PRODUCED BY: Front Line Genomics

THE SEQUENCING BUYER'S GUIDE

6TH EDITION

SPONSORED BY

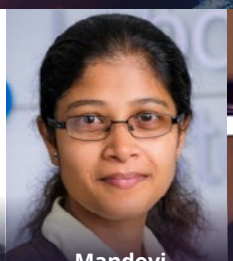
Agilent biomodal illumina INTEGRA

Novogene QIAGEN Source Genomics Source BioScience Group TECAN. TWIST BIOSCIENCE

CONTRIBUTORS

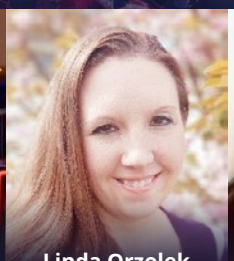
Alex Couto Alves
Head of the Bioinformatics Core Facility & Lecturer in Bioinformatics and Statistical Genomics
University of Surrey

David Baker
Head of Sequencing
The Quadram Institute



Mandovi Chatterjee
Director, Single-cell Core
Harvard Medical School

Josh Fienman
Scientist, Genomics (NGS Technology Center)
Pfizer

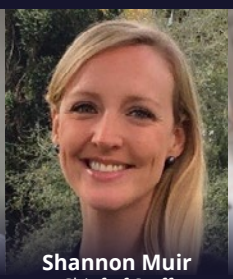


Linda Orzolek
Director Single Cell & Transcriptomics Core
John Hopkins University

Devjanee Swain Lenz
Director, Sequencing and Genome Technologies
Duke Center for Genomic and Computational Biology

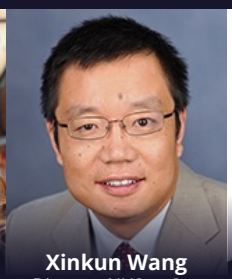
Nancy Cox
Professor of Medicine and Director of the Division of Genetic Medicine
Vanderbilt University

Howard Mcleod
Professor of Medicine and Biology & Director, Centre for Precision Medicine
Utah Tech University

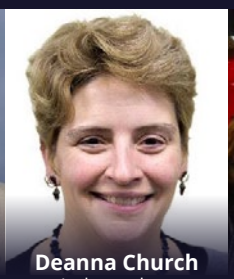


Shannon Muir
Chief of Staff
Latino Cancer Institute

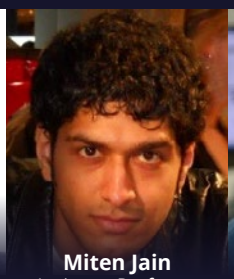
Marylyn Ritchie
Director of Institute for Biomedical Informatics
University of Pennsylvania



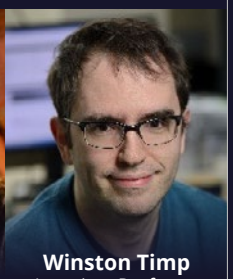
Xinkun Wang
Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology
Northwestern University



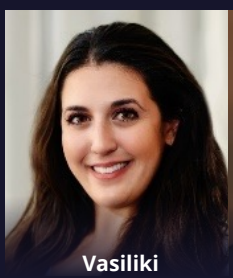
Deanna Church
Independent Consultant and Executive-in-Residence
Dmchurch Bio, llc and General Inception



Miten Jain
Assistant Professor
Northeastern University

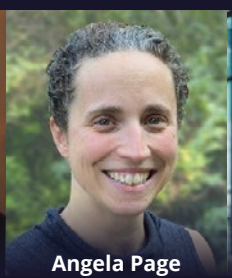


Winston Timp
Associate Professor
John Hopkins University

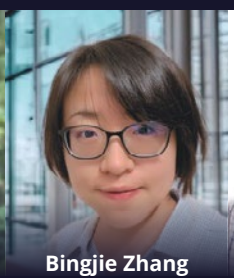


Vasiliki Rahimzadeh
Assistant Professor, Center for Medical Ethics and Health Policy
Baylor College of Medicine

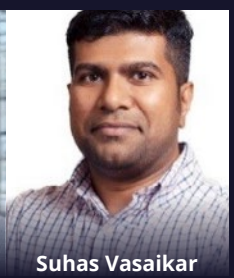
Bradley Malin
Co-Director, Health Data Science Centre
Vanderbilt University



Angela Page
Director of Strategy and Engagement (GA4GH)
The Broad Institute of MIT and Harvard



Bingjie Zhang
Postdoctoral Research Fellow, Satija Lab
New York Genome Center



Suhas Vasaikar
Principal Scientist, Clinical Biomarker and Diagnostics
Seattle Genetics (Seagen)

Matt Higgs

Science Writer
Front Line Genomics

Ashleigh Davey

Science Writer
Front Line Genomics

FOREWORD

NEXT GENERATION SEQUENCING IS FUNDAMENTAL TO GENOMICS AND SCIENTIFIC PROGRESS. IT IS ALSO AN EXCITING SPACE FOR TECHNOLOGICAL DEVELOPMENT AND INNOVATION. WHETHER IN A LARGE-SCALE GENOMIC LAB, IN THE CLINIC OR IN THE FIELD, ACCESS TO SEQUENCING TECHNOLOGY IS BECOMING ESSENTIAL.

This is the rationale behind the Sequencing Buyer's Guide. In an exciting technological market with a variety of options, a resource that provides some clarity is much needed.

Within these pages, you will find a guide for how to select a sequencer and detailed comparisons of the sequencers, both long-read and short-read. You will also find comparisons of the library prep kits available to you and a guide for how to prepare your samples for sequencing. Finally, there will be separate considerations of single-cell sequencing, ESG and accessibility in genomics, outsourcing and some of the latest and exciting developments in sequencing such as solid-state nanopore and multi-omics.

Furthermore, by collating panel discussions and insights from a series of experts in the field, we have gained unique insights and guidance, which have shaped this resource. Excerpts from our discussions with these experts, and discussions between the experts themselves are found throughout the chapters. Within them, you will find advice on how to get the most out of your sequencing, hard fought wisdom gained from working with these technologies, as well as perspectives and views on current topics in sequencing.

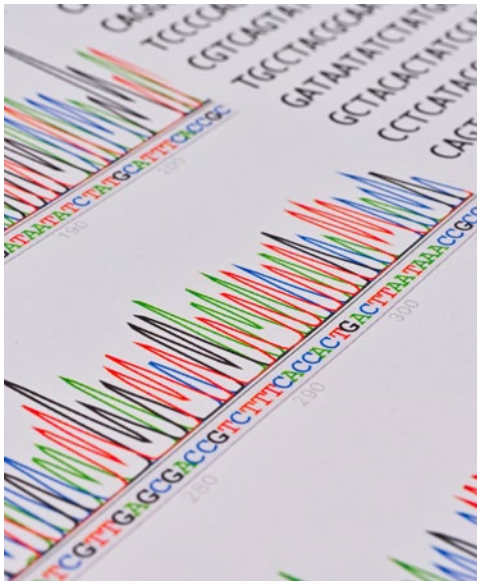
We would like to take this opportunity to thank all of our contributors for their time and insights when writing this Buyer's Guide.

We would also like to thank the sponsors of this report – Illumina, biomodal, Novogene, Source Genomics, Agilent Technologies, QIAGEN, Tecan, Integra and Twist Bioscience.

We hope you find this Buyer's Guide a helpful resource.

Thank you for reading.

CONTENTS



FIRST STEPS - SAMPLE PREPARATION

NEXT-GENERATION SEQUENCING (NGS) HAS REVOLUTIONIZED GENOMICS BY ENABLING HIGH-THROUGHPUT AND COST-EFFECTIVE ANALYSIS OF NUCLEIC ACID SEQUENCES. THIS CHAPTER WILL START AT THE BEGINNING, HOW TO PREPARE SAMPLES FOR DNA EXTRACTION AND HOW TO PREPARE LIBRARIES FOR OPTIMAL SEQUENCING PERFORMANCE.

Good sample preparation is paramount for the success of NGS as it directly influences the quality and reliability of data generated, whilst also minimising biases that may be introduced during sample processing. Therefore, proper sample preparation is essential for obtaining reliable sequencing results, as any errors or inconsistencies introduced at this stage can propagate throughout the entire sequencing workflow. Moreover, high-quality samples with consistent library preparation are crucial for achieving uniform coverage across the genome, reducing biases and enhancing the sensitivity and specificity of variant detection - critical considerations for applications such as whole genome sequencing and clinical diagnostics.

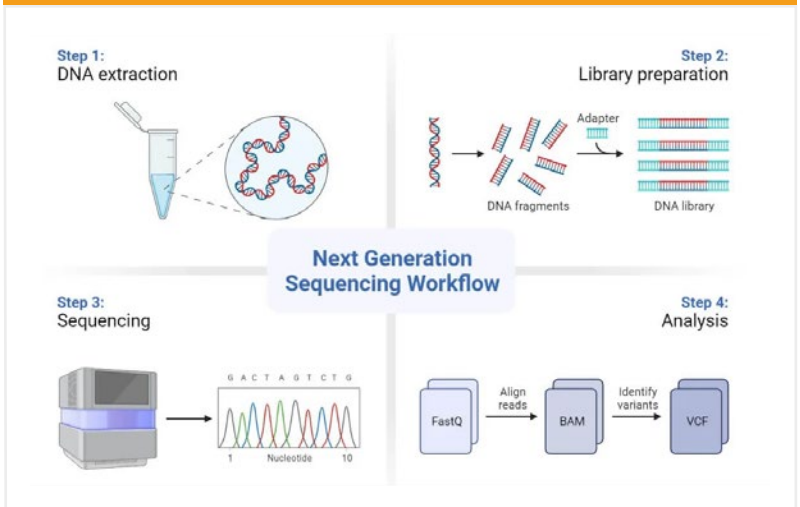
As NGS technologies continue to advance, with an increasing focus on single-cell, multi-omics and ultra-high-throughput sequencing, the significance of robust sample preparation will become even more pronounced. The continued expansion of the sequencing market has led to the introduction of many NGS sample preparation products, which are now capable of creating sequencing libraries from challenging samples for diverse genomics applications. Although a variety of sample preparation products are now commercially available in the NGS space, there is no one-size-fits-all solution; each sample preparation workflow must be carefully optimised for the sequencing instrument and application at hand.

From the initial sample collection to the assembly of genomic data, the typical NGS workflow comprises of multiple steps that can largely be grouped into four major processes; nucleic acid extraction, library preparation, sequencing and data analysis (see Figure 1.1). As good sample preparation lays the foundations for high quality sequencing data, this chapter will delve into steps 1 and 2 of this process, including considerations for nucleic acid extraction of traditional and specialised samples, and the key steps of NGS library preparation.

Curious about the later stages of the NGS workflow? This comprehensive buyer's guide provides key details about sequencing instruments, data analysis and more. Skip ahead to ***Chapter 3: Meet the Sequencers*** or ***Chapter 6: NGS Data Analysis and Data Management*** for more information on these topics.

FIGURE 1.1: BASIC OVERVIEW OF THE NEXT GENERATION SEQUENCING WORKFLOW.

Generally, the NGS workflow is split into 4 main steps: DNA (or RNA) extraction, library preparation, sequencing and bioinformatic analysis (e.g., genome assembly). At several stages of this workflow, the quality and quantity of the DNA is checked to ensure it is suitable for downstream applications.



Droplets make it happen

INTEGRA

LESS PREP MORE PROGRESS

A game-changing microfluidics platform for fully automated NGS prep protocols

NEW

MIRO CANVAS

MIRO Cartridge

MIRO Dropgloss

integra-biosciences.com

Nucleic acid extraction

The first step of the NGS workflow is the extraction of nucleic acids (genomic DNA or RNA) from the biological sample. This process aims to obtain high quantity and purity nucleic acids, whist also removing contaminants that may interfere with downstream sequencing reactions. Although the choice of extraction method depends on the nucleic acid and sample type – commonly, this process utilises animal tissues, cell lines or bodily fluids – extraction frequently uses methods such as phenol-chloroform extraction, silica-based column purification or magnetic bead-based technologies.

The process typically begins with cell lysis, where cellular membranes are disrupted, and nucleic acids are released into solution. For many applications, nucleic acid extraction is done through chemical methods, often involving reagents that break cell membranes and denature proteins (e.g., sodium dodecyl sulfate (SDS) detergent). Although chemical methods are suitable for extraction from cell lines, more structured samples may require additional methods – including physical grinding or enzymatic treatment – to release the nucleic acids from the cells. Notably, RNA extraction often includes an additional step to inhibit RNA-degrading enzymes, such as RNases, to preserve the integrity of the RNA molecules.

Following lysis, proteins, lipids and other contaminants are removed through centrifugation or bead-based methods, leaving purified nucleic acids. The removal of cellular debris is a key step in the extraction process, as failure to do so may affect the purification and downstream applications. Subsequently, the nucleic acid is isolated from the cleared lysate, which is often done via precipitation or silica-binding technologies. The extracted nucleic acids then serve as the starting material for subsequent steps in NGS library preparation, ensuring that the genetic material is representative and suitable for downstream high-throughput sequencing applications.

Rigorous quality control measures are employed during nucleic acid extraction to assess quantity, purity and integrity, ensuring the reliability of the entire NGS workflow. Nucleic acid quantification can be carried out in a variety of ways, most commonly by UV spectroscopy where measuring the absorbance of the sample allows the yield and purity to be calculated.

Specialised samples need careful consideration

For many sequencing experiments, the typical sample includes tissue or bulk cells which can be processed using standard reagents and protocols. However, certain samples (see below) may require nucleic acid extraction and library preparation products that are tailored to create high library complexity from challenging starting materials.

- Blood (plasma/serum)
- FFPE tissue
- Low input/degraded samples
- Viruses
- Single cells

One notable product in this area is the xGen cfDNA & FFPE DNA Library Preparation Kit (Integrated DNA Technologies). Designed specifically for use with cfDNA, FFPE and low input/degraded DNA samples, the kit requires only 1 – 250 ng of DNA input whilst also producing a higher library yield and mean target coverage than other kits on the market[1]. To see additional information on products suitable for low input or sensitive samples, check out ***Chapter 2: Finding Your NGS Library Preparation Kit***.

Library preparation

After nucleic acid extraction, the next step in the sequencing workflow is the creation of a sequencing library. Defined as a collection of nucleic acid (DNA or RNA) fragments of a defined length distribution with adapters attached[2], sequencing libraries can be easily prepared using commercially available kits that perform all the required processing, amplification and purification steps (see below). Importantly, the sequencing library created must be compatible with the desired sequencing instrument, with many platforms having specific requirements for library insert sizes and adapters.

NGS LIBRARY PREPARATION STEPS

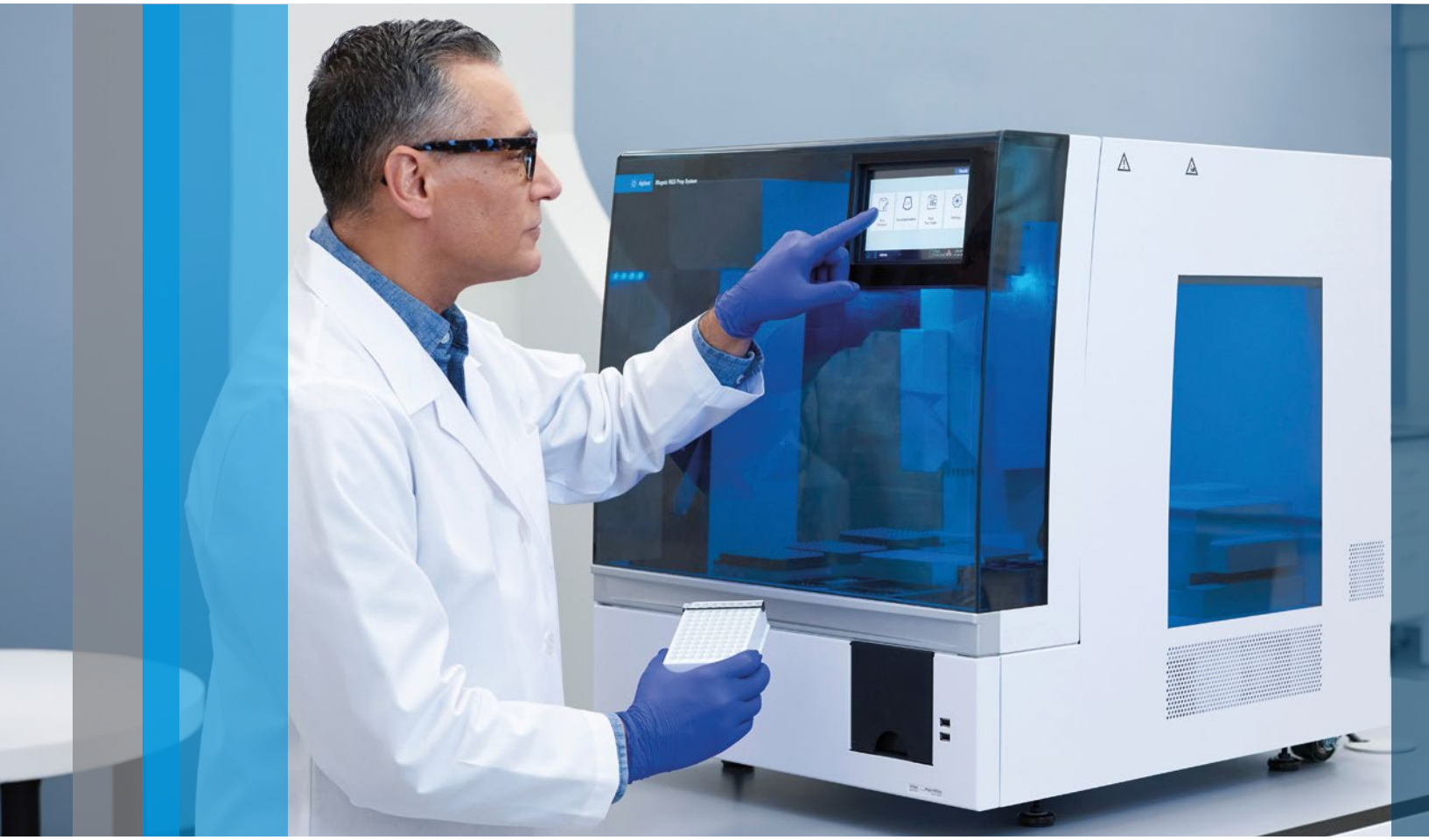
1. ***Fragmentation:*** First, the genomic DNA must be broken into smaller fragments for analysis, with fragments falling within a desired size range. Typically, this is done via either physical (e.g., acoustic shearing/sonication), chemical (e.g., heating with a divalent metal cation) or enzymatic (e.g., restriction endonucleases) methods. The approach selected depends on the DNA insert size required, sample DNA input and equipment availability. Generally, physical and enzymatic fragmentation methods are popular for DNA sample preparation, whilst chemical fragmentation is more often applied to RNA fragmentation (see section: ***RNA library preparation)***.
2. ***End-repair:*** After fragmentation is completed, the nucleic acid pieces contain 5' and 3' overhangs that must be repaired. This is done using enzymes (polymerases) that create blunt ends. Subsequently, the 5' ends are phosphorylated so they are compatible for the adapter ligation process. Some sequencing platforms (including Illumina instruments) may also require the addition of a single adenine base to create a 3' overhang before adapter ligation.
3. ***Adapter ligation:*** Adapters are synthetic oligonucleotides that are attached to both ends of the DNA fragments. These short sequences interact with oligonucleotides on the sequencing instrument's flow cells to ensure that the library is recognised and sequenced. Adapters may also be used as a barcode, with some adapters featuring a short index sequencing thus allowing multiplexing capabilities. One kind of barcode, unique molecular identifiers (UMIs), are often included before amplification steps to help identify PCR biases and rectify sequencing errors.
4. ***Amplification:*** This optional step enables the sequencing library to be amplified via polymerase chain reaction (PCR), which allows lower sample inputs to be used for library preparation. Although helpful in some cases, PCR amplification can introduce GC bias, duplicates or artifacts that can hinder downstream sequencing analysis. For situations where library amplification is not required, PCR-free protocols may be preferrable to ensure high library complexity and more readily enable specific applications (including whole genome sequencing and SNP detection).
5. ***Purification:*** The final step of library preparation involves refining the library to remove unwanted products, leaving only nucleic acid fragments suitable for sequencing. Size selection – often conducted by agarose gel or magnetic bead purification – allows fragments of the correct size to be isolated, resulting in a more uniform library. Additionally, this process removes unwanted products (including excess adapter oligos, dimers and primers) that may reduce sequencing efficiency. The optimal library size is determined by the application and sequencing platform selected.
6. ***Quality control:*** Finally, the library must be analysed to determine that the quantity and the quality of the DNA meets the requirements of the sequencing instrument. Specifically, quality control workflows assess the quantity and size distribution of the library, ensuring that the library fragments fall within the desired range and preventing coverage biases. Additionally, the purity of the library may also be examined to determine if any contamination from adapter dimers is present that may affect the sequencing output.

For more detailed information about NGS library preparation kits currently available, head to ***Chapter 2: Finding Your NGS Library Preparation Kit***.

SEQUENCING LIBRARIES CAN BE EASILY PREPARED USING COMMERCIALLY AVAILABLE KITS THAT PERFORM ALL THE REQUIRED PROCESSING, AMPLIFICATION AND PURIFICATION STEPS"

Agilent

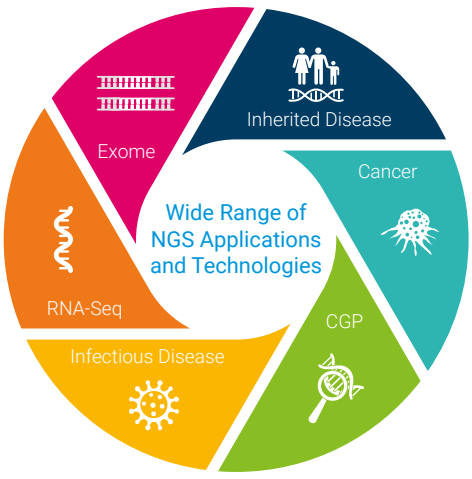
Trusted Answers



Powerful Does Not Mean Complicated

Agilent Magnis SureSelect XT HS2 DNA and RNA Reagent kits

- Walkaway automation for NGS library prep and enrichment, including enzymatic fragmentation and cDNA conversion
- Ultimate ease-of-use, only 15 minutes of hands-on time with the Magnis NGS Prep system
- Low input range with FFPE samples
- Fully validated and optimized SureSelect protocols and reagent kits



www.agilent.com/genomics/ngs_automation

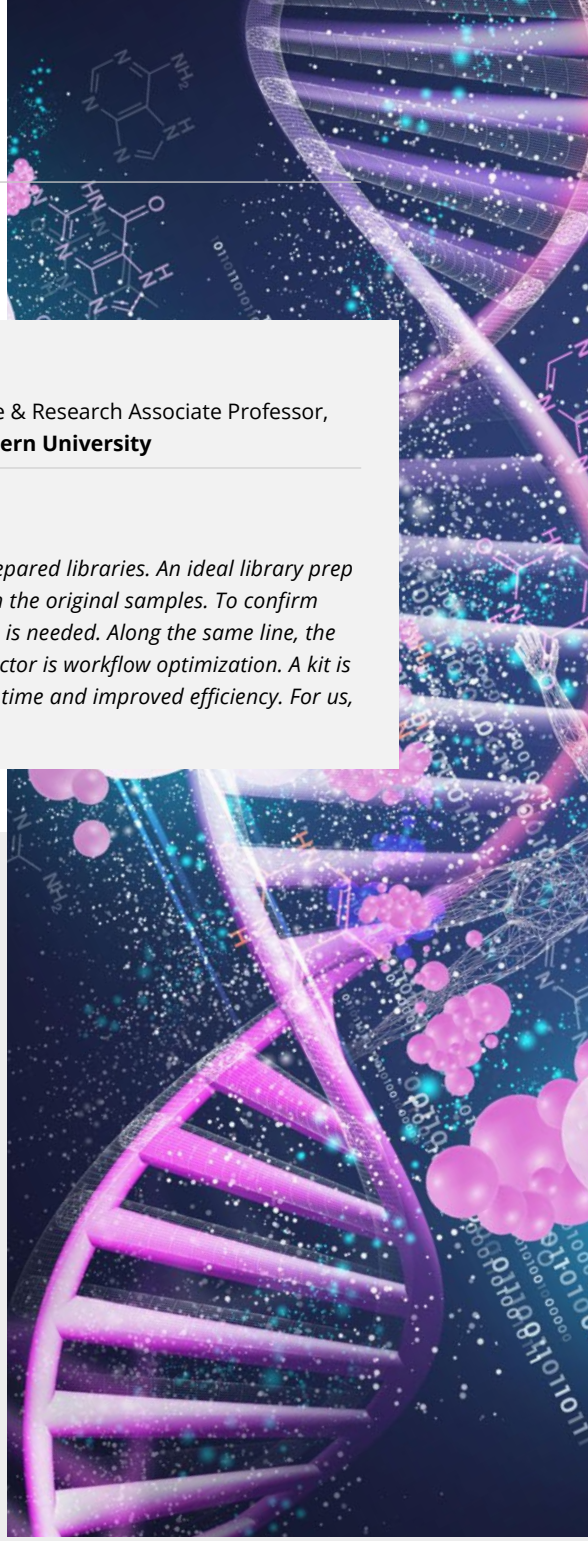
PR7001-1149

XINKUN WANG

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology, **Northwestern University**

FLG: What factors matter when choosing a library prep kit?

Xinkun: *The primary deciding factor is the quality of the data generated from the prepared libraries. An ideal library prep kit needs to produce libraries that faithfully represent the composition of molecules in the original samples. To confirm data quality, a post-NGS validation step using orthogonal technologies, such as qPCR, is needed. Along the same line, the performance of the kit needs to be replicated by different sequencing labs. Another factor is workflow optimization. A kit is always welcome that provides a well streamlined workflow, leading to reduced bench time and improved efficiency. For us, availability of additional help for workflow automation is another welcomed factor.*



Start at the end: an overview of sequencing applications

After taking your sample type and desired sequencing instrument into account, there may still be an overwhelming number of library preparation kits that meet your criteria. A simple way to narrow down your search is to consider the downstream applications of your sequencing experiment.

WHOLE GENOME SEQUENCING (WGS)

One of the most common applications for NGS is WGS, which can be applied to small microbial genomes and large animal or plant genomes alike. This technique delivers a comprehensive overview of the entire genome and can be used for a range of applications, including structural variant detection and de novo genome assembly. Thanks to advancements in sequencing technology, human WGS is possible on both short and long-read platforms. When selecting a WGS library preparation kit, you may want to consider the kit input requirements and insert size generated.

WHOLE EXOME SEQUENCING (WES)

WES is a popular targeted sequencing technique whereby only the protein-coding regions (exons) of the genome are sequenced. This cost-effective approach enables researchers to focus on key areas of the genome that may be relevant to specific diseases, therefore, WES is a fundamental method in areas such as clinical diagnostics. Additionally, the smaller datasets produced by WES may also be desirable, as the sequencing data analysis tends to be faster and less expensive than WGS. Notably, only library preparation kits that contain an exon enrichment step are suitable for WES.

METHYLATION SEQUENCING

Detecting the presence of specific cytosine methylation modifications within the genome is now a popular sequencing application, thanks to the growing knowledge of the role of epigenetics in health and disease development. Methylation profiling is commonly carried out on short-read platforms via bisulphite sequencing, a process whereby unmethylated cytosines are converted into uracil. Post-sequencing, these bases can be identified and the amount of methylated cytosines in the original sample can then be calculated. Alternatively, the methylation status of bases can be directly analysed using Oxford Nanopore Technologies or PacBio long-read instruments, without bisulphite conversion.

MRNA SEQUENCING

It can be argued that one of the most in-demand subcategories of RNA sequencing is mRNA-seq, for transcriptome analysis. This method allows researchers to quantity gene expression by specifically selecting RNA transcripts containing a poly(A) tail – a common feature of mRNAs and some non-coding RNAs. Many companies offer NGS library preparation kits that include poly(A) enrichment, which may or may not also include an additional rRNA depletion step.

SMALL RNA SEQUENCING

In recent years, the interest in small non-coding RNA species (e.g., miRNAs and snoRNAs) and their regulatory roles has been increasing, perhaps due to research uncovering their usefulness as biomarkers for various diseases[3]. Therefore, library preparation kit specifically targeting these populations are becoming increasingly popular. These kits feature a size selection step – often referred to as size fractionation – to isolate RNA that corresponds to the desired length.

DAVID BAKER

Head of Sequencing
The Quadram Institute

FLG: What factors matter when choosing a library prep kit?

Xinkun: *Application and genome size. For larger genomes requiring a lot of data, it's worth paying the extra buck to make a decent library. However, for smaller genomes and lower resolution sequencing, using a cheaper albeit vendor-supplied kit with reduced volume reaction will suffice. In the 1990's when we were making Big Dye Sanger libraries, you'd never do a full reaction, but you'd do an 1/8 or 1/16 to reduce costs without impacting the data. Availability of additional help for workflow automation is another welcomed factor.*

RNA library preparation

Although featuring many of the same steps as DNA library preparation, the creation of RNA sequencing libraries features a few key differences. Typically, most RNA sequencing protocols first require that the extracted RNA is converted into cDNA. This conversion is necessary as many sequencing instruments require PCR-amplified libraries; therefore, RNA must first be transformed into cDNA via reverse transcription. Subsequently, the RNA can be fragmented into smaller pieces suitable for library preparation, which are then subject to adapter ligation and amplification (if required).

During this process, researchers may opt to perform RNA enrichment strategies to enrich the library for the RNA species of interest. Although the RNA in a sample consists of a multitude of different RNA species with varying translational and regulatory roles, the majority of RNA is in fact ribosomal RNA (rRNA) - comprising up to 90% of the total RNA found within cells[4]. Therefore, many RNA-seq applications require rRNA removal (otherwise known as rRNA depletion) to increase the relative abundance of the RNA species of interest for sequencing. In order to do this, rRNA can be isolated and removed by hybridisation capture using rRNA-specific probes, which bind to the rRNA and are then isolated from the RNA sample using magnetic bead separation. This approach is particularly helpful when sequencing long non-coding RNAs that lack additional structural features suitable for enrichment.

For the specific isolation of mRNA transcripts, rRNA depletion may be used in conjunction with poly(A) enrichment – a process which selects RNAs containing a polyadenylated tail using oligo primers. This method is now a staple of mRNA-seq, where poly(A) selection may be used to quickly isolate mRNA transcripts – plus other RNA species containing poly(A) tails - without the requirement for rRNA depletion. For some smaller RNAs containing poly(A) tails such as snoRNAs, size selection enrichment may instead be carried out.

Chapter 1 references

1. Integrated DNA Technologies, Inc. **xGen TM cfDNA & FFPE DNA Library Prep v2 MC Kit.** https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/flyer/xgen-cfdna-ffpe-dna-library-prep-v2-mc-kit.pdf?sfvrsn=3a54fd07_14 (2022).
2. Hess, J. F. *et al.* **Library preparation for next generation sequencing: A review of automation strategies.** *Biotechnol. Adv.* **41**, 107537 (2020).
3. Watson, C. N., Belli, A. & Di Pietro, V. S**mall non-coding RNAs: New class of biomarkers and potential therapeutic targets in neurodegenerative disease.** *Front. Genet.* 10, 364 (2019).
4. O'Neil, D., Glowatz, H. & Schlumpberger, M. **Ribosomal RNA depletion for efficient use of RNA-seq capacity.** *Curr. Protoc. Mol. Biol.* Chapter 4, Unit 4.19 (2013).

QIAseq® – revealing the biological answers you seek

Your research quest is a complex, multi-faceted journey. Next-generation sequencing (NGS) can help you unlock intricate genomic and transcriptomic details.

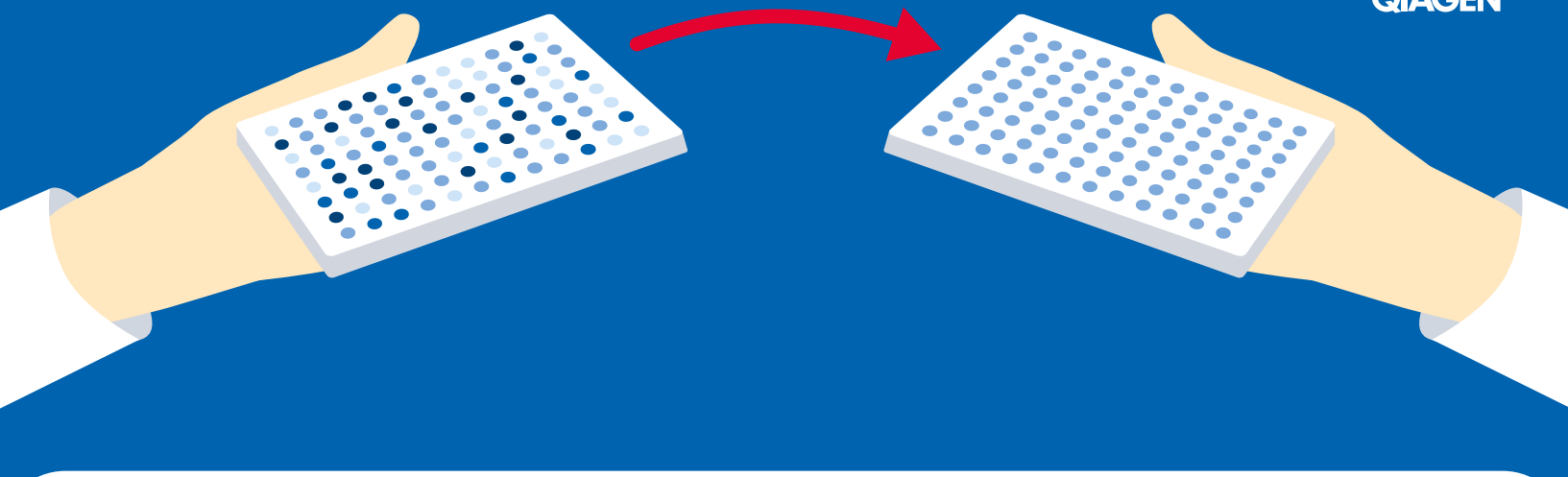
Arm yourself with QIAseq NGS technologies to empower your journey from Sample to Insight®.

Whether you're sequencing DNA or RNA, or even both, QIAseq NGS solutions provide:

- Accurate results from the most challenging of samples
- Streamlined, convenient and automation-friendly workflows
- Meaningful insights from NGS data with integrated bioinformatics
- Flexible customization and fast turnaround times

The secret lies in the sequence. Let QIAseq reveal the answers.

Discover more at **www.qiagen.com/NGS**



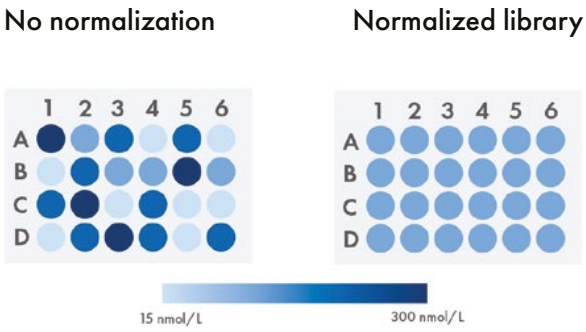
Take the shortcut – normalize libraries without quantification

For cost-efficient multiplexed sequencing, you need a method that can reliably normalize libraries across a wide range of concentrations in a fast and easy manner.

Skip library quantification - take the shortcut with our new QIAseq Normalizer.

QIAseq Normalizer Kits makes use of modified library amplification primers and a chemistry based on magnetic beads to efficiently bind and release a defined amount of library molecules in a tightly controlled fashion. With a target concentration of approximately 4 nmol/L, normalized libraries are tightly adjusted for balanced library pooling and optimal clustering on Illumina flow cells.

- Streamlined, 30-min benchtop protocol
- Gold-standard quality of qPCR without the long workflow
- Works with a broad range of concentrations from 15 to over 300 nmol/L
- Compatible with most Illumina libraries
- Automation-friendly workflow



Learn More: www.qiagen.com/QIAseq-Normalizer

CHAPTER 2

FINDING YOUR NGS LIBRARY PREPARATION KIT

IN THE EVER-EVOLVING LANDSCAPE OF GENOMICS, THE SELECTION OF AN APPROPRIATE LIBRARY PREPARATION KIT FOR NEXT-GENERATION SEQUENCING (NGS) IS A PIVOTAL DECISION THAT CAN SIGNIFICANTLY IMPACT THE QUALITY AND RELIABILITY OF THE DATA GENERATED. THIS CHAPTER COMPARES THE SPECIFICATIONS OF LIBRARY PREP KITS FROM VARIOUS SOLUTION PROVIDERS TO HELP YOU CHOOSE THE RIGHT KIT FOR YOUR APPLICATION

Unsurprisingly, as the number of sequencing platforms has rapidly expanded so too have the number of library preparation kits available. Many third-party kits are now being sold alongside core library preparation kits from sequencing instrument providers. The diverse range of products available ensures that researchers have a wealth of library preparation options, regardless of their sequencing requirements.

Additionally, the development of single-cell analysis techniques has transformed the sequencing space, with kits now being compatible with incredibly low input amounts for library generation. It is easier than ever to sequence precious samples at high sensitivity – essential for the analysis of rare and low abundance nucleic acid samples such as cell-free DNA. Many kits offer the ability to use a wide range of nucleic acid input amounts, meaning that one kit can be applied to a multitude of biological applications, including whole genome sequencing (WGS), transcriptome analysis, targeted sequencing of specific genomic loci and variant detection.

In this chapter, we highlight the various NGS library preparation kits (for both DNA and RNA sequencing) with a special focus on new products launched in the last year. However, new is not always better, so we have carefully curated this chapter to also include tried-and-true assays that have become reliable favourites within the field. By providing a succinct overview of the library preparation kit options available, this chapter aims to guide researchers in selecting a kit that best suits their experimental needs.

For more detailed information on sequencing instrument options, please refer to ***Chapter 3: Meet the Sequencers.***

DNA Library Preparation

As DNA is the most well-established target for NGS analysis, a range of DNA library preparation kits have entered the market that are suited to various input quantities, genome types or applications. Like sequencing instruments, the kits can be divided into two main categories: those compatible with short-read sequencers and those designed for use with long-read platforms. Therefore, the selection process begins with choosing a kit that is complementary to the type of sequencing and instrument selected. Subsequently, a number of factors must be considered to ensure that the kit chosen can generate a robust library suitable for downstream sequencing and analysis.

It all adds up: factors to consider when purchasing a DNA library preparation kit

ASSAY SIMPLICITY

In a world where high-throughput sequencing is becoming increasingly more common, the creation of NGS libraries with fewer steps, less reagents and simple instructions makes the sequencing process more manageable – especially when multiple samples are being processed in parallel. One prominent example of this is magnetic bead-based DNA purification, a labour-intensive step present in some kits that can result in errors in library preparation[1]. Therefore, selecting a kit with minimal pipetting steps or one with processes that can be automated (see below) may result in less errors and higher reproducibility.

One of the most prominent factors of complexity is time – researchers often desire an efficient workflow with minimal hands-on time required. In response to this demand, newer NGS library preparation products with streamlined workflows have been launched that can create functional libraries in just a couple of hours.

AUTOMATION

Preparing hundreds or thousands of samples at once is now a reality for some research labs. Luckily, many vendors – including Twist Bioscience, Illumina, New England Biolabs and Qiagen – now offer kits compatible with automation that allow theefficient processing of samples to produce NGS libraries with less human intervention. Such automation systems typically include liquid handling instruments, capable of autonomously running the full library preparation protocol. In addition to reducing hands-on time for researchers, automation also decreases contamination errors, improves sample quality and allows laboratories to scale-up their sequencing preparations as needed[2].

DNA INPUT

For some samples, the amount and type of DNA required for library preparation needs to be considered. Many kits are now compatible with low input amounts (often as little as 1 ng or less). Although this small amount may affect the quality or coverage of downstream sequencing data, this low input requirement is helpful when dealing with uncommon or low abundance samples.

Additionally, samples with damaged or fragmented DNA are more challenging to process for NGS. One prominent example of this is FFPE samples - preserved biological tissue samples that are commonly used in immunohistochemistry analysis. More recently, cell-free DNA has also entered the limelight as an important NGS target that is often present in low abundance or highly fragmented, which can also pose a problem for library preparation using standard kits.

