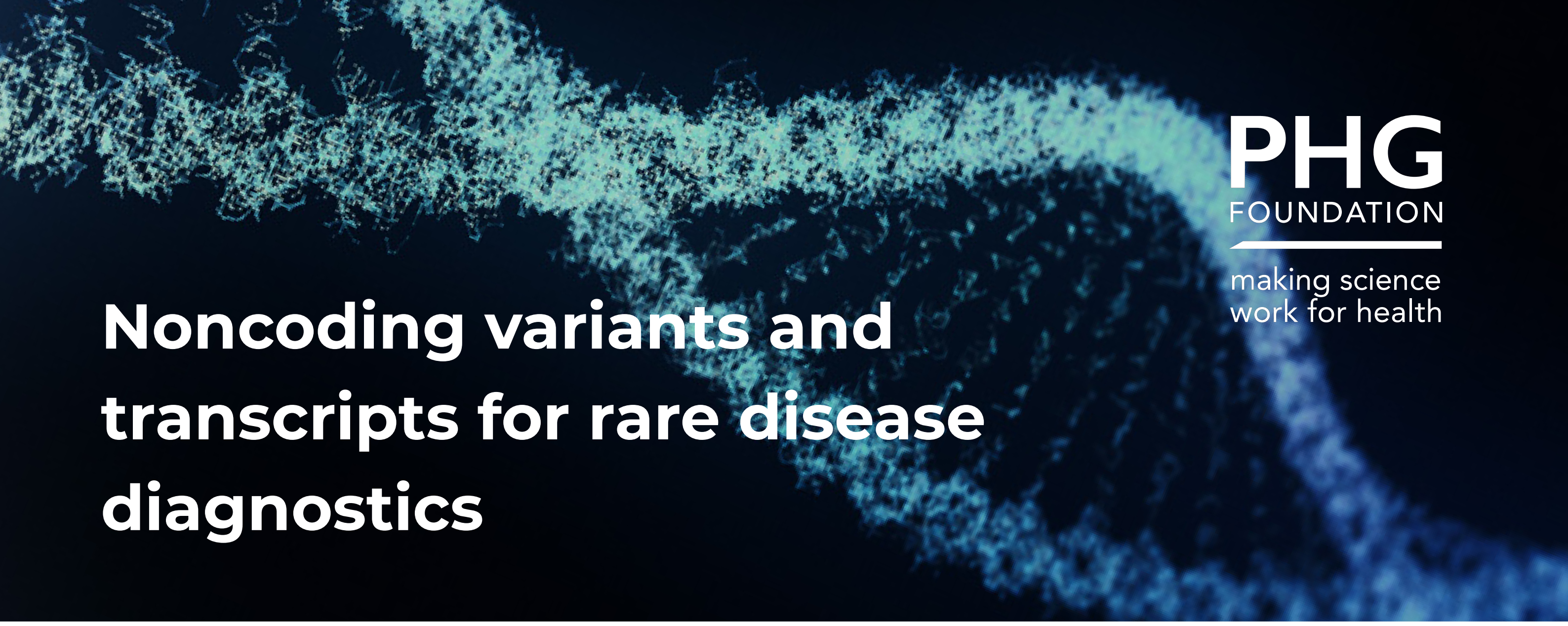

PHG
FOUNDATION
making science
work for health

Noncoding variants and transcripts for rare disease diagnostics

Summary

- Recent advances in genomic methods have allowed the study of the 98% of the genome previously considered as non-functional or 'noncoding'
- Integrating the analysis of noncoding regions holds promise for improving rare disease diagnosis; the combination of noncoding genomics and transcriptomics offers the greatest improvement in diagnostic yield
- Further research is needed to discover the functions of some of these 'noncoding' regions
- To use noncoding genome data in clinical practice, whole genome sequencing data and total RNA-Seq data are required, along with a scale-up of data infrastructure to support data analysis

Expanding rare disease diagnosis

The majority of rare diseases have a genetic basis. However, despite advances in genomic analysis methods, diagnosing rare diseases remains a challenge. One significant contributing factor is that clinical diagnostic pathways focus on well-researched protein-coding regions, which account for only 2% of the human genome [1]. While these targeted analyses are cost-effective and yield clinically interpretable results, expanding the scope to include the remaining 98% of the genome, known as 'noncoding' regions, holds the promise of addressing the diagnostic gap in rare diseases.

