

Concorso pubblico, per titoli ed esami, per la copertura di n. 2 posti di Tecnologo di I livello (ex posizione economica EP1), con rapporto di lavoro subordinato a tempo determinato, per la durata di 18 mesi, con regime di impegno a tempo pieno, per le esigenze del Dipartimento di Bioscienze, Biotecnologie e Ambiente, nell'ambito del progetto "Genoma mEdiciNa pERsonalizzatA – GENERA" presso l'Università degli Studi di Bari Aldo Moro (CUP H93C22000500001), indetto con DDG n. 1005 del giorno 19/06/2025

Domande elaborate dalla commissione esaminatrice per l'espletamento della prova orale.

### **DOMANDE BUSTA N. 1**

Il/La candidato/a illustri:

- le tecnologie di seconda generazione e le loro principali applicazioni biotecnologiche
- la struttura di un foglio di calcolo Excel
- il ruolo e la composizione del Senato Accademico

Il/La candidato/a legga e traduca l'*abstract* del manoscritto "Transcriptomics in the era of long-read sequencing" (Nat. Methods 2025 doi: 10.1038/s41576-025-00828-z), di cui si allega la prima pagina.

### **DOMANDE BUSTA N. 2**

Il/La candidato/a illustri:

- le tecnologie di terza generazione e le loro principali applicazioni biotecnologiche
- la modalità per inserire una formula in una cella di Excel
- il ruolo e la composizione del Consiglio di Amministrazione

Il/La candidato/a legga e traduca l'*abstract* del manoscritto "Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment" (Nat. Rev. Genet. 2025 doi: 10.1038/s41592-025-02631-4), di cui si allega la prima pagina.

### **DOMANDE BUSTA N. 3**

Il/La candidato/a illustri:

- i benefici dell'utilizzo delle piattaforme terza generazione per studi genomici
- il significato di "formattazione" in un documento Office
- i principali compiti del Rettore

Il/La candidato/a legga e traduca l'*abstract* del manoscritto "Comprehensive genome analysis and variant detection at scale using DRAGEN" (Nat. Biotech. 2025 doi: 10.1038/s41587-024-02382-1), di cui si allega la prima pagina.

### **DOMANDE BUSTA N. 4**

Il/La candidato/a illustri:

- i principali vantaggi del deep sequencing in ambito biotecnologico e ambientale
- la modalità di inserimento di immagini in una presentazione PowerPoint
- il ruolo e la composizione del Consiglio di Dipartimento

Il/La candidato/a legga e traduca l'*abstract* del manoscritto "Notable challenges posed by long-read sequencing for the study of transcriptional diversity and genome annotation" (Genome Res. 2025 doi: 10.1101/gr.279865.124), di cui si allega la prima pagina.

Di seguito il documento in lingua inglese.

Il Segretario della Commissione  
Sig. Fabio CORPOSANTO

## Perspective

# Notable challenges posed by long-read sequencing for the study of transcriptional diversity and genome annotation

Carolina Monzó,<sup>1</sup> Adam Frankish,<sup>2</sup> and Ana Conesa<sup>1</sup>

<sup>1</sup>Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC), Paterna 46980, Spain;

<sup>2</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus Hinxton, Cambridge CB10 1SA, United Kingdom

Long-read sequencing (LRS) technologies have revolutionized transcriptomic research by enabling the comprehensive sequencing of full-length transcripts. Using these technologies, researchers have reported tens of thousands of novel transcripts, even in well-annotated genomes, while developing new algorithms and experimental approaches to handle the noisy data. The Long-read RNA-seq Genome Annotation Assessment Project community effort benchmarked LRS methods in transcriptomics and validated many novel, lowly expressed, often times sample-specific transcripts identified by long reads. These molecules represent deviations of the major transcriptional program that were overlooked by short-read sequencing methods but are now captured by the full-length, single-molecule approach. This Perspective discusses the challenges and opportunities associated with LRS' capacity to unravel this fraction of the transcriptome, in terms of both transcriptome biology and genome annotation. For transcriptome biology, we need to develop novel experimental and computational methods to effectively differentiate technology errors from rare but real molecules. For genome annotation, we must agree on the strategy to capture molecular variability while still defining reference annotations that are useful for the genomics community.

