic infections

Many fungi, including those that cause human disease, have had their genomes sequenced. However, the use of this information in understanding and managing fungal infections is not as developed as for bacterial and viral infections. This is partly due to the more complex physiology and biochemistry of fungi presenting a greater challenge in terms of translating knowledge of their genomes into insights into their function and behaviour as pathogens.

Most fungal infections are also opportunistic, and thus host (*i.e.* human) factors, such as genetic susceptibility or acquired immune deficiency as a result of chemotherapy or HIV infection, affect to a great extent whether a fungal infection transforms into a clinically significant disease. While fungal infections are on the increase, due to increased numbers of immunocompromised patients, they also do not present as great a disease burden as viral or bacterial infections, and so pressure to develop new treatments is not as great. One of the best understood fungi in the context of human disease is the yeast *Candida albicans* (and related species), the cause of candidiasis in humans. Genomic research on *C. albicans* is currently focused on understanding the biology of the organism, finding potential therapeutic targets, and understanding how it causes disease in humans.

The malaria parasite is the best studied of the parasitic infections and was the first to have its genome sequenced, in 2002[119]. Many genomic studies have been carried out to better understand the parasite's biology, with a focus on understanding resistance and susceptibility to therapies. While initiatives exist to sequence the genomes of other disease causing parasites and understand their biology, this research is less advanced and progress is slow due to a combination of factors including poor funding and complex biology. This is despite acknowledgement that many parasitic infections represent significant areas of unmet global health need – of the seventeen neglected tropical diseases listed by the WHO, eleven are caused by protozoa or helminths[120]. There is currently an international project underway, coordinated by the Wellcome Trust Sanger Institute, to produce reference genomes for the 50 most common helminth infections. These genomes are already being made available to researchers to drive forward understanding of these diseases and development of new treatment approaches.

While initiatives exist to sequence the genomes of other disease causing parasites and understand their biology, this research is less advanced and progress is slow due to a combination of factors including poor funding and complex biology. This is despite many parasitic infections representing significant areas of unmet global health need. Of the seventeen neglected tropical diseases listed by the WHO, eleven are caused by protozoa or helminths.

While there are many valuable research initiatives underway to understand the genomics of these other important human pathogens, it is unclear when and how the results of these studies will be translated into clinical or public health interventions. Consequently they are not considered further in this report.

9.3 Host-pathogen interactions, personalisation of vaccines and therapies

The manifestation of infectious disease is not only a function of the behaviour of the pathogen causing the disease, but also the underlying physiology, and in particular immunology, of the human that it infects. Human genomics has also, therefore, the potential to inform infectious disease management by illuminating our understanding of how each person's genomic variation affects their response to the pathogen, and indeed to any vaccine or drug used to prevent or treat infections. For example, infection rates for many pathogens tend to be much higher than observed disease rates, suggesting that the population susceptibility to the effects of infection, which are governed by host immunity and ultimately host genomics, is highly variable.

The genetic architecture of this susceptibility remains, however, to be fully determined. Most likely there is a spectrum of genetic impact, ranging from rare, highly penetrant single mutations that result in severe immunodeficiency and increased susceptibility to infection, through to more common variations that in aggregate may modulate susceptibility to infection to a much less dramatic extent. While information on the genetic susceptibility to infection is potentially very informative - for example resistance to HIV infection can in some cases be attributed to a mutation in a receptor protein (CCR5) that the virus uses to enter immune cells - there are as yet few examples of where this information has a clear impact on clinical and public health practice. In the case of CCR5-mediated HIV resistance, this knowledge has been used to develop experimental treatments but the genetic information has not yet resulted in a breakthrough treatment or vaccine. Another disease where host genetic variation has been shown to influence response to infection is dengue virus – mutations in MHC complex proteins have been shown to confer susceptibility to dengue shock syndrome, the most severe form of the disease[121]. However this research is still at an early stage and the functional basis of these mutations is currently not known.

One disease where host genetics has influenced clinical practice is hepatitis C, where variants of the *IL28B* gene lead to different clinical outcomes in patients with hepatitis C infections. Those patients with two copies of the 'C' gene variant have better response to therapy and are also more likely to spontaneously clear their infection, information which is already being used for diagnostic decisions[122, 123].

In conclusion, while understanding of the human genomic contribution to the manifestations of infectious disease should in principle enable the personalisation of risk prediction and of preventive or therapeutic interventions, these benefits remain largely hypothetical and thus are far from being realised in clinical and public health practice.

9.4 Reverse vaccinology

This approach uses genomic and bioinformatic approaches to identify candidate components for vaccines. Reverse vaccinology was first applied to the development of a vaccine to protect against infection with meningococcus B, which causes bacterial meningitis. The development of vaccines for this particular bacterium was complicated by the polysaccharide outer shell of the bacterium (a common vaccine target) being identical to that of a polysaccharide found on human cells. It was also unclear which of the other surface molecules (antigens) on the bacterium would be suitable for use in a vaccine instead. Following sequencing of the *Meningococcus B* genome, bioinformatic methods were used to scan the genome for potential antigens, and to identify those that were sufficiently different to human proteins for testing as vaccine candidates. The most promising of these were then used in prototype vaccines and ultimately, following further refinements and clinical trials, a new meningitis B vaccine was produced that is now licensed for use within the EU, which the NHS in England has introduced into the childhood vaccination programme, available from September 2015.

9.5 Drug development and synthetic biology

The use of genomic analysis of the genomes of both pathogenic and non-pathogenic microorganisms can, in principle, impact on drug development in two distinct ways:

- Enabling the identification of new candidate targets for antibiotic drugs or antibiotic molecules themselves

- Understanding biosynthetic pathways within bacteria and fungi to enable re-engineering of these pathways to produce drug molecules, so called 'synthetic biology'

Fungi are considered to be a rich source of potential new therapeutics, and drugs developed so far from fungi include antibiotics, immunosuppressants, anti-fungals, chemotherapy drugs and statins. Genomic technology has increased the reach of drug discovery programmes in fungi by highlighting genes whose functions will result in the production of potentially therapeutic molecules, but are rarely observed under common laboratory conditions.

Using genomic data from bacteria to identify potential targets against which to design novel antibiotics remains however a significant challenge, in particular because of the difficulties of predicting function and expression of proteins directly from their genes, and has yet to yield any notable successes.

The use of genomic technology to synthesise new drugs has also proved to be challenging as this requires not only an understanding of the biochemical pathways needed to produce the drug, but also an understanding of the organism that could be used to produce it. The earliest and best known example is the genetic modification of *E. coli* by insertion of the human insulin gene, a process that has been used since 1982 to produce insulin to treat diabetes. In 2013 large-scale production started of artemisinic acid, a precursor of the anti-malarial drug artemisinin, using genetically modified yeast[124]. Artemisinic acid is currently extracted from the sweet wormwood plant, meaning that supplies can be unpredictable, disrupting the availability of drugs to those in greatest need. While this is a good example of using genomic knowledge and technology to produce a much needed drug, development was time-consuming due to the scale of the research effort needed to refine the process in order to make it suitable for large scale production. This included taking a whole organism approach to fully understand the synthesis process within the wormwood plant, and how the inserted genes operated in the yeast, which was first reported in 2006[125].

Thus, while it is clear that greater understanding of the genomes of both pathogenic and non-pathogenic organisms should, in principle, catalyse efforts to create new therapeutics to treat infectious diseases, this aim is far from being realised in practice in all but a few cases.

10 How could pathogen genomics contribute to the management of infectious disease?

In Part II of this report we asked whether genomic technology and knowledge can *currently* address the microbiological questions central to infectious disease management in a way that is sufficiently superior to existing methods (as measured by reduced cost, increased speed, increased sensitivity or specificity) to warrant its introduction into mainstream clinical and public health microbiological practice at this time. Below we set out briefly the answers to this question arising from the research and analysis presented in the preceding chapters.

10.1 Genomics cannot currently improve upon existing diagnostic microbiological investigations (except for *M. tuberculosis)*

In principle the use of genomic information could contribute more to improving the effectiveness of current diagnostic investigations (identification and AST) of bacteria than viruses. This is largely due to clinical diagnostic virology having already transitioned to molecular methods of virus identification and drug susceptibility testing, whereas diagnostic bacteriology continues to rely more on phenotypic methods for these investigations and so has more to gain from the increased discriminatory power of genomics. However, in practice whole genome sequencing cannot in its current form compete in either case with existing methods for the purposes of pathogen identification and AST. This is largely because the cost and time taken to obtain genomic information is significantly greater than that associated with current testing for viral and bacterial pathogens. As noted in chapter 5 and chapter 6, the one clear exception to this 'rule' is *M. tuberculosis*, for which phenotypic AST is significantly slower than AST using whole genome sequencing.

10.2 Genomics could in principle significantly improve on existing infectious disease surveillance and outbreak control investigations

Chapter 7 sets out the theoretical advantages of genomic methods over existing typing methods for the investigation of infectious disease outbreaks, and for the performance of disease surveillance. It also presents published examples of how these advantages have manifested themselves in practice, albeit in research-based and predominantly retrospective settings. We conclude on the basis of these that there is a strong case **in principle** to pursue the clinical and public health implementation of pathogen whole genome sequencing as a tool for outbreak control and surveillance, as part of a wider system of infectious disease management. However, evaluation of pathogen whole genome sequencing against the criteria described earlier for an optimal typing test with which to investigate outbreaks or undertake surveillance (table 10.1) highlights that there are still many challenges to be overcome in turning this promising new technology into a tool that can be used reliably and routinely as part of clinical and public health microbiological practice. Furthermore, there are significant challenges to ensuring that the wider health systems responsible for managing infectious disease are appropriately configured, resourced and developed to realise the benefits of pathogen WGS. We highlight how at the present time WGS performs in a clinical settings if assessed against the requirements for an effective typing test (table 10.1), but acknowledging that many existing tests do not necessarily meet these 'ideal' standards (chapter 7). A coordinated national strategy for pathogen genomics, as outlined in the following chapters, will facilitate the necessary improvements and developments needed for WGS to meet if not exceed necessary quality standards for an effective test and enable it to outperform existing methods.

10.3 Conclusions

Pursuing the implementation of pathogen whole genome sequencing into clinical and public health microbiology services can, given the current state of technology, be justified for the purposes of outbreak control and surveillance, and exceptionally the diagnostic management of *M. tuberculosis*. There remain, however, significant barriers to be overcome in realising the effectiveness of implementing this technology. These include the need for further method and knowledge development - to ensure that WGS-based investigations have sufficient analytical validity and clinical utility - and also the need to optimise the configuration of the health systems and services in which pathogen genomic information will be both produced and used.

Finally, while the current state of genomic technology and knowledge mean that its application to the majority of diagnostic microbiology remains impractical, a case can be made that genomic information could be of significant use in this area. Accordingly, improvements in existing genomic technologies and our underlying knowledge of pathogen genomes should continue to be actively pursued with a view to developing new genomics-based diagnostic devices and methods that are more suited as tools for diagnostic microbiology than those currently available.

Table 10.1 A comparison of WGS against criteria for useful bacterial typing tools

Requirement for an effective typing test	Current performance of WGS in a clinical context	Potential future performance of WGS in a clinical context
Discriminatory power sufficient to address the clinical or epidemiological question	High. Exceeds that of existing phenotypic and molecular tests	Expected to remain the most discriminatory test
Reproducibility over time and across different laboratories	Variable. Depends on use of consensus analytical approaches between laboratories and these may evolve with time changing results obtained	Can improve substantially if stakeholders and service providers collaborate to establish consensus standards and share best-practice (chapter 14, chapter 17)
Short assay time to minimise delays in producing actionable results in outbreak situations	Variable. Highly dependent on method to which it is being compared and the pathogen being investigated	Could improve and exceed other existing methods if adequate informatics infrastructure and analytics support are established (chapter 11, chapter 13, chapter 16)
Available at costs affordable to healthcare systems to maximise universality of use	Low. Cost of sequencing platforms and IT infrastructure remain prohibitively expensive for most laboratories	Costs could reduce, and accessibility improve if services are configured to realise economies of scale and by establishing and organising infrastructure in way that promotes efficiencies (chapter 11, chapter 14, chapter 15, and chapter 16); Also by demonstrating evidence of cost effectiveness (chapter 18, chapter 19)
Easy to perform assays that do not require highly specialised training and equipment	Low. Assays require specialised equipment and highly skilled scientists for testing and analysis	Techniques and their application could become more accessible as technologies advance, and through workforce training, and the development of automated analytical tools (chapter 3, chapter 4, and chapter 16)
Easy to interpret results that are quantitative and unambiguous	Variable. Produces quantitative and reproducible data, but interpretation may still be ambiguous due to lack of underlying knowledge of genotype-phenotype relationships or use of probabilistic phylogenetic methods	Could improve substantially, but only if efforts are undertaken to consolidate quality-controlled data and knowledge; through sharing and establishment of best practice and standards; and through continued development and refinement of analytic techniques (chapter 14, chapter 15, chapter 16, chapter 17)
Standardised classification nomenclature enabling portability of results and reports between laboratories	Low. Different laboratories using different analytical methods resulting in varying degrees of correspondence in classification systems	Can improve through coordinated and collaborative efforts at a national and ideally international level to agree and establish standard conventions and interoperable formats (chapter 14, chapter 17)

Part III

Part II of this report sets out examples of the potential utility of pathogen genomics, many of which have been supported by translational research funding from the UK government. The focus for policy makers and those tasked with delivering improvements in infectious disease management in England, must now shift towards delivering the benefits of this technology in the real world of the health system, while maintaining a focus on the continued need for translational research and development. Factors that will determine whether the benefits of pathogen genomics can be realised include:

- Navigating the current complex landscape of microbiology service provision in England
- Developing the scientific, clinical and economic evidence base on which a 'case for implementing pathogen genomics services' can be made to health service commissioners
- Configuring and commissioning pathogen genomics services that meet local clinical needs while delivering anticipated benefits for national public health surveillance
- Developing data management systems and policies required to collate, integrate and benefit from the vast quantity of genomic and clinical data that both pathogen genomics services and on-going research and development activities will generate

In Part III of the report we describe and analyse each of these in detail. We also present recommendations to support policy makers responding to the challenges presented in each case, maximising effective development and delivery of pathogen genomics informed infectious disease management services.

11 Genomics in the evolving landscape of microbiology services in England

While translational research and implementation activities are predominantly undertaken in regional or national specialist PHE laboratories collaborating with academic centres of excellence in basic and translational research, the vast majority of routine microbiology services are offered through NHS microbiology laboratories operating outside these highly specialised environments.

11.1 Introduction

The management of infectious diseases in England involves a broad range of organisations and individuals that variously have roles:

- Setting strategic policy direction *e.g.* on healthcare associated infections
- Commissioning and implementing services
- Providing frontline care to patients and populations
- Providing safe environments that minimise human exposure to pathogens
- Monitoring the effectiveness of all of these activities in the context of the continually changing landscape of infectious disease 'threats'

The Department of Health, working with Public Health England and NHS England, set the direction and priorities for the nationwide management of infectious diseases. There is, however, regional and local variation in the prevalence of different infections and the susceptibilities and exposures of populations to these infections. The need to respond to local health needs is reflected in the *Health and Social Care Act 2012*, which requires local authority Health and Wellbeing boards to work with their regional and local PHE and NHS led health services to ensure they meet the particular public health needs of their population, including with respect to infectious diseases.

The benefits of genomics will only be realised in full through the establishment of coordinated services that cut across the organisational, professional and geographical boundaries.

The emphasis on response to local needs is also manifested through the devolution of commissioning of the majority of health services provided by community and secondary care settings *e.g.* NHS Trust hospitals, to general practitioners through local clinical commissioning groups (CCGs).

The ability of these organisations to manage infectious diseases depends on the availability of high quality clinical and public health microbiology services. These services must be able to respond to cases or outbreaks of disease, to monitor patterns of disease incidence, prevalence and geographical distribution and to advise on appropriate clinical or public health actions required to minimise the threat of infectious disease to the relevant population. In the near future we anticipate they will also be tasked with delivering genomics-informed analysis and advice to their client organisations to improve the quality of infectious disease management in England. Consequently, understanding how microbiology services are configured and function is essential to the effective implementation of pathogen genomics in the health system. This requirement is particularly pressing, as the benefits of genomics will only be realised in full through the establishment of coordinated services that cut across the organisational, professional and geographical boundaries that characterise current microbiology provision in England.

11.2 The configuration of microbiology services in the English health system

11.2.1 Public health microbiology services

In England public health microbiology services are provided by Public Health England through a network of specialist microbiology laboratories embedded within NHS hospital trusts and a number of national reference laboratories that are mostly located in dedicated PHE run facilities but are occasionally located within regional hospital-based facilities. The principal functions of the public health microbiology services offered by PHE, through a combination of specialist regional and national reference laboratories, include:

- Delivery of specialist microbiological investigations to support diagnosis of infections, outbreak control and ongoing national disease surveillance programmes in community and healthcare settings
- Monitoring of antibiotic resistance and advising on actions to reduce further development and transmission of resistance
- Maintaining surveillance of vaccine uptake and effectiveness and advising on necessary changes in vaccination programmes
- Provision of advice to government on infectious disease policy
- Rapid response and management of novel or re-emerging disease threats, including large scale events such as flu pandemics

- Delivery of microbiological testing of environmental, food and water samples to organisations including the Food Standards Agency, port authorities and local authorities

- Research, development and implementation of new techniques to improve the effectiveness of microbiological investigations *e.g.* genomics

11.2.2 NHS clinical microbiology services

In England each of the 160 acute NHS hospital trusts, 34 community care providers and approximately 8,000 GP practices require full-time access to clinical microbiology services in order to support the diagnosis and clinical management of patients and the management of suspected outbreaks within both community and hospital settings. NHS clinical microbiology services are typically provided through hospital based microbiology laboratories staffed by a combination of medical microbiologists, virologists, clinical and biomedical scientists. A typical NHS clinical microbiology laboratory offers identification of a range of the most common pathogens and will undertake determination of the drug susceptibility of these pathogens to inform the most effective selection of treatment for patients. They play a crucial role in leading infection control activities within their hospitals and supporting these in the community. These include screening for healthcare associated infections, undertaking outbreak investigations and directing the implementation of infection control procedures. Clinical care of patients, based on laboratory microbiology investigations, is typically supported by medical microbiologists, virologists and infectious disease physicians depending on the infection and the clinical complexity.

Specialist testing, usually of less common pathogens or pathogens that require more complex molecular investigations that are beyond the scope of routine NHS laboratory provision is then undertaken by sending samples away either to the relevant regional PHE-led specialist microbiology laboratory or, in some cases, directly to a PHE national reference laboratory.

NHS microbiology laboratories also play a significant role in contributing to public health microbiology practice. They may test samples as part of a PHE-led outbreak investigation, undertake routine detection and reporting of notifiable pathogens of public health significance *e.g. M. tuberculosis* and provide routine surveillance information on the incidence of infections, in particular those resistant to antibiotics, within the local areas that they serve.

11.2.3 The role of private providers

A number of NHS hospital trusts now use private companies to provide microbiology services. Examples include the recent joint venture partnership between The Doctors Laboratory (TDL), University College London Hospital NHS Foundation Trust and The Royal Free London NHS Foundation Trust. Through this partnership TDL is contracted to provide a wide range of pathology services, including microbiology to these hospitals and by extension their CCGs. Other companies involved in these types of arrangements include Serco and Spire Health.

Public-private partnerships delivering NHS microbiology services should not be treated differently to those provided wholly by the NHS trusts themselves, or through contracts with PHE. NHS trusts must therefore ensure that their obligations to participate in public health related activities are met by the terms of the contract agreed with any external provider.

Recommendation 1

PHE will need to work with all microbiology service providers, both public and private sector, to ensure that they participate fully in meeting requirements to contribute to national infectious disease surveillance, through appropriate contributions to the implementation and development of pathogen genomics services.

11.2.4 The commissioning of microbiology services

The sources and flows of funding to support microbiology service delivery and development mirror the complexity of the organisational structures that underpin these activities. From a national perspective, funding for microbiology service delivery flows from the Department of Health via two main organisations, NHSE and PHE.

- **PHE -** PHE receives funds to pay for reference and some specialist microbiology service delivery both from central government, in order to fulfil their statutory health protection responsibilities, and through the sale of its services both to the NHS and to other national and international customers (private or public sector). NHS funding constitutes the majority (up to 90%) of the budget for PHE regional specialist microbiology laboratories, as they function principally as service providers to NHS primary, secondary and tertiary care.

- **NHSE -** receives funding from central government the majority of which is disbursed via local Clinical Commissioning Groups (CCGs). These CCGs procure microbiology services from NHS hospital trust microbiology laboratories as part of block pathology contracts with these hospitals for the provision of direct access testing for patients in primary care, and as part of the bundled tariff costs of episodes of hospital care *e.g.* the microbiology testing required as part of the care for a patient admitted to hospital with pneumonia.

From the perspective of the different types of laboratories (NHS Trust, PHE regional and PHE reference) the flows of funding are as follows:

- **NHS Trust -** NHS microbiology laboratories receive their funding from their host NHS Trusts. They can also receive income from other Trusts where it is part of a pathology network (see below) in return for providing microbiology services both directly to primary care physicians in the community and to secondary and tertiary care services. Each laboratory will in turn have to pay its associated regional PHE specialist laboratory for the provision of specialist testing *e.g.* certain virology services or molecular tests, the price of which is determined by a Service Level Agreement (SLA) between the relevant NHS and PHE laboratories. NHS hospital laboratories do not typically receive central funds from PHE for their routine role in support of public health microbiology activities, except where PHE identifies a specific outbreak of concern to their local health protection team or other acute need for particular testing to be undertaken for public health purposes. In these circumstances PHE will reimburse the cost of such testing where it is undertaken in local NHS laboratories.

- **PHE regional -** These laboratories interact with their host NHS trusts on a similar basis to a directly NHS controlled laboratory. Thus they receive a block of funding as part of a SLA with the host trust whose level is dependent on expected volumes of testing. In additional regional PHE specialist laboratories also have SLAs with other NHS trusts within their region to provide specialist and, where regional or local service consolidation has occurred, also routine testing. Collectively these two sources make up the majority of their funding. The remainder comes from central PHE funding to the laboratory to support new test development, the provision of tests of public health importance free of charge to the NHS and to support other ongoing public health needs, such as the provision of surge capacity in the case of a large scale epidemic or pandemic.

- **PHE reference -** The laboratories, and in particular PHE Colindale and PHE Porton provide diagnostic testing for rare or nationally significant pathogens to both local NHS and PHE regional laboratories, typically without charge, as part of their public health responsibilities. The costs of this activity and also their many secondary testing activities, such as typing, surveillance, vaccination evaluation *etc.* are met predominantly from central PHE funding, but also from external revenues from providing these services to other organisations in the UK and beyond.

11.2.5 The impact of these commissioning and funding arrangements on the provision of genomics services

Currently, who pays for any microbiological investigation is determined by a range of factors:

- The public health importance of the organism being investigated

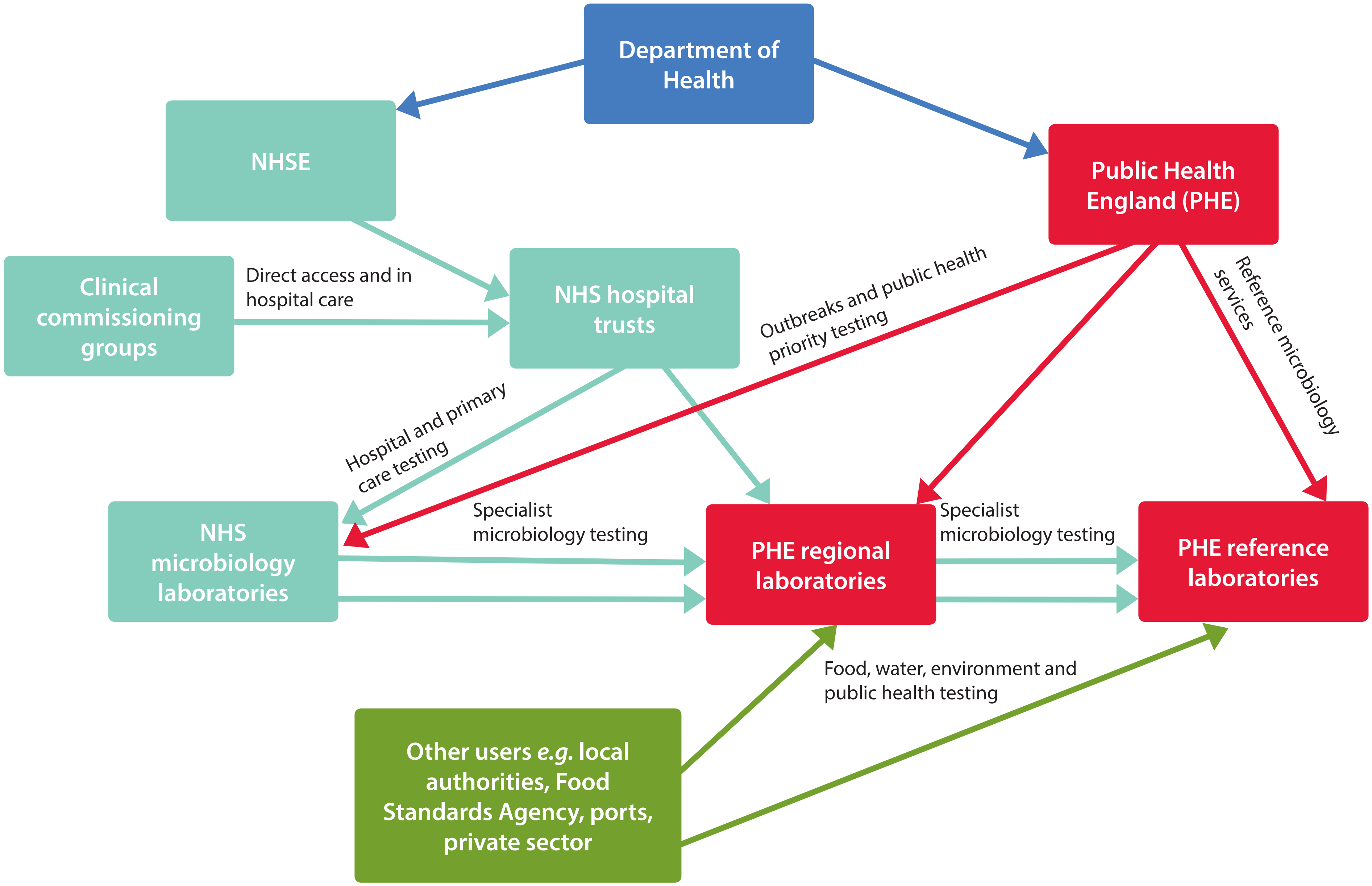
- The primary purpose of the test *e.g.* diagnosis, outbreak investigation, routine surveillance

- The test referrer *e.g.* primary care physician, public health epidemiologist, secondary care physician

- The designation of a test as specialist *versus* reference

- The location of the infected person

Figure 11.1 Funding flows for support microbiology services in England


Department of Health
NHSE
Public Health England (PHE)
Clinical commissioning groups
Direct access and in hospital care
NHS hospital trusts
Outbreaks and public health priority testing
Reference microbiology services
Hospital and primary care testing
NHS microbiology laboratories
Specialist microbiology testing
PHE regional laboratories
Specialist microbiology testing
PHE reference laboratories
Food, water, environment and public health testing
Other users e.g. local authorities, Food Standards Agency, ports, private sector

The introduction of genomic testing can be expected to pose a significant challenge to the application of these criteria as it is a single assay that, having been performed once, provides data that can be used for multiple tests of both diagnostic and public health significance. The question will therefore arise as to who should pay when WGS pathogen testing is undertaken within an NHS microbiology laboratory for the purpose of routine diagnostic investigation that also yields information of significant value to public health microbiology. This is particularly significant where availability of genomic information from a diagnostic laboratory obviates the need to undertake costly secondary testing within specialist or reference microbiology laboratories. In such cases investment in genomics services may be a net cost to the NHS but lead to net savings for PHE. Mechanisms will be needed to ensure that costs and benefits of developing and delivering genomic testing of shared clinical and public health utility are themselves shared across the organisations involved.

Recommendation 2

Agreement needs to be reached between PHE and NHSE with regards to funding for service development and delivery where the pathogen genomics services have a dual clinical and public health benefit.

11.2.6 NHS pathology reconfiguration and its impact on microbiology services

Over the past ten years there has been increasing recognition - stimulated by the Carter reports of 2006 and 2008 - that the highly atomised nature of NHS pathology services, with each individual NHS trust attempting to sustain their own broad portfolio of laboratories across most if not all specialities, was neither the most cost-effective way to provide pathology services to the NHS nor the best way to provide responsive, high quality and innovative services to patients. The Carter reports articulated very clearly the need for NHS pathology services to instead consolidate into large networks, where efficiencies of scale, streamlining of management and operations and a greater focus on quality and innovation could be achieved. This drive towards pathology consolidation has continued until the present day despite increasing internal competition between NHS trusts arising from various policy initiatives. It is therefore an important factor to consider when devising policy to support the effective implementation of genomics into the health service. Indeed, enabling pathology services to benefit from innovations such as genomics was one of the principal drivers behind Lord Carter's recommendations.

The effective use of genomics in microbiology is likely to require significant investment in infrastructure to support:

- Sequencing and analysis
- The availability of scientific and clinical expertise to exploit genomic information
- The aggregation, exchange and storage of genomic and clinical data
- Access to genomic and clinical data by a wide range of users

Recommendation 3

The initial implementation of pathogen genomics services should be focused in laboratories providing consolidated microbiology services, as these are most likely to be able to realise necessary economies of scale and to achieve the concentrations of expertise and efficient data management required.

As pathology consolidation is currently occurring at a variable pace across England, pathogen genomics will inevitably be implemented in England amidst a dynamic and mixed landscape of service configurations including:

- Single NHS trust microbiology laboratories that deliver in-house services to their own hospital patients and a geographically restricted local primary care population
- Small scale locally consolidated microbiology services that perhaps merge two local hospital laboratories onto one site
- Large scale regional microbiology services delivering across broad geographical areas such as that recently developed in the East of England (centred around the PHE regional laboratory in Cambridge University Hospitals NHS Foundation Trust)
- Unified pathology services that may operate within and between trusts, sometimes including private providers or managers and deliver microbiology services as part of a wider unified package of pathology services including biochemistry, genetics, histopathology *etc*.

Each of these will likely be operating on different business models in response to variations in underlying user demand, population and internal organisational needs and will necessarily take different views on the utility and feasibility of investing in the development of or access to genomics services for microbiology.

11.3 How and where are pathogen genomics services for clinical or public health use currently being developed?

The development of pathogen genomics services requires a number of steps to be undertaken prior to a new service being deployed:

- Development of the proof of principle evidence, through primary research studies published in peer-reviewed journals, that genomic information can be useful in investigating infectious disease
- The development of the underlying knowledge base required to inform interpretation of the results of genomics based investigations

- The acquisition, development and validation of the infrastructure, analytical and laboratory techniques and staff capabilities required to deliver whole genome sequencing and its interpretation within clinically accredited facilities

- Consultation with end users and commissioners to ensure that proposed new laboratory services will meet their needs and that the patient / public health pathway is configured to ensure benefits can be delivered

- Piloting of potential genomics services using real world clinical and public health infrastructure to assess the validity, utility and potential outcomes of service delivery

- Health economic assessment of the cost-effectiveness of any proposed service development and the development of business cases to support the investment required to establish services

In England these activities have been distributed across a range of organisations, including those, such as the PHE reference laboratory, whose primary purpose is service delivery, and those, such as academic institutions, whose primary purpose is to undertake research. In this section we describe briefly the principal initiatives underway deliver this service development work. In the next chapter we focus in more depth on some of the pilot programmes, and fully implemented services that have arisen from these initiatives.

11.3.1 The Healthcare Innovation Challenge Fund initiatives

Recognising the need to accelerate the translation and implementation of technologies such as pathogen genomics into mainstream health services the Department of Health and Wellcome Trust established a collaborative funding programme, the Healthcare Innovation Challenge Fund (HICF), to support structured programmes of translational and implementation focused research and service development across a range of healthcare technologies. Four separate grants have been made to consortia under this funding framework to develop microbiology services based on the application of pathogen genomics (See www.hicfund.org.uk[126] for details). These are:

- **Translating whole genome sequence technology into diagnostic and public health microbiology –** this is a University of Cambridge led initiative to develop the knowledge, tools and processes necessary to implement active genomic surveillance for outbreaks of infectious disease. It is currently focused - through collaboration with the Cambridge University Hospitals NHS Foundation Trust (CUHFT), the regional PHE laboratory based within CUHFT and the Wellcome Trust Sanger Institute - on developing genomic surveillance and outbreak investigation services, delivered through the regional PHE laboratory, for key priority pathogens for their local hospital populations.

- **Implementation of microbial whole genome sequencing for individual patient care, local outbreak recognition and national surveillance –** this University of Oxford led initiative is also focused on developing the knowledge, tools and processes necessary to implement genomics-informed microbiological services. They have an emphasis on the creation of a network of NHS and PHE operated service laboratories (in Oxford, Leeds, Brighton and Birmingham) that are able to deliver responsive, locally-based genome sequencing. This initiative currently has a particular focus on the development and piloting of a networked service for genomics based TB diagnostics and epidemiology.

- **Fully integrated, realtime detection, diagnosis and control of community diarrhoeal disease clusters and outbreak –** this University of Liverpool led initiative focuses specifically on the role that genomics can play in the management of diarrhoeal disease. It aims to develop a service that operates at a community healthcare level to detect diarrhoeal disease and use genomics to undertake surveillance, investigate outbreaks and guide public health action in real-time. A particular feature of this programme is the focus on integrating human and animal disease surveillance into their service model.

- **Infection response through viral genomics –** this is a University College London led initiative to develop services that exploit the power of whole genome sequencing of viruses to improve patient management, through treatment stratification, infection control within hospital environments and also control of epidemics of viral disease. They are targeting a range of pathogens for service development, including HIV, HCV, measles, influenza and norovirus.

Each of these programmes is taking a distinctive approach to developing pathogen genomics services that could be adopted by NHS and PHE and used in routine patient and population care. The service models being developed by the two most advanced programmes - in terms of their longevity and progress towards delivering viable frontline services - are described in the following chapter.

11.3.2 PHE Central Genomics Service

In parallel to the investments by the Department of Health and Wellcome Trust in the HICF programmes, the Health Protection Agency (now part of PHE) undertook in 2012, with funding from the Department of Health, to develop a centralised genomics facility at PHE Colindale in London to deliver a genomics based microbiology service to meet public health needs. The objective of this service was to provide cost-effective, resilient and accredited genomics capabilities to PHE. This was viewed as crucial to enabling 'self sufficiency' for PHE in case of national incidents or emergencies, but also to providing cost-effective access to genomics-informed services across PHE's infectious disease functions (including field epidemiology and communicable disease control services).

The PHE central genomics service has focused on developing services for pathogens of particular public health importance, including *Salmonella*, *S. aureus*, *S. pneumoniae* and influenza. This service was launched in April 2014, and is currently delivering internally validated (but as yet not clinically accredited) genomic data and interpretation to customers within PHE, mainly to inform outbreak investigations. The configuration of this service is described in more detail in the subsequent chapter.

11.4 Managing the transition from service development to service adoption, diffusion and delivery

11.4.1 Supporting adoption and diffusion through effective communication

As outlined above, there are already several locations and organisations across which pathogen genomics informed microbiology services are being developed. As the benefits of these services are realised the number can be expected to increase. This expansion places an onus on the pioneers in this field - and their funding organisations - to ensure the experience, expertise and knowledge they obtain is effectively shared across the network of potential service providers. This will minimise the need for each individual service to 'reinvent the wheel' through the empirical determination of the best methods for sequencing, analysis and interpretation where these have already been determined by other service providers.

The expansion of pathogen genomic informed microbiology services places an onus on the 'pioneers' in this field - and their funding organisations - to ensure the experience, expertise and knowledge they obtain is effectively shared across the network of potential service providers.

This will require a shift from communicating such information through peer review publication - expected during the translational research led phase of development - to the institution of mechanisms similar to those used by the NICE Health Technologies Adoption Programme in which learning from centres that are early adopters of a technology is collated and shared in the form of 'adoption support guides'. In some cases this may involve mobilising staff from one centre to train those at another, or may simply take the form of written guidance. The particular benefits of such an actively managed approach to supporting the adoption of genomics in microbiology services could include:

- **Opportunities to ensure consistent service quality –** it is in the interest of both PHE and NHSE that consistent, high quality genomics services are developed to ensure high quality equitable patient care and the generation of genomic data of a standard and completeness suitable for public health decision-making

- **Maximise efficiency of diffusion –** the availability of high quality guidance on how to develop effective genomics informed microbiology services could significantly shorten the time taken, and reduce the costs involved in their establishment

- **Addressing market failure –** unlike other technologies such as mass spectrometry that have been offered to the microbiology laboratories by commercial companies as 'ready to use' diagnostic devices, genomics devices are likely to consist of varying combinations of in-house developed sequencing methods (on 'research use only' platforms), analytical software and human expertise for interpretation. Whilst there is a financial incentive for commercial companies to provide support in setting up and running services using their products, the motivation for public sector service developers (who may principally be research and development oriented) to provide this support is likely to be less as they will not benefit from expending time and effort in diffusing their innovations across the health service

11.5 Conclusions

It is notable that the translational research and implementation activities outlined above are being undertaken in a limited number of locations, predominantly regional or national specialist PHE laboratories collaborating with academic centres of excellence in basic and translational research. By contrast the vast majority of routine microbiology services are offered through NHS microbiology laboratories operating outside these highly specialised environments. While it is necessary for service development to be undertaken initially within existing centres of expertise, it is also vital to consider how, in the near future, the services under development within these centres can be deployed most effectively for the management of infectious diseases nationwide.

In the following chapters we discuss the various ways in which this might be achieved, taking into account the variation in the underlying health needs across the population, access to genomics expertise and infrastructure across the health system and demands and expectations of users and commissioners. We explore the range of service delivery models currently under consideration and development within expert centres, and how these might be adapted and / or adopted to best deliver the ultimate objective of making insights from pathogen genomics available for the management of infectious disease across the country.