To remedy this, there are several kits on the market that are specialised for dealing with damaged, fragmented or low-quality DNA samples – such as the Twist cfDNA Library Prep Kit from Twist Bioscience. As a result, specialised products like these allow researchers to rescue valuable sequencing data from rare sources.

FLG: How much of your process is automated? Is there anything people bear in mind when trying to automate sequencing prep?

David Baker

Head of Sequencing
The Quadram Institute

When thinking about automation, always think about the human resource needed to run it. These things are rarely 'click and go', you usually need somebody to watch. Also, you need somebody who can program and look after the automation. Operating many applications might be an issue. If you were doing the same thing day in and day out, then it might be worth the investment. Finally, it is also rare that a liquid handler is quicker than a person, so throughput is also important.

Xinkun Wang

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology
Northwestern University

We try to automate as much of the library prep process as possible. Some steps that require human input or intervention cannot be easily automated, including sample quality assessment and PCR amplification steps that require thermal cyclers are not available on-board. The rest of the steps are essentially all automatable. As automation systems cannot replicate every aspect of human operator-executed library prep, the library yield from such systems is typically lower than that from manual prep, which can be resolved by using more input DNA/RNA.

PCR VS NO PCR

Library amplification via PCR is often a required step for many library preparation kits. Although PCR allows researchers to sequence samples with low DNA content, PCR may introduce GC bias, amplification bias and duplicates that may hinder downstream genome assembly or data analysis. To counteract this problem, many vendors have created PCR-free kits that offer reduced assay times and increased coverage across genomic regions that are traditionally challenging to sequence. One well-established example is Illumina's TruSeq DNA PCR-Free kit, which shows impressive coverage improvement for G-rich, high GC and promoter regions when compared to data from their TruSeq DNA kit featuring PCR amplification[3].

MULTIPLEXING

Sample multiplexing (occasionally referred to as indexing) is the process of tagging each DNA fragment to identify which sample that DNA fragment originated from. By doing this, researchers can pool the libraries for multiple samples together and sequence several samples in parallel. After sequencing, the unique tag (often referred to as a barcode or index) can then be used to group the data into their respective samples before analysis takes place.

Multiplexing is an attractive way to conduct high-throughput sequencing and save both time and money. There are now many options for high-throughput adapter sets that enable large-scale sample multiplexing, with varying platform compatibility and number of adapters. One versatile example is Twist Bioscience's Universal Adapter Systems, which consist of Twist's Universal Adapters (compatible with all "T-A" overhangs workflows) and either Twist Unique Dual-Indexed (UDI) Primers or Twist HT UDI Primers (3,072 indexes). Alternatively, users may instead opt for the latest full-length UDI adapters (1,536 indexes), which are perfect for PCR-free WGS applications.

COST

Ultimately, the best library preparation kit is one that produces the highest quality data for the lowest price. Hidden reagent costs, expenses associated with analysis, researcher time and kit usage (e.g., wasting resources due to reagent expiration) can all play a factor in the total cost per reaction – in addition to the initial expense of the library preparation kit itself. If the experiment requires a small number of samples, or if sequencing is not regularly required, labs may instead opt to outsource sequencing to NGS service providers. For a comprehensive overview of sequencing outsourcing, we direct you to ***Chapter 7: What About Outsourcing?***

In the following sections, we list the newest and most popular products from the leaders within the NGS library preparation space. Notably, kits for highly specialised starting materials (e.g., tumour tissue) and gene panels have been omitted for brevity. Instead, we focus on kits that are relevant to a wide variety of experimental setups. Although not covered here, many of the suppliers listed also sell more specialised sequencing products, such as Illumina's AmpliSeq panels and Thermo Fisher Scientific's Oncomine assays.

DNA library preparation kits for short-read systems

Owing to their immense popularity, the leaders of the short-read industry, Illumina, currently dominate the library preparation market (see Table 2.1). In addition to the various library preparation kits supplied by Illumina, many third-party companies also offer products that are compatible with Illumina sequencing instruments. In contrast, library preparation kits suitable for other short-read platforms (such as those offered by MGI and PacBio) can be purchased directly from the supplier.

Notably, many sequencing platforms (such as Singular Genomics' G4 instrument and Element Biosciences' AVITI system) offer library conversion kits, custom adapters and indices, which enable users to perform their library preparation using a wide range of kits from various third-party suppliers. This allows the sequencing instruments to easily slot into any preexisting NGS workflows.

TABLE 2.1: DETAILS FOR SELECTED DNA LIBRARY PREPARATION KITS COMPATIBLE WITH SHORT-READ SEQUENCING PLATFORMS.

Information was sourced from the supplier's website or handbook associated with the product. Details provided were correct at the time of publication (January 2024), please check supplier's website for the most up-to-date specifications and data. Abbreviations: TCS – targeted capture sequencing, WES – whole exome sequencing, WGS – whole genome sequencing.

Supplier	Kit	System Compatibility	Assay time	Input Quantity	Target Insert Size	PCR Required	Applications
Twist Bioscience	cfDNA Library Preparation Kit	Illumina instruments	2 hours	1 ng - 20 ng	300-400 bp	Yes	Targeted sequencing, WGS
	Library Preparation Enzymatic Fragmentation Kit 2.0	Illumina instruments	3 hours	1 ng – 500 ng	Flexible	Yes	Targeted sequencing, WGS
	Mechanical Fragmentation Library Preparation Kit	Illumina instruments	3 hours	1 ng – 500 ng	200 bp – 250 bp	Yes	Targeted sequencing, WGS
Agilent	SureSelect XT HS2 DNA Reagent Kit	Illumina instruments Element instruments (library conversion required)	9 hours (TCS)	10 - 200 ng of total DNA (from intact or highly fragmented FFPE samples)	Customisable	Yes	DNA targeted enrichment
Illumina	Illumina DNA PCR-Free Prep	Illumina: MiSeq, MiSeqDx in Research Mode, NextSeq 500, NextSeq 550, NextSeq 550Dx in Research Mode, NovaSeq 6000, NovaSeq X, NovaSeq X Plus	1.5 hours	25 ng - 300 ng	450 bp +/- 75 bp	No	De novo assembly, WGS
	Illumina DNA Prep	Illumina: HiSeq 3000, HiSeq 4000, HiSeq X Five, HiSeq X Ten, iSeq 100, MiniSeq, MiSeq, MiSeqDx in Research Mode, NextSeq 1000, NextSeq 2000, NextSeq 550, NextSeq 550Dx in Research Mode, NovaSeq 6000, NovaSeq X, NovaSeq X Plus	3 - 4 hours	Small genomes (e.g. microbial): 1-500 ng DNA Large genomes (e.g. human): 100-500 ng DNA	~350bp	Yes	Amplicon sequencing, De novo assembly, WGS
	Nextera XT DNA Library Preparation Kit	Illumina: iSeq 100, MiniSeq, MiSeq, NextSeq 1000, NextSeq 2000, NextSeq 500, NextSeq 550	5.5 hours	1 ng	300 bp – 1.5 kb	Yes	16S rRNA sequencing, amplicon sequencing, De novo assembly, WGS
	TruSeq DNA Nano	Illumina: HiSeq 3000, HiSeq 4000, HiSeq X Five, HiSeq X Ten, MiniSeq, MiSeq, MiSeqDx in Research Mode, NextSeq 2000, NextSeq 500, NextSeq 550, NovaSeq 6000, NovaSeq X, NovaSeq X Plus	6 hours	100 ng	350 bp or 550 bp	Yes	Genotyping, WGS
	TruSeq DNA PCR-Free	Illumina: HiSeq 3000, HiSeq 4000, HiSeq X Five, HiSeq X Ten, MiniSeq, MiSeq, MiSeqDx in Research Mode, NextSeq 1000, NextSeq 2000, NextSeq 500, NextSeq 550, NovaSeq 6000	5 hours	1 ug	350 bp or 550 bp	No	Genotyping, WGS

TABLE 2.1: DETAILS FOR SELECTED DNA LIBRARY PREPARATION KITS COMPATIBLE WITH SHORT-READ SEQUENCING PLATFORMS. CONTINUED.

Supplier	Kit	System Compatibility	Assay time	Input Quantity	Target Insert Size	PCR Required	Applications
Integrated DNA Technologies	xGen DNA EZ Library Prep Kit	Illumina instruments	<2 hours	100 pg – 1 µg	550 bp	Yes	Genotyping, WES, WGS
	xGen DNA Library MC Kit	Illumina instruments	2 hours	1 ng – 1 µg	350 bp	Yes	WGS, WES, metagenomic sequencing
	xGen ssDNA & Low-Input DNA Library Prep Kit	Illumina instruments	2 hours	10 pg – 250 ng	200 bp or 350 bp	Yes	Sequencing of low-quality degraded DNA/ ssDNA
MGI	MGIEasy Duplex UMI Universal Library Prep Set	MGI: DNBSEQ-G400, DNBSEQ-G50, DNBSEQ-T7	7 hours	10 ng – 1 µg	200 bp – 600 bp	Yes	Low frequency variant detection
	MGIEasy Exome FS Library Prep Set V2.0	MGI: DNBSEQ-G400	7 hours	50 ng –400 ng	200 bp – 600 bp	Yes	WES
	MGIEasy Fast FS DNA Library Prep Set	MGI: DNBSEQ-G400, DNBSEQ-G50, DNBSEQ-T7	2 - 2.3 hours	1 ng – 200 ng	200 bp – 600 bp	Yes	Microbial genome sequencing, metagenome sequencing
	MGIEasy Fast PCR-FREE FS Library Prep Set	MGI: DNBSEQ-G400, DNBSEQ-G50, DNBSEQ-T7	1.5 - 1.8 hours	50 ng – 300 ng	200 bp – 600 bp	No	Low pass WGS, viral WGS, long amplicons sequencing
	MGIEasy FS DNA Library Prep Set	MGI: DNBSEQ-T7, DNBSEQ-G400, DNBSEQ-G50	5.5 hours	5 ng – 400 ng	200 bp – 600 bp	Yes	WGS
	MGIEasy PCR-Free Library Prep Set	MGI: DNBSEQ-G400, DNBSEQ-G50, DNBSEQ-T7	3.5 hours	80 ng –200 ng	350 bp –400 bp	No	WGS
New England Biolabs	NEBNext Enzymatic Methyl-seq Kit	Illumina instruments	7.5 hours	10 ng – 200 ng	Customisable	Yes	Methylome Analysis
	NEBNext Fast DNA Fragmentation & Library Prep Set for Ion Torrent	Ion Torrent instruments	<2 hours	10 ng – 1 µg	Enzymatic Fragmentation; Customisable	Yes	WGS, WES
	NEBNext Fast DNA Library Prep Set for Ion Torrent	Ion Torrent instruments	<2 hours	10 ng – 1 µg	Customisable	Yes	WGS, WES
	NEBNext FFPE DNA Library Prep Kit	Illumina instruments	2.8 – 3.8 hours	5 ng – 250 ng	Customisable	Yes	FFPE DNA
	NEBNext Ultra II DNA Library Prep Kit for Illumina	Illumina instruments	1.7 – 3.2 hours	500 pg – 1 µg	Customisable	Yes	WGS, WES, ChIP-seq
	NEBNext Ultra II FS DNA Library Prep Kit for Illumina	Illumina instruments	1.3 – 3.2 hours	100 pg – 500 ng	Enzymatic Fragmentation; Customisable	Yes	WGS, WES
	NEBNext Ultra II DNA PCR-free Library Prep Kit for Illumina	Illumina instruments	<2 hours	250 ng – 1 µg	400 bp peak size range	No	WGS, WES
	NEBNext Ultra II FS DNA PCR-free Library Prep Kit for Illumina	Illumina instruments	<2 hours	50 ng – 500 ng	Enzymatic Fragmentation; 300 bp, 350 bp, 450 bp peak size ranges	No	WGS, WES
	NEBNext UltraExpress DNA Library Prep Kit	Illumina instruments	110 minutes	10 ng – 200 ng	Customisable (200 bp peak size by default)	Yes	WGS, WES
	NEBNext UltraExpress FS DNA Library Prep Kit	Illumina instruments	105 minutes	10 ng – 200 ng	Enzymatic Fragmentation; Customisable (150 bp peak size by default)	Yes	WGS, WES
	NEBNext Ultrashear FFPE DNA Library Prep Kit	Illumina instruments	3.25 – 4.25 hours	5 ng – 250 ng	Enzymatic Fragmentation; Customisable	Yes	FFPE DNA

TABLE 2.1: DETAILS FOR SELECTED DNA LIBRARY PREPARATION KITS COMPATIBLE WITH SHORT-READ SEQUENCING PLATFORMS. CONTINUED.

Supplier	Kit	System Compatibility	Assay time	Input Quantity	Target Insert Size	PCR Required	Applications
PacBio	Onso DNA Library Prep kit	PacBio: Onso	3 hours	10 ng – 1 µg	Application specific	PCR-free and amplified workflows available	WGS, WES, targeted sequencing
	Onso fragmentation DNA library prep kit	PacBio: Onso	3 hours	10 ng – 500 ng	Application specific	PCR-free and amplified workflows available	WGS, WES, targeted sequencing
Qiagen	QIAseq FX DNA Library Kit	Illumina instruments, Element AVITI	2.5 hours	20 pg – 1 µg	Flexible	Optional	WGS< WES, metagenomic sequencing
	QIAseq Ultralow Input Library Kit	Illumina instruments, Element AVITI	2.5 hours	10 pg –100 ng	Not specified	Optional	Sequencing of low-input/ degraded DNA
Roche	KAPA EvoPlus Kits	Illumina instruments	2 hours	10 ng – 500 ng	Flexible	Optional	WES, WGS
	KAPA HyperPlus Kits	Illumina instruments	2.5 hours	1 ng – 1 µg	Flexible	Optional	WES, WGS
	KAPA HyperPrep Kits	Illumina instruments	2.75 hours	1 ng – 1 µg	Not specified	Optional	Amplicon sequencing, WES, WGS,
Thermo Fisher Scientific	Collibri PCR-free PS DNA Library Prep Kits for Illumina Systems	Illumina: iSeq, HiSeq 1000/1500/2000/2500/3000/4000/X, MiSeq, MiniSeq, NextSeq 500/550, NovaSeq 6000	2 hours	500 ng – 1 µg	150 bp – 1000 bp	No	WGS
	Collibri PS DNA Library Prep Kits for Illumina Systems	Illumina: iSeq, HiSeq 1000/1500/2000/2500/3000/4000/X, MiSeq, MiniSeq, NextSeq 500/550, NovaSeq 6000	~2.5 hours	1 ng – 1 µg	150 bp –1000 bp	Yes	WGS
	Ion Plus Fragment Library Kit	Ion Torrent: Ion PGM, Ion Proton, Ion S5, Ion S5 XL, Ion GeneStudio™ S5 Series	~2 hours	100 ng or 1 µg	100 bp –600 bp	Yes	WGS

In 2023, several new products were launched in the short-read DNA library preparation area. One notable highlight is the release of New England Biolab's new NEBNext UltraExpress™ DNA Kit, compatible with Illumina platforms. Announced in November 2023, the kits are designed with speed in mind – with DNA libraries being successfully generated within 2 hours. The kits also follow a simplified protocol using the same adapter concentration and PCR cycle numbers regardless of input amount used, making library preparation easier whilst maintaining sample input flexibility.

Additionally, 2023 was a milestone year for PacBio with the launch of their Onso short-read platform. The novel system was launched in August last year alongside two library preparation kits, designed to create libraries from high-molecular weight or fragmented/damaged samples. The system aims to generate high (Q40+) accuracy data that supports all major short-read applications.

DNA library preparation kits for long-read systems

Over in the long-read arena, the industry leaders – Oxford Nanopore Technologies and PacBio – have a number of library preparation kits available to purchase directly (see Table 2.2). Interestingly, third-party suppliers of products for long-read library preparation are less common than in short-read sequencing. However, one notable outlier is New England Biolabs, which have partnered with Oxford Nanopore Technologies on their NEBNext® Companion Module for use in Oxford Nanopore Technologies' ligation sequencing protocols.

Interested in the possible benefits of long-read technology? For a more in-depth overview of short and long-read sequencing, check out ***Chapter 3: Meet the Sequencers.***

TABLE 2.2: DETAILS FOR SELECTED DNA LIBRARY PREPARATION KITS COMPATIBLE WITH LONG-READ SEQUENCING PLATFORMS.

Information was sourced from the supplier's website or handbook associated with the product. Details provided were correct at the time of publication (January 2024), please check supplier's website for the most up-to-date specifications and data. Abbreviations: WES – whole exome sequencing, WGS – whole genome sequencing.

Supplier	Kit	System Compatibility	Assay time	Input Quantity	Applications
Illumina	Illumina Complete Long Read Prep, Human	Illumina: NovaSeq 6000, NovaSeq 6000Dx, NovaSeq X, NovaSeq X Plus	7.5 – 8.5 hours	50 ng DNA	WGS
Oxford Nanopore Technologies	16S Barcoding Kit	Oxford Nanopore: MinION Mk1B, MinION Mk1C, GridION, Flongle	25 minutes (plus PCR time)	10 ng gDNA	Bacterial 16S rRNA sequencing, metagenomics, targeted sequencing
	Ligation Sequencing Kit	Oxford Nanopore: Flongle, MinION Mk1B, MinION Mk1C, GridION, PromethION	1 hour	1000 ng gDNA or 100-200 fmol for amplicons	WGS, methylation analysis, metagenomics, single-cell sequencing, targeted sequencing
	Rapid PCR Sequencing Kit	Oxford Nanopore: MinION Mk1B, MinION Mk1C, GridION, PromethION, Flongle	15 minutes (plus PCR time)	1 ng – 5 ng gDNA	WGS, metagenomics
	Rapid Sequencing Kit	Oxford Nanopore: Flongle MinION Mk1B MinION Mk1C GridION PromethION	10 minutes	50 ng – 100 ng gDNA	WGS, metagenomics, methylation analysis
	Ultra-long DNA Sequencing Kit	Oxford Nanopore: MinION Mk1B, MinION Mk1C, GridION, PromethION	~ 3.5 hours (plus overnight elution)	6M cells	WGS, methylation analysis
PacBio	SMRTbell prep kit 3.0	PacBio: Revio and Sequel II/IIe	4 hours	300 ng - 5 µg	Metagenomics, methylation analysis, targeted sequencing, WGS

Illumina threw their hat into the ring of long-read sequencing in 2023, with the debut of their novel Complete Long Read Prep kit. The novel technology makes long-read sequencing possible on their powerful NovaSeq instruments, which are commonly used for short-read sequencing. Following a simple one-day protocol, users can generate sequencing reads of 5-7 kb (with some reads >10 kb) that enable complete coverage of genomic regions that may be missed using short-read technology alone.

Last year also saw PacBio announce their collaboration with leading library preparation automation partners, ensuring that their HiFi long-read sequencing can be carried out faster and with higher throughput. The sequencing giants informed customers that several automation providers – Hamilton, Integra, Revvity and Tecan – have successfully created fully automated protocols for PacBio's Revio and Sequel instruments, enabling users to easily scale up their sequencing to thousands of genomes per year[5].

RNA library preparation

As a relative newcomer to the sequencing world, RNA sequencing (RNA-seq) has recently exploded in popularity, with transcriptomics and spatial analysis proving themselves to be essential tools in deciphering cellular function. Now, RNA-seq is more accessible than ever before, with a multitude of library preparation options for both short-read and long-read sequencing platforms. Unlike DNA sequencing, RNA-seq often relies on complimentary DNA (cDNA) generation during library preparation before sequencing can take place. For more information on processing samples for RNA-seq, head to ***Chapter 1: First Steps – Sample Preparation.***

RNA LIBRARY PREPARATION KITS FOR SHORT-READ SYSTEMS

Much like in DNA library preparation, the market for short-read RNA library preparation kits is heavily focussed on products compatible with Illumina sequencers. However, as with the DNA kits, Singular Genomics and Element Biosciences instruments are compatible with many popular RNA library preparation kits from companies such as Bio-Rad, Integrated DNA Technologies and Lexogen. Below, we have listed some of the newest and most reputable library preparation kits currently being used for RNA analysis (see Table 2.3). Notably, kits for rRNA depletion or specialised sample processing have been excluded.

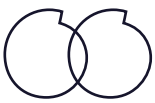
TABLE 2.3: DETAILS FOR SELECTED RNA LIBRARY PREPARATION KITS COMPATIBLE WITH SHORT-READ SEQUENCING PLATFORMS.

Information was sourced from the supplier's website or handbook associated with the product. Details provided were correct at the time of publication (January 2024), please check supplier's website for the most up-to-date specifications and data. Input quantity listed assumes RNA is extracted from whole cells/tissues. For RNA from other sources (e.g. blood plasma), see supplier's protocols for recommended input quantity. Abbreviations: TCS – target capture sequencing, WTS – whole transcriptome sequencing.

Supplier	Kit	Compatibility	Assay time	Input Quantity	Applications
Twist Bioscience	RNA Library Preparation Kit	Illumina instruments	<5 hours	1 ng – 1 µg total RNA	mRNA-seq, RNA-seq (including lncRNA), targeted RNA-seq
Agilent	SureSelect XT HS2 RNA Reagent Kit	Illumina instruments	6.5 hours (WTS) TCS (11 hours)	WTS: 10 - 1000 ng	WTS, RNA targeted enrichment
Bio-Rad	SEQuoia Complete Stranded RNA Library Prep Kit	Illumina instruments	<4 hours	100 pg –1 µg total RNA	RNA-seq, mRNA-seq, small RNA-seq
	SEQuoia Express Stranded RNA Library Prep Kit	Illumina instruments	3 hours	1 ng – 1 ug total RNA	mRNA-seq, lncRNA sequencing
Illumina	Illumina RNA Prep with Enrichment	Illumina: HiSeq 3000, HiSeq 4000, iSeq 100, MiniSeq, MiSeq, NextSeq 1000, NextSeq 2000, NextSeq 500, NextSeq 550, NovaSeq 6000, NovaSeq X, NovaSeq X Plus	<9 hours	10 ng total RNA from fresh/frozen samples	mRNA-seq, targeted RNA-seq
	Illumina Stranded mRNA Library Prep Kit	Illumina: HiSeq 3000, HiSeq 4000, NextSeq 1000, NextSeq 2000, NextSeq 500, NextSeq 550, NovaSeq 6000, NovaSeq X, NovaSeq X Plus	~ 6.5 hours	25 ng – 1 µg total RNA	mRNA-seq
	TruSeq RNA Library Prep Kit v2	Illumina: Genome Analyzer IIx, HiSeq 3000, HiSeq 4000, MiSeq, NextSeq 500, NextSeq 550	~10.5 hours	100 ng – 1 ug total RNA	mRNA-seq
	TruSeq Small RNA Library Prep Kit	Illumina: iSeq 100, MiniSeq, MiSeq, NextSeq 1000, NextSeq 2000, NextSeq 500, NextSeq 550	1 day	1 µg total RNA	Small RNA-seq
	TruSeq Stranded Total RNA Library Prep Kit	Illumina: HiSeq 3000, HiSeq 4000, NextSeq 1000, NextSeq 2000, NextSeq 500, NextSeq 550, NovaSeq 6000	11.5 hours	100 ng - 1 µg total RNA	RNA-seq (including mRNA and lncRNA)
Integrated DNA Technologies	xGen Broad-Range RNA Library Prep Kit	Illumina instruments	4.5 hours	10 ng – 1 µg total RNA	RNA-seq (including mRNA and lncRNA)
	xGen RNA Library Preparation Kit	Illumina instruments	3.5 hours	100 ng – 1 µg total RNA	RNA-seq, mRNA-seq
Lexogen	CORALL RNA-Seq V2	Illumina: HiSeq 2000/2500/3000/4000, NextSeq 500/550/2000, MiSeq, iSeq, NovaSeq 6000	4.5 hours	1 ng – 1 µg total RNA	RNA-seq, ribosomal RNA-depleted RNA-seq
	LUTHOR High-Definition Single-Cell 3' mRNA-Seq Library Prep Kit	Illumina: iSeq 100, MiniSeq, NextSeq 500 – 2000, HiSeq 2000, 2500/3000/4000, NovaSeq 6000 (v1.0 and 1.5 reagent kits), NovaSeq X	~6 hours	10 pg – 1 ng	Single-cell RNA-seq
	QuantSeq 3' mRNA-Seq V2 Library Prep Kit FWD	Illumina: iSeq 100, MiniSeq, NextSeq 500 – 2000, HiSeq 3000 and 4000, NovaSeq 6000 (v1.5 reagent kits)	4.5 hours	1 ng – 500 ng total RNA	mRNA-seq
	Small RNA-Seq Library Prep Kit	Illumina: HiSeq 2000/2500/3000/4000, NextSeq 500/550, MiSeq, MiniSeq, Genome Analyzer	5 hours	100 ng – 1 µg total RNA	Small RNA-seq

TABLE 2.3: DETAILS FOR SELECTED RNA LIBRARY PREPARATION KITS COMPATIBLE WITH SHORT-READ SEQUENCING PLATFORMS. CONTINUED.

Supplier	Kit	Compatibility	Assay time	Input Quantity	Applications
MGI	MGIEasy RNA Directional Library Prep Set	MGI: DNBSEQ-G400, DNBSEQ-G50	7 hours	10 ng – 1 µg total RNA	RNA-seq, mRNA-seq
	MGIEasy RNA Library Prep Set	MGI: DNBSEQ-G400, DNBSEQ-G50	7 hours	10 ng – 1 µg total RNA	RNA-seq, mRNA-seq
	MGIEasy Small RNA Library Prep Kit	MGI: DNBSEQ-G400	Not specified	10 ng – 1 µg total RNA	Small RNA-seq
New England Biolabs	NEBNext® Multiplex Small RNA Library Prep Set for Illumina®	Illumina instruments	7 hours	100 ng–1 µg total RNA	Small RNA-seq (< 200 nt)
	NEBNext Single Cell/Low Input RNA Library Prep Kit	Illumina instruments	6 – 7 hours	Single cells or 2 pg – 200 ng RNA	Single-cell RNA-seq, low input samples
	NEBNext Ultra II Directional RNA Library Prep Kit	Illumina instruments	5.5 – 5.7 hours with poly(A) enrichment, 6.6 – 6.8 hours with rRNA depletion	10 ng – 1 µg total RNA	RNA-seq for gene expression analysis, alternative splicing, and transcriptome analysis
	NEBNext UltraExpress RNA Library Prep Kit	Illumina instruments	3 hours	25 – 250 ng total RNA	RNA-seq for gene expression analysis, alternative splicing, and transcriptome analysis
Qiagen	QIAseq miRNA Library Kit	Illumina instruments, Thermo Fisher NGS instruments, Element AVITI, Singular Genomics G4	8 hours	1 ng – 500 ng total RNA	miRNA-seq
	QIAseq Stranded RNA Library Kits	Illumina instruments, Element AVITI, Singular Genomics G4	4 – 5 hours	1 ng – 1000 ng of total RNA; 100 ng of total RNA (or greater) is recommended	RNA-seq, strand-specific RNA-seq, whole transcriptome analysis
	QIAseq UPX 3' Transcriptome Kit	Illumina instruments	Not specified	10 pg – 10 ng total RNA	3' mRNA-seq
Revvity (formerly affiliated with PerkinElmer)	NEXTFLEX Rapid Directional RNA-Seq Kit 2.0	Illumina instruments	~ 7 hours	5 ng – 5 µg total RNA	RNA-seq
	NEXTFLEX® Small RNA Sequencing Kit V4	Illumina instruments	~5 hours	1 ng – 1 µg total RNA	Small RNA-seq
Roche	KAPA mRNA HyperPrep Kit	Illumina instruments	6 hours	50 ng – 1 µg total RNA	mRNA-seq
	KAPA RNA HyperPrep Kit	Illumina instruments	4.5 hours	1 ng – 100 ng total RNA	RNA-seq, targeted RNA-seq
Takara Bio	SMARTer Target RNA Capture	Illumina instruments	2 days	10 ng – 1 µg total RNA	Targeted RNA-seq
	SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing	Ion Torrent and Illumina instruments	Not specified	1,000 cells or 10 pg – 10 ng of total RNA	Single cell RNA-seq, transcriptome analysis
Thermo Fisher Scientific	Ion Total RNA-Seq Kit v2	Ion Torrent: Ion GeneStudio S5, Ion PGM, Ion Proton	~ 5 hours	1 ng – 500 ng of poly(A) RNA or 10–500 ng of rRNA-depleted total RNA	mRNA-seq, small RNA-seq



"MUCH LIKE IN DNA LIBRARY PREPARATION, THE MARKET FOR SHORT-READ RNA LIBRARY PREPARATION KITS IS HEAVILY FOCUSSED ON PRODUCTS COMPATIBLE WITH ILLUMINA SEQUENCERS."

RNA LIBRARY PREPARATION KITS FOR LONG-READ SYSTEMS

Unsurprisingly, Oxford Nanopore Technologies and PacBio are the key players in the long-read RNA library preparation market (see Table 2.4). Currently, Oxford Nanopore Technologies are the only providers of direct RNA sequencing, which allows the base modifications of native RNA molecules to be identified.

TABLE 2.4: DETAILS FOR SELECTED RNA LIBRARY PREPARATION KITS COMPATIBLE WITH LONG-READ SEQUENCING PLATFORMS.

Information was sourced from the supplier's website or handbook associated with the product. Details provided were correct at the time of publication (January 2024), please check supplier's website for the most up-to-date specifications and data.

Supplier	Kit	Compatability	Assay Time	Input Quality	Applications
Oxford Nanopore Technologies	cDNA-PCR Sequencing Kit V14	Oxford Nanopore: MinION Mk1B, MinION Mk1C, GridION, PromethION	~3.75 hours (excluding PCR)	10 ng poly(A)+ RNA or 500 ng total RNA	RNA-seq
	Direct RNA Sequencing Kit	Oxford Nanopore: MinION Mk1B, MinION Mk1C, GridION, PromethION	~1.5 hours	300 ng poly(A)+ RNA or 1 µg total RNA	RNA-seq without cDNA conversion, base modification analysis
PacBio	Iso-Seq express 2.0 kit and Kinnex full-length RNA kit	PacBio: Revio, Sequel II/IIe	1.5 days	300 ng total RNA	RNA-seq
	Kinnex single-cell RNA kit	PacBio: Revio, Sequel II/IIe	3 days	15 ng - 75 ng single-cell cDNA	Single-cell RNA-seq
	Iso-Seq express 2.0 kit and SMRTbell prep kit 3.0	PacBio: Revio, Sequel II/IIe	8 hours	300 ng total RNA	RNA-seq

The RNA-seq field had some impressive technological developments in 2023, starting with Oxford Nanopore Technologies announcing their newest Direct RNA Sequencing Kit (SQK-RNA004), complete with improved RNA sequencing technology. First announced at London Calling – the company's annual conference – in early 2023, the new kit promises significantly increased single molecule raw read accuracy and output, with additional improvements RNA modification detection[6].

However, Oxford Nanopore Technologies are not the only ones bringing innovation to the long-read RNA-seq field. PacBio also made waves in late 2023 with the release of their Kinnex kits, specifically designed for long-read RNA-seq at high resolution. The company have launched several new products for various key applications, including single cell, full-length RNA and 16S rRNA sequencing, which are compatible with both their Revio and Sequel II systems. The major advantage of the Kinnex system is increased throughput, with a documented 16-fold throughput increase using Kinnex products on the Revio instrument[7] without any detrimental effects on sequencing accuracy or quality.

Chapter 2 references

1. Hess, J. F. *et al.* **Library preparation for next generation sequencing: A review of automation strategies.** *Biotechnol. Adv.* **41**, 107537 (2020).
2. Socea, J. N., Stone, V. N., Qian, X., Gibbs, P. L. & Levinson, K. J. **Implementing laboratory automation for next-generation sequencing: benefits and challenges for library preparation.** *Front Public Health* **11**, 1195581 (2023).
3. Illumina Inc. **TruSeq™ DNA PCR-Free.** https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_dna_pcr_free_sample_prep.pdf (2017).
4. New England Biolabs. **New England Biolabs® launches new NEBNext UltraExpress™ DNA and RNA Kits for faster, easier NGS library prep workflows.** https://www.neb.com/en-gb/about-neb/news-and-press-releases/new-england-biolabs-launches-new-nebnext-ultraexpress-dna-and-rna-kits-for-faster-easier-ngs-library-prep-workflows (2023).
5. **PacBio announces collaboration with leading library preparation automation partners.** *PacBio* https://www.pacb.com/press_releases/pacbio-announces-collaboration-with-leading-library-preparation-automation-partners/ (2023).
6. **Oxford Nanopore announces breakthrough technology performance and pathway to enabling sequencing by anyone, anywhere in London Calling tech update.** *Oxford Nanopore Technologies* https://nanoporetech.com/oxford-nanopore-london-calling-23-technology-update (2023).
7. Sullivan, M. **Kinnex launch promises to revolutionize RNA research.** https://www.pacb.com/blog/kinnex-launch-promises-to-revolutionize-rna-research/.

YOUR GO-TO NGS PARTNER.

Ready for wherever you aim to go.

NGS LIBRARY PREP SOLUTIONS – FLEXIBILITY SETS BREAKTHROUGHS FREE.

Our NGS library preparation portfolio provides multiple solutions with proven technology and expertise from reagents, automation, and consumables to application support across different throughputs.

- **MagicPrep™ NGS System**
  Simplify your NGS workflows with library prep at the press of a button.
- **DreamPrep® NGS Compact & DreamPrep® NGS**
  Scale up a wide variety of NGS workflows with walk-away solutions.
- **Tecan's Reagents**
  Streamline your NGS workflows with various sample types.
- **Custom Systems for OEM Partners**
  Designed to your requirements and your regulatory needs.

No matter where you aim to go, our NGS library prep solutions are designed to enable your needs and accelerate your science. So leave the library prep to us – you're free to focus on your next breakthrough.

Learn more

www.tecan.com/ngslibraryprep

TECAN.

CHAPTER 3

MEET THE SEQUENCERS

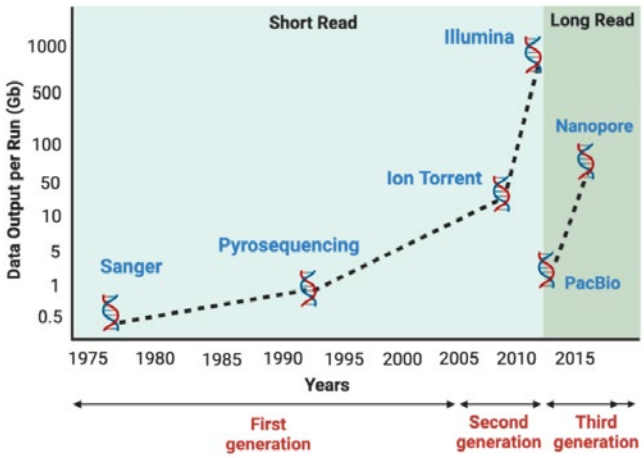
WHILE THE LANDSCAPE OF SEQUENCING HASN'T CHANGED DRAMATICALLY THIS YEAR, THERE IS A GRADUAL SHIFT IN MARKET SHARE AS NEW AND EXISTING PLAYERS SOLIDIFY THEIR POSITIONS.

For example, patents that Illumina has held for the majority of the sequencing era have ended, or are coming to an end, with some due to expire in 2024. This sets up the sequencing market to become even more competitive, and, ultimately, for prices to drop and the accessibility of sequencing to increase. This chapter presents our yearly review of the sequencing market, the sequencing platforms available and their specifications.

A Brief History of sequencing and sequencers

The history of sequencing and sequencers has seen a progressive improvement in the amount of DNA material that can be sequenced (see Figure 3.1) and the total cost per genome. This can be captured in three generational changes in sequencing chemistry.

FIGURE 3.1. THE AMOUNT OF DATA OUTPUT POSSIBLE BY EACH MAJOR TYPE OF SEQUENCING.
Image Credit: Satam, et al. [1]



FIRST-GENERATION

Determining the order of nucleic acid residues in biological samples is essential practice in healthcare, pharma and research. And the sequencing instrument is the most significant investment a lab will make in their journey to profile DNA and RNA in this way.

When reflecting on the original breakthrough Sanger sequencing, first deployed in 1972, we've come a long way. Sanger's chain termination methodology works by disrupting the DNA synthesis reaction using dideoxynucleotides with radioactive isotope labels. Gel electrophoresis can then determine the DNA sequence by the position of the electrophoretic band. This process allows 300-1000bp fragments to be read.

Sanger sequencing is still performed today and, while it has evolved considerably, the core principles remain unchanged. The major advantage of this method is the affordability and suitability for small-scale projects. This method tends to achieve a Q20 threshold, which is not as accurate as 2nd and 3rd generation sequencers, and the high labour and innability to scale makes it an unsuitable method for whole-genome level sequencing and large-scale projects.

SECOND-GENERATION AND THE COST OF A GENOME

Second generation sequencing, or next generation sequencing (NGS), really began with the Roche GS20 from 454 Life Sciences, released in 2005. This pyrosequencing NGS technology was the first platform in the market and greatly increased the amount of DNA that could be sequenced at once.

DNA synthesis releases pyrophosphate that then produces a light signal when converted to ATP. This pyrosequencing method worked by measuring the released light signal, which was proportional to the amount of pyrophosphate produced, and unique for each DNA base.

This technology allowed the sequencing of thousands to millions of DNA fragments. This, in turn, unlocked the ability to effectively sequencing whole genomes. While 2003 saw the first human genome sequenced (to 90% completion), 2008 saw the first NGS sequenced human genome, the genome of James Watson, which took just four months and under $1.5 million to sequence[3].

Several new methodologies for NGS were released over the subsequent years, including methods that detect the release of hydrogen ions as modelled by the Ion Torrent released in 2011, and the sequencing by synthesis (SBS) approach taken by the Illumina instruments, first released in 2007. By 2010, Illumina instruments had dropped the cost of sequencing the human genome to the $10,000 mark. By 2014, the launch of Illumina's HiSeq X Ten saw the first $1,000 dollar genome.

In late 2022 and early 2023, sequencing records were broken again. Both Illumina and Element Biosciences claimed the ability to produce a $200 genome on their new instruments. Ultima Genomics and MGI went one step further and claimed that their cost-saving sequencers could produce a genome for as little as $100. This progress massively outstrips Moore's Law (see Figure 3.2). However, it is worth noting that 'cost per genome' is a metric that can be misleading. While the HiSeq X Ten's $1,000 genome included DNA extraction and library prep etc., modern discussions of 'cost per genome' only takes into account the sequencing costs.

THIRD-GENERATION

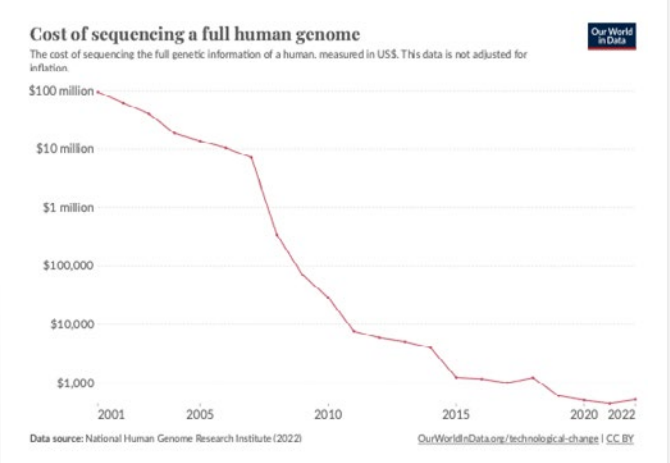
Finally, the third generation sequencing or single molecule sequencing platforms was first widely available through Pacific Biosciences (now PacBio) in 2011 with the single-molecule real time (SMRT) sequencing. Compared to second-gen, this sequencing does not require DNA amplification, and can sequence extremely long single molecules of DNA with a high level of accuracy. Oxford Nanopore Technologies commercialised a second methodology to sequence long-reads when they produced their first biological nanopore-based platform in 2014.

The utility of long reads will be highlighted below. But, through long-read nanopore technology, another milestone was reached in 2022 - the first complete human genome. Nanopore sequencing accounted for the missing 8% of the genome that short-read technologies cannot effectively sequence, but long-read technologies can[4].