[Supplemental material is available for this article.]

Long-read sequencing (LRS) technologies, such as those developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have revolutionized genomic and transcriptomic research. Their ability to generate very long reads has enabled significant advancements, including complete sequencing of human chromosomes (Nurk et al. 2022) and full-length sequencing of single-molecule transcripts spanning kilobases (Sharon et al. 2013; Weirather et al. 2017; Soneson et al. 2019). This unprecedented capability earned LRS recognition by Nature Methods as the Method of the Year in 2022, highlighting its transformative impact on both fields (Marx 2023). One of the most significant contributions of long-read methods to the study of transcription is their capacity to uncover alternative isoforms with a confidence not present in short-read methods, which has led to the discovery of tens of thousands of novel transcripts even in well-annotated organisms (Roach et al. 2020; Glinos et al. 2022; Veiga et al. 2022; Zhang et al. 2022), and represents a data source of great value for the de novo annotation of the Earth BioGenome Project (Lawniczak et al. 2022). Despite its strengths, LRS presents several shortcomings. The quality of long-read RNA sequencing (lrrNA-seq) can be compromised by factors such as RNA degradation, biases introduced during library preparation, sequencing errors, and inaccurate bioinformatic processing during mapping, transcript assembly, and quantification, which may lead to the incorrect identification of transcript models, i.e., computational representations of transcripts depicting their transcription start and termination sites (TSS and TTS) and intron composition (Amarasinghe

et al. 2020; Marx 2023). Most lrrNA-seq experiments rely on cDNA libraries, as they provide high sequencing throughput and accuracy. However, reverse transcription may introduce errors driven by specific sequences present in the RNA's primary sequence. These sequences can promote single-nucleotide errors and mispriming, resulting in faulty cDNA molecules (technical artifacts) that inaccurately represent structural variations (Verwilt et al. 2023). The ONT direct RNA-seq method can potentially overcome these issues while also identifying RNA modifications in the native molecule. However, sequencing throughput from current direct RNA protocols is still relatively low compared to cDNA-based protocols (~20 M reads in direct RNA protocols [Oxford Nanopore Technologies 2019] vs. ~130 M reads in cDNA-based protocols [Aguzzoli Heberle et al. 2024]), which compromises transcript identification. Despite these shortcomings, direct RNA holds great potential for improving transcript identification in the future.

A significant challenge in the analysis of lrrNA-seq data is the accurate identification of novel transcripts while effectively distinguishing them from artifacts introduced by the technology. To address this, various software tools have been created for reconstructing transcript models from LRS, and recently the technology has been subjected to rigorous benchmarking (Križanovic et al. 2018; Soneson et al. 2019; Kuo et al. 2020; Dong et al. 2023; Su et al. 2024; Pardo-Palacios et al. 2024b). The most comprehensive study to evaluate lrrNA-seq methods to date is the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP), a community effort aimed at systematically evaluating library preparation, sequencing platforms, and analysis tools for the identification

**Corresponding author:** [ana.conesa@csic.es](mailto:ana.conesa@csic.es)

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.279865.124>. Freely available online through the *Genome Research* Open Access option.

© 2025 Monzó et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

# Transcriptomics in the era of long-read sequencing

Carolina Monzó <sup>1,2</sup>✉, Tianyuan Liu <sup>1,2</sup> & Ana Conesa <sup>1</sup>✉

## Abstract

Transcriptome sequencing revolutionized the analysis of gene expression, providing an unbiased approach to gene detection and quantification that enabled the discovery of novel isoforms, alternative splicing events and fusion transcripts. However, although short-read sequencing technologies have surpassed the limited dynamic range of previous technologies such as microarrays, they have limitations, for example, in resolving full-length transcripts and complex isoforms. Over the past 5 years, long-read sequencing technologies have matured considerably, with improvements in instrumentation and analytical methods, enabling their application to RNA sequencing (RNA-seq). Benchmarking studies are beginning to identify the strengths and limitations of long-read RNA-seq, although there remains a need for comprehensive resources to guide newcomers through the intricacies of this approach. In this Review, we provide a comprehensive overview of the long-read RNA-seq workflow, from library preparation and sequencing challenges to core data processing, downstream analyses and emerging developments. We present an extensive inventory of experimental and analytical methods and discuss current challenges and prospects.