12 Pathogen genomics in action

A small number of microbiology laboratories in England have begun to introduce pathogen whole genome sequencing into their existing clinical and public health microbiology practice. In this section we describe how different units are piloting the implementation of WGS. These projects exemplify a range of different approaches being taken to the acquisition, interpretation and use of genomic data for microbiological investigations. They also demonstrate the different levels at which genomic data can be collected and used, from single hospitals, localised investigations and diagnostics to national networks for wider surveillance and centralised service provision.

12.1 Case study - PHE Colindale Central Genomics Service

One of the key health protection functions of Public Health England is to provide microbiological investigation services that contribute to the management of infectious disease. These services include functions of strategic national importance, such as provision of resilience in response to large scale outbreaks of infection *i.e.* epidemics or pandemics, the management of rare or particularly threatening pathogens (*e.g.* Ebola) and nationwide surveillance of infectious disease both to assess the efficacy of existing management strategies such as vaccination, and to detect potentially significant outbreaks of disease that require public health intervention.

In 2012 it was anticipated that genomics based methods could potentially be more cost effective to deploy than at least some existing microbiological methods (particular those used in specialised and reference microbiology). Evidence was also emerging that genomic methods could be more accurate and informative than existing microbiology approaches and that deployment of genomics based microbiological investigations could potentially have a transformative effect on infectious disease management. On the basis of these assumptions a successful business case was made to develop a central genomics service within the reference microbiology laboratories at PHE Colindale.

Aims and objectives

The stated objectives of the Central Genomics Service are to:

- To provide a cost-effective, resilient and accredited NGS capability for PHE
- To provide cost-effective and easy access to NGS platforms across PHE in order to facilitate rapid adoption of NGS for microbiology services and activities
- To ensure PHE is 'self-sufficient' in terms of NGS capability for response to national incidents and emergencies
- To enhance response to microbiological threats through development of emergency response capacity in genomics

During the development phase of this project a focus was placed on meeting these objectives for the following specific pathogens (or groups of pathogens) prioritised by service users and providers within PHE:

- *Salmonella spp*
- *S. aureus*
- *S. pneumoniae*
- Influenza
- Blood borne virus (BBV) antiviral resistance

Participants and structure of interactions

This service has been developed on a centralised model in which all functions are provided in-house within PHE Colindale. Delivery and development are being undertaken predominantly through collaboration between the Genomics Services unit, Bioinformatics units, and Information and Communications Technology services within PHE Colindale, with input from wider reference and specialist microbiology service providers across other PHE locations.

Service configuration and operation

Overview

The central genomics service is configured to provide high throughput, automated handling of DNA samples received from submitting laboratories, taking the samples from raw genomic DNA through the steps outlined in chapter 3 to the production of raw genomic data (in the form of FASTQ files) and associated metadata (relating to quality control, sample identity *etc*.). These data are then provided either directly to the requesting microbiology laboratory for their own analysis, or can be analysed and interpreted using the standardised and automated or more *ad hoc* and bespoke software being developed by the bioinformatics unit according to customer requirements.

Sequencing and IT hardware

In line with their mandate to build capacity to deliver genomic information, this service operates two HiSeq2500 and Two MiSeq sequencing machines. The sample preparation processes required prior to sequencing are automated using robotics wherever possible. An integrated LIMS system for tracking samples through the laboratory has been developed.

Significant investment has also been made in IT infrastructure to ensure that there is computational and storage capacity to handle both data analysis and archiving required to deliver the service. This IT infrastructure includes a dedicated high performance computing cluster and dedicated high performance data storage (both local and distributed).

Staffing

The wet laboratory processing of samples is undertaken by molecular biologists with varying skill levels, and the data management and analysis is undertaken by a combination of bioinformaticians, informatics experts and experts in IT infrastructure. There is a core twelve person virtual team drawn from across these groups of experts who lead on implementation.

There is a standalone bioinformatics unit (currently consisting of approximately 20 people) who contribute directly to the development and delivery of the central genomics service through developing analytical tools and databases and also undertaking in-house analysis for client laboratories. This team includes systems administrators, software developers, bioinformaticians and computer scientists.

Progress and evaluation

Salmonella *pilot*

The central genomics service has sought to validate their processes through running a pilot *Salmonella* typing project in which genomics-based typing was compared to existing serotyping methods in parallel on the same samples to determine the relative performance of the two methods. The primary outcome measure was concordance between serotype derived from genomic data and the phenotypically derived serotype, with >90% concordance being achieved in the first stage of validation.

The service is also piloting the more detailed SNP based analysis of a subset of *Salmonella* genomes with a view to enhancing the detection of outbreaks compared to existing methods by enabling the determination of relatedness of isolates of the same serotype. This may be of particular value where a large cluster of a particularly common *Salmonella* serotype is observed and current methods are unable to determine whether the cluster constitutes an outbreak or is a random spike in the background level of unrelated *Salmonella* cases. These genomic epidemiological methods remain under development, but are being tested in real world outbreak investigations, albeit on a research basis.

The central service aims to cease phenotypic *Salmonella* testing and use genomic testing as the first line test in April 2015.

Current status of the central service

While the central genomics service was launched in April 2014 and has been receiving samples for processing from clients within PHE and the NHS since this time, the service has yet to be accredited by UKAS. This means the results arising from its services must be treated as 'for research purposes only' and cannot be relied upon for clinical and public health decision-making. In order to achieve this accredited status, there remains significant internal validation work to be undertaken to ensure reliability and reproducibility for all assays. This encompasses the validation of laboratory workflows and bioinformatic processes, and an organism specific basis where these steps cannot be applied generically. See chapter 14 for further discussion of accreditation challenges.

Further developmental work required

The effectiveness of this - and indeed all genomics services - will continue to be limited until further work is carried out to develop comprehensive datasets describing the genomic diversity and architecture of each organism being investigated. Interpretation of the significance of any genomic variation observed is hindered without good prior knowledge of what to expect in outbreak and non-outbreak scenarios, and how this expectation varies within and between species.

From an analytical perspective, the tools for automating high throughput genomic analysis remain under development within the bioinformatics service at PHE and elsewhere. There is also a continuing need to identify the most effective and appropriate comparator genomes, analytical methods with which to perform genomic epidemiological investigations, and mechanisms for delivering consistent robust pipelines across PHE.

Conclusions

This centralised genomics service has several potential advantages to offer in delivering genomics-informed microbiological investigations to customers across the health service. By operating a high throughput facility it can achieve cost efficiencies that may not be accessible to smaller units, and physical centralisation eases the challenges of accessing and coordinating the expertise required to develop the complex laboratory and analytic workflows required to achieve pathogen-to-result services. However, as with all centralised pathology services, these advantages must be traded off against a potentially reduced ability to deliver rapid results, interact effectively with widely distributed frontline users of these results, and respond to varied local needs and priorities. Given these considerations this service is likely to remain most well suited to supporting specialist national surveillance and outbreak responses for which PHE is responsible and for developing genomic methods and the infrastructure required to support their wider adoption.

This service has successfully demonstrated that there is both potential demand, and a business case to be made, for the replacement of some existing specialist and reference microbiology testing services *e.g. Salmonella* serotyping with genomics. However, whilst this technology replacement may be cost saving at the laboratory level (for some tests at least) it remains to be demonstrated (and validated) that in the real world the results of the genomic investigations being undertaken lead to comparable or superior outcomes to those based on the existing methods. Only when the service is fully accredited (the target for this is in 2015) and able to deliver usable information to clinicians and public health practitioners on a routine basis can it be subjected to rigorous evaluation based on actual outcomes such as improved sensitivity and specificity in outbreak detection and response, and ultimately patient and population health outcomes.

12.2 Case study - Developing a specialist teaching hospital-centred infection control service (PHE Cambridge)

The pre-existing illnesses or injuries of many hospital patients may make them particularly susceptible to becoming infected with a range of different pathogens that circulate amongst the patient populations and in healthcare facilities. Reduction in the incidence of such healthcare associated infections (HCAIs) is an important component of hospital infectious disease management services and is a national strategic health policy. Whilst the focus on improving infection control practice within hospitals has led to dramatic falls in the incidence of these infections in the past ten years (90% reduction in MRSA bloodstream infections), they remain significant causes of morbidity and mortality across hospitals in the UK. Therefore, there still is a demand for hospital infection control teams to develop more sensitive and specific methods to detect outbreaks of such infections, investigate their sources and routes of transmission, and thus minimise the number of affected patients.

By comparing whole genome sequences of pathogens sequenced from different patients, fine scale differentiation of bacterial strains can be determined and the movement of the pathogen between patients can be traced. This high-precision epidemiological tool can facilitate more effective interventions that ultimately lead to more contained outbreaks and fewer outbreaks in the future, thereby reducing the number of people affected. Additionally, determining the source of an outbreak can help assign responsibility and accountability for the outbreak.

Following on from the success of the Health Innovation Challenge Fund (HICF) programme in which the University of Cambridge have shown how the application of genomics methods can have significant clinical utility in real world scenarios (see MRSA case study below), the PHE Clinical Microbiology and Public Health Laboratory (CMPHL) are now developing a clinical microbial genomics service that aims to provide whole genome sequencing (WGS) of selected pathogens for infection control purposes.

Aims and objectives

The microbial genomics service under development aims to assist hospital infection control teams in the management of the designated infections through the provision of WGS and analysis of bacterial samples isolated from infected patients. The success of this service would be measured as a reduction in the incidence of these infections, rationalisation of infection control procedures within the hospital and potentially cost savings to the client hospital trusts through reduced cost of treatment and infection control measures such as isolation, reduced length of patient stays and reduced application of financial sanctions by NHSE associated with excess cases of HCAIs.

The following specific pathogens will be the focus of the pilot due to their prevalence in hospitals, and the uncertainty of whether or not co-occurring cases in a hospital constitute outbreaks or independently acquired infections.

- **MRSA –** causes both skin and soft tissue and potentially fatal bloodstream infections
- ***C. difficile*** – causes severe diarrhoea, particularly in patients receiving antibiotic treatment
- ***M. abcessus*** – associated with cystic fibrosis and other lung conditions

Note that since the methodologies of pathogen genome sequencing and analysis are relatively generic, this programme could be expanded later to include other pathogens.

Participants and structure of interactions

Following a recent consolidation of microbiology services across the East of England, the PHE Clinical Microbiology and Public Health Laboratory (CMPHL) provides both specialist and routine microbiology services to Cambridge University Hospitals NHS Foundation Trust (CUHFT) and other hospitals in the East of England. The laboratory is a key member in the collaborative network of University and research institute based scientists, clinicians and clinician-scientists that are part of the University of Cambridge led HICF programme that is developing methods for the sequencing and analysis of pathogen genomes for clinical use.

Service configuration and operation

The genomics service under development leverages a range of existing genomics and microbiology infrastructure, capacity and expertise available within the University of Cambridge, PHE Cambridge and CUHFT:

- **Specialist microbiology –** all standard infection control and microbiological investigations will be performed by the relevant clinical and laboratory staff in accordance with current practice (see chapter 14). However, where a judgement is made that genomic information might assist in ruling in or out the presence of an outbreak, or in investigating sources or routes of transmission of infection where an outbreak has been identified, these standard procedures will be followed by the extraction of DNA from relevant bacterial isolates for WGS analysis

- **Genome sequencing –** CUHFT houses a regional human molecular genetics laboratory that offers a range of diagnostic NGS based tests to patients with rare inherited disease. In the PHE Cambridge service model, extracted bacterial DNA is sent to this clinical genetics laboratory for whole genome sequencing to be performed under clinically accredited conditions

- **Genome data analysis and interpretation –** WGS data will then be fed back to the microbiology laboratory for analysis and interpretation, using automated software under development by bioinformaticians and software engineers working at the Wellcome Trust Sanger Institute, in the context of relevant clinical and epidemiological information supplied by clinical staff. A report will be generated by the relevant clinical microbiologist that can be stored in patient electronic health records and also be used to inform infection control measures undertaken by clinical staff

Progress and evaluation

Proof of principle studies

Through the existing research funding available to the consortium developing this service they have undertaken several WGS based investigations of outbreaks within their constituent hospitals that have had positive outcomes for infection control. These include the MRSA investigations described earlier in section 7.7 of this report, investigations into a suspected outbreak of *M. tuberculosis,* and the investigation of a cluster of *M. abcessus* infections amongst cystic fibrosis patients[8,127,128].

Current status of the service

This service is under development and not yet fully operational. A pilot phase is being supported by CUHFT during which time the appropriate clinical or epidemiological triggers for deciding to use genomics as part of a prospective surveillance programme or reactive investigation will be determined. Accreditation is also being sought for the additional DNA extraction methods being introduced in the microbiology laboratory.

Further developmental work required

Methods for integrating and automating the analysis of genomic and clinical / epidemiological information, which will also require accreditation, remain under development. As part of the current pilot, health economic analysis is also being undertaken to determine the cost effectiveness of genomics as an adjunct to existing microbiological investigations, with a view to informing the most appropriate use of the technology and building a business case for longer term adoption of the service.

Currently the success of this service model depends on:

- The existence of an active research grant-supported collaborative network of experts drawn from academia and the NHS
- Local availability of genomic sequencing infrastructure that is already paid for and accredited for clinical use
- Local availability of expert bioinformaticians to construct and maintain analysis pipelines
- Local availability of high performance computing and data storage facilities
- Existing expertise in molecular microbiological methods and relevant infrastructure within the regional PHE microbiology facility
- A sufficiently wide client base of hospitals to sustain what may initially be a low volume service

Conclusions

The service model being developed in this pilot has particular advantages that arise from its efficient use of existing genomics resources available within the NHS trust and surrounding University facilities. Other similarly configured hospital trusts, where there is a concentration of relevant expertise and infrastructure, particularly in delivering clinical genomic analysis (either through existing molecular genetics or molecular pathology services) and a sufficiently large demand for genomics informed microbiology investigations within either an NHS- or PHE-led microbiology service may find this model a useful template in designing their own services. However, for hospitals that do not have existing expertise in genomics, and who currently outsource specialist molecular microbiology to their regional microbiology laboratory establishing an in-house service using this model is likely to be less feasible due to the cost and complexity of acquiring the necessary sequencing and analytical infrastructure and expertise. In this latter case it may be desirable for hospitals to send their samples to a centralised centre rather than conduct local analysis. The range of possible genomic service configurations, and processes to support decisions to commission or procure such services, are discussed in detail in subsequent chapters.

12.3 Case study - Developing a whole genome sequencing approach to HIV treatment management (PHE Cambridge)

The human immunodeficiency virus (HIV), the pathogen underlying acquired immune deficiency syndrome (AIDS), is one of the largest global health burdens, and approximately 100,000 people with HIV are estimated to be living in the UK. Appropriate treatment of HIV with antiviral drugs can improve the length and quality of lives of infected patients. However, there are issues of drug resistance, where certain viral strains are resistant to particular antiviral drugs.

There are around 25 different available drugs to treat HIV infections, and any given HIV infection could be resistant to a combination of these different drugs. Therefore, an important component of effective treatment is to ensure that the drugs given to the patient are tailored towards the specific strain of HIV.

The current protocol for treating an HIV patient in most UK hospitals is as follows:

- HIV genotyping is first performed when a patient appears to have an HIV infection

- Drug treatment is then chosen and given to the patient with the aim of suppressing the HIV viral load

- If the patient develops symptoms or the HIV viral load rises then genotyping is performed again

The second genotyping test hopes to reveal if patients have relapsed due to either issues with drug adherence (where the patient is not taking their prescribed drugs as instructed) or issues of drug resistance. In the former case, genotyping would reveal the prevalence of the wild-type viral strain. In the latter case resistance mutations could be detected, helping tailor the choice of the most appropriate antiviral drugs to give to the patient.

However, conventional genotyping approaches for detecting HIV antiviral resistance-based on capillary sequencing technology only capture the resistance profile for a subset of these drugs while deep NGS and WGS methods can provide a full prediction across all the drugs, thus permitting the most suitable drug treatment for each patient. WGS data could also be used for epidemiological purposes to potentially trace the transmission of HIV among individuals. Using their expertise in microbial genomics, and building on their existing microbiology services, the Cambridge led PHE team are planning to develop an HIV deep WGS assay to be deployed in selected PHE laboratories.

Aims and objectives

The aim of this project is to develop a validated WGS assay for diagnostic HIV resistance testing and to provide WGS data for epidemiological purposes that will be deposited in national databases. This will fulfil the 100,000 Genomes Project objective of producing WGS data to accompany resistance test assays.

Participants and structure of interactions

The assay development for the project has been divided between PHE Cambridge - who are integrated with academics and clinicians at the University of Cambridge as part of their HICF consortium - and the PHE laboratories in Birmingham and Colindale. It is intended that a clinically accredited HIV WGS assay - developed through collaboration across these three sites - should eventually be deployed across three to four PHE regional specialist microbiology laboratories and a subset of NHS microbiology laboratories wishing to offer HIV genotyping.

Service configuration and operation

PHE aims to deliver an accredited HIV WGS service by August 2015 therefore the exact details of how the service will be implemented, configured and operated are still under development, and the impact of the assay on existing HIV care pathways is yet to be determined.

Progress and evaluation

The assay development and evaluation processes for this pilot are underway across the three centres involved. The effectiveness of these assays in improving patient outcomes will be assessed once a preferred assay methodology has been established and evaluated as part of the existing HIV testing pathway.

Further developmental work required

NGS based HIV genotyping (where only a proportion of the HIV genome is analysed in depth) is already available as a clinically accredited service via the PHE Cambridge laboratory. It remains to be determined whether the current evolution and development of this capability into a deep whole genome sequencing service will identify new biomarkers of drug resistance and hence potentially improve future treatment selection for patients, and consequently lead to better clinical outcomes. It has also been proposed that information from such a WGS based service could also be used for epidemiological purposes, in the population-wide management of HIV, but the utility of this approach in practice remains to be determined, and given the sensitive nature of sexually transmitted infections such as HIV, such work would need to consider potential ethical, legal and social issues associated with this approach.

12.4 Case study - Tuberculosis genomics service pilot in Birmingham, Brighton, Leeds and Oxford

Tuberculosis (TB) is a UK government priority. Incidence of the disease in the UK, and particularly in England, is at its highest since the 1980s and higher than most Western European countries. In England in 2013 there were 7,290 reported TB cases (13.5 per 100,000 population), most of which occurred in urban areas with a high population density. Many cases are due to reactivated latent infection in individuals born outside the UK. However, UK-born patients are more likely to have pulmonary disease and be involved in transmission. Efforts to significantly decrease the number of TB cases have not yet been effective, and if current trends continue in two years there will be more TB cases in the UK than in the United States. Although the number of cases is relatively low, TB diagnosis and treatment use disproportionately more resources than other infectious diseases due to the difficulty of diagnosis and length of treatment, which is six months for standard drug-sensitive cases.

In response to the number of TB cases in England, PHE, NHS England and other stakeholders released a strategy in January 2015 outlining a collaborative project to strengthen the health system's response to TB infections. There are a number of challenges associated with diagnosing, monitoring and treating TB, which were highlighted by the report:

- **Improving TB diagnostics -** Mycobacteria grow extremely slowly (dividing approximately once every 24 hours) therefore traditional culture based methods cause delays in confirming diagnosis and in detecting antibiotic resistance. The European Centre for Disease prevention and Control has set a target for the proportion of pulmonary TB cases that are confirmed by growth in culture at 80%. The figure for England is currently 68.7%

- **Reducing diagnostic delay -** this can be due to (1) patients not entering the health system quickly enough, and (2) diagnosis via microbiology which can take up to two months or more in drug resistant cases. These delays leave patients untreated or on the wrong drug regime, increasing transmission risk

- **Reducing drug resistant TB -** although there were only 74 cases of MDR-TB in the UK in 2013, it is very resource intensive to treat, in terms of the drugs needed, likelihood of hospital admission, and a two-year treatment time

- **Reducing TB transmission -** understanding TB transmission in detail is vital to promote efforts to reduce incidence, and to trace and treat patients to prevent the spread of the disease

Aims and objectives

The aim of the TB pilot is to use whole genome sequencing (WGS) to:

- Create a database of all isolates such that outbreaks and transmission chains can be determined, and drug resistance monitored

- Speed up diagnosis: once the organism has been successfully cultured, WGS will speed up diagnosis and determination of resistance to 1–2 weeks versus 2–8 weeks or more

- Greater resolution of genetic information will lead to improved epidemiological information, tracking of transmission and improved surveillance

Participants and structure of interactions

Mycobacterium tuberculosis has been selected by Genomics England as a pathogen suitable for WGS method implementation for patient care and this task has been entrusted to PHE. The initial phase of the pilot is taking place in four PHE and NHS (PHE collaborating) microbiology laboratories: Birmingham, Brighton, Oxford and Leeds. This approach relies on expertise in different geographical locations, developed largely through the University of Oxford led HICF project, and makes use of many of the procedures already in place for laboratory safety, tracking and transporting specimens, therefore incurring minimal costs to set up new systems.

Service configuration and operation

Overview

Bacterial cultures grown in these laboratories from patients with tuberculosis will be sequenced locally in the laboratory. The project has developed standardised procedures for extracting and purifying DNA from bacterial cultures, which are easily reproducible between laboratories. The genomic data are then analysed using a semiautomated system by a team in Oxford, with the results returned to the laboratory that submitted the sequence.

Sequencing and IT Hardware

Sequencing is taking place in laboratories in Birmingham, Brighton, Oxford and Leeds. The sequencing data are uploaded by the laboratories to an online facility, BaseSpace, from which it is downloaded and analysed by a team based in Oxford. They carry out sequence assembly and analysis using a semiautomated system that allows species identification, resistance prediction, genomic matching to near neighbours in the database and data storage. A report is sent back to the laboratory that submitted the sequence data, which is shared with the clinical team who plan the patient's treatment. Interpretation of the report is carried out by the microbiology and public health teams.

Steps have been put in place for the management of the data. Minimal clinical and laboratory data, collected through LIMS, must be associated with each sample submitted for sequencing. The pilot will also provide links with the enhanced tuberculosis surveillance system (hosted by PHE Colindale), and explore new ways of capturing clinical and pathology data. There is an integrated informatics system for registering and keeping track of specimens, with an automated workflow, hosted between Oxford and PHE Colindale.

Staffing

The processing of samples is undertaken by biomedical scientists, while the analysis is undertaken by bioinformaticians and computer scientists, supported by systems administrators and software developers.

Progress and evaluation

TB pilot - current status

The pilot is supported by a database of more than 2,500 WGS TB isolates including all retrospective sequential and consecutive TB isolates for Oxford (2007-present) and Birmingham City (2009-present). All subsequently sequenced isolates are added to this database.

Currently all sequencing is carried out in parallel with conventional pathways for diagnosing and monitoring TB, which includes MIRU-VNTR typing for determining genotype of isolates. The laboratory procedures and processes for sequencing and analysing TB are currently not accredited, therefore the two services are running in parallel until the genomic approach is fully validated and accredited as being suitable for frontline clinical use.

Further developmental work required

The effectiveness of the pilot and its subsequent development into a fully-fledged WGS service will depend on the resolution of a number of key factors:

- The relationship of the TB sequencing service to other PHE laboratories, and integration of this service into current workflows
- Accreditation of workflows to national and global standards
- Automation of processes to make them easy to use in a clinical context
- Data linkage and storage and how these data link to current databases storing clinical, genotypic and surveillance information on TB
- Managing information and maintaining expertise, *e.g.* determining how poorly understood mutations lead to resistance, and maintaining a catalogue of these and known resistance mutations
- Cost effectiveness analysis of WGS performance versus current procedures

Conclusions

In principle, the implementation of whole genome sequencing has a positive impact on the diagnosis, treatment and infection control of tuberculosis. The benefits can include:

- Simplified workflows – one experimental procedure (sequencing) identifies the pathogen and its strain/s, as well as drug susceptibility, and more quickly than current practice

- Better quality data – WGS provides comprehensive information on the genetics of the disease, allowing for easier identification

- Increased volume of data about the pathogen, including genetic, epidemiological and clinical information, which can be used to form a strong national surveillance system

- Information from sequencing means that patients are more likely to be given the correct therapy upfront, reducing the chance of hospital admission and development of drug resistance

The advantages of the networked approach being taken in the pilot are:

- Expertise and expensive equipment can be concentrated in a few key locations, ensuring most efficient use of resources. Given the number of cases of TB in the UK each year (around 8,000) it is likely to be more cost-effective to send samples to a number of centres. This can include using labs that carry out sequencing for other pathogens such as Hepatitis C or HIV

- As the pilot develops into a fully-fledged service, the networked approach gives flexibility to the system, for example new labs can link into existing accredited workflows, and new analysis pipelines can be made available online

However these advantages are mitigated by the following disadvantages, which will need to be resolved in addition to the further work listed above:

- The use of WGS affects the role of the reference lab, which traditionally carries out strain typing and susceptibility testing

- Hospitals in areas which have low incidence of TB or lack the resources to set up a service will benefit from having access to remote but well-resourced regional centres for sequencing. However, in some cases this networked approach and remote working could result in delays, for example through challenges in communication or external circumstances delaying the arrival of samples, and the receipt of results

- There are implications for the workforce including training staff in new technologies, and access to bioinformatics expertise

Although there is evidence to suggest that using WGS will confer significant benefits to the management of TB in terms of time savings and optimal use of resources, full accreditation and rigorous testing in a real-world setting should determine whether WGS is cost-effective and delivers better outcomes compared to conventional methods.

13 Configuring pathogen genomics services I: a frontline view

Effective use of pathogen genomic information to improve outcomes for patients and populations will depend on the optimal configuration of the microbiology and wider infectious disease services through which this information is generated and utilised.

13.1 Introduction

In this chapter we describe the circumstances under which the implementation of genomics informed infectious disease management services, particularly for outbreak control and surveillance, might be considered beneficial for patients and / or populations and the ways in which they might optimally be configured.

13.2 Detection and investigation of outbreaks of infectious disease

Both public health microbiology services and hospital-focused microbiology services undertake programmes of infectious disease surveillance and investigation which aim to prevent outbreaks by prompting pre-emptive infection control measures and / or to investigate suspected outbreaks to enable rational decisions to be taken about whether to undertake additional infection control measures.

The sensitivity and specificity of these outbreak detection and investigation process are, however, limited (chapter 8) by the quality of the microbiological information (phenotypic or genotypic) available to discriminate between related and unrelated cases of infections that appear to cluster in time and place. This is not the only limitation on the effectiveness of outbreak detection, as the completeness and accuracy of epidemiological information *e.g.*the precise circumstances or location in which infections occur, are also often far from perfect. Nevertheless, it is clear from many published studies that in principle, where microbiological resolution of the relatedness of isolates is the limiting factor in detection and resolution of suspected outbreaks, the use of genomic information should enable improved sensitivity and specificity of these processes (see chapter 8 for details of how).

13.2.1 Deciding whether to commission a genomics-informed outbreak detection and investigation service

The decision to implement outbreak surveillance and investigation services informed by pathogen genomics will depend a range of factors including:

Drivers	Questions and notes
Need	• What is the underlying prevalence of the infection in the settings covered by the service? • What is the expected frequency of suspected outbreaks that would benefit from genomics-informed investigation? • What volume of testing activity is anticipated?
Clinical / public health significance	• Does the infection being investigated represent a significant burden in terms of patient or population morbidity and / or mortality? • Is the infection a priority with respect to preventing the spread of antibiotic resistance?
Strategic policy	• Are there local, national or international policy priorities associated with effective detection and control of the proposed infection? • Are there specific performance management targets, and financial penalties / incentives associated with control of the infection to be investigated?
Cost of procurement	• What are the costs of procuring genomic information for this service? • What throughput is expected, and can sufficient economies of scale be achieved? • Will services be established in-house or procured from an external provider?
Impact on cost and effectiveness of current testing and care pathway	• What is the relative expense of current microbiological testing for outbreak surveillance / investigation vs WGS? • Will WGS led approaches significantly increase sensitivity and specificity of outbreak detection / investigation? • Will improvements in sensitivity / specificity increase or decrease the number of outbreak investigations and interventions required? • Will any change in the frequency or duration of outbreaks lead to significant reductions in cost of managing infectious disease in the setting?
Impact on outcomes for patients and populations	• Are there the mechanisms, capacity and resources within existing infection control pathways to respond effectively to genomic information and to realise improved outcomes through more rapid or rational implementation of infection control measures? • How likely are measurable improvements in patient and population outcomes to be delivered?

The relative weight and importance placed on each of these factors will depend on the underlying characteristics of the organisation making the decision on whether or not to implement such genomics-informed surveillance services. Ultimately, they will be determined by an assessment of the evidence of the cost-effectiveness, and the underlying willingness of the health service to pay for any anticipated benefits. These health economic considerations are explored in chapter 19.

13.2.2 Configuring an effective genomics-informed outbreak and investigation service

Where genomics-informed outbreak surveillance and investigation services for particular pathogens are commissioned, this should be done only following the establishment of a robust and effective pathway, incorporating genomic information. Optimal configuration of such a pathway would include:

Aspect of service design	Explanation
Scope and entry criteria	• The pathogen or range of pathogens for which the services is to be used should be defined • The criteria that would trigger the use of genomics informed surveillance or outbreak investigation should be defined
Definition of test characteristics	• A consistent definition of the genomic information that would trigger an outbreak investigation should be defined and agreed for each pathogen and clinical / public health setting • Clear criteria, underpinned by reproducible informatics processes, for refuting / confirming transmission events should be agreed
Data quality and test performance criteria	• Minimum standards for data quality, both epidemiological and genomic, required for robust surveillance and investigation must be agreed. • Existing best practice guidelines should be followed or exceeded • Genomic testing should be appropriately accredited and subject to external quality assurance where possible
Resources for data collection and interpretation	• The availability of resources to collect the quantity and quality of genomic, epidemiological and clinical data required to provide the agreed level of service must be in place • The availability of in-house or procured expertise to analyse and interpret data to the required standard must be in place

Aspect of service design	Explanation
Availability and suitability of routes to communicate results and underlying data	• Reporting tools and mechanisms to ensure understandable and actionable information reaches frontline staff must be in place • Infrastructure and processes to ensure underlying genomic and clinical / epidemiological data are appropriately shared for wider strategic public health needs should be established
Resources for downstream clinical / public health response	• The availability of resources to enable the appropriate infection control response *e.g.* hospital beds for isolation, clinical capacity to treat additional cases, appropriately trained staff to perform contact tracing, must be in place

Each of these considerations will need to be underpinned by a robust evidence base demonstrating a rationale for the definitions and criteria used in designing the service, and appropriate performance evaluation and validation systems to determine the effectiveness of the chosen configuration. During the early implementation of these services it is expected that the level of evidence available to categorically define important test criteria, such as what genomic data would constitute support or refutation of a possible transmission event, may be limited. In these cases, it is reasonable to work to the best evidence available at the time, assuming minimum standards for test performance can be met. This situation will however require that the service providers and users engage actively in efforts to improve the quality of this evidence base, to ensure the robustness and quality of the service can be raised over time.

Recommendation 4

A defined pathway, encompassing test referral mechanisms, sequencing, analysis and interpretation must be developed for each pathogen and each application of genomics. Implementation of these pathways will require a coordinated approach.

Recommendation 5

Robust and effective prioritisation processes will need to be developed for new service developments. These must be informed by consultation including frontline end user groups.

13.3 Informing infection control policies and evaluating their effectiveness

Pathogen genomics can provide high resolution information about the population dynamics of infections over time, in particular the origins, the extent and the mechanism of their transmission. Such information could contribute significantly to informing preventive infection control practices in two ways:

- **Prospective -** providing information to guide the optimal configuration of new infection control measures

- **Retrospective -** providing an audit of the effectiveness of preventive measures once they are in place

This information could be of use to local health service providers, national public health agencies seeking to optimise their infection control practices and central or local government and health service policy makers with a role in setting overall targets and priorities for infection control.

Examples of potential uses include:

- The decision on whether or not the introduction of screening to detect and prevent transmission of a putative healthcare associated infection, or indeed to introduce financial measures against hospitals with 'excess' cases of healthcare associated infections, depends on underlying knowledge of whether these infections are being acquired within the hospital and are, therefore, amenable to improved control through enhanced compliance with hygiene and other infection control practices in the hospital itself

- The introduction of new vaccine programmes, or the evaluation of the effectiveness of current programmes, where knowledge of the origins, mechanisms and extent of spread of the infection being targeted are crucial to determining prospectively whether a vaccine programme is likely to be effective, and retrospectively whether it is achieving its stated aims

13.3.1 Deciding whether to implement a genomics-informed service for the evaluation of infection control policies and procedures

It is conceivable that genomics services for the prospective or retrospective evaluation of infection control policies could be commissioned either locally, by individual health service organisations with a responsibility for infection control in a defined setting or range of settings, or nationally, by organisation such as PHE or NHSE wishing to review and rationalise their nationally applied infection control policies. While it is not clear who would be responsible for providing and funding such services, it is nevertheless important to address the questions in the following table in deciding whether or not they could be warranted.

The configuration of services to deliver infection control policy analysis would have broadly similar requirements to those covered for outbreak detection and investigation, which are described in the table below.

Driver	Questions
Need	• How frequently are infection control policies subject to review and evaluation and what triggers these? • Are reviews and evaluations generalizable across settings, or would individualised analyses be required for each healthcare facility or community? • What is the underlying perceived utility of genomics in informing infection control policy?
Costs of service provision	• Who will provide these services? • Can necessary genomic data be procured as a 'by product' of other genomic investigations or would separate data collection and analysis be required? • How is demand, and therefore throughput, likely to change over time?
Impact on policy design and evaluation	• To what extent will genomic information impact on infection control policies that have wider social and political determinants? • Does genomic information constitute significantly higher quality evidence for policy change than that available from current non-genomic methods of audit and evaluation?
Impact on health outcomes for patients and populations	• Will changes in infection control policy consequent to use of genomic information significantly improve outcomes for patients or populations? • Are the resources and capabilities available to undertake any change in infection control practice indicated by a genomics-informed investigation of current policy?
Impact on cost of infection control provision	Do the results of any genomics informed investigation indicate changes to infection control policy that would lead to a net increase or decrease in cost? How might costs of infection control be redistributed between healthcare organisations if more accurate information on the source of infections becomes available through genomics informed investigations?

13.4 Optimising the distribution of genomic sequencing and analysis services

Implicit in the consideration of the configuration of different genomics-informed microbiology services is the need to first determine the most effective way of delivering the necessary information to underpin these services to the end-users. The question to be addressed is what is the most effective and efficient way of delivering the necessary pathogen genomic information from these services to their end users? Supplementary to this is how different parts of the service pathway are distributed *i.e.* whether sequencing, analysis and interpretation must necessarily be undertaken in a single location as part of an integrated service or is more effectively distributed, with some parts being centralised and others placed closer to the end users.

The question to be addressed is what is the most effective and efficient way of delivering the necessary pathogen genomic information from genomics-informed microbiology services to their end users?

The current lack of evaluations of the effectiveness of different potential service configurations, combined with the widely varying underlying circumstances of the service users and the different service requirements for different pathogens, means that it is not possible to propose a definitive model for the optimal distribution of genomic sequencing and analysis provision in England. Instead we will describe the criteria that must be considered, on a user-by-user and pathogen-by-pathogen basis, when deciding on how to procure genomics services of the types described above.

13.4.1 Whole genome sequencing capacity

A user wishing to procure pathogen whole genome sequences for clinical or public health purposes has several options, including:

- Acquire a sequencing instrument for their own microbiology laboratory and the necessary skills and accreditation to operate it for clinical or public health purposes

- Gain access to existing genomic sequencing facilities within their own organisation that can offer suitably accredited sequencing

- Send away samples for genomic sequencing by an external organisation offering a suitably accredited service

The choice between these three options will depend on the following factors:

- **Volume and distribution of demand –** this will influence whether the most cost effective way of procuring sequencing is to acquire a dedicated in-house instrument, or pay for use of other internal or external sequencing capacity from a larger regional or national laboratory running high throughput, relatively low cost per sample machines, or a local molecular genetics laboratory with spare capacity that can be procured on their existing machines.

- **Desired turnaround time –** if rapid turnaround times are required then the decision may be skewed towards more localised sequencing where there may be greater control of when runs are initiated and over sample delivery. This will however have to be balanced with effective utilisation of machines, as the expense of 'random access' sequencing runs are significantly increased where the capacity of the machine is underutilised.

- **Relative costs and available finance –** where significant capital costs for procuring sequencing platforms and associated laboratory equipment have already been met through prior grant funding (from research agencies) the marginal costs of establishing an in-house clinical service will be significantly lower than if significant new capital investment is required. Conversely, where such funding is not available, the cost of procuring sequencing services from other internal or external laboratories, particularly those where the costs of procuring and maintaining sequencing equipment have already been met from other sources, may be significantly lower.

- **Availability and location of existing expertise and equipment –** implementing in-house sequencing services will be significantly easier in environments where significant expertise in developing and delivering genomic medicine are already available than where such expertise has to be sourced externally. Given the fragmented nature of pathology services, it must however be noted that physical proximity to existing expertise is not *per se* sufficient to support new in-house service development. Support is likely to be more dependent on effective working relationships between practitioners within and across disciplinary boundaries, wherever they are physically located.

Given that different organisations, across the NHS and PHE, are likely to be differentially influenced by the above factors, we conclude that in the initial stages of implementation of pathogen genomics across the healthcare system a mixed model of sequencing provision is likely to prevail, determined by the needs and circumstances of individual users. This has significant implications for ensuring the standardisation and quality of provision required to deliver equitable local services and also to deliver the genomic information required for nationwide public health programmes. These requirements will be discussed in more detail in subsequent sections.

13.4.2 Optimising the distribution of analytical and interpretive capacity

Acquisition of genomic sequencing is only the first step in obtaining genomic information with which to inform services. Arguably, the bioinformatic processing and clinical / epidemiological interpretation of the data are more challenging steps, and are certainly crucial to success. Where and by whom these steps are most optimally undertaken, similarly to the sequencing itself, is subject to a number of considerations (more extensive rationale for these is covered in the preceding chapters on data management):

- Availability of software tools and expertise for analysis and interpretation

- Access to support and expertise in the development and maintenance of analytical and interpretation services

- Access to necessary IT infrastructure to undertake computational analysis

- Access to necessary IT infrastructure to transfer and store genomic data

There is unlikely to be a 'one size fits all' model for the delivery of genome analysis and interpretation services within microbiology. A number of models could be adopted, with preference depending on local circumstances whether or not there is access to in-house clinical and bioinformatics expertise, or access to a local high performance computing service (often through linkage with co-located university IT facilities).