The current state of affairs in 2nd and 3rd generation sequencing is by no means the peak. For the first time, sequencing costs are being driven to general affordability. Innovation and new sequencers will be coming. In fact, you can track the development of new sequencers here, at Shawn Baker's blog. A cursory look shows that several companies are working on their own spin on Illumina's SBS chemistry, and several more are working to produce their version of nanopore single molecule sequencing. Furthermore, you can see an NGS necropolis, which highlights how precarious the development of sequencers can be for companies.

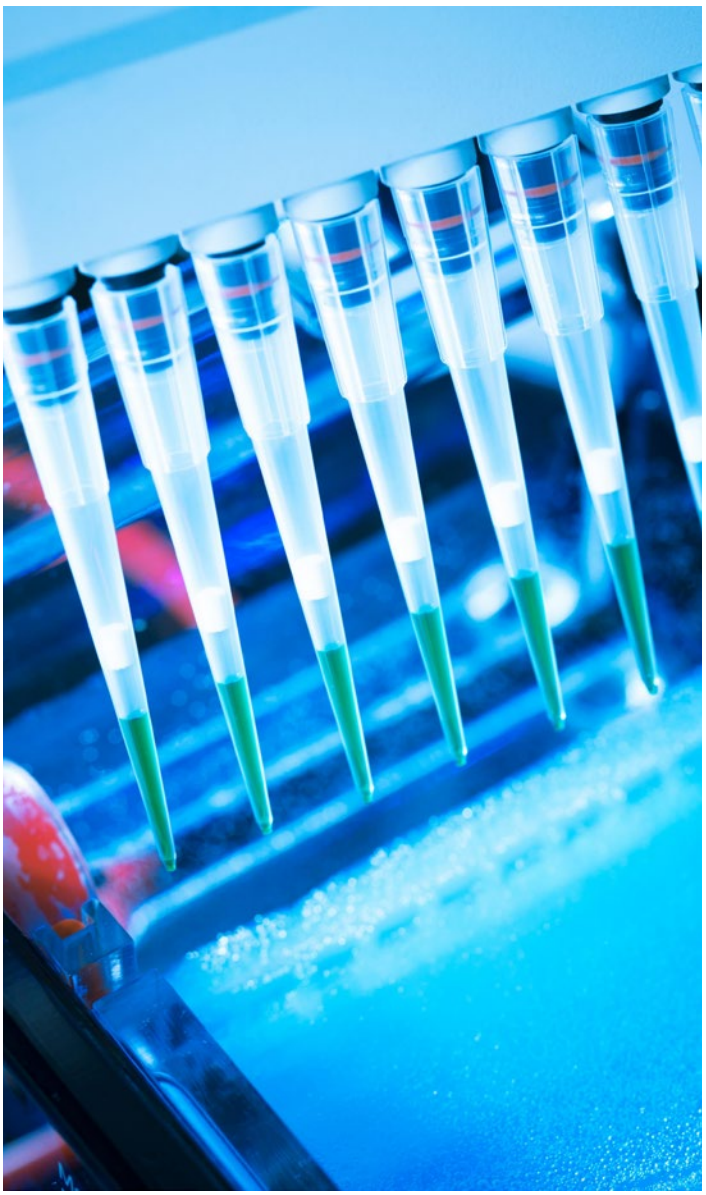
FIGURE 3.2. THE COST OF SEQUENCING A FULL HUMAN GENOME FROM 2001 TO 2022.

Image Credit: Our World in Data



Right now, the market has a varied (but limited) portfolio of both short-read (2nd Gen) and long-read (3rd Gen) sequencers. Here, we will review the instruments available for both. One thing that you may notice is missing from this portfolio is pricing. While some companies are very clear with their pricing, such as the transparent price list for Oxford Nanopore Technologies, other companies offer bespoke pricing and, in general, the lack of standardised pricing is an issue for purchasers of sequencing equipment. Furthermore, displaying prices for a sequencing run is heavily dependent on the library prep, how loaded the flow cell is, and how many runs are evaluated at once.

Short-Read Sequencers



For short reads, the principal chemistry in the market is Illumina's sequencing by synthesis (SBS) chemistry. This works by coupling the four DNA bases to fluorescent markers alongside a terminator chemical group that pauses DNA synthesis. While DNA is being synthesized, each fluorescent marker is optically verified before the tag and terminator are removed, and the next step in the sequence is recorded.

Of the other chemistries in the market, MGI is an established player, and uses a similar process (but one that is more labour intensive at lower costs) to convert DNA templates into arrays of DNA nanoballs. Another established chemistry is Thermo Fisher Scientific's Ion Torrent, which relies on ion semiconducting sequencing, a method that uses an ultrasensitive pH chip to detect hydrogen ions released when nucleotides are incorporated during DNA synthesis.

Of the newcomers, Singular Genomics' G4 is most similar to Illumina, but has a novel flow-cell design. This makes it much easier to run multiple sequencing experiments simultaneously, aiming to outperform Illumina when it comes to flexibility. Ultima Genomics' UG100 has a very distinctive flow cell, in which sequencing reagents are applied to the exposed surface of a spinning disc, centrifugally distributing the reagents evenly. This lowers cost and increases the capacity of their machine. Finally, Element Biosciences' AVITI uses sequencing by binding (SBB), by which fluorescently labelled nucleotides are not permanently incorporated into the newly synthesized DNA, but bind transiently. They are then imaged, washed and replaced by unlabelled nucleotides. This is a more natural DNA synthesis process.

Some of the strengths of short reads are the relatively high accuracy, the comparatively lower cost and the comparatively high throughput over long reads. This allows researchers to identify small genetic variations that may impact diseases like cancer, and to sequence multiple whole genomes, transcriptomes and exomes. However, an inherent limitation for sequencing shorter stretches of DNA using NGS is the need to fragment and amplify the short sequences. A continuous sequence must be assembled from the clone fragments. This can introduce biases into the samples and an insufficient overlap in fragments can occur, impairing the ability to assemble. Furthermore, large genetic alterations, such as inversions, translocations and indels, may be missed by short-read.

We will now review the specifications for the major short-read sequencers currently available to the reader.

MEET THE SEQUENCERS

Illumina: iSeq100™

Size	Benchtop (425 mm x 305 mm x 330 mm, W = 15.9 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	144 Mb - 1.2 Gb
Number of paired-end reads per run	Up to 8 million
Max read length	2 x 150 bp
Max run duration	19 hours
Quality score	80% bases higher than Q30 (PE150)
Typical applications	Small whole-genome sequencing, targeted gene sequencing, gene expression profiling

Illumina: MiniSeq™

Size	Benchtop (456 mm x 480 mm x 518 mm, W = 45 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	1.65Gb - 7.5Gb
Number of paired-end reads per run	14 million - 50 million
Max read length	2 x 150 bp
Max run duration	~ 24 hours
Quality score	> 80% bases higher than Q30 (PE150)
Typical applications	Small whole-genome sequencing, targeted gene sequencing, targeted gene expression profiling, 16S Metagenomic Sequencing

Illumina: MiSeq™

Size	Benchtop (686 mm x 565 mm x 523 mm, W = 57.2 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	540 Mb - 15 Gb
Number of paired-end reads per run	24 million - 50 million
Max read length	2 x 300 bp
Max run duration	24 hours (PE150), 56 hours (PE300)
Quality score	80% bases > Q30 (PE150), 70% bases > Q30 (PE300)
Typical applications	Small whole-genome sequencing, targeted gene sequencing, targeted gene expression profiling, 16S metagenomic sequencing.

Illumina: MiSeqDx™

Size	Benchtop (686 mm x 565 mm x 523 mm, W = 54.5 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	≥ 5 Gb
Number of paired-end reads per run	≥ 15 million
Max read length	2 x 150 bp
Max run duration	24 hours
Quality score	≥ 80% bases higher than Q30 (PE150)
Typical applications	In vitro diagnostic use. cystic fibrosis testing, IVD assay development, targeted enrichment and cancer companion diagnostics

Note: *Values above are for Diagnostic mode. In Research (RUO) mode, the Dx has the same performance specifications as the MiSeqDx™.*

Illumina: NextSeq™ 550

Size	Benchtop (533 mm x 635 mm x 584 mm, W = 83 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	16 Gb – 120 Gb
Number of paired-end reads per run	260 million – 800 million
Max read length	2 x 150 bp
Max run duration	29 hours
Quality score	> 75% bases higher than Q30 (PE150)
Typical applications	Small whole genome sequencing, small whole exome sequencing, targeted gene sequencing, transcriptome sequencing.

Illumina: NextSeq™ 550Dx

Size	Benchtop (540mm x 690 mm x 580 mm, W = 84.4 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	≥ 90 Gb
Number of paired-end reads per run	600 million
Max read length	2 x 150 bp
Max run duration	< 35 hours
Quality score	> 75% bases higher than Q30 (PE150)
Typical applications	In vitro diagnostic use. IVD assay development, targeted enrichment, non-invasive prenatal testing, comprehensive genomic profiling.

Note: *Values above are for Diagnostic mode. In Research (RUO) mode, the Dx has the same performance specifications as the NextSeq™ 550.*

Illumina: NextSeq™ 1000 & 2000

Size	Benchtop (550 mm x 650 mm x 600 mm, W = 141 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	10 Gb - 500 Gb
Number of paired-end reads per run	200 million – 3.4 billion
Max read length	2 x 300 bp
Max run duration	48 hours
Quality score	>= 85% bases higher than Q30 (PE150)
Typical applications	Small whole-genome sequencing, exome & large panel sequencing, targeted gene sequencing, targeted gene expression, chromatin analysis, methylation sequencing, single-cell profiling, transcriptome sequencing, metagenomic profiling, proteogenomics, 16S Metagenomic sequencing.

Illumina: NovaSeq™ 6000

Size	Production-Scale (800 mm x 945 mm x 1656 mm, W = 481 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	65 Gb – 6 Tb
Number of paired-end reads per run	1.3 billion – 40 billion
Max read length	2 x 250 bp
Max run duration	~ 44 hours
Quality score	≥ 85% bases higher than Q30 (PE150)
Typical applications	Large whole-genome sequencing, exome & large panel sequencing, single-cell profiling, transcriptome sequencing, chromatin analysis, methylation sequencing, metagenomic profiling, proteogenomics, cell-free sequencing & liquid biopsy analysis.

illumina®

New P4 flow cell on the NextSeq™ 2000...

...will enable the highest output and most cost-efficient runs on an Illumina benchtop platform.

Learn more

Illumina: NovaSeq™ 6000Dx

Size	Production-Scale (800 mm x 945 mm x 1656 mm, W = 481 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	1 Tb - 6 Tb
Number of paired-end reads per run	6.67 billion - 40 billion
Max read length	2 x 150 bp
Max run duration	≤ 45 hours
Quality score	≥ 85% bases higher than Q30 (PE150)
Typical applications	In vitro diagnostic use. Targeted enrichment, IVD assay development.

Illumina: NovaSeq™ X and X Plus

Size	Production-Scale (864 mm x 933 mm x 1588 mm, W = 532 Kg)
Sequencing chemistry	Sequencing by synthesis
Output per run	165 Gb – 16 Tb
Number of paired-end reads per run	3.2 billion – 104 billion
Max read length	2 x 150 bp
Max run duration	~48 hours
Quality score	≥ 85% bases higher than Q30 (PE150)
Typical applications	Large whole-genome sequencing, whole exome & large panel sequencing, single-cell profiling, whole transcriptome sequencing, chromatin analysis, methylation sequencing, metagenomic profiling, proteogenomics, cell-free sequencing & liquid biopsy analysis.

Note: *Values above are for Diagnostic mode. In Research (RUO) mode, the Dx has the same performance specifications as the NovaSeq™ 6000*

Thermo Fisher Scientific: Ion Torrent GeneXus™

Size	Benchtop (815 mm x 1065 mm x 1678 mm, W = 68 Kg)
Sequencing chemistry	Ion semiconductor sequencing
Output per run	15 - 60 Gb
Number of reads per run:	15 million - 60 million
Max read length	400 bp
Max run duration	24 hours
Quality score	> 99% bases higher than Q30
Typical applications	Whole genome sequencing, whole exome sequencing, targeted genome sequencing.

Thermo Fisher Scientific: Ion GeneStudio™ S5

Size	Benchtop (542 mm x 806 mm x 509 mm, W = 63.5 Kg)
Sequencing chemistry	Ion semiconductor sequencing
Output per run	10 Gb - 15 Gb
Number of reads per run:	60 million - 80 million
Max read length	600 bp
Max run duration	19 hours
Quality score	> 99% bases higher than Q30
Typical applications	Targeted DNA sequencing, exome sequencing, targeted RNA sequencing, targeted transcriptome sequencing, whole transcriptome sequencing.

Thermo Fisher Scientific: Ion GeneStudio™ S5 Prime

Size	Benchtop (542 mm x 806 mm x 509 mm, W = 63.5 Kg)
Sequencing chemistry	Ion semiconductor sequencing
Output per run	20 Gb - 25 Gb
Number of reads per run:	100 - 130 million
Max read length	600 bp
Max run duration	8.5 hours
Quality score	> 99% bases higher than Q30
Typical applications	Exome sequencing, targeted transcriptome sequencing, whole transcriptome sequencing.

MGI: DNBSEQ-E25

Size	Benchtop (348 mm x 312 mm x 257 mm)
Sequencing chemistry	DNA nanoball
Output per run	7.5 Gb
Number of paired-end reads per run	25 million
Max read length	2 x 150 bp
Max run duration	20 hours
Quality score	>80% bases higher than Q30
Typical applications	Plug-and-Play style operations, remote site sequencing, pathogen and microorganism identification, small WGS, targeted DNA/RNA panels.

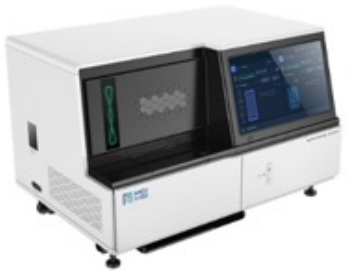
MGI: DNBSEQ-G50

Size	Benchtop (654 mm x 489 mm x 545 mm, W = 85 Kg)
Sequencing chemistry	DNA nanoball
Output per run	150 Gb (PE150-FCL)
Number of paired-end reads per run	100 million (FCS) - 500 million (FCL)
Max read length	2 x 150 bp
Max run duration	28 hours (PE150-FCS) - 40 hours (PE150-FCL)
Quality score	>80% bases higher than Q30 (PE150-FCL)
Typical applications	Small whole-genome sequencing, targeted DNA/RNA panels, lowpass whole-genome sequencing.

MGI: DNBSEQ-G99

Size	Benchtop (607 mm x 680 mm x 640 mm, W = 140Kg)
Sequencing chemistry	DNA nanoball
Output per run	96 Gb (PE300)
Number of paired-end reads per run	80 million x 2
Max read length	2 x 300 bp
Max run duration	12 hours (PE150), 30 hours (PE300)
Quality score	>85% bases higher than Q30 (PE300)
Typical applications	Small whole-genome sequencing, targeted DNA/RNA panels, lowpass whole-genome sequencing, transcriptome sequencing.

MGI: DNBSEQ-G400



Size	Benchtop (1086 mm x 756 mm x 710 mm, W = 200 Kg)
Sequencing chemistry	DNA nanoball
Output per run	1,440 Gb (PE200-FCL)
Number of paired-end reads per run	1.8 billion x 2
Max read length	2 x 300 bp (FCS)
Max run duration	37 hours (PE150-FCS), 109 hours (SE400-FCL)
Quality score	>85% bases higher than Q30 (PE150-FCS)
Typical applications	Whole-genome sequencing, whole exome sequencing, transcriptome sequencing.

MGI: DNBSEQ-T7

Size	Production-Scale (1656 mm x 903 mm x 1815 mm, W = 765 Kg)
Sequencing chemistry	DNA nanoball
Output per run	7 Tb
Number of paired-end reads per run	5.8 billion x 4
Max read length	2 x 150 bp
Max run duration	< 24 hours
Quality score	> 85% bases higher than Q30 (PE150)
Typical applications	Deep whole-genome sequencing, deep exome sequencing, transcriptome sequencing, targeted panel projects.

MGI: DNBSEQ-T10x4RS

Size	Production-Scale (7200 mm x 5000 mm x 1950 mm, W = 10,000 Kg)
Sequencing chemistry	DNA nanoball
Output per run	76.8 Tb (PE150)
Number of paired-end reads per run	32 billion - 45 billion x 8
Max read length	2 x 150 bp
Max run duration	106 hours (PE150)
Quality score	>85% bases higher than Q30 (PE150)
Typical applications	Large-scale population studies, ultra-high-depth whole genome sequencing.

MGI: DNBSEQ-T20x2

Size	Production-Scale (4200 mm x 4800 mm x 2000 mm, W = 3700 Kg)
Sequencing chemistry	DNA nanoball
Output per run	72 Tb (PE150)
Number of paired-end reads per run	40 billion x 6
Max read length	2 x 150 bp
Max run duration	80 hours (PE150)
Quality score	>80% bases higher than Q30 (PE150)
Typical applications	Large-scale population studies, ultra-high-depth whole genome sequencing.

Singular Genomics: G4

Size	Benchtop (982 mm x 583 mm x 813 mm, W = 135 Kg)
Sequencing chemistry	Rapid sequencing by synthesis
Output per run	Up to 480 Gb
Number of reads per run:	Flexible; 200 million - 3.2 billion (F2: 200M reads, F3: 400M reads, F3 + Max Reads: 800M reads, per FC) 4 FC can be loaded and sequenced in parallel
Max read length	2 x 150 bp
Max run duration	30 hours
Quality score	80 – 90% bases higher than Q30
Typical applications	RNA gene expression; single cell RNAseq; total RNA-seq; exome sequencing; target enrichment, whole genome sequencing, shotgun metagenomics.

PacBio: Onso

Size	Benchtop (940mm x 686mm x 762mm, W = 123.3 Kg)
Sequencing chemistry	Sequencing by binding
Output per run	120 - 150 Gb (PE150)
Number of paired-end reads per run	800 million – 1 billion (PE150)
Max read length	2 x 150 bp
Max run duration	48 hours (PE150)
Quality score	≥ 90% of bases Q40
Typical applications	Cancer research, gene editing, whole exome sequencing, single-cell analysis.

Element Biosciences: AVITI System

Size	Benchtop (295mm x 376mm x 285mm, W = 155.1 Kg)
Sequencing chemistry	Avidity Base Chemistry (ABC)
Output per run	300 Gb (600 Gb - dual flow cell run) (PE150)
Number of reads per run:	1 billion
Max read length	2 x 150 bp or 2 x 300 bp
Max run duration	38 hours (PE150 at 1B reads), 60 hours (PE300)
Quality score	>90% bases higher than Q30 (PE150)
Typical applications	Single-cell RNA sequencing, whole genome sequencing, whole exome sequencing, targeted sequencing, SARS-CoV-2 sequencing.

Note: *AVITI LT is an alternative version of the system that runs low and medium throughput kits and Elembio Cloud is an online platform to monitor sequencing runs with real-time updates*

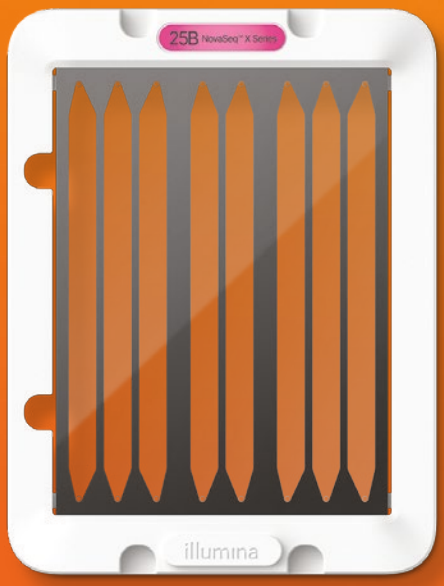
Ultima Genomics: UG100™

Size	Production-Scale
Sequencing chemistry	Mostly natural sequencing-by-synthesis
Output per run	2.4 Tb per wafer (1 x 300 bp)
Number of reads per run:	~ 10 billion
Max read length	1 x 300bp
Max run duration	<20 hours
Quality score	>85% bases higher than Q30
Typical applications	Single-cell RNA sequencing, whole genome sequencing, whole exome sequencing, multi-omics, clinical

Note: *Details acquired from non-peer-reviewed publication from the Ultima Genomics team[5]. Public launch of the UG100 occurred in February 2024*

illumina®

Scale up your sequencing with the new 25B flow cell



Exclusively on the NovaSeq™ X Series

Learn more

Short-read market overview

2022 saw major short-read releases, such as the Illumina NovaSeq X Series and PacBio's Onso. 2023 might be seen as the year of delivering promises, contemplating 2022's releases, and incremental small gains.

Illumina continues to dominate the short-read market. In the last decade it was estimated that over 90% of the world's sequencing data is generated on Illumina machines. One estimate in 2022 shows it still to be over 90%. The recent release of the 25B flow cell for the NovaSeq X Series creates a new kind of scale, 26 billion reads per run at a cost of $0.64 dollar per million reads. One strength for Illumina is their breadth of offerings, covering the highest throughput instruments (summarized above) to low-cost instruments such as the iSeq.

Challengers who have risen over the last year and a half (e.g., Element Biosciences and Ultima Genomics) are carving a small space in the market and beginning to sell sequencers. Ultima Genomics is still operating in relative stealth (having only publicly launched their sequencing platform in February 2024), after their ground-breaking announcement of the $100 genome, but are delivering units selectively (see here for a very recent blog giving insider information on the UG100). Element Biosciences are now on the second generation of their chemistry – Avidity Cloudbreak - that runs 20% faster with improved data quality at the end of each read or greater accuracy and early indexing for real-time sequencing run management. . They have achieved the $200 genome and have passed 100 commercial orders on their instrument, as of September 2023.

At the end of 2022, Pacific Biosciences (PacBio), who have traditionally developed long-read technologies (see next section), announced the Onso, their first short-read sequencer. It uses sequencing by binding (similar to AVITI), in which fluorescently tagged nucleotides are not directly incorporated into the newly synthesized strands on the flow-cell. The first Onso systems were shipped in Q3 2023, and feedback will soon be available.

Long-Read Sequencers

Third generation (or long-read) sequencing focuses on reading DNA sequences far longer than the 25 bp – 200/300 bp that are currently the standard for short-read. Long-read could sequence reads from 5 kb (5000 bp) up to 4 mb (4,000,000 bp) in length. This offers a different angle than short-read when it comes to sequencing.

By sequencing a single long molecule, you overcome the amplification bias in short reads and create better length of overlap for more comprehensive sequence assembly. Furthermore, it is estimated that long-read sequencing gives you access to an additional 5% - 10% of the genome that could never be unambiguously addressed with short-read alone, although these returns diminish after ~20Kb reads (see this blog). This allows for better overall resolution of highly repetitive genomic regions and allows for the assembly of large and complex genomes.

Downsides to long-read sequencing are mainly associated with the cost and the accuracy per read. Per genome, the cost of long-read sequencing is still around five times more expensive than short-read (e.g., it's $995 on PacBio's Revio Sequencer). We expect that cost to come down, but it may always be the case that long reads cost more. Errors are also more present in long-read technologies than their short-read comparisons, as we will discuss below.

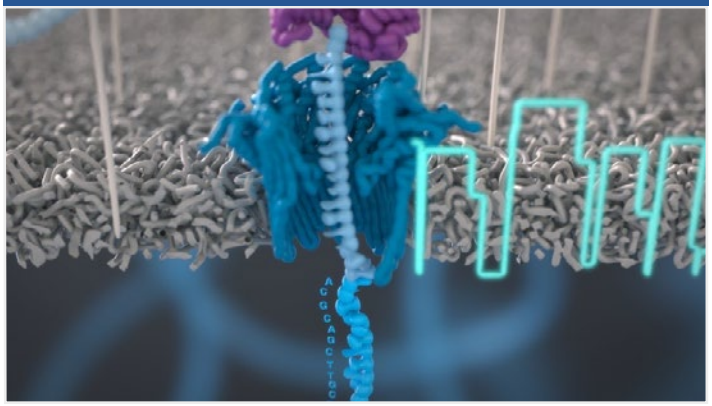
As has already been mentioned, two sequencing providers currently provide the main long-read sequencers in the market.

PacBio are the proprietors of the Single Molecule Real-Time (SMRT) sequencing technology. This technology is based on a single DNA polymerase attached to a zero-mode waveguide chip (a nanostructure for fluorescence detection). Within each zero-mode waveguide chip, two adapters are attached to both ends of the DNA molecule to form a circular single-stranded structure. DNA polymerase is used to sequence a complementary strand and the fluorescence is measured to identify the corresponding nucleotides.

SMRT technology is able to obtain data from 10kb up to 25kb in length. The latest platform, Revio, is capable of producing 360Gb of HiFi reads per run, which takes under 24 hours, the equivalent of 1,300 human genomes a year.

MEET THE SEQUENCERS

FIGURE 3.3. DEPICTION OF NANOPORE SEQUENCING IN ACTION.
Image Credit: Oxford Nanopore Technologies.



Oxford Nanopore Technologies (ONT) are the dominant nanopore long-read sequencing company and have a variety of sequencers in their arsenal. Nanopore sequencing works by measuring the disruption in electrical current across a membrane when a specific nucleotide passes through a protein nanopore (see Figure 3.3). Each nucleotide has a distinct signal, and this means that sequencing can occur in real time as each base is decoded from the disturbance.

This rapid, accurate mode of sequencing has seen nanopore tailor their offerings across a range of applications, including providing the only set of portable sequencers that can be used for fieldwork and on-site pathogen detection. Nanopore sequencing does have a high error rate due to the inability to control the speed of DNA molecules passing through the pore. SMRT sequencing has completely random errors through the DNA polymerase becoming damaged through use, while nanopore's are higher and systematic.

Finally, several companies are set to enter into the long-read market, specifically through nanopore sequencing approaches. Qitan Technologies has produced two nanopore sequencers, the QNome-3841 and the QNome-3841hex, currently only available in China. Furthermore, companies such as Armonica Technologies, Genvida, INanoBio and Nooma Bio are all currently developing nanopore style sequencers. This all suggests that the long-read market will begin to see more competition, and hopefully more innovation.

The specifications for the main long read sequencing platforms currently available are below/on the next page.

PacBio: Sequel II / IIe

Size	Production-Scale (927mm x 864mm x 1676mm, W = 362Kg)
Sequencing chemistry	Single molecule real-time (SMRT)
Output per run	~30 Gb HiFi
Number of HiFi reads per run	Up to 4 million
Max read length	15 kb – 20 kb
Max run duration	Up to 30 hours
Quality score	HiFi reads >99% accuracy
Typical applications	Whole genome sequencing, RNA sequencing, targeted sequencing, complex populations, epigenetics.

PacBio: Revio

Size	Production-Scale (927mm x 914mm x 1745mm, W = 465 Kg)
Sequencing chemistry	Single molecule real-time (SMRT)
Output per run	90 Gb x 4
Number of reads per run:	6 million - 10 million
Max read length	20 kb - 25 kb
Max run duration	24 hours (15 kb -20 kb), 30 hours (20 kb -25 kb)
Quality score	99.95% bases >= Q33 (15 kb - 20 kb)
Typical applications	Human genome analysis, single-cell transcriptome analysis, small genome analysis, targeted sequencing, epigenetics.

Oxford Nanopore Technologies: MinION MK1B

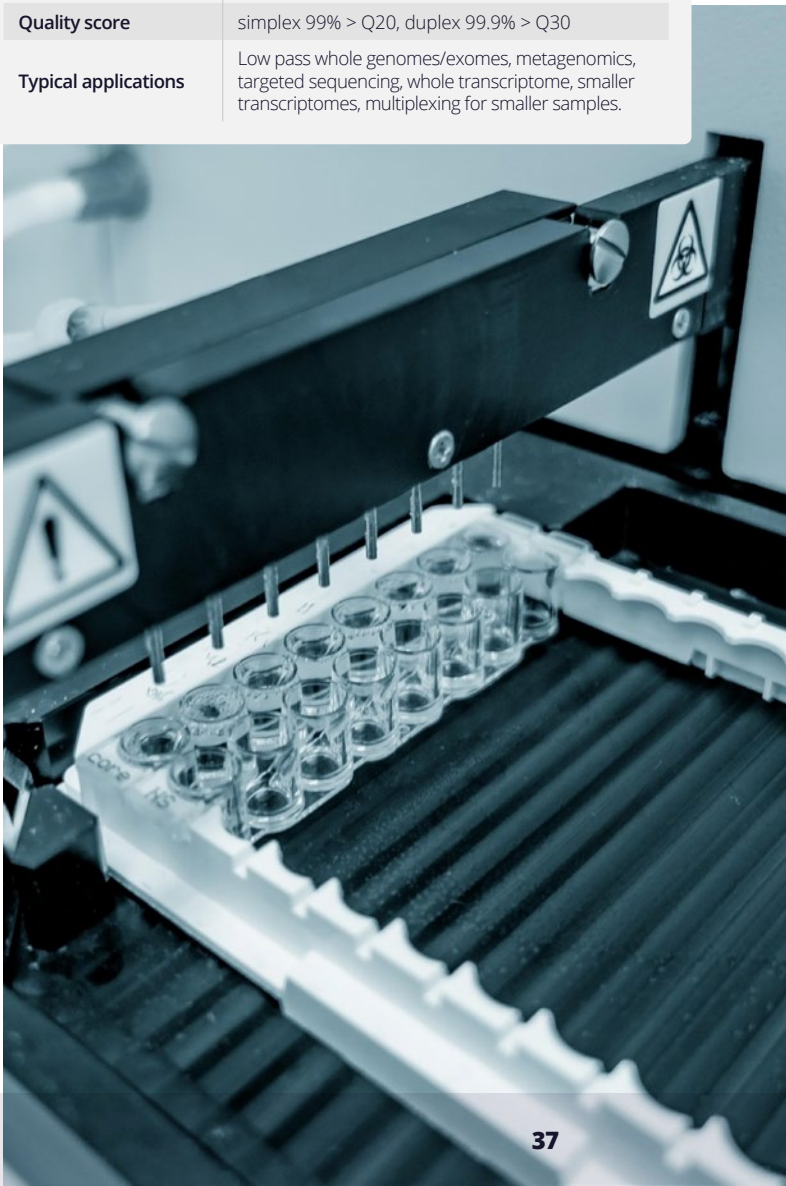
Size	Portable (105 mm x 23 mm x 33 mm, W = 0.087 Kg)
Sequencing chemistry	Nanopore
Typical output per run	35 Gb
Number of reads per run:	Use case dependent
Max read length	20 bp – 4 Mb
Max run duration	Up to 72 hours
Quality score	simplex 99% > Q20, duplex 99.9% > Q30
Typical applications	Low pass whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome, smaller transcriptomes, multiplexing for smaller samples.

Oxford Nanopore Technologies: MinION MK1D

Size	Portable
Sequencing chemistry	Nanopore
Typical output per run	35 Gb
Number of reads per run:	Use case dependent
Max read length	20 bp – 4 Mb
Max run duration	Up to 72 hours
Quality score	simplex 99% > Q20, duplex 99.9% > Q30
Typical applications	Low pass whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome, smaller transcriptomes, multiplexing for smaller samples.

Oxford Nanopore Technologies: GridION

Size	Benchtop (364 mm x 220 mm x 360 mm, W = 11 Kg)
Sequencing chemistry	Nanopore
Typical output per run	175 Gb
Number of reads per run:	Use case dependent
Max read length	20 bp – 4 Mb
Max run duration	Up to 72 hours
Quality score	simplex 99% > Q20, duplex 99.9% > Q30
Typical applications	Low pass whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome, smaller transcriptomes, multiplexing for smaller samples.



Oxford Nanopore Technologies: PromethION 2/PromethION 2 Solo

Size	P2 Solo = Portable (110mm x 87mm x 152mm, W = 1.5Kg)
Sequencing chemistry	Nanopore
Typical output per run	400 Gb
Number of reads per run:	Use case dependent
Max read length	20 bp – 4 Mb
Max run duration	Up to 72 hours
Quality score	simplex 99% > Q20, duplex 99.9% > Q30
Typical applications	Human, plant and animal whole genome sequencing, large genome assembly, targeted sequencing, transcriptomics, single cell, highly multiplexed sequencing, high depth metagenomics, epigenetics.

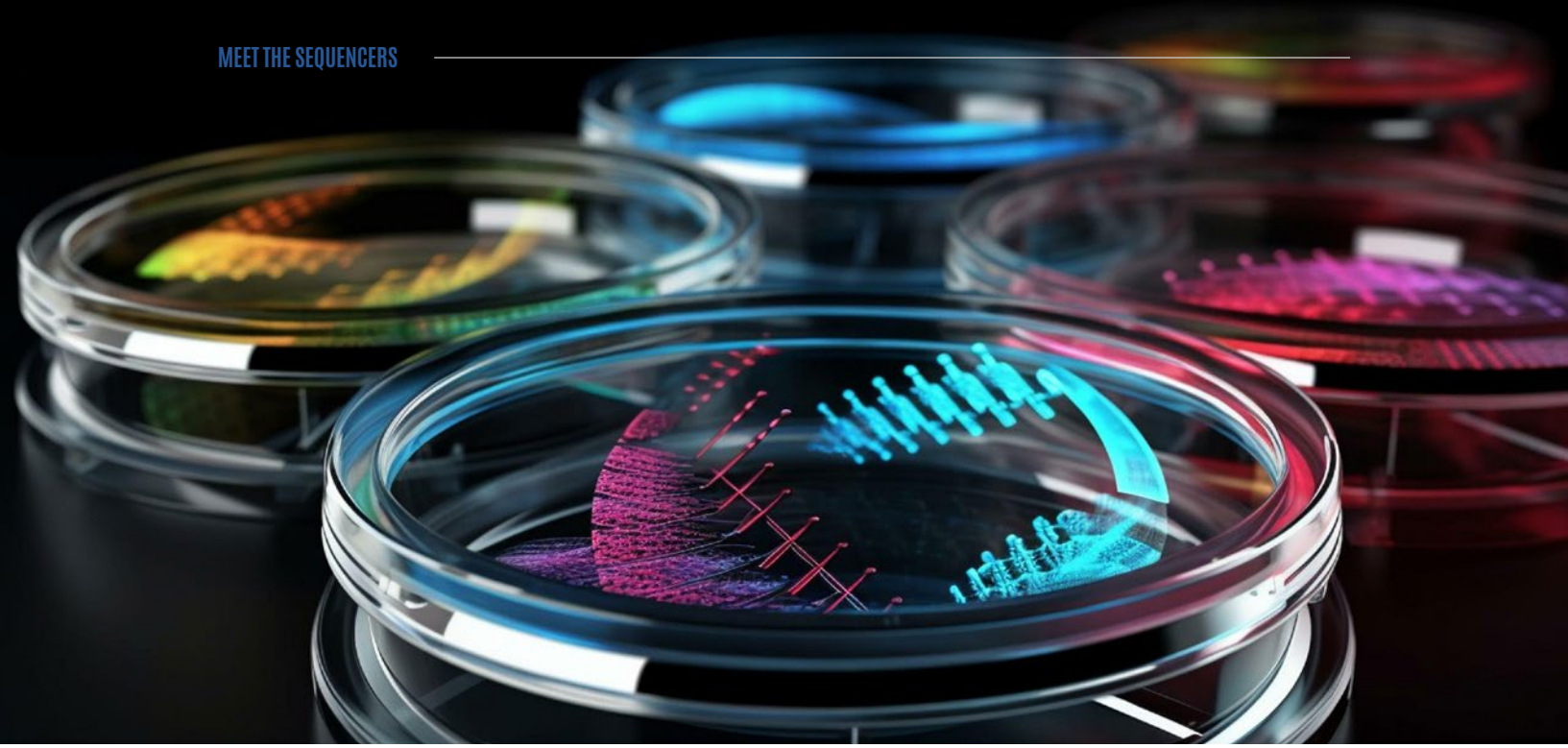
Oxford Nanopore Technologies: PromethION 24

Size	Benchtop (Sequencing module - 590 mm x 190 mm x 430 mm, W = 28 Kg, Computing module = 178 mm x 440 mm x 470 mm, W = 25 Kg)
Sequencing chemistry	Nanopore
Typical output per run	4.8 Tb
Number of reads per run:	Use case dependent
Max read length	20 bp – 4 Mb
Max run duration	Up to 72 hours
Quality score	simplex 99% > Q20, duplex 99.9% > Q30
Typical applications	Human, plant and animal whole genome sequencing, large genome assembly, targeted sequencing, transcriptomics, single cell, highly multiplexed sequencing, high depth metagenomics, epigenetics.

Oxford Nanopore Technologies: PromethION 48

Size	Benchtop (Sequencing module - 590 mm x 190 mm x 430 mm, W = 28 Kg, Computing module = 178 mm x 440 mm x 470 mm, W = 25 Kg)
Sequencing chemistry	Nanopore
Typical output per run	9.6 Tb
Number of reads per run:	Use case dependent – Theoretical max – 261 million
Max read length	20 bp – 4 Mb
Max run duration	Up to 72 hours
Quality score	simplex 99% > Q20, duplex 99.9% > Q30
Typical applications	Human, plant and animal whole genome sequencing, large genome assembly, targeted sequencing, transcriptomics, single cell, highly multiplexed sequencing, high depth metagenomics, epigenetics.

"THIS RAPID, ACCURATE MODE OF SEQUENCING HAS SEEN NANOPORE TAILOR THEIR OFFERINGS ACROSS A RANGE OF APPLICATIONS, INCLUDING PROVIDING THE ONLY SET OF PORTABLE SEQUENCERS THAT CAN BE USED FOR FIELDWORK AND ON-SITE PATHOGEN DETECTION."



Long-read market overview

This year has seen PacBio and Oxford Nanopore Technologies both increase their market share, representing the increasing interest in long-read technologies. Both have new models under development. PacBio only recently released the Revio, but are already producing an ultra-high throughput production-scale sequencer and a lower throughput benchtop model. Oxford Nanopore Technologies has two new products under development, the MinION Mk1D and the SmidgION. The latter would be the smallest sequencing device so far and is designed to be used with a smartphone anywhere.

Furthermore, short-read focused companies have produced long-read kits to enable the benefits of longer reads on their platform. Illumina launched Complete Long Reads for NovaSeq in March 2023. This kit tagments long-single-molecule fragments and can generate contiguous long-read sequences around 5-7kb in length, with some reads greater than 10kb. Element Biosciences also have the LoopSeq™ for the AVITI™, released late in 2022, which barcodes longer sequences (up to 5kb) before sequencing. This expands the use of AVITI to applications where longer reads are valuable, such as understanding microbial diversity, viral genomes and the immune repertoire. MGI's MGIEasy stLFR kit does much the same as LoopSeq and allows long read information to be gained from their short-read platforms through barcoding.

With the successful crossover from Illumina, PacBio, MGI and Element Biosciences into both short and long reads, a balanced and flexible sequencing offering from companies could be the future of the sequencing market.

Chapter 3 references

1. Satam, H. *et al.* **Next-Generation Sequencing Technology: Current Trends and Advancements.** *Biology* **12**, 997 (2023).
2. Sanger, F., Nicklen, S. & Coulson, A.R. **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* **74**, 5463-7 (1977).
3. Wadman, **M. James Watson's genome sequenced at high speed.** *Nature* **452**, 788-788 (2008).
4. Nurk, S. *et al.* **The complete sequence of a human genome.** *Science* **376**, 44-53 (2022).
5. Almogy, G. *et al.* **Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform.** *bioRxiv*, 2022.05.29.493900 (2022).

MIRO CANVAS

NGS Prep System

Advances in NGS have made it possible to sequence entire genomes far faster and more cost effectively than with previous sequencing methods. However, manual sample and library preparation for NGS workflows is time consuming and error prone, limiting lab productivity and throughput. INTEGRA Biosciences has launched **MIRO CANVAS** – a compact digital microfluidics platform for fully automated NGS sample preparation – to help accelerate genomics discoveries. This intuitive system offers full automation to help scientists in academic, research and clinical laboratories translate discoveries into solutions.

