OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of Data Mining in Genomics & Proteomics

Perspective

Ongoing Challenges to Finalize the Catalog of Human Genes and Transcripts

Zhougeng Liu*

Department of Biotechnology, College of Medicine, National Cheng Kung University, Tainan, Taiwan

DESCRIPTION

One of the fundamental challenges in cataloging human genes lies in defining what exactly constitutes a "gene." In the early days of molecular biology, a gene was considered a simple unit of heredity, encoding a specific protein. However, advances in genomics have revealed that genes are not as straightforward as once thought. Genes can produce multiple transcripts through mechanisms such as alternative splicing, alternative promoters and polyadenylation sites, leading to a diverse array of mRNA and protein isoforms. The question of whether these isoforms should be considered part of the same gene or distinct entities adds complexity to gene classification.

Additionally, there are many non-coding genes that do not produce proteins but generate functional RNA molecules, such as long non-coding RNAs (lncRNAs), microRNAs (miRNAs) and other small RNAs. These RNAs play an important regulatory roles in cellular processes, but defining their boundaries and functional significance is difficult. The expanding scope of what is considered a gene, including pseudogenes, gene fragments and regulatory elements, complicates the task of creating a comprehensive gene catalog.

The complexity of alternative splicing and transcription

Alternative splicing is a main mechanism through which a single gene can produce multiple mRNA transcripts, each encoding different protein isoforms. It is estimated that more than 90% of human genes undergo alternative splicing, which significantly expands the diversity of the proteome. However, determining the full extent of alternative splicing events and accurately annotating them remains a major challenge.

The effective nature of transcription adds further complexity. Genes can have multiple transcription start sites and transcripts can be processed in a tissue-specific or condition-specific manner. As a result, a single gene may give rise to a multitude of transcripts, each with different functions or regulatory roles. Capturing the diversity of these transcripts across different cell types, tissues and developmental stages is an enormous challenge, as transcript expression can vary widely depending on context.

Furthermore, advances in technologies like single-cell RNA sequencing (scRNA-seq) have revealed an even greater level of transcriptomic complexity than was previously appreciated. Each individual cell can express a unique set of transcripts and this cellular heterogeneity adds another layer of difficulty in finalizing a comprehensive catalog of human transcripts.

Incomplete understanding of non-coding RNAs

While the identification of protein-coding genes has been relatively successful, the cataloging of non-coding RNAs (ncRNAs) remains incomplete and challenging. Non-coding RNAs, which do not translate into proteins, are now known to play critical roles in gene regulation, chromatin organization and cellular signaling. However, many ncRNAs are expressed at low levels or in specific tissues, making them difficult to detect and characterize.

Long non-coding RNAs (lncRNAs), in particular, represent a large and diverse class of ncRNAs, many of which remain poorly understood. Unlike protein-coding genes, lncRNAs often have less well-defined functional roles and their expression can be highly tissue-specific, further complicating efforts to annotate them. The relatively low abundance and weak evolutionary conservation of many lncRNAs also make them harder to identify and validate using traditional genome annotation methods.

MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are smaller non-coding RNAs that have well-established roles in gene regulation, but their biogenesis and target specificities are complex. Furthermore, novel classes of small RNAs are continually being discovered, suggesting that the catalog of non-coding RNAs is far from complete.

Correspondence to: Zhougeng Liu, Department of Biotechnology, College of Medicine, National Cheng Kung University, Tainan, Taiwan, E-mail: liugengz.biotech@126.com

Received: 24-Aug-2024, Manuscript No. JDMGP-24-27278; **Editor assigned:** 26-Aug-2024, PreQC No. JDMGP-24-27278 (PQ); **Reviewed:** 09-Sep-2024, QC No. JDMGP-24-27278; **Revised:** 16-Sep-2024, Manuscript No. JDMGP-24-27278 (R); **Published:** 23-Sep-2024, DOI: 10.4172/2153-0602.24.15.354

Citation: Liu Z (2024). Ongoing Challenges to Finalize the Catalog of Human Genes and Transcripts. J Data Mining Genomics Proteomics. 15:354.

Technological limitations and data integration

The technological advancements in sequencing, especially Next-Generation Sequencing (NGS) and third-generation sequencing platforms like nanopore and Single-Molecule Real-Time (SMRT) sequencing, have significantly improved our ability to detect and sequence genes and transcripts. However, these technologies also present challenges that affect the completeness of gene catalogs.

Short-read sequencing: While short-read sequencing technologies, like Illumina, have been the foundation of many genomic studies, they often struggle to resolve complex genomic regions and cannot fully capture long-range transcript isoforms. As a result, certain alternative splicing events or transcript variants may be missed.

Long-read sequencing: Long-read sequencing technologies, such as those offered by Oxford Nanopore and Pacific Biosciences, have improved the ability to capture full-length transcripts and resolve complex genomic regions. However, these technologies are still relatively costly and can have higher error rates compared to short-read sequencing, which complicates data interpretation and annotation.

Data integration challenges: Integrating different types of sequencing data (DNA, RNA, epigenomic) and technologies is a major challenge. The need to coordinate data from different platforms, experimental conditions and biological samples requires complicated computational tools and bioinformatics pipelines. Ensuring consistency and accuracy in gene and transcript annotation across different datasets is important for producing a comprehensive and reliable catalog.

Evolutionary and population variability

Human genetic variation adds another layer of complexity to the task of cataloging genes and transcripts. Different individuals carry unique sets of genetic variations, including Single Nucleotide Polymorphisms (SNPs), Copy Number Variations (CNVs) and structural variants, which can affect gene structure and expression. The identification of gene variants that are specific to certain populations or individuals further complicates the task of defining a universal gene catalog.

Additionally, many genes exhibit evolutionary variability and their homologs in other species may have different structures or functions. Comparative genomics, which seeks to identify conserved genes across species, provides important insights but also highlights the challenges of determining which genetic elements are functionally important in humans.

Functional annotation and validation

Cataloging genes and transcripts is not only about identifying sequences but also understanding their functional significance. Functional annotation involves determining what a gene or transcript does, its role in cellular processes and how it contributes to health and disease. However, many genes and transcripts identified through sequencing remain poorly characterized and their functions are unknown. Experimental validation through techniques such as CRISPR gene editing, RNA interference (RNAi) and proteomics is essential for confirming the biological roles of newly identified genes and transcripts.