Despite the lack of an obvious single preferred model for accessing analytical and interpretive services, we can however conclude that, at least in the early phases of implementation, bioinformatics expertise and computational capacity are likely to remain 'rate limiting' factors in the decentralisation of genomics services, and considerable reliance will remain on more centralised models for data analysis and processing than might be desired in the longer term.

Recommendation 6

The location of sequencing and analysis services should not be pre-determined, and a mixed model should be allowed to develop which makes optimal use of available resources and takes account of local / national demand for genomics: variables include the cost, throughput achievable at each location, and turnaround time.

13.5 Conclusions

This chapter highlights the wide range of factors that will need to be taken into consideration when commissioning, developing and delivering microbiological services informed by pathogen genomics. Whilst we have focused on the introduction of services that can be broadly categorised as supporting infection control, we anticipate that similar factors will define the appropriate configuration of services based on the application of genomics to other areas of microbiological investigation further into the future. What is clear from the above discussion is that, from an individual service provider and user perspective, there is not currently a viable 'one size fits all' model of service delivery that can be recommended. Instead, a range of services configurations are likely to evolve in response to differing demands across the health service. Indeed, examples of potential service configurations are already emerging through the various pilot and implementation initiatives, described in chapter 12.

In the next chapter we consider how, against this background of diverse and need led service configuration, equitable, high quality and nationwide provision of pathogen genomics informed clinical management can be achieved. We also consider how to realise the wider strategic public health importance of genomic data generated by individual microbiology services, including through interaction with animal health services and global partners engaged in similar efforts to implement pathogen genomics in their own settings.

14 Configuring pathogen genomics services II: a whole system view

A whole system view of the implementation of pathogen genomics is necessary to ensure that collectively individual services provide high quality care, available nationwide on an equitable basis and that each service is configured and resourced in a way that enables it to participate in the delivery of wider public health benefits.

14.1 Introduction

In the previous chapter we concluded that, particularly during the initial implementation of pathogen genomics services, a mixed operational model would necessarily predominate in which individual service users and providers would commission and develop differently configured services according to their particular needs and circumstances. While such mixed provision allows for multiple versions of 'local optimality' to be determined, this must be balanced against health system-wide requirements to ensure:

- Consistently high quality and equitable service provision
- Interoperability, responsiveness and capacity to serve strategic public health needs
- Ability to respond to and exploit developments in genomic knowledge and technology

In this chapter we will explore the mechanisms that will be required to achieve the above goals and consider how these will intersect with the factors underlying effective individual service delivery described in the previous chapter.

14.2 Quality assurance for pathogen genomics services

Like all pathology services microbiology is subject to multiple forms of quality assurance. Their purpose is to provide internal assurance to providers that they are producing a high quality and safe service, and external assurance to commissioners or users that the service they are procuring meets relevant national or international standards. It also has an important role in providing reassurance to patients that the diagnostic tests on which much of their care depends are reliable and accurate. A recent review of pathology quality assurance in the NHS in England has been undertaken by Dr Ian Barnes, and contains many general recommendations on improvements in pathology QA that are relevant to microbiology services.

We focus here on three aspects of quality assurance that are particularly pertinent to supporting the development and delivery of high quality pathogen genomics services that meet the Barnes report's criteria of 'reliability, robustness and responsiveness'. These are laboratory accreditation, external quality assurance and best practice guideline development.

14.2.1 Laboratory accreditation

Clinical laboratory accreditation is a voluntary process in the UK that provides independent recognition of a laboratory's competence to perform specific tests. The vast majority of pathology laboratories, including microbiology laboratories, submit themselves for accreditation by the United Kingdom Accreditation Service (UKAS), which currently assesses their competences using the international standard ISO 15189: 2012. This accreditation provides vital assurance to users and commissioners that the laboratory offering a particular test is competent to do so.

In the context of the introduction of genomics, the transition from the previous regime of the Clinical Pathology Accreditation (CPA) scheme to ISO15189 offers some particular advantages in that the latter places a greater emphasis on the laboratory to show evidence of continuous improvement – likely to be particularly important in implementing a rapidly developing technology – and also on the accreditation of both the pre- and post- analytical phases of testing and of clinical effectiveness. In light of earlier discussions on the need to ensure that new genomics services are implemented and evaluated as part of a complete pathway of care, the emphasis on clinical effectiveness as a relevant quality metric is particularly welcome.

One important way in which UKAS assesses the competence of laboratories to perform specific tests is through their successful participation in external quality assurance schemes. These will be discussed in the section below.

Recommendation 7

All laboratories providing clinical pathogen genomics services need to be accredited to the appropriate national / international standards.

14.2.2 External quality assurance (EQA)

External quality assurance schemes are established to enable laboratories to regularly evaluate their performance of specific tests against a set of criteria agreed by a relevant body of experts in the conduct of that test. These schemes are voluntary, but as noted above, participation in them is often an essential component of achieving accredited status for a laboratory's testing activity. EQA schemes typically involve the distribution of a standardised set of samples, or in some cases data arising from samples analysed elsewhere, to participating laboratories. These samples or data are then analysed and interpreted by the receiving laboratories and returned to the EQA assessors for evaluation. Results of the evaluations are then provided to the laboratories to enable them to determine whether their procedures are performing to the necessary standard.

EQA is an effective way of ensuring that individual laboratories conform to the minimum standards expected by their peers for the quality of the results and interpretation they offer their users. Importantly, they also enable comparative analysis of performance across multiple laboratories enabling inconsistencies in the quality of provision to be highlighted and responded to by the relevant professional groups. This is particularly important where the assessment is being made of the qualitative interpretation of a test result rather than the quantitative measurement of the test itself, as qualitative judgements are more susceptible to variability, with significant potential impact on the clinical effectiveness of the services provided on the basis of these judgements. For example, if a genomic service for investigating outbreaks is established, it will be important not only to assess the accuracy with which variation in the genomes between bacteria isolated from patients implicated in the outbreak is measured, but also the way in which this information is interpreted to reflect possible chains of transmission between patients.

It will be vital for the microbiology NEQAS to develop relevant schemes to assess not only the technical performance of the assays underlying these tests i.e. *the whole genome sequencing process itself, but also the interpretation of the test results.*

Currently a national external quality assurance service (NEQAS) for microbiology is operated by Public Health England. This service offers a number of EQA schemes focused on different testing activities in microbiology laboratories, including a range of molecular tests, particularly in the area of virology.

As various laboratories implement genomics services for routine clinical and public health use it will be vital for the microbiology NEQAS to develop relevant schemes to assess not only the technical performance of the assays underlying these tests *i.e.* the whole genome sequencing process itself, but also the interpretation of the test results. As with other areas of microbiology, it is likely that separate schemes may be required for applications of WGS to different pathogens and also for application of WGS analysis to different clinical or public health questions *e.g.* outbreak epidemiology *vs* antibiotic susceptibility testing. The PHE central genomics services laboratory have initiated discussions with NEQAS with a view to developing such EQA schemes, but wider input from the range of organisations and laboratories with expertise in genomics (within and beyond microbiology) will be required to ensure these schemes meet the needs of the service users and laboratories alike.

Recommendation 8

Evaluation and comparison of test performance should span the whole process from sample extraction to clinical report, encompassing assessments of both analytical and clinical validity and clinical utility.

Recommendation 9

The clinical and public health microbiology 'community' needs to work with UKAS and NEQAS to establish standards that can be used to develop appropriate accreditation processes.

14.2.3 Developing and sharing best practice in pathogen genomics investigations

Inevitably, in the development and implementation of new technologies into the health system knowledge and expertise are distributed in a highly uneven manner. Centres of expertise in pathogen genomics, likely to be both the earliest providers and users of genomics services, have a crucial role to play therefore in ensuring that their experience is exploited for wider benefit of the health service. In practical terms this means ensuring that where development of optimised and validated protocols for sequencing, analysis and interpretation have been undertaken, these are assessed by appropriate experts to establish 'best practice' The results should then be made available within the health system to those who may wish to implement such services.

Consistent with this idea is the statement from the recent Barnes review of pathology quality assurance that '*There is a need for a national approach by the professional bodies to produce agreed testing protocols which local laboratory services should implement to reduce […] variation and promote more standardised testing for patients'*.

The existence of such best practice guidelines serves several functions, including:

- Ensuring knowledge of best practice, and opportunity to benefit from it, is available to all laboratories on an equal basis

- Efficiency by avoiding unnecessary duplication of effort in method assessment and selection by laboratories

- Providing a minimum standard of service against which quality assurance evaluations can be made

- Providing a benchmark quality of service that can be used in assessment of the service provision to different patient populations and to inform commissioners

- Supporting the generation of high quality data for wider public health and service development purposes

There are also several challenges to both developing and sharing best practice:

- **Agreeing what constitutes best practice** – where multiple laboratories have independently devised genomic methods for achieving a common goal, a process will be required to compare performance and validity of these methods to determine which, if any, can be said to constitute best practice that ought to be followed by a wider community of practitioners

- **Agreeing who carries the expertise and authority to adjudicate on best practice –** authority is likely to derive from the professional status and scientific expertise of those making the judgement, but agreement on a relevant panel of experts for each pathogen and application of genomics may be required

- **Encouraging uptake without inhibiting service development or local adaptation –** best practices are usually conceived of as advisory rather than mandatory in order to enable their adaptation to meet local needs, and also to enable laboratories able to develop services that exceed the minimum requirements

Clinical and public health microbiology already has a body of professional best practice guidelines, known as Standards in Microbiological Investigation (SMIs), which cover a wide range of laboratory services. These guidelines, which are prepared by sub-groups of experts in the relevant field, are subject to consultation with a wider community of microbiologists prior to finalisation, and which are overseen by the Standards Unit of Public Health England, would appear to be useful templates from which to build best practice guidelines for the use of genomics in clinical and public health microbiology. Indeed SMIs are highlighted as examples of best practice in guideline development by the recent Barnes review of pathology quality assurance. There are also guidelines available for the use of next generation sequencing technology and analysis in human genomic medicine, from which useful information for the construction of microbiology-specific guidelines might be derived.

Recommendation 10

In order to ensure that services are of sufficiently high quality, and delivered in a consistent manner, guidelines (equivalent to SMIs) establishing minimum standards for pathogen genomics services must be developed.

14.2.4 Dynamism *versus* standardisation in genomic service development and delivery

Given that the sequencing and analysis of pathogen genomes are currently both subject to almost continual development and optimisation in the centres of expertise described in chapter 12 and many other locations worldwide, the establishment of meaningful and stable quality assessment processes for pathogen genomics in the short term may prove particularly challenging. Nevertheless it is essential, from a patient safety and service quality perspective, that where multiple services are available from which a user or commissioner can chose, there is a mechanism for assessing which of these services are offering an acceptable standard of care.

To achieve such consensus it will be vital that a body of experts, drawn from across research, public health and clinical domains, and including representatives of relevant professional groups, is established to develop principles for what will constitute best practice in pathogen genomics investigations and metrics for appropriate EQA schemes. Their deliberations may be limited initially to consideration of broad statements about ways of working and the necessary components of any service standards of reporting or documentation of methodology. Over time such a group will be vital to provide a considered view on the validity and utility of novel method and application developments, as these become more established in practice, with the aim of informing wider clinical and public health user groups as to their suitability for use.

Recommendation 11

Develop a national collaborative network of pathogen genomic service providers to share knowledge and best practice, collaborate on service and methodology development and agree standards for clinical and public health service delivery.

14.3 Meeting strategic national public health needs

14.3.1 What is the strategic public health utility of pathogen genomic data collected for other purposes?

As discussed in chapter 12 and chapter 13 the structure of the English health system, with its focus on clinically led commissioning of services suitable for local populations, is likely to drive the establishment of a variety of configurations of genomics services each suited to meeting the needs of their primary users. It is vital however, that in all cases these genomics services view national public health authorities *i.e.* PHE as one of the primary users of the results, even if they are not the direct commissioners of the service or requesters of specific tests. This requirement stems from the view that while the primary purpose of a pathogen genomics based test may be to investigate an outbreak in a given population, or to inform the treatment of a particular patient, that embedded within the data generated through that test may be information of significant public health utility. Uses of such data include:

- **National and international surveillance –** this includes surveillance of novel pathogenic strains and the genetic determinants of their pathogenicity, changes in distribution of antibiotic resistance or its genetic determinants and identification of genomic correlates with vaccine response or escape

- **Outbreak detection –** aggregated genomic and epidemiological data across multiple local services may allow for the more accurate detection of geographically distributed outbreaks of national or even international public health significance

- **Policy evaluation and development –** accurate genome-based surveillance of infectious disease epidemiology, gathered from aggregated local testing laboratories, could contribute significantly to the evaluation of existing national and international policies for the management of these diseases and the more evidence-based development of new policies

- **Enabling new test, technology and therapeutic development –** to facilitate the expansion envisaged in the range of genomics-based test provision into pathogen identification and drug susceptibility testing (beyond tuberculosis), a vast quantity of clinically validated data about the relationship between genome sequence, pathogen identification and antibiotic susceptibility testing by existing phenotypic methods, and clinical outcomes will be required. There is a significant public health interest in the aggregation of this data and its interrogation for the purposes of developing novel tests, based on emerging or currently established genomic technologies, and also novel therapeutics and vaccines.

14.3.2 What will it take to realise the population health utility of pathogen genomic data collected for other purposes?

Realising the benefits for population health of the introduction of individual genomics services that serve particular local needs will require significant coordination of activity across all service providers, and the establishment of significant additional infrastructure to support this endeavour. Fundamentally, this is a challenge of data sharing, aggregation, interpretation and evaluation. These challenges are addressed in detail in chapter 15, chapter 16 and chapter 17, here we highlight the key principles that will need to be adhered to by laboratories contributing to the strategic public health aims described above:

- Ensuring timely access to relevant genomic and clinical data to enable necessary public health actions and future development of therapeutics or preventive interventions

- Ensuring that genomic data is appropriately linked to epidemiological and clinical data required to contextualise its significance

- Ensuring data collected is of sufficient quality, and in standardised, interoperable formats that allow for its aggregation and analysis

- Collation and archiving of relevant clinical samples, isolates and / or extracted DNA for future use in research studies, therapeutic and vaccine development and retrospective outbreak investigations

Implicit in these requirements is a pressing need for consensus to be developed across providers of pathogen genomics services as to how these principles can be codified and transformed into action. This will include what infrastructure is necessary to support them, what governance policies and standards will be required to support their fair and effective implementation and what incentives and sanctions should be put in place to ensure adherence to them. Furthermore, funding mechanisms will be urgently required to support the development of the necessary infrastructure and capacity within and beyond public health microbiology services to translate the resulting information and knowledge into both immediate public health responses and longer term programmes of therapeutic, vaccine and healthcare policy development.

Lessons can be learned from the management of existing strategic public health initiatives in microbiology, from infectious disease notification (where legislation enables financial penalties to be imposed in response to failures in compliance) to voluntary and mandatory surveillance schemes for healthcare associated infections and a range of other pathogens, which are monitored variously by NHS England, Public Health England and the Department of Health. From a behavioural point of view, consideration must also be given to how participation in such strategic public health initiatives can be made as easy as possible by minimising the burden of reporting and sharing information, and by maximising the utility of this effort for the contributors through the provision of feedback and access to analysis of their data relevant to their service. Perhaps a suitable model for comparison is the National Cancer Registration Service, where great emphasis is placed on making mandatory monthly data submission as simple as possible and on ensuring cancer services are provided with meaningful feedback from the data about their own performance.

While responsibility for the coordination and oversight of these strategic public health initiatives should be the responsibility of PHE, it will also be vital that requirements to contribute to these are built into the mandates and commissioning arrangements of all of the relevant organisations.

Recommendation 12

Realisation of the strategic public health benefits of the implementation of pathogen genomics services will require coordinated action amongst providers and users to develop underpinning policies and procedures to support co-operation and inter-operation of services. These efforts should be led by Public Health England but be explicitly supported by all relevant health service and policy making organisations.

Recommendation 13

Criteria must be established to decide under what circumstances sequenced pathogen isolates (or related clinical materials) must be stored for future public health use, timescales for any storage requirements and sources of funding to ensure sustainability of any sample archives created.

Recommendation 14

Additional investment will be required, above that envisaged for the development of individual pathogen genomics services, to build the infrastructure and capacity required to realise the broader and longer term public health benefits of the implementation of pathogen genomics for disease surveillance, treatment and prevention.

14.4 Beyond pathogen genomics in humans – the One Health approach

Tackling infectious disease requires an integrated approach that considers the contribution of humans, animals, plants and other environmental domains - an approach articulated by the 'One Health' movement.

To this point in our report we have discussed the impact of pathogen genomics purely in the context of investigating infections in humans and the delivery of clinical and public health services directly to those human populations. It is important, however, when considering the wider strategic impact of pathogen genomics on individual and population health to consider the interactions between humans and their environment, and particularly the animals and animal based products *e.g.* foods with which they come into contact. Animals contribute to human disease in a number of ways, primarily:

- Reservoirs of infections that can be transmitted to humans (zoonoses)
- Source of development of antimicrobial resistance
- Carriage of pathogens that may enter the human food chain

Tackling infectious disease therefore requires an integrated approach that considers the contribution of humans, animals, plants and other environmental domains. This approach is articulated by the 'One Health' movement, which aims to encourage the collaborative working of organisations with interests in each of these domains *e.g.* medical services, veterinary services and environmental organisations to address significant threats to global health, including but not limited to infectious disease.

In the UK context, the recognition by policy makers of the need to take a 'One Health' approach to tackling infectious disease is best exemplified by the recently published antimicrobial resistance strategy, a joint policy documents of the Department of Health and the Department for Food, the Environment and Rural Affairs. This policy, co-sponsored by the Chief Medical Officer and the Chief Veterinarian, sets out clearly that the problem of increasing antimicrobial resistance originates in both human and animal health practice, and therefore the solutions will require coordinated and concerted action in both spheres.

In the context of this report on pathogen genomics, there is evidence that pathogen whole genome sequencing can be informative in the surveillance and control of infectious disease at the interface of animals and humans in a number of ways:

- Identifying sources of emerging antimicrobial resistance in pathogens, and their transmission between animals and humans
- Surveillance and outbreak investigation of foodborne infection that encompasses both the infected humans and the underlying food supply chain
- Surveillance and early identification of potentially zoonotic infections circulating in animal populations

Surveillance and investigation of infectious disease across the boundaries of human, animal and food health is already supported by a number of organisations in the UK, including Public Health England, the Food Standards Agency and the Animal and Plant Health Agency (formerly the Animal Health and Veterinary Laboratories Agency). For example they operate a group on human and animal infections and risk surveillance (HAIRS) whose purposes include facilitating horizon scanning, risk identification, risk management and communication.

While PHE has made great strides in the development of genomics services for the investigation and surveillance of infectious diseases in humans, and work is underway to develop similar services within the APHA, there is clearly a significant opportunity for these organisations to develop a coordinated strategy to ensure that as each develops the use of pathogen genomics for its own primary purpose *i.e.* the epidemiological investigation and surveillance of human or animal disease, opportunities to share information and experience pertinent to the zoonotic transmission of disease to humans, the development of antimicrobial resistance, and foodborne infections are identified and supported through necessary policies, agreements and infrastructure.

The benefits from such an approach should manifest themselves in improved public health, through earlier identification of potentially significant emerging zoonoses, a greater understanding of the dynamics of antimicrobial resistance development and thus the development of more effective policies with which to control it. This should result in increased effectiveness of investigations into foodborne outbreaks where clearer epidemiological linkages to the sources of infection in humans are made by the genomic sequencing of bacterial isolates from both infected patients and potentially contaminated foods.

Recommendation 15

Existing links between the infectious disease aspects of animal and human health services should be exploited and strengthened to ensure that synergies in the developments of their genomics programmes are realised and a 'One Health' approach to managing infectious disease threats can be developed where appropriate.

14.5 International dimensions to the realisation of the benefits of pathogen genomics

The strategic public health benefits that can be realised from the introduction of pathogen genomics discussed above are all contingent, to some extent, on external factors that prevail beyond the borders of the British Isles. Infectious disease does not respect national boundaries, and international travel, globalised food, livestock, plant and commodity markets all provide ample opportunity for the importation and exportation of infections. In this context, effective surveillance, epidemiological investigation of outbreaks and detection of emerging infectious disease threats rely on effective transnational organisations. Through their networks and routes of communication they can coordinate the efforts of national bodies in tackling these challenges.

The realisation of the benefits of pathogen genomics, in tackling infectious disease threats that have a major international dimension, such as emerging zoonotic infections, development of antimicrobial resistance and foodborne disease, will depend to a significant extent on the coordination of pathogen genomic information from microbiology services in countries around the world. While it is unrealistic to expect that all countries with whom we share infectious disease risks will be in a position to implement genomic surveillance and outbreak control as part of their management of infectious disease, it will nevertheless be vital that where such services are put in place coordination is achieved, at least at the level of exchange of data and expertise to ensure that transnational threats can be rapidly identified and responded to.

In anticipation of this requirement an international nongovernmental organisation, the Global Microbial Identifier (GMI), has been established whose long term aim is to develop a platform through which countries can share and access pathogen genomic data from one another. This information, along with sufficient metadata to interpret its significance, will enable 'real time global genomic epidemiology' for infectious disease. The GMI project, run by volunteers from scientific and public health organisations from around the world, is also engaged in developing community standards for the generation and reporting of genomic data - standards that will be essential in ensuring that when dealing with international outbreaks, data generated in individual countries is of sufficient quality to be relied upon by microbiologists in other countries and is in formats that are interoperable and that support data aggregation and analysis. While there remain significant political, social, regulatory and logistical challenges to achieving the goal of the GMI project, its principles are nevertheless important to recognise and act upon.

Recommendation 16

Organisations leading on the development and delivery of pathogen genomics in the UK should work with and show leadership within transnational organisations and specific international genomics focused initiatives to ensure that best practice is shared and sufficiently standardised, or at least interoperable datasets are developed and regulatory barriers to effective genomic and metadata exchange are addressed.

14.6 Conclusions

There are two key reasons why it is essential to take a whole system view of the implementation of pathogen genomics, as well as considering how it can be optimally achieved at the level of individual services. It is important to ensure that collectively the individual services established provide high quality care, available nationwide on an equitable basis. It is also equally important to ensure that these individual services are configured and resourced in a way that enables them to participate in the delivery of wider public health benefits, and that policies are in place to ensure frontline pathogen genomics service providers within the NHS and PHE interact with other public and private sector organisations, as appropriate, whose participation is required to realise more fully the potential benefits of this technology to improve population health.

15 The necessity of data sharing: principles and practicalities

Pathogen sequence data can be viewed as a multifaceted resource, with each individual genome sequence having many potential uses in both clinical and public health settings. Realisation of the benefits of this 'multi use' attribute of pathogen genomic data has the potential to transform microbiology services by replacing several processes that occur across diagnostic and reference laboratories with one technology.

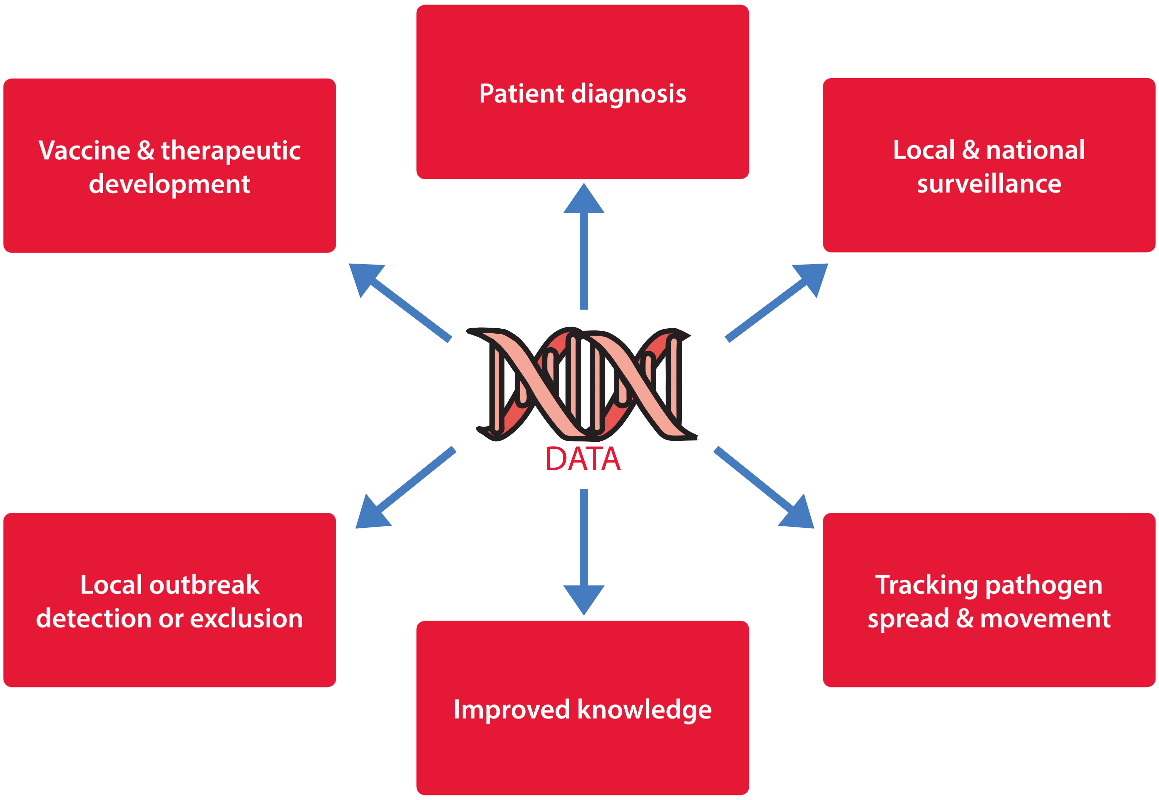
15.1 Introduction

As discussed in earlier chapters, a single pathogen genome can provide information on:

- Pathogen and strain identity
- Drug susceptibility
- Vaccine efficacy
- Epidemiological relationships with other isolates in outbreaks
- Emergence of new infections
- Emergence or transfer of antibiotic resistance
- New therapeutics and vaccine development

Crucially, different stakeholders across the network of organisations and individuals involved in managing infectious disease will therefore require access to overlapping elements of the same pathogen genomic information for the delivery of these diverse, yet complementary functions (figure 15.1). From a clinical and public health perspective end-users will extend across microbiology and molecular diagnostics, hospital infection control, epidemiology, public health surveillance and management. Academics and researchers in the commercial settings will also derive utility from access to pathogen sequence data; for example through improved understanding of pathogens (*e.g.* their evolution, transmission), or for vaccine and therapeutic development.

Figure 15.1 The multifaceted applications of pathogen genomic data


Patient diagnosis
Vaccine & therapeutic development
Local & national surveillance
DATA
Local outbreak detection or exclusion
Tracking pathogen spread & movement
Improved knowledge

15.2 Why genomic data sharing and aggregation is essential for developing and delivering maximally effective services to manage infectious disease

The theoretical advantage offered by the use of genomic data for microbiological investigations is that it is capable of providing a higher resolution and more accurate description of the clinically and epidemiologically relevant properties of an individual pathogen, and also an understanding of how these relate to the properties of other relevant pathogen samples; for example within a patient at a single point in time; within a patient over time; in an outbreak; or within a community / population. It is important to appreciate, however, that the extent to which this theoretical benefit is translated into a real world advantage over existing methodologies in terms of the sensitivity and specificity of diagnostic or discriminatory microbiological tests depends on the depth and breadth of genomic data available to support the investigation being undertaken.

Below we outline how aggregating and sharing data support the two key activities of characterising individual pathogens, and determining the relatedness of similar pathogens:

- **Characterising individual pathogens –** the sensitivity and specificity with which patterns and relationships between genomic features and clinical or phenotypic characteristics in single pathogen species can be identified is closely related to the amount of background data available for analysis. For example, as more samples are available for comparison it becomes easier to differentiate between variants that are causally related to the appearance of resistance to a particular drug, or animal to human transmissibility, and others that are associated by chance.

  Potential outcomes:

  - More sensitive and specific detection of emerging or evolving pathogens that may pose a particular threat to human health

  - Research and development leading to more accurate genome guided diagnosis, prognosis and treatment selection algorithms

  - Improved evidence base to guide AMR and wider infection control policies

- **Determining relatedness of similar pathogens –** the sensitivity and specificity with which outbreaks can be detected and their true extent delineated depends particularly on the number of pathogen isolates available for genomic analysis. For example, aggregating genomic data on samples collected over time and different locations within a hospital allows putative outbreak isolates to be placed in historical context to determine whether they are significantly more similar to one another than epidemiologically unrelated isolates, which would support their identification as an outbreak cluster. At both a local and national level, the aggregation and comparison of genomic data collected over time and across a range of locations from isolates collected as part of routine infectious disease surveillance activities will also support the identification of outbreaks that span wider ranges of time and geography, and which may otherwise have gone undetected by conventional epidemiological methods.

  Potential outcomes:

  - Accurate tracking of routes of pathogen transmission and identification of sources of outbreaks

  - Timely detection and delineation of outbreaks which will expedite investigations and response measures *i.e.* real-time genomic epidemiology

  - Earlier recognition of changes in effectiveness of vaccines and vaccination programmes

- A resource for the development of new or improved outbreak analysis tools, particularly for the analysis for large data sets

The sharing and aggregation of data is essential if the full potential of pathogen genomics in healthcare is to be realised. As the amount and availability of data increases, so do the probabilities of identifying emerging patterns and important new associations. For example, more data on a particular pathogen (depth) will enhance and enable the associations between genotype and clinical features (*e.g.* the correlation of gene x with resistance to a given drug). The collation of data will also facilitate vital scientific advances that will feed into the improved clinical application of pathogen genomics. This includes the refinement of available reference genomes for supporting analysis, improvement in the evidence-base on the molecular evolution rate of pathogens, and so increased accuracy of predictions of relatedness of pathogens in suspected outbreaks. Additionally increased availability of data across broader geographic spaces (breadth of data), will be essential in identifying connections relevant to local, national (and international) outbreak investigations (*i.e.* determining whether an isolate in one location is related to isolates in other regions).

15.3 Developing policies and mechanisms to ensure data 'sharing'

Realising the outcomes described above will require the construction of a comprehensive, accurate and curated catalogue of genomic data on all sequenced pathogens, as well as relevant phenotypic, clinical and epidemiological data associated with each sample. Such a catalogue is essential to enable both the immediate and longer term improvements outlined above in both clinical care and public health.

Simply put, pathogen genomes, when analysed in isolation from wider genomic and clinical data sets are almost entirely uninformative. Conversely the robustness and effectiveness of any pathogen genomics guided infectious disease management service will be proportional to its ability to place each individual analysed genome in the most complete and high resolution genomic and clinical context possible.

By extension from the above statements, the success of any envisaged system of pathogen genomics guided infectious disease management service will rely entirely on the ability and willingness of those generating pathogen genomic data to share this information. For those who will use the data, timely and uncomplicated access at the appropriate time, accompanied by adequate data analysis / visualisation tools (chapter 16) will be necessary and this will facilitate improved infectious disease management. A coordinated and effective mechanism will be required for managing data arising from a variety of clinical and public health sequencing initiatives.

15.4 Definitions

Here we set out a number of definitions pertinent to discussions in the subsequent sections, particularly around sharing of different elements of data, and the interpretation of terminology around data 'sharing' itself.

The term 'data' in the context of this chapter encompasses a number of constituents of information that may arise as a consequence of pathogen genomic investigations. More specifically 'data' can denote any of:

- *Pathogen genomic data,* including:
    - **Raw sequence data** - the unprocessed genomic data obtained directly from the sequencing machine often obtained in a file format called FASTQ
    - **Assembled or aligned sequence data** - a representation of the genomic sequence of an organism as predicted using 'assembly' software. This would include both *de novo* assemblies, as well as reference guided assemblies (chapter 4)
- *Metadata* or *accessory data*: used interchangeably in this chapter, refers to 'data' about 'data' –*i.e.* information about the sequence data. Metadata can be further divided to denote:
    - **Contextual data –** includes information on the sequenced sample, *i.e.* what organism, the sampling source, and environment, associated clinical information about the patient from which it was obtained, or epidemiological data on its suspected relatedness to other samples
    - **Sequencing metadata –** details of the sequencing process, *e.g.* routes of data flow, sequencing method
    - **Secondary** or **derived data –** inferences about the data resulting from analysis

References to data sharing may encompass:

- **Data sharing –** denotes the distribution of data by the 'data generators' to other users who did not generate the datasets
- **Data release –** here implies the sharing of data in an open access manner (*i.e.* sharing in the public domain without restrictions on access)
- **Data deposition or data submission –** implies the sharing of data within a specific domain and / or for a specific purpose *e.g.* sharing with authorities for public health service delivery

15.5 Maximising the utility of data with an effective data sharing strategy

15.5.1 Sharing both genomic data and clinical / epidemiological metadata – the challenge

An effective data sharing strategy to support genomics informed management of infectious disease, must include both genomic data and additional data about the genomic data (metadata; see15.4 for definitions); for example location of where the pathogen was isolated, the date of isolation. The greater the detail and availability of the metadata that accompanies any genomic dataset, the more effective any clinical or public health intervention based upon its analysis are likely to be.

However as clinical and epidemiological metadata potentially include information from which individuals with infectious diseases may be identified, risking their right to medical confidentiality, the breadth of access to these data will necessarily be more restricted than what is required for pathogen genomic data, which in isolation from clinical or epidemiological metadata will not generally reveal confidential health information about the individuals from whose infections it was obtained.

The extent of the genomic data and metadata that need to be shared and with whom these need to be shared, in order to deliver the benefits outlined in figure 15.1, varies according to the application of genomics and the pathogen to which it is being applied. Three core principles can, however, guide this decision in each case:

1. The more data (in terms both of number of data items and detail of information within each data item) that can be shared and aggregated into linked datasets, the greater the likelihood that any subsequent clinical or public health investigation or research programme based upon these datasets will yield benefit to patients or populations

2. The broader the range of professional and expert groups who are able to access the data, the greater the likelihood of clinical or public health benefit from pathogen genomics being realised

3. It will be necessary to restrict with whom different data items are shared, particularly in the case of sensitive clinical and epidemiological metadata, in order both to meet existing regulatory requirements with respect to patient confidentiality and to ensure that trust is established and maintained amongst commissioners, providers and users of potential pathogen genomics services

Balancing the potential public health benefits of data sharing (principles 1 and 2 above), with the risks to patient confidentiality associated with this practice (principle 3) is the key challenge to any effective data sharing strategy. The benefits of data sharing are set out in 15.2. The ethical, legal and social context underlying the risks associated with data sharing are set out below.

15.6 Deciding which elements of data to share and with whom: The ethical, legal and social considerations

Data sharing raises a number of distinct ethical, legal and social challenges. The extent to which pathogen genomic data and metadata can legitimately be shared depends on the associated risks and benefits. Although the different components of pathogen genomic data and associated metadata are well-defined (15.4), clear consensus on how widely each component should be shared (*e.g.* with no restrictions and publically available, or restricted to only authorised healthcare / public health workers), has yet to be established and formalised, as have acceptable use policies for data shared wider than public health networks.

15.6.1 Data sharing for healthcare and health protection: The legal and regulatory context

Since UK data protection legislation is framed in terms of the type of data, the extent to which it is identifiable to the data controller and the application to which it is put, interpreting how the regulatory framework will apply to the sharing of pathogen genomic data and associated metadata is not straightforward. Firstly the Data Protection Act 1998 applies to 'personal identifiable data' and not to data that is de-identified or anonymised. Moreover, the degree to which data can be identified, will depend upon how different data are linked together. UK law establishes special safeguards for certain categories including health data, but exemptions for processing data for health purposes also apply which allow data to be processed for medical purposes, including medical diagnosis and treatment, and shared between health professionals or between those who owe a similar duty of confidentiality [DPA Schedule 3 paragraph 8]. Other legal concepts like the common law of confidentiality or the protections in the Human Rights Act 1998 are also not absolute, and depend heavily on context and on professional judgement. For example, the common law recognises that a person's confidentiality may be breached if the purpose of the disclosure is to warn another person who is imminent danger of serious harm [GMC Confidentiality guidance]. Similarly, Article 8 of the Human Rights Act 1998, which establishes a right to respect for privacy and family life provides a right interference by governments 'where necessary in a democratic society'… for 'the protection of health and morals'.

15.6.2 Risks of sharing different types of data

The risks of sharing genomic, clinical and epidemiological data arising from the management of infectious disease include:

- Infected individuals being identified and their privacy invaded, resulting in discrimination or stigmatisation. In the worst cases, this could involve being denied employment or insurance

- From an institutional perspective, identification could result in increasing numbers of legal claims or a wider loss of public confidence or public trust *e.g.* through association of healthcare facilities or food manufacturers with implied responsibility for causing individual infections or outbreaks

As discussed above, some elements confer greater risk of identifiability than others and the distinction between genomic data and metadata, as well as the categories of metadata, is therefore pertinent to considerations of how openly data ought to be shared and consequently the choice of data storage and access solutions. In the context of this chapter, the risks / benefits attached to the release of different data types forms the basis of judgement around with whom and when genomic data and metadata might be shared.

- **Raw sequence data derived from purified microbial cultures –** this data is unlikely to contain any contaminating human genomic sequences, and so its public release in isolation of any metadata carries little risk of patient or host individuals' identification

- **Sequence data derived from uncultured samples (metagenomics) –** where the clinical samples are of human origin, the metagenomic sequences reads will comprise a significant proportion of human genomic sequences intermingled with pathogen sequences. Safeguards would be needed to ensure human sequence reads are robustly identified and discarded prior to the release of data into the public domain

- **Metadata –** in the context of pathogen genomics, metadata (particularly clinical information) are normally the most sensitive elements of data, as their release into public domains may compromise an individual's or organisation's privacy. The level of sensitivity attached to metadata varies according to:

    - The category, and specific piece of metadata; for example information on the quality of the sequence data is unlikely to compromise individual privacy, whereas contextual information on geographic source and host details (age, sex), can potentially do so

    - The pathogen in question, as certain categories of pathogens (*e.g.* sexually transmitted infections such as gonorrhoea and HIV) carry greater social stigma than others

More work is needed to determine the risks attached to the release of different elements of metadata within a pathogen specific context while the safeguards necessary to mitigate against these risks are yet to be clearly determined.