Full walk-away automation of NGS prep and target enrichment protocols

Setting up a run on **MIRO CANVAS** takes only 15 minutes. Simply pipette sample and reagents into the MIRO cartridge and walk away with confidence. The system fully automates all necessary steps, including thermal cycling and magnetic bead operations. Samples can be processed as they are received, without the need for batching, which can be critical for clinical applications.

One system for multiple NGS platforms

MIRO CANVAS provides pre-installed and verified NGS sample preparation protocols designed to offer accuracy, precision and reliability for both short- and long-read sequencing platforms, including PacBio and nanopore sequencing. The gentle liquid handling technology used in the instrument is particularly suited to preserving long DNA fragments, ensuring high quality end results for any sequencing application.

Up to 75 percent reduction in reagent usage

MIRO CANVAS uses digital microfluidics technology to move droplets within the cartridge over an array of electrodes. This allows reagent volumes to be reduced by up to 75 percent for lower running costs.

Minimal training required

The large, full-colour, onboard touchscreen of the **MIRO CANVAS** guides you through setting up a run and, when ready, the system will prompt you to move onto the next step.

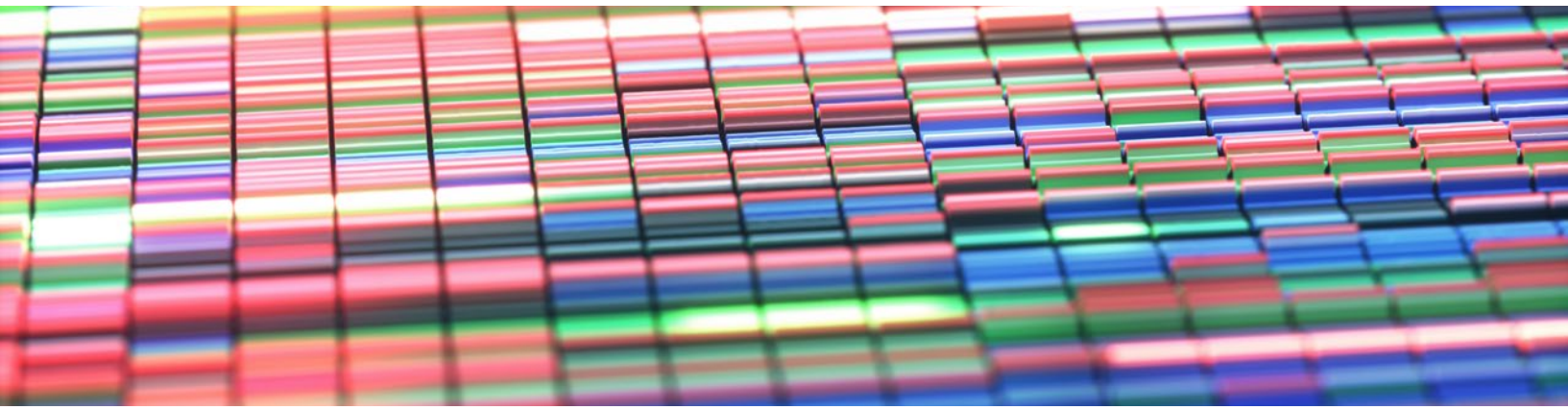
To learn more, visit **https://www.integra-biosciences.com/en/ngs-automation/miro-canvas**

INTEGRA

CHAPTER 4

FINDING YOUR SEQUENCING TECHNOLOGY

THE TECHNICAL SPECIFICATIONS OUTLINED IN THE PREVIOUS CHAPTER ARE ONLY HALF THE STORY WHEN IT COMES TO CHOOSING A SEQUENCING TECHNOLOGY. YOU ALSO NEED TO MATCH YOUR APPLICATION AND NEEDS TO THE VARIOUS SEQUENCERS THAT ARE AVAILABLE. SHOULD YOU USE LONG-READ OR SHORT-READ? DO YOU NEED HIGH-THROUGHPUT OR HIGH ACCURACY? THIS CHAPTER WILL HELP GUIDE YOU THROUGH THIS DECISION-MAKING PROCESS



Short or long-read?

The sequencing market is essentially divided into two types of instruments, those that sequence short-reads (up to 600 bp in length) and those that can sequence much longer genomics reads (up to 4 mb). For many applications (such as genome assembly), both types of sequencing are valid to use, but it is important to consider the advantages of each type and whether it is worth investing in the latest short-read or long-read platforms.

A valid question to ask yourself is whether your genomic regions of interest are readily accessible with short-read sequencing? If you want to investigate diseases that require the analysis of repetitive or GC-rich regions, or necessitate phasing, modern long-read platforms may provide answers. If your genomic areas of interest are suitable for sequencing by short-read technologies, it may be more cost-effective to utilise a short-read approach instead.

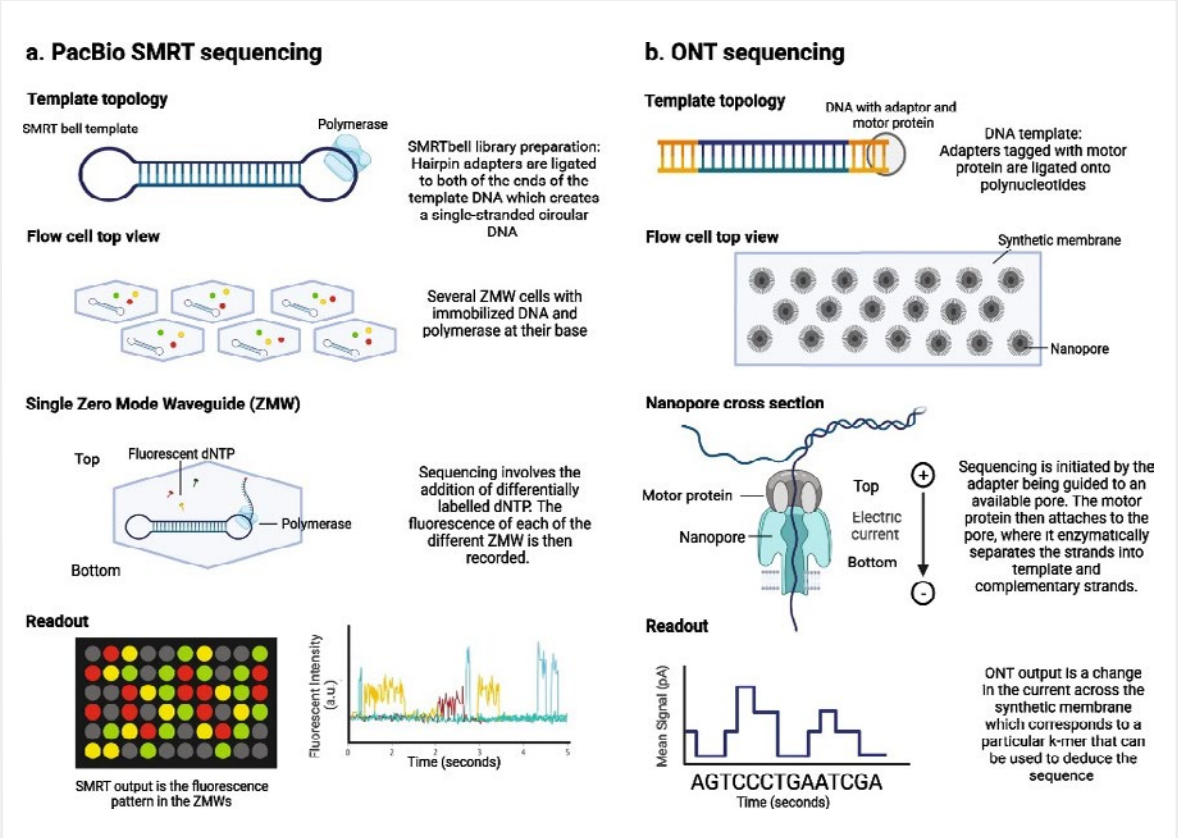
As it currently stands, it is about 5 times cheaper to sequence a human genome with the latest short-read technology compared to the latest long-read platforms ($200 on the Illumina NovaSeqX/Element Biosciences' AVITI vs. $995 on PacBio Revio). However, long-read sequencing does generally provide more information than short-read approaches. This raises a question within the genomics community - if long-read got cheap enough, could it replace short-read as the default sequencing approach?

Long-read sequencing has come a long way since its launch, and was recently announced as the Nature Method of the year for 2022[1]. The two types of long-read technology (see Figure 4.1) are now at the stage where they are enabling researchers to explore genomes at unprecedented resolutions. This has led to direct improvements for genome assemblies, such as the new, complete human genome that was published in 2022[2], and the draft human pangenome in 2023[3]. For these kind of applications (accurate, large-scale genome assembly), long-read is gaining prominence[4] and readers should refer to this review that highlights the variety of applications that long-read sequencing has improved when compared to existing short-read technology[5].

In clinical and research spaces, it is important to not use long-read sequencing for the sake of using a popular platform that may not suit your sequencing requirements. You need to ask yourself – are you going to get novel information, or a more accurate diagnosis, using this chosen platform? Your sequencing approach should be selected or updated based on the expected results, with the approach you've selected providing demonstrably better outcomes. PacBio have set up a consortium to ask what additional diagnostic yield is gained from using long-read in clinical genetics, and what does short-read sequencing miss? Furthermore, a recent review on the topic recognizes the unique advantages of long-read sequencing in the clinic, noting its use for precise SNP detection and haplotype phasing, whilst also noting the limitations of applying it to all situations[6].

If your application involves attaining perfect genome coverage, many groups are now considering using a hybrid approach of both long-read and short-read sequencing. You can leverage the lower cost per base, high depth and high quality short-read data and then layer on top the long-read data to resolve complex structural variants and phase haplotypes. This is a hugely beneficial approach for de novo genome assembly or rare disease variant sequencing.

FIGURE 4.1. SCHEMATIC DIAGRAM OF A PACBIO SMRT SEQUENCING AND B OXFORD NANOPORE TECHNOLOGIES.
Image Credit: Oehler, et al.[6]



Recently the 'perfect' bacterial genome was assembled using both Oxford Nanopore Technologies and Illumina sequencing[7]. Compared to the previous 'short-read first' approaches[8,9], the improvements we've seen in long-read yield and accuracy encouraged researchers to adopt a long-read first approach. This meant making a long-read-only assembly and then polishing with short reads to produce a 100% accurate genome (see Figure 4.2).

Choosing between sequencers

FLG: What factors matter when choosing a sequencing technology for a project?

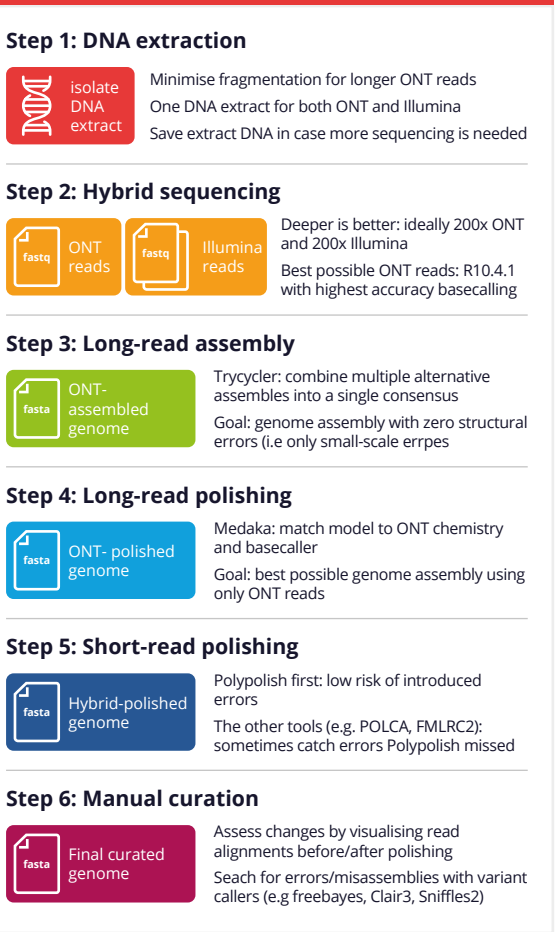
Xinkun Wang
Director, NUSeq Core Facility, Center for Genetic Medicine & Research
Associate Professor, Department. of Cell & Developmental Biology
Northwestern University

To choose an appropriate technology, the needs of the project should be thoroughly evaluated first. This process should encompass various technical aspects, including required data throughput, read length and sequencing accuracy. Other technical aspects, such as physical separation of lanes, the ability to run multiple flow cell types and sequencing run time, may also be important factors to consider. As sequencing technologies continue to evolve, when comparing different technologies, besides considering their current capabilities we also try to assess their potential for future advancements. Financial factors to consider include costs for initial instrument acquisition, sequencing reagents and instrument maintenance (including service contract). As run problems and instrument breakdowns inevitably happen, the technology provider's technical support ability should also be factored in. We need a support team that can get problems solved quickly when they arise.

David Baker
Head of Sequencing
The Quadram Institute

Think about what similar researchers in your field are using in your field. Using new technologies may cause issues with informatics too. Saying that, I'm a believer that it's too easy to go for tried and tested technology, and I always try to contemplate switching. Never listen to the vendors. Go to the users for their experiences.

FIGURE 4.2. ILLUSTRATED OVERVIEW OF THE RECOMMENDED APPROACH TO PERFECT BACTERIAL WHOLE-GENOME ASSEMBLY FROM WICK, *ET AL.*[7]
Image Credit: Wick, et al.[7]



APPLICATION

In previous versions of this guide, sequencing applications have consistently emerged as the biggest consideration when choosing a sequencing platform. Alongside whether long-read or short-read is most appropriate for your project (see above), there are further questions to ask; do you need a high throughput per run, specific read-length, paired-end sequencing or extreme read accuracy? All these considerations help shape your decision.

If high throughput is important then production-level sequencers should be considered. While they are expensive and much larger than their benchtop counterparts, they are essential for large-scale genomics work, and these instruments tend to form the backbone of core sequencing facilities. However, for most groups, benchtop and portable sequencers are much more appealing for everyday use.

Competitors to Illumina in the short-read market have tried to distinguish themselves by application. Singular Genomics' G4 sequencer is focused on flexibility, with a flow-cell design that makes it easy to run multiple distinct sequencer experiments in parallel, which may be valuable for certain use-cases.

Element Bioscience's AVITI has exceptionally low error rates, low costs and flexible applications, putting itself forward as the all-round benchtop sequencer for those looking for a broad range of smaller scale uses. Furthermore, its workflow is compatible with Illumina's (while others including MGI and Ultima Genomics require extra-steps).

Ultima Genomics deploys a chemical process that promises to be significantly cheaper than competitors. Recent data has shown that the sequencing performance from the UG100 was 'very similar' to Illumina sequencing for a single-cell sequencing experiment[10]. The suggestion is that the UG100 could be particularly useful for cost-effective, large-scale sequencing/scRNA-seq projects, beneficial for those looking for high-throughput applications.

For long-read sequencing, PacBio's HiFi process allows for higher accuracy than Oxford Nanopore Technologies. The process reads the same segment of DNA multiple times, ironing out random base-calling errors. This tends to be at the drawback of higher costs and lower throughput (although Revio addresses both these concerns).

Oxford Nanopore Technologies sequencers, on the other hand, offer versatility and portability, which can flexibly be applied to very short and extremely long reads. They are cheaper to run than PacBio (and most short-read sequencers of similar capacity), have a higher throughput, and can be used in the field. However, the error rates from Oxford Nanopore Technologies sequencers are higher, although computational solutions exist to counter this (see below).

Recent comparisons of long-read methodologies for specific purposes are available[11]. If high accuracy, longer read lengths and complex genomic variations are the primary focus, PacBio sequencing is often preferred. However, if rapid turnaround time, real-time analysis, portability and lower input material requirements are critical, Oxford Nanopore Technologies sequencing can be the more suitable option.

READ ACCURACY

The second biggest consideration in choosing a sequencer is read accuracy. This measure of error rate can impact the utility of sequencing data including coverage uniformity, the mappability of reads and the phasing of genome copies. There are two types of accuracy for sequencing – read accuracy (the inherent error rate for a read) and consensus accuracy. The latter is built through combining information from multiple reads in a dataset, so deeper coverage generally equals a higher consensus accuracy.

When evaluating sequencers, read accuracy is what matters. All sequencers perform to a minimum standard, but some sequencers are more geared towards read accuracy than others. PacBio's Onso and Element Biosciences' AVITI claims an error rate of Q40[12], equivalent to one error in 10,000 bases. Illumina and most other short-read sequencers regularly achieve Q30 (1 in 1000).

Long-read technologies tend to have higher error rates, but company-led improvements in accuracy and modern self-correction methods[13-15] have brought this error rate closer to short-read levels - with PacBio's HiFi technology meeting the 1 in 1000, Q30 error rate and Oxford Nanopore Technologies achieving Q28.

COSTS

There is a continued demand for more transparency regarding the complete costs associated with running a sequencer. Oxford Nanopore Technologies are particularly transparent with their costings, but they are the exception, not the rule. Once the above considerations about application and suitability of the sequencer for your project have been evaluated. The next consideration is whether the sequencer is worth the monetary investment, and whether you could get a similar quality performance from a cheaper, alternative, sequencer.

FINDING YOUR SEQUENCING TECHNOLOGY

When weighing up sequencing costs, it's important to think about the different types of expenses rather than just the upfront price of the sequencing instrument. Below we outline some questions to ask when it comes to properly working out the cost of running a sequencer:

Operational

- What is the cost per sample?
- How much are consumables and library prep kits?
- Production-scale instruments tend to only be cost-effective when they are running at almost full capacity, can you meet that?

Hands-on Labour

- How labour-intensive is the process? The more hands-on time a sequencing protocol needs, the more it costs in time and money.
- Will users need specialist training to use the instrument?

Ancillary Equipment

- Will you need additional equipment or sequencer add-ons?

Data costs

- Will the data be stored on-site or on the cloud?
- Will additional computational equipment be needed to complete the experiment?
- Will data analysis need to be outsourced or can it be completed in-house?

Furthermore, it can be expected for owners to pay 10% of an instruments' base cost every year for service contracts[16], which needs to be kept in mind if the budget is being stretched to account for the upfront sequencer cost alone.

XINKUN WANG

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology, **Northwestern University**

FLG: How does upkeep vary between sequencing technologies?

Xinkun: *While we have recently adopted multiple sequencing technologies, our upkeep experience is mostly on Illumina sequencers - although upkeep does not usually vary much between technologies. Such upkeep tasks mostly include instrument washes to maintain top working condition, monitoring of run performance and software updates. Keeping good records on loading conditions of different library types and sources (if made elsewhere) and managing sequencing reagent inventory may also be considered as part of the upkeep process. Service agreements cover preventative maintenance, so it is not usually performed by facility staff.*

BIOINFORMATICS CAPABILITY

One huge piece of advice consistently given out by sequencing operators is to understand your informatics requirements before you conduct your experiments. It will not only tell you whether you have the computational capacity to perform your experiment (see Chapter 6: NGS Data Analysis and Data Management) but this will also shape your sample collection process and sequencer selection. Some sequencing and third-party companies provide companion software for their sequencers alongside support and advice for handling the data, which could be an important consideration before making a purchase.

CONVENIENCE

Convenience could be a consideration for many people. If you are familiar with one type of sequencer, you might not want the potential hassle of learning a new system. Furthermore, it is worth considering whether the sequencer you are looking at is 'plug in and play'. Or whether you want a sequencer with better performance and a steeper learning curve for the workflow. For example, MGI's DNA nanoball technology tends to produce higher quality data than the comparable Illumina sequencer. On the flip side, the preparatory process for MGI sequencers tends to be much more labour intensive.

TRUST

The value of trust and reputation is sometimes understated. For many people using and buying sequencers, the high performance of Illumina and the continual innovation (bolstered by the fact that 90% of the world's sequencing is performed on Illumina platforms[16]) will play a big factor in deciding to stay with Illumina platforms. Newer companies may offer better niche uses or a novel and interesting sequencing chemistry, but for many people, this has not been enough to shift them away from Illumina instruments for short-read sequencing.

For example, a recent poll from the ASeq Newsletter substack saw that 72% of people wouldn't leave Illumina for an instrument that was identical but 10% cheaper per run. This changed to 76% willing to swap if the instrument was 50% cheaper. This highlights the trust that Illumina have built within the sequencing market.

XINKUN WANG

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology, **Northwestern University**

FLG: When would it be time to upgrade/buy a newer sequencer? If money was no object, what sequencer would you buy, and why?

Xinkun: *The best time to upgrade or acquire a newer sequencer is when it is needed. If money was no object, another good time is after its release, although being an early adopter is not without risk because a brand-new machine usually has bugs that the manufacturer is not aware of. Strategically, it might be worthwhile to wait to get the instrument once the bugs have been removed by the first wave of early users.*

As to what sequencer I'd buy now if money was no object, this would be those that can produce large volumes of long reads. At the current time, this is both the Oxford Nanopore PromethION 48 and PacBio Revio systems. We have projects that require both of them, and they are complementary since the P48 has higher throughput while the Revio has better accuracy.

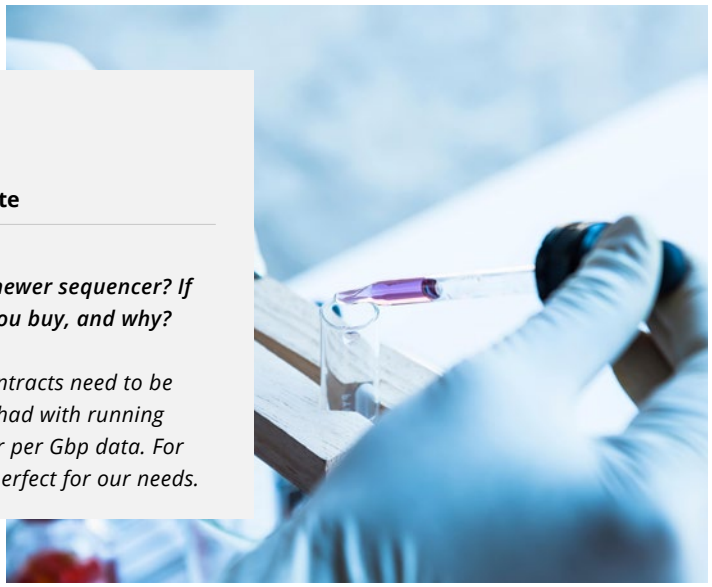
"AS TO WHAT SEQUENCER I'D BUY NOW IF MONEY WAS NO OBJECT, THIS WOULD BE THOSE THAT CAN PRODUCE LARGE VOLUMES OF LONG READS. AT THE CURRENT TIME, THIS IS BOTH THE OXFORD NANOPORE PROMETHION 48 AND PACBIO REVIO SYSTEMS."

DAVID BAKER

Head of Sequencing
The Quadram Institute

FLG: When would it be time to upgrade/buy a newer sequencer? If money was no object, what sequencer would you buy, and why?

David: *These things aren't cheap and service contracts need to be considered too. There is always a balance to be had with running internal sequencing and outsourcing for cheaper per Gbp data. For my current throughput, I find the Nextseq2000 perfect for our needs.*



We conclude this chapter with a panel-style discussion between several of our sequencing experts. Within, they discuss several topics including: which new short-read platform looks most exciting, at what point might long-read overtake short-read, and what considerations and trade-offs should be taken into account when deciding which sequencing platform to invest in. The discussion was between:

- ***Deanna Church,*** who is an Executive in Residence at a company called General Inception that partners with scientific founders to build companies.
- ***Miten Jain,*** who is an Assistant Professor at Northeastern University, and his work focuses on developing innovative sequencing technologies and analysis methods to decode the complexities of the genome. His group is developing methods for ultra long DNA, direct RNA and single-cell applications. The overarching goal for the group is to combine genome, transcriptome and proteome analysis methods to help improve our understanding of genome structure and function.
- ***Winston Timp,*** who is an Associate Professor at Johns Hopkins University. He focuses on the development and application of sequencing technologies to gain a deeper understanding of biology and a more accurate set of clinical tools for human disease. He has a keen interest in gene regulation and expression, sequencing the genomes and transcriptomes of a variety of organisms, from SARS-CoV-2 to hummingbirds, to Giant sequoias, as well as humans. His most recent work includes epigenetic investigation of the T2T human genome and characterisation of mitochondrial disorders.

> A RECENT POLL FROM ASAQ NEWSLETTER SUBSTACK SAW THAT 72% OF PEOPLE WOULDN'T LEAVE ILLUMINA FOR AN INSTRUMENT THAT WAS IDENTICAL BUT 10% CHEAPER PER RUN."

PANEL DISCUSSION :

THE NEW AND WHO OF ADVANCED SEQUENCING TECHNOLOGIES - 2023

THE CONTENT USED HERE IS A SHORTENED, EDITED TRANSCRIPT FROM A SESSION AT THE FESTIVAL OF GENOMICS BOSTON 2023

Deanna Church
Independent Consultant and Executive-in-Residence, **Dmchurch Bio, llc and General Inception**

Miten Jain
Assistant Professor
Northeastern University

Winston Timp
Associate Professor
John Hopkins University

Deanna Church: Both Miten and Winston have recently been involved in a publication demonstrating scalable approaches for using Nanopore sequencing to profile haplotype resolved variation and methylation and are at the forefront of developing sequencing technologies and analysis methods. So, I think that there will be a lot of information to be had here.

I'm just going to jump right in. Both of you, in particular, tend to focus on the development of long read sequencing technologies. But recently, there has been an uptick in some short read technologies. ***And I'm wondering, of the new short read platforms, which are the ones that you guys think are most exciting?***

Winston Timp: That's a great question. The reason that I'm most interested in short read, just as a side note, is for things that are counting applications. Long read is great, and there have been a lot of innovations from both Oxford Nanopore Technologies and PacBio to up their counts, but we're still talking tens or hundreds of millions of reads. Short read, whether it's Illumina or Element Biosciences or Singular Genomics, often generates billions of reads. And lot of applications like CUT&RUN, CHIP-seq, RNA-seq, ATAC-seq, some of these other tools are really about counts.

I like the kind of instruments that can give me lots of short-ish reads, 50 base long reads. I'm not interested in those middle-of-the-road, couple of hundred base long reads. I want short reads, but lots of them. And I want it to be frictionless for my lab. I want my lab to be able to do it. So, I'm looking at these benchtop, portable, relatively low cap investment, short read platforms that can generate billions of reads. What we've been looking at a lot right now in my lab is the Element Biosciences AVITI instrument, which can generate around a billion reads for about $1,000. We have also been assessing the Singular Genomics instruments and the NextSeq, but I'm most excited right now about Element Biosciences.

Miten Jain: I haven't tested anything in our lab. But much like Winston said, the molecular counts are important. We're a little bit different from their group; we do use short reads to test some of the finesses in accuracy that we might miss from the long nanopore reads that we work with, and we haven't worked directly with it, but our collaborators at Google Health have been testing the Element Biosciences AVITI platform. The results that I've seen have been very impressive, and that's the one I'm looking forward to.

Deanna Church: Great, thank you. I'm going to ask maybe a provocative question. You guys made the point about these counting applications and short reads being very useful. But we've also heard a lot about the advantages of long reads for genome analysis. And long reads have been around for quite a while. ***But I would still say the vast majority of genomes are sequenced using short reads. So, what do you think is the inflection point? Or what needs to happen to make long reads more accessible for more people?***

Winston Timp: Frankly, Deanna, I think we hit it this year. The fact that it's now about $1,000 to generate something like a 30x coverage genome from either Oxford Nanopore Technologies or PacBio, means that it's now tractable. And that's what we're starting to see. For example, the 'All Of Us' consortium has a long read element, and the Broad, U-Dub, Baylor and ourselves are generating long read data for All Of Us. And there's other work, such as the work that you were referring to that Miten and I are involved in from CARD (NIH Intramural Center for Alzheimer's and Related Dementias), that's generating hundreds of samples. That's only now possible. In the old days of PacBio (RSII) and the initial MinION, the yield, cost and accuracy were all issues, but now we're at that inflection point, and I think that we're starting to see the data be generated. So, the papers will take a while to come out, but I think we've just recently reached an inflection point where we can now apply this at scale. And I'm excited about it!

Miten Jain: I agree, I think scale is what really changes the game. For years, it was a boutique application. Suddenly, with the advent of Revio and the PromethION, any lab can start generating these genomes. And not one at a time, tens of samples at a time or tens of samples per week, at a scale that's quite affordable. So, suddenly, you have this influx of data, and the discussion has moved from 'can I generate long reads?' to 'how do I make use of long reads?' And some of the bigger consortiums, like CARD and All Of Us, are really leading in terms of recognising that value.

Part of the inflection comes from adoption. Whether it was Pangenome and CARD initially, and now All of Us, a large part of that was being ready to swiftly move and adopt the ultra-long methods. And also realising that the information that comes from those is significantly comparable, if not better, than from all other methods. Most of the genome inference that is happening - you do get all the information with methylation; you do get all the substantially improved information with phasing. Variant calling has been one of the areas where suddenly long reads are overtaking short reads. With PacBio, certainly there have been areas where you actually can use long reads alone, essentially using HiFi for assembly and then HiFi for polishing.

So, the fact that all of these assimilate in this timeline, coupled with the adoption from some of these larger groups that are doing production-scale sequencing, hits that tipping point, and it hopefully has happened. I agree with Winston that we think it's happened. But we are not the target consumer, in the sense that we already work with these methods. So, the hope is that this inflection point that we think is being hit is going to really tip over for most other users over time, and hopefully over the next year.

Deanna Church: I'm gonna hit you guys with a follow up question. I would say that algorithms haven't actually changed that much for assembly during all of this time. And I agree with you, in fact, in that I think in terms of data production and quality, that tipping point is there, but are the tools available to let most people use them? ***You guys are at the forefront of this approach, and you can whip up your own tools as you need them. But are the tools ready and fully-baked for someone who is not an aficionado to start using them?***

Winston Timp: Fully-baked is a very strong word! I'd say they're half-baked. To keep the baking analogy, you don't have to go and make the cake from scratch. You can buy a cake mix from the store, but you still have to add the eggs, the oil, set your oven to 350. It's not like you're just buying the cake from the store and taking it home. There's still stuff to do. But arguably, you still have to do that for short read interpretation anyway. To take the analogy too far, it's a complicated cake mix from the store, it's angel food cake or something hard. But you're not making it from scratch, we're not on Great British Bake Off!

We have help. And I think that these tools are advancing rapidly. I would direct your attention to, for example, the LRGASP work by Angela Brooks, where they're trying to assess what you can do with long cDNAs and transcriptomic work. The tools are rapidly maturing for doing epigenetics from long reads, both from PacBio and Oxford Nanopore, and for variant calling, things like PEPPER-Margin-DeepVariant are maturing rapidly. There are more and more cake mixes coming into the market all the time, and people are getting used to using them and figuring out how to do it.

Miten Jain: I agree, the methods are coming. There's a lot of interest from the developer community to develop some of these new innovative methods, or taking an old algorithm that can be adopted now that the data can be handled. The key is, again, making it production scale. So, are the methods there? Yes. Are they there where advanced users can use them? Certainly. In the CARD pipeline, for example, we've seen that some of this methodology is being adopted for larger scale projects. However, it is still developers using those. Where that tipping point changes, again, is in maturity and time. The example I've used in the past is GATK, a fantastic initiative that the Broad Institute did. But it took a good few years to mature to a level where it became, quote-unquote, the pipeline to use with Illumina data. With long reads, I think we are starting to get in that range, we just need a little bit more time for these tools to mature and then have a set of tools for assembly and analysis methods over time.

"MOST OF THE GENOME INFERENCE THAT IS HAPPENING - YOU DO GET ALL THE INFORMATION WITH METHYLATION; YOU DO GET ALL THE SUBSTANTIALLY IMPROVED INFORMATION WITH PHASING."

Deanna Church: I think you read my mind, because my follow up question was going to be 'what is the GATK for long reads?' But I think you just addressed that, in terms of the amount of time and effort. And I do think it's exciting that many of these projects are now adopting long reads because I think we got things like GATK, and many of the tools we use, from projects like the 1000 Genomes Project. And so, having these large projects can really help push forward the technology.

Winston Timp: What's also interesting about some of these large projects is that they're coming along at the same time as tools like AnVIL, and similar platforms. So, as the workflows are being developed, we're a little bit more mature than I would argue we were in the GATK days, so it'll be a little easier for people to follow along in the, now well-trodden, path.

Deanna Church: Well, there's been a lot of blood, sweat and tears when understanding how to analyse the genome with short reads, and I think the long read community is also benefiting from that.

Miten Jain: To add to Winston's point, the discussion is now starting to move into not only methods, but the scale of data and the scale of computing that's required for these applications. And that's not trivial by any measure, we really are getting to that point where genomics is generating data that's orders of magnitude higher than most other internet applications. That's an element that, pre-long read era, was not as much of a concern, but now with the volume and the scale, it is also becoming a concern. So, software development has to take that into account over time.

Deanna Church: I think that's a great point. I'm going to move on to what is possibly one of the questions I hear most often from people. ***If you're starting a project, how do you think about the trade-offs regarding which sequencing technology you use?*** There are cases where you might have a mix of long reads and short reads, how do you think about that trade-off with the number of samples, the complexity of analysis? What are the things you think about in your lab when you're setting up a project like that?

Miten Jain: It's complicated, because everything has a different requirement. If money is no concern, then the right answer is what Adam Philippy calls the 'kitchen sink approach.' You get everything from everything, and it's fantastic. But clearly, from an academic perspective, and for most projects, there is a concern about scale. Then it comes down to, what's the ultimate goal? If the goal is to get a genome assembly with as high a quality as possible, then in our group, at least, we'll think more along the lines of Nanopore Ultra-Long and HiFi, or Nanopore Ultra-Long with duplex Nanopore data. If we want to do something more in the transcriptome space, then we'll go with one long read platform, which in our case happens to be Nanopore, and then something on the short read side like Illumina or Element Biosciences. It really comes down to the fact that necessity drives the requirements and drives the innovation.

Most of the time, we recommend to our collaborators that if you were to do a brand-new project, and the scope is somewhat manageable, to go for a long read platform coupled with a short read platform. And then you have to consider what scale you need. If you need to do 10 genomes, then that's one thing. If you need to do 100 genomes, then you will have to recalibrate your expectation as to what you can actually do. In which case, you think about what devices you have access to, the timeframe of the project itself and how much money you can spend.

Thankfully, some of these points are starting to be less of an issue. For example, with long reads, PromethION and REVIO can both do lots of genomes very quickly at a similar enough cost. Short reads have more competition, which is good from a researcher's standpoint. But a lot of times, it's really a more bespoke and custom thought process as to what the project is, coupled with what you will do for that.

Winston Timp: I have a slightly different answer. I figure out the simplest thing that will answer the question. A lot of the time people approach me and say, 'I want to do long reads.' And I say, 'No, you don't. You want Illumina 3-prime EST tags to tell you gene level information,' or, 'You want qPCR.' Let's not forget that qPCR is still a thing, and it's a $3 assay. We don't need to always choose the most complicated thing.

So, the simplest thing that will answer the question is where I start and, as Miten said, it really depends on your question. It depends on what you're doing. It also depends on the sophistication of the analysis you're going to need to do, and if you're going need to develop a whole new analysis method to do this, when there's already something that exists in the short read space that will answer your questions. For example, if you're doing SNP calling, you could use PacBio's HiFi, which gives extremely accurate reads. But you could also just do Illumina and put it in the GATK and get your SNPs and get a VCF out. Do you need to make your life more difficult, submit it to a queue, figure out the high molecular weight, get through the REVIO data? It will be beautiful data, but if you're just calling germline variation, maybe you don't need it.

But, of course, being a method developer, in my own lab we're usually doing the most complicated things because I want to do something that's new and fun. There's a balance there. Oftentimes, it really depends on what you want out of it, you've got to ask what is the easiest, simplest assay or test that will answer your biological question of interest, and go from there.

Miten Jain: One thing that I'll add, and this is something that we've started doing more in the last year and a half, is also to tell collaborators that generating data is the easiest part of the problem these days. For the most part, if you are doing conventional sequencing, generating data is not a problem, both in terms of scale and cost. Because one way or the other, it's figured out. Winston's right in that you have to consider how you handle the data, if the methods are in existence, and, even if they are, is the research group you are collaborating with able to work with those? That's an important consideration, because you can't just give somebody 20 terabytes of data and say, 'Here's the tool, go do it.' Most people are not accustomed to that either, because not everything is that well established.

Deanna Church: I love those answers. I would just modify Winston's a little bit - if somebody just wants to do SNP calling, I don't know that short read is necessarily the go-to, because I think it may depend on where you want to SNP call.

Winston Timp: There's TaqMan assays for qPCR too. I'm not trying to be a qPCR salesman here!

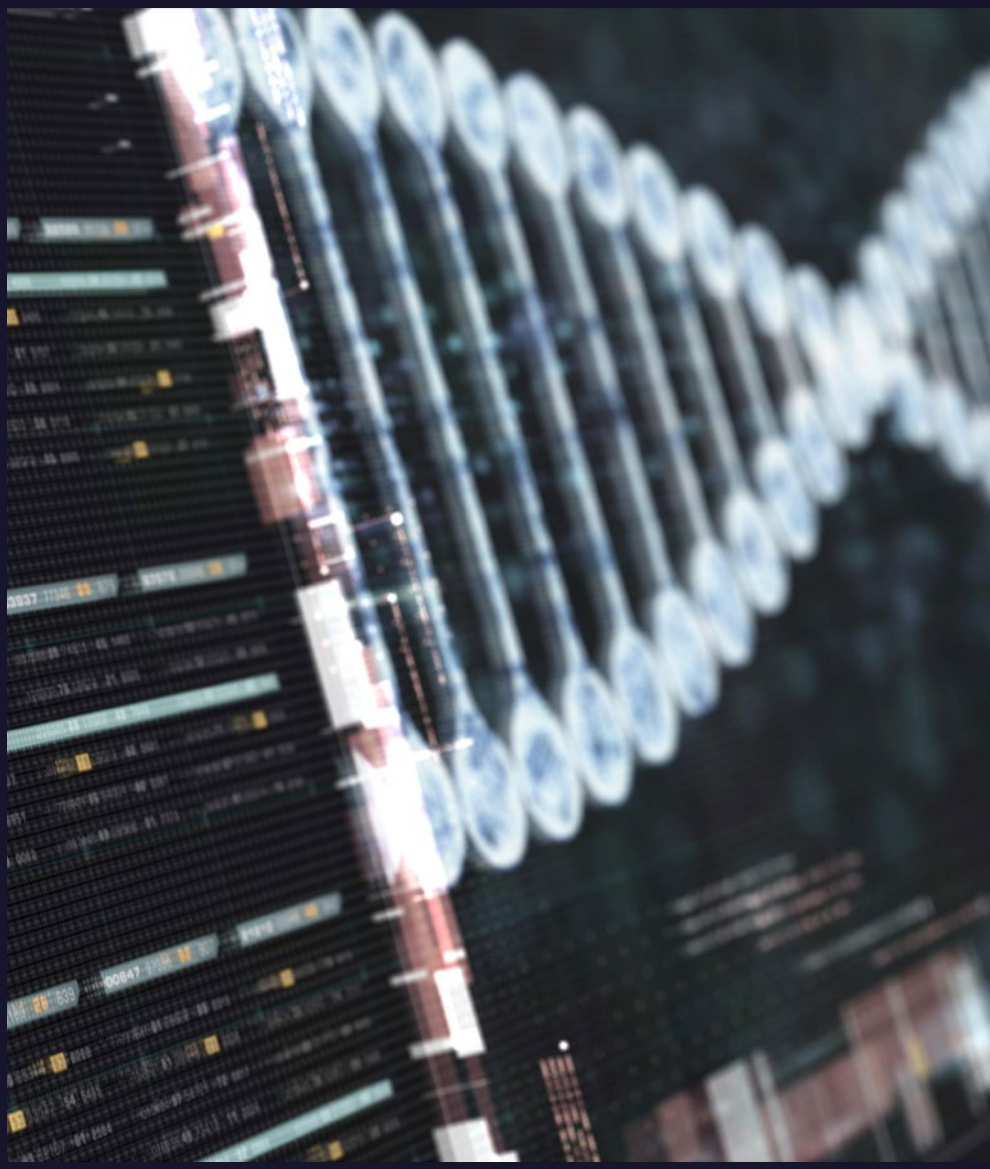
Deanna Church: Well, you're doing a good job! But there are also regions of the genome that, even if you want to do SNP calling, they're not going to be very accessible to short reads.