## Sections

Introduction

Experimental and sequencing design

Core data processing of lrRNA-seq data

Downstream computational analysis

Emerging lrRNA-seq applications

Conclusions and future perspectives

<sup>1</sup>Institute for Integrative Systems Biology, Spanish National Research Council, Paterna, Valencia, Spain. <sup>2</sup>These authors contributed equally: Carolina Monzó, Tianyuan Liu. ✉e-mail: [carolina.monzo@csic.es](mailto:carolina.monzo@csic.es); [ana.conesa@csic.es](mailto:ana.conesa@csic.es)

# Comprehensive genome analysis and variant detection at scale using DRAGEN

Received: 24 December 2023

Accepted: 8 August 2024

Published online: 25 October 2024

 Check for updates

Sairam Behera<sup>1,5</sup>, Severine Catreux<sup>2,5</sup>✉, Massimiliano Rossi<sup>2,5</sup>, Sean Truong<sup>2</sup>, Zhuoyi Huang<sup>2</sup>, Michael Rühle<sup>2</sup>, Arun Visvanath<sup>2</sup>, Gavin Parnaby<sup>2</sup>, Cooper Roddey<sup>2</sup>, Vitor Onuchic<sup>2</sup>, Andrea Finocchio<sup>2</sup>, Daniel L. Cameron<sup>2</sup>, Adam English<sup>1</sup>, Shyamal Mehtalia<sup>2</sup>, James Han<sup>2,6</sup>✉, Rami Mehio<sup>2,6</sup>✉ & Fritz J. Sedlazeck<sup>1,3,4,6</sup>✉

Research and medical genomics require comprehensive, scalable methods for the discovery of novel disease targets, evolutionary drivers and genetic markers with clinical significance. This necessitates a framework to identify all types of variants independent of their size or location. Here we present DRAGEN, which uses multigenome mapping with pangenome references, hardware acceleration and machine learning-based variant detection to provide insights into individual genomes, with ~30 min of computation time from raw reads to variant detection. DRAGEN outperforms current state-of-the-art methods in speed and accuracy across all variant types (single-nucleotide variations, insertions or deletions, short tandem repeats, structural variations and copy number variations) and incorporates specialized methods for analysis of medically relevant genes. We demonstrate the performance of DRAGEN across 3,202 whole-genome sequencing datasets by generating fully genotyped multisample variant call format files and demonstrate its scalability, accuracy and innovation to further advance the integration of comprehensive genomics. Overall, DRAGEN marks a major milestone in sequencing data analysis and will provide insights across various diseases, including Mendelian and rare diseases, with a highly comprehensive and scalable platform.

Over the last decade, the advent of genomic sequencing as a common methodology in genomics, genetics and medical applications has enabled multiple discoveries and insights for diseases, population diversity, evolutionary mechanisms and personalized medicine strategies<sup>1–4</sup>. This was made possible in large part due to improvements in next-generation sequencing (that is, Illumina) in terms of costs, high data quality and scalability<sup>1</sup>. Highly accurate methods for the detection of single-nucleotide variations (SNVs) and smaller (<50 bp) insertions or deletions (indels) have been at the forefront of variant detection

and interpretation. Despite the amount of attention that SNVs have garnered, they are not the only variant type that differentiates two genomes<sup>5,6</sup>. Recently, an increasing number of studies incorporate structural variation (SV)<sup>7–9</sup> into their analysis. SVs are often defined to be 50 bp or larger and lead to deletions, insertions, amplifications or rearrangements of a genome<sup>7</sup>. Copy number variation (CNV) is another genomic variation that arises from deletions (loss of copies) or duplications (gain of copies) of a specific DNA segment<sup>7</sup>. Another understudied variant type is short tandem repeat (STR) expansions that are mainly