A balance therefore needs to be drawn between minimising risks to individuals or organisations and releasing enough metadata to enable the greatest possible health benefits for patients and populations. More work is needed to determine the risks attached to the release of different elements of metadata within a pathogen specific context while the safeguards necessary to mitigate against these risks are yet to be clearly determined. Different protocols will need to be adopted for each pathogen, as the balance of risks to patients and public health benefits is likely to be affected by the characteristics of the pathogen (*e.g.* in terms of likely morbidity and mortality: infectivity; treatability and drug resistance). Criteria for the release of minimal metadata will need to be defined, minimal metadata being the level and types of metadata that can be made available in the public domain to maximise utility of the accompanying genomic data yet minimising risks to individuals. Data that cannot be released into public domains, but is needed by authorised healthcare and public health professionals for service delivery should remain within a suitable secured access database.

Recommendation 17

When considering data release to a publicly accessible database, stakeholders should adopt proportionate safeguards that balance the need to protect the interests of data subjects, particularly relating to privacy and confidentiality, against the likely benefits of proceeding with data sharing.

Recommendation 18

Raw genomic data and minimal metadata ought to be shared as widely as possible (following appropriate QC and assuming public release is approved) preferably through public data repositories to ensure long term sustainability.

Recommendation 19

Criteria for defining what minimal data sets are appropriate for release to publicly accessible databases should be developed, with risk assessments being undertaken to identify in particular which elements of metadata can be released publicly for each pathogen. PHE (and their Office of Data Release) would be best placed to deliver on this, along with NHS input.

In the following section we consider two models for managing the balance between the benefits of sharing as much data as widely as possible, with the risks associated with this from an individual and organisational perspective.

15.7 Data sharing to deliver effective pathogen genomics informed public health services

15.7.1 A restricted access model

The primary uses of pathogen genomic data by public health practitioners will be in the domains of management of outbreaks of infectious disease (genomic epidemiology) and longitudinal surveillance of infectious disease to inform vaccination programmes, monitor AMR *etc*. The effectiveness of these activities will depend on maximising the availability of genomic, clinical and epidemiological data, which may be generated in a range of both public health and clinical *i.e.* NHS settings, to enable decision making by the relevant professionals.

Given risks to individuals and organisations associated with combined genomic and metadata release, the simplest model of data sharing that would ensure the effectiveness of these applications would be to require all public health and clinical professionals generating genomic and associated clinical / epidemiological metadata to share this in its entirety within a closed, restricted access database that was only accessible by authorised healthcare professionals with a demonstrable need to use the data to fulfil either patient care or population level prevention of infectious disease.

The principal advantage of this model is:

- Data is shared only between those who have a pre-existing obligation of confidentiality

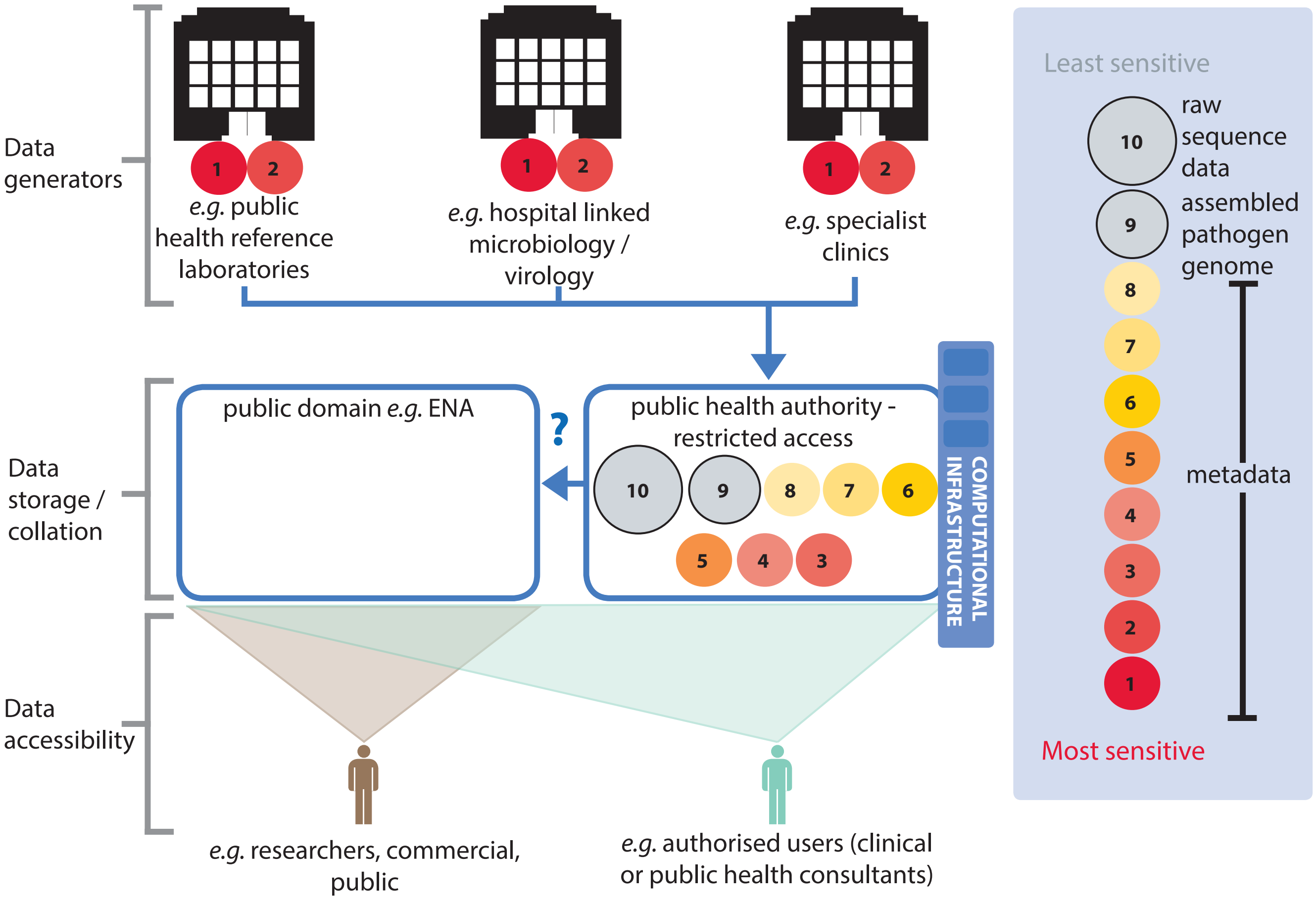
- Potential concerns about confidential data (about patients or organisations) being placed in the public domain are allayed, making implementation easier to achieve from a regulatory perspective

The principal disadvantages are that:

- Failure to share these data beyond the immediate network of healthcare professionals in England that would be generating these data would stifle the research and innovation required to expand the utility and potential applications of pathogen genomics in the longer term

- Lack of access to the data by overseas healthcare professionals would inhibit efforts at establishing transnational genomic surveillance and epidemiology initiatives, such as GMI, that seek to address the key challenge of managing the spread of infectious disease between countries

Figure 15.2 A restricted access model for data sharing



Data generators
1
2
e.g. public health reference laboratories
1
2
e.g. hospital linked microbiology / virology
1
2
e.g. specialist clinics
Data storage / collation
public domain *e.g.* ENA
?
public health authority - restricted access
10
9
8
7
6
5
4
3
COMPUTATIONAL INFRASTRUCTURE
Data accessibility
e.g. researchers, commercial, public
e.g. authorised users (clinical or public health consultants)
Least sensitive
10
raw sequence data
9
assembled pathogen genome
8
7
6
5
4
3
2
1
metadata
Most sensitive

The size of circles (not to scale) are indicative of the relative data storage burden (computational disc space), of the different subsets of data. Raw genomic data will consume the greatest disc space (therefore cost more to store than other data types).

15.7.2 A two-tier data sharing model that maintains data confidentiality where necessary and enables public access to less sensitive data

Our consultation with pathogen genomic data experts, identified a two-tier data sharing model that maximises the utility of genomic data and associated metadata whilst minimising the risks to patients and organisations as preferable to the 'restricted access only' model described above (figure 15.2).

The key feature of this model is that it only proposes to limit access to more sensitive levels of metadata, whilst placing all other genomic and less sensitive metadata, that pose minimal threats to patient or organisational confidentiality, in the public domain where it can be accessed by the widest possible range of researchers and public health and clinical practitioners. The sensitive metadata would only be accessible to the subset of regulated healthcare professionals who have a justifiable need to use them in discharging their clinical or public health duties of care, for example through conducting detailed outbreak investigations that relied on the ability to link genomic data with patient identifiable information such as their precise location and clinical symptoms.

The principal advantages of this model are:

- Public-level data sharing will maximise the number of potential end-users of the data and therefore improves the chances and pace of delivering the wide-ranging benefits pathogen genomics

- Data sharing in the public domain would complement existing international initiatives aiming for effective and responsible data sharing, for example the GMI, and the Global Alliance for Genomics and Health. Moreover these initiatives are already investing substantial effort into tackling the technical, practical, ethical, and legal barriers to (international) data sharing

- Since large-scale public DNA sequence repositories are already in existence, these could potentially be configured to hold the data being generated (as is the case for the GMI initiative), from clinical and public health sources, and could take the economic burden of storing some data, particularly genomics data which consumes the most disc storage space

- Public sharing of data funded by the taxpayer would be consistent with the UK governments open data strategy

For a two-tier access model to operate the following barriers would have to be overcome:

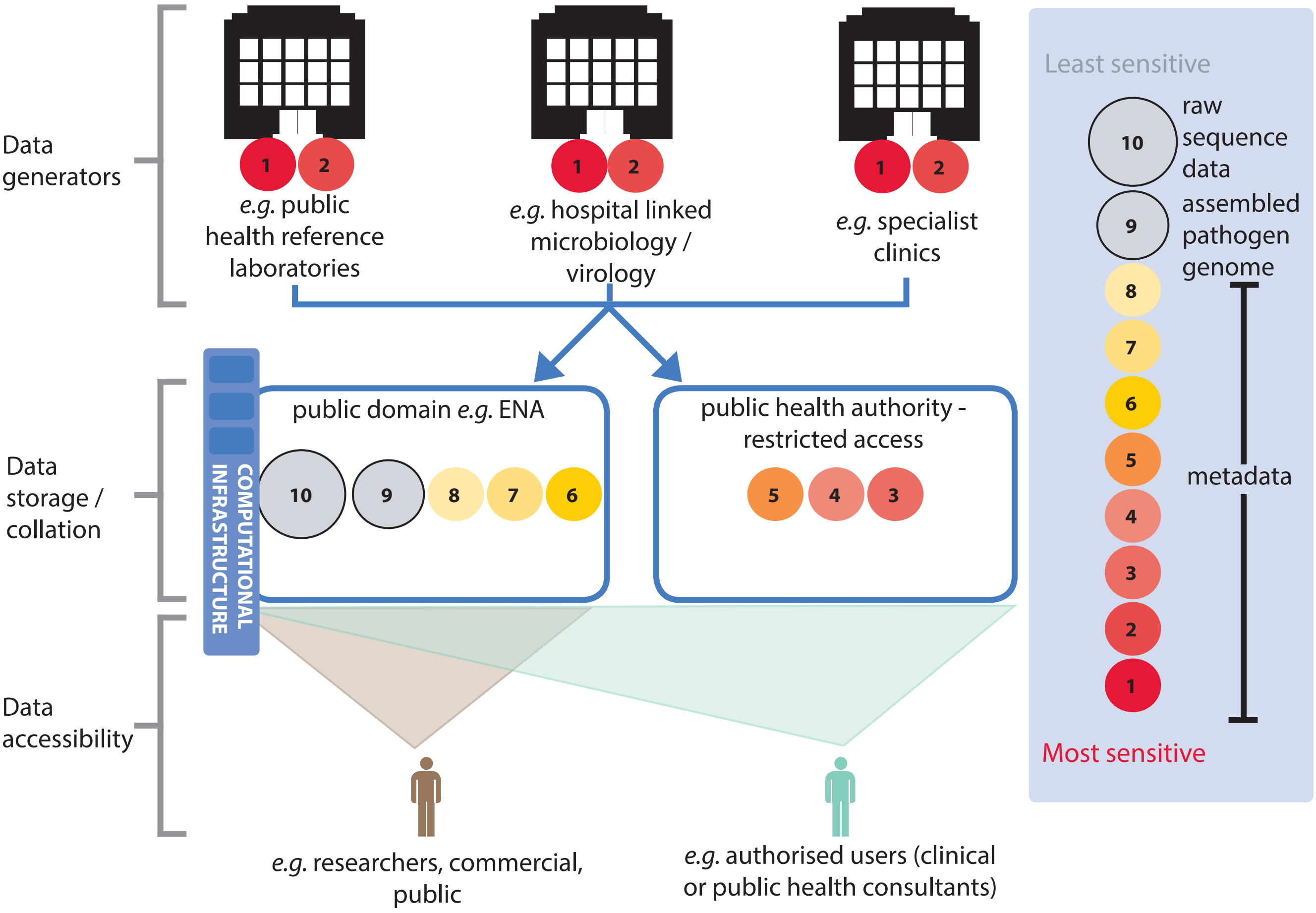
- Reaching agreement on the types of metadata that should be subject to restricted access *versus* public release for each pathogen (15.7 – 15.8) and application of genomics to its analysis.

- Establishment of effective infrastructure and appropriate service level agreements (between public health authorities and any public database resource) to support the collation and analysis of genomic data arising from healthcare investigations.

- Ensuring interoperability of the public access and restricted access databases to allow the accurate linkage of genomic data and metadata across the two tiers

- Creating sufficient computation capacity to facilitate real-time analysis of data for outbreak detection applications

Figure 15.3 A simplified vision of a two-tier data sharing strategy



Data generators
1
2
e.g. public health reference laboratories
e.g. hospital linked microbiology / virology
e.g. specialist clinics
Data storage / collation
COMPUTATIONAL INFRASTRUCTURE
public domain e.g. ENA
10
9
8
7
6
public health authority - restricted access
5
4
3
Data accessibility
e.g. researchers, commercial, public
e.g. authorised users (clinical or public health consultants)
Least sensitive
raw sequence data
assembled pathogen genome
metadata
Most sensitive

The size of circles (not to scale) are indicative of the relative data storage burden (computational disc space), of the different subsets of data. Raw genomic data will consume the greatest disc space (therefore cost more to store than other data types), and so its longer term storage would be better suited in a consolidated repository build for high volume data storage.

15.7.3 Deciding which data sharing model should be adopted

There was a clear preference amongst our consultees and stakeholders that from an effectiveness perspective, the two-tier model of data sharing represents the optimal solution to maximising the benefits of the development and delivery of pathogen genomics services. However, it is also clear that from a logistical and regulatory perspective such a model will be more challenging and time consuming to build than a simpler restricted access system that could be constructed, controlled and maintained within the health service. Given that pathogen genomic data are already being generated by clinical and public health microbiology practitioners in England, there is a clear need to move quickly to ensure this data is collated and shared to maximise the effectiveness of the services from which it is originating.

A decision will, therefore, have to be taken by health policy makers as to whether to rapidly construct a simpler, restricted access database that could at least facilitate this limited level of data sharing in the short term, albeit with limitations in its utility in the longer term, or whether to focus on overcoming infrastructure and regulatory barriers to achieving the optimal two-tier model. If the latter approach is taken, a transitional solution, enabling data sharing in the short term, would still be essential to avoid inhibiting progress that is already being made.

Regardless of which model of data sharing is adopted it will be necessary to determine:

- How to ensure data is submitted to any identified repositories (incentivising and supporting data sharing (15.8)

- Appropriate timing of data sharing (15.9-15.10)

- How to undertake and enable data collection and sharing (15.11)

- Where data ought to be collated (*i.e.* which database/s), and who would be best placed to manage and maintain these repositories. (15.12-15.13)

- Whether (or what level of) quality control procedures should be applied to data collected for the aforementioned and generated for clinical / public health decision making (17.2)

15.8 Incentivising and supporting data sharing

We anticipate that implementation of pathogen genomics is likely to result in data being generated in range of locations and organisations from smaller-scale diagnostic facilities to larger-scale public health laboratories, and commercial providers. The distribution of data generation, and variation in the nature of data collected and the methods used to do so across locations, pose significant challenges to the effective aggregation of data. Those performing sequencing will have varying:

- Levels of informatics capacity and expertise to manage the timely and cost effective transfer of data

- Incentives / disincentive to share data

- Methods and technologies for generating raw sequence data

- Methods for processing raw data into assembled genomes, performing further analysis and quality controlling this process (see 17.2)

Policies to incentivise timely data deposition by all providers of clinical and public health pathogen sequencing are therefore needed, particularly to enable effective genomic-based national surveillance and outbreak detection. Experience has demonstrated that compliance with data sharing guidelines in healthcare and research is suboptimal[129, 130] unless there are clear mandates to share data and / or sanctions for those not complying. Without a clear mandate to submit data, there will be little incentive for providers to invest in the time, training and infrastructure required for data transfer and deposition. Moreover if a data sharing mandate for public benefit is not implemented promptly, commercial interests might otherwise dominate and result in proprietary genome sequence databases being developed, limiting access to key data sets for public health surveillance activities[131]. Ultimately genomic and clinical data is likely to become trapped in silos, or its release to authorities delayed, unless active measures are taken to prevent this.

Recommendation 20

It must be mandatory for all providers of NHS or PHE pathogen genomic investigations to make sequence data and all other necessary clinical and epidemiological data available for use by legitimate NHS healthcare and public health professionals within agreed timeframes, for the purpose of delivering their stipulated functions. A mandate needs to be implemented urgently to prevent data that is currently being generated from being lost in silos.

Longer term approval of a data deposition mandate is contingent upon the compliance with the mandate generating or demonstrating value to the organisations being asked to comply. For example if hospitals routinely have to submit data to public health authorities, then feedback on how this data is being used and it's bearing on the hospital's (infection control) performance is more likely to generate endorsement of a mandate and compliance with its terms. Similarly the ability of individuals and organisations to comply with a data sharing mandate will depend on the availability of adequate infrastructure to support the curation of data and its transfer into the stipulated repository, as well as the availability of an appropriate repository itself.

Recommendation 21

The benefits of data collation and risks of not aggregating data should be articulated to those being mandated to submit data. A feedback or reward strategy should be developed to gain longer term accord with and practical support for a data sharing mandate, and investment made in adequate infrastructure to enable data deposition at the practical level.

As other governmental bodies such as the Food Standards Agency, or the Animal and Plant Health Agency advance with pathogen sequencing initiatives, consideration should be given as to how cross-organisation data exchange might be implemented, particularly in the interests of detecting and managing infectious disease spread across species.

15.9 Timing of data sharing

15.9.1 Timing of data sharing for immediate delivery of patient care and population health protection

From a public health perspective, an appropriate time frame for depositing genomic data and metadata in shared access databases is fundamental to the ability to inform management of infectious disease. The more rapid the deposition of data, the sooner outbreaks can be detected and the higher the chances are of any public health intervention significantly limiting onward transmission and reducing morbidity and mortality.

The more rapid the deposition of data, the sooner outbreaks can be detected and the higher the chances are of any public health intervention significantly limiting further spread.

Barriers to timely sharing of data include:

- **The incentive structure for scientific advancement -** this concerns the delayed release of data by the group generating it, in order to assure first academic publication of any analysis
- **Ethical concerns, legal and regulatory issues –** risks to personal privacy and confidentiality (see 15.7)
- **Uncertainty surrounding data ownership –** this concerns the ambiguity around who has (first) rights to benefit commercially, financially, and academically from data generated by one group or organisation, while making the data available for others to access and utilise
- **Financial / resource restraint –** in term of funding for infrastructure and personnel to curate and deposit data

There are two levels of consideration regarding the timing of data sharing; one is sharing of data with public health authorities for public health benefit (*e.g.* for national surveillance), the other is the timing of data sharing into publically accessibly repositories. Where data is generated for clinical or public health investigations, public health needs must always take precedence over academic or individual interests and therefore data must be shared as soon as possible and / or within an agreed time frame with the relevant public health authorities (see recommendation 20).

15.10 Timing of data release to public domains – risks and benefits

The deposition of data into public repositories can, in principle, be beneficial for the public's health, both by enabling international, collaborative genomic epidemiology in response to outbreaks, and also by informing research and development activities with longer term benefits in terms of developing new knowledge and technology. As research and development activity is less time sensitive than immediate public health delivery, decisions on the timing of public data release is likely to be driven primarily by the balance between the value placed on wider access to data for genomic epidemiology purposes and the risks associated with this activity.

Prompt public release of data, as occurred during an *E.coli* outbreak focused in northern Germany in 2011, can facilitate global collaborative analysis that provides insight into the outbreak[96] Such open approaches are not, however, without risks. Analyses undertaken outside of accredited public health laboratories may be more likely to generate erroneous conclusions due to variations in data quality or inaccurate interpretation arising from mis- or under-informed analysis. The repercussions of erroneous conclusions *e.g.* about the source of an outbreak, or the mode of its transmission, may do significant harm to individuals or organisations if placed in the public domain.

Other concerns about releasing data to publicly accessible databases include that it might be misappropriated for harmful purposes. This could include the manipulation of stocks of naturally occurring pathogens (such as smallpox) or utilising knowledge about pathogen sequences to adapt existing viruses to make them more virulent or resistant for use as bioweapons. However, to date, the technical difficulties of achieving this mean that such efforts have not surfaced.

In conclusion a strategy for the timing of genomic and metadata release into public domains is needed. PHE are currently seeking to develop a policy on genomic data release for a number of reasons. Firstly the organisation is already accruing genomic data and must determine how to respond to the data release challenge. An interim 'data release policy' has been drafted for consideration by the office of data release. Secondly PHE need to respond to the UK Governments 'open data strategy'[132,133].

Recommendation 22

All pathogen genomic data and associated metadata required by healthcare and public health professionals to maximise the effectiveness of their management of infectious disease in individual patients and populations should be submitted to the designated database without delay.

Recommendation 23

Where data release into the public domain is envisaged / considered, a strategy for the timing of genomic data and limited metadata release that takes into account a balance between the need to serve wider public health benefit and the rights of individuals and organisations, should be devised. Provision should be made for access by researchers, companies, and healthcare and public health professionals outside the UK.

15.11 Enabling data sharing at the technical level

Adequate infrastructure is needed for a data sharing mandate to function, regardless of whether the degree of sharing is within closed networks only or also incorporates the public sharing of some data. The infrastructure should include a data management system that enables regulated and efficient data submission, storage and access to all legitimate users. Ideally the eventual solution should allow fulfilment of all the potential functions of genomic and clinical data outlined above (figure 15.1). The organisation/s responsible for the construction of this data sharing system will need to decide the optimal configuration for any data management system. There are essentially two options for the storage of genomic data and metadata as part of a clinical and public health focused data management system. The first is to construct a new data management system that meets optimal criteria, the second is to adapt and develop existing databases (which predominantly function as publicly accessible repositories of genomic data generated through research) to meet the needs of this new data sharing initiative.

Fundamental questions that need to be addressed include:

- Whether new data management infrastructure is required, or whether existing databases can be reconfigured to meet their needs

- Who will be responsible for the governance and operation of any data management system

- How the availability of different elements of data should be organised within any system to balance maximisation of their potential utility against minimisation of any harms associated with public data release

15.12 Optimal criteria for designing a data sharing system

Data sharing of the type required for the effective operation of pathogen genomics informed infectious disease management services can be achieved in a number of ways. We propose a series of optimal criteria for configuration by which the likely effectiveness of any system being designed can usefully be judged:

- **Coordinated data collation and storage efforts –** a unified database management system, consisting of one or more databases that are subject to common standards of construction, interoperation and regulation, and with a single mechanism of governance / oversight and a single source of financial sustainability. This approach has two key advantages:

    - o **Analytical –** maximise data utility by ensuring carefully regulated access to the greatest available depth and breadth of high quality genomic data and metadata to underpin delivery of immediate clinical and public health benefits, and longer term research and development of new services and products

    - o **Economic –** minimises resource costs associated with construction of multiple parallel data sharing systems, inefficiencies associated with poor interoperability across independently managed systems and the opportunity cost of missed chances to improve patient and population outcomes associated with data being locked in silos

- **Ease of data submission –** the timely deposition of data necessitates a submission system which alleviates as far as practicable the burden of data deposition, in terms of 'man hours' and 'computer hours' required to complete the process. Real time data submission would be ideal, whereby sequence data is streamed directly from the sequencing machine to the repository, and associated metadata is extracted from laboratory information management systems (LIMS) and paired with sequence data. Challenges to achieving this aim include:

- The varied levels of technical expertise and robustness of informatics infrastructure available in each submitting location
- The need to maintain appropriate standards of quality control over data being submitted
- Wide variation in the LIMS and other IT platforms used locally for the management of genomic and clinical data in microbiology laboratories and their host institutions, and their interoperability with systems likely to be used for data submission

So whilst automated data deposition should be a medium-long term objective, in the short-medium term other mechanisms for data submission will need to be devised. These solutions should be accommodating of the capabilities and resource limitations of the data submitters. To minimise resource costs those being asked to submit data should only need to do so once, to a single designated location, for a given dataset.

- **Ease of data access –** ideally those requiring access to data should be able to locate all relevant information in a single repository, rather than having to piece together segments of data from multiple sites. This will likely require data infrastructure with public and restricted access partitions, to enable the variable levels of access to different portions of the data envisaged for different groups of individuals and organisations. Adequate security systems will be needed to regulate access to patient identifiable data, clinical data, and other sensitive metadata. Each element might require different safeguards to be put in place, to protect against identification. Additionally some safeguards might apply to the entire dataset, such as the requirement for potential data users not to seek to re-identify data or to only use them for a discrete set of purposes.

- **Co-located data and computational capacity –** sufficient computational power should ideally be physically available alongside the collated data (to be analysed), as this eliminates or at least reduces the time taken to transfer data to the computational infrastructure (used to undertake the analysis), reduces the cost associated with data movement, and enables more rapid analysis (17.2).

  Co-located computational capacity could also provide a platform for providing informatics tools and programmes that can enable users to query and analyse the data remotely, further improving the user experience.

To minimise resource costs those being asked to submit data should only need to do so once, to a single designated location, for a given dataset.

- **Sustainability –** ultimately any infrastructure used for collating and storing data from clinical and public health investigations should be robustly constructed (from a data security perspective) and sustainably funded to ensure stable, long term availability of the data.

15.13 Where could pathogen genomic data and associated metadata be stored?

Our consultations with experts in the field of genomic data management clearly indicate that it would be preferable to adapt and develop existing databases rather than construct a new data management system from scratch, particularly if a two-tier restricted-and-public access model is adopted. Below we discuss the advantages of this approach, and also the issues that will need to be overcome to implement this type of data management solution, in the context of a suitable database option for a UK centred system such as the European Nucleotide Archive (ENA). The ENA (based at EMBL-EBI in Cambridgeshire) is the European node of a three way international operation to exchange sequence data, collectively known as the International Nucleotide Sequence Database Collaboration (INSDC), and so currently data deposited into these resources becomes accessible to the broadest possible audience.

Advantages:

- **Hardware capacity and technical expertise –** creating and running a database, particularly one holding raw sequence data, is a significant informatics operation, requiring substantial hardware and technical expertise. Existing large-scale public sequence databases such as the ENA already have appropriate hardware, expertise in running a database and managing access to it. Such predominantly publically funded resources also identify sequence data storage, including data from clinical and public health operations, as being within their mandate, and could therefore provide a single unified source of genomic data storage.

- **Economic efficiency –** data storage is cheapest per unit if it is purchased at scale. Hence a single centralised repository would be the most cost effective way of procuring storage capacity.

- **Sustainability –** in the case of the ENA, which formed in 2008 and is jointly funded by the European Molecular Biology Laboratory, the European Commission and the Wellcome Trust, it is envisaged that this resource will be a sustainable long term storage solution for sequence data.

- **Maximising access –** an existing major global initiative to share genomics data internationally for the management of infectious diseases at global scale; the GMI, are already collaborating with the INSDC members to develop formats for capturing data to enable international collaboration and response to infectious diseases. An FDA driven project in the USA for the detection of outbreaks of foodborne illnesses, utilised the American (National Centre for Biotechnology Information database) based node of the INSDC for their storage and sharing of the pathogen sequence data.

Challenges:

- **The need to develop separate metadata database –** deposition of pathogen genomic data within public repositories would simultaneously meet the need to maximise access to this data by clinical and public health practitioners and the desire to make this data publicly available for re-use for research, development and public health purposes in other countries.

However, as noted above, a publicly accessible databases are not an appropriate storage location for the level of metadata required to enable clinical and epidemiological analysis for the purposes of providing patient and population care. Thus construction of a separate database – most likely managed and governed by appropriate public health authorities – would still be required to collate and store this detailed metadata in a location to which access to data could be limited to users with a legitimate clinical or public health need to use it. Methods would have to be devised to link this 'private' metadata store to the 'public' genomic data stored in the public repository and to enable the deposition in the public database of a 'minimal metadata set', approved for public release, alongside each genomic data set to maximise utility for research and development.

- **User friendliness –** currently, data submission to public repositories, initially built for academic needs, is not necessarily optimised for speed and ease of use. At present, submission mechanisms require a level of informatics expertise, and are sufficiently slow, to pose a significant barrier to compliance with data deposition mandates by frontline clinical service laboratories. Work with the public repository owner (*e.g.* ENA) would have to be undertaken to develop an interface for data submission that meets the accessibility and 'real time' requirements of a useful clinical public health data management system.

- **Co-locating data and computational capacity –** the existing INSDC databases are designed and optimised for large scale storage but do not currently provide co-located computational power to its users. While ENA envisages locating analytical tools and computational capacity alongside its genomic data, this will take time to implement and complete.

- **Timescales for action –** pathogen genomic data from clinical and public health investigations are already being generated and a suitable system for sharing this data and associated metadata, along with computational power for performing analysis on collated data, is needed now. As this functionality cannot be provided immediately by public repositories, transitional data sharing systems – which may be limited in scale while data volumes are relatively low – will need to be constructed to support the effective implementation of pathogen genomics services in the short term.

A suitable system for sharing genomic data from clinical and public health investigations and associated metadata, along with computational power for performing analysis on collated data, is needed now.

15.14 Implementing data collection and consolidation – conclusions

Data sharing is fundamental to the clinical and public health application of pathogen genomics. Currently the most immediate tangible benefits of pathogen genomics – infection control and surveillance – cannot be delivered effectively at a national or even local scale without the agreement to share and infrastructure to exchange data. In order to realise the medium-to-longer term benefits, data should be shared in the public domain (with reasonable safeguards). Yet there are a range of practical, legal and ethical challenges and considerations associated with public release, which need to be addressed, and should be done so in due course. The time taken to resolve these challenges should not however impede the sharing of data with appropriate authorities.

Accordingly authorities will require sufficient storage and computational provisions for their purposes, and will need to collect genomic data and relevant metadata to enable genomic based surveillance and service development. Whether or not data sharing extends beyond closed networks and into the public domain, investment in appropriate infrastructure to store and exchange data will be necessary.

Existing public repositories such as the ENA, a sustainable, non-profit resource with substantial experience in storing high throughput sequence data, and an integral function with international sharing initiatives, can play an essential role in the long term storage of raw sequence data, particularly where data sharing within the public domain is envisaged. From a practical perspective deposition within a public data repository would enable data to be shared widely and presents an economical solution for the storage of data that consumes the most disc space (raw sequence data). The framework for managing data access, could be strengthened and developed to provide additional safeguards, were it to be expanded to take on this role. For categories of data (metadata and patient data) that cannot be shared in the public domain solutions and infrastructure will be needed to store and manage this data securely.

The time taken to resolve practical, legal and ethical challenges and considerations associated with public release should not impede the sharing of data with appropriate authorities.

Recommendation 24

A public health authority such as PHE should be responsible for the collation and storage of all genomic data and metadata for the purposes of clinical and public health service delivery, and to support the development of new clinical and public health applications of genomics in the early stages of implementation until solutions can be developed in collaboration with databases such as ENA to provide access to the necessary storage and expertise to build and maintain an optimal sharing system in the longer term.

16 Genome analysis in practice: developing the IT infrastructure and bioinformatics expertise

Translating genomic data into information and knowledge that can be interpreted to guide the management of infectious disease and improve outcomes for patients depends critically on the availability of high performance analytical software and computational infrastructure and on the availability of people with the skills to develop, operate and maintain these resources.

16.1 Introduction

In this chapter we describe these crucial elements of pathogen genomics services in detail and consider current and future mechanisms for provision and the relative merits thereof.

16.2 Software tools and analysis pipelines

An abundance of software tools is available for performing the specific data processing steps involved in extracting relevant information from raw sequence data, for example genome assembly, or genome annotation (see chapter 4). To date these tools have been almost exclusively developed in academic research settings and to serve research studies undertaken by scientists with relevant bioinformatic skills and expertise. Consequently their user interface, support, and design are not attuned to use in a non-academic, clinical or public health setting, where the following factors are more significant:

- **Robustness, reproducibility and validation**

  In the academic setting analysis of sequence data has conventionally followed an iterative process of testing, evaluation and optimisation[131]. This is because the availability of tools and the nature of sequence data and biological knowledge are constantly evolving, and academic objectives are concerned with developing novel or innovative techniques to keep pace with such changes. Consequently academic tools and analysis pipelines are often regularly being updated and modified, and may therefore be less stable or simple to operate.

By contrast the principal priorities for clinical and public health implementation are robust analytical tools that produce reproducible results in accordance with clinical laboratory accreditation requirements. This requires standardised operating procedures to be in place for the configuration and use of analytical tools and extensive validation of analysis pipelines, in which multiple tools may be connected together in series or in parallel, to be undertaken prior to deployment as part of a clinical service and following any subsequent changes made to the pipeline itself. Validation refers to the process of evaluating the performance of a new instrument / tool or test methodology, with the goal of providing objective evidence that the evaluated method will show acceptable reproducibility and accuracy so as to be clinically applicable. Whenever the conditions under which an original validation was done change (*e.g.* changes to the programming code that operates the analysis software), analytical methods then need revalidation before their introduction into routine use. The constant manipulation of tools and analysis pipelines can therefore become challenging in a clinical setting.

- **User-friendly interface**

  A lack of user friendly and automated analysis software has been cited as a major barrier to the routine use of pathogen sequencing in the clinical and public health context, where existing laboratory staff may not have the programming skills required to create, manipulate and maintain their own analysis tools, tasks commonly undertaken using command line based development of customised scripts[94]. Even where analytical tools with graphical user interfaces have been developed, these are typically designed with an emphasis on flexibility of parameter selection and / or the selection of which procedures and tools to use to construct an analysis pipeline. While these offer useful flexibility to experienced users undertaking exploratory, research-based analyses, they require substantive understanding of the underlying principles of genomic analysis and expert knowledge to judge which settings or components are appropriate. Efforts are underway to adapt analysis tools into more user friendly software packages that would be operable by existing laboratory, clinical and epidemiological professionals with minimal support or training from bioinformatics experts[94]. Until these become widely available it remains the case that genomic analysis for microbiological investigations will remain the domain of specialist trained bioinformaticians able to work with and adapt existing analytical tools for clinical use.

- **Attuned to throughput requirements of the laboratory**

  Results from the analysis of sequence data need to be available in a timeframe that is useful for the intended purpose, for example to decide on the course of a patient's anti-microbial treatment, or to confirm or rule out a disease outbreak. This will require analysis services to be resilient to high throughput requests and fluctuations in demand. Automation of data processing reduces configurable options and combines the numerous analysis steps into a single process, thereby reducing user-training requirements and affording high throughput data processing. Automated pipelines alone are not, however, sufficient to assure scalable analytical service provision, this will also require sufficient computational resources (16.3). Nevertheless automation is central to the delivery of simple, reproducible, efficient, and time-responsive analyses.

A lack of user friendly and automated analysis software has been cited as a major barrier to the routine use of pathogen sequencing in the clinical and public health context...

Recommendation 25

Accessible interfaces or software tools must be developed that meet the needs of clinical users by enabling straightforward access to the information in genomic and metadata databases and to facilitate the ability of legitimate users to perform analyses on underlying data.

16.3 Accessing computational power for genome analysis

As the application of pathogen genomics becomes routine, there will be increasing demands for compute power to perform analysis. These demands can be met by the establishment of the type of centralised data management system described in the previous chapter, which would include both data storage capacity and computational resources that could be made available across a network of provider laboratories. Although capacity for data storage and computational power need not be supplied by the same physical infrastructure, there are advantages of co-locating storage and computational capacity. Having the data as close as possible and accessible to the computational infrastructure eliminates – or at least reduces – the time taken to transfer data. This is turn enables more rapid analyses and reduces the cost associated with data movement.

Below we describe mechanisms by which pathogen genomics providers could access centralised computational power for pathogen genomic data analysis, and their relative merits.

16.3.1 Virtual machines

Virtual computing is the process of simulating infrastructure resources, including computing environments, operating systems, or storage, instead of actually procuring physical versions of those resources. Put simply, users have the benefit of additional programmes or hardware without having to purchase or install their own computer or software tools. These additional resources are provided through virtual machines which run on existing computers and in essence make the computers perform like different, or even more powerful machines.