Winston Timp: 100%. But then you get those regions and the annotation and the interpretation, which is something I'm very interested in, may not be complete enough. Back to Miten's point, what are you going to do with the data once you get it? So, you get this SNP in this previously undiscovered area of the human genome...now what? And if you have a 'now what?' then definitely do that long read method. But if you don't have a 'now what?', well, think it through to the end.

Miten Jain: Just to quickly add to that, when we look at variant calling methods, we will always have that easy-to-map and hard-to-map comparison. The reason that easy-to-map exists is because you could do it any other way, for the most part. And that's I think what Winston's point is - if that's the goal, then it really is more about what you can do fastest and easiest.

Deanna Church: I 100% agree. I think I was more just trying to be provocative to make people think about these decisions about what technology to use. You want to be cognizant of both what you can do and what you can't do, right? Everything is going to be a trade-off and that's okay. But keeping in mind what you can't do is pretty important.

I DO THINK IT'S EXCITING THAT MANY OF THESE PROJECTS ARE NOW ADOPTING LONG READS BECAUSE I THINK WE GOT THINGS LIKE GATK, AND MANY OF THE TOOLS WE USE, FROM PROJECTS LIKE THE 1000 GENOMES PROJECT."



Your last answer, Miten, leads into the next question. I think one of the things that is very interesting in the sequence developer space is all these benchtop instruments. There's really a push towards not sending your samples out for sequencing and having a sequencer in your lab and doing it yourself, and these benchtop sequencers enable that. What are the things you think someone looking to buy one of these should consider? Because historically, I feel like when we talk about sequencing, we do talk about the data generation being the easy part. Now, the other parts are not really talked about. ***So, what are the things that people should really consider before they go all in and buy a benchtop sequencer?***

Miten Jain: Infrastructure, certainly. Especially as you go into more of the long read methods with Nanopore and PacBio, if you have the infrastructure in terms of, as the data stream out, having a place to store them and a place to analyse them. Now, taking a step back, in terms of being able to run these sequencing platforms, there are things that are very standard. For example, with Nanopore, 30KB or 50KB read in or where distributions are around 40-50KB length are trivial. You can do it any day in the lab, walk in and do the experiment - fine. If that's the goal, it's easy to make a choice, pick a PromethION or P2 Solo or what have you.

If your goal is to do ultra-long sequencing, then that is something you will need to get some training in. You'd have to be willing to get the expertise in your hands. Can it be done? Yes. Does every lab have that, a-priori? Possibly. I would say it's something that comes with experience in training.

In terms of getting the benchtop devices, you have to consider the infrastructure itself; can you handle it? But also consider if you are able to run those devices well, and then what's your amortisation cost. Some of these sequences require a fair bit of investment upfront, some of the others don't as much. So, the barrier to entry is low. Depending on that, and what your scales are, with some of the short read platforms, historically, you have to feed the beast. If you buy a nice benchtop sequencer, it's great. If you can't keep it running, then that poses a problem. As a result, you have to also be mindful of how much you need. But for the most part, at least from our group's perspective, given that we have been a Nanopore developer and technology innovator group for a long time, we do like the idea of having the ability in our own hands. Most of our students and staff can think about innovative experiments, execute them at the benchtop, take the data, look at the computational analysis, and then feed the intuition both ways on the wet lab side and the dry lab side. And there's something to be said about being able to operate in that space.

So, from my perspective, I typically am one of those people who will say that you should have a sequencer in your lab, because you will be able to do things that you are otherwise not able to. And you always have access to a core if you need large scale stuff.

Winston Timp: I would agree. One of the first purchases in my lab was an Illumina MiSeq, and that's because I wanted it to be frictionless for students to be able to try things. Students don't want to drop off to the core, figure out how to submit the forms, go to the drop-box and put the samples in. They're more willing to just try a run, and then they also will understand more about how the sausage is made. So, the data flows faster. I think that, again, with innovations in targeted sequencing approaches, with these platforms, you can generate this data really quickly and easily. I was selling qPCR machines before, but now I'm selling tabletop sequencers. Why just do qPCR when you could get genome-wide data on the same question? That kind of thing is powerful.

And I'm not necessarily thinking about using the PromethION or whole genome sequencing. I'm thinking about targeted or specific assays. I'm thinking about the MinION. It's the cheapest sequencer, it's $1,000, and it hooks up to a laptop. You do need to understand how to do the computational interpretation, but you don't necessarily need a big infrastructure, it literally can plug into a laptop. The Illumina iSeq is also a very cheap sequencer; it's very straightforward, generates a million reads every time, right on target, and it can generate small assays, amplicon essays. And let's realise that the Ion Torrent was placed in a lot of pathology labs, because they had nice amplicon panels that all the clinicians loved. Those kinds of targeted assays are really powerful and can answer a lot of questions. So, when I think about tabletop, smaller-scale instruments, that's what I think about.

Miten makes a great point that some of these other instruments that are larger scale, like Element, Singular, the Next-seq, are even bigger. They are still benchtop, if you have a big bench, but when I'm thinking about tabletop or benchtop instruments, I'm really thinking down at the bottom of the market. I'm thinking about the MiSeq, the iSeq, the MiniSeq and the MinION. Those can answer a lot of your questions if you think about things in a targeted way. And here's a little bit of pitch about the adaptive sampling that you can do with Nanopore, where you can programmatically target regions of the genome you want, and from a MinION you can get 30x coverage for a gene panel. You can answer a lot of questions with that.

Miten Jain: I'll add to that. With the MinION, something else that happens that's fabulous, is that when you work with students, especially undergraduate students and early career graduate students, the fact that you can have a MinION sitting in front of them connected to a laptop with a flongle on a small chip, where the experiment is only costing $100 or a couple $100... suddenly, the engagement is fantastic. And it's really interesting when some of those students with unabashed imagination have access to those devices, and then you give them interesting problems. They come up with very creative solutions.

Chapter 4 references

1. Marx, V. **Method of the year: long-read sequencing.** *Nature Methods* **20**, 6-11 (2023).
2. Nurk, S. *et al.* **The complete sequence of a human genome.** *Science* **376**, 44-53 (2022).
3. Liao, W.-W. *et al.* **A draft human pangenome reference.** *Nature* **617**, 312-324 (2023).
4. Warburton, P.E. & Sebra, R.P. **Long-read DNA sequencing: recent advances and remaining challenges.** *Annual Review of Genomics and Human Genetics* **24**(2023).
5. van Dijk, E.L. *et al.* **Genomics in the long-read sequencing era.** *Trends in Genetics* **39**, 649-671 (2023).
6. Oehler, J.B., Wright, H., Stark, Z., Mallett, A.J. & Schmitz, U. **The application of long-read sequencing in clinical settings.** *Human Genomics* **17**, 73 (2023).
7. Wick, R.R., Judd, L.M. & Holt, K.E. **Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing.** *PLOS Computational Biology* **19**, e1010905 (2023).
8. Fuselli, S. *et al.* **A new hybrid approach for MHC genotyping: high-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate Alpine chamois (Rupicapra rupicapra).** *Heredity (Edinb)* **121**, 293-303 (2018).
9. Wick, R.R., Judd, L.M., Gorrie, C.L. & Holt, K.E. **Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.** *PLoS Comput Biol* **13**, e1005595 (2017).
10. Simmons, S.K. *et al.* **Mostly natural sequencing-by-synthesis for scRNA-seq using Ultima sequencing.** *Nature Biotechnology* **41**, 204-211 (2023).
11. Pardo-Palacios, F.J. *et al.* **Systematic assessment of long-read RNA-seq methods for transcript identification and quantification.** *bioRxiv* (2023).
12. Arslan, S. *et al.* **Sequencing by avidity enables high accuracy with low reagent consumption.** *bioRxiv,* 2022.11.03.514117 (2022).
13. Hu, J. *et al.* **An efficient error correction and accurate assembly tool for noisy long reads.** *bioRxiv,* 2023.03. 09.531669 (2023).
14. Tang, T. *et al.* **Integration of hybrid and self-correction method improves the quality of long-read sequencing data.** *Briefings in Functional Genomics,* elad026 (2023).
15. Zhang, H., Jain, C. & Aluru, S. **A comprehensive evaluation of long read error correction methods.** *BMC genomics* **21**, 1-15 (2020).
16. Eisenstein, M. **Innovative technologies crowd the short-read sequencing market.** *Nature* **614**, 798-800 (2023).

CASE STUDY: MULTIOMIC 5-LETTER SEQUENCING ALLOWS READING OF MODIFIED CYTOSINE BASES AND SIMULTANEOUS MEASUREMENT OF GENOMIC MUTATIONS IN CANCER CELLS

RESEARCHERS IN DR SAM APARICIO'S GROUP AT THE BRITISH COLUMBIA CANCER RESEARCH CENTRE (BCCRC) AND THE UNIVERSITY OF BRITISH COLUMBIA, BC, CANADA, UTILISED 5-LETTER SEQUENCING TECHNOLOGY, DUET MULTIOMICS SOLUTION +MODC, TO INVESTIGATE 'EPIGENETIC REWIRING' IN BREAST CANCER CELLS.

In this study, duet multiomics solution +modC helped to reveal:

- the epigenetic landscape of untransformed diploid breast epithelial cells with wild-type, p53-/-BRCA1-/- and p53-/-BRCA2-/- genetic backgrounds
- significant activation of stem cell enhancers through reduced DNA methylation in p53-/-BRCA1-/- cells only
- similar activation of stem cell enhancers in a triple negative breast cancer (TNBC) patient xenograft sample
- the epigenetic rewiring caused by BRCA1-/-, identifying it as a crucial gene for this type of cancer pathogenesis

Challenge

In this case study, we highlight the Aparicio group's research on decoding the relationship between genomic mutational background and epigenomic, or nongenomic, transcriptomic contributions to the fitness of cancer cells. This research study measures both the state of the genome in cancer cells, as well as decodes the state of the epigenome and the transcriptome. Making it vital to capture the epigenetic information encoded in modified cytosine bases in DNA, as a component of their investigations.

Solution

OVERCOMING CHALLENGES TO PERFORM SIMULTANEOUS GENOMIC AND EPIGENETIC SEQUENCING OF BREAST CANCER CELLS

To overcome previous challenges associated with the utilisation of low-sample volumes and multiple workflows, Dr Gurdeep Singh, postdoctoral fellow, at the Aparicio lab employed **duet multiomics solution +modC** to simultaneously investigate the genetic and epigenetic landscape of breast cancer cells known to exhibit homologous recombination deficiency (HRD) which drives genomic instability and cancer pathogenesis. This cancer cell trait or driving mechanism is involved in both triple-negative breast cancer (TNBC) and high-grade serous ovarian cancer.

duet multiomics solution +modC enabled Dr Singh to investigate epigenetic rewiring - known to play a critical role in cancer pathogenesis, cancer advancement, and cancer drug resistance - in breast tumour cells.

HRD mutational instability is known to be dependent on BRCA and p53 mutations. In this study, Dr Singh used untransformed diploid 184hTERT breast epithelial cell lines deficient in genes often mutated in HRD cancer (p53−/−BRCA2−/− and p53−/−BRCA1−/−), with wild-type genomic background (WT184hTERT) cells as a control.

"Having a single-workflow method that allows reading of modified cytosine bases and simultaneous measurement of genomic mutations is a game-changer for us."

Dr Sam Aparicio

About BCCRC

The BC Cancer Research Centre's mission is to pursue world-class research that aims to transform the lives of patients by exploring basic mechanisms and technology developments in all areas of cancer research including cancer control, clinical studies and trials, cancer surveillance, and population health and services. The research portfolio also supports facilities and platforms in genomics, bioinformatics, imaging, drug development, and tissue banking.

The Aparicio group studies the genomic and phenotypic behaviour of breast and other cancers. They integrate leading technologies to support their efforts to better understand how cancer clones evolve and to identify novel strategies for cancer treatment and predictors of response.

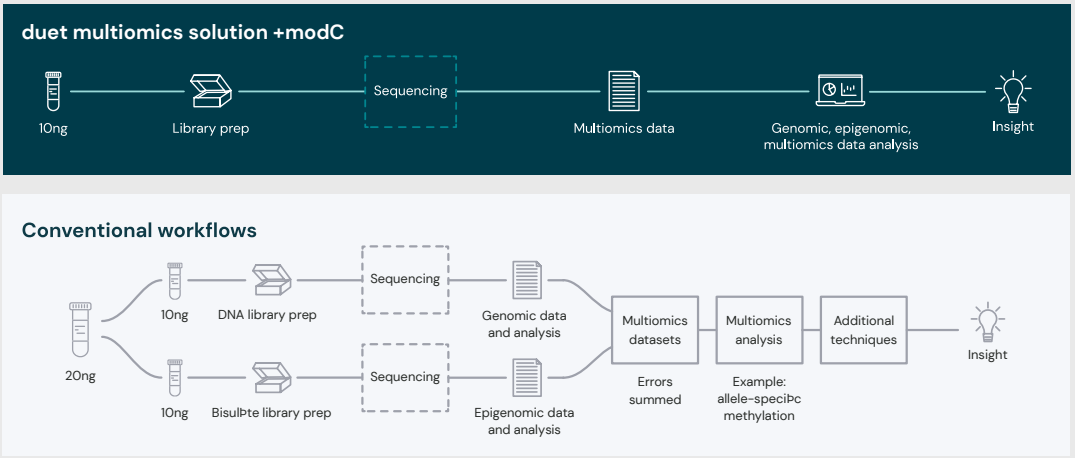


Figure 1. The duet multiomics solution +modC single workflow vs conventional genomic and epigenomic sequencing workflows. (Top panel) Following the pre-sequencing workflow, and device-agnostic sequencing, the post-sequencing bioinformatics pipeline aligns epigenetic and genomic sequencing data for analysis, interrogation, and insight.

(Bottom panel) Conventional epigenomic and genomic sequencing methods require multiple workflows, are more prone to errors, require more DNA sample (20ng), and multiple datasets to gain insights.

Method

DECODING EPIGENETIC REWIRING USING DUET MULTIOMICS SOLUTION +MODC

Firstly, DNA methylation using long-read sequencing was used to compare and confirm the DNA methylation landscape seen with **duet multiomics solution +modC** for WT184hTERT. The resulting data revealed strong Pearson correlation.

The next step was to interrogate the genomic and epigenetic landscape of all the cell types using **duet multiomics solution +modC.** The single workflow approach enabled researchers to glean greater insights from small amounts of sample DNA (Fig1).

Interestingly, using **duet multiomics solution +modC**, revealed that only the p53-/-BRCA1-/- 184hTERT, and not the untransformed diploid 184hTERT breast epithelial cell lines (WT184hTERT) or the p53-/-BRCA2-/- 184hTERT, showed significant activation of stem cell enhancers through reduced DNA methylation, and hence cancer-associated epigenetic reprogramming.

In a second step, Dr Singh used **duet multiomics solution +modC** on a reference TNBC patient-derived xenograft (PDX) sample, which also showed significant activation of stem cell enhancers through DNA methylation changes.

Results

EMPOWERING GAME-CHANGING RESEARCH IN A SINGLE WORKFLOW

In this study, the Aparacio lab used duet multiomics solution +modC to analyse in vitro breast cancer cell lines, then compared these findings to cells from a patient biopsy. They found strong correlation and alignment from the resultant comparative genomic and epigenetic data and were able to inform their research on cancer progression in triple-negative breast cancer. The findings illustrate that BRCA1-/- is crucial for HRD-specific cancer pathogenesis, where it also drives genomic instability signatures, and while BRCA2-/- drives genomic instability, it alone may not be able to drive the necessary epigenetic rewiring for cancer progression.

Researcher Spotlight

Dr Sam Aparicio, BM, BCh, PhD, FRCPath, FRSC

Dr Samuel Aparicio is the Nan & Lorraine Robertson Chair in Breast Cancer Research, holds the Canada Research Chair (Tier 1) in Molecular Oncology, and is the recipient of the 2014 Aubrey J Tingle Prize. He is also Head of the Department of Breast and Molecular Oncology at BC Cancer Research, part of the Provincial Health Services Authority, and a Professor in the Department of Pathology and Laboratory Medicine at UBC.

Dr Gurdeep Singh, PhD

Dr Gurdeep Singh is Post-Doc in Dr Samuel Aparicio's lab at BC Cancer, decoding the epigenetic basis of cancer pathogenesis and drug-resistance using CpG methylation & epigenomic landscape, and defining/testing the responsible transcriptional regulators. Dr Singh received his PhD in 2021 from The University of Toronto where he identified the genome sequence code that confers enhancer activity in embryonic stem cells, and other tissues, using functional genomics experiments and computational approaches.

CHAPTER 5

SINGLE-CELL AND SPATIAL SEQUENCING

SINGLE-CELL AND SPATIAL ASSAYS HAVE BEEN MAKING WAVES IN THE SEQUENCING SPACE FOR OVER A DECADE. THIS CHAPTER WILL EXPLORE HOW SINGLE-CELL SEQUENCING AND SPATIAL ANALYSIS HAS CHANGED HOW WE APPROACH SEQUENCING, AND THE VALUE OF CELL SPECIFICITY AND SPATIAL RELATIONSHIPS WHEN IT COMES TO SEQUENCING OMICS.

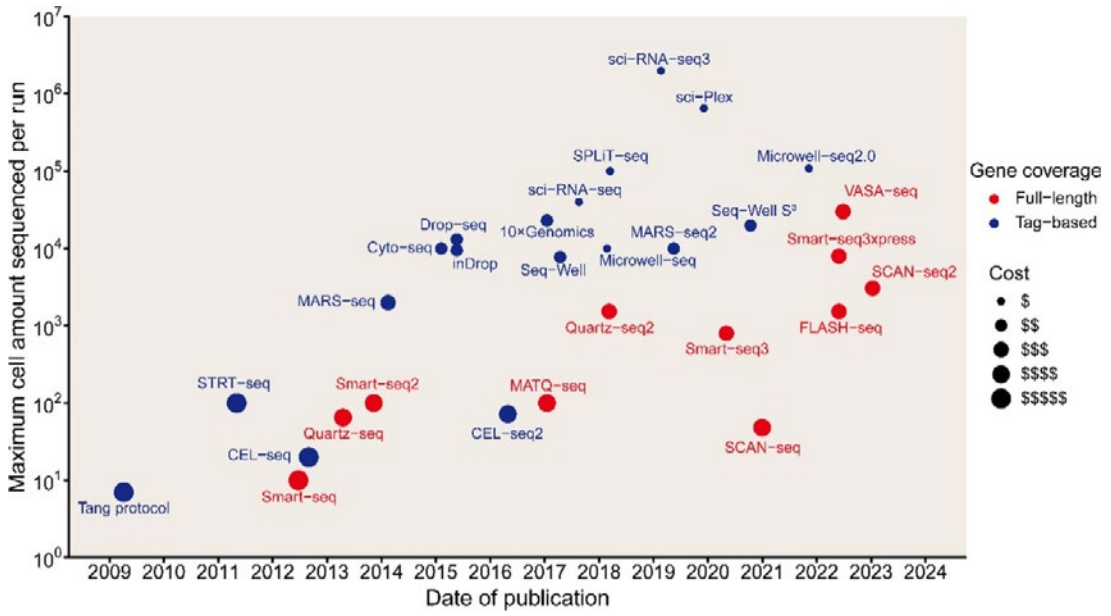
Single-cell sequencing - advantages and applications

Sequencing has historically been performed on bulk tissues, and bulk sequencing is still used today with great utility. When sequencing a whole genome, bulk data works excellently. However, biological processes and diseases are often complex and heterogenous, and an understanding of this can only be gained from methods that capture this nuance. This is the advantage of single-cell sequencing.

Primarily used for RNA sequencing as opposed to DNA sequencing, single-cell methods deconvolute bulk tissues into individual cells that can then be separately sequenced. Although this results in only a small amount of genetic material per cell, you gain an appreciation of each unit of the tissue as opposed to an aggregate reading across it. Hence, single-cell RNA sequencing allows researchers to understand a disease from the level of each individual cell, which is the level at which diseases tend to act. It allows us to ask how gene expression differs between healthy and diseased cell types and what might this mean for the disease profile and progression.

FIGURE 5.1. DEVELOPMENT OF SINGLE-CELL RNA SEQUENCING TECHNOLOGY.

Timeline and throughput of various scRNA-seq methods. Scatterplot depicts the published date and throughput of sequencing for each technology. The colour indicates the different gene coverage. Size indicates the cost per sequenced cell of scRNA-seq methods. Image Credit: Huang, et al. [1]



Single-cell DNA sequencing is also valuable. While healthy cells in a tissue will have the same genome, single-cell DNA sequencing can identify somatic or germline mutations in specific cellular populations, which can help with investigations in cancer, ageing and neurodegeneration. The issue with this methodology arises from the fact that there are very small amounts of genomic material in a single cell. Whole genome DNA methods, such as multiple displacement amplifications (MDA)[2], multiple annealing and looping-based amplification cycles (MALBAC)[3] and degenerate oligonucleotide primed PCR (DOP-PCR)[4], allow this analysis at genome-scale by amplifying low abundancies of DNA.

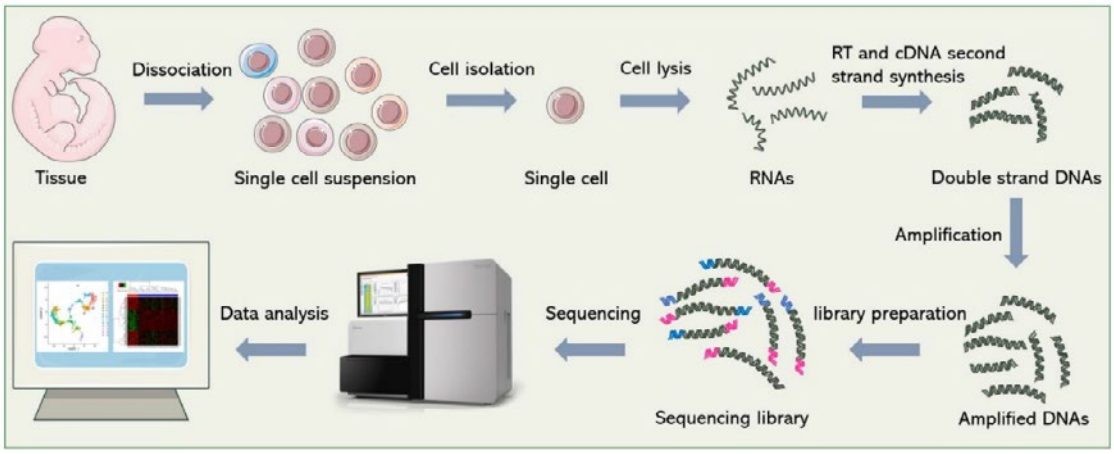
Single-cell sequencing still relies on the sequencers covered in Chapter 3: Meet the Sequencers.. This methodology benefits from the advancement in sequencing technology capabilities. However, there have also been advances in the tools used to prepare cells for these methods. Figure 5.1 shows how this improvement in technology has resulted in huge increases in the number of cells that can be processed in a single experiment, alongside the subsequent decrease in cost per cell to run single-cell methodologies.

This process has advanced, from the early days of working with individual cells, to the first fluidic circuits that allowed the processing 100s of cells[5,6], and eventually to droplet methods (inDrop[7] and Drop-seq[8]), allowing 10s of 1000s of cells to be processed. Finally, combinatorial indexing methods (sci-RNA-seq[9,10] and SPLiT-seq[11]) have brought us to the current era, in which million-cell experiments are now a viable reality for most researchers.

Single-cell sequencing - A brief look at the solutions

To perform single-cell sequencing, you do not require a new sequencer. Rather, you simply require tools to effectively partition, barcode and pool individual cells (see Figure 5.2). As previously mentioned, this was historically done by manually isolating individual cells. However, modern methods allow the processing of 1,000s of cells at a time.

FIGURE 5.2. BASIC WORKFLOW OF SINGLE-CELL SEQUENCING
Image Credit: Pan, et al. [12]



TO PERFORM SINGLE-CELL SEQUENCING, YOU DO NOT REQUIRE A NEW SEQUENCER. RATHER, YOU SIMPLY REQUIRE TOOLS TO EFFECTIVELY PARTITION, BARCODE AND POOL INDIVIDUAL CELLS."

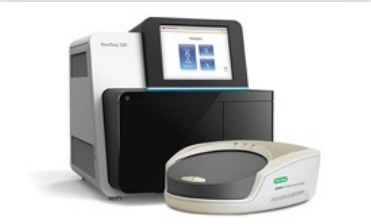
In fact, a variety of single-cell sequencing instruments, kits and solutions are available in the market for effective cell preparation. We will briefly review a selection of these below.

The most well-known single-cell instrument is the Chromium Controller from 10x Genomics. The Chromium uses advanced microfluidics to partition single cells and barcode them on a large scale (up to 128 samples and a million cells). It can be used for multi-omics and on fixed or frozen samples using the Flex kit.

The BD Rhapsody from BD Biosciences is another end-to-end system that allows you to visualize all steps in the single-cell capture workflow, with analytic metrics available at each stage. The latest HT Xpress System has increased capacity to allow million-cell studies.

Mission Bio's Tapestri platform deploys a two-step microfluidic workflow, which makes DNA and protein information accessible from single-cells. This makes it the only single-cell system capable of providing genotypic and proteomic information isolated from individual cells across thousands of cells per run.

The ICELL8 cx Single-Cell System from Takara Biosciences can process hundreds of cells or nuclei of all sizes/ shapes with eight samples in parallel. The system uses imaging to find cell-containing wells over empty or duplicate wells.



Bio Rad's ddSEQ Isolator is a droplet-based single-cell isolation-based system for single-cell chromatin accessibility studies using their SureCell ATAC-seq Library Prep Kit.

Singleron produces the Matrix single-cell processing system alongside several single-cell kits (GEXSCOPE) to take you from tissue preparation to sequencing. It can take up to 4 samples and can analyse 500,000 samples per run, with a 38-minute run time.

Also worth considering are the relatively new instrument-free single-cell sequencing solutions offered by companies such as Parse Biosciences, Fluent Biosciences, Scale Biosciences and Honeycomb Biosciences. These kits do not require external hardware like the above solutions, meaning the single-cell isolation process can occur within the test tube before being passed to a sequencer. These kits perform very well in terms of sequencing depth and genes per cell, but are more limited in throughput. However, they are relatively cheap, making them advantageous for those looking to perform infrequent experiments.

- Parse Biosciences' Evercode V2 uses the cell or nucleus as the reaction vessel; no hardware is needed to produce the RNA for sequencing. Since the kits work on fixed cells, it is possible to process up to 1 million cells across 96 samples collected across time using the Evercode WT Mega kit. The Evercode™TCR kit also allows T cell receptor profiling.
- PIPseq from Fluent Biosciences is similar and has one of the highest cell captures (85%) and better sensitivity. The kit also offers a scalable and cost-effective sequencing solution, compatible with Illumina NGS sequencing instruments. Their T100 kit is the first single tube solution with the capacity for 100,000 cells.
- Scale Biosciences presents their own technology to escape the need for cell partitioning instruments in single-cell sequencing, through their Scale Bio™ RNA kit. Their product is scalable, affordable and allows deep profiling of cells. The single-cell ATAC kit sets Scale Biosciences apart as offering instrument-free epigenomic pre-indexing.
- Honeycomb Biotechnologies released HIVE CLX, which has 160,000 picowells in their distinctive array. This allows for integration of sample storage and single-cell profiling without specialized instrumentation. Cells are captured quickly and effectively, allowing rare cell type capture, and can be stored as you go - meaning samples can be collected across a week without batch effects.

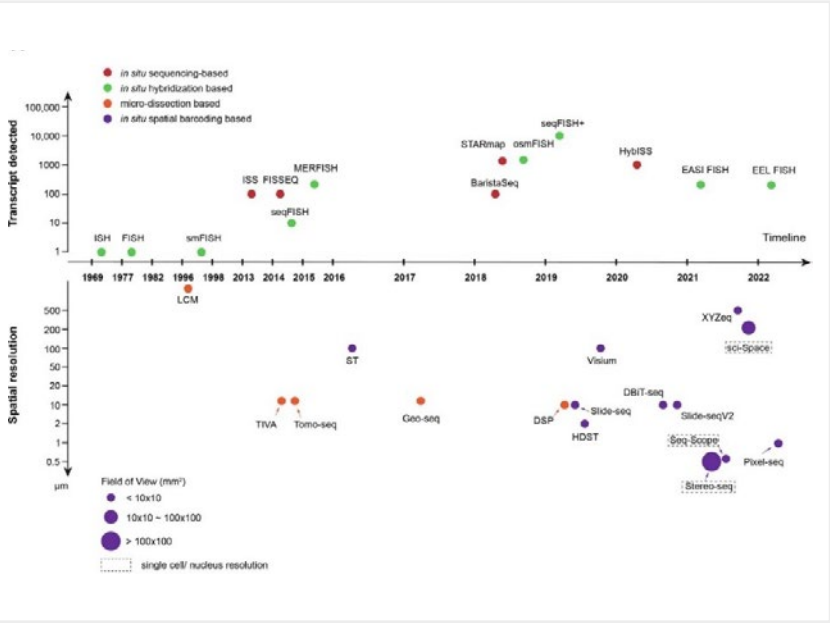
Spatial sequencing

Spatial omics encompasses several methodologies (see Figure 5.3), all with the goal of visualizing omic information in spatial context. If there is a mutation, or aberrant RNA expression, these methods can localize the feature to an area of the tissue.

Many popular spatial methods 'tag' omics material in the tissue of interest using probes with fluorescent markers. Examples of this technology include Vizgen's MERSCOPE and ACD Biotechne's RNAscope. These methods are popular and tend to capture genomic information at very high resolutions, so you know where your transcripts are localized with accuracy. However, these methods are limited in number of targets and rely on imaging technologies instead of NGS, so wouldn't be considered a 'sequencing' approach.

FIGURE 5.3. TIMELINE OF MAJOR LANDMARKS IN SPATIAL GENOMIC METHODS.

Methodologies are categorised based on type (red, in situ sequencing-based method; green, in situ hybridization based method; orange: micro-dissection based method; purple: in situ spatial barcoding based method). In-situ spatial barcoding approaches are quantified by spatial resolution and by Field of View (size of circle). Dashed box indicates single-cell/nuclei resolution. Image Credit: Cheng, et al. [25]



Alternatively, there are spatial approaches that attempt a 'sequencing' approach or rely on NGS. One example is in-situ sequencing[13] (ISS) in which padlock probes are used to sequence targeted genes in tissue sections. This original approach could acquire subcellular resolution, but is limited to ~100 targets. The more recent versions, such as HybISS[14], have a much larger target limit using sequencing-by-hybridisation instead of by ligation. 2023 saw the publication of Improved ISS[15] (IISS), which uses new probing, barcoding and imaging for better gene profiling. Other methods of in-situ sequencing, including FISSEQ[16], published in 2015, use fluorescence to capture genome-wide RNA in an unbiased manner, but, again, do not have the capacity for whole transcriptome level sequencing. More recent methods, such as STARmap[17], use padlocks without reverse transcription, and DNA nanoballs to target an expanded range of targets.

Alternatively, it appears to be the in-situ barcoding methods that are truly achieving spatial 'sequencing'. These approaches use barcodes for transcripts within cells that can then be extracted onto a slide with positional information recorded. These transcripts can then undergo library construction before being sequenced using NGS. Examples of these approaches include 10x Genomics' Visium (originally Spatial Transcriptomics[18]) , NanoString's GeoMx DSP and Curio Biosciences' Curio Seeker (based on Slide-seq[19,20]). These technologies allow transcriptome-wide sequencing, but this was typically at the expense of high spatial resolution and accuracy.

However, recent barcoding techniques, such as Seq-Scope[21], Pixel-seq[22] and Stereo-seq[23], represent the cutting edge of this method and typically allow spatial resolution of under 1 micrometer. Stereo-seq uses DNA nanoballs to achieve this and is the first technology to have achieved both subcellular resolution and a centimetre-scale field of view.

In a similar vein, new technologies such as Slide-tags[24] allow the best of both worlds, using single-nuclei sequencing with spatial information. Slide-tags first barcodes nuclei within tissues and records their location at a high spatial resolution. These nuclei can then be dissociated and isolated, meaning the mature single-nuclei technology can be used to sequence them. However, the nuclei still have the spatial barcodes and can be mapped back to their tissue context.

We conclude this chapter with a discussion and Q&A with four of our single-cell and spatial experts. In this discussion, four single-cell and spatial experts discuss the current state of both the single-cell and spatial landscapes, the advice that they give to clients when setting up assays, and their opinions of the technology available for single-cell and spatial sequencing.

PANEL DISCUSSION :

THE RESOLUTION REVOLUTION: GOING BEYOND SINGLE-CELL ANALYSIS AS YOU KNOW IT

THE CONTENT USED HERE IS A SHORTENED, EDITED TRANSCRIPT FROM A SESSION AT THE FESTIVAL OF GENOMICS BOSTON 2023

Mandovi Chatterjee
Director, Single-cell Core
Harvard Medical School

Josh Fienman
Scientist, Genomics (NGS Technology Center)
Pfizer

Linda Orzolek
Director Single Cell & Transcriptomics Core
John Hopkins University

Devjanee Swain Lenz
Director, Sequencing and Genome Technologies
Duke Center for Genomic and Computational Biology

Josh Fienman: I'm Josh Fienman. I work in the Systems Immunology group at Pfizer. We use 10x solutions – 3', 5', Flex – and we also have been using Parse's solutions. And we also do a little bit of bulk RNA-seq and a bit of spatial transcriptomics as well.

Devjanee Swain Lenz: I am Devjanee Swain Lenz. I am Faculty at Duke University in the Department of Molecular Genetics and Microbiology. And I'm also the Director of Sequencing and Genomic Technologies at Duke. We work with Duke researchers and other academic researchers for a wide variety of things. For single-cell, we work with various cores throughout the University who have the fancy equipment, and then we do more of the plate-based methods and Tapestri DNA sequencing ourselves.

Mandovi Chatterjee: I'm Mandovi Chatterjee, I'm the Director of the Single-cell core at Harvard Medical School. We offer our services to all the academic labs and industry labs in the Greater Boston area and from other parts of the country and overseas as well. We support single-cell applications across many different technologies - that includes 10x Genomics, Parse Biosciences, BD Rhapsody, Fluent Biosciences. And we're also expanding in the spatial transcriptomics area.

Linda Orzolek: My name is Linda Orzolek. I'm the Director of the Single-cell and Transcriptomics Core at Johns Hopkins University. So, at Hopkins, similar to Harvard, we're offering our services both internal and external. We offer all 10x services from standard 3' assays through to Xenium in situ spatial. We also support Parse Biosciences, Mission Bio, we are introducing Curio Biosciences for their spatial transcriptomics, Bioskryb and other one-cell-per-well-plate methods that are available also.

I know, personally, as a Core Director, both my

favourite and most frustrating thing to do is have conversations with our clients, because it's exciting to show people what is now available that they weren't really aware of. So, ***what are some of the most crucial bits of information that you find yourself consistently giving your clients when it comes to preparing for a single-cell or a spatial assay?***

Mandovi Chatterjee: We work with many different kinds of tissues across many different organisms. And in my experience, users come with different levels of expertise. Some people need quite a bit of hand holding, and in which case, from start to beginning, they need a lot of guidance. And some people are very savvy, not in terms of just bioinformatic analysis, but also, they're quite familiar with the sample prep part of the workflow, and associated molecular biology bit. Regardless of what the expertise is, we start with learning about their project and what their experimental design is, what kind of system they're working with and what their needs are. And based on that, we sometimes suggest a suitable approach. Many times, they come decided which platform or technology they want to use, but sometimes we see that there is a better option available out there. Then we try to lay that out in front of them and ultimately, it's their choice what they want to go with.

Devjanee Swain Lenz: I find that the advice I end up giving the most is to define your question. I think a lot of people think 'I want to use this exact platform.' And based on what their actual biological question is, we can give them better advice, sometimes a lot cheaper. For instance, if you're choosing between Nanostring or 10x Genomics, depending on how deep you want to go, and how many samples you have, you're going to choose one or the other. And those are two different price points, right? So, my advice is to define your biological question and talk to your biostatisticians before you actually plan your experiment. Understand that each experiment is unique, every tissue type and every species is different and will come with its own journey.

Josh Fienman: I would echo my colleagues' impressions there that we are very, very big on getting all parties from the scientific team, from the technical team to computational biologists, all in one room together to discuss the question and make sure everyone's on the same page. And then we defer to each other's expertise. We can sometimes guide you to the right platform based on your question – it's our job to put the puzzle together of what best fits your question. But we find it very important and very helpful to have all parties together to hash things out at the very beginning, before you start anything. And also, to determine how much it's going to cost. I started saying, 'I'm not telling you no, I'm telling you how much do you want to spend.' Because these things can be very expensive, and you want to make sure it's worth the investment and that you're going to get out what you need.

Linda Orzolek: Obviously, there's a common theme here - talking to people before you start. We have had – and I'm sure you have too - way too many times where people would show up and say, 'I have my cells, let's go.' And in the end, they're going to be disappointed with the answers. They're going to be disappointed with the waste of money, because the project wasn't properly planned. So, project development and those discussions are some of the most crucial steps.

Now, Josh pointed out a good note there on the cost. We all know single-cell sequencing is going to be expensive. And I've had clients come to me and say, 'It's not fair, we're too small to be able to do this.' As a result, I've had people who have been scared away by the dollar amounts. How do you approach that conversation, and get people to look past the dollar sign and take a chance on single-cell when it's their first assay, and they realise, 'Hey, my pilot study might cost me $20,000'?

Devjanee Swain Lenz: I typically ask if they've actually done bulk RNA-seq upfront. People, I think, just jump into single-cell. If they've done that cheaper bulk RNA-seq project in the past couple of years, that gives them a little bit more confidence that they can actually design their project well. And then with that, and I'm not the biggest fan of doing two replicates, but you can always do two replicates, and then write a grant. And we do offer grant support. So, that usually is where that goes.

Josh Fienman: I think we've also had similar conversations about right-sizing your experiment to, unfortunately, the money and the resources you have at hand. But again, there's sometimes a cheaper way to do something, if that still answers your question. I think that's a common barrier, depending on a department's finances at the time, knowing that it's going to be expensive, but also making sure that you're not sacrificing your experiment just to save money, and then you end up spending a lot of money on nothing. That's the other side of that coin, I guess.

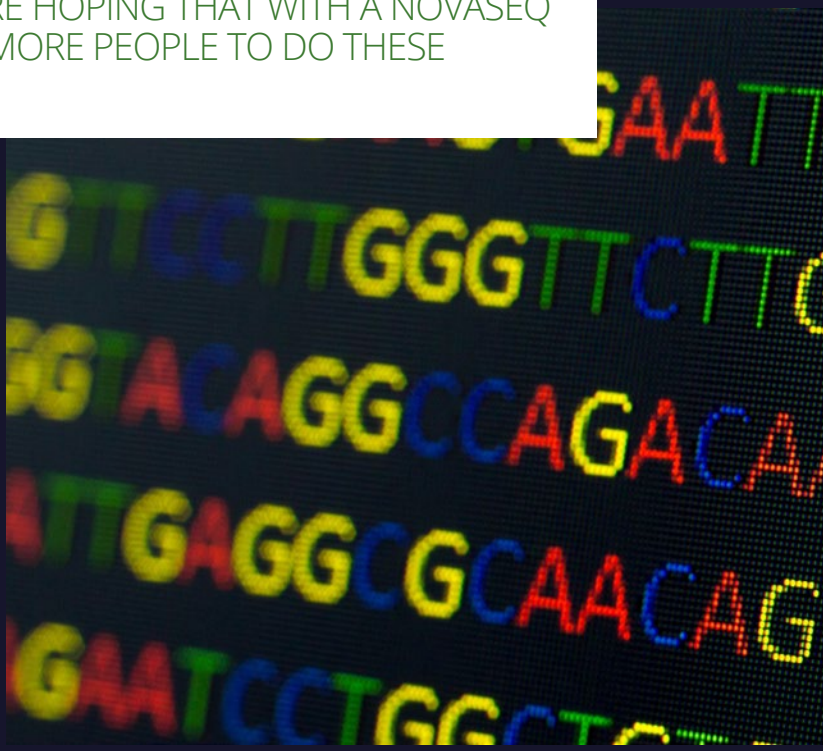
Mandovi Chatterjee: Yes, single-cell experiments are not going to be cheap. That's the bottom line. But in my opinion, the most important decision-making factor should not be the price, but the quality of the experiment. Obviously, some balance needs to be drawn; it's not possible for every single lab to have 10 biological replicates for one condition or profile one million cells per sample. So, that preliminary conversation is very important, where we can understand what your needs are, what your biological question is, whether there are less expensive options available or some trick that can be applied. For example, when working with human samples, you can pool multiple samples in one, without any hash tagging approaches, and you can bioinformatically demultiplex those samples based on SNPs. There are caveats to this approach too. This is a trick that can be applicable to certain cases. However, it's important to know about the project before we can suggest such or other tricks, which is why the initial conversation about the project in detail is crucial.