What is 'noncoding'?

Traditionally, the term 'noncoding' encompasses any genomic region or transcript that cannot code for a protein such as introns – genomic regions located between exons – and intergenic regions which are stretches of DNA between genes. Initially, it was believed that only protein-coding regions had a functional effect. However, progress in omic-based methods has facilitated the detection and analysis of low abundance transcripts and proteins from noncoding regions. These discoveries challenged conventional use of the term 'noncoding' and highlighted that a genomic region can be functional by exerting its influence solely through its transcripts.

WYNG
FOUNDATION

The views set out in this briefing note do not necessarily represent those of the WYNG Foundation

Noncoding regions are often referred to as the 'dark genome' to underscore how limited current understanding of the functions of many of these regions are. While certain noncoding regions may appear to lack functionality, others might be incorrectly labelled as being non-functional due to the limitations of methods currently used to study them. Consequently, the term 'noncoding' does not imply non-functionality, nor should it impede the investigation of noncoding regions in disease contexts.

Why analyse noncoding regions?

Analysing noncoding variants significantly expands the search space for identifying disease-associated genomic variants. Moreover, the temporal, tissue, cell type or disease-specificity of noncoding transcripts makes them more useful as biomarkers compared to the widespread but relatively consistent expression of protein-coding transcripts [2], [3]. For example, blood expression levels of noncoding transcripts H19 and RP11-445H22.5 have proven to be more reliable for breast cancer diagnosis than conventional glycoprotein markers [2].

Work on proteins produced by noncoding regions is still predominantly research-oriented and in its infancy. In this briefing note, we focus on evaluating noncoding-based diagnostics at the variant-level (information obtained from DNA) and transcript-level (information obtained from RNA). Additionally, it explores a multiomics approach that combines both these aspects.

Challenges in noncoding analyses

The most prominent challenges in noncoding variant and transcript analyses include:

- **functional interpretation of the vast number of disease-associated noncoding variants and transcripts that can be identified** - as high as 4.4 million variants
- **non-actionable findings** resulting from lack of functional understanding of disease-associated noncoding variants, including variants of uncertain significance, and noncoding transcripts. This issue is further exacerbated when the genomes from non-European ancestry populations are analysed against standard reference genomes
- **limited availability of patient samples** with the same rare disease hinders comprehensive transcriptomic analyses

Strategies for noncoding analyses

Various strategies can be used to support noncoding variant and transcript analyses:

- DNA/RNA sample collection: as noncoding transcript expression is tissue specific, RNA may be collected from disease-associated tissues. To support multiomic analysis, DNA and RNA extraction can be done simultaneously from tissue samples.
- DNA/RNA sequencing: to maximise variant detection, whole genome sequencing can be used. Low-pass whole genome sequencing could also be an acceptable clinical approach to control cost. To maximise noncoding transcript detection, total RNA-Seq (without selecting for poly-A tails that noncoding transcripts may lack) may be used. RNA Capture long-read sequencing can also be used for targeted detection of noncoding transcripts with low abundance.
- identifying disease-associated variants and transcripts: up-to-date reference genomes with both protein-coding and noncoding annotations should be used for noncoding analyses. Additionally, cloud-based servers may be considered to support data analysis.

Specialised analysis tools for noncoding regions may be used to perform:

- variant filtering and prioritisation e.g., RegulomeDB, DeepSEA, GWAVA
- low-sample size transcriptomic analyses e.g., OUTRIDER, FRASER, LeafCutterMD
- prediction of transcript function from its sequence using computational tools such as 'InterProSCAN'

Restricting manual variant interpretation to variants in regulatory regions, evolutionarily conserved regions, or disease-associated genes – as these are likely to give definitive results – and assessing multiple variants simultaneously for regulatory functions using approaches such as massively parallel reporter assays can help elucidate functions of noncoding variants. Databases that contain specific information on noncoding variants and transcripts, such as ncVarDB, 3DSNP, NONCODE, IncRNAdb and FANTOM5, or existing databases integrated with noncoding information, such as gnomAD, ClinGen, and PanelApp, could also be used to support noncoding analyses.