The virtual machines are underpinned by a physical hardware that contains and controls the additional resources (*e.g.* processing power, memory, storage, operating systems), and allocates them to the virtual machines. In terms of enabling genomic data analysis virtual machines offer the advantages of:

- Potentially circumventing the need for local investment in infrastructure, or at least reducing the hardware demands on the user's side
- Reducing complexity for end users in choosing and installing appropriate software and hardware, as well as eliminating the overhead of managing these components
- Facilitating and ensuring the use of clinically accredited analysis pipelines if for example the physical hardware underpinning the virtual machines could be centrally administered and the virtual interface and resources provided 'locked' to meet accreditation standards

The potential complications and considerations of implementing a network of virtual machines include:

- The complexity of configuring the virtual machines to work across the wide array of IT systems and existing computational infrastucture in use across the microbiology network. Hence some level of informatics expertise would still be required at the local level to set-up virtual machines

- The upfront investment in installing, managing and then maintaining a virtual machine infrastructure

- Software and analysis pipelines would have to be designed or reconfigured to work in this environment

- Given the above factors, a time lag before a virtual machine infrastructure could be established would be likely

16.3.2 The cloud

Virtualisation and cloud computing share a similar concept of maximising computing resources. Although the two systems are sometimes conflated, there are some subtle but important differences. Virtual computing delivers resources through physical infrastructure often owned by the organisation who will distribute these resources, whereas cloud computing delivers resources via the internet. Users' computers would need to be able to run software that enables them to interact with the cloud computing system; this can be as simple as a web browser. The computational resources (*e.g.* memory, computational power, software) are provided via the cloud's network of computers. It is generally accepted that for cloud computing to be defined as such it must provide the flexibility to grow on demand with automated 'self-service' provisioning of resources (*e.g.* to increase use of computational resources without requiring 'human' interaction with the cloud provider).

Given their similarities, cloud and virtual computing share the same advantages of, reducing hardware and software demands on the user's side and reducing the overhead of installing and managing these components. The key additional advantage of cloud computing is the ability to increase and add capacity on demand, in real time, without having to invest in new infrastructure or training personnel. So for example, a microbiology laboratory providing pathogen genomics services, can avoid heavy capital investment to setup computational resources, and instead take advantage of on-demand online cloud computing services when they need to increase analytic throughput (*e.g.* during a suspected outbreak). As with virtualisation, cloud computing can give organisation-wide access to computer applications, and so a means of controlling and ensuring the use of accredited, validated pipelines.

The demand for improved capacity and infrastructure in medical bioinformatics has seen major research investments for the development of cloud-based services including the 'cloud infrastructure for microbial bioinformatics initiative'[134] There is currently no dedicated cloud based resource purely for the analytic needs of clinical and public health providers of microbial genomics, arguably because implementation of pathogen genomic analysis is still in the infancy stage in these settings compared to academia. Some key considerations when delivering analytic services via the cloud include:

- Any analysis software and pipelines must be configured to function in the cloud environment. It is not straightforward to transpose existing analysis pipelines built to initially run on local infrastructure, to run in the same way on different infrastructure including the cloud. This requires testing, and even possible significant rewriting of programmes. Some popular bioinformatics applications for sequence analysis, already have cloud based versions, including Galaxy[135] and Cloud Virtual Resource[136]. Yet as pathogen and application specific, standardised and automated analysis software systems begin to emerge, they will require adaptation to function in the cloud environment

- Depending on the nature of the data to be submitted over the internet and the location of the network of computers which make up the cloud, there will be data confidentiality challenges which will need to be addressed and assessed before cloud based infrastructure is utilised for clinical applications. For example 'public' clouds and cloud based systems where the infrastructure is based offshore, will require service level and data privacy agreements, and clear guidance from appropriate authorities on the use of such infrastructure where the potentially third-party holding of patient derived data is concerned

- Given the above considerations, cloud based infrastructure for clinical and public health applications is unlikely to be available for conventional use in the immediate future, since access to this infrastructure would have to be procured, appropriate service level agreements established, and software tools built and tested to function in this environment

16.3.3 The need to develop a centralised computational infrastructure for pathogen genomics services

Currently there is no centralised provision, based on cloud or virtual computing, for clinical and public health laboratories wishing to undertake pathogen genome analysis. Those groups involved in developing and piloting pathogen genomics services (described in chapter 12) have developed local in-house solutions to enable access to both compute and storage capacity. These have mostly involved accessing existing University-based high performance computing services, or in the case of PHE, developing their own high performance computer cluster.

These localised solutions are unlikely to be sufficiently scalable or sustainable to provide computational capacity as demand for pathogen genomics services increases and the number of centres wishing to provide these services similarly expands. It will, therefore, be vital that organisations involved in providing these services identify mechanisms to provide large scale, sustainable computational resources. This may be achieved through development of infrastructure within existing genomics data centres such as ENA or the Genomics England data centre, or as a standalone facility.

Recommendation 26

Pathogen genomics service providers will need to invest in developing and maintaining, or procuring remote access to, sufficient computational capacity to enable their data analysis.

16.4 (Bio)informatics expertise and staffing

16.4.1 What is a bioinformatician?

Bioinformatics is a multi-and interdisciplinary field combining principles from computer science, mathematics and biology, and practitioners bring different skills and competencies often depending on their prior specialities. For example some practitioners may focus on the development of analytic pipelines, others on the construction and curation of tools and databases, and others on the mining and analysis of data using existing tools. The introduction of pathogen genomics into healthcare is likely to require multiple bioinformaticians with different skill sets at different stages of implementation and service delivery. In addition to bioinformaticians, there are other types of expertise that will be needed to facilitate data analysis and management, such as software developers to build user friendly analytical tools and systems engineers to manage and maintain informatics infrastructure.

16.4.2 Meeting the challenge of increased demand for bioinformatic expertise

Currently, most bioinformatic expertise resides within the academic and private sectors, with the exception of the bioinformatics service within the PHE reference laboratories which has existed for almost 12 years and expanded more recently in response to the demands for pathogen genomic analysis. As regional PHE microbiology laboratories and local NHS hospital laboratories begin to implement genomics services they will need to consider how to access the bioinformatic and computational expertise required to operate these services.

Addressing the current shortage, within the parts of the healthcare and public health workforce dedicated to the management of infectious disease, of skilled bioinformaticians able to undertake genome analysis using the existing tools available must, therefore, be a priority for those tasked with delivering pathogen genomics services. In the following section we outline the two principal ways in which this shortfall can be overcome through the training of existing staff and recruitment to the health sector of bioinformaticians balanced with a focus on the development of software tools that will reduce reliance on this finite pool of bioinformaticians.

16.4.3 Training and retraining of the public health and healthcare workforce

The current drive towards mainstreaming of genomic medicine within the health system has led Health Education England (HEE) to establish a number of new training programmes and initiatives both to provide more specialists able to deliver genomic analysis and bioinformatics to the health service and to broaden understanding by existing healthcare professionals of how to utilise genomic information in the course of their practice. These initiatives include:

- The introduction of an NHS scientific training programme (STP) in bioinformatics (launched in 2013) designed to be a route for bioinformaticians in clinical specialities

- Online introductory courses to genomics and bioinformatics for the wider NHS workforce

- The establishment of a clinical bioinformatics task force by HEE in collaboration with the NHS and PHE and other stakeholders, with a remit to identify the training requirements to support the 100,000 Genomes Project in the short term and adoption of genomic medicine in the long term

Notably these initiatives are focused predominantly on the implementation of human genomic medicine and on training for the NHS workforce to deliver this in their relevant specialist areas. Responsibility for delivery of training in bioinformatics and genome analysis as they relate to pathogens has however been largely delegated to PHE. This is in line with their lead role in implementing pathogen genomics services and in delivering the pathogen genomics component of the 100,000 Genomes Project.

Within PHE, different staff groups are receiving a level of training appropriate to their role in delivering infectious disease management services informed by genomics. For example, a cadre of expert bioinformaticians whose role is to develop analysis pipelines, computational tools and perform genomic analysis as part of the central genomics service is being developed within the PHE microbiology division. This expert group is also cascading their own knowledge and skills to their user groups, predominantly epidemiologists and clinical scientists wishing to undertake their own genomic data analysis as part of their role in informing infectious disease management and through the provision of courses in NGS data analysis for public health.

Beyond the central genomics service developing within the microbiology laboratory, PHE is also providing training to its frontline field epidemiologists and other public health practitioners to ensure they are able to interpret genomic information provided to them in the course of their work managing infectious disease. An example of this training is the ePathGen online learning resource (http://public-health-genomics.phe.org.uk), which includes tutorials and case studies illustrating how genomic information can be used to investigate infectious disease cases and outbreaks.

Recommendation 27

PHE and HEE should continue to work together to ensure that education and training are provided to support the development of the bioinformatics workforce and the analytical and interpretive skills of frontline users of pathogen genomics services.

16.4.4 Focusing scarce bioinformatics resources on the development of user friendly, standardised software for genome analysis

Current demand for bioinformaticians to undertake pathogen genome analysis for clinical and public health purposes is driven largely by the fact that the software tools available for this task can only be configured, operated and maintained by this small group of scientists who both understand the principles of genome analysis and are experienced in computer programming and software development. Bioinformaticians will continue to add significant value to the development of pathogen genomics services over the coming years by undertaking research, development of novel analytical methods and establishment and curation of databases that support genomic analysis. It is not, however, realistic to expect this limited resource to be able to expand rapidly enough (through training and recruitment) to meet the demand for clinical pathogen genome analysis in the next few years. It would also be an inefficient use of their individual skills, and this scarce resource as a whole, to employ bioinformaticians independently within each new pathogen genomics service established within the health service, and expect them to build, maintain and even operate their own genome analysis pipelines. This would entail much duplication of effort, depend on sustaining and growing rapidly this small pool of scientists, and lead to the parallel development of multiple approaches to genome analysis of varying quality and interoperability.

Instead, it would be significantly more efficient to focus this group of scientists on developing and delivering genome analysis software that is configurable and operable by existing laboratory and clinical staff with minimal additional training. This approach capitalises on their expertise by building, maintaining and supporting the delivery of high quality, standardised genome analytical software that can be distributed across an expanding number of pathogen genomics services. In this way, a small pool of skilled bioinformaticians can act as catalysts for the more efficient conversion of data in into interpretable results. This will ultimately lead to a reduced need for these multiskilled specialists to perform 'hands on' genome analysis themselves, freeing them to focus on improving the quality and scope of the tools themselves.

This approach is currently being developed with the University of Cambridge led pathogen genomics implementation programme, where academic experts in pathogen genome analysis and in the development of accessible user interfaces for interpreting genome data are collaborating with the aim of developing software tools that will enable existing clinical and laboratory staff to undertake analysis of pathogen genomes without the need to employ specialist bioinformaticians to support this activity.

Another approach to reducing the need for frontline microbiology laboratories to establish bioinformatics expertise is being explored by the University of Oxford led pathogen genomics implementation programme. Pathogen genome sequence data generated in a network of clinical laboratories is uploaded to a cloud environment where it can be analysed by a central pool of bioinformaticians using a single standardised tool, with results and interpretation being returned to the submitting laboratories without the need for them to have undertaken the analysis themselves.

Experts working in the University of Cambridge led pathogen genomics implementation programme are developing software tools to enable existing clinical and laboratory staff to undertake analysis of pathogen genomes without the need to employ specialist bioinformaticians to support this activity.

Recommendation 28

Additional investment to increase the availability of bioinformaticians able to develop and deliver pathogen genome analytical services will be required, at least in the short term, until analytical tools operable by the existing laboratory and clinical workforce are developed.

16.5 Conclusions

In the early phases of implementation, bioinformatics expertise, computational capacity, and accessible analytic tools are all likely to remain 'rate limiting' factors in the delivery of genomics services. Centralised or networked models for sharing in the procurement, development and delivery of these vital service components are likely to be both more efficient and more effective, by concentrating expertise, enabling standardised high quality services to be developed and deployed across the network of providers and by achieving economies of scale where infrastructure investment is required.

Such centralised or networked models of analytical service provision could be delivered by virtual or cloud computing platforms. A central authority could have oversight of the analytic pipelines and tools available via these platforms, and as discussed in the previous chapter, they could be placed alongside genomic and clinical metadata sets as part of a wider, unified analysis and data management system to serve the needs of pathogen genomics services nationwide.

Ultimately the large-scale national use of pathogen genomics will depend on the availability of user friendly and automated software for data analysis and scalable and sustainable computing infrastructure. Mechanisms to deliver solutions to address these needs should be prioritised, particularly so that pathogen genomics informed services can be accessible across the country on an equitable basis.

Recommendation 29

A PHE led strategy for the organisation of access to computational infrastructure and bioinformatics expertise will be required to ensure access to genome analysis services is not an impediment to the implementation of genomics services.

17 Assuring the quality of genomic disease management

Genomics informed microbiological investigations, like all clinical laboratory investigations, will have to meet agreed minimum standards of accuracy, robustness, reproducibility and usability, and will have to be accredited. Accreditation requires every step in the process, from sample receipt to data reporting, to meet stipulated standards.

17.1 Introduction

Genomics informed microbiological investigations, like all clinical laboratory investigations, will have to meet agreed minimum standards of accuracy, robustness, reproducibility and usability, and will have to be accredited. Accreditation requires every step in the process, from sample receipt to data reporting to meet stipulated standards. In this chapter we discuss the aspects of genomic data generation, analysis, interpretation and reporting to which standards will need to be applied not only to ensure the consistency and safety of genomics based infectious disease management services, but also that their effectiveness is maximised.

17.2 Standards to control the quality and format of raw genomic data and metadata

The quality of the genome sequence data and clinical metadata generated for a given sample analysed within a laboratory must meet minimum standards required by their internal validation processes to ensure the analytical validity of any clinical test or public health investigation to which they contribute. Unlike the raw data underlying many other pathology tests, however, pathogen genome data and associated metadata, will be shared, reanalysed and reused by other laboratories for a range of different purposes. There is, therefore, an additional requirement that single, universally agreed standards for data quality and format must be applied across all laboratories undertaking pathogen genomic analysis. Without this level of consistency in approach, the ability of clinical and public health professionals to combine, reanalyse and reuse this data (once deposited in a shared repository) will be severely limited, as individual data sets originating from different laboratories will not be interoperable and confidence in their quality will be insufficient to meet necessary standards for the clinical accreditation of any investigations on which they would be based.

Recommendation 30

Agreement is required on the standards for genomic data quality and format across laboratories undertaking pathogen genomic analysis for clinical and public health investigations. There should also be mechanisms for standardising descriptive clinical and epidemiological information relating to genomic data to maximise the interoperability, and therefore the utility, of data collected across different locations.

17.3 Benchmarking performance of genome analysis methods

Given that the development of methods for clinical pathogen genome analysis is being undertaken in parallel across a number of institutions, and that the number of such institutions is likely to increase as more become engaged in developing pathogen genomics services, it is highly plausible that multiple methods will emerge for single analytical tasks *e.g.* determining the relatedness of a group of *Salmonella* isolates. Whilst this may encourage innovation and drive improvements in quality, from the perspective of clinical service delivery, assurance that whatever method is being used at least meets a minimum agreed benchmark performance standard is essential. This not only ensures that patients are receiving the highest standard of test available regardless of where it is performed, but also that results emerging from these analyses can be reused by clinicians and public health professionals in other locations with confidence in their integrity.

Achieving this will require benchmarking of the performance of any new analytical method being considered for clinical or public health use against a standard currently agreed by the relevant professional group to be optimal for this purpose. This would enable laboratories to determine whether their new method could perform at least as well as, if not better than, any existing equivalent approaches. Performance could be measured using metrics most relevant to service delivery such as accuracy, reliability, speed and cost *i.e.* focusing on outcomes rather than the configuration of the underlying method itself.

Recommendation 31

Mechanisms need to be developed by relevant professional groups for benchmarking the performance of equivalent genome analysis methods, and for ensuring that methods used in service settings meet minimum standards.

17.4 Standardising the format of results reported from genome analysis

For every envisaged application of pathogen genomics – from outbreak control to drug susceptibility testing – there are multiple methods available to extract the clinically and epidemiologically relevant information from a genome. These methods will produce different results and expressed in different formats which, whilst internally consistent, are often not directly comparable, and are therefore not interoperable. As discussed above, this situation will not be conducive to the combining, reuse and reinterpretation of results that it is anticipated will become a standard part of pathogen genomics based outbreak investigations. Consistent characterisation of the strain and relatedness of multiple, often geographically dispersed, isolates of a pathogen are required to make meaningful epidemiological inferences about them. Furthermore, the existence of multiple analytical approaches and formats will complicate decision making for clinicians and epidemiologists, even where the results formats are not combined in a single investigation, simply by requiring them to be trained to interpret the multiple formats of answers that will arise from the same initial 'question' depending on the laboratory to which the test was referred.

Ultimately, agreement on the approach and semantic standards (nomenclature) used for characterising pathogens will be necessary, to ensure that results are communicated in a format that is understandable, comparable and portable between different locations.

The more wide-spanning the adoption of agreed nomenclature, the greater the geographic scope for performing surveillance and epidemiological analysis. This becomes particularly significant for pathogens of international significance, for example food-borne diseases susceptible to spread as a result of global trade. Therefore nationally, and ideally internationally standardised nomenclature is critical for the convenient and unambiguous exchange of information on the characteristics (including relatedness) of genomes in order to facilitate, local, national and even global surveillance and epidemiological analysis of infection.

The more wide-spanning the adoption of agreed nomenclature, the greater the geographic scope for performing surveillance and epidemiological analysis.

Recommendation 32

In order to support greater interoperability of data generated across the health system there should be mechanisms (preferably international) established for standardising nomenclature for genomic characteristics of pathogens and their relatedness.

17.5 Curating high quality genotype and phenotype relationships

A longer term benefit and utility of implementing pathogen genomics into health services is the potential to accelerate the development or refinement of new applications of genomics (see chapter 9). Over time the accumulation of genomics data with the appropriate subset of metadata, clinical data and phenotypic observations in the microbe will enable new insights to be generated (*e.g.* a particular genotype is associated with drug resistance, or response to treatment, or clinical outcome), which in turn will facilitate the development of new or improved services. Realisation of this benefit will require resources (databases) into which such information can be curated for each pathogen relevant to health, with quality control procedures to ensure the information recorded are sufficiently consistent and reliable to support clinical decision making. Currently the curation of pathogen genomic information resources is predominantly led by academic groups with special interests in given pathogens and for the purpose of supporting research endeavours. Consequently data within these resources have generally not had the stringency of quality assessment that is likely to be necessary for clinical applications (see chapter 4); for example in a format where healthcare workers can be assured a recorded genetic variation is definitively associated with resistance to a given drug, and therefore their patient should undergo a particular course of treatment / clinical management. An exception includes the Stanford HIV drug resistance database[91], which does appraise data to be entered into the resource, and is indeed used routinely in HIV genotyping in clinical settings. Going forward, 'clinical-grade' resources capturing genotype to phenotype relationships will be needed for all other clinically relevant pathogens. To ensure both the integrity and long term sustainability of organism specific genotype-phenotype databases:

- Academic and clinical groups with expertise in a specific organism should be identified and expert panels formed to inform the curation and assessment of data collated into these resources

- These resources should be established with oversight from a central health authority and may even be integrated into the data sharing repository (chapter 15)

- Adequate phenotypic evidence, including pathogen isolates, or tissue collections should be collected in concert with genotypic data, to permit the validation of genotype-phenotype correlations as evidence for these accumulates over time

Recommendation 33

Curation of each organism specific genotype-phenotype database and analytical pipeline, and archiving of isolate / tissue collections must be under the control of a designated responsible authority. Each authority should operate with PHE oversight and funding to support their sustainability.

17.6 Setting standards for data and information management

The ability of laboratories to transmit and receive data and results and reports arising from genomic analysis in the standardised formats required to maximise their utility will depend to a great extent on the configuration and capabilities of their existing Laboratory Information Management Systems (LIMS), electronic (or paper) patient records and IT infrastructure such as network capacity. Variability in the configuration and capabilities of these systems across and even within different healthcare and public health organisations contributing to the management of infectious disease risks undermining any attempt to achieve the standardised data submission to centralised repositories and standardised reporting to end users. This is essential for an effective system to deliver pathogen genomics informed clinical management services. Indeed this variability in local data management and communications infrastructure already significantly reduces the effectiveness with which information is shared between clinical and public health microbiology and infectious disease management services. The adoption of genomics based tests will place further pressure on these systems. It will require their reconfiguration to adapt to the changing type, scale, direction and volume of data and information that needs to be transmitted and received by laboratories and healthcare or public health providers.

17.7 The role of LIMS in providing standardised data and reporting

The key objectives of deploying a LIMS are to make sample and associated data management more efficient and rigorous in high throughput laboratories, and to minimise the risk of human error. The latter for instance includes safeguarding against the possibility of accidental swapping or mislabelling of samples, through use of unique barcode identifiers that are assigned to each sample to be analysed and are tracked throughout the testing procedure. Most clinical and public health microbiology laboratories use commercially developed LIMS to manage their data and while each is system is likely to collect and track similar types of data (*e.g.* date, location of referral, phenotypic information, results) the formats of the files in which they present, store and transmit this data vary widely. Given the anticipated volume of genomic data that will be generated by pathogen genomics services within these laboratories, it is essential that each LIMS is able to automatically export and import this data (and relevant metadata) to and from any relevant repositories, in the standardised formats required to enable their combination, comparison, analysis and reuse by others.

Not only is standardisation required for the purposes of data exchange between laboratories and repositories, it is also equally important at the level of test requisitioning and reporting. In the latter case, adherence to standards currently being developed by the National Laboratory Catalogue of Medicine could be used to ensure that genomic tests are consistently described across the network of providers, and reports provided in formats that are consistent and thus easily interpretable regardless of where they originate.

Achieving this level of coordination and consistency will be extremely challenging. It is clearly impractical for all organisations involved in the delivery of pathogen genomics to unify to a single LIMS and IT system as a means of standardising data transfer and communication across the network. Instead, mechanisms that allow conversion of data formats arising from different platforms to meet the agreed common standards will need to be identified.

Recommendation 34

The challenges of integrating clinical and genomic data, enabling data interoperation and delivering user friendly service requisitioning and reporting interfaces across different LIMS and IT systems need to be addressed. This will require agreement on data management standards between all organisations involved in delivering or using pathogen genomics services.

17.8 Conclusions

Because so much of the value of genomic data for the management of infectious disease arises from it being shared, combined, reused and reanalysed by different professionals for different purposes across different locations, standardisation of the ways in which is produced, analysed and communicated are essential. This standardisation is required not only to assure all users of the data and results of its quality, but also to ensure consistency in the quality of the services they provide.

Achieving this level of standardisation will require coordination and co-operation amongst the diverse range of professionals involved in generating, managing and using genomic and associated clinical data. It is these groups that will need to define standards, ensure compliance and review their suitability, as they possess the expertise and authority to do so. It should be emphasised, however, that standardisation of the quality of output (data or results) required for a clinical service does not equate to homogenisation of methods used to generate that output. In a field as dynamic as pathogen genomics it should continue to be expected, and indeed encouraged, that those involved in service delivery will also continue to innovate through the development and adoption of new and improved methods that deliver higher quality outcomes. Clearly, it is vital that any standards and benchmarks established are reviewed frequently to ensure that methods leading to improvements in service quality can be rapidly identified and adopted by all providers.

18 Building the evidence base I: developing, demonstrating and evaluating clinical utility

In this section of the report we discuss approaches to ensuring that the necessary evidence base is built to enable expansion of the range of organisms and applications for which pathogen genomics could be used. We also emphasise the importance of providing evidence of their effectiveness in clinical practice to enable a case to be made for uptake of this technology beyond the centres of expertise in which they will initially be developed.

18.1 Introduction

The utility of pathogen genomic information in supporting clinical or public health microbiological investigation depends entirely on the ability to accurately interpret the meaning and significance of that information in the context of the infection it causes in a patient or group of patients. In this respect, the huge diversity of genome sequences between and within bacterial and viral species, and the speed with which these genomes undergo mutation makes accurate interpretation a particular challenge. The utility of pathogen WGS in clinical and public health microbiology is, therefore, limited to scenarios in which there is sufficient knowledge about how genome variation for particular organisms correlates with clinically or epidemiologically relevant parameters to enable prediction of the latter from the former.

Whilst for some pathogens there is sufficient information about the genomic correlates of clinically and epidemiologically important features, such as *S. aureus*, to consider implementing clinical investigations based on their genome sequences, for many other clinically significant organisms there is as yet insufficient scientific evidence to do this. Furthermore, even for organisms where the relationship between genome and functional characteristics is well understood in principle and the feasibility of interpreting these relationships has been demonstrated in research settings, the clinical performance of tests based on this knowledge, and their impact on cost and clinical outcomes in frontline clinical and public health services remains to be established and articulated.

18.2 Sources of genomic knowledge development

The genomic knowledge on which the current implementation of pathogen genomics into practice is based has predominantly been developed through academic research initiatives. Thus while some organisms have been studied intensively and much is known already about their genomes, others that may be of less interest from a research perspective, but which are highly significant in clinical or public health terms, remain less well understood. In the future it will therefore be vital to consider how best to develop and capture knowledge and understanding of the genomes of all clinically relevant pathogens, not just those that have already been the subject of intensive academic study. Here we propose three routes through which this could be achieved.

While some organisms have been studied intensively and much is known already about their genomes, others that may be of less interest from a research perspective, but which are highly significant in clinical or public health terms, remain less well understood.

- **Generating new data through funding of targeted and centrally prioritised research initiatives –** for pathogens that are considered to be of national clinical or public health priority, but for which crucial information to link genotype to phenotype and clinical outcome are lacking, directed investment may be required. In these cases teams may be funded through a top-down, nationally determined process to sequence existing legacy culture collections (*e.g.* collection of type cultures at PHE, *Salmonella* or the TB back catalog) or to gather prospective collections of cultures and associated clinical information for analysis. Comparison of genome sequences from such cultures with clinical, other microbiological and epidemiological data such as strain typing would allow new algorithms to be developed for these organisms that can address relevant clinical and epidemiological questions directly from the genome sequence.

- **Mining data from routine practice –** as genomic sequencing becomes introduced into clinical and public health practice for the management of certain infections, the data collected can begin to be mined and shared across organisations to ensure the maximum health benefit is obtained from its acquisition. For example, if sequencing of all phenotypically antibiotic resistant *S. aureus* cultures obtained in routine screening became standard practice , these sequence data could also be mined to develop and refine our knowledge of the genotypes that are contributing to antibiotic resistance in each case. This genotype to phenotype knowledge base could then be used in the development of genotype driven testing for antibiotic resistance (using whole genome, or targeted molecular analysis approaches). Furthermore, sharing such whole genome data sets, in particular with epidemiologists and microbiologists specialising in antimicrobial resistance surveillance, will allow changes in patterns and spread of antibiotic resistance mechanisms to be detected and understood, better guiding public health actions. Mechanisms for achieving this are discussed in chapter 15.

- **Sharing data from and with academic research –** in parallel to the expansion of our knowledge of pathogen genotype-phenotype relationships through the mechanisms described above, it should also be anticipated that academic research will continue to contribute significantly to their development. Research-derived genomic data have been instrumental in underpinning the development of analytical methods and reference databases upon which future clinical services are currently being built. Ensuring continued access to the genomic data that arise from small and large scale pathogen sequencing and analysis projects will require vigilance to ensure that obligations established by funders to ensure timely deposition of this data in publicly accessible archives are adhered to.

While the description above highlights the potential sources of data from which new knowledge can be developed to support the wider and more effective implementation of pathogen genomics into practice, it must be recognised that many of the processes and structures required to ensure these data sets can be extracted, curated, stored and analysed do not yet exist. The issues associated with managing pathogen genomics and clinical data are explored in more detail in chapter 15, chapter 16 and chapter 17.

18.3 Demonstrating and evaluating clinical utility

18.3.1 Background

In parallel to the development of the underlying knowledge base from which new applications of pathogen genomics can be developed, it is also vital that knowledge is developed of 'what works' in clinical and public health practice. Whilst the theoretical effectiveness of pathogen genomics as a tool for a range of microbiological investigations has been established (see Part II) the three key questions that must be asked of all new healthcare innovations have yet to be explicitly addressed through research and evaluation:

- Does it provide a significant improvement in individual or population health outcomes?

- Are these worth paying for?

- Does it reduce overall costs?

The answers to these questions will guide the decisions of payers across the health service on whether or not to fund the introduction of tests based on pathogen whole genome sequencing.

Inherent in addressing these questions are several as yet unresolved challenges and tensions in the process by which new diagnostics are evaluated for their utility in the UK health service. Central amongst these is the ability of health service laboratories to develop new tests in-house. Tests developed in-house undergo rigorous analytical validation in order to ensure compliance with the relevant accreditation requirements and external quality assurance (see chapter 14), but do not require the type of prospective evaluations of clinical utility that are applied to other medical interventions such as pharmaceutical products. Thus a new pathology test that meets the required standards of analytical validity and robustness may be introduced into clinical practice prior to any transparent evaluation of whether or not its use significantly improves outcomes over the current standard of care.

This may be justifiable where the new test functions simply as a like-for-like replacement for an existing test, providing the same output measurements for lower cost or faster turnaround. However, where entirely new analytical processes are being introduced, and where different parameters are being measured (*e.g.* DNA sequence rather than bacterial growth rate in the presence of antibiotic) rigorous evaluation of both the analytical validity and the clinical utility must be undertaken to determine whether the sensitivity and specificity achievable in the laboratory are sufficient to achieve comparable or improved outcomes at the level of patient and population.

It is not clear where responsibility for such evaluation lies, or even at which stage in the implementation process it ought to be undertaken. The lack of commercial incentives – typically linked to meeting regulatory requirements – to undertake such evaluations prior to test marketing means that where evaluations are undertaken, they often occur sometime after a new test has been introduced into clinical care, when sufficient data can be acquired through audit of routine practice to undertake a retrospective evaluation. Such evaluations may however be limited in the insight into clinical utility that they can offer, as where one test has replaced another, no head-to-head comparison of effectiveness, randomised or otherwise, can be made.

18.3.2 Evaluate early implementation programmes rigorously

How can the microbiology and infectious disease community establish the effectiveness of using genomic testing in practice? In the first instance this can be achieved by careful evaluation of implementation pilots, such as those described earlier in this part of the report. Such evaluations should, however, aim to compare WGS with existing standard practice, and to account for as much of the patient / population pathway as possible.

The importance of such evaluations is highlighted by the recent report into the effectiveness of introducing molecular strain typing into the standard investigation and management practices for TB in England. Results of an evaluation completed three years after the service was implemented, showed that despite the testing being delivered as required, it had no detectable impact on TB incidence, due to suboptimal performance of other important parts of the TB care pathway that meant the system as a whole was unable to capitalise on the useful information provided by this new diagnostic. The costs and benefits of this test were estimated at over £90,000 per QALY, which is significantly above the value considered cost-effective for new healthcare interventions by NICE.

It was also noted in the TB strain typing report that failure to establish baseline data for the performance of the service prior to undertaking the evaluation significantly hampered the ability to draw conclusions about changes in its effectiveness. Thus it is imperative that centres involved in early efforts to implement WGS have good baseline data, or continue their existing service for a period after the introduction of the new test, as is envisaged by PHE for the transition from MIRU-VNTR to WGS for TB, to ensure that useful comparator metrics remain available.

18.3.3 Collect and aggregate data from routine practice for audit purposes

In addition to data collected and evaluated from implementation pilots, there will be a need to collect as much information as possible on the performance of pathogen whole genome sequencing once it enters routine use in the early adopter centres. Routine collection of longitudinal data on patient outcomes and changes in the incidence and prevalence of key indicator infections (*e.g.* TB or HCAIs) following the introduction of new patient and public healthcare pathways that incorporate WGS will enable health system managers to perform audits and evaluate whether or not these changes are having their intended effects.

Crucially, as the incidence of certain infections, and particularly the frequency of outbreaks of infection (where WGS may be expected to have the most significant effect on outcomes) in any one locality may be low, aggregation and sharing of outcome data across all centres using WGS for comparable purposes will be essential to ensure there is enough information available to assess the success or otherwise of these changes in practice. For example, the effect of introducing WGS based cluster investigation on the incidence of MRSA outbreaks in hospitals may be difficult to detect in any one hospital where such an outbreak may only be expected to occur once every few years, but building a national picture across multiple participating institutions may allow any reduction to be more confidently detected.

18.3.4 Conduct evaluation trials

The conduct of trials of the use of pathogen genomics designed specifically with patient and public health outcome measures as their primary outputs should also contribute significantly to improving our understanding of the effectiveness of pathogen genomics in practice. Advantages of prospective trials over retrospective evaluations or audits are that they offer the opportunity to test different approaches head to head under more controlled circumstances in which the real effects of the intervention are more likely to be detected. Such trials might involve comparing outcomes from parallel microbiological investigations of the same outbreak with or without the use of WGS. Alternatively they could take advantage of natural experiments that can be conducted by comparing outcomes in settings that have adopted WGS with those that have not, or indeed by comparing outcomes between management of different pathogens within an individual setting where one uses a WGS guided approach and the other does not. While fully randomised controlled trials may be considered too expensive and time consuming to conduct, relative to the benefits expected, there are still clearly many opportunities to design robust tests of effectiveness of WGS-guided microbiology investigations when compared to those they are replacing.

18.4 Conclusions

The effective and appropriate introduction and establishment of pathogen genomics across the UK health service will depend heavily on the appropriate collection and utilisation of data, knowledge and evidence. Upstream, in test development terms, there is a great deal to be gained by the systematic collection and aggregation of genomic and clinical data from which new tests, methodologies and applications for pathogen genomics can be developed. Downstream it is equally important to rigorously test whether these tests are effective in achieving their goals in terms of improved patient and population health. Balanced against arguments about the effectiveness of pathogen genomics are issues of cost, with decisions to implement such new technologies dependent on the balance between the two. Whether the introduction of pathogen genomics into microbiological investigations and infectious disease management pathways is cost effective will depend on a wide range of factors. These will be explored in more detail in the next chapter.

Recommendation 35

Evidence for clinical and public health utility and cost effectiveness will need to be clearly demonstrated prior to funding and adoption of pathogen specific genomics services by clinical and public health end-users.

19 Building the evidence base II: cost effectiveness of pathogen genomics services

The NHS is constantly facing pressure to deliver good quality healthcare, free at the point of use for anyone who is a UK resident, within a budget funded through taxation and set yearly by parliament. The principal rule in economics is that resources are finite. Healthcare economics is the field of study concerned with how such scarce resources are allocated within the healthcare sector in order to inform equitable and efficient decision making.

19.1 Economics of healthcare

Economic evaluation is a key tool used by health economists and is defined as '*the comparative analysis of alternative courses of actions in terms of both their costs and consequences*' [137] Given that resources are scarce, economic evaluation is vital for two main reasons. First, it deals with both the costs and consequences of evaluated interventions (both new interventions and existing interventions). Not only is it necessary for something to work well (both efficacy and effectiveness) but it must also provide value for money. Second, it allows decision makers an explicit set of criteria upon which choices amongst competing uses for scarce resources can be based. When a decision is made to use resources to treat one patient it means that another patient cannot be treated with those same resources. A further reason why economic evaluation is used in healthcare according to Morris, Devlin and Parkin[138] is '*to contain costs and manage demand*' with the introduction of new technologies often seen as high cost either through having a high unit cost or by leading to changes in the overall care of their target population which leads to increased overall costs.

19.2 Economics of genomics in infectious disease

Given the current pressure to find efficiency savings, the NHS is less likely to adopt new technologies that require substantial upfront investment for benefits and savings that may accrue well into the future. However, the cost of whole genome sequencing (WGS) has fallen dramatically over the last five years and given the relatively small size of pathogen genomes compared to humans (see chapter 2), much interest is now focusing on whether the time is right for introducing genomics into clinical and public health microbiology as a cost-effective use of resources. The generic question facing decision makers is: *will genomics provide a more efficient means of using existing resources than current practice in infectious diseases*? The answer will depend on weighing the costs against the outcomes, comparing this to either current practice or gold standard, and then determining whether genomics represents a better use of existing resources. The answer will also be highly context specific, depending on factors including the application of genomics being considered and the pathogen to which it will be applied – it is unlikely therefore to be generalisable. Genomics may be cost-effective in certain scenarios and not in others. Costs will also be significantly different depending on whether the technological infrastructure and expertise already exists within the hospital (both for sequencing and analysis) or whether such services require setting up or 'buying in'. The level of optimisation of existing infectious disease management pathways will also impact on the cost-effectiveness of using WGS, as will the current incentives from government to reduce the number of infections through financial penalties. These may be some of the key drivers in determining whether the use of WGS is cost-effective.

19.3 Lack of evidence

Despite an extensive body of literature surrounding the transformative potential of pathogen genomics in clinical and public health microbiology, there are few publications describing economic evaluation of using pathogen genomics in clinical and public health practice. There are several reasons for this. First, there is generally very limited literature covering the economic evaluation of next generation sequencing technologies regardless of the application. Second, WGS is a diagnostic technology with the primary health impact measure of interest being the number of infections potentially averted. In infectious diseases the probability of getting infected in the first place depends on several factors including pathogen virulence and infectivity, behaviour, environment, level of risk of infection (which can incorporate the first two points) and importantly the actual prevalence of the disease. These factors are challenging when designing a prospective evaluation using randomised control trials (RCTs) – the gold standard study design for assessing the efficacy and safety of interventions. Whereas the development of drugs by necessity requires RCTs in order to determine efficacy and effectiveness, diagnostic interventions are rarely evaluated using the RCT.

The academic research community is incentivised towards developing an intervention and determining whether it works (both efficacy and effectiveness) and less towards the actual implementation of interventions within healthcare services...

The RCT methodology provides a direct platform for conducting economic evaluations – necessary for generating high quality evidence. It also may not be ethical to conduct an RCT to evaluate the effectiveness of genomics informed infectious disease management approaches given that there is a strong *a priori* belief that they will provide more accurate and informative results than current methods. This is not to say that RCTs cannot be conducted, but as they are expensive and time-consuming, this places limitations on their use in determining the patient benefit of laboratory-based diagnostic investigations. Furthermore, the academic research community is incentivised towards developing an intervention and determining whether it works (both efficacy and effectiveness) and less towards the actual implementation of interventions within healthcare services, where economic evidence is often an essential part of the decision making processes. In the absence of such evidence, this chapter now outlines approaches to establishing the evidence base.