Linda Orzolek: One of the big things that we also talk about is price per data point. One of the questions we always get is, 'How many replicates do I need?' And I always laugh at that question because five or six years ago, you only needed one replicate because every cell was a replicate. And that argument existed because everything was too expensive for people to justify doing things with the actual number of biological replicates that you need. So, now that sequencing costs and library prep costs are all coming down, we're in a position where we can generate a lot more data for less cost than we did five years ago, and we can justify the need for biological replicates now. But if you look at it and you do stick with that kind of assessment that 'every cell is a replicate', every cell is an entity that you're sequencing and looking at a transcriptional profile for. So, every cell cluster you get is another data point.

If you're looking at spatial resolution, like when we talk about *in situ* sequencing at subcellular levels, and you're looking at 400 transcripts across these large areas, we'll think about some of the older methods because a new method might cost you $5,000-$10,000. Whereas an RNAscope might be cheaper than that, but you're getting a much smaller number.

"SINGLE-CELL EXPERIMENTS ARE NOT GOING TO BE CHEAP. THAT'S THE BOTTOM LINE. BUT IN MY OPINION, THE MOST IMPORTANT DECISION-MAKING FACTOR SHOULD NOT BE THE PRICE, BUT THE QUALITY OF THE EXPERIMENT."

> WE DON'T HAVE PEOPLE DOING A MILLION CELLS YET, BECAUSE WE ONLY HAVE A NOVASEQ 6000. WE'RE HOPING THAT WITH A NOVASEQ X, WE COULD START SUPPORTING MORE PEOPLE TO DO THESE LARGER SCALE EXPERIMENTS"



So, it's the throughput that you also have to consider, and looking at the big picture when it comes to what the cost is actually going to generate for you. Getting into spatial, I know that, for us, it's been about a year and a half since we started integrating spatial technologies into our services. And we're starting to see that everything ramps up, especially in conjunction with the in-situ aspects.

Are you guys seeing the shift towards spatial? Do you see that spatial is reducing the number of single-cell assays you're doing? Are people doing it in conjunction together? Or is it bringing out new researchers who are just getting into these areas?

Mandovi Chatterjee: We are offering both Visium and MERSCOPE. Visium is an NGS-based approach and MERSCOPE is an image-based approach. We actually adopted it pretty slowly. The interest is there, but I feel that people are slightly deterred by the cost of spatial transcriptomics. It's still early days in spatial transcriptomics; a lot of things are not very well understood, there are challenges and the bioinformatic solutions require a lot of improvement too. So, I think people are slightly hesitant about it.

Devjanee Swain Lenz: Duke has Visium [in the Molecular Genomics Core Facility] and that took off really quickly. And then [my colleagues said] the people who had grant money used up their grant money on Visium. So, yes, the people who have the grant money are going spend it, and then they can't anymore, and then the people who want to use it, can't.

Josh Fienman: We're also seeing a relatively slow uptake, but for us it doesn't seem like it's cost related. It's almost like we need to market it a little bit better, because we don't get broad 'everyone wants to do Visium' type of projects. But the people who do single-cell will usually want to do spatial, and if you get people in that realm, we kind of view them as complementary, where you'll do single-cell and then spatial, and then someone who did spatial wants to come back and start doing the in-situ methods. So, groups that are really, really into it seem to go all in on all three different types of readout. But we do view them as complementary. So usually, we'll try to convince someone to do Visium and single-cell up front, so you get better resolution and spatial context.

Linda Orzolek: ***Do you think that cost or bioinformatics is the most daunting aspect of single-cell or spatial analysis for your clients?***

Devjanee Swain Lenz: I'm going to say it's the bioinformatics. It's not just at Duke, it's not just in academia, it's not just in industry. It's not even just in science. There's a bioinformatics bottleneck with all of the data that we are producing. So, I'm going to go with bioinformatics on that one.

Mandovi Chatterjee: I agree. The experiments are designed by biologists, who are, most of the time, not experts in bioinformatics. Half of the workflow is wet bench work, and the rest is bioinformatic analysis. So, it's important to build a team of both biologists and bioinformaticians before you dive into the experiment. I have seen people who have generated the data, or libraries, and they've been sequenced, but the sequencing data is just lying there somewhere without getting analysed for months.

Josh Fienman: I would actually argue that we're having the opposite problem. We try to make sure we enlist the help of our very talented computational biology colleagues at the beginning. Usually, we keep them informed. So, we don't see the files just sitting in storage waiting to be analysed so much. The cost proposal, I think, is probably the bigger issue for us. But that might just be because we have a lot of very talented computational biology colleagues that are ready to tackle this kind of stuff.

Question: Do you think scalability will be a daunting task for bioinformatics going forward? Is that a concern for the users; they can generate a billion cells, but will they be able to analyse them? Is there infrastructure for that?

Linda Orzolek: If we're debating whether bioinformatics is more unapproachable, then what is the impact of generating these much larger scale studies? For us, unfortunately, we are not a bioinformatics core, we can't support it. We don't have people doing a million cells yet, because we only have a NovaSeq 6000. We're hoping that with a NovaSeq X, we could start supporting more people to do these larger scale experiments, but then, does that hold anything up on the bioinformatics side? Josh, maybe you have a better approach to that, since you seem to have more of that computational support.

Josh Fienman: I do not necessarily have the computational support in this case, but I think the tools are evolving to accommodate those datasets. I feel like a year or so ago, the infrastructure wasn't there. And you could generate the data, but you couldn't analyse it – there was that type of problem. But I think we're seeing tools that are slowly maturing to be able to do these kinds of experiments and our analysts seem to be embracing them as they come along. So, this hasn't been quite as much of an issue. To be fair, we've proposed a couple of million-cell experiments, but I don't think we've actually executed them just yet. So, it's to be determined.

Mandovi Chatterjee: I guess it's the cost factor. That is a bigger challenge than the bioinformatics challenge here, because if you barcode a billion cells, you have to sequence a billion cells as well, which can increase the cost of your experiment quite a bit.

Linda Orzolek: I think the underlying hope would be that as all of these pipelines are being developed that they will be adaptable for the obvious changes that are coming. The expansion of these projects, the time investment, the computational power that's going to be required to address them; this is all going to increase exponentially. But the hope, I think more on the bioinformatics side, is to develop things that are adaptable.

Question: When it comes to the different single-cell platforms that are out there, what are the factors that you should consider if you want to select one?

Linda Orzolek: I think a lot of it comes down to what your sample type is, what your sample availability is, whether you can collect samples at one time and process immediately or if fixation is going to be beneficial. For example, if you're doing a time course study, or you're doing patient studies where you may have one patient in the OR every couple of months. So, it really comes down to the details of your project, as we talked about before, and working out the biological question that you want to ask. And how can you go about setting that up?

So, do you need RNA? Is it just for gene expression? Or is a multi-omics approach going to be necessary for you? Do you need to look at ATAC? Do you need to look at DNA or protein? It's a very complex question. I think, realistically, you should rely on the support of core services and people who have their hands in a little bit of everything. That's what our jobs are, that's what we're here for, to talk about those details. That's my perspective of the main questions that need to be answered.

Question: Sometimes when single-cell data goes through QC, you lose a lot of cells. Maybe because of high doublet, maybe because of ambient RNA. And when using different methods, you get a lot of differences. Moving forward, is there going to be a way to improve this? For example, to get high quality data so that you don't lose a lot of cells, especially the QC level.

Mandovi Chatterjee: How good is your sample quality? Regardless of the technology you use, sample quality plays a huge role in data quality. There's a very common term, 'garbage in, garbage out.' If you have a good quality sample, then usually the data quality is very good. When the sample quality is bad, you cannot expect good quality data. A good quality sample means good single-cell suspension, with high viability, and in case of nuclei, good lysis and intact nuclei.

Linda Orzolek: The other thing to consider when you're looking at data quality is not throwing out data just because it doesn't reach some threshold. We've had projects where people have called us up in a panic, because there's such high mitochondrial levels in all of their cells, and they have filtered it out to the point that they targeted 10,000 cells, and they have 1,000 left. And as a service facility, we're trying to figure out what we did wrong, and we look back and it's muscle tissue. So, there's biological relevance for what you're seeing there. Again, that's where bioinformatics comes in - making sure that we're not setting up standard thresholds, and that the data are actually being analysed correctly, because low quality data can also be very biologically relevant.

I think, in addition to looking at the obvious sample input quality, what have you done to your samples to start with? If you're looking at immune repertoire, have you activated your T cells by some treatment that they have undergone? Are you cell-sorting, or have they been sitting out on ice for a while? Cells undergoing something that is not biological will always influence the data, and it's a matter of processing cells in a gentle and appropriate manner, so that we're not triggering transcriptional changes that are going to ultimately cause your data to be filtered out. For example, triggering apoptosis, so all of a sudden, we're starting to see, maybe not necessarily dead cells, but they're starting to die. So, the transcriptional profiles are changing. All of that ultimately comes down to sample quality, but also evaluating it in within the biological context.

Chapter 5 references

1. Huang, D. *et al.* **Advances in single-cell RNA sequencing and its applications in cancer research**. *Journal of Hematology & Oncology* **16**, 98 (2023).
2. Spits, C. *et al.* **Whole-genome multiple displacement amplification from single cells.** *Nature Protocols* **1**, 1965-1970 (2006).
3. Chapman, A.R. *et al.* **Single cell transcriptome amplification with MALBAC**. *PLoS One* **10**, e0120889 (2015).
4. Blagodatskikh, K.A. *et al.* **Improved DOP-PCR (iDOP-PCR): A robust and simple WGA method for efficient amplification of low copy number genomic DNA**. *PLoS One* **12**, e0184507 (2017).
5. DeLaughter, D.M. **The use of the Fluidigm C1 for RNA expression analyses of single cells**. *Current protocols in molecular biology* **122**, e55 (2018).
6. Brennecke, P. *et al.* **Accounting for technical noise in single-cell RNA-seq experiments**. *Nature Methods* **10**, 1093-1095 (2013).
7. Klein, A.M. *et al.* **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells**. *Cell* **161**, 1187-1201 (2015).
8. Macosko, E.Z. *et al.* **Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.** *Cell* **161**, 1202-1214 (2015).
9. Cao, J. *et al.* **Comprehensive single-cell transcriptional profiling of a multicellular organism.** *Science* **357**, 661-667 (2017).
10. Domcke, S. *et al.* **A human cell atlas of fetal chromatin accessibility.** *Science* **370,** eaba7612 (2020).
11. Rosenberg, A.B. *et al.* **Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.** *Science* **360**, 176-182 (2018).
12. Pan, Y., Cao, W., Mu, Y. & Zhu, Q. **Microfluidics facilitates the development of single-cell RNA sequencing.** *Biosensors* **12**, 450 (2022).
13. Ke, R. *et al.* **In situ sequencing for RNA analysis in preserved tissue and cells.** *Nat Methods* **10**, 857-60 (2013).
14. Gyllborg, D. *et al.* **Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue.** *Nucleic acids research* **48**, e112-e112 (2020).
15. Tang, X. *et al.* **Improved in situ sequencing for high-resolution targeted spatial transcriptomic analysis in tissue sections.** *Journal of Genetics and Genomics* (2023).
16. Lee, J.H. *et al.* **Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues.** *Nature Protocols* **10**, 442-458 (2015).
17. Wang, X. *et al.* **Three-dimensional intact-tissue sequencing of single-cell transcriptional states.** *Science* **361**(2018).
18. Ståhl, P.L. *et al.* **Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.** *Science* **353**, 78-82 (2016).
19. Rodriques, S.G. *et al.* **Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution.** *Science* **363**, 1463-1467 (2019).
20. Stickels, R.R. *et al.* **Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2.** *Nat Biotechnol* **39**, 313-319 (2021).
21. Cho, C.-S. *et al.* **Microscopic examination of spatial transcriptome using Seq-Scope.** *Cell* **184**, 3559-3572. e22 (2021).
22. Fu, X. *et al.* **Polony gels enable amplifiable DNA stamping and spatial transcriptomics of chronic pain.** *Cell* **185**, 4621-4633. e17 (2022).
23. Chen, A. *et al.* **Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays.** *Cell* **185**, 1777-1792. e21 (2022).
24. Russell, A.J.C. *et al.* **Slide-tags: scalable, single-nucleus barcoding for multi-modal spatial genomics.** *bioRxiv*, 2023.04.01.535228 (2023).
25. Cheng, M. *et al.* **Spatially resolved transcriptomics: A comprehensive review of their technological advances, applications, and challenges.** *Journal of Genetics and Genomics* (2023).

CHAPTER 6

NGS DATA ANALYSIS AND DATA MANAGEMENT

MODERN SEQUENCING TECHNOLOGIES GENERATE A LOT OF COMPLEX DATA THAT REQUIRES SOPHISTICATED STATISTICAL ANALYSIS TO MAKE SENSE OF IT. IN THIS CHAPTER, WE WILL COVER SOME OF THE LATEST ADVANCES IN NGS DATA ANALYSIS, DATA MANAGEMENT AND DATA STANDARDS.

Many computational tools have been developed over the last few decades to perform analysis on raw NGS data. One might think that, by 2024, the issue of data analysis might be resolved with standardise pipelines. However, with machines capable of producing terabytes of data in a single run, both large-scale data analysis and data management have become prominent issues for sequencing. With the advances in deep learning models as well, there are new opportunities to make your data analysis even more sophisticated than ever before.

XINKUN WANG

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology, **Northwestern University**

FLG: What data analysis challenges still exist for sequencing data?

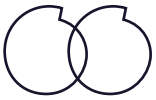
Xinkun: *While better tools are always needed from the bioinformatics, biostatistics and data science communities, especially for the integration of sequencing data with other -omics data, the most pressing challenge is the shortage of data analysts who can help biologists to analyse the data at hand. One solution for this challenge is to train biologists to enable them to perform their own analysis and, in the long run, educate the next generation of biomedical research workforce on sequencing data analysis skills. My textbook, "Next-Generation Sequencing Data Analysis", is an effort to meet this challenge to train both practitioners and students. From the enthusiastic responses I have received from the community, including those for the recently updated second edition, I have played a small role in helping meet this challenge with the rest of the bioinformatics community.*

Data analysis and AI

Raw data from NGS instruments needs to be processed, analysed and interpreted. To do this, we have a selection of computational methods, algorithms and tools to handle preprocessing, alignment and variant calling. Once processed, meaningful biological information can be extracted with techniques such as reference-based mapping, transcriptomic analysis and genetic variations can be identified such as SNPs and CNVs. Table 6.1 highlights a selection of tools (both established and novel) that perform various functions for sequencing analysis.

TABLE 6.1. BIOINFORMATIC STEPS AND TOOLS USED FOR NGS DATA ANALYSIS.
Table Credit: Satam, et al. [1]

Analysis	Commonly Used Tools
Common Analysis	
Quality check of sequences	FastQC, FASTX-toolkit[2], MultiQC[3]
Trimming of adaptors and low-quality bases	Trimmomatic[4], Cutadapt[5], fastp[6]
Alignment of sequence reads to reference genome	BWA[7], Bowtie[8], dragMAP
Reports visualization	MultiQC[3]
Whole-Genome Sequencing/Whole-Exome Sequencing/Targeted Panel	
Removal of duplicate reads	Picard, Sambamba[9]
Variant calling (single-nucleotide polymorphisms and indels)	GATK[10], freeBayes[11], Platypus[12], VarScan[13], DeepVariant[14], Illumina Dragen
Filter and merge variants	bcftools[15]
Variant annotation	ANNOVAR[16], ensemblVEP[17], snpEff[18], NIRVANA
Structural variant calling	DELLY[19], Lumpy[20], Manta[21], GRIDDS[22], Wham[23], Pindel[24]
Copy number variation (CNV) calling	CNVnator[25], GATK gCNV[26], cn.MOPS[27], cnvCapSeq (targeted sequencing)[28], ExomeDepth (CNVs from Exome)[29]
Transcriptomics	
Alignment of reads to reference	Splice-aware aligner such as TopHat2[30], HISAT2[31], and STAR[32]
Transcript quantification	featureCounts[33], HTSeq-count[34], Salmon[35], Kallisto[36]
Differential gene expression analysis enrichment of gene categories	DESeq2[37], EdgeR[38], DAVID[39], clusterProfiler[40], Enrichr[41]
16s rRNA seq	
16S rRNAseq analysis pipelines	QIIME2[42], mothur[43], USEARCH[44]
Ribosomal RNA databases	Greengenes[45], Silva[46], RDP[47]
Shotgun Metagenomics	
Taxonomic classification	MetaPhlAn4[48], Kaiju[49], Kraken[50]
Assembly of metagenomic reads	metaSPAdes[51], metaIDBA[52]
Protein databases for taxonomic classification	NCBI non-redundant protein database[53]
Gene annotation	Prokka[54], MetaGeneMark[55]
Databases for functional annotation of genes	COG[56], KEGG[57], GO[58]



ONE MIGHT THINK THAT, BY 2024, THE ISSUE OF DATA ANALYSIS MIGHT BE RESOLVED WITH STANDARDISE PIPELINES. HOWEVER, WITH MACHINES CAPABLE OF PRODUCING TERABYTES OF DATA IN A SINGLE RUN, BOTH LARGE-SCALE DATA ANALYSIS AND DATA MANAGEMENT HAVE BECOME PROMINENT ISSUES FOR SEQUENCING."

An overview of an NGS bioinformatics workflow can be found in Figure 6.1. For a helpful resource on NGS data analysis, a very user-friendly bioinformatics guide from 2023 can be found here[59]. Furthermore, textbooks such as the Next-Generation Sequencing Data Analysis are incredibly valuable for understanding how to build your workflow from raw data to multi-levelled output.

Thermo Fisher Scientific, Illumina and Roche all provide NGS data analysis support. Qiagen is perhaps the best option for researchers with less experience in bioinformatics. There are several different versions of their software, most of which are licenced as a subscription and act as a service.

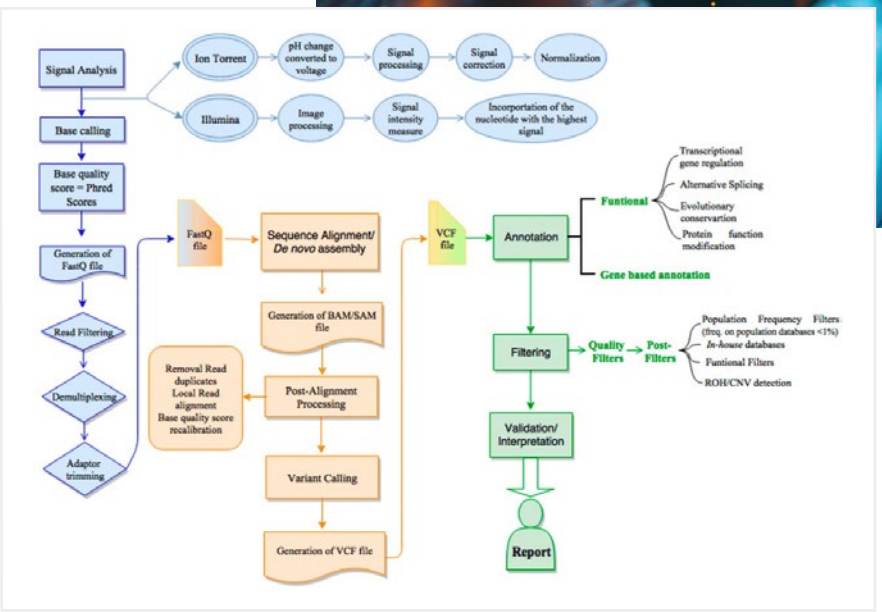


FIGURE 6.1. AN OVERVIEW OF THE NEXT GENERATION SEQUENCING (NGS) BIOINFORMATICS WORKFLOW.

The bioinformatic workflow is subdivided in the primary (blue), secondary (orange) and tertiary (green) analysis. The primary data analysis consists of the detection and analysis of raw data. Then, on the secondary analysis, the reads are aligned against the reference genome (or de novo assembled) and the calling is performed. The last step is the tertiary analysis, which includes the variant annotation, variant filtering, prioritization, data visualization and reporting. Abbreviations: CNV—copy number variation, ROH—runs of homozygosity, VCF—variant calling format. Image and Caption Credit: Pereira, et al. [60]

Furthermore, AI and machine learning are a promising area of development for sequencing data analysis. The unbiased and powerful methodology of these approaches can lead to computational tools with much greater speed and accuracy than their predecessors.

Recent examples of AI models being used to improve bioinformatic analysis include a deep neural network to validate genetic variants[61], a process which typically requires multiple variant callers[62]. Moreover, a recently developed model called DeepSelectNet[63], another deep neural network, allows for selective sequencing in Oxford Nanopore Technologies sequencing data, improving accuracy by 12% over existing methods. These tools join a wide selection of deep learning tools that have been released in the last few years that continually improve sequencing analytical capabilities[64].

ALEX COUTO ALVES

Head of the Bioinformatics Core Facility & Lecturer in Bioinformatics and Statistical Genomics
University of Surrey

FLG: What data analysis challenges still exist for sequencing data that your teams are trying to address?

Alex: *Sequencing data provides many opportunities to understand life, evolution, and disease. We are particularly interested in representations of genetic factors beyond k-mers, and variants anchored in a position. We believe we can understand genomes by dropping locality and linearity. We are looking into the dynamics of variation of viral and bacterial genomes, and thinking of new ways to predict patterns in dynamic pangenome graphs that underpin the transition of totally harmless microbes into pathogens with epidemic potential.*

FLG: What factors determine the bioinformatic tools you recommend to researchers who look for bioinformatics assistance? Do you have internal gold standards for specific pipelines?

Alex: *We recommend tools based on our own empirical experience, our understanding of the rigour of the methodology, and the needs of the project. We need to balance the sample size, prior information, the questions and goals of the project, the available resources including time and staff, etc. Methods that, on paper, should provide better results do not perform as expected. This is one of the reasons that I would recommend involving a bioinformatician or biostatistician scientist in your project, just the same way you involve a proteomics or metabolomics scientist to generate these data.*

FLG: Could you give examples of some of the innovations and bespoke tools that your team is deploying to process data efficiently?

Alex: *We are interested in understanding how genetic effects change with age. Ageing and development are associated with regulatory clocks, such as DNA methylation clocks. The effects of genetic variants change with age as these clocks alter the patterns of gene expression. One of the challenges we have in genetic epidemiology is dealing with dynamic changes in the genetic effects caused by these molecular changes. We are developing Bayesian models for longitudinal data non-uniformly sampled with nonlinear trajectories. As our models become more complex and more realistic, the computational complexity and the computational costs increase dramatically. We are looking into new ways of amortizing the computational cost and incorporating prior information to reduce the actual time to fit these models to data.*

FLG: Is there something researchers ask for that you wish you could do?

Alex: *I am often asked to analyse data generated by projects with flawed experimental design. Sometimes, the only thing we can do is a post-mortem of what failed. Every year, funding agencies around the world throw millions of pounds in the bin because modelling, data analysis and experimental design are not taken as seriously as they could be. As funding is tighter than ever, maybe this is an area where gains of efficiency can help funding agencies increase the value for money.*

Data storage

How to efficiently store and access NGS data is a prominent topic in the sequencing community. Many genomic facilities were quite simply not set up to handle the excesses of data that can now be produced by a sequencer in a single day. A single human genome at 37x coverage is over 100 GB of data. Today's large-scale projects are sequencing thousands of genomes. meaning that one experiment could involve tens of terabytes worth of data. For this issue, there are solutions available either in local or cloud-based storage.

> IT MAY COST 2-3 PENCE/CENTS PER GB PER MONTH TO STORE DATA ON THE TOP TIER WITH EASY ACCESSIBILITY, BUT FOR LONG-TERM STORAGE, IT CAN COST AS LITTLE AS 0.2 PENCE/CENTS PER GB PER MONTH."

When it comes to storing the data, it is sometimes advantageous to use a local solution (such as large hard-drives, solid-state drives or high-performance computing clusters (HPCCs)). While this option may have a high startup cost or high maintenance fees, you tend to have more control over the data and more control over data security. You can also access your data at your convenience, offline. However, not all groups will have access to a local solution or the capital to afford to maintain one.

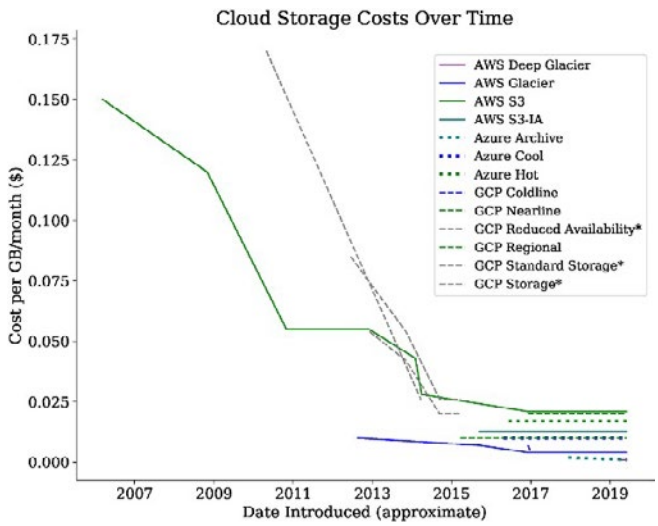
The alternative – cloud computing - is perhaps the only storage model capable of providing a widely accessible storage solution for large-scale sequencing experiments. Cloud computing is an online backup space that maintains a repository of your data on multiple servers across different locations. This tends to be more cost efficient and keeps your data secure and legally compliant, while remaining easily accessible.

The world-leading cloud platforms, such as Google Cloud Platform (GCP), Amazon Web Services (AWS) and Microsoft Azure, provide high-quality cloud storage options while achieving and maintaining compliance with complex regulatory requirements, frameworks and guidelines. These options also allow easy scalability compared to purchasing additional storage hardware.

Furthermore, these platforms have a tiered structure in which deeper layers are cheaper to store data within, but are less easily accessible. Your data can be moved between layers starting in the expensive, easier to access layer, before moving deeper over time as accessibility becomes less of a concern. It may cost 2-3 pence/cents per GB per month to store data on the top tier with easy accessibility, but for long-term storage, it can cost as little as 0.2 pence/cents per GB per month[65]. In a similar pattern to the cost of sequencing the human genome, the cost of storing a genome on the cloud has dramatically reduced with time (see Figure 6.2).

FIGURE 6.2. HISTORICAL PRICES OF CLOUD STORAGE ACROSS MAJOR CLOUD VENDORS AND PRODUCTS OVER TIME.

Image Credit: Krumm and Hoffman [65]



These cloud providers are now offering complete platforms for genomic data, where it can be stored, accessed and analysed. Terra and Cromwell on GCP is a good starting point for middle-scale NGS or GWAS data analysis projects. Terra allows researchers to execute many analysis workflows on the Cromwell engine, and it also offers a workflow reuse and exchange environment for research reproducibility, without taking the ownership of the computational infrastructure and its management. Moreover, Google launched a cloud computing service for companies to store genomic data in 2014. Formerly called Google Genomics, it is now named Cloud Life Sciences. Microsoft has a similar service named Microsoft Genomics.

XINKUN WANG

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology, **Northwestern University**

FLG: How are developments in AI, parallel computing and cloud computing helping address challenges in sequencing computation?

Xinkun: *Developments in computer science, mathematics and statistics have fundamentally propelled sequencing data analysis forward. Cloud computing has largely solved the problems associated with storing, transferring, and even analysing the extremely high volumes of sequencing data generated. Parallel computing leads to significant acceleration in steps that are inherently "embarrassingly parallel," such as aligning millions of reads to a reference genome. The power of AI in sequencing data processing has just begun to be realized, but even at this early stage, its power has already been demonstrated in some key steps, such as variant calling by DeepVariant and clinically relevant variant prioritization by Moon.*

ALEX COUTO ALVES

Head of the Bioinformatics Core Facility & Lecturer in Bioinformatics and Statistical Genomics **University of Surrey**

FLG: How are developments in AI, parallel computing and cloud computing helping address challenges in sequencing computation?

Alex: *I think cloud computing / parallel computing are absolutely fantastic technologies for addressing very large optimisation problems, such as the ones posed by large language models applied to sequence data. We regularly use parallel computing and the increase in computational power will make more accessible very complex combinatorial problems in phylogenetics, Bayesian statistics, multidimensional integration of data, etc.*

Data standards and data access

On a broader note, the reproducibility and accessibility of NGS data is an ever-present concern. Access to sequencers is becoming increasingly open and there is an increasing pressure to include the data processing scripts and give access to the raw data when publishing results. We are also seeing an increasing push to establish standard gold practices in data analysis for various genomic processes.

Recent publications have detailed the best standards for specific applications, including quality control for genome and exome sequencing analysis[66], clinical whole genome sequencing[67], and imputing genotypes from low-coverage sequencing data[68,69]. Furthermore, platforms such as AnVil are setting the standard for accessible, high-quality data with over 4 petabytes of sequencing data from 600,000 individuals. This allows easy access to datasets and allows large-scale integrated analysis.

We conclude this chapter with a short discussion amongst some of our contributors concerning some questions on this topic; for instance, what is a good dataset, who should have access to the data, and what are the implications of dataset diversity?

PANEL DISCUSSION :

FROM GIGABASES TO GIGABYTES - HOW CAN GENOMIC INITIATIVES SERVE THE RESEARCH COMMUNITY THROUGH WORLD CLASS DATASETS?

THE CONTENT USED HERE IS A SHORTENED, EDITED TRANSCRIPT FROM A SESSION AT THE FESTIVAL OF GENOMICS BOSTON 2023

Vasiliki Rahimzadeh

Assistant Professor, Center for Medical Ethics and Health Policy
Baylor College of Medicine

Bradley Malin

Co-Director, Health Data Science Centre
Vanderbilt University

Angela Page

Director of Strategy and Engagement (GA4GH), **The Broad Institute of MIT and Harvard**

Vasiliki Rahimzadeh: My name is Vasiliki Rahimzadeh. I'm based at the Centre for Medical Ethics and Health Policy at Baylor College of Medicine. Here, we're talking about how to enhance both the scientific and social value of genomic datasets. We'll talk a little bit more about some of the ethical, legal and social issues, and making that data accessible to scientists, researchers and the broader scientific community, while not compromising the rights, interests and values of the communities and individuals from which that data relates. So, we'll talk a little bit about some of the empirical research we do, some of the research consortia we're a part of, and how we've been thinking about some of those issues collectively for the genomic science community at large, both now and in the future.

Angela Page: Hello. I'm Angela Page, I'm Director of Strategy and Engagement for the Global Alliance for Genomics and Health (GA4GH). I am based at the Broad Institute, here in Cambridge, our organisation has staff in the US, UK and Canada. For those of you who are not familiar with GA4GH, we were founded in 2013, about 10 years after the first human genome was sequenced. It was founded as a response to the recognition that - with the decreasing costs of sequencing, the increasing potential use cases and the promise of genomics - if we were going to do this effectively and ethically then we really needed to bring data together from all over the globe, and that was going to require interoperability and harmonisation and collaboration.

GA4GH is really a pre-competitive consensus-driven consortium for defining standards in genomic data research and, increasingly, clinical care and clinical practice. GA4GH was founded on the human right to benefit from science. This is really important, I think, for the context of this discussion. Right around the same time that GA4GH was founded, there was a paper

in the magazine Science by Chapman and Windham about the human right to science, which was outlined in the 1947 Declaration of Human Rights but wasn't widely recognized by the world. Around the time that GA4GH was founded, the leads of our regulatory and ethics workstream noticed this paper and took it as the foundation for everything that GA4GH does now.

If all of this data is coming up around the globe, because we have all of these individual research and national-level genomic data strategies, if we can't figure out a way to bring it together and actually search and analyse across it, then we are actually not delivering on that right for all humans to benefit from that science. And the other side of it is for those who are contributing to the work to be recognised for those contributions. In that sense, we're looking at the participants who are contributing their data and the researchers who are contributing to the science.

Vasiliki Rahimzadeh: ***Brad, could you tell us a little bit about your work on the 'All Of Us' program, who you are, what your role is, and how that's contributing to some of those goals on the United States front?***

Bradley Malin: I'm Brad Malin, I'm a Professor of Biomedical Informatics, Biostatistics and Computer Science at Vanderbilt University. I was involved initially with the Electronic Medical Records and Genomics Network (the eMERGE Network), which the NIH had sponsored, probably about 15 years ago now. It was built to collect electronic medical record data and tie it to biospecimens so that we could do genomic and phenomic association studies using natural observational data. And there were a lot of questions about how to make that type of information available on a broader scale. This was right around the same time that the Database of Genotypes and Phenotypes at the NIH (dbGaP) was established.

So, when we first started on All of Us, almost eight years ago, when we first did the pilot, we started looking at the friction associated with trying to get access to data and trying to use that data in a scalable fashion. And we enumerated a couple of things. One of them was the fact that if you go to dbGaP and you ask for the data, you have to go through a long review process, and that review process is overseen by a bunch of people in the federal government. And that was a concern.

Moreover, once you get the data, you just get the data. And that's really it. There's no infrastructure supporting analysis of that data. That was one of the first things that we said - we want the infrastructure that's going to give us the ability to support large scale analytics. That was when we entered a partnership initially for the program, with the Verily team and with the Broad to use the Terra System to manage all that data. But we then said, 'what is going to speed up access?' Not just the infrastructure. And we shifted from a specific project research request to what we refer to now as the 'researcher passport'. In this respect, we vet the person and not the project. That really changed the way that people started interacting with the data. We also we also made it so that it was easy, in my opinion, to build workspaces that could be shared across different groups, and you could work within somebody's workspace.

Now, the challenge associated with this was, how do you address some of the governance issues associated with this data? Because when we built the system, there were questions about who's going to get to have access to it. And we began in what we think of it as a beta format, with access for academic universities, nonprofits and groups that were all in the United States. This was a test where we wanted to see just what people were going to do with the system. How much were they going to hammer on the computational resources? Do we think that our privacy protections were going to hold up for that system? We found that it seemed to be the case. And we did this initially without genomic data. It was just medical records, data survey data. We added whole genome sequencing data a couple of years ago, we're now up to several hundred thousand people. The question now is, how do we start expanding access? Because we're getting lots of questions from pharma and from life sciences companies, and we're getting requests from groups outside the United States. So, we've got several pilots that we are just now starting to work on, where we're working on that expansion. And we're doing it small, we're doing it step by step to try and see how this is going to play out over time.

"IT'S NOT JUST A MATTER OF 'TAKE THE DATA AND THEN FIGURE OUT WHO GETS ACCESS TO THE DATA.' AND I WORRY THAT THIS IS ACTUALLY A DIRECTION THAT WE'VE HEADED IN ACROSS TIME, WHERE WE WANT OPEN ACCESS, AND WE WANT DATA TO BE MADE AS BROADLY AVAILABLE AS POSSIBLE."

Vasiliki Rahimzadeh: I think this concept of managed access is where our work really aligns. As someone who's really interested in the processes of gatekeeping access to data, a lot of my own recent empirical research has looked at what institutional bodies are responsible for what types of governance across the research data lifecycle. We recently did an international survey of data access committee members. These are institutional bodies that sit at the helm of data repositories, academic and medical centres and research institutions. For-profit companies and commercial companies can sometimes also have similar data access committees. We're really interested in the ways that they've set up their policies and practices for managed access. What we found was really striking on three levels. The first is that these data access committees (DAC) vary quite broadly. They can be a DAC of one, so a researcher who's produced the dataset manages the access themselves. Then there are large data and resource access boards, similar to the All of Us model that you just heard about. Within these committees, the members differ with respect to their expertise and the scope of roles and responsibilities. This was an interesting finding, too, because it matters who's making the decisions about who gets access to the data and for what purposes.

The second really interesting finding was that they used different intake processes and used different criteria to decide which researchers could have access and for what purposes. The criteria for judging data access requests were quite broad across the DAC that we surveyed. This has implications for consistency of decisions; you should expect that requests for the same dataset for the exact same purpose should be reviewed the same way. This is not what we found in our surveys, and it highlighted the lack of standardisation as an element of governance that we have to work on. That led us to look at semi-automated decision support tools for data access. That's where some of my empirical research has left off and draws on broader questions that we're talking about here today, about ***what makes a good dataset. I think there's the quality aspect of the data, but there's also the rigor, effectiveness and responsibility of the governance mechanisms attached to that quality data. I would love to hear both of your perspectives on that question.***

Angela Page: I think there are a couple of characteristics of a good dataset. An important one is the idea of it being as open as possible while being as closed as necessary. We know that identifiability is going to be an issue, but when is that an issue? And when is it not an issue? Depending on what data you're actually putting out there – such as sensitive health information about an individual - then perhaps that should be more closed off and have more controlled access. If it's knowledge generated from looking at the genome writ large, perhaps that does not need to be as tightly locked down.

GA4GH recently released a policy called the Diversity in Datasets policy, which is kind of responding to these many calls for increasing diversity in genomic data. In 2016, there was a paper that said, I think, 81% of genomic data that had been sequenced around the globe was coming from people of European descent. And then five years later, that number had actually gone up. So, we're actually doing worse five years later. I don't know what it is now.

But what does that mean and why is diversity important? You need to think about the research question that you're asking in order to answer what axes of diversity are important. When you're building a 'good' data set, especially it's for a rare disease, then that dataset needs to be huge. This is so you can actually find another person on the other side of the globe who has the same phenotype and the same variant in order to say, 'this is causal.' On the other hand, if you're looking at polygenic scores, then genetic ancestry diversity is important. Maybe as, if not more, important is the environmental/geographic data. It's not sufficient to just look at a bunch of people in the US who have diverse genetic ancestry.

Finally, for data quality, is the whole FAIR principle – is it findable, accessible, interoperable and reproducible? For me, coming from GA4GH, I'm biased, but I think it's really dependent on standards and harmonisation, and making sure that we're all working together in a collaborative fashion to the best extent that we can. And collaborating on interoperability and then competing on implementation is kind of one of our bylines.

Vasiliki Rahimzadeh: Brad, I know All of Us has taken a really prospective look at diversity in the data. And it was a foundational principle in the samples collected and the data it makes available. ***So, in addition to diversity within that data set, are there implications for that having that kind of diversity, beyond the scientific implications?***

Bradley Malin: I want to take a step back for a second. We're talking a lot about data, but I want to talk about the people - there are people behind the data. And I think one of the things we recognised was that it's not just about who gets access to the data, but who plays a role in governance. You were alluding to this with the industry boards, but one of the things that we did at All of Us was that we took the participants, and we put them on all of the boards. So, we've got Science Boards, we've got Access Boards, we've got boards that are looking at publications, and there's this question about partnership. It's not just a matter of 'take the data and then figure out who gets access to the data.' And I worry that this is actually a direction that we've headed in across time, where we want open access, and we want data to be made as broadly available as possible, which is great, but at the same time, there's this question of respect and ensuring that there's an understanding and that expectations are not violated.

For anybody who's a computational person, that was a learning experience for me. And it takes time to get used to it because it does slow the research process initially. But I feel it's critical. Because if you break that relationship with the representatives, you will kill the project. I think that we've seen that in certain projects as well. That was a big take home message for a lot of this. And that's a take home message for any project that's moving forward.

Angela Page: I totally agree. And I think it really depends on the population that you're talking about. You need to have that trust and that communication with the population to answer those questions. At our recent meeting, just the other week, we had a patient-parent panellist who was talking about his son, a rare disease patient who passed away several years ago. He was shocked to know that data-sharing was not routine. He assumed that we would be doing this for the good of patients like his son. And then on the other hand, you also have indigenous communities where data sovereignty is a very important thing. So, it really depends on who it is that you're talking about and with.