<sup>1</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. <sup>2</sup>Illumina, Inc., San Diego, CA, USA. <sup>3</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. <sup>4</sup>Department of Computer Science, Rice University, Houston, TX, USA. <sup>5</sup>These authors contributed equally: Sairam Behera, Severine Catreux, Massimiliano Rossi. <sup>6</sup>These authors jointly supervised this work: James Han, Rami Mehio, Fritz J. Sedlazeck. ✉e-mail: [scatreux@illumina.com](mailto:scatreux@illumina.com); [jhan6@illumina.com](mailto:jhan6@illumina.com); [rmehio@illumina.com](mailto:rmehio@illumina.com); [fritz.sedlazeck@bcm.edu](mailto:fritz.sedlazeck@bcm.edu)

# Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment

Received: 15 March 2024

Accepted: 16 February 2025

Published online: 28 March 2025

 Check for updates

Sam Kovaka<sup>1</sup>✉, Paul W. Hook<sup>2</sup>, Katharine M. Jenike<sup>3</sup>, Vikram Shivakumar<sup>1</sup>, Luke B. Morina<sup>2</sup>, Roham Razaghi<sup>2</sup>, Winston Timp<sup>2,3</sup> & Michael C. Schatz<sup>1,3,4</sup>

Nanopore signal analysis enables detection of nucleotide modifications from native DNA and RNA sequencing, providing both accurate genetic or transcriptomic and epigenetic information without additional library preparation. At present, only a limited set of modifications can be directly basecalled (for example, 5-methylcytosine), while most others require exploratory methods that often begin with alignment of nanopore signal to a nucleotide reference. We present Uncalled4, a toolkit for nanopore signal alignment, analysis and visualization. Uncalled4 features an efficient banded signal alignment algorithm, BAM signal alignment file format, statistics for comparing signal alignment methods and a reproducible de novo training method for *k*-mer-based pore models, revealing potential errors in Oxford Nanopore Technologies' state-of-the-art DNA model. We apply Uncalled4 to RNA 6-methyladenine (m6A) detection in seven human cell lines, identifying 26% more modifications than Nanopolish using m6Anet, including in several genes where m6A has known implications in cancer. Uncalled4 is available open source at [github.com/skovaka/uncalled4](https://github.com/skovaka/uncalled4).

Long-read single-molecule sequencers from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) have increasing utility in generating complete genomes and transcriptomes by improving resolution of complex DNA and RNA sequences<sup>1–3</sup>. These sequencers can also detect nucleotide modifications without any specialized library preparation, enabling genome-wide epigenetic profiling including within highly repetitive regions that could not be accurately aligned to with short reads<sup>4</sup>. Nanopore sequencing is unique in not relying on sequencing-by-synthesis, instead measuring electric current that varies over time as nucleotides pass through a pore. While many analyses only use the basecalled sequence, inclusion of the electric current can improve fidelity in several applications, including error correction<sup>5,6</sup>, real-time targeted sequencing<sup>7,8</sup> and nucleotide modification detection<sup>9</sup>. Furthermore, ONT is currently the only commercially available platform for directly sequencing RNA without generation

of complementary DNA (cDNA), enabling detection of epitranscriptomic modifications. Over 150 known RNA modifications are known to exist, although only a few can be comprehensively detected at the single-nucleotide level, with varying accuracy<sup>10</sup>.

Early nanopore sequencers exhibited a high error rate, which could be improved via signal-based polishing<sup>5</sup> or advanced basecalling algorithms. However, a combination of improvements to sequencing chemistry and computational methods have decreased the average ONT DNA sequencing error rate to nearly 1%, making signal-based polishing largely unnecessary for DNA. This was achieved, in part, by a recent major DNA chemistry update to the r10.4.1 pore, which features two 'reader heads' rather than the one present in the previous standard, r9.4.1 (Fig. 1a). Direct RNA accuracy has lagged behind, where signal-based polishing can still improve splice site identification<sup>6</sup>. On the software side, modern basecallers use neural networks trained

<sup>1</sup>Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. <sup>2</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. <sup>3</sup>Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA. <sup>4</sup>Department of Biology, Johns Hopkins University, Baltimore, MD, USA. ✉e-mail: [skovaka1@jhu.edu](mailto:skovaka1@jhu.edu)