19.4 Where will the evidence come from?

If RCTs are unlikely to be undertaken to assess the cost-effectiveness of genomics based infectious disease investigations then economic evaluation can instead be pursued using a model based approach. The use of mathematical modelling to define a set of possible consequences for a pair of alternative interventions under evaluation is particularly powerful where there is great uncertainty around the epidemiology of the condition of interest *e.g.* what the underlying prevalence of infections or outbreaks of a particular disease are, and how they change over time. Models can be used to make estimates of how successful different interventions can be at averting infections whilst varying the parameters of interest, such as underlying prevalence of the infection, to determine the predicted effectiveness under different conditions. Furthermore, the transmissible nature of infectious diseases requires any economic evaluation to account for the impact of interventions not only on the infected patients whose samples are being analysed but also on the reduction in onward transmissions. In the case of any evaluation of genomics this approach is particularly important, as breaking chains of transmission and averting outbreaks that are costly in both health outcome and financial terms are key objectives of many genomics based investigation. Modelling used in these scenarios is termed 'dynamic transmission modelling'. By contrast, where genomic technology is being used simply to guide a change in the treatment pathway of the infected patient which in turn leads to a change in health outcome to that patient but little secondary benefit through interrupted transmission, then a more traditional decision model can be used. It is expected that both of these approaches will be required in evaluating the cost-effectiveness of pathogen genomics. Another strength of using a model-based approach is that evidence can be based on a variety of sources including non-randomised study designs such as controlled before-and-after studies and cohort studies, which are more likely to be available in the published literature. Furthermore, models can be used to evaluate actual practice and the potential divergences between efficacy and effectiveness and between different settings and contexts which could not be done using the RCT study design.

19.4.1 The role of modelling in economic evaluation

A key strength of modelling in economic evaluations is that they can be used to highlight the impact of uncertainty in the available evidence on the reliability of the results generated. While the goal of an economic evaluation is to try and reduce the uncertainty in both resource use and outcomes by using the best available evidence, the reality is that there are often limitations in the completeness or quality of the evidence base that leave significant amounts of uncertainty in any analysis. Using a model-based approach allows some of the effects of this uncertainty to be accounted for, by using evidence collated from many independent sources, undertaking several comparisons within a single model, and using longer time horizons than can be achieved using traditional trial-based designs.

Understanding the impact of uncertainty on economic evaluation also aids generalisability, although it should be noted that models are usually built to answer a specific question and so usually need to be modified to address different, albeit related questions *e.g.* a model created to investigate using genomics in managing TB may not be directly generalisable to management of MRSA, but many of the inputs and decision problems may still be relevant. All good models need to be appropriate to the decision problem, understandable by the decision maker and believable.

A limitation of taking a model-based approach is that of building your decision problem into a model. Models are often constructed based on making various assumptions that can contribute towards uncertainty whilst well designed and conducted RCTs avoid this problem. RCTs have a well-defined methodology whereas there is not a single set way in which a model must be built and how a model is built, is often dependent on our understanding of the decision problem. A key recommendation is to keep models as simple as possible to aid understanding without removing key parameters that influence decision making. Furthermore, all models are limited by their inputs but we can be explicit about the evidence and data used, the structure of the model, and determine the influence of key parameters that drive the decision making process.

19.4.2 The value of sensitivity analyses

All good economic evaluations require a sensitivity analysis – regardless of whether they are trial-based or model-based – in order to determine how changes in certain parameters impact on the results to determine how robust the conclusions drawn from the analysis are. This will particularly be the case for evaluations of pathogen genomics because of the uncertainty surrounding both the costs and the benefits of using this technology in different infectious disease scenarios, and the highly variable nature of the underlying epidemiology of the diseases themselves. Furthermore, assumptions are made in areas where evidence is lacking and so sensitivity analyses are particularly important to determining the effects of variation in these assumptions and how they would impact on decisions. Given the lack of evidence currently available and the variability in costs attached to pathogen genomics and the wider test and clinical pathway (given the lack of nationally set tariffs associated with microbiology investigations), the level of uncertainty is expected to be large. However, decisions are still required regardless of how much uncertainty there is, and this is where appropriate modelling and sensitivity analysis can help.

19.5 Conducting an economic evaluation of genomics to aid outbreak investigation

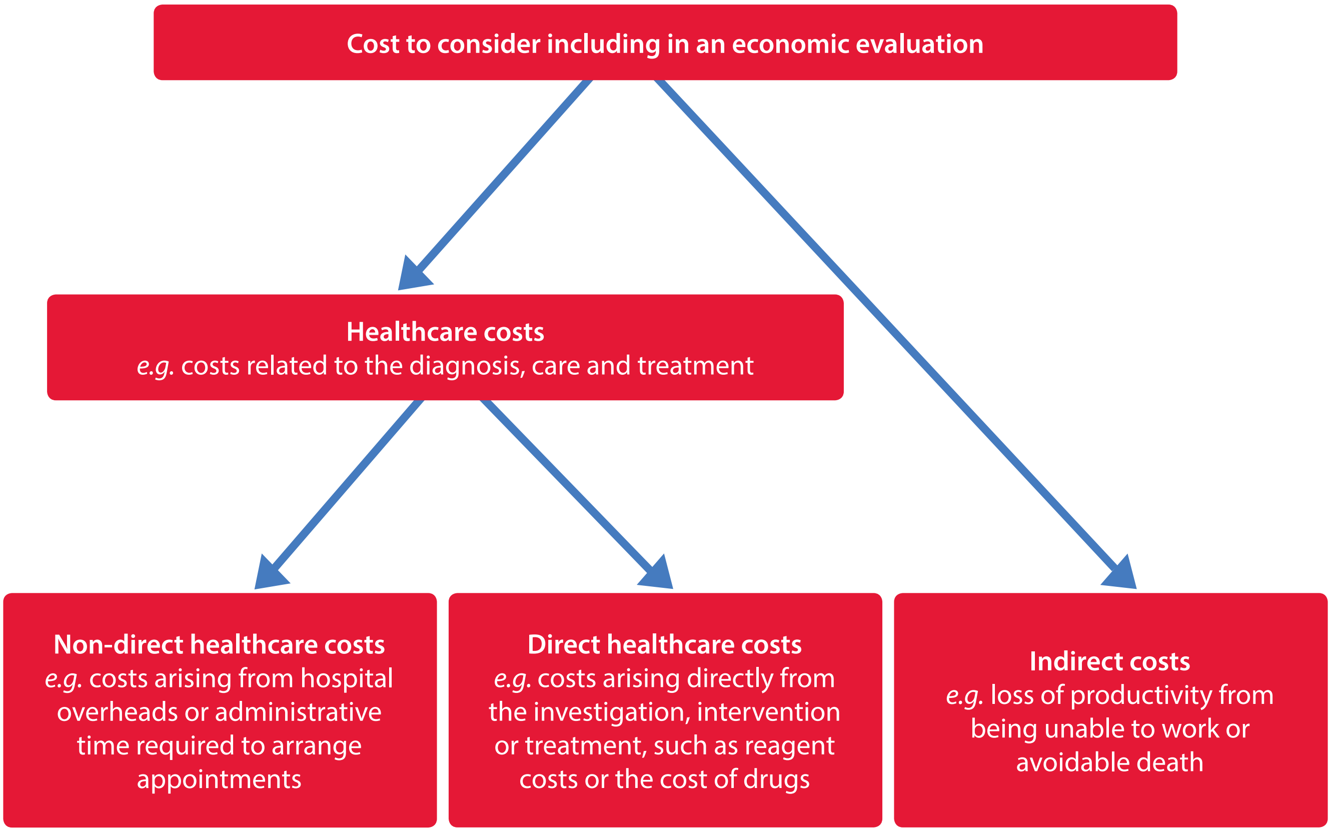
19.5.1 What do you need to measure?

At their most basic level, all economic evaluations need to identify, measure, value and then compare the costs and benefits of the interventions being studied. But before costs and outcomes are considered the issue of perspective needs clarifying as this has implications on both. Economic evaluations are normally used to evaluate the relative efficiency of alternative healthcare interventions and consequently the perspective is usually that of the healthcare payer. However, health economists argue that because we are concerned with the welfare of society a societal perspective should be taken. Of course, it can be argued that a true societal perspective is too broad and so it is often limited to the immediate impacts on patients and families. Regardless of the stance taken, with both perspectives having their own merits within a large body of literature, as a minimum the perspective should be explicitly stated for an economic evaluation. Depending on the question, a different perspective may be required.

19.5.2 Costs

Briefly, costing involves identifying resources that the alternative interventions impact upon, measuring their use, and attaching a value to these resources so that the cost of an intervention can be calculated by multiplying resource use and their value. Costs can be split into direct and non-direct healthcare costs and non-healthcare costs (figure 19.1). Costing can be undertaken from a top-down macro-costing approach (using average costs allocated to patients) or from a bottom-up micro-costing approach (estimating costs individually for each patient and then summing up at the end). For economic evaluation the micro-costing approach tends to be favoured given the improved level of precision and the ability to undertake more sophisticated statistical analyses based on the data collected. Both costs incurred (positive costs) and savings generated (negative costs) are captured to generate a net figure.

Figure 19.1 Direct and non-direct healthcare costs and non-healthcare costs



Cost to consider including in an economic evaluation

Healthcare costs
e.g. costs related to the diagnosis, care and treatment

Non-direct healthcare costs
e.g. costs arising from hospital overheads or administrative time required to arrange appointments

Direct healthcare costs
e.g. costs arising directly from the investigation, intervention or treatment, such as reagent costs or the cost of drugs

Indirect costs
e.g. loss of productivity from being unable to work or avoidable death

The costs involved in genome sequencing are becoming well defined and understood and are summarised in our previous report *Next Steps in the Sequence*. These costs cover the use of WGS as the assay and include, for example, the cost of equipment and consumables, labour costs, informatics and clinical expertise in interpreting the data to achieve a meaningful test result. The costs relating to the clinical outcome consequent to the test result also require defining, capturing and valuing. For example the following should be considered for a model evaluating the use of pathogen genomics in outbreak management:

- **Cost of inpatient stay –** length of stay and hospital costs incurred per patient during their stay, often depending on their ward type

- **Cost of actual care received –** initial treatment plus any change to treatment following test result, incorporating both costs of tests / drugs and staff (for the duration of treatment), including outpatient care received

- **Costs of using WGS** and the costs associated with any change to clinical pathways specific to use of WGS (for example the inclusion of additional confirmatory tests or additional screening interventions following the use of WGS)

- **Other financial penalties** associated with incidence of cases of infection, outbreaks, or exceeding cumulative infection control targets for the incidence of certain infections

Accounting for these costs in the model would allow estimates to be made of the change in costs associated with using genomics versus current practice, and in turn enable determination of which would be more expensive and in which scenarios. Whilst the expectation is that the number of actual outbreaks will decrease and hence the costs associated with that should drop, it should be noted here that it is possible that using genomics may result in an increase in the number of potential outbreaks investigated compared to current methods, with more activity required in investigation and management, all increasing the costs to the health system.

19.5.3 Patient / population outcomes

As set out in part II of this report, pathogen genomics based investigations can be used to fulfil different roles within the management of patients or in the surveillance and control of infectious diseases at a population level. There is a need, therefore, to identify, capture and value the multiple potential health outcomes of interest when modelling the effect of genomics. For example in an outbreak management scenario these could include:

- Number (or estimation of percentage) of susceptible patients

- Number (or estimation of percentage) of infected patients

- Number (or estimation of percentage) of immunised patients, if relevant

- Number (or estimation of percentage) of patient deaths

- Length of hospital stay for each type of patient (perhaps depending on ward type or clinical category)

- Severity of clinical symptoms or capturing different stages of disease if they impact on either costs, interventions or outcomes

- Estimation of number (or percentage) of patients referred to A&E as a result of an infection following discharge from hospital

Inclusion of measurements or estimations of these outcomes in any model would allow it to predict the change in clinical outcome (for example the number of infections averted) for using genomics versus current practice and thus to determine its impact on the number of primary and secondary infections relative to current methodology. Notably, the primary benefit of using WGS being captured in such a model for outbreak management is the reduction in the size of an outbreak *i.e.* fewer patients becoming infected, rather than improvements in the health outcome of individual infected patients.

19.5.4 Efficiency in production

There are situations in which the use of pathogen genomics may not be expected to lead to any change in clinical outcomes, but has been proposed to be cost-effective on the basis that it would replace more costly methodologies currently in use while maintaining the same clinical performance. In this situation the decision problem of any model can be simplified and the issue of productive efficiency is brought to the fore – will WGS allow us to maximise the health outcome (*e.g.* produce more diagnoses) without increasing the budget.

19.6 Wider challenges to the economic evaluation of genomics in infectious disease management

Many of the challenges presented in this next section are similar to those previously described in our report *Next Steps in the Sequence*, which addressed human genomic analysis, but remain equally relevant here. As discussed elsewhere in this report, genomics and the use of WGS has the potential to impact across many parts of microbiological investigations and this is a key strength of using WGS. We discuss some of the challenges in capturing these strengths, such as its ability to be used as a single technological sequencing platform across different tests or applications within genomics.

19.6.1 Evaluation of complex genomic technologies with multiplex use

The information arising from a single whole genome analysis can have multiple applications in a number of areas of infectious disease management. This brings with it the challenge of trying to incorporate wider infrastructure costs and benefits, potentially across multiple healthcare organisations, into a single analysis. Not only can one technology be used for analysing multiple pathogens – a situation already common in microbiology laboratories using mass spectrometry – it can also be used for sequencing human genome samples as well when not being used for pathogen genomes. This can allow economy of scale, by running the machine at maximum throughput to bring down the per test cost of using this single machine across different laboratories within a single hospital setting. Other costs such as overheads and laboratory staff could also be shared.

19.6.2 Evaluating a dynamic technology

Although the dramatic reduction in the price of genome sequencing has slowed down recently, next generation sequencing is still a dynamic, fast evolving technology. This poses the question, *when should we undertake an economic evaluation of the test if its specification and cost are constantly evolving*? Not only are costs of genomics expected to get even lower, but the quality of sequencing is expected to improve and the wealth of knowledge surrounding pathogen genomics is also growing. New technologies when first implemented also have 'learning curves' whereby the efficiency improves over time. The current bottleneck is expected to be the development of working bioinformatics pipelines to streamline the analysis. A further complication is that different institutions may implement such technology in slightly different ways challenging attempts to make comparisons and undertake evaluations that would be directly relevant to aiding decision making about adoption within the wider NHS. This is where a model-based approach would be favoured with the potential to easily adapt and update data for models which already exist.

> *New technologies when first implemented also have 'learning curves' whereby the efficiency improves over time. The current bottleneck is expected to be the development of working bioinformatics pipelines to streamline the analysis.*

19.6.3 Wider economic implications / training needs

Not only is genomic technology itself required to deliver pathogen genomics based investigations, but expertise is also needed by staff to efficiently and effectively run these services. This extends from the laboratory staff to the bioinformatics support required to undertake analyses and the creation of databases and analysis pipelines to produce results that can be easily used by clinical microbiologists in routine practice. Although the genome size of pathogens is significantly smaller than that of humans, computing power and data storage facilities are still expected to be significant and needs to be factored into economic evaluations. Currently there is no ideal clinical pathway by which to provide these services and as such these issues will differ between different laboratories where the service model will be different *e.g.* some hospitals may do all tests in-house whereas others may have private providers undertaking the laboratory testing.

19.7 Can macroeconomic considerations support a political case for investing in the implementation of pathogen genomics services?

The above discussion seeks to explore the ways in which micro economic modelling can be used in a relatively localised and specific context to make decisions about whether or not to invest scarce and finite resources in the establishment of a pathogen genomics service as part of a wider programme of infectious disease management. Importantly, we assume that such decisions will be made within the constraints of existing local NHS or PHE budgets.

There is however the wider question to consider of what price is placed, in a national strategic policy context, on the control of infectious disease to the economy, and whether the anticipated benefits of implementing pathogen genomics (in terms of reducing this burden) warrant additional strategic investment, rather than reallocation from within existing local budgets, to ensure it occurs. Making this type of economic case for national investment depends on the existence of estimates of the economic burden of infectious disease, and projections of the effect of genomics upon these. Unfortunately both of these parameters are extremely difficult to measure (in the case of the former), or predict (in the case of the latter) in monetary form.

An attempt was made to estimate the economic burden of infectious disease in England in the second volume of the Chief Medical Officer's annual report for 2011, *Infections and the rise of antimicrobial resistance* [139]. Whilst a figure of £30 billion each year (this includes costs to the healthcare service, to the labour market and to individuals themselves) was arrived at, its provenance is unclear and it is well known that such studies are frequently limited by the data available and either over- or under-estimate the benefits or values attached to such resources[140]. While they can act as descriptive indicators to support policy advice and political decision making they are of limited use in making quantitative analysis-based prioritisations of interventions from within a finite budget, where other forms of health economic evaluation (such as cost-effectiveness analysis) that undertake direct comparative evaluation of alternative courses of action are more useful.

Rather than focusing on the almost inestimable figure of the total economic burden associated with infectious disease, it might be more plausible to estimate the economic impact of particularly significant infectious disease related events. In current public policy terms the two most prominent events or processes are the emergence of antimicrobial resistance (AMR) – as highlighted by the award of the £10 million Longitude 2014 prize fund to help solve the problem of global antibiotic resistance – and the occurrence of a flu pandemic[141]. The Department of Health has sponsored an analysis by Smith and Coast[142] of the available literature examining the economic impact of AMR. Strikingly, this analysis, which compared the highest estimate of the burden of antimicrobial resistance ($55 billion) to other selected conditions within the US, found that antimicrobial resistance rated fairly low down the order of health related costs to the economy. Smith and Coast argue, therefore, that the use of antibiotics is so ingrained in the way we currently deliver healthcare, the consequences and the costs to health services and human health are so difficult to envisage that the true economic burden of AMR remains inestimable. Despite these difficulties, a recent review of AMR published in December 2014 and chaired by Jim O'Neill included two economic burden studies (conducted by RAND Europe and KPMG) which, whilst acknowledging the 'severe lack of data', estimated the cost at a 100 trillion US dollar reduction off the world GDP by 2050 in failing to address AMR[143]. Furthermore, they estimated some of the secondary impacts of AMR in the related knock-on effects to healthcare, bringing the total reduction to world GDP by 2050 of 210 trillion US Dollars (*i.e.* 210 trillion US dollars reduction in the world GDP over the next 35 years).

Smith and Coast suggest treating investment in further AMR research as insurance against a worst case scenario, for example where we lose the ability to use antibiotics.

19.8 Conclusions

It is well acknowledged that infectious disease poses a real and immediate threat to human health. The literature surrounding the potentially transformative nature of pathogen genomics in tackling this threat is growing although the economic evidence to support contentions of its cost-effectiveness currently lags behind. The Department of Health and the Wellcome Trust have responded by including health economic evaluation as a key component of the HICF projects they have funded to support the implementation of pathogen genomics in the UK (see chapter 12). Once these HICF programmes define services that can be economically evaluated it should become possible for them to address the lack of economic evidence around pathogen genomics. Following this, the challenge will be using this evidence to make the case for diffusion and adoption of pathogen genomics based services across the NHS. This will remain difficult, because although the HICF funded work will start to address the lack of evidence it will by no means address all of the issues and uncertainties around the economics of pathogen genomics. It should, however, allow the identification of key economic drivers where further research can be directed in order to better understand the complex systems within which pathogen genomics work and reduce uncertainty. And this is before we start to consider the wider complexities of incorporating pathogen genomics into the One Health Initiative Framework (see chapter 8) given that 70% of emerging infectious diseases are zoonotic *i.e.* they are spread from animals to humans. Smith and Coast[142] suggest treating investment in further research in this area as insurance against a worst case scenario, for example where we lose the ability to use antibiotics with subsequent effects on our entire healthcare system, rather than treating it simply as a cost.

20 Ethical, legal & social issues when implementing pathogen genomics

Beyond specific concerns about data sharing, it is unlikely that pathogen genomics will change the underlying tensions between the interests of individuals and wider society that manifest in the management of infectious disease more generally. Nevertheless, pathogen genomics has the potential to exacerbate a number of existing issues, and to raise a smaller number of novel issues which will need to be addressed if these technologies are to be introduced in a responsible and efficient manner.

20.1 Background about the regulatory context and the type of ELSI issues that arise

The implementation of pathogen genomics will be influenced by the regulatory context within which infectious disease services operate. The ethical, legal and social issues arising from infectious diseases are sometimes challenging because they reflect the tension between societal interests (in public health) and other rights, duties and responsibilities that arise at the level of the individual. This may include biomedical ethical principles which operate at the level of the individual through concerns about individual privacy and confidentiality and the ethical requirement to promote autonomous choice. An example of this conflict is where societal concerns about transmission of a novel and highly infectious disease may justify placing potentially infected individuals into enforced quarantine even against their will.

The most significant ethical, legal and societal challenges that arise from the implementation of pathogen genomics as a new investigatory tool relate to how data might be processed, shared and linked in ways that could identify their source, or lead to breaches of confidentiality or privacy resulting in discrimination or stigmatisation. Since the ability to share data effectively and securely is a pre-requisite for the implementation strategies recommended in this report, we review the ethical, legal and societal challenges arising from data sharing and use in chapter 15.

20.2 Identifying sources of infection and chains of transmission

As discussed in chapter 7, pathogen whole genome sequencing technologies can significantly enhance the ability of microbiologists to discriminate between similar bacterial and viral strains, allowing more sensitive and robust inferences to be made about the source of infection and the direction and mode of transmission of infections. The potential for pathogen genomics to enhance the accuracy of transmission chain analysis raises the possibility that these technologies might be used to attempt to establish a civil or criminal claim under UK law. The potential for legal actions already exists, although increasing availability of pathogen genomic evidence could potentially encourage more litigation. We are not aware of any current cases in the UK where these technologies have been used for such a purpose. However, it is possible that an individual who is infected by their sexual partner, could bring a claim against them in criminal or civil law, or that a claim might be made against a healthcare provider (if the infection was acquired during healthcare).

Existing case law suggests that it might be possible to bring a criminal case concerning the transmission of infectious diseases such as HIV under the Offences Against the Person Act (1861) if the infected person was reckless as to whether they would infect their sexual partner and had caused them grievous bodily harm in the form of a serious illness.

A civil case might be easier to prove, because of the differing evidential requirements (that the facts have to be proved on the balance of probability rather than beyond reasonable doubt). In a civil case, a complainant would have to prove various elements against the defendants in order to bring a successful negligence claim in tort, namely that the defendant owed them a duty of care, that there was breach of that duty and that the complainant suffered damage as a result. Establishing causation (that the negligent act caused the damage) is often difficult. Pathogen sequencing techniques could help to establish when and whether infection occurred, but are unlikely to be conclusive enough to establish liability because sequence data comparison would not be sufficient to prove a transmission link, or to provide confidence about the directionality of transmission. Additional epidemiological or temporal information would be needed to establish causation. Thus in the short term, these technologies are likely to be used to exclude liability rather than prove it. Nevertheless, pathogen sequence results could be used as forensic evidence to support or refute litigation claims.

If pathogen genomics can enhance the sensitivity and specificity of inferences about transmission chains, another consequence is that employers may seek to be more proactive about testing their employees regularly, to ensure that they are not infected by or carrying specific organisms, and therefore presenting a risk of infection to others. Employees could also be tested on an *ad hoc* basis in response to an infection outbreak, both to establish the hospital as the source of infection or to exclude it and suggest a source within the community. It also seems plausible that employers may seek to test their employees more regularly, if only to address employer responsibilities under the Control of Substances Hazardous to Health Regulations (2002). An inherent challenge of both workplace and community based testing is to ensure that consent to testing is both adequately informed, and has been given voluntarily.

Another context in which these technologies might be used is where food products are responsible for causing foodborne infection by pathogens such as *Salmonella*[144]. Pathogen genomic technologies might help bolster evidence of transmission chains (from raw product through to the consumer) especially if supplemented by improved methods of interpretation and databases of isolates. Currently similar caveats about the role of these technologies in establishing legal liability apply: pathogen genomics techniques will be more robust in excluding liability rather than proving it because of existing legal thresholds in establishing causation. In all cases, it is important that legitimate use of these technologies for health benefit is not compromised by concerns about how they might be misused in litigation. This may require safeguards to be put in place and wider education across stakeholder groups[145].

20.3 Metagenomic analysis

Metagenomics (discussed in chapter 5) is the application of whole genome sequencing to 'raw' uncultured, and hence unpurified biological samples (taken from people or environments) that contain the genomes of multiple organisms. In the context of clinical applications of metagenomics, samples typically contain a mixture of host genomic material and genomic material from the vast array of microorganisms that inhabit the host tissue from which the sample was taken. Metagenomics currently has limited clinical or public health utility, outside of the occasional investigations into the identity of novel or extremely rare pathogens causing disease noted in chapter 5. It is anticipated, however, that as this technology matures and the analytical methods on which it depends for its success are refined, that it may play a more significant role in mainstream diagnostic microbiology. It may also, in the future, be used to characterise the 'microbiome' of individuals *i.e.* the diversity of commensal microorganisms with which they are colonised in different states of health and disease. Such information has been proposed, to have potentially predictive value in identifying individuals at higher or lower risks of poor health and disease, although evidence for this proposition remains weak and the prospect of applying such information clinically remains a distant prospect.

Metagenomics is the application of whole genome sequencing to 'raw' uncultured, and hence unpurified biological samples (taken from people or environments) that contain the genomes of multiple organisms.

It is vital to consider the ethical, legal and social issues that may arise from the use of metagenomic analysis in advance of any steps towards implementing it as part of clinical diagnostic or preventive services to avoid there becoming barriers to implementation. Many of the ethical, legal and social issues raised by metagenomic analysis are not, however, novel: these include the need to ensure privacy and confidentiality; consent; ownership of samples and data; generating unexpected or incidental findings; return of results; ensuring equitable access and wider concerns about governance and accountability.

The breadth of metagenomic analysis means that it is possible that unexpected or incidental findings may be detected. These could include serious conditions such as chronic infections with blood borne viruses including hepatitis or HIV, which although treatable may have adverse health outcomes. Incidental findings of a different type may occur if non-germline samples (such as faecal samples) are contaminated with germline cells which could potentially reveal predictive information about developing inherited disease[146]. This could be addressed if adequate filtering and removal of human sequence was undertaken. Nevertheless, the results of metagenomic analysis for both diagnostic and preventive purposes could lead to discrimination and stigmatisation. Public acceptability of the procedure may also be low, especially in some cultures, where sample collection is invasive and may offend against some cultural norms. However, in future, returning these types of analysis results to participants might allow preventative interventions that could improve health outcomes. Ensuring that the informed consent process is sufficiently comprehensive so that participants are adequately informed about the scope of testing, the potential for health related findings to occur, and the impact and management of results will be a challenge.

21 Delivering safe and effective genomics services in a dynamic technology and knowledge environment

Throughout our research and engagement with stakeholders one of the most dominant recurring themes in discussions was the rapid pace of development of both the technologies underpinning genomic analysis and genomic knowledge on which interpretative services are based.

21.1 The next five years – what can we expect?

There is a widespread view that further advances within the next five years will enable genomes to be sequenced faster, analysis and interpretation of pathogen genomic data to become increasingly effective and accurate, and culture free metagenomic sequencing to become more accessible in both cost and technical terms. Below we describe each of these anticipated advances in more detail:

- **Sequencing technology development –** there is a strong expectation of significant improvements in genomic sequencing platforms. This includes the development of devices that sequence using nanotechnology-based approaches, taking advantage of microfluidic (so-called 'lab-on-a-chip') engineering, to develop portable sequencing platforms that can be operated outside of traditional laboratory environments and closer to the point of care. Rapid progress is also anticipated in simplifying or even eliminating much of the sample preparation that precedes sequencing by synthesis approaches, and which currently places significant limitations on the clinical utility of WGS for identifying pathogens and drug susceptibility by extending turnaround times beyond those achievable by conventional techniques. Together, such innovations have the potential to reduce assay costs, and extend the utility of genomic assays into areas of clinical and public health microbiological investigation currently out of reach for existing genomic technologies.

There is a widespread view that the next five years will see faster sequencing of genomes and increasingly effective and accurate analysis and interpretation of pathogen genomic data.

- **Analytical and interpretative software and method development –** availability of robust, stable and user-friendly software packages to enable the end-to-end analysis and interpretation of genomic data for clinical and public health investigations is currently a limiting factor in the adoption of pathogen genomics. There are however many efforts currently underway to address this issue and it should be expected that within the next five years software will become available that is able to receive raw genomic and epidemiological data and provide actionable reports, without the user requiring any level of computational or bioinformatic expertise beyond that expected of current healthcare practitioners. The analytical methodologies underlying these software tools are also expected to improve, as the interpretation of phylogenetic relationships is refined through experience and further research, and as underlying data on drug susceptibility variants improves.

- **Expansion in our knowledge of genotype to phenotype relationships –** one of the great long term benefits of implementing pathogen genomics into clinical and public health services now, and ensuring appropriate periods of overlap with existing phenotypic methods, is that it will create a virtuous circle in which the insights generated can be used to accelerate the development and refinement of new applications of genomics. For example, genomic data obtained during outbreak investigation or routine surveillance will include genomic determinants of antimicrobial resistance that can be used, in combination with appropriate phenotypic and clinical data, to develop new algorithms for the prediction of drug susceptibility directly from genomic sequences. These may then be implemented using either improved versions of existing sequencing platforms, or some of the many portable point of care devices currently under development for rapid drug susceptibility testing outside the laboratory.

- **The development of metagenomic analysis for the clinic –** the advances described above in sequencing platform technology and method development should enable the introduction of culture free genomics, metagenomics, into the clinic. Currently the most significant barriers to this are the complexity of analysing multiple pathogen genomics simultaneously, and the cost of obtaining sufficient pathogen genomic sequence from uncultured patient samples such as stools or sputum. Research has already demonstrated that culture free techniques can detect tuberculosis from sputum and *E. coli* infections from stool without culture, and so if anticipated improvements in the cost and throughput of sequencing are made, and our ability to interpret complex genomic data improves, translating these results into the clinic should occur within five years.

These advances will be delivered through a combination of approaches, including ongoing private sector innovation in sequencing technology and ongoing public and private sector translational and basic research aimed at developing knowledge of genotype-phenotype relationships and analytical methodologies to exploit this knowledge in the clinic. These will, in turn, require the development and maintenance of an effective genomic and clinical data sharing system that facilitates appropriately regulated access to researchers.

21.2 Adapting and adopting – realising the benefits of future innovation in pathogen genomics

Pathology services, and microbiology in particular, have a record of successfully adopting innovations in a timely and effective manner. In the context of microbiology services this has included the transition towards molecular *i.e.* DNA or RNA based testing for a wide range of viral and bacterial diagnostics, and the widespread introduction of mass spectrometry as a tool for rapid bacterial pathogen identification from cultures. The underlying mechanisms already exist, therefore, within clinical and public health microbiology to adopt and adapt to any further developments in genomics that may ensue over the next five years. The challenge will be to balance the costs incurred against potential gains in clinical effectiveness, and to maximise the pace with which this takes place whilst minimising threats to patient safety and care. Below we outline some of the ways in which these challenges might be met:

- **Risk sharing and appropriate service configuration –** if current sequencing technology is expected to become rapidly redundant, potential service developers and users might benefit from pooling their risks, and investing in shared large scale sequencing facilities, which benefit from economies of scale and efficiencies of utilisation, rather than each making individual investments in their own sequencing platforms which are more likely to be underutilised, leaving laboratories unlikely to recover their costs (see earlier discussion in chapter 13). This would be consistent with our earlier proposals in chapter 13 that for many other reasons, during this period of implementation and rapid technological change, a smaller number of high throughput genomics services may be the most appropriate model for service configuration.

- **Establishment of clear minimum standards for service delivery –** there are often concerns about how well the validity, utility and safety of innovative technologies must be established prior to their use in clinical service. A balance must be struck between appropriate regulation of new testing methodologies and the stifling of innovation in service provision. Ideally this can be achieved by the establishment of agreed minimum performance standards, based on external quality assurance, that are auditable and verifiable. Where laboratories wish to introduce novel, innovative services, this can be done without hindrance so long as they are able to demonstrate performance that matches or exceeds the current standard of care as encapsulated in the standards for EQA and the analytical validation and performance requirements of their test accreditation processes.

- **Development of mechanisms and infrastructure for creating a 'virtuous circle' of innovation –** the successful adoption of many of the technological and knowledge advances described in previous sections will depend on the effective implementation of current genomic technologies. To create the virtuous circle, in which knowledge and experience accruing from the current wave of genomics implementation can be used to catalyse the introduction of subsequent technologies and methodologies, will require infrastructure to ensure knowledge is both captured and disseminated within, and crucially, beyond the organisations in which it is generated. For example, as clinical microbiology laboratories accrue data that can be used to determine the genotype to phenotype relationships of genomic variation to antibiotic susceptibility, this information must be extracted, consolidated and made available to researchers and to those developing new testing platforms and analytics that may enable future rapid point of care identification of infections and their drug susceptibility. These future innovations will then feed back into the clinic, where they will enable improved effectiveness of clinical microbiology practice, potentially at significantly reduced cost to the health service.

- **Continued financial support of translational research and implementation activities –** targeted funding of research and implementation activities, through mechanisms such as the Healthcare Innovation Challenge Fund, should be sustained in order to develop, deliver and exploit the next wave of technological and methodological advances for clinical application in microbiology. Specific funding should also be allocated to enable the trialing and evaluation of these new applications in real world healthcare and community settings. In addition to 'one off' investments in translational research and implementation programmes, sustained longer term funding may also be required to enable the development, curation and maintenance of resources such as phenotype to genotype relationship databases that will underpin much of the interpretative capability of both current and future genomic (or metagenomic) services. This funding should be centrally managed by the health service, and independent of specific research grants, in order to ensure that the facilities it provides are sustainable and fit for clinical and public health use.

- **Advanced consideration of relevant ethical, social and legal issues –** as the range of potential applications of genomics in managing infectious disease broadens so do the ethical, social and legal issues with which they may be associated. Crucially, if innovation is to be undertaken responsibly and with the support of the patients and populations whom it is intended to help, the ethical, social and legal impacts must be considered in parallel with the development and implementation processes, not as an afterthought. Such responsible approaches to innovation allow services to be developed in a way that will maximise their public acceptability once they reach the point of implementation. It also allows opportunities for relevant adaptations to be made during the development process to ensure compliance with relevant regulatory frameworks. A clear example of where this approach will be important is in the future use of metagenomic approaches to microbiological investigations, in which untargeted sequencing mean that both human and pathogen genome data will be generated from patient samples. Establishing ethically sound regulations to govern the generation and use of this data will be essential to assure both patients and health services that this new technology can be delivered in a legally and socially acceptable manner.

If innovation is to be undertaken responsibly... the ethical, social and legal impacts must be considered in parallel with the development and implementation processes. Such responsible approaches to innovation allow services to be developed in a way that will maximise their public acceptability once they reach the point of implementation.

21.3 Conclusions

In this chapter we have explored a subset of the most plausible innovations that can be envisaged for pathogen genomics in the next five years. This list is deliberately conservative in scope, and any estimate of the timescales over which such innovations might yield improvements in patient care and population health are highly uncertain. It is important to recognise that the likelihood of these innovations having a positive impact on the management of infectious disease depends to a great extent on a wide range of factors, many beyond the control of the health system and the policy makers who govern it. Nevertheless, these organisations and individuals can, through the actions described above, provide significant logistical, financial, regulatory and political support to those wishing to develop and implement future innovations in genomics.

Part IV

In this final part of the report, we first summarise our conclusions on the current state of pathogen genomics in England. We will reflect on how the current state of genomic technology, knowledge and related analytical methods currently circumscribe the scope of applications of pathogen genomics in infectious disease management at this point in time. We will also briefly outline the range of initiatives underway to implement pathogen genomics services, based on this limited initial range of applications.

In the final chapter of this report we will set out our vision for how the results of our research can be used to:

'*Support the development and delivery of genomics informed infectious disease services that are evidence based, high quality, available population wide, and on an equitable basis*'

This is achieved through the presentation of a roadmap. This will illustrate how the various recommendations made throughout this report map onto the processes that need to be undertaken, and the systems that need to be built, to transform pathogen genomics from a promising emerging technology on the verge of limited implementation in the health service, into a highly effective, well evidenced, mainstream part of infectious disease management services that delivers significant patient and population benefit at a cost the health service can afford.

22 Where are we now?

In this section we examine the current state of pathogen genomics as a tool for public health and clinical use in England.

22.1 How can pathogen genomics be used in clinical and public healthcare right now?

There are three main areas in which pathogen genomics is currently sufficiently well developed as a technological, scientific and analytical discipline to deliver potential clinical and public health utility in infectious disease investigations:

1. **Outbreak management –** the resolution provided by genomic analysis can be used to detect, delineate and investigate outbreaks of infection. This is particularly useful for outbreaks of infections where current microbiological methods are insufficiently sensitive to determine whether cases of infection which appear similar, implying either transmission or a common source, are in fact part of an outbreak. This application of genomics has been demonstrated to be effective for guiding control of a small number of nosocomial (hospital acquired) infections (including MRSA, *P. aeruginosa* and *C. difficile*) and those occurring predominantly in the community such as *M. tuberculosis* and S*almonella spp*.

2. **Management of tuberculosis –** unlike for most fast growing bacteria and viral pathogens, genomic analysis of *M. tuberculosis* and other related mycobacterial infections has been demonstrated to be an effective diagnostic tool. Compared to current microbiological methods, which rely on extremely slow growth of these bacteria over many weeks in culture, genome sequencing offers significantly more rapid species identification, identification of some (if not all) markers of antibiotic resistance, and utility for detecting, delineating and investigating outbreaks.