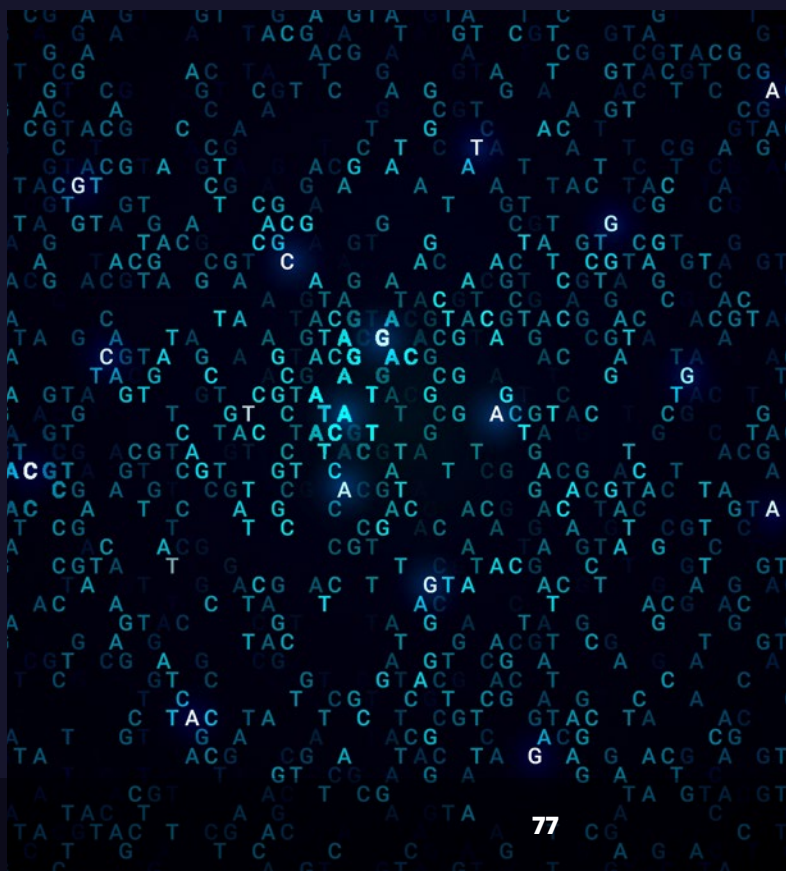
Bradley Malin: There's definitely 'no one size fits all' model. I think what works in the United States isn't necessarily what works in other countries either. And the way that we've gone with All of Us is different than what UK Biobank has done with their program. It's just about different relationships with your subject pool.

Vasiliki Rahimzadeh: ***Can you share the types of issues that your Community Board members brought up that may have been a blind spot for the scientists on the committee?***

Bradley Malin: We spent a very long time debating the concept of stigmatising research, and it's still something that we talk about. When we established All of Us, we told all the participants they were going to become a member of this program, and that the data can be used to study anything - anything that is in your medical record, anything that's in your genome, any type of mobile computing information that you give to us, it can be studied. And we asked, do you want to be a partner? Do you want to be part of this?

As we started moving forward, we said, there are ways in which the data could be exploited, that would just not be consistent with the expectations of being a good human, you know? So, we've created educational materials for anybody who comes into the system. You really have to go through what makes good stewardship, what makes you a good researcher in our industry. And if you violate those terms, we do reserve the right to kick you out. So, it's a community of researchers in that regard. When you do that, though, you have to question what it is that you consider to be taboo. There's a dividing line.

We didn't draw a hard line, we created examples and we worked with the community to define those. For example, for those of you who are not familiar with the Havasupai case, when you study the origins of a population in terms of where they come from, but you only tell them that it's going to be used for studying certain types of clinical phenotypes, then that's a violation of expectation. Or if you use the All of Us data in order to demonstrate that you have one particular sub-population that appear to be intellectually more well-off, without any grounding for doing that type of research. That would be stigmatising research. And so, we have an Access Board, and we tell all of the potential users that if they have any concerns about what they're going to do in terms of studies, and if it's considered questionable, come to the Access Board first. They will do a 360 review on this, and they will provide feedback. But there will also be post hoc analysis of the workspaces that people are building, so that there can be an investigation to determine if this is going to be considered acceptable. I don't think that that there's ever been a circumstance where we've said, 'this is a complete violation.' But we do use it as an opportunity to have a discussion with the researchers and use these examples.



Chapter 6 references

1. Satam, H. *et al.* **Next-generation sequencing technology: current trends and advancements.** *Biology* **12**, 997 (2023).
2. Iyer, R. **Isolation and molecular characterization of a novel pseudomonas putida strain capable of degrading organophosphate and aromatic compounds.** *Advances in Biological Chemistry* **3**, 564 (2013).
3. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics* **32**, 3047-3048 (2016).
4. Bolger, A.M., Lohse, M. & Usadel, B. **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* **30**, 2114-2120 (2014).
5. Martin, M. **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet. journal* **17**, 10-12 (2011).
6. Chen, S., Zhou, Y., Chen, Y. & Gu, J. **fastp: an ultra-fast all-in-one FASTQ preprocessor.** *Bioinformatics* **34**, i884-i890 (2018).
7. Li, H. **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** *arXiv preprint arXiv:1303.3997* (2013).
8. Langmead, B. & Salzberg, S.L. **Fast gapped-read alignment with Bowtie 2.** *Nature methods* **9**, 357-359 (2012).
9. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. **Sambamba: fast processing of NGS alignment formats.** *Bioinformatics* **31**, 2032-2034 (2015).
10. McKenna, A. *et al.* **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome research* **20**, 1297-1303 (2010).
11. Garrison, E. & Marth, G. **Haplotype-based variant detection from short-read sequencing.** *arXiv preprint arXiv:1207.3907* (2012).
12. Rimmer, A. *et al.* **Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications.** *Nature genetics* **46**, 912-918 (2014).
13. Koboldt, D.C. *et al.* **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** *Bioinformatics* **25**, 2283-2285 (2009).
14. Poplin, R. *et al.* **A universal SNP and small-indel variant caller using deep neural networks.** *Nature biotechnology* **36**, 983-987 (2018).
15. Li, H. **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* **27**, 2987-2993 (2011).
16. Wang, K., Li, M. & Hakonarson, H. **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Research* **38**, e164-e164 (2010).
17. McLaren, W. *et al.* **The Ensembl Variant Effect Predictor.** *Genome Biology* **17**, 122 (2016).
18. Cingolani, P. *et al.* **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff.** *Fly* **6**, 80-92 (2012).
19. Rausch, T. *et al.* **DELLY: structural variant discovery by integrated paired-end and split-read analysis**. *Bioinformatics* **28**, i333-i339 (2012).
20. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. **LUMPY: a probabilistic framework for structural variant discovery.** *Genome Biology* **15**, R84 (2014).
21. Chen, X. *et al.* **Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.** *Bioinformatics* **32**, 1220-1222 (2015).
22. Cameron, D.L. *et al.* **GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly.** *Genome Res* **27**, 2050-2060 (2017).
23. Kronenberg, Z.N. *et al.* **Wham: Identifying structural variants of biological consequence.** *PLoS Comput Biol* **11**, e1004572 (2015).
24. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.** *Bioinformatics* **25**, 2865-2871 (2009).
25. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. **CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.** *Genome Res* **21**, 974-84 (2011).
26. Babadi, M. *et al.* **GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data.** *Nature Genetics* **55**, 1589-1597 (2023).
27. Klambauer, G. *et al.* **cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate**. *Nucleic Acids Research* **40**, e69-e69 (2012).
28. Bellos, E. *et al.* **cnvCapSeq: detecting copy number variation in long-range targeted resequencing data.** *Nucleic Acids Research* **42**, e158-e158 (2014).
29. Plagnol, V. *et al.* **A robust model for read count data in exome sequencing experiments and implications for copy number variant calling.** *Bioinformatics* **28**, 2747-2754 (2012).
30. Kim, D. *et al.* **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biology* **14**, R36 (2013).
31. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.** *Nature Biotechnology* **37**, 907-915 (2019).
32. Dobin, A. *et al.* **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* **29**, 15-21 (2012).
33. Liao, Y., Smyth, G.K. & Shi, W. **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* **30**, 923-930 (2013).
34. Anders, S., Pyl, P.T. & Huber, W. **HTSeq—a Python framework to work with high-throughput sequencing data.** *Bioinformatics* **31**, 166-169 (2014).
35. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. **Salmon provides fast and bias-aware quantification of transcript expression.** *Nature Methods* **14**, 417-419 (2017).
36. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. **Near-optimal probabilistic RNA-seq quantification.** *Nature Biotechnology* **34**, 525-527 (2016).
37. Love, M.I., Huber, W. & Anders, S. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* **15**, 1-21 (2014).
38. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *bioinformatics* **26**, 139-140 (2010).
39. Dennis, G. *et al.* **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biology* **4**, R60 (2003).
40. Wu, T. *et al.* **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.** *The Innovation* **2**(2021).
41. Chen, E.Y. *et al.* **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinformatics* **14**, 128 (2013).
42. Caporaso, J.G. *et al.* **QIIME allows analysis of high-throughput community sequencing data.** *Nature methods* **7**, 335-336 (2010).
43. Schloss, P.D. *et al.* **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.** *Applied and environmental microbiology* **75**, 7537-7541 (2009).
44. Edgar, R.C. **UPARSE: highly accurate OTU sequences from microbial amplicon reads.** *Nature methods* **10**, 996-998 (2013).
45. McDonald, D. *et al.* **An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.** *The ISME journal* **6**, 610-618 (2012).
46. Quast, C. *et al.* **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic acids research* **41**, D590-D596 (2012).
47. Cole, J.R. *et al.* **Ribosomal Database Project: data and tools for high throughput rRNA analysis.** *Nucleic acids research* **42**, D633-D642 (2014).
48. Blanco-Míguez, A. *et al.* **Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4.** *Nature Biotechnology*, 1-12 (2023).
49. Menzel, P., Ng, K.L. & Krogh, A. **Fast and sensitive taxonomic classification for metagenomics with Kaiju.** *Nature communications* **7**, 11257 (2016).
50. Wood, D.E. & Salzberg, S.L. **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome biology* **15**, 1-12 (2014).
51. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. **metaSPAdes: a new versatile metagenomic assembler.** *Genome research* **27**, 824-834 (2017).
52. Peng, Y., Leung, H.C., Yiu, S.-M. & Chin, F.Y. **Meta-IDBA: a de Novo assembler for metagenomic data**. *Bioinformatics* **27**, i94-i101 (2011).
53. Pruitt, K.D., Tatusova, T. & Maglott, D.R. **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic acids research* **33**, D501-D504 (2005).
54. Seemann, T. **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics* **30**, 2068-2069 (2014).
55. Zhu, W., Lomsadze, A. & Borodovsky, M. **Ab initio gene identification in metagenomic sequences**. *Nucleic acids research* **38**, e132-e132 (2010).
56. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. **A genomic perspective on protein families.** *Science* **278**, 631-637 (1997).
57. Kanehisa, M. & Goto, S. **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* **28**, 27-30 (2000).
58. Consortium, G.O. **The Gene Ontology (GO) database and informatics resource**. *Nucleic acids research* **32**, D258-D261 (2004).
59. Larson, N.B., Oberg, A.L., Adjei, A.A. & Wang, L. **A clinician's guide to bioinformatics for next-generation sequencing.** *Journal of Thoracic Oncology* **18**, 143-157 (2023).
60. Pereira, R., Oliveira, J. & Sousa, M. **Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics.** *Journal of Clinical Medicine* **9**, 132 (2020).
61. Vaisband, M. *et al.* **Validation of genetic variants from NGS data using deep convolutional neural networks.** *BMC Bioinformatics* **24**, 158 (2023).
62. Koboldt, D.C. **Best practices for variant calling in clinical sequencing**. *Genome Medicine* **12**, 1-13 (2020).
63. Senanayake, A., Gamaarachchi, H., Herath, D. & Ragel, R. **DeepSelectNet: deep neural network based selective sequencing for oxford nanopore sequencing.** *BMC Bioinformatics* **24**, 31 (2023).
64. Schmidt, B. & Hildebrandt, **A. Deep learning in next-generation sequencing**. *Drug Discovery Today* **26**, 173-180 (2021).
65. Krumm, N. & Hoffman, N. **Practical estimation of cloud storage costs for clinical genomic data**. *Pract Lab Med* **21**, e00168 (2020).
66. Gajapathy, M., Wilk, B.M. & Worthey, E.A. **QuaC: A ippeline implementing quality control best practices for genome sequencing and exome sequencing data.** *bioRxiv*, 2023.03. 06.531383 (2023).
67. Austin-Tse, C.A. *et al.* **Best practices for the interpretation and reporting of clinical whole genome sequencing.** *npj Genomic Medicine* **7**, 27 (2022).
68. Watowich, M.M. *et al.* **Best practices for genotype imputation from low-coverage sequencing data in natural populations.** *Molecular Ecology Resources* **n/a**.
69. Rubinacci, S., Hofmeister, R.J., Sousa da Mota, B. & Delaneau, O. **Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes**. *Nature Genetics* **55**, 1088-1090 (2023).

WHAT ABOUT OUTSOURCING?

BY NOW, IT SHOULD BE CLEAR TO THE READER THAT SEQUENCING TECHNOLOGY IS RELENTLESSLY MARCHING TOWARDS AFFORDABLE AND ACCESSIBLE EXPERIMENTATION.HOWEVER, IF YOU ARE CONSIDERING BUYING A SEQUENCER, IT IS STILL WORTH SERIOUSLY WEIGHING UP WHETHER YOU ARE BETTER SUITED TO OUTSOURCING YOUR SEQUENCING EXPERIMENT AND AVOIDING THE UP-FRONT FINANCIAL INVESTMENT.

Introduction to outsourcing

Outsourcing is when sequencing and/or analysis is performed by an external group, department or company. But why would someone outsource their sequencing experiment rather than invest in their own set up?

As you will see below, there are several factors at play to determine whether outsourcing is genuinely the better option when it comes to sequencing. Although many individuals would embrace the opportunity to own a sequencer, and to have complete control over the process, it's not always the most practical or economical decision.

Advantages of outsourcing NGS and third-gen sequencing

COST

Whether it is for academic research, drug discovery or diagnosis, sequencing DNA and RNA comes at a cost. This is in large part due to the high price of sequencing instruments, which can cost millions of dollars. For many, the price is justified, given that they are designed to be used many times, making them a long-term investment. However, this upfront cost is further increased by consumables costs and upkeep costs.

Ultimately, buying a sequencer is a cost vs. use decision: will you use the machine enough to justify the purchase? If not, it generally is far cheaper to outsource the occasional sequencing experiment.

OPTIONS

Since each sequencer has specialised use cases and specific operation parameters, alongside the increasing desire to combine short-read and long-read technology in one experiment, buying one sequencer can leave you with limited options. Outsourcing services are in the fortunate position to be able to afford a variety of sequencers, from different suppliers covering both short- and long-reads. This means that the latest sequencing options could be within reach for you.

Moreover, sequencing is still evolving. New instruments are released almost annually with better yields, better coverage, better throughput and faster turn-around times. In such a dynamic market, you should consider if it is worth investing in a machine if you cannot afford to keep up with this pace. Outsourcing services continually invest in new technologies meaning that you can take advantage of the latest sequencers with every project.

TIME

Time is a factor in the decision for two reasons. Firstly, a significant advantage of outsourcing is freeing up your time as a scientist/clinician to engage in other activities. The preparatory and sequencing work is time consuming and, ultimately, will not be performed better by you than by an outsourced sequencing expert. While you lose the flexibility and ultimate control over the sequencing workflow, what you gain back is time to allocate in any way you see fit, knowing that the sequencing is taken care of.

Secondly, it can actually take less time to outsource sequencing compared to doing it within your lab/facility. Sequencing tends to be performed by junior members of a team, which requires training and often many failed attempts before the sequencing can be reliably performed. And if you choose to use a core facility, there can often be inconsistency in turnaround times. This can depend on machines breaking down, staff illness and absence, whether there are enough projects to fill up all lanes of a flow cell, or simply a queue of other samples. For reliable and generally fast turn-around times, outsourcing may be a better option.

EXPERTISE

As eluded to above, it takes time to become an expert in sequencing prep and sequencer operation. Properly designed sequencing libraries are the foundation of a successful sequencing experiment, and the value of in-house knowledge cannot be underestimated when it comes to performing this manual task. Just one mistake in sequencing could result in costly errors and loss of precious, sometimes irreplaceable, samples.

For reliable results, leaning into the expertise of an outsourcing option can be less 'painful' than learning the process and dealing with inevitable failed runs. Furthermore, this accessible service can allow for the scaling of research or clinical sample analysis, allowing many more samples to be analysed than could be achieved in-house with the same machine.

OVERALL

Ultimately, if you plan to perform a lot of sequencing - enough sequencing to keep a machine busy >80% of the time - then investing in a machine for your lab may be a logical decision. However, it is likely that many people reading this report will not be performing that much sequencing. Even if you do a lot of sequencing, the full economic cost is significantly greater than the advertised price of consumables, so you need to be certain that the space, light, power, staff, extra equipment, failed runs, maintenance contracts, repairs and IT/servers is accounted for. Outsourcing ultimately provides a cheaper, reliable alternative to access the latest sequencing options.

Outsourcing data analysis

Much of the above relates to the outsourcing of wet lab sequencing experiments., Another equally important part of the process, however, is the dry lab analysis.

The data output by sequencers is at such a scale that it requires advanced protocols to handle it. There are three major factors when considering whether you are set up to analyse the data:

- Do you have the expertise to choose the right pipeline for the project?
- Do you have familiarity with the right software?
- Do you have access to the necessary hardware?

While core facilities often provide bioinformatics support, this tends to be specialised and might not support the exact goals of your project. For example, while DNA sequencing read alignment is fairly standard, there are multiple methods for aligning methylation sites from bisulphite-seq and, as such, external expertise can be useful for deciding the best approach.

By engaging an outsourcing service, you can receive a consultation in designing the analysis pipeline and gain access to a large selection of tools and packages for analysis. You will often get options for primary and secondary data analysis as well as downstream data analysis workflows. This can take the form of analysis of gene expression, variant calling, genome assembly, ChIP-seq, Methyl-seq or some other form of custom analysis e.g. O-Link.

Furthermore, by consulting a BioIT service before even starting the wet lab experiment, you can ensure that a number of aspects of your experiment are correct such as number of samples and type of samples for your intended analysis.

A final data outsourcing consideration is storage. The largest sequencers produce tens of terabytes of data per run. That could be over a petabyte of data across a year from one sequencer alone, with the inevitable decision to be made as to whether to invest in expensive storage or delete raw data.

Source Genomics

Olink proteomics services

Next generation sequencing

Sanger sequencing services

Source Genomics

Consistently high customer satisfaction

Multiomics services

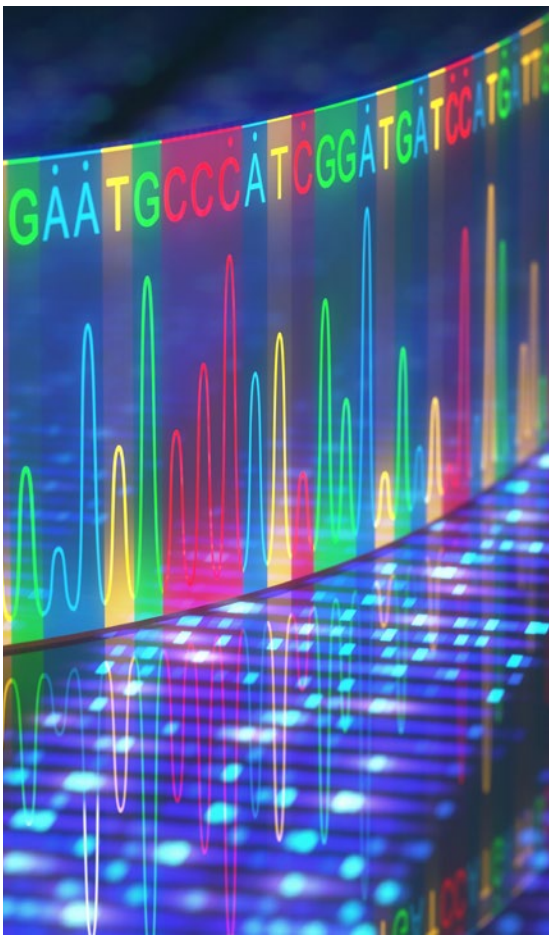
Visit website

sourcebioscience.com

ISO 9001
Electronic Quality Management System

Outsource or insource?

Before going straight to outsourcing, it is valid to consider whether it would be better to insource to a different group within your own institution or go directly to an external outsourcing service. Below, we overview the advantages and disadvantages of each.



INSOURCING TO A COLLABORATOR LAB

Pros:

- Can be the cheapest option for the individual researcher
- Personal relationship with the insourcing group
- Easily transfer samples if internal or nearby

Cons:

- Can get stuck with collaborator workflows
- Limited by minimal collaborator sequencer options
- No guarantee of expertise
- Have to balance your request with their work – could result in slow turnaround time

INSOURCING TO A CORE FACILITY IN YOUR INSTITUTION

Pros:

- Tends to be cheaper than outsourcing externally
- Easily transfer samples to internal facility
- Likely have a selection of sequencers
- More expertise and reliability than an individual research group

Cons:

- Have to share access to the facility with others – slow turnaround times
- Can be limited by facilities' expertise, not all options are available
- Can be difficult to contact and difficult to work things out with

OUTSOURCING TO AN EXTERNAL RESOURCE

Pros:

- Widest range of sequencing options and workflows
- Dedicated customer care team to support you
- Guaranteed place in the queue
- Has the fastest turnaround on average – and there is often a 'priority' or rapid turnaround option
- Highest level of expertise – maximise likelihood of success and tends to result in consistent quality
- Can operate as a 'one-stop-shop' with access to a variety of other technologies (See Table 7.1)

Cons:

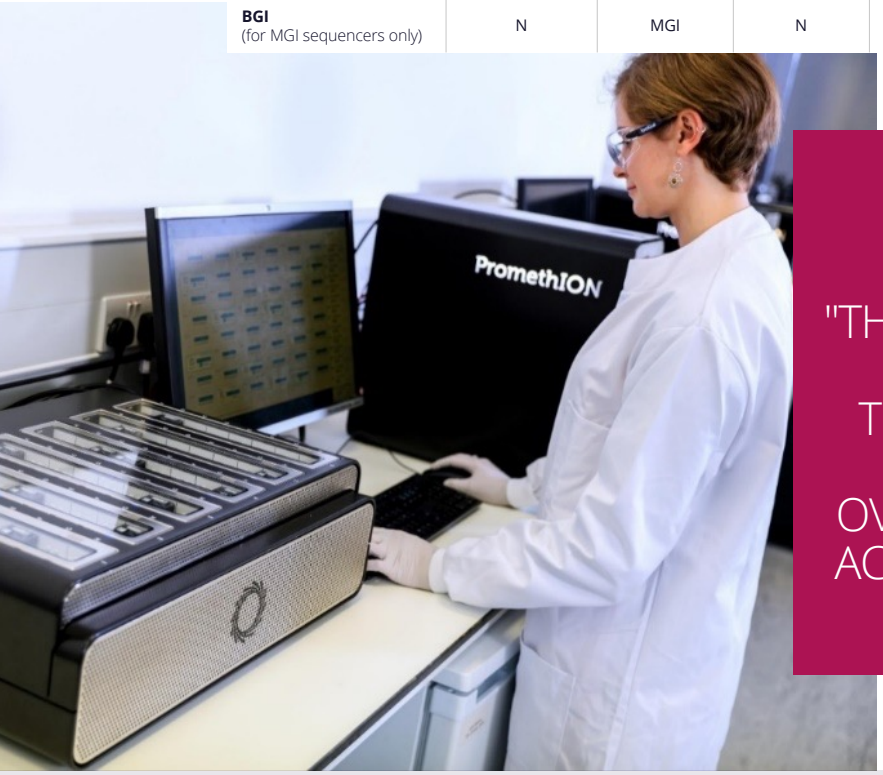
- Likely to be a more expensive option
- Have to send samples externally using dedicated couriers, which is easy but not risk-free
- While up-front agreements ensure that customers receive exactly what they are expecting/require, that may make changes to plans more of a challenge

Outsourcing options

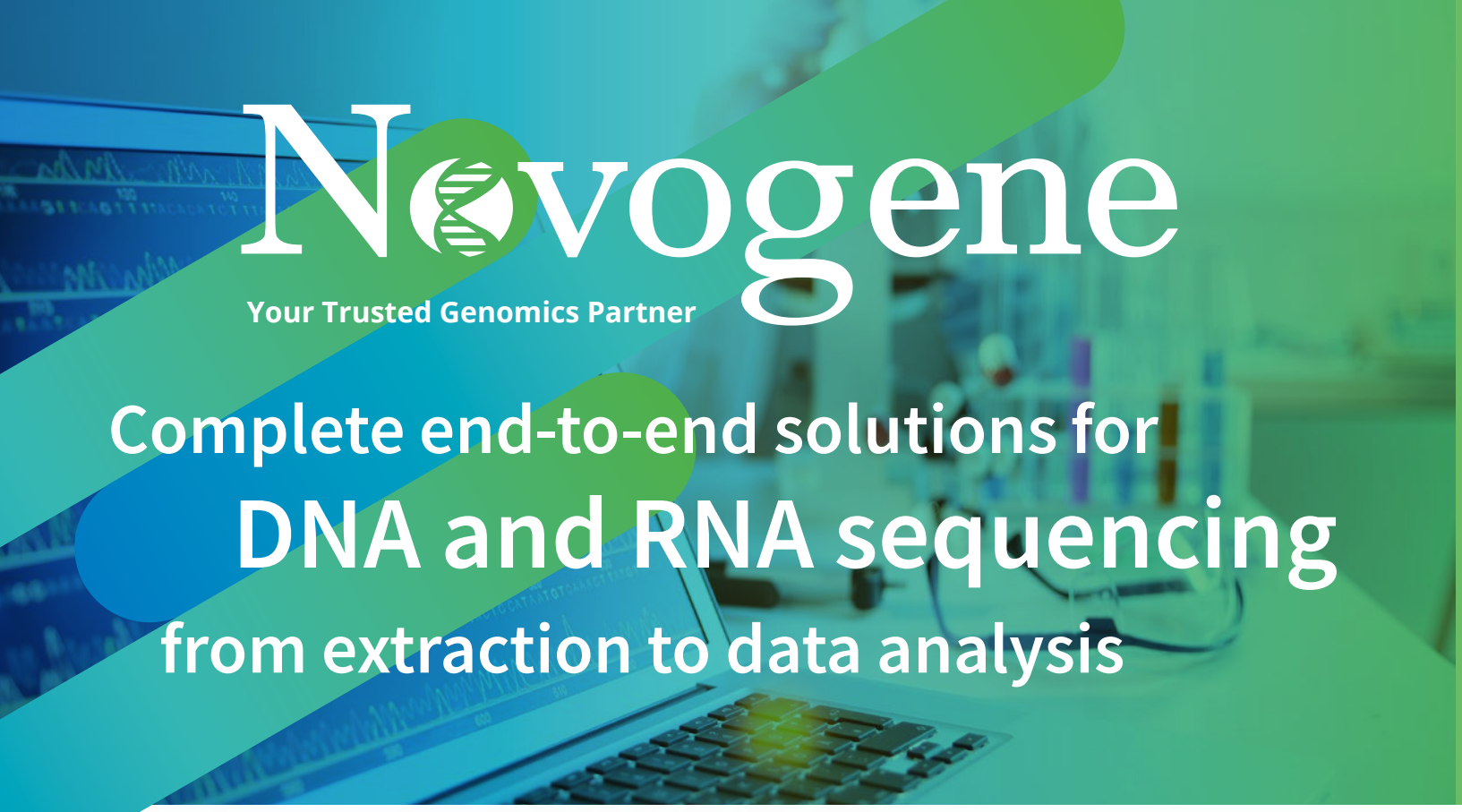
Naturally a number of outsourcing providers exists for you to engage with. To finish this chapter, we have highlighted a selection of outsourcing options in Table 7.1 along with an overview of some of their relevant services.

TABLE 7.1. EXAMPLE SEQUENCING OUTSOURCERS AND A SELECTION OF THE SERVICES THEY OFFER.
Do they provide Sanger sequencing? Which platforms do they offer for short and long read sequencing? Do they offer single-cell sequencing? Where are they located? And what other notable services do they offer? It should be noted that all companies provide dedicated bioinformatic support.

Example Outsource Options	Sanger sequencing?	NGS options	Long-Read options	Single-cell capacity?	Location	Notable Other Services
Source Genomics	Y	Illumina	ONT	Y	UK	Spatial - 10X Genomics, Proteomics - Olink, Optical Genome Mapping - BioNano
GENEWIZ	Y	Illumina	PacBio	Y	UK, Europe, USA, China	Spatial - Nanostring, Proteomics - Olink, Epigenome sequencing
Novogene	N	Illumina	PacBio ONT	Y	China, USA, Europe	Epigenome sequencing
Eurofins Genomics	Y	Illumina	ONT	N	Germany	Proteomics - Olink
Sampled	N	Illumina	PacBio	Y	USA, UK	Proteomics - Olink, Methyl-seq
Macrogen	Y	Illumina	PacBio ONT	Y	Korea	Spatial - 10X Genomics/ Nanostring, Proteomics - Olink, Epigenome sequencing
GenomeScan	N	Illumina Thermo	PacBio ONT	Y	Netherlands	Proteomics - Olink, Epigenome sequencing
CeGaT	N	Illumina	PacBio	Y	Germany	-
CD Genomics	Y	Illumina MGI	PacBio ONT	Y	USA	Spatial - 10x Genomics, Epigenomic sequencing
Illumina (for Illumina sequencers only)	N	Illumina	N	N	USA and selected other regions	-
BGI (for MGI sequencers only)	N	MGI	N	Y	China	Epigenomic sequencing



"THE LARGEST SEQUENCERS PRODUCE TENS OF TERABYTES OF DATA PER RUN. THAT COULD BE OVER A PETABYTE OF DATA ACROSS A YEAR FROM ONE SEQUENCER ALONE,"



Novogene

Your Trusted Genomics Partner

Complete end-to-end solutions for DNA and RNA sequencing from extraction to data analysis

Highest Quality
Illumina NovaSeq Q30 ≥ 85% guarantee

Largest Capacity
Unmatched automation and throughput

Latest Technology
Illumina, PacBio, and Oxford Nanopore platforms

Fast Turnaround
Faster than local cores and the vast majority of service providers

Expert Analysis
"Publication-ready" data provided by expert bioinformaticians

Novogene is a leading provider of genomic services and solutions with cutting edge Next Generation Sequencing (NGS) and bioinformatics expertise. With one of the largest sequencing capacities in the world, we utilise our **deep scientific knowledge, first-class customer service and unsurpassed data quality** to help clients realise their research goals in the rapidly evolving world of genomics. With almost 2,000 employees, multiple locations around the world, 62 NGS related patents and nearly 20,000 publications in top tier journals such as Nature and Science, we have rapidly become a world-leader in NGS services.

For more details on our services and to download product flyers scan the QR code below.

Search 'Novogene Europe'

www.novogene.com

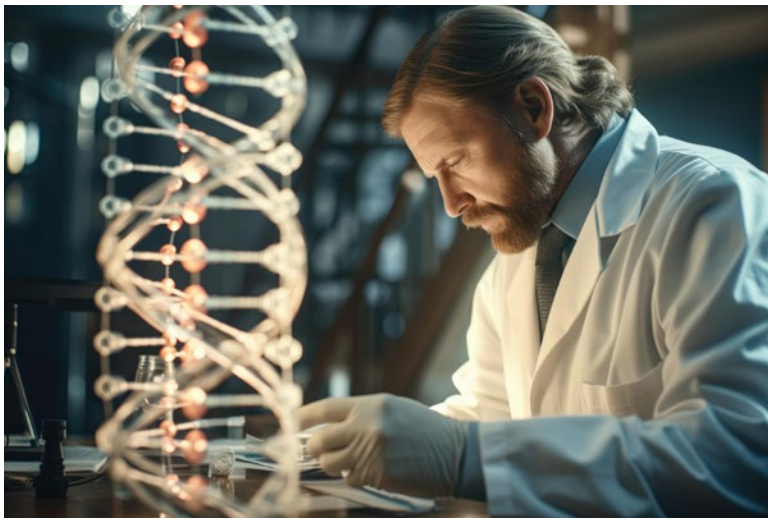
CHAPTER 8

ACCESSIBILITY AND ESG IN SEQUENCING

THE SPECIFICATIONS OF THE TECHNOLOGY AND THE SUITABILITY TO YOUR PROJECT ARE VALUABLE CONSIDERATIONS WHEN PLANNING A SEQUENCING EXPERIMENT. EQUALLY IMPORTANT IS THE SOCIETAL IMPACT OF THESE EXPERIMENTS – NAMELY THE ACCESSIBILITY OF GENOMICS TECHNIQUES AND THE ESG POLICIES OF THE COMPANIES IN THIS SPACE.

This chapter shifts from the specs, procedures and kits to explore some of the other considerations when it comes to sequencing. Specifically, we will cover the accessibility of sequencing in the clinic, how access to sequencing around the world is changing, how diversity can be increased in genomics data and, finally, the environmental, social and corporate governance of sequencing companies.

Accessibility in genomics



Implementing sequencing in routine clinical care is a challenging but necessary aspiration. A person's genome can be an essential tool for providing early diagnosis and intervention in a clinical setting.

For the latest sequencers to make their way into the clinic, they need to meet the cost, reliability, preference and infrastructure requirements of patients and providers. Sequencing approaches in the clinic not only need to be sensitive, specific and accurate, but be able to provide results fast enough to have real-time impact.

The Human Genome Project cost $2.7 billion and took 13 years to complete, but we have now entered an era of the $200 genome (perhaps even $100, see Chapter 3: Meet the Sequencers). This can be done using machines capable of sequencing 10s of 1,000s of genomes per year. Furthermore, many NGS instruments can sequence whole genomes in under a day, meaning that this technology is becoming increasingly viable for use in the clinic.

However, in resource-limited areas both locally and worldwide, accessibility to sequencing is still challenging. The expenses of maintaining a sequencer (let alone a facility) far outstrips the funds available in most developing countries. Furthermore, keeping up with the pace of development, the cost of training staff and keeping up to date with literature locked behind pay walls all work to prevent accessibility to the latest sequencing information and technology.

The advantages of wider accessibility to genomics should be obvious, and the COVID-19 pandemic was a stark reminder of the importance of deploying these technologies worldwide. For instance, the need for improved pathogen surveillance in the Global South is essential for future pandemic prevention, yet many of these nations cannot support such a program. A recognition of this need came from Illumina, who launched a Global Health Access Initiative in November 2023 to support pathogen sequencing in low- and middle-income countries.

To consider the accessibility of genomic methods in medicine, we brought together a selection of our contributors to discuss the state of genomic medicine in 2023. You can see their discussion on how AI might change genomics in the clinic, how to overcome the financial barrier of genomic technologies and how to deploy this technology in the clinic at scale.

ROUNDTABLE DISCUSSION:

GENOMIC MEDICINE IN 2023 - ADVANCES & OPPORTUNITIES

THE CONTENT USED HERE IS A SHORTENED, EDITED TRANSCRIPT FROM A SESSION AT THE FESTIVAL OF GENOMICS BOSTON 2023

Nancy Cox
Professor of Medicine and Director of the Division of Genetic Medicine
Vanderbilt University

Howard Mcleod
Professor of Medicine and Biology & Director, Centre for Precision Medicine
Utah Tech University

Shannon Muir
Chief of Staff
Latino Cancer Institute

Marylyn Ritchie
Director of Institute for Biomedical Informatics
University of Pennsylvania

Shannon Muir: My name is Shannon Muir. I am the Chief of Staff for the Latino Cancer Institute. Prior to this, I was the Co-Director of Precision Medicine in the California Governor's Office of Planning and Research, where I ran the California Initiative to Advance Precision Medicine for four years prior to switching over.

Howard Mcleod: I'm Howard McLeod and I am Director of the Centre for Precision Medicine at Utah Tech University, and a Professor of Medicine and Biology.

Nancy Cox: I'm Nancy Cox. I'm a Quantitative Human Geneticist. I lead the Vanderbilt Genetics Institute and the Division of Genetic Medicine at Vanderbilt University Medical Centre.

Marylyn Ritchie: My name is Marylyn Ritchie. I'm a Professor of Genetics at the University of Pennsylvania, I also direct the Institute for Biomedical Informatics.

Shannon Muir. So, a question for everyone. First, AI has obviously come to the forefront in 2023. I know at the Latino Cancer Institute, we're concerned about biases that are introduced in AI and in the training sets, and, of course, some of the ethics related to AI. ***But I'm wondering what each of you see as some of the strengths and weaknesses of using AI in the clinic. And how can we make sure that we use AI with omics to make precision medicine faster and better?***

Howard Mcleod: I think that there are a lot of possibilities with AI. The joke I like to use for AI is 'there's too much A, not enough I.' And it's very true that right now, there are so many tools for AI and most of them are completely useless. Right now, the AI application that's most appealing and most used is around prioritising which tissues to focus on. So, it's really a human prioritisation tool. Eventually, it will actually take on some heavy lifting. But on the clinical side, at least, we're still in the proving stage. But on the research side, there's a lot of promise.

Nancy Cox: I'm old, and there's been at least seven different waves of AI where we've thought it was going to revolutionise biological science and medicine. But I'm a little blown away, not so much by the commercial things like ChatGPT, but by some of the things we're actually seeing in the lab and analysis. The latest generation of AI is qualitatively different than anything we've been able to do up to this point. But conversely, our need to understand how it picks up these signals is much greater in clinical medicine than it is in omics and in lots of other areas, where we can just show that we can accurately predict and that's good enough. In clinical medicine, we have to be really careful and thoughtful about the signatures that are driving good quality prediction. Physicians are really loath to use black boxes, they're really not happy about true black box prediction.

That said, there are a lot of things we use in medicine that we don't fully understand, and that has always been true. So, the one thing that I try to get non-clinicians to pay attention to is how problematic our base data is from healthcare centres. Research quality data from the UK Biobank, where everybody answered all questions and where the same laboratory measures were used, is one thing. But in every medical centre, there are different practices - physicians choose what tests to order for patients based on their idiosyncratic experiences through their whole medical lives. And that means our data is full of really important biases. That allows you to draw signal from bias in a way that can be very problematic, with the really powerful tools that extract all these little things and lead you to nonsense findings. And that's what we're trying to guard against in medicine. It's not just that the methods - it's because the methods are better that we now have to worry about the very serious and idiosyncratic biases in the actual data.

Marylyn Ritchie: These have all been really good points. The other points I would make; one of them is remembering what AI is really good at - identifying patterns.

Whatever patterns are in the data, it will pull them out. There's a great publication, and then some articles, about imaging. Imaging is a place where AI has worked really well. It has been shown that AI can diagnose skin cancer really quickly and really effectively just from photographs. The article said, 'you don't need a dermatologist to actually look at a patient anymore, you can just have the AI look at the pictures.' But upon further review, what the AI was doing was finding the ruler in the picture, because every picture of an actual skin cancer lesion had a ruler next to it. Whereas a normal mole didn't have a ruler, because you didn't need to measure the size of the mole. They only had images that were skin cancer with a ruler, and it it's not skin cancer, no ruler. So, what the AI learned is that the presence of a ruler is predictive of cancer. That is not actually helpful.

If we think about using AI in genomic medicine, the examples that I think could be successful are gold standard use case examples, where we know ground truth and the AI could learn the pattern in the truth. For example, people who have known pathogenic mutations in a disease-causing Mendelian gene, you could take clinical data from their electronic health record (EHR) and try to learn if there is a pattern that is different in people who have that mutation in comparison to people who don't. Could that pattern of clinical data be predictive of other people who should have genetic testing? Because perhaps they have that Mendelian syndrome and just hadn't been tested. And I know there are some studies ongoing around the country that are doing just that, using genetics and EHR data as the pattern for the AI to learn. I think it could be really successful. But again, back to the points about bias, if your EHR data is really biased and is missing certain global or socio-economic populations, and you don't have full, complete data, that's where you might find patterns that are really predictive of a specific group.