Diagnostic potential of noncoding analysis for rare diseases

Disease-associated noncoding variants and transcripts have been identified in rare diseases such as developmental disorders, hyperinsulinism, systemic sclerosis and inherited retinal degeneration. Across a range of rare disease cohorts, the combined analysis of variants and transcripts, which included noncoding variants and transcripts, was found to improve the diagnostic yield by up to 36% [3]. Notably, the identification of noncoding transcripts implicated in the pathophysiology of systemic sclerosis has contributed to the development of a new drug, remlarsen, which mimics the function of a noncoding transcript that is dysfunctional in the disease [4].

Initiatives in noncoding-based diagnostics for rare diseases

The emergence of several initiatives illustrates a growing interest in noncoding-based diagnostics. In the USA, the Noncoding variants program (NoVa) was set up to support the development of new noncoding variant-based risk prediction models. An EU-funded project 'Solve-RD' was set up as a collaborative network to share rare disease patient genomic data and medical expertise. Reanalysis of undiagnosed rare disease cases relying on these collaborative networks led to 511 new diagnoses – an improvement in diagnostic yield by 8.5%.

Large-scale projects with rare disease patient data such as EU's 1+ million genomes initiative and UK10K can support discovery research on noncoding variants and transcripts. While whole genome sequencing is not widely used in China, an Expert Consensus in 2018 published recommendations for new whole genome sequencing protocols and standards to support rare disease diagnosis and treatment in children. This could increase the appetite for noncoding analysis in clinical diagnostics.

How close is it to clinical implementation?

Currently, clinical provision for noncoding-based diagnostics is rare disease specific. For example, the UK's 100,000 Genomes project rare disease pilot study showed a 13% increase in rare disease diagnosis using noncoding variants and, consequently, the NHS test directory offered whole genome sequencing as first-line diagnosis for certain rare diseases [5]. However, in most cases, noncoding-based diagnostics is a second-line option, complementing routine protein-coding-based diagnostics.

What could help clinical uptake?

Measures needed to support clinical uptake of noncoding variant analysis are:

- standardised clinical data collection practices to support multiomic patient data collection
- standardised data analysis pipelines and skilled bioinformaticians trained in noncoding analysis to process variant and transcriptomic data
- guidelines to interpret noncoding variants and incorporate functional analysis of noncoding transcripts in their interpretation
- preparing clinical data processing and storage systems to handle the increased data load from noncoding analysis

Appropriate safeguards around genomic data access that would not hinder patient care and collaborative research must be implemented. Current research indicates that classifying the human genome into protein-coding and noncoding relies on ambiguous boundaries requiring continued updates to genome annotations. How this changing information and additional variants of uncertain significance are communicated to patients and research participants while minimising confusion or stress must be addressed.

Conclusions

The current landscape supports increased use of noncoding genome-based diagnoses in clinical care to complement traditional protein-coding region-based approaches. To ensure that a complete and up-to-date knowledge base is readily available for clinical diagnosis, collaborations between research bodies and clinicians are essential. Moreover, capturing and storing data for potential future re-analysis, can help to ensure that valuable insights from noncoding regions are not missed.

References

1. J. M. Ellingford et al., ‘Recommendations for clinical interpretation of variants found in non-coding regions of the genome’, Genome Med., vol. 14, no. 1, p. 73, Jul. 2022, doi: 10.1186/s13073-022-01073-3.
2. C. Badowski, B. He, and L. X. Garmire, ‘Blood-derived lncRNAs as biomarkers for cancer diagnosis: the Good, the Bad and the Beauty’, Npj Precis. Oncol., vol. 6, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41698-022-00283-7.
3. S. B. Montgomery, J. A. Bernstein, and M. T. Wheeler, ‘Toward transcriptomics as a primary tool for rare disease investigation’, Cold Spring Harb. Mol. Case Stud., vol. 8, no. 2, p. a006198, Feb. 2022, doi: 10.1101/mcs.a006198.
4. J. Henderson, J. Distler, and S. O’Reilly, ‘The Role of Epigenetic Modifications in Systemic Sclerosis: A Druggable Target’, Trends Mol. Med., vol. 25, no. 5, pp. 395–411, May 2019, doi: 10.1016/j.molmed.2019.02.001.
5. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report’, N. Engl. J. Med., vol. 385, no. 20, pp. 1868–1880, Nov. 2021, doi: 10.1056/NEJMoa2035790

Author: Dr Chaitanya Erady

Published: November 2023

UNIVERSITY OF
CAMBRIDGE