3. **Longer term public health surveillance of infections –** genome sequencing offers enhanced resolution, and in some cases reduced cost, compared to existing methods used by specialist and reference microbiology services to undertake monitoring and characterisation of pathogens of particular public health importance. Applications of genomics in this area include:

    a. Monitoring the spread, and understanding the mechanisms, of antibiotic resistance within and between bacterial species

    b. Monitoring patterns and mechanisms of vaccine escape

c. Informing national infection control policy by determining mechanisms of infection transmission within populations and detecting changes in these patterns that might require shifts in public health action

d. Detecting emerging viral pathogens *e.g.* coronaviruses, using culture free approaches

22.2 What potential applications of pathogen genomics currently remain out of reach of clinical and public healthcare?

The vast majority of clinical management of infectious diseases is undertaken on the basis of an identification of the pathogen causing infection and determination of its susceptibility to the drugs available for its treatment. For reasons discussed below, pathogen genomics is currently unable to 'compete' with existing methods in the following areas of diagnostic microbiology:

1. **Primary pathogen identification –** excepting Mycobacteria (see above) there are existing phenotypic or molecular methods available that achieve this task quicker and at dramatically lower cost than is currently possible using genomic methods

2. **Determine drug susceptibility –** excepting Mycobacteria (see above) genomic methods are not sufficiently timely or accurate to supplant existing phenotypic or other molecular methods (which may be NGS based) for determining the susceptibility of bacteria or viruses to antibiotics

3. **Culture free pathogen identification –** whilst genomics has been used to identify rare or emerging viral pathogens directly from clinical samples (where selective culture and amplification are not possible), this approach is not sufficiently well-developed for routine diagnostic use

22.3 To which pathogens is genomic analysis currently applicable in a clinical or public health context?

The laboratory process of pathogen whole genome sequencing is generic and can in principle be applied to any pathogen from which sufficient high quality genomic DNA can be extracted. However, whilst analysis of the genomic data arising from this process can similarly be applied to any pathogen, in practice the current utility of this approach varies widely between pathogens and depends largely on:

- The current state of understanding of their genomic architecture

- Understanding of the the ways in which their genome varies

- Availability of analytical methods with which to detect genomic variation and interpret its significance

Collectively these features determine whether sufficiently robust and accurate whole genome analysis can be performed for any given pathogen, and therefore whether such analysis could even be considered for clinical or public health use.

To date, the translational research community has focused on developing the requisite knowledge base and methods to address the analysis of key nosocomial and community acquired bacterial infections including (but not limited to):

- *S. aureus*
- *C. difficile*
- *M. abscessus*
- *P. aeruginosa*
- *M. tuberculosis*
- *Salmonella* species
- *Streptococcus* species
- *K. pneumoniae*

WGS can also in principle be applied to any viral pathogen. While current focus within the translational research community is on developing the knowledge and methods required to apply whole genome sequencing to HIV and HCV, the extent of any additional clinical or public health utility of analysing the whole genomes of these and other viral pathogens over current approaches that analyse the selected parts of the genome, using NGS based or PCR based methods, is currently unclear.

22.4 What other limitations, intrinsic to genomic technology, hold pathogen genomics back from wider use in current microbiological practice?

There are two key features intrinsic to the current state of genomic technology that significantly restrict the scope of its application.

- **Cost –** whilst the cost of sequencing and analysing the whole genome of a pathogen is ten to twenty fold lower than that of a human whole genome, it is still extremely expensive when compared to most standard microbiological techniques currently used in the investigation and management of infectious disease. It is possible to obtain all clinically relevant information from the majority of samples sent to a microbiology laboratory using culture-based or simple molecular techniques for less than £1 per sample. Pathogen whole genome sequencing currently costs between £50-150 per sample. This restricts its utility to replacing current tests that are even more costly *e.g* some serological typing assays for pathogens such as *Salmonella* that are undertaken in reference or specialist laboratories or to situations in which the overall costs that can be saved across the care pathway exceed the costs of the genomic investigation *e.g.*

where the costs of managing an outbreak of infection outweigh the costs of using genomics to achieve early detection and resolution of the outbreak.

- **Turnaround time –** timeliness in diagnostic microbiology is crucial to effective management of infectious disease, either in the context of initiating appropriate patient treatment, or in the context of enacting appropriate infection control measures to prevent onward transmission of infections. Consequently a significant premium is placed on the use of microbiological tests with rapid turnaround times. For most viral and bacterial pathogens, identification and drug susceptibility testing is achieved within 48 hours and often much less. Excepting Mycobacteria, current whole genome sequencing methods cannot compete with these timescales, with the most optimistic calculations (assuming there is no batching or delays in the laboratory processes) suggesting a total turnaround time from patient sample to interpretable results of at least 72 hours.

22.5 Pathogen genomics and health policy – converting potential into reality

Given the potential for pathogen genomics to transform the management of infectious disease, the outstanding challenge to those charged with securing the health of the nation is how to convert this potential into reality.

Given the potential for pathogen genomics to transform the management of infectious disease, the outstanding challenge to those charged with securing the health of the nation is how to convert this potential into reality. This will mean delivering the changes in policy and practice required both to integrate this new technology into existing systems for microbiological investigation and to ensure that wider and more complex health systems involved in infectious disease management are optimised to capitalise upon on the genomic information they will receive. Achieving this will, in turn, require significant high level strategic commitment to deliver necessary investment in infrastructure and skills, to drive the adaptation to the configuration of health services and to ensure the cross-organisational coordination likely to be critical for success.

Within England, this commitment has manifested most recently in the development of the 100,000 Genomes Project, which whilst focused predominantly on sequencing and utilising information arising from human genomes, also supports Public Health England's efforts to implement pathogen genomics as part of their infectious disease management function. More generally, this project signals the importance placed by government on genomics as a driver of improved health, and potentially wealth, to the English population.

It is, therefore, an ideal time for those within the health system charged with delivering improved management of infectious disease to capitalise on this conducive political climate and develop the necessary support at a strategic and operational level, to realise the benefits of genomics in this domain.

22.6 What is the current state of implementation of pathogen genomics services in the UK?

While there are a number of translational research initiatives working towards the implementation of pathogen genomics in clinical and public health management of infectious disease (Part III), there are as yet no commissioned and fully operational services able to offer accredited diagnostic pathogen whole genome sequencing for use within either clinical or public health settings.

The various translational research and in-house service development initiatives underway across England are at different stages of development. For the most part they have demonstrated the necessary method and knowledge development to underpin the applications of genomics for the range of pathogens they envisage their services will initially target. They have also published numerous proof of principle studies demonstrating the scientific validity of these methods, and in some cases evidence of their clinical validity and utility has also been published.

These initiatives have also established important infrastructure and expertise, including sequencing, computational and analytical capacity, as part of their activities. It should, however, be noted that in many cases this part or all of this infrastructure and expertise resides within academic centres rather than public health or clinical facilities, and so its long term accessibility, sustainability and scalability for service delivery purposes is not clear.

The most advanced of these initiatives are currently undertaking service pilots, but remain to be accredited and evaluated for short and long term cost-effectiveness. The delivery and evaluation of these ongoing pilots, will be crucial to developing and embedding important infrastructure and expertise in the small number of labs they involve. In addition, this will provide the evidence base with which to support future wider implementation if they are successful.

22.7 Conclusions

The current scope of application for pathogen genomics in the real world is significantly more limited than that which could be achieved in principle, largely due to limitations in current technology and knowledge. There are, however, several important areas of infectious disease management in which there is a demonstrable, realistic and immediate prospect that the application of pathogen genomics could have significant positive impact at both the individual patient and population level.

The aim of this report is to support the development and delivery of genomics informed infectious disease services that are evidence based, high quality, available population wide, and on an equitable basis.

To meet this goal, given the current state of the field of pathogen genomics, there are therefore two key objectives to be met:

1. The widest possible implementation of pathogen genomics services for which there is already demonstrable scientific and clinical validity and clinical utility. This will include the maximisation of the effectiveness and impact of their operations on patient and population outcomes.

2. The long term development and delivery of an expanded range of infectious disease management services based on genomics that will significantly improve their quality and effectiveness.

In the next chapter we set out in our roadmap how the policy and practice recommendations presented in this report can be used to support the achievement of the above objectives. We emphasise what needs to be done – by strategic policy makers, frontline service providers and users – to ensure appropriate genomics services are developed and delivered in the short term. We also consider how the rapid pace of genomic technology and knowledge development can be harnessed effectively to ensure that future innovations in this highly dynamic field are capitalised upon as rapidly as possible to bring their benefits to patients as soon as practicable. Finally we present the case for developing a 'catalyst' that, if implemented by policy makers, could significantly increase the chances of success in achieving these objectives.

23 Where do we go from here? The roadmap

The recommendations in this report aim to overcome barriers that risk limiting progress in service delivery and compromising benefits to patients and to population health. Here, we propose a roadmap of these recommendations leading to the development of a 'catalyst' as a mechanism for integrating and accelerating service development.

23.1 Context

23.1.1 The complexity of infectious diseases and the systems that manage them

The structure and function of the systems and pathways that underpin the management of infectious diseases are complex and diverse, depending on (amongst many factors) the characteristics of the pathogen causing the infection, the magnitude and nature of the effects it has on the health of individual patients and the ways in which it is transmitted. For example, infectious disease management services have to address challenges as diverse as the long term management of epidemics of chronic viral diseases such as HIV and HCV, the burden of acute (but usually) self-limiting gastrointestinal and respiratory infections in the community, prevention of outbreaks of healthcare associated infections amongst highly susceptible populations of patients, and rare but potentially lethal cases of imported tropical diseases. Tackling such diverse threats requires a similarly diverse range of organisations and professionals to work in a coordinated fashion to deliver the range of services – including screening, vaccination, microbiological investigation, clinical care and infection control – necessary to varying extents to manage each clinical scenario.

23.1.2 The universality of the genome – reducing complexity, increasing accuracy and effectiveness

The universality of the genome as the blueprint that determines the behaviour of all pathogens (in combination with their environment) offers an unprecedented opportunity to simplify much of this vast complexity by encoding most of the key characteristics relevant to the management of any pathogen in a single, machine readable 'language'. Reading and interpreting the genomes of pathogens could potentially become the principal method through which information about organisms causing infections is gathered, synthesised and deployed to enable the reduction in their impact on the health of our population. This approach would not eliminate completely the complexity inherent in the management of infectious diseases. Much of this is determined by non-genomic factors such as human behaviour, our environment and the characteristics of available strategies to manage infection. However, by simplifying the parts of the services that rely on accurate characterisation of the pathogens themselves, genomics could significantly improve the efficiency and effectiveness of their delivery.

The benefits of using genomics arise not only from the unifying simplicity of applying a single analytical technology to a diverse range of organisms, but also from the potential to provide higher resolution analysis of the characteristics of these organisms in each case than is achievable by current methods (Part II).

> *The benefits of using genomics arise not only from the unifying simplicity of applying a single analytical technology to a diverse range of organisms, but also from the potential to provide higher resolution analysis of the characteristics of these organisms in each case than is achievable by current methods.*

23.1.3 Delivering on the potential of genomics

In the concluding parts of this report we outline how, through meeting the objectives outlined at the end of the previous chapter, the potential of genomics to enhance the effectiveness of infectious disease management services can be achieved. We set out:

- The principles that must underlie any successful genomics-informed approach to infectious disease management
- The roadmap as our blueprint for how this report's recommendations support the development of processes and systems needed to embed genomics in infectious disease management
- The catalyst as a mechanism for integrating and accelerating service development and delivery that will maximise both service effectiveness in the short term, and the rate of expansion of services and their impact in the longer term, principally through the sharing of data, knowledge, expertise and strategic activities
- The key messages for each of the three key stakeholder groups we have identified as having the ability to deliver on the recommendations

23.2 Principles for delivering a high quality, evidence based approach to genomics enabled infectious disease management systems

The aim of our report is to emphasise four key characteristics of the genomics enabled infectious disease management system that we envisage being delivered in the UK: high quality, evidence based, population wide availability and equity of access.

Any approach to delivering such a system will need to be:

- Driven by open and effective data and knowledge management, exchange and access
- Underpinned by a strong scientific, clinical and health economic evidence base
- Delivered by healthcare professionals with the right expertise
- Responsive to fluctuating local needs and the variability in user characteristics across the healthcare system
- Able to operate at a scale that can deliver access nationwide and according to need
- Focused on continuously improving the effectiveness, capacity, and scope of services
- Able to integrate emerging knowledge and technology into services rapidly and effectively
- Rigorously evaluated to ensure patient and population benefit is achieved
- Based on coordinated policies developed across organisations involved in managing infectious disease to maximise effectiveness and value for money

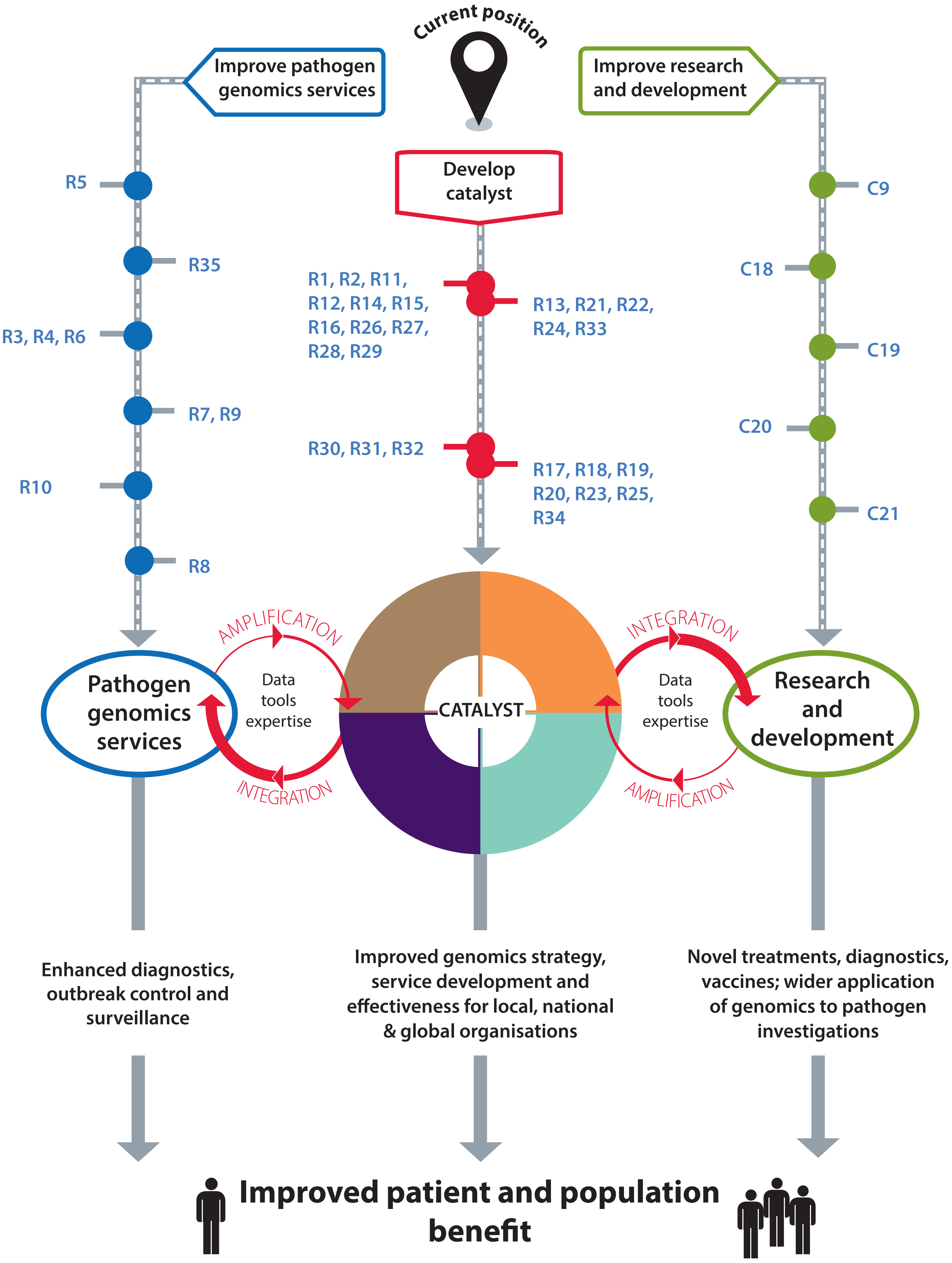
23.3 The roadmap

23.3.1 Introduction

Throughout our report we have developed a series of recommendations for policy and practice to meet the dual objectives of maximising the effectiveness of services that can be implemented now and accelerating the rate at which these services are developed to expand their scope and improve their quality in the longer term. In the graphic below we illustrate how our recommendations 'map' onto the paths to achieving these objectives. We identify two distinct pathways:

- **Service delivery –** this path focuses on the steps that need to be taken to establish and deliver an effective genomics enabled service that aims to bring benefits to patients and population health
- **Research and development –** this path emphasises the actions that need to be undertaken to accelerate the development of genomic technology, methodology and knowledge that will underpin the expansion in scope and improvement in quality of services that are to be delivered to patients

Figure 23.1 Roadmap for pathogen genomics strategy



Current position
Improve pathogen genomics services
Develop catalyst
Improve research and development
R5
R35
R3, R4, R6
R7, R9
R10
R8
R1, R2, R11, R12, R14, R15, R16, R26, R27, R28, R29
R13, R21, R22, R24, R33
R30, R31, R32
R17, R18, R19, R20, R23, R25, R34
C9
C18
C19
C20
C21
AMPLIFICATION
Data tools expertise
INTEGRATION
Pathogen genomics services
CATALYST
INTEGRATION
Data tools expertise
AMPLIFICATION
Research and development
Enhanced diagnostics, outbreak control and surveillance
Improved genomics strategy, service development and effectiveness for local, national & global organisations
Novel treatments, diagnostics, vaccines; wider application of genomics to pathogen investigations
Improved patient and population benefit

* R = recommendation / C = chapter

Path	Recommendations / chapters	What they refer to
Improve pathogen genomics services	R5	Assess need
	R35	Establish validity and utility
	R3, R4, R6	Configure
	R7, R9	Accredit
	R10	Quality assure
	R8	Evaluate
Develop catalyst	R1, R2, R11, R12, R14, R15, R16, R26, R27, R28, R29	Strategic coordination and investment
	R13, R21, R22, R24, R33	Ensure collation and curation of samples and data
	R30, R31, R32	Standardise data formats
	R17, R18, R19, R20, R23, R25, R34	Enable and regulate data access
Improve research and development	C9	Wider applications of genomics in infectious disease
	C18	Building an evidence base I: developing, demonstrating and evaluating clinical utility
	C19	Building an evidence base II: cost effectiveness of pathogen genomics services
	C20	ELSI when implementing pathogen genomics
	C21	Delivering safe and effective services in a dynamic technology and knowledge environment

23.3.2 The need for integration and acceleration

As described throughout this report, a small number of groups of scientists and clinicians are already pursuing the actions set out along both the service delivery and research and development paths within their own research and health delivery organisations. The challenges that we aim to address with our roadmap are, therefore, how to maximise the effectiveness, accelerate the rate of development and broaden the scope of these activities.

Our analysis and research has identified several barriers that must be overcome if these objectives are to be achieved:

- The negative effect of a highly fragmented and unstable health innovation and delivery system on the effectiveness of implementation efforts

- Absence of mechanisms and infrastructure to facilitate integration, development and exchange of data, knowledge and expertise

- The need for multiple organisations with varying levels of awareness, engagement and capability with respect to pathogen genomics to develop coordinated strategic approaches to ensure effective development and delivery of services

Unless these are successfully addressed, only limited progress towards service delivery can be made. Progress in research and development will doubtless continue. However, the rate at which the results of these activities are converted into patient and population health benefit nationwide will be severely compromised.

In the next section we propose the development of a 'catalyst' as a mechanism to achieve the integration required at every level, from strategy development to frontline service delivery, that can overcome these challenges and support the realisation of the aims set out in this report.

23.4 The catalyst

23.4.1 Definition

The catalyst is: '*A set of real or virtual structures that amplifies and integrates the current activities in pathogen genomics to accelerate and increase the effectiveness of their impact on patient and population health.*'

23.4.2 What should the catalyst do?

The catalyst has four principal functions that collectively address the need to establish an integrated approach to the development and delivery of genomics enabled infectious disease management services:

1. Repository function

- As highlighted in chapter 15, there is a clear need for infrastructure to be developed that enables all data generated during the delivery of genomics informed services to be captured, integrated and shared for the purposes of enhancing the effectiveness of the infection services that are being delivered locally and the performance of nationwide public health surveillance and control of infection

- An integrated repository of genomic data, clinical and epidemiological data and biological sample archives will also be a valuable resource that, through managed access, can catalyse the development of new diagnostic genomic technologies, analytical methodologies and therapeutics or vaccines that will enhance the long term effectiveness of infectious disease management services

- A repository does not have to manifest as one physical piece of infrastructure. Storage of data and samples is likely to be centralised in some cases *e.g.* all genomic data may be placed in a public repository such as ENA, and distributed in others *e.g.* the maintenance of local sample archives. What is important is that mechanisms are established for responsible oversight, managing and curating the data and samples collected and in particular for facilitating deposition and access to the contents of the repository by all legitimate providers and users

- Provision of this repository function will require investment and co-ordination across a range of organisations, including PHE, NHSE and various national / international research organisations. Overarching strategic leadership will therefore be required for its delivery and could be provided by the Department of Health.

2. Collaborative function

- Accelerating the rate at which knowledge is generated from the data deposited in the repository and translated into new products and services that can be deployed in the health service depends on the effectiveness of the interaction of those involved

- Convening mechanisms need to be established, or strengthened where they already exist, to ensure that those engaged in research and development in both public sector and private sector organisations are engaged with frontline service providers and strategic policy makers in PHE and the NHS to ensure that research and development activity is focused on and responsive to the needs of the health service, and to ensure that results are effectively communicated and used by these services

- Collaboration between groups engaged in research and development must also be improved, through deliberate knowledge brokering activities, in order to maximise the efficient use of resources, particularly where multiple groups are working towards common objectives and there are opportunities to reduce duplication of effort, to more rapidly converge on optimal solutions to technological and analytical challenges, and more effectively translate these into new service developments

3. Standardisation and expertise diffusion function

- In chapter 14 we identify the need to develop best practice guidance and standards for the delivery and use of genomics services. Uptake of these will underpin the nationwide availability of the highest quality services based on genomics by providing a route to ensure that knowledge of 'gold standard' laboratory processes and clinical pathways and advice on their adoption are available regardless of location

- Standards established will also form a critical part of ensuring the quality of the data submitted to the repositories, which in turn will underpin both effective service delivery and development

- Standards and guidance will need to be developed by a range of expert scientific and health professional groups working collaboratively. A key function of the catalyst is, therefore, to provide convening mechanisms through which these groups can come together to develop, review and communicate the guidance and standards

- The catalyst can also facilitate the development and diffusion of analytical expertise by facilitating collaborative development and validation of analytical methods that can then be distributed to all those wishing to undertake genome analysis as part of their infection management services

4. Strategic coordination and development function

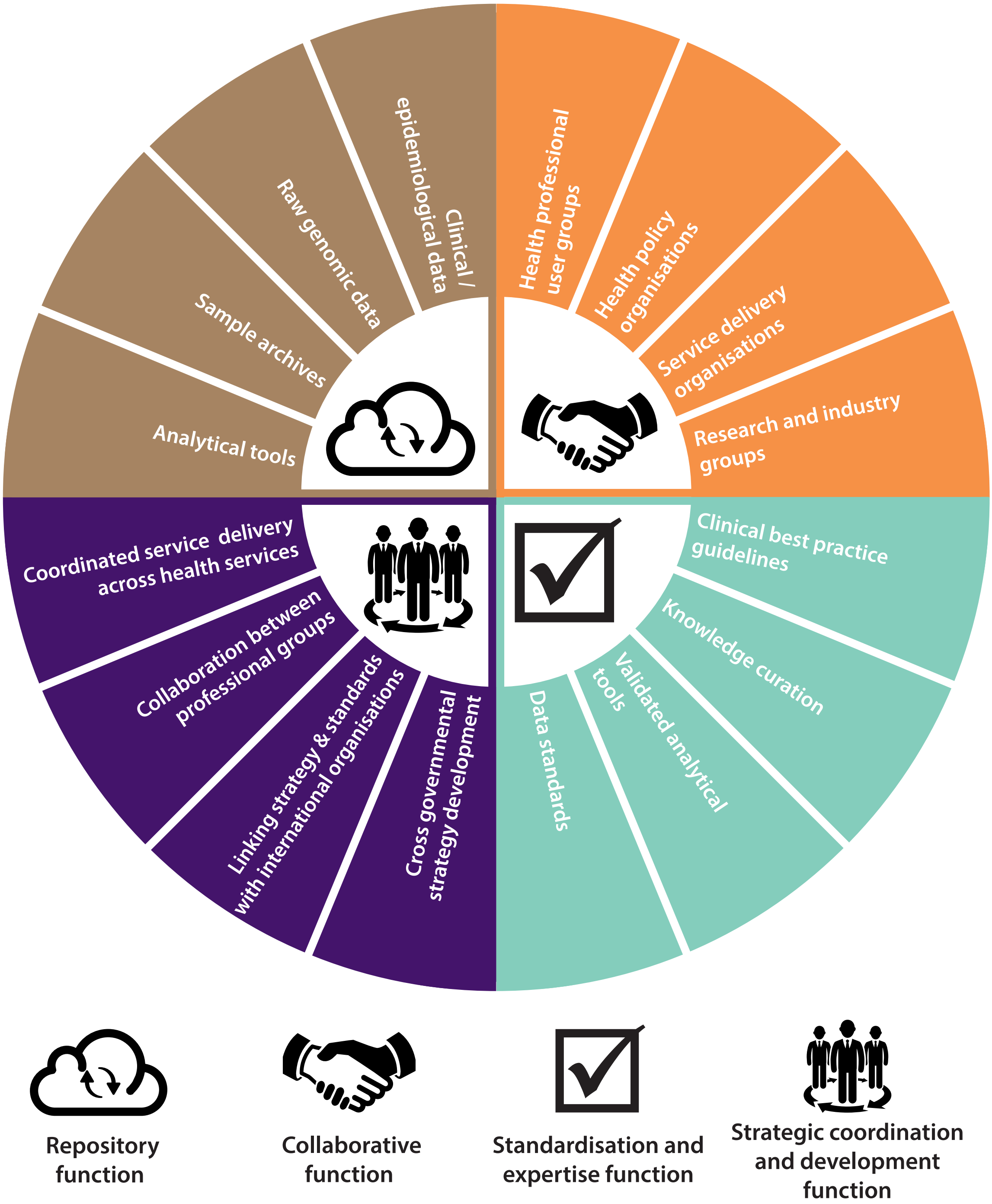
- A wide range of organisations are potentially involved in the development and delivery of genomics enabled services. In addition to the bottom up collaborative approaches described above the catalyst should contain 'top down' leadership and strategic coordination functions to ensure the delivery of the other functions and enhance their effectiveness

- A leadership group, which could be led by the Department of Health and including representatives from all relevant organisations, professions and groups, must be established to oversee and drive the development and delivery of the repository and collaborative functions of the catalyst

- This leadership group will need to ensure that where multiple organisations are involved in service delivery and development (*e.g.* PHE, NHS, APHA and FSA) or research and knowledge generation (*e.g.* research funders, industry partners and academia) they co-develop strategies that ensure coordinated delivery of their objectives to maximise health benefit and use of resources

- This leadership group should also seek to engage at a high level with international governmental and non governmental organisations involved in managing infectious diseases to ensure that efforts to develop genomic approaches in the UK are well aligned with those being undertaken in other countries, and to lead in, or at least participate any efforts to harmonise and standardise these approaches

Figure 23.2 Catalyst



Clinical / epidemiological data
Raw genomic data
Sample archives
Analytical tools
Health professional user groups
Health policy organisations
Service delivery organisations
Research and industry groups
Coordinated service delivery across health services
Collaboration between professional groups
Linking strategy & standards with international organisations
Cross governmental strategy development
Clinical best practice guidelines
Knowledge curation
Validated analytical tools
Data standards
Repository function
Collaborative function
Standardisation and expertise function
Strategic coordination and development function

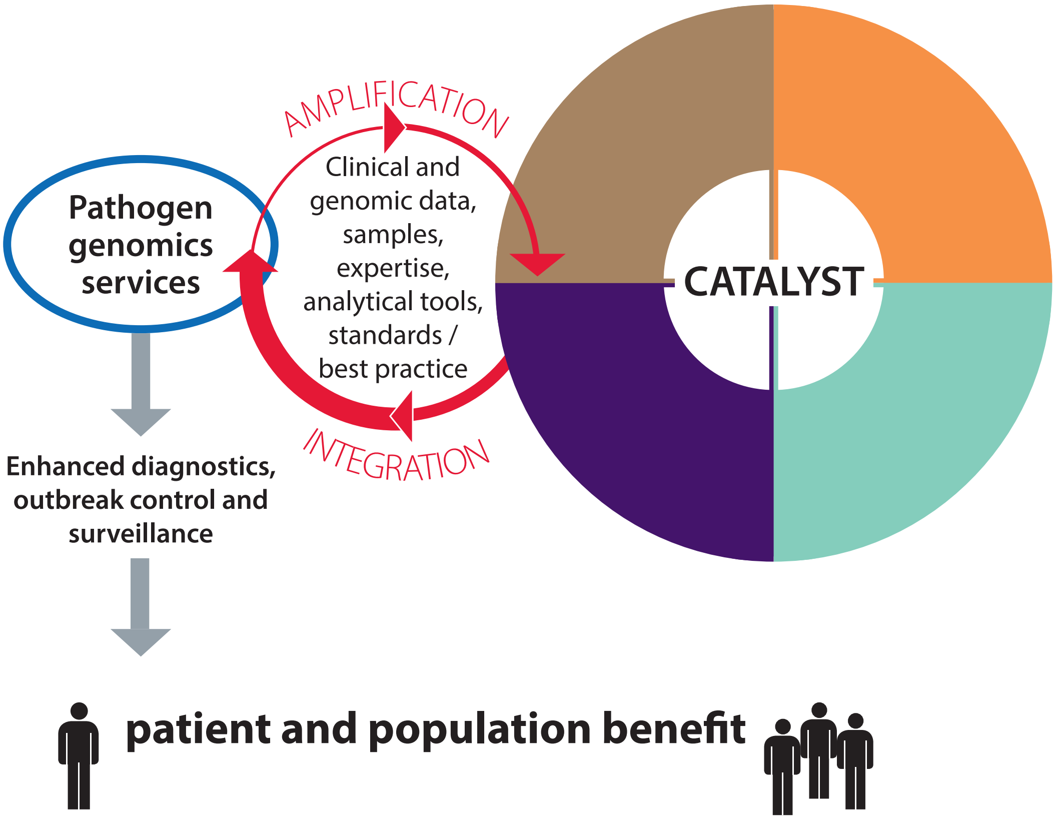
24.4.3 How can the catalyst accelerate the delivery of patient and population benefit from pathogen genomics services?

We envisage that establishment of the catalyst will accelerate the delivery of patient and population benefit within two positive feedback loops that can be defined within the roadmap:

1. Current service delivery

Those organisations involved in the early waves of pathogen genomics service delivery will deposit data in the repository and share expertise, analytical methods and best practice as they develop. Data deposited in the repository can be used to enhance outbreak detection, refine diagnostics and inform surveillance for all participating service delivery organisations. Expertise and methods shared through the catalyst will support continuous improvement in the quality of services delivered by each participating organisation. Together these two self-reinforcing processes will have a significant positive impact on the benefits delivered by these services to patients and populations.

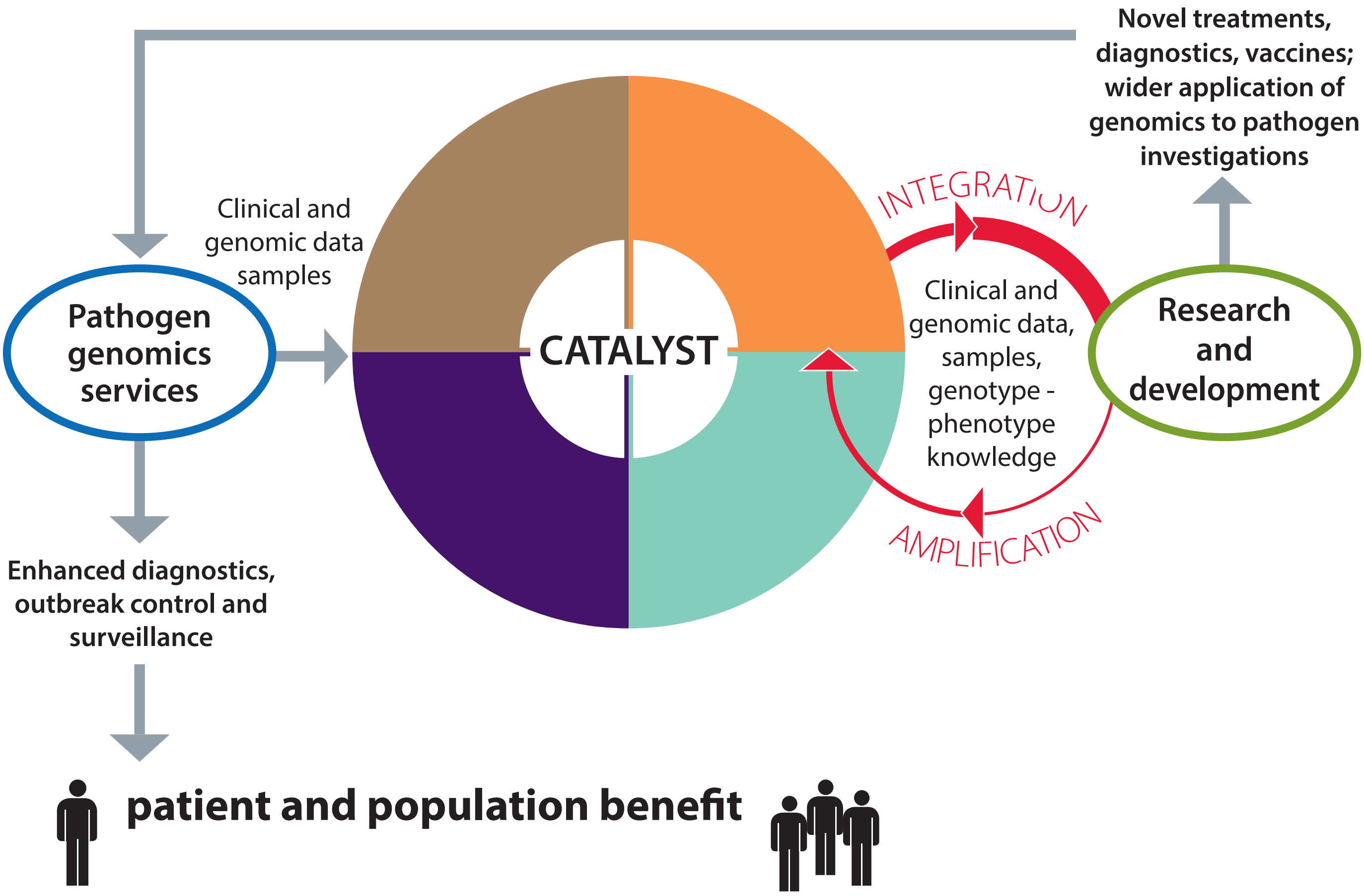
Figure 23.3 Catalyst impact on current service delivery


AMPLIFICATION
Pathogen genomics services
Clinical and genomic data, samples, expertise, analytical tools, standards / best practice
CATALYST
INTEGRATION
Enhanced diagnostics, outbreak control and surveillance
patient and population benefit

2. *Future service development*

In this context genomic and clinical data, samples and knowledge are deposited and shared within the repository by both service delivery organisations and research and development groups. The catalyst will facilitate access to data and knowledge by a wide range of research and development organisations, and through enabling collaborative interactions between them drive the development of knowledge and products that can then be fed back into service delivery organisations for evaluation and implementation. This loop is also self-reinforcing, with the amount and quality of data and knowledge fed into the catalyst by service delivery organisations and research and development groups driving up the quality and capacity of the services that can be delivered, and thus the data and knowledge they generate. Most importantly these increases in service quality, scope and capacity will drive improvements in patient and population health outcomes.

Figure 23.4 Catalyst impact on future service delivery


Novel treatments, diagnostics, vaccines; wider application of genomics to pathogen investigations
Clinical and genomic data samples
Pathogen genomics services
CATALYST
INTEGRATION
Clinical and genomic data, samples, genotype - phenotype knowledge
AMPLIFICATION
Research and development
Enhanced diagnostics, outbreak control and surveillance
patient and population benefit

24.5 Key messages to stakeholders

This report contains a number of recommendations that will be relevant to varying extents to the different stakeholders involved in organising, providing and using services involving pathogen genomics. Beyond these individual recommendations, our concluding messages to the three key stakeholder groups (policy makers, providers / users and purchasers) are as follows:

Health policy makers – it is vital that you ensure the building of the catalyst, whilst also continuing to resource the implementation and research and development arms of the roadmap, and work strategically across all organisations involved in managing infectious disease to ensure implementation is coordinated and effective.

Frontline practitioners (providers and users) – follow our recommendations on service configuration and link to the catalyst to ensure effective service development within your branch of the health system. Emphasise working together to develop standards and best practice. Pioneers must take responsibility for building evidence base and networking with new providers to demonstrate feasibility and cost-effectiveness of genomics services.

Frontline health service managers / commissioners (purchasers) – be aware that this technology is almost ready for implementation, start to assess your needs and consider how its use could benefit you, predominantly in managing nosocomial infections, but in future for a much wider range of frontline diagnostic applications too. You will need to link to the catalyst to access the necessary resources to develop new services. Work with PHE specialist laboratories to access control of infection services in the short term, and follow our recommendations to support effective development and utilisation of services in your own context in the longer term.

Recommendations

Recommendation no.	Recommendation	Report section
1	PHE will need to work with all microbiology service providers, both public and private sector, to ensure that they participate fully in meeting requirements to contribute to national infectious disease surveillance, through appropriate contributions to the implementation and development of pathogen genomics services.	11.2
2	Agreement needs to be reached between PHE and NHSE with regards to funding for service development and delivery where the pathogen genomics services have a dual clinical and public health benefit.	11.2
3	The initial implementation of pathogen genomics services should be focused in laboratories providing consolidated microbiology services, as these are most likely to be able to realise necessary economies of scale and to achieve the concentrations of expertise and efficient data management required.	11.2
4	A defined pathway, encompassing test referral mechanisms, sequencing, analysis and interpretation must be developed for each pathogen and each application of genomics. Implementation of these pathways will require a coordinated approach.	13.2
5	Robust and effective prioritisation processes will need to be developed for new service developments. These must be informed by consultation including frontline end user groups.	13.2
6	The location of sequencing and analysis services should not be pre-determined, and a mixed model should be allowed to develop which makes optimal use of available resources and takes account of local / national demand for genomics: variables include the cost, throughput achievable at each location, and turnaround time.	13.4
7	All laboratories providing clinical pathogen genomics services need to be accredited to the appropriate national / international standards.	14.2
8	Evaluation and comparison of test performance should span the whole process from sample extraction to clinical report, encompassing assessments of both analytical and clinical validity and clinical utility.	14.2
9	The clinical and public health microbiology 'community' needs to work with UKAS and NEQAS to establish standards that can be used to develop appropriate accreditation processes.	14.2
10	In order to ensure that services are of sufficiently high quality, and delivered in a consistent manner, guidelines (equivalent to SMIs) establishing minimum standards for pathogen genomics services must be developed.	14.2
11	Develop a national collaborative network of pathogen genomic service providers to share knowledge and best practice, collaborate on service and methodology development and agree standards for clinical and public health service delivery.	14.2

Recommendation no.	Recommendation	Report section
12	Realisation of the strategic public health benefits of the implementation of pathogen genomics services will require coordinated action amongst providers and users to develop underpinning policies and procedures to support co-operation and inter-operation of services. These efforts should be led by Public Health England but be explicitly supported by all relevant health service and policy making organisations.	14.3
13	Criteria must be established to decide under what circumstances sequenced pathogen isolates (or related clinical materials) must be stored for future public health use, timescales for any storage requirements and sources of funding to ensure sustainability of any sample archives created.	14.3
14	Additional investment will be required, above that envisaged for the development of individual pathogen genomics services, to build the infrastructure and capacity required to realise the broader and longer term public health benefits of the implementation of pathogen genomics for disease surveillance, treatment and prevention.	14.3
15	Existing links between the infectious disease aspects of animal and human health services should be exploited and strengthened to ensure that synergies in the developments of their genomics programmes are realised and a 'One Health' approach to managing infectious disease threats can be developed where appropriate.	14.4
16	Organisations leading on the development and delivery of pathogen genomics in the UK should work with and show leadership within transnational organisations and specific international genomics focused initiatives to ensure that best practice is shared and sufficiently standardised, or at least interoperable datasets are developed and regulatory barriers to effective genomic and metadata exchange are addressed.	14.4
17	When considering data release to a publicly accessible database, stakeholders should adopt proportionate safeguards that balance the need to protect the interests of data subjects, particularly relating to privacy and confidentiality, against the likely benefits of proceeding with data sharing.	15.6
18	Raw genomic data and minimal metadata ought to be shared as widely as possible (following appropriate QC and assuming public release is approved) preferably through public data repositories to ensure long term sustainability.	15.6
19	Criteria for defining what minimal data sets are appropriate for release to publicly accessible databases should be developed, with risk assessments being undertaken to identify in particular which elements of metadata can be released publicly for each pathogen. PHE (and their Office of Data Release) would be best placed to deliver on this, along with NHS input.	15.6
20	It must be mandatory for all providers of NHS or PHE pathogen genomic investigations to make sequence data and all other necessary clinical and epidemiological data available for use by legitimate NHS healthcare and public health professionals within agreed timeframes, for the purpose of delivering their stipulated functions. A mandate needs to be implemented urgently to prevent data that is currently being generated from being lost in silos.	15.8

Recommendation no.	Recommendation	Report section
21	The benefits of data collation and risks of not aggregating data should be articulated to those being mandated to submit data. A feedback or reward strategy should be developed to gain longer term accord with and practical support for a data sharing mandate, and investment made in adequate infrastructure to enable data deposition at the practical level.	15.8
22	All pathogen genomic data and associated metadata required by healthcare and public health professionals to maximise the effectiveness of their management of infectious disease in individual patients and populations should be submitted to the designated database without delay.	15.10
23	Where data release into the public domain is envisaged / considered, a strategy for the timing of genomic data and limited metadata release that takes into account a balance between the need to serve wider public health benefit and the rights of individuals and organisations, should be devised. Provision should be made for access by researchers, companies, and healthcare and public health professionals outside the UK.	15.10
24	A public health authority such as PHE should be responsible for the collation and storage of all genomic data and metadata for the purposes of clinical and public health service delivery, and to support the development of new clinical and public health applications of genomics in the early stages of implementation until solutions can be developed in collaboration with databases such as ENA to provide access to the necessary storage and expertise to build and maintain an optimal sharing system in the longer term.	15.14
25	Accessible interfaces or software tools must be developed that meet the needs of clinical users by enabling straightforward access to the information in genomic and metadata databases and to facilitate the ability of legitimate users to perform analyses on underlying data.	16.2
26	Pathogen genomics service providers will need to invest in developing and maintaining, or procuring remote access to, sufficient computational capacity to enable their data analysis.	16.3
27	PHE and HEE should continue to work together to ensure that education and training are provided to support the development of the bioinformatics workforce and the analytical and interpretive skills of frontline users of pathogen genomics services.	16.3
28	Additional investment to increase the availability of bioinformaticians able to develop and deliver pathogen genome analytical services will be required, at least in the short term, until analytical tools operable by the existing laboratory and clinical workforce are developed.	16.4
29	A PHE led strategy for the organisation of access to computational infrastructure and bioinformatics expertise will be required to ensure access to genome analysis services is not an impediment to the implementation of genomics services.	16.4

Recommendation no.	Recommendation	Report section
30	Agreement is required on the standards for genomic data quality and format across laboratories undertaking pathogen genomic analysis for clinical and public health investigations. There should also be mechanisms for standardising descriptive clinical and epidemiological information relating to genomic data to maximise the interoperability, and therefore the utility, of data collected across different locations.	17.2
31	Mechanisms need to be developed by relevant professional groups for benchmarking the performance of equivalent genome analysis methods, and for ensuring that methods used in service settings meet minimum standards.	17.3
32	In order to support greater interoperability of data generated across the health system there should be mechanisms (preferably international) established for standardising nomenclature for genomic characteristics of pathogens and their relatedness.	17.4
33	Curation of each organism specific genotype-phenotype database and analytical pipeline, and archiving of isolate / tissue collections must be under the control of a designated responsible authority. Each authority should operate with PHE oversight and funding to support their sustainability.	17.5
34	The challenges of integrating clinical and genomic data, enabling data interoperation and delivering user friendly service requisitioning and reporting interfaces across different LIMS and IT systems need to be addressed. This will require agreement on data management standards between all organisations involved in delivering or using pathogen genomics services.	17.5
35	Evidence for clinical and public health utility and cost effectiveness will need to be clearly demonstrated prior to funding and adoption of pathogen specific genomics services by clinical and public health end-users.	18.4

References

1. Palazzo AF and Gregory TR. The case for junk DNA. PLoS Genet. 2014 May 8; 10(5): e100435.

2. Conlan S, Thomas PJ, Deming C *et al*. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med. 2014 Sep 17; 6(254): 254ra126.

3. Improved data reveals higher global burden of tuberculosis. [Internet] 2014. Available from: www.who.int/mediacentre/news/notes/2014/global-tuberculosis-report/en Accessed 14 May 2015.

4. Tuberculosis mortality and mortality rate, England and Wales, 1913-2013. [Internet] 2013. Available from: www.gov.uk/government/uploads/system/uploads/attachment_data/file/363056/Tuberculosis_mortality_and_mortality_rate.pdf Accessed 14 May 2015.

5. Cole ST, Brosch R, Parkhill J *et al*. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998 Jun 11; 393(6685): 537-44.

6. Deaths Involving MRSA: England and Wales, 2008 to 2012. [Internet] 2012. Available from: www.ons.gov.uk/ons/dcp171778_324558.pdf Accessed 14 May 2015.

7. Holden MT, Feil EJ, Lindsay JA *et al*. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci U S A. 2004 Jun 29; 101(26): 9786-91.

8. Harris SR, Cartwright EJ, Török ME *et al*. Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. Lancet Infect Dis. 2013 Feb; 13(2): 130-6.

9. Deaths Involving *Clostridium difficile,* England and Wales, 2012. [Internet] 2012. Available from: www.ons.gov.uk/ons/rel/subnational-health2/deaths-involving-clostridium-difficile/2012/stb-deaths-involving-clostridium-difficile-2012.html Accessed 14 May 2015.

10. Sebaihia M, Wren BW, Mullany P *et al*. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet. 2006 Jul; 38(7): 779-86.

11. Eyre DW, Cule ML, Wilson DJ *et al*. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med. 2013 Sep 26; 369(13):1195-205.

12. Williams ÇH and Stanway G. Viruses: Genomes and Genomics. eLS. 2009.

13. Number of deaths due to HIV/AIDS. [Internet] 2014. Available from: www.who.int/gho/hiv/epidemic_status/deaths_text/en/ Accessed 14 May 2015.

14. HIV in the United Kingdom: 2010 Report. [Internet] 2010. Available from: www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1287145367237 Accessed 14 May 2015.

15. HIV in the United Kingdom. [Internet] 2014. Available from: www.gov.uk/government/uploads/system/uploads/attachment_data/file/401662/2014_PHE_HIV_annual_report_draft_Final_07-01-2015.pdf Accessed 14 May 2015.

16. Hepatitis C [Internet]. 2014. Available from: www.who.int/mediacentre/factsheets/fs164/en Accessed 14 May 2015.

17. General Information on hepatitis C. [Internet] 2013. Available from: www.gov.uk/government/collections/hepatitis-c-guidance-data-and-analysis Accessed 14 May 2015.

18. Cabinet Office. National Risk Register of Civil Emergencies. [Internet] 2013. Available from: www.gov.uk/government/publications/national-risk-register-for-civil-emergencies-2013-edition Accessed 14 May 2015.

19. Dawood FS, Iuliano AD, Reed C *et al*. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis. 2012 Sep; 12(9): 687-95.

20. Cumulative number of confirmed human cases for avian influenza: A(H5N1) reported to WHO, 2003-2012. [Internet] 2012. Available from: www.who.int/influenza/human_animal_interface/EN_GIP_20120810CumulativeNumberH5N1cases.pdf Accessed 14 May 2015.

21. Fricke WF, Rasko DA and Ravel J. The role of genomics in the identification, prediction, and prevention of biological threats. PLoS Biol. 2009 Oct; 7(10): e1000217.

22. Duffy S, Shackelton LA and Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008 Apr; 9(4): 267-76.

23. Sanger F, Coulson AR, Friedmann T *et al*. The nucleotide sequence of bacteriophage phiX174. J Mol Biol. 1978 Oct 25; 125(2): 225-46.

24. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan; 11(1): 31-46.

25. Dark MJ. Whole-genome sequencing in bacteriology: state of the art. Infect Drug Resist. 2013 Oct 8; 6: 115-23.

26. Bentley DR, Balasubramanian S, Swerdlow HP *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov 6; 456(7218): 53-9.

27. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan; 11(1): 31-46. {Internet] Figure at: www.nature.com/nrg/journal/v11/n1/fig_tab/nrg2626_F2.html Accessed 14 May 2015.

28. Karow J. Ion Torrent Systems Presents $50,000 Electronic Sequencer at AGBT. GenomeWeb In Sequence. [Internet] 2010. Available from: www.genomeweb.com/sequencing/ion-torrent-systems-presents-50000-electronic-sequencer-agbt Accessed 14 May 2015.

29. Rothberg JM, Hinz W, Rearick TM *et al*. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011 Jul 21; 475(7356): 348-52.

30. Vogel U, Szczepanowski R, Claus H *et al*. Ion torrent personal genome machine sequencing for genomic typing of *Neisseria meningitidis* for rapid determination of multiple layers of typing information. J Clin Microbiol. 2012 Jun; 50(6): 1889-94.

31. Sherry NL, Porter JL, Seemann T *et al*. Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. J Clin Microbiol. 2013 May; 51(5): 1396-40.

32. Mellmann A, Harmsen D, Cummings C *et al*. Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. PloS ONE. 2011 07/20; 6(7): e22751.

33. Harris SR, Török M, Cartwright EJP *et al*. Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks. Nat Biotech. 2013 Jul 9; 31(7): 592-4.

34. [Internet] Available from: www.genomics.cn/en/navigation/show_navigation?nid=2640 Accessed 14 May 2015.

35. Eid J, Fehr A, Gray J *et al*. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2; 323(5910): 133-8.

36. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan; 11(1): 31-46. [Internet] Figure at: www.nature.com/nrg/journal/v11/n1/fig_tab/nrg2626_F4.html Accessed 14 May 2015.

37. Aksimentiev A, Heng JB, Timp G *et al*. Microscopic Kinetics of DNA Translocation through Synthetic Nanopores. Biophysical Journal. 2004 Sept; 87(3): 2086-97.

38. Garaj S, Hubbard W, Reina A *et al*. Graphene as a subnanometre trans-electrode membrane. Nature. 2010 Sept 9; 467(7312): 190-3.

39. Howorka S, Cheley S and Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. Nat Biotech. 2001 Jul print; 19(7): 636-9.

40. Kircher M and Kelso J. High-throughput DNA sequencing--concepts and limitations. Bioessays. 2010 Jun; 32(6): 524-36.

41. Branton D, Deamer DW, Marziali A *et al*. The potential and challenges of nanopore sequencing. Nature Biotechn. 2008 Oct; 26(10): 1146-53.

42. Ashton PM, Nair S, Dallman T *et al*. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotech. 2014 Dec 8: 2015; 33: 296-300.

43. Loman NJ and Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics. 2014 Dec 1; 30(23): 3399-401.

44. Quick J, Quinlan AR and Loman NJ. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. Gigascience. 2014 3: 22.

45. Watson M, Thomson M, Risse J *et al*. poRe: an R package for the visualization and analysis of nanopore sequencing data. Bioinformatics. 2015 Jan 1; 31(1): 114-5.

46. Schaffer A. 10 Breakthrough Technologies. MIT Tech Rev. [Internet] 2012 June. Available from: www2.technologyreview.com/article/427677/nanopore-sequencing Accessed 14 May 2015.

47. Loman NJ, Misra RV, Dallman TJ *et al*. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotech. 2012 May; 30(5): 434-9.

48. Quail MA, Smith M, Coupland P *et al*. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13: 341.

49. Stewart EJ. Growing unculturable bacteria. J Bacteriol. 2012 Aug; 194(16): 4151-60.

50. Koser CU, Fraser LJ, Ioannou A *et al*. Rapid single-colony whole-genome sequencing of bacterial pathogens. J Antimicrob Chemother. 2014 May; 69(5): 1275-81.

51. Wilson DJ. Insights from genomics into bacterial pathogen populations. PLoS Pathog. 2012 Sep; 8(9): e1002874.

52. Nagarajan N and Pop M. Sequence assembly demystified. Nat Rev Genet. 2013 Mar; 14(3): 157-67.

53. Williams D, Trimble WL, Shilts M *et al*. Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC Genomics. 2013; 14: 537.

54. Rainey PB and Bailey MJ. Physical and genetic map of the *Pseudomonas fluorescens* SBW25 chromosome. Mol Microbiol. 1996 Feb; 19(3): 521-33.

55. Lupski JR and Weinstock GM. Short, interspersed repetitive DNA sequences in prokaryotic genomes. J Bacteriol. 1992 Jul; 174(14): 4525-9.

56. van Belkum A. Short sequence repeats in microbial pathogenesis and evolution. Cell Mol Life Sci. 1999 Nov 30; 56(9-10): 729-34.

57. Miller JR, Koren S and Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010 Jun; 95(6): 315-27.

58. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010 Sep; 11(5):473-83.

59. Chain PS, Grafham DV, Fulton RS *et al*. Genomics. Genome project standards in a new era of sequencing. Science. 2009 Oct 9; 326(5950): 236-7.

60. Fraser CM, Eisen JA, Nelson KE *et al*. The value of complete microbial genome sequencing (you get what you pay for). J Bacteriol. 2002 Dec; 184(23): 6403-5.

61. Galardini M, Biondi EG, Bazzicalupo M *et al*. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med. 2011; 6: 11.

62. Chin CS, Alexander DH, Marks P *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 Jun; 10(6): 563-9.

63. Edwards DJ and Holt KE. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp. 2013; 3(1): 2.

64. Genome Project Standards in a New Era of Sequencing; Science 9 October 2009: vol. 326 no. 5950 236-23 [Internet] Figure at: www.sciencemag.org/content/326/5950/236/F1.expansion.html Accessed 14 May 2015.

65. Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001 Jul; 2(7): 493-503.

66. Kumar K, Desai V, Cheng L *et al*. AGeS: a software system for microbial genome sequence annotation. PloS ONE. 2011; 6(3): e17469.

67. Van Domselaar GH, Stothard P, Shrivastava S *et al*. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 2005 Jul 1; 33(Web Server issue): W455-9.

68. Pareja-Tobes P, Manrique M, Pareja-Tobes E *et al*. BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. PloS ONE. 2012; 7(11): e49239.

69. Besemer J and Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 2005 Jul 1; 33(Web Server issue): W451-4.

70. Markowitz VM, Chen IM, Chu K *et al*. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 2012 Jan; 40(Database issue): D123-9.

71. Markowitz VM, Chen IM, Palaniappan K *et al*. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res. 2010 Jan; 38(Database issue): D382-90.

72. Markowitz VM, Chen IM, Palaniappan K *et al*. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012 Jan; 40(Database issue): D115-22.

73. Cantarel BL, Korf I, Robb SM *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan; 18(1): 188-96.

74. Holt C and Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011 12: 491.

75. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul; 30(14):2068-9.

76. Aziz RK, Bartels D, Best AA *et al*. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008; 9: 75.

77. Wang S, Sundaram JP and Spiro D. VIGOR, an annotation program for small viral genomes. BMC Bioinformatics. 2010 11: 451.

78. Wang S, Sundaram JP and Stockwell TB. VIGOR extended to annotate genomes for additional 12 different viruses. Nucleic Acids Res. 2012; Jul; 40(Web Server issue): W186-92.

79. Chaudhuri RR, Loman NJ, Snyder LA *et al*. xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res. 2008 Jan; 36(Database issue): D543-6.

80. Chaudhuri RR and Pallen MJ. xBASE, a collection of online databases for bacterial comparative genomics. Nucleic Acids Res. 2006 Jan 1; 34(Database issue): D335-7.

81. Zankari E, Hasman H, Cosentino S *et al*. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012 Nov; 67(11): 2640-4.

82. Tang MW, Liu TF and Shafer RW. The HIVdb system for HIV-1 genotypic resistance interpretation. Intervirology. 2012; 55(2): 98-101.

83. Center For Genomic Epidemiology - Overview of Services [Internet]. Available from: http://cge.cbs.dtu.dk/services Accessed 14 May 2015.

84. Naamati G, Askenazi M and Linial M. ClanTox: a classifier of short animal toxins. Nucleic Acids Res. 2009 Jul; 37(Web Server issue): W363-8.

85. Naamati G, Askenazi M and Linial M. A predictor for toxin-like proteins exposes cell modulator candidates within viral genomes. Bioinformatics. 2010 Sep 15; 26(18): i482-8.

86. Zhou CE, Smith J, Lam M *et al*. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. 2007 Jan; 35(Database issue): D391-4.

87. Cosentino S, Voldby Larsen M, Moller Aarestrup F *et al*. PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PloS ONE. 2013 8(10): e77302.

88. Hasman H, Saputra D, Sicheritz-Ponten T *et al*. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol. 2014 Jan; 52(1): 139-46.

89. Underwood A and Green J. Call for a quality standard for sequence-based assays in clinical microbiology: necessity for quality assessment of sequences used in microbial identification and typing. J Clin Microbiol. 2011 Jan; 49(1): 23-6.

90. Tatusova T, Ciufo S, Fedorov B *et al*. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. 2014 Jan 1; 42(1): D553-9.

91. Rhee SY, Gonzales MJ, Kantor R *et al*. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. 2003 Jan 1; 31(1): 298-303.

92. Maiden MC, Bygraves JA, Feil E *et al*. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998 Mar 17; 95(6): 3140-5.

93. Thompson JD, Linard B, Lecompte O et al. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS ONE. 2011; 6:e18093.

94. Wyres KL, Conway TC, Garg S *et al*. WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare? Pathogens. 2014; 3(2):437-58.

95. Loman NJ, Constantinidou C, Christner M *et al*. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. JAMA. 2013 Apr 10; 309(14): 1502-10.

96. Rohde H, Qin J, Cui Y *et al*. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N Engl J Med. 2011 Aug 25; 365(8): 718-24.

97. Grard G, Fair JN, Lee D *et al*. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS Pathog. 2012 Sep 27; 8(9): e1002924.

98. Wilson MR, Naccache SN, Samayoa E *et al*. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med. 2014 Jun 6; 19; 370(25): 2408-17.

99. Fischer N, Rohde H, Indenbirken D *et al*. Rapid metagenomic diagnostics for suspected outbreak of severe pneumonia. Emerg Infect Dis. 2014 Jun; 20(6): 1072-5.

100. Stoesser N, Batty EM, Eyre DW *et al*. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother. 2013 Oct; 68(10):2234-44.

101. Chewapreecha C, Harris SR, Croucher NJ *et al*. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014 Mar; 46(3): 305-9.

102. Croucher NJ, Finkelstein JA, Pelton SI *et al*. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013 Jun; 45(6): 656-63.

103. Croucher NJ, Harris SR, Barquist L *et al*. A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog. 2012 8(6): e1002745.

104. Golubchik T, Brueggemann AB, Street T *et al*. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. Nat Genet. 2012 Jan 29; 44(3):352-5.

105. Gardy JL, Johnston JC, Ho Sui SJ *et al*. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. N Engl J Med. 2011; 364: 730-9.

106. Roetzer A, Diel R, Kohl T *et al*. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. PLoS Medicine. 2013 03/12; 10: e1001387.

107. Walker TM, Ip CL, Harrell RH *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013 Feb; 13(2):137-46.

108. Koser CU, Holden MT, Ellington MJ *et al*. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012 Jun 14; 366(24): 2267-75.

109. Snitkin E, Zelazny A, Thomas P *et al*. Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. Sci Transl Med. 2012 08/22; 4(148): 148ra16-ra16.

110. Walker JT, Jhutty A, Parks S *et al*. Investigation of healthcare-acquired infections associated with *Pseudomonas aeruginosa* biofilms in taps in neonatal units in Northern Ireland. J Hosp Infect. 2014 Jan; 86(1): 16-23.

111. Department of Health. HTM 04-01 - Addendum: *Pseudomonas aeruginosa* – advice for augmented care units. [Internet] 2013. Available from: www.gov.uk/government/uploads/system/uploads/attachment_data/file/140105/Health_Technical_Memorandum_04-01_Addendum.pdf Accessed 14 May 2015.

112. Quick J, Cumley N, Wearn CM *et al*. Seeking the source of *Pseudomonas aeruginosa* infections in a recently opened hospital: an observational study using whole-genome sequencing. BMJ Open. 2014 4(11): e006278.

113. Firth C and Lipkin WI. The genomics of emerging pathogens. Annu Rev Genomics Hum Genet. 2013; 14: 281-300.

114. Taubenberger JK and Morens DM. 1918 Influenza: the mother of all pandemics. Emerg Infect Dis. 2006 Jan; 12(1): 15-22.

115. Smith GJ, Vijaykrishna D, Bahl J *et al*. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009 Jun 25; 459(7250): 1122-5.

116. Gire SK, Goba A, Andersen KG *et al*. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014 Sep 12; 345(6202): 1369-72.

117. Azhar EI, El-Kafrawy SA, Farraj SA *et al*. Evidence for camel-to-human transmission of MERS coronavirus. N Engl J Med. 2014 Jun 26; 370(26): 2499-505.

118. Adney D, Doremalen N, Bushmaker T *et al*. Replication and Shedding of MERS-CoV in Upper Respiratory Tract of Inoculated Dromedary Camels. Emerging Infect Diseases. 2014; 20(12) [Internet]: dx.doi.org/10.3201/eid2012.141280 Accessed 14 May 2015.

119. Gardner MJ, Hall N, Fung E *et al*. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002 Oct 3; 419(6906): 498-511.

120. Neglected Tropical Diseases [Internet]. Available from: www.who.int/neglected_diseases/diseases/en Accessed 14 May 2015.

121. Khor CC, Chau TN, Pang J *et al*. Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. Nat Genet. 2011 Nov; 43(11): 1139-41.

122. Ge D, Fellay J, Thompson AJ et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 2009 Sept; 461:399-401.

123. Thomas DL, Thio CL, Martin MP et al. Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature. 2009 Oct; 461:798-801.

124. Paddon CJ and Keasling JD. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nat Rev Microbiol. 2014 May; 12(5): 355-67.

125. Ro DK, Paradise EM, Ouellet M *et al*. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature. 2006 Apr 13; 440(7086): 940-3.

126. [Internet] Available from: www.hicfund.org.uk/HICFundPortfolio/Theme5.aspx Accessed 14 May 2015.

127. Török M, Reuter S, Bryant J *et al*. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. J Clin Microbiol. 2013 Feb 1; 51(2): 611-4.

128. Bryant JM, Grogono DM, Greaves D *et al*. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. Lancet. 2013 May 4; 381(9877): 1551-60.

129. Savage CJ and Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. PloS ONE. 2009; 4(9): e7078.

130. Raza S , Luheshi L, Hall A. Sharing clinical genomic data for better diagnostics. PHG Foundation. Oct. 2014.

131. Fricke WF and Rasko DA. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. Nat Rev Genet. 2014 Jan; 15(1): 49-55.

132. HM Government. Open Data White Paper: Unleashing the Potential. 2012.

133. Department of Health. The power of information: Putting all of us in control of the health and care information we need. 2012.

134. CLIMB: Cloud Infrastructure for Microbial Bioinformatics [Internet]. Available from: www.climb.ac.uk/cloud Accessed 14 May 2015.

135. Afgan E, Baker D, Coraor N *et al*. Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics. 2010; 11 Suppl 12: S4.

136. Angiuoli SV, Matalka M, Gussman A *et al*. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics. 2011; 12:356.

137. Drummond MF, Sculpher, MJ, Torrance, GW, O'Brien, BJ, Stoddart, GL Methods for the economic evaluation of health care programmes. Third Edition. Oxford University Press; 2005.

138. Morris S, Devlin, N, Parkin, D, Spencer, A. Economic Analysis in Health Care. Second Edition ed. John Wiley & Sons Ltd; 2012.

139. Annual Report of the Chief Medical Officer 2011: Infections and the rise of antimicrobial resistance. Department of Health, 2013. [Internet] Available from: www.gov.uk/government/uploads/system/uploads/attachment_data/file/138331/CMO_Annual_Report_Volume_2_2011.pdf Accessed 14 May 2015.

140. Mishan EJ, Quah, E. Cost-Benefit Analysis. Fifth Edition ed. Routledge; 2007.

141. Smith RD, Keogh-Brown MR and Barnett T. Estimating the economic impact of pandemic influenza: An application of the computable general equilibrium model to the UK. Social Science & Medicine. 2011 7; 73(2): 235-44.

142. Smith R and Coast J. The true cost of antimicrobial resistance. BMJ. 2013; 346: f1493.

143. Review on Antimicrobial Resistance. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations. 2014. [Internet] Available from: http://amr-review.org/sites/default/files/Report-52.15.pdf Accessed 14 May 2015.

144. Bergholz TM, Moreno Switt AI and Wiedmann M. Omics approaches in food safety: fulfilling the promise? Trends in Microbiology. 2014 5; 22(5): 275-81.

145. Pillay D, Rambaut A, Geretti AM *et al*. HIV phylogenetics. BMJ: 2007 Sept 6; 335(7618): 460-1.

146. McGuire AL, Achenbaum LS, Whitney SN *et al.* Perspectives On Human Microbiome Research Ethics. J Empir Res Hum Res Ethics. 2012 Jul; 7(3):1-14.

Acknowledgements

We would like to gratefully acknowledge the contributions of the following people who have generously provided advice and information during the course of this project[*]:

Steering group members

Dr Tom Barlow **(from March 2014 to July 2014)**	Scientist, Health Science and Bioethics, Department of Health, London
Dr Paul Cosford	Director for Health Protection and Medical Director, Public Health England
Professor Saheer Gharbia	Head of Genomics Research, Public Health England, London
Dr Dorian Kennedy **(to March 2014)**	Deputy Director of Genetics & Scientific Development, Department of Health, London
Professor David Lomas	Dean, Faculty of Medical Sciences, University College London
Dr Christine McCartney	Director, Microbiology Services, Public Health England
Professor Julian Parkhill	Head of Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge
Mr Colin Pavelin **(from July 2014)**	Health Science Policy Lead, Department of Health, London
Professor Sharon Peacock **(to September 2014)**	Professor of Clinical Microbiology, Department of Medicine, University of Cambridge
Professor Tim Peto	Professor of Medicine, Co-Leader for the Infection Theme of the Oxford Biomedical Research Centre, Experimental Medicine Division NDM, John Radcliffe Hospital, Oxford
Dr Estée Török **(from September 2014)**	Senior Research Associate & Honorary Consultant Physician, University of Cambridge, Addenbrooke's Hospital, Cambridge

Consultees

Dr David Aanensen	Senior Lecturer in Bioinformatics, Imperial College, London
Mr Ion Agirrezabal	Senior Analyst, HTA and Health Economics Division, Costello Medical Consulting Ltd, Cambridge
Dr Cath Arnold[**]	Unit Head, Genomics services, Public Health England
Dr Nick Brown[**]	Consultant Medical Microbiologist, Public Health England, Clinical Microbiology & Public Health Laboratory, Addenbrooke's Hospital, Cambridge
Dr Kim Brugger	Head of Bioinformatics, Medical Genetics Laboratories, Addenbrooke's Hospital, Cambridge
Dr Meera Chand[**]	Consultant Microbiologist, Public Health England, Guy's & St Thomas' Hospitals, London
Dr Guy Cochrane	Head of European Nucleotide Archive, EMBL-EBI, Wellcome Trust Sanger Institute, Cambridge

Dr Tom Connor	Senior Lecturer, Cardiff University / CLIMB, Cardiff School of Biosciences, Cardiff
Mr Chris Dunne	General Manager – Informatics, The Health Informatics Service (THIS)
Ms Lucy Eddowes	Senior Analyst, Costello Medical Consulting Ltd, Cambridge
Dr Jonathan Edgeworth	Consultant Microbiologist & Medical Director Viapath, Guy's & St Thomas' Hospital, Department of Infectious Diseases, St Thomas' Hospital, London
Professor Martin Fisher[†]	Professor of HIV Medicine and Honorary Consultant, University of Sussex
Dr Tom Fowler	Director of Public Health, Genomics England, Queen Mary University of London
Mr Malcolm Goodwin	Service Delivery Manager, Infection Sciences, Viapath, King's College Hospital, London
Dr Jane Greatorex[**]	Senior Research Scientist, Public Health England (HPA), Clinical Microbiology, Addenbrooke's Hospital, Cambridge
Professor Jonathan Green[**]	Head of MS Informatics, Public Health England, London
Ms Mandy Griffin	Chief Operating Officer, The Health Informatics Service (THIS), NHS
Mr Chris Guest	Microbiology Operations Manager, Viapath, St Thomas' Hospital
Mrs Nicky Hall	IT and Data Manager, Flitwick GP Surgery
Dr Tim Harrison	Head, Respiratory & Vaccine Preventable Bacteria Reference Unit (RVPBRU), Public Health England, Microbiology Reference Services, London
Dr Carolyn Hemsley	Consultant in Infectious Diseases, and Microbiology and Clinical Lead, Department of Infectious Diseases, Guy's and St Thomas' Hospital Foundation Trust, London
Mrs Annette Jeanes	Director of Infection Prevention and Control, Consultant Nurse Infection Control, University College London NHS Foundation Trust
Professor Paul Kellam	Viral Genomics Group Leader & Senior Investigator, Wellcome Trust Sanger Institute, Cambridge
Dr Mike Kelsey	Consultant Microbiologist, Whittington Health, Whittington Hospital, London
Dr Zisis Kozlakidis	ICONIC Project Manager and Innovation Fellow, University College London Hospitals
Dr Nick Loman	Independent Research Fellow, Institute of Microbiology & Infection, University of Birmingham
Dr Howard Martin	Clinical Scientist, East Anglian Medical Genetics Service, Molecular Genetics Laboratories, Addenbrooke's Hospital, Cambridge
Dr Mike Millar	Consultant Medical Microbiologist, Barts Health NHS Trust
Professor Mark Pallen	Professor of Microbial Genomics, Division of Microbiology & Infection, Warwick Medical School, University of Warwick, Coventry
Professor Julian Parkhill[**]	Head of Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge
Dr John Paul[**]	Lead Public Health Microbiologist - SE Region, Public Health England, Royal Sussex County Hospital, Brighton

Dr Mark Reacher	Consultant Epidemiologist, Public Health England, Eastern Field Epidemiology Unit, IPH, Cambridge
Dr Sandra Reuter**	Bioinformatician and Postdoctoral Research Fellow, Wellcome Trust Genome Campus, Cambridge
Dr Geoff Smith	Vice President, Technology Development, Illumina, Essex
Dr Grace Smith**	Consultant Microbiologist, Midlands Regional Centre for Mycobacteriology and Heart of England NHS Foundation Trust, Birmingham
Dr Melvyn Smith	Principal Clinical Scientist, Infectious Diseases, Viapath, King's College Hospital, London
Professor Richard Smith**	Professor of Health System Economics, London School of Hygiene & Tropical Medicine, London
Mr Ash Sykes	Senior Software Developer, X-Lab
Dr Andrea Szendroi	Clinical Scientist, ViaPath, King's College Hospital, London
Dr Lucy Thomas	Consultant Epidemiologist, Public Health England, London
Dr Estée Török**	Senior Research Associate & Honorary Consultant Physician, University of Cambridge, Addenbrooke's Hospital, Cambridge
Dr Anthony Underwood**	Lead for Bioinformatics, Public Health England, London
Dr Tim Walker	Specialist Registrar in Infectious Diseases and Microbiology, University of Oxford
Ms Sally Wellsteed	Antimicrobial Resistance & Healthcare-Associated Infection Team Leader, Department of Health, London
Mr Alisdair Wotherspoon	Head of Science Delivery, Food Standards Agency, London
Dr Hongyi Zhang	Consultant Virologist & Deputy Clinical Services Director, Public Health Laboratory – Cambridge, Public Health England

*Affiliations were correct at time their contribution to the report was made

**We would like in particular to thank these people for discussing and fact checking various sections of this report. While their contribution has been invaluable, all responsibility for the final content of the report rests with the PHG Foundation authors.

†Deceased.

The PHG Foundation are grateful to visiting interns Philippa Crane (University of Sussex) and Aled Jones (Trainee Clinical Bioinformatician – Guy's and St Thomas' NHS Foundation Trust) for their research contributions to this project.

Abbreviation	Full name
A	Adenine
AIDS	Acquired Immune Deficiency Syndrome
AMR	Antimicrobial Resistance
APHA	Animal and Plant Health Agency
ART	Antiretroviral Therapy
AST	Antimicrobial Susceptibility Testing
BPs	Base Pairs
BBV	Blood Borne Virus
C	Cytosine
CCGs	Clinical Commissioning Groups
CCR5	C-C cehmokine receptor type 5
CMO	Chief Medical Officer
XDR	Extensively Drug Resistant
Contigs	Contiguous Sequences
CPA	Clinical Pathology Accreditation
CUHFT	Cambridge University Hospitals NHS Foundation Trust
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
EIDs	Emerging Infectious Diseases
ENA	European Nucleotide Archive
EMBL–EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
EQA	External Quality Assurance
FSA	Food Standards Agency
FDA	US Food and Drug Administration
G	Guanine
GDP	Gross Domestic Product
GN broth	Gram Negative broth
GMI	Global Microbial Identifier
HAIRs	Human and Animal Infections and Risk Surveillance
HCAIs	Healthcare-associated Infections
HEE	Health Education England
HICF	Healthcare Innovation Challenge Fund
HPA/PHE	Health Protection Agency/ Public Health England
HCV	Hepatitis C Virus
H	Hemagglutinin
H+	Hydrogen Ions
HICF	Healthcare Innovation Challenge Fund
HIV	Human Immunodeficiency Virus
Ion PGM	Ion Personal Genome Analyser
INSDC	International Nucleotide Sequence Database Collaboration
LIMS	Laboratory Information Management System
MDR	Multidrug-resistant

Abbreviation	Full name
MERS	Middle East Respiratory Syndrome
MLST	Multi Locus Sequence Typing
MLVA	Multi Locus Variable Number Tandem Repeat Analysis
MRSA	Methicillin-resistant *Staphylococcus Aureus*
MSA	Multiple Sequence Alignment
NT	Nucleotides
N	Neuraminidase
NEQAS	National External Quality Assurance Service
NGS	Next Generation Sequencing
NHS	National Health Service
NHSE	National Health Service England
PCR	Polymerase Chain Reaction
PFGE	Pulsed Field Gel Electrophoresis
PHE	Public Health England
QC	Quality Control
RCTs	Randomised Control Trials
RNA	Ribonucleic Acid
RNAse	Ribonuclease
RVPBRU	Respiratory and Vaccine Preventable Bacteria Reference Unit
SARS	Severe Acute Respiratory Syndrome
SLA	Service Level Agreement
SLST	Single Locus Sequence Typing
SMIs	Standards in Microbiological Investigation
spp.	Species
T	Thymine
TB	Tuberculosis
TDL	The Doctors Laboratory
THIS	The Health Informatics Service
UKAS	United Kingdom Accreditation Service
U	Uracil
WGS	Whole Genome Sequencing
WHO	World Health Organisation
XDR	Extensively Drug Resistant

phg
foundation
making science
work for health

About the PHG Foundation

The PHG Foundation is a pioneering independent think-tank with a special focus on genomics and other emerging health technologies that can provide more accurate and effective personalised medicine. Our mission is to make science work for health. Established in 1997 as the founding UK centre for public health genomics, we are now an acknowledged world leader in the effective and responsible translation and application of genomic technologies for health.

We create robust policy solutions to problems and barriers relating to implementation of science in health services, and provide knowledge, evidence and ideas to stimulate and direct well-informed discussion and debate on the potential and pitfalls of key biomedical developments, and to inform and educate stakeholders. We also provide expert research, analysis, health services planning and consultancy services for governments, health systems, and other non-profit organisations.

NHS
National Institute for Health Research

About the National Institute of Health Research

The National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre is a partnership between the University of Cambridge and Cambridge University Hospitals Foundation Trust. They receive substantial levels of funding from the NIHR to translate fundamental biomedical research into clinical research that benefits patients and improves healthcare provision.

978-1-907198-18-2

CAMBRIDGE UNIVERSITY
Health Partners
Knowledge-based healthcare

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
T +44 (0) 1223 761 900

www.phgfoundation.org

phg
foundation
making science
work for health