I think the power is there and it's really exciting. It's quite frankly a little scary! Some of the things that you ask these chat bots, and you're like, 'why do you know the answer to that? You shouldn't know that!' But they also get it wrong a lot of the time. So, I think it's a really exciting time, but we need to do a lot more research and understand what's really happening.

Shannon Muir: Thank you all so much, I think those are really great points. About issues related to funding and reimbursement. ***In my work, I'm very keenly aware of how cost can be a barrier to access and lead to health disparities. I'd love to hear each of your thoughts on the challenges and possible solutions to the financial side of genomic medicine.***

Marylyn Ritchie: As a lot of health systems in the United States are moving towards more of a value-based care environment - where instead of billing for each specific thing you do, there's a package for a whole group of procedures - I could imagine bundling the costs of genetic testing in with those. That could be really beneficial for some conditions. For example, in pharmacogenomics, which is an area I do a lot of work in, often we prescribe a medication without doing any genetic testing, and it's the first-line therapy that we give everyone. But in some patients, it doesn't work. So, the patient goes on the drug, they have their stay in the hospital and go home, then they have an adverse side effect, and they end up coming back to effect, or for another hospital admission.

In a value-based care setting, the health system has to pay for that. I'm just going to make up a number and say this whole thing is going to cost $10,000, and if the genetic testing was part of that, it would be bundled in. But if the patient comes back into the ER and is then in the ICU, the health system has to eat the cost for that extra time, because that procedure was $10,000 and that's what the insurance paid. If we could add genetic testing as part of that, to make sure we put patients on the right medication or at the right dose, then they're less likely to come back into the ER or for another hospital stay. I think that's one approach.

The other approach, I would say, is if health systems would start to actually think through some of the costs. The genetic testing costs are really low now, but even in our health systems that take care of the uninsured, often what we end up doing is just treating their symptoms. So, they keep coming back. For some of those patients, if you did the genetic testing, you might find out that they're a familial hypercholesterolemia patient, or they really need this procedure or this medication. Instead, we just keep treating symptoms as they arise. So, I think in some cases, health systems might start to see that for certain conditions, it might just make sense to do the genetic testing for a couple of hundred dollars. Then they could treat the patient really effectively instead of having these repeat patients that come back again and again. I'm optimistic that, that we're going to see this in the future, but we need those cost effectiveness analyses to be done.

Nancy Cox: I'm a hopeless optimist. My view on this is coloured very much by what I believe are health inequities, major institutionalised health inequities, that can't be fixed without genetics. So, I'm cautiously optimistic that as those kinds of things become better understood and more widely appreciated that there will be no choice but to add genetics to a very basic and minimal aspect of health care. Because our health care systems will not work appropriately and fairly for everyone unless that is true. And it's not just the clear inequities that income inequality generates, it's truly institutionalised health disparities that are historic. Our perception of health equity will not allow us to continue to fail to have this information, to offer the right kind of minimal quality healthcare to everybody. I really do believe that.

Howard Mcleod: Going back to your original question around cost and reimbursement, I can't think of a single leading clinical genomics program in the United States that requires reimbursement before it starts. Most of them have either donor money or some mechanism to start and then get paid. If you wonder why 99% of the hospitals out there do not have a genomics program, it's because their model is sharing payment, and then starting. So, at the institution I work at now and the institutions I've worked at in the past, they have all had a mechanism to get started, knowing that they're not going to get paid for the first 200 tests, while you battle to get paid in the healthcare system that that we're stuck with at the moment.

The models of free market and the models of capitalism need to be suspended when starting genomics, because they're not what is being followed in practice. They can eventually come into play. But the ability to get started is really what's holding a lot of places back.

Places like the Broad are now eliminating that because you don't have to get started, they're already going. As we go to the masses, it's has to be going to the masses, not having masses come to us. And it's going to take some creative solutions, both on the technical side, but also on the finance side, in order to really make this work. Because we're just not structured in a way where we do good and then do well, it's really not the way we're structured.

Shannon Muir: Thank you so much. I want to shift a little bit to talk about education and training. Marylyn, we know that genetic counsellors and geneticists are sort of at capacity. ***So, what do you think some strategies are to address this personnel bottleneck that a lot of institutions run into?***

Marylyn Ritchie: A lot of systems are starting to create genetic counselling programs to try to train more, but that isn't scalable either. I think one of the things that that we're trying, and I think we'll see more of moving into the future, is to develop infrastructure within the electronic health record system, to teach and guide physicians and providers who are not trained in genetics to order genetic testing and figure out what to do with the results when they come back from the lab.

At Penn Medicine, we have an R01 funded right now for 10 different conditions, in cardiology, neurology and neuroendocrine. And we are building clinical decision support and nudges, to basically nudge the clinician. So, these are patients who have been diagnosed with a condition where genetic testing should be done as part of medical management. But we find lots of patients aren't getting genetic testing, and it's because they haven't been referred to genetics or a genetic counsellor, and the clinicians don't know what tests to order, don't know when to order it, or don't know how to order it.

We've built the ordering into the EHR, so it's just a couple clicks, and now we're doing a clinical trial of nudging the provider and also nudging patients to ask for the test, and building in the education so that the provider knows what to do. And then when the results come back, if they do have a pathogenic mutation, they will get referred to genetic counsellors. But they don't need a genetic counsellor just to order the test and get the data back. And for many patients, most patients, they won't have a pathogenic variant. These are rare. So, for the majority of patients, they need the test to

MY VIEW ON THIS IS COLOURED VERY MUCH BY WHAT I BELIEVE ARE HEALTH INEQUITIES, MAJOR INSTITUTIONALISED HEALTH INEQUITIES, THAT CAN'T BE FIXED WITHOUT GENETICS."

> 35% OF MEDICAL CENTRES DON'T HAVE A COURSE IN GENETICS. I THINK WE ARE GOING TO BE IN THIS FRACTURED EDUCATION SPACE FOR A VERY LONG TIME."

identify the few that have the variant, and then those patients could see genetic counselling.

If we can think more and more about how to use the infrastructure that we have and use that to train the physicians who aren't genetic specialists, so that they can be part of that workforce. And I know medical education is also changing. Now, first year medical students are getting this as part of their training, and it happens again in year two, it happens again in year four. So, 10 years from now, the practicing physicians will have some of this, but we have to train the workforce who did not get this in their training.

Shannon Muir: That's a really interesting point. Right now, at the Latino Cancer Institute, we are working with UC Davis and Gilead on a project that's actually examining, not the science behind it, just the workflow piece. You have somebody walk in the door at the clinic, whose job is it to do what? And it has really opened my eyes to the complicated nature of that. Where do the arrows go from one box to the next and how does information move around? So, I really appreciate your point there.

Nancy, just over to you. As far as the timeline goes, we've talked about a lot of challenges and maybe some solutions. ***What do you think is the overall timeline until this all comes to fruition, and we see genomics in the clinic in a way that's working favourably?***

Nancy Cox: It really will come before we're actually all ready for it. It's one of those 'be careful what you wish for' things. I really do believe everybody should have a genome available for their health care management. It should be there for everybody, and it should be there now. In terms of cost versus value, I don't think there's anything in medicine that we would do that would be as valuable as that. But 50% of practising physicians have never had a course in genetics, and that's crazy. 35% of medical centres don't have a course in genetics. I think we are going to be in this fractured education space for a very long time. But the truth is, physicians don't have to know anything but how to order the test and what to do when they get the result back. They didn't need to understand the Framingham equation to use it very effectively to treat patients with cardiovascular disease, right? They didn't have to understand the details of it, they just had to know how to implement it well. And that's what we need to do, get to a stage where we can implement it. I think it'll be faster than any of us believe it can possibly happen. And we won't be ready for it.

Diverse genomes

The vast majority of genomic data available today is from individuals of European ancestry. To be precise, around 80% of published datasets use European data[1]. While this is representative of global access to sequencing, it is not an acceptable status quo that the current 'default' genome does not represent the genetic diversity and worldwide allelic variants. This means that polygenic risk scores and targeted therapies could be biased towards European genomes, leaving many individuals unrepresented.

This issue is well recognised and there are efforts being made to address this, through programs such as the 'All of Us' program in the USA and the H3Africa Consortium.

The 'All of Us' program is building one of the largest biomedical resources of its kind, collecting genomes from a diverse group of participants across the USA. So far, the program has 750,000 participants, over 45% of which are from racial and ethnic minority groups.

H3Africa is a large-scale collaborative effort funded by The Wellcome Trust and the NIH, led by African investigators to cover the genetic diversity across different regions in Africa.

Updates regarding genomic diversity from 2023 came from the Human Pangenome Reference Consortium (HPRC). The consortium published a draft pangenome in mid-2023 comprised of 47 phased diploid assemblies from genetically diverse individuals, sequenced using PacBio and Oxford Nanopore Technologies[2]. The publication of this resource marks the beginning of the process to identify structurally complex and variable loci that are missing in the current reference genome (Figure 8.1). The consortium plans to produce a 350 assembly pangenome by mid-2024, further increasing the power of this resource and significantly addressing the problem of European-centric genomics.

A little bit more about these consortia, they naturally want to understand the reference genome, but they also understand the need to modernize it. All technology modernizes, such as phones and laptops, and the human reference genome is no different.

FIGURE 8.1. THE NEW PANGENOME REFERENCE IS A COLLECTION OF DIFFERENT GENOMES FROM WHICH TO COMPARE AN INDIVIDUAL GENOME SEQUENCE.

Like a map of the subway system, the pangenome graph has many possible routes for a sequence to take, represented by the different colours. Image Credit: Darryl Leja, NHGRI

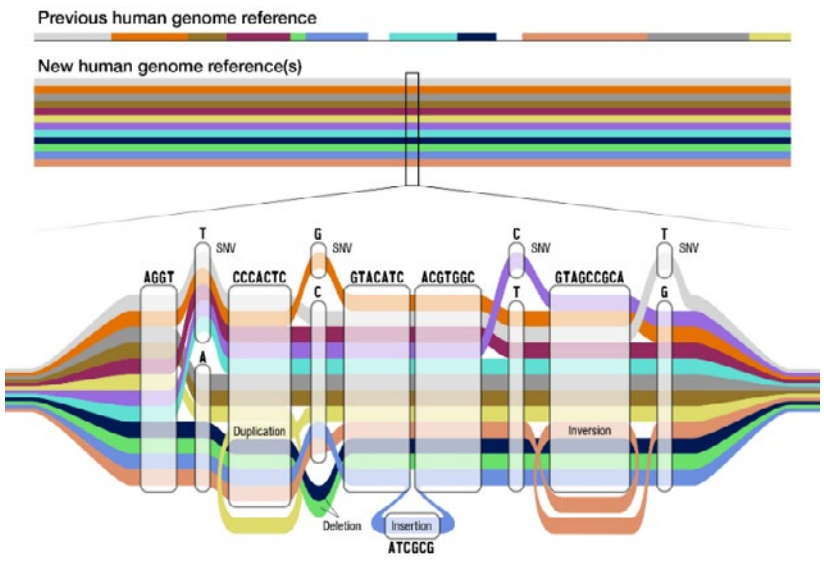
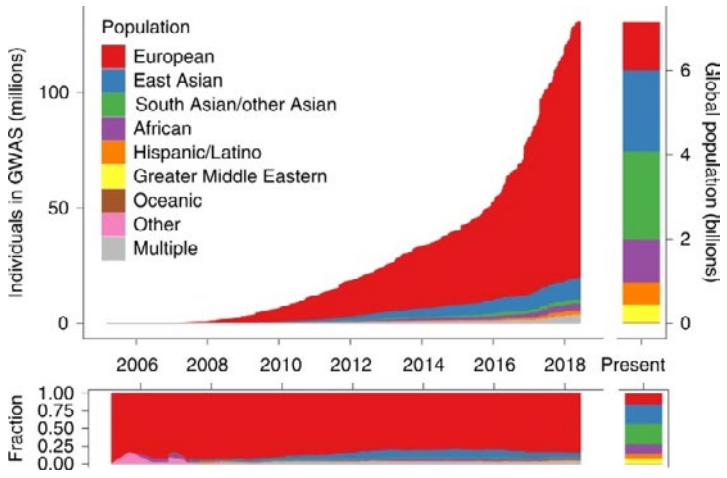


FIGURE 8.2. ANCESTRY OF GWAS PARTICIPANTS OVER TIME, AS COMPARED WITH THE GLOBAL POPULATION.

Image Credit: Martin, et al. [4]



The previous human genome, HG-38, was a linear haplotype and mostly represents a single individual. This introduces biases, which means that genetic variation cannot be detected, monitored and studied in a meaningful way.

Modernizing the genome has been occurring in two waves.

Firstly, the goal was to map the human genome in full. In April 2022, the Telomere 2 Telomere (T2T) consortium published the first human complete reference genome[3], introducing about 200 billion bases that were not present in HG-38. Two long-read platforms were the cause of this advancement, the PacBio HiFi and Oxford Nanopore sequencers, which enabled systematic read acquisition of 100,000 kb on more.

The second goal has been to use this technology for not just one genome but for other genomes too, since one genome cannot fully represent the genetic diversity of the human population. This is the goal of the Human Pangenome Reference Consortium mentioned above.

While diversifying the reference genome is important when it comes to completing the genome and getting the full picture, it is also important for ensuring equitable healthcare. The previous reference genome represents one American individual and as a result, the healthcare standard is mostly based on one American. As can be seen from Figure 8.2, genomes of European ancestry still make up the bulk of what is sequenced, hence we are not capturing the allele frequencies around the world, and we are missing the global genetic diversity.

The challenge for HPRC is how to structure and upgrade this new reference genome to capture all of this information from across ancestries. The initial goal is to generate the reference pangenome with at least 1000 diverse human haplotypes. And, instead of a linear reference, this is visualized more as a subway map (Figure 8.1), where regions that are similar to one another can be grouped into nodes.

ESG in sequencing

As a society, we are becoming more conscious of where our products come from and the environmental and social impact of what we consume. Why would this be any different for sequencing? When looking for equipment, the environmental, social and corporate governance (ESG) policy of the manufacturer can provide important detail. To reflect this, we have detailed the areas in which leading sequencing companies have made ESG positions and pledges (see Figure 8.3)

FIGURE 8.3. AN ILLUSTRATION OF THE TYPICAL TYPES OF INFORMATION THAT MAY BE USED TO ASSESS AN ESG CRITERIA.

For example, if a company's social contribution was being measured, human rights and child labour might be considered. Likewise, compliance and shareholder democracy may be considered if an organisation's governance was being analysed. Image credit: anevis

ILLUMINA

In many ways, Illumina have led the way for ESG policies within sequencing companies. Their dedicated hub for ESG can be found here. You can find the yearly ESG reports (2022 version here), starting from 2019, alongside data and resources covering aspects of their ESG policy.

For the environment, Illumina have goals of being Net Zero on carbon emissions by 2050. Furthermore, by 2030, they aim to reduces greenhouse gas emissions by 46%, reduce packaging by 75% and to divert landfill by 90%.

For social issues, Illumina aim to increase gender and ethnicity in leadership, to cover 2 billion lives with access to sequencing, create a positive company culture and give back in the form of $20 million philanthropic support for sustainability initiatives and 100,000 hours of volunteered time, all by 2030. They aim to do this while achieving the $100 genome.

Finally, for governance, Illumina aim to have 20% of their yearly spend with diverse suppliers and to become the most trusted company when it comes to genomic privacy and ethics.

2023 marks Illumina's 25th anniversary and there is no pause in their commitment to their ESG goals. The 2022 reports shows that the NovaSeq™X has a 61% climate impact reduction compared to the Novaseq™6000, a 5% increase in female executive leadership and $17.7 million in donations and $11 million in-kind donations to support the Pathogen Genomics Initiative highlighted above.

THERMO FISHER SCIENTIFIC

Thermo Fisher Scientific are another company with a dedicated ESG portal and they also provide a yearly Corporate Social Responsibility Report. Thermo Fisher Scientific are also aiming for Net Zero by 2050. They have reduced emissions by 25% since 2018 and plan for 50% reduction by 2030. Environmentally, they are aiming for 90% of suppliers to have science-based climate targets by 2027, a value that increased by 7% in 2022. They are designing products to have less waste, less packaging and to be more efficient.

Socially, they have achieved a 9% increase in women in leadership roles, reaching 47%. Thermo Fisher Scientific have also given 120,000 volunteered hours and donated $5.8 million to critical causes globally. It's important to note that these targets are business wide, and it is difficult to assess the environmental impact of the sequencing branch of the company specifically. However, just like Illumina, Thermo Fisher Scientific have a variety of pledges across the ESG sphere with many similar targets.

NOW, 59% OF SHIPPED FLOW CELLS ARE RETURNED AND MANY CAN BE RE-USED. THEIR GOAL TO SCALE RESPONSIBLY, BY PROTECTING THE PLANET THROUGH ENERGY EFFICIENT PRODUCT DESIGN AND ENSURING SUPPLIERS COMPLY TO ONT STANDARDS."

PACBIO

PacBio produced their first ESG report in 2022. PacBio is a smaller company than Illumina (<1,000 employees vs. > 10,000), and as such the goals and scope of the report are more focused with fewer targets. They adopted an environmental policy in July 2022 to minimize their use energy, water and manage their waste and emissions responsibly.

Socially, PacBio are transparent with their board makeup (33% female) and invest in their employees' safety and health, and in equal opportunities. They also have policies in place to ensure ethical supply chain management, ethical marketing and good business ethics within the company.

OXFORD NANOPORE TECHNOLOGIES (ONT)

Oxford Nanopore Technologies (PacBio's main long-read competitor) have also begun to produce ESG reports, their first one being published in 2022. For a company similar in size to PacBio, ONT have a clear description of their current position and distinct goals for the future. .

Environmentally, ONT achieved a 25.47% reduction in CO2 emissions in 2022 as well as using only LED lighting in their facilities. 79% of their packaging comes from recycled materials (91 tonnes). Furthermore, their flow cells are recyclable and the recycling scheme that ONT run saw a 19% increase in flow cells being returned. Now, 59% of shipped flow cells are returned and many can be re-used. Their goal to scale responsibly, by protecting the planet through energy efficient product design and ensuring suppliers comply to ONT standards.

Socially, a 2% increase in female employment sees the total come to 42%. Furthermore, 8,832 hours of training were undertaken by their employees. They have general policies for inclusivity, safety and personal development, with the goal of promoting a culture that is inclusive and prioritises development and wellbeing.

While there was not a report or clear set of policies, BGI/MGI have achieved some significant ESG initiatives in 2022,which they were recognised for by Fortune China. These initiatives included, the BGI Group's work during the COVID-19 crisis, their health screening services in rural and impoverished areas (providing more than 13 million screenings), their work to prevent hemoglobinopathies like sickle cell disease and their social responsibility work to increase public awareness of health campaigns.

We could not find similar pledges and transparent ESG pledges from newer players such as Element Biosciences, Singular Genomics and Ultima Genomics, but given the increasing pressure for companies to be transparent with their practices, it is only a matter of time before they make pledges themselves should they continue to grow in market share.

Chapter 8 references

1. Fatumo, S. *et al.* **A roadmap to increase diversity in genomic studies.** *Nat Med* **28**, 243-250 (2022).
2. Liao, W.-W. *et al.* **A draft human pangenome reference.** *Nature* **617**, 312-324 (2023).
3. Nurk, S. *et al.* **The complete sequence of a human genome.** *Science* **376**, 44-53 (2022).
4. Martin, A.R. *et al.* **Clinical use of current polygenic risk scores may exacerbate health disparities.** *Nature Genetics* **51**, 584-591 (2019).

CHAPTER 9

WHAT'S NEW AND WHAT'S NEXT IN SEQUENCING?

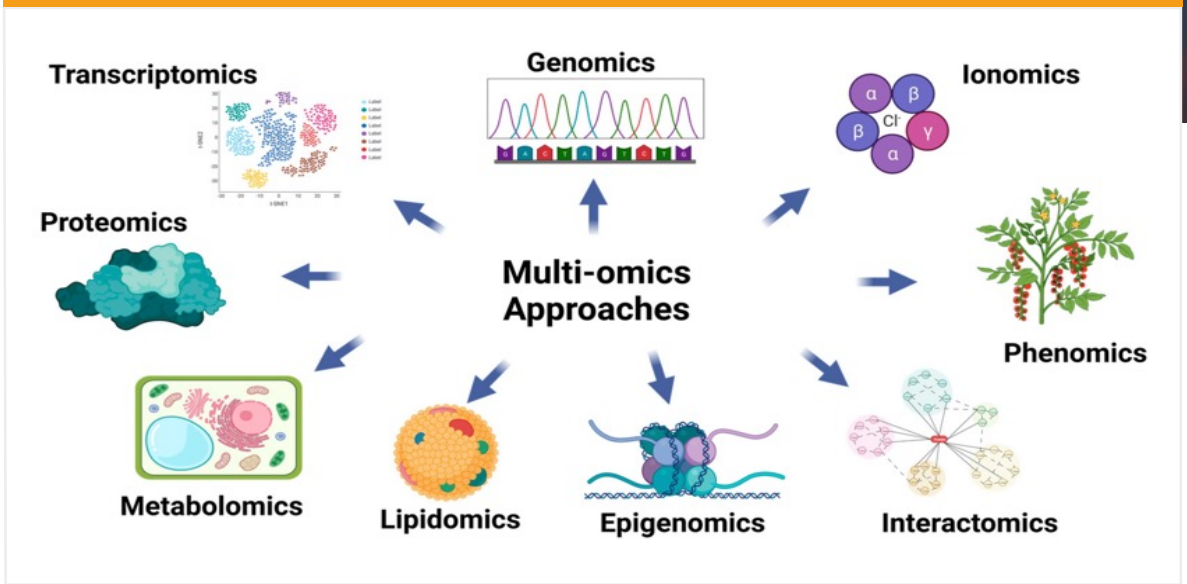
THIS FINAL CHAPTER HIGHLIGHTS SOME OF THE LATEST AND MOST EXCITING INNOVATIONS IN SEQUENCING. WHETHER IT'S PROTEIN SEQUENCING, MULTI-OMICS, THE RACE TO Q40, OR SOLID-STATE NANOPORE SEQUENCING, YOU'LL FIND OUT ABOUT IT HERE.

Multi-omics technology

If you want to understand a disease, or define a cell type, is sequencing DNA or RNA enough? The answer, quite often, is no[1,2]. This is why multi-omics sequencing is becoming an increasingly valued methodology for scientists[3-5]. Furthermore, these methods are being developed at high resolution, with exciting single-cell and spatial methods released in the last 12-18 months[6-8].

Multi-omics involves collecting multiple 'omics' measurements from a single sample, or even a single cell. This multi-modal data is then integrated together using sophisticated computational methods. Now, relationships between DNA, RNA, proteins and more (see Figure 9.1) can be explored, and we can construct multimodal profiles of diseases. These are more robust than their mono-omic counterparts, and can be used to help untangle heterogeneity of disease progression and treatment response.

FIGURE 9.1. A SELECTION OF THE MULTI-OMICS APPROACHES THAT ARE CURRENTLY AVAILABLE TO RESEARCHERS.
Image credit: Roychowdhury, et al.[9]



SUHAS VASAIKAR

Principal Scientist, Clinical Biomarker and Diagnostics
Seattle Genetics (Seagen)

FLG: From your experience, what are some of the latest/exciting things happening in the multi-omics field?

Suhas: *There are several exciting developments happening in the field of multi-omics that have the potential to transform our understanding of complex biological systems. Here are some examples:*

- ***Integration of single-cell data:*** *Single-cell omics technologies are rapidly advancing, and researchers are now able to generate multi-omics data from individual cells. This has the potential to provide unprecedented insights into cellular heterogeneity, cell-to-cell communication and disease mechanisms.*
- ***Multi-omics data visualization:*** *As the amount of multi-omics data being generated continues to increase, there is a growing need for effective data visualization tools. New visualization methods, such as interactive network-based visualization platforms, are now being developed to help researchers gain insights from complex multi-omics data sets.*
- ***Multi-omics biomarker discovery:*** *Integrating data from multiple omics technologies can help identify biomarkers that are more accurate and reliable than those identified using a single technology. These biomarkers can be used for disease diagnosis, prognosis and treatment.*
- ***Deep learning approaches:*** *Deep learning approaches, such as deep neural networks and convolutional neural networks, are now being applied to multi-omics data sets to identify complex patterns and relationships between different omics data types. These methods have the potential to reveal new insights into disease mechanisms and identify novel therapeutic targets.*

The principal challenge for multi-omics sequencing comes from the inherent difficulty in integrating omic data, which exist at different data scales, noise ratios and 'completeness' (amounts of missing data). Computational methods present a solution to this problem and deploy either algorithm-based or machine learning models to effectively match the omics data within a sample/cell[10]. The latest of these methods can perform sophisticated mosaic integration, linking omics data from the same sample and from different samples alike[11,12].

NOW, RELATIONSHIPS BETWEEN DNA, RNA, PROTEINS AND MORE CAN BE EXPLORED, AND WE CAN CONSTRUCT MULTIMODAL PROFILES OF DISEASES. THESE ARE MORE ROBUST THAN THEIR MONO-OMIC COUNTERPARTS, AND CAN BE USED TO HELP UNTANGLE HETEROGENEITY OF DISEASE PROGRESSION AND TREATMENT RESPONSE."

BINGJIE ZHANG

Postdoctoral Research Fellow, Satija Lab
New York Genome Center

FLG: Why do you think multi-omics integration is still such a challenge? Are there any approaches to integration that you particularly like?

Bingjie: *Integration is challenging simply because we are measuring two different modalities, and our understanding of how they correlate with each other is not very clear. Mapping RNA-seq to open chromatin accessibility might be relatively easier because of the underlying assumption that actively transcribed genes should have greater open chromatin accessibility, which is a correlation we can model. However, for other modalities, such as RNA-seq and protein data, the most abundant protein may not correlate with high gene expression. This disconnect makes integration very difficult. Moreover, sensitivity remains an issue. A gene detected at the RNA level may simply be missing in the ATAC dataset. In terms of the scale, while scRNA-seq can profile thousands of genes, current proteomic methods may only measure a limited spectrum, often restricted to hundreds of proteins. Our solution is bridge integration[11], a method developed by Yuhan Hao in the Satija Lab, which utilizes a bridge dataset as a 'dictionary.' This dataset is a multimodal dataset that measures the two modalities we wish to integrate within the same cell, allowing us to use it as a biological "translator" to establish connections between those two modalities.*

As for other methods, I would recommend GLUE from Ge Gao's group[13]. They employ an entirely different method from ours, incorporating graph-based methods to integrate prior knowledge into the model, thereby inferring a connection between the two modalities.

Please refer to the Front Line Genomics Multi-Omics Playbook for an in-depth overview of the latest multi-omics methods, an array of applications and the most exciting integration methods available.

Single-molecule proteomic sequencing

An up-and-coming topic within sequencing is the rapid developments in the capacity for proteomic sequencing. Detecting and identifying proteins has been possible for decades using fluorescent antibodies (immunoassays) or using mass spectrometry-based methods. It is only very recently that methods have emerged that allow for high-throughput proteomics.

2023 has seen some impressive examples of large-scale proteomics work. For example, large-scale proteomics work in six human cell lines identified 1 million peptides across 17,717 human proteins (using the Thermo Fisher Scientific Orbitrap, see below) and has built a catalogue of peptides for future work[14]. Other examples have revealed unique proteomic patterns from thousands of proteins in blood plasma for lung cancer using Seer's Proteograph platform[15] and across cancer types using the Olink Proteomics Explore technology[16].

We will now briefly overview some examples of commercial proteomic sequencing platforms and kits that are available to the reader.

Quantum-Si's Platinum single-molecule protein sequencing platform works on digested individual peptides that are immobilized in wells. Fluorescently labelled N-terminal amino acid (NAA) recognizers then bind the individual peptides, and the resulting unique fluorescence signal is recorded onto a chip. By sequentially cleaving the NAA, the next amino acid is exposed, and the sequence is recorded. This allows unbiased single-molecule resolution and shows post-translational modifications, as well as detecting low abundance proteins in complex mixtures.

Quantum-Si's Platinum

SomaLogic's SomaScan kit uses Slow Off-rate Modified Aptamers (SOMAmer), which provide greater specificity compared to antibodies. From a 55 µL sample, their platform can measure 11,000 proteins using the extensive SOMAmer library and the platform has a high throughput of 1,000 samples a day.

Olink Proteomics has an Explore kit based on the Proximity Extension Assay (PEA) technology that can sequence over 5,400 proteins from 2 µL of sample. Matched antibodies carrying DNA tags bind to proteins in the sample and form a dual bond. Only these bound proteins will then have their DNA tag hybridized, this DNA is amplified and can then be read by NGS.

Seer's Proteograph

Seer's Proteograph technology harnesses engineered nanoparticles, which consist of a magnetic core and a surface that binds to proteins within a biofluid. A single nanoparticle can bind to a broad range of proteins and hence a panel of diverse nanoparticles can profile a dynamic range of proteins via mass spectrometry. Output-wise, this results in quantifying thousands of proteins in hours rather than days or weeks.

Thermo Fisher Scientific's Orbitrap Astral Mass Spectrometer is setting new standards for mass spectrometers, capable of analysing one sample in 8 minutes and identifying over 8,000 protein groups in one run. With more time, over 15,000 proteins can be detected, and this technology works for single cells too.

Orbitrap Astral Mass Spectrometer

Bruker's timsTOF Ultra uses trapped ion mobility spectrometry (TIMS) and quadruple Time-Of-Flight (TOF) technology to produce 4D-Proteomics™. This mass spectrometer can identify >5000 protein groups and >55,000 peptides at single-cell sensitivity with very high confidence (<1% FDR)

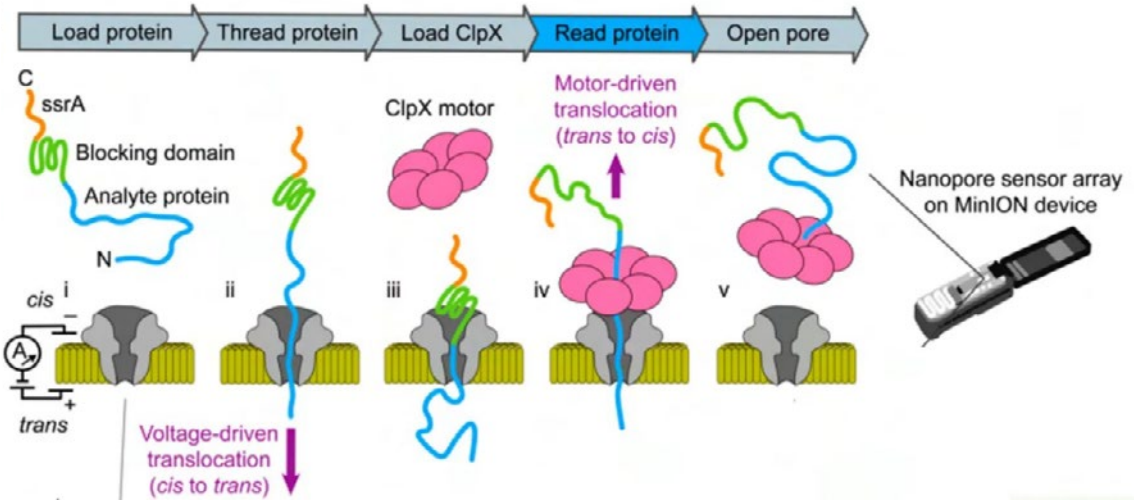
timsTOF Ultra

NANOPORE PROTEOMIC SEQUENCING

Much interest has also been expressed in repurposing nanopore technology for proteomic sequencing[17-19], among other purposes[20]. This has struggled to be developed for several reasons. Typically, nanopores are too big and lack the sensitivity required to discriminate between amino acids[21]Theoretically a better nanopore could help, but it's challenging to get high accuracy protein nanopore sequencers. Furthermore, peptides translocate too quickly through the nanopore for individual amino acids[17]

The ideal approach would be one in which proteins are unfolded, linearly translocated through the nanopore amino acid by amino acid, and the individual amino acids are recognized by the specific current signatures they produce.

FIGURE 9.2. NANOPORE PROTEIN READING USING A CLPX UNFOLDASE
Schematic of cis-based unfoldase approach on the MinION platform. Image Credit: Motone, et al. [22]



In October 2023, a multi-pass, single-molecule nanopore sequencing paper was published on bioRxiv, where long protein strands were sequenced with single-amino-acid sensitivity[22]. Their approach uses a ClpX motor to ratchet proteins through the nanopore while reading individual amino acids, with each region being read multiple times through, causing the ClpX to slip and eventually dissociate from the nanopore (see Figure 9.2). This approach seems highly promising and closer to the ideal approach of drawing proteins through the nanopore, but it does have its drawbacks (see here).

Solid-state nanopore sequencing

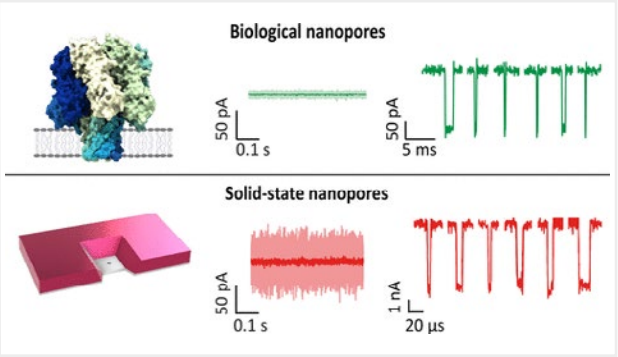
As hinted in the proteomics section above, nanopores should ideally have dimensions comparable to the analyte of interest, so that measurable changes in ionic current amplitude can be measured above the noise level when the analyte passes through20. Biological nanopores (formed by protein subunits or DNA scaffolds) have precise dimensions (1-10 nm) enabling the recognition of certain biomolecules. However, being biological, they have a relative shelf life, limited reuse potential and they're difficult to engineer[23].

Solid-state nanopores address many of these concerns since they are crafted from inorganic/plastic membranes (e.g., Si3N4). Pores are artificially created in single atom thick sheets of material and the pores can have diameters up to hundreds of nanometres wide, allowing large biomolecules and complexes to pass through. These nanopores can be constructed with several methods such as electron milling[24] and laser-based optical etching[25].

Solid-state nanopores have other benefits. For example, they could dramatically increase signal to noise ratios - one study found this to be on the order of 160-fold[26] (see Figure 9.3). With recent advances in solid-state nanopores and protein nanopore engineering, it is now possible to build artificial systems that recapitulate biological pores in vitro. The hope is for solid-state nanopores to become a powerful single-molecule detection platform that is agnostic to the nature of the sequenced molecule[27].

There are currently no viable commercial solid-state nanopore sensors, mostly due to the challenges in making the manufacturing process cost-effective[28]. But the promise of these nanopores is obvious. Oxford Nanopore Technologies recently acquired Northern Nanopore Instruments, who specialise in innovative solid-state nanopore fabrication technology. This reflects their understanding that solid-state nanopores are perhaps the future of nanopore-based sequencing. Furthermore, solid-state nanopores are an important part of the toolkit to achieve Oxford Nanopore Technologies' mission – to enable the analysis of anything, by anyone, anywhere.

FIGURE 9.3. BIOLOGICAL VS. SOLID-STATE NANOPORES SIGNAL-NOISE RATIO.
Image Credit: Fragasso, et al. [26]



Advancements in sequencing

While advancements in sequencing have been covered throughout the chapters of this report, we would like to take a moment to consider some aspects.

For one, we've seen a gradual improvement in sequencing accuracy over recent years. As already mentioned in ***Chapter 4: Finding your Sequencing Technology***, long-read methods have seen significant improvements in accuracy, bringing them to comparable accuracy as standard short-read methods (see here and here). However, some short-read sequencers have gone beyond the Q30 standard and are racing to Q40 and beyond.

Q40 is the equivalent of one error in 10,000 bases, an order of magnitude better than Q30 and is currently routinely achieved by Element Biosciences' AVITI and PacBio's Onso. In fact, this jump in quality caused initial disturbances in pipelines, such as 10x Genomics' Cell Ranger, which was aborting runs due the quality parameters being outside the standard guidelines (which was fixed very quickly with one line of code). Furthermore, Element Biosciences' very recently announced they had achieved 70% Q50 on their AVITI sequencer, and will be releasing a commercial kit to achieve this - Cloudbreak UltraQ.

Higher accuracy might make NGS more valuable for rare variant detection in cancer and improve performance in low-pass/shallow whole genome sequencing. The shift to Q40 will necessitate some computational changes, but on the whole, higher accuracy represents progress. The T2T, Human Pangenome Reference and Genome in a Bottle consortia have their sights on a loftier goal in a project nicknamed the Q100 project. This project aims to create a genome benchmark with near complete accuracy (1 error in 10 billion bases).

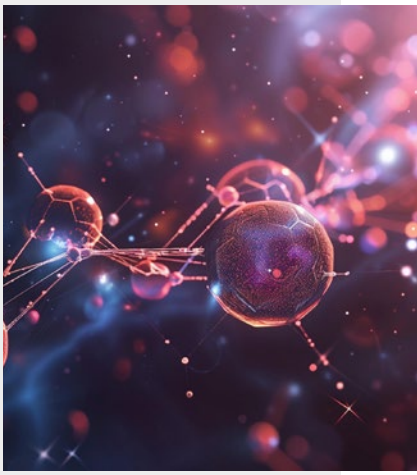
Looking ahead more broadly, a blog post from Nava Whiteford from this year details a few interesting directions that the future of sequencing could take to improve the experience of everyday scientists.

- First, is his hope for ***sequencers to become boring like qPCR machines*** - reliable and easily available. Considering how important sequencing is to single-cell, spatial, liquid biopsy etc., there would be several quality-of-life improvements if having a brand-new sequencer was no longer a big deal. Namely, one would having access to numerous vendors for consumables and easy access to the sequencing instruments (perhaps even second-hand).
- Second is for further developments in automation beyond microfluidics and pipetting robots. Instead, ***sequencing could adopt automation workflows that operate sample-to-answer***. Perhaps, platforms could integrate the automated preparation, meaning samples could be introduced into the sequencer to produce a whole genome without any hands-on preparation.

Spatial Profiling on a Sequencer?

At AGBT 2024, both Singular Genomics and Element Biosciences announced a merger of spatial biology with their sequencers. The Singular Genomics G4X will allow direct RNA sequencing, multiplex proteomics and digital H&E on the same FFPE section. The Element Biosciences AVITI[24] will allow in situ DNA, RNA, Protein and morphology measurements across 1.3 million cells per flow cell.

These platforms could mark a new era of sequencing, in which one platform is the basis of two distinct technologies.



- Third, and following on from the second, is an ***increasing use of sequencing for diagnostics***. Illumina reports that clinical applications are ~50% of their market. Harnessing the human genome is arguably the future of healthcare, and in recognition of this, research published this year from the 100,000 Genomes cancer program has objectively showed that NGS yielded more comprehensive information than cancer panels across 33 types of tumour[29]. Additionally, the UK Biobank recently released data for ½ million genomes for worldwide use, in recognition of the clinical value of these data.

We asked a few of our contributors what they thought might be on the horizon that will change the sequencing world.

XINKUN WANG

Director, NUSeq Core Facility, Center for Genetic Medicine & Research Associate Professor, Department of Cell & Developmental Biology, **Northwestern University**

FLG: Is there something on the horizon that could change the sequencing world?

Xinkun: *There are three trends that could change the sequencing world. Firstly, the emerging adoption of sequencing for areas other than genomics is very exciting. For example, systems such as those from Olink and SomaLogic use sequencing for detection of thousands of proteins simultaneously. Secondly, the further reduction in short-read sequencing cost to under $100 per genome will continue to democratize sequencing to have a far-reaching societal impact. Last but not least, long-read sequencing technologies continue to improve on both data throughput and accuracy. Once long-read sequencing reaches the same throughput, accuracy and cost effectiveness as short-read sequencing, it will change the sequencing world even more, based on the fact that long-reads carry more bio-information than short reads.*

DAVID BAKER

Head of Sequencing
The Quadram Institute

FLG: Is there something researchers ask for that you wish you could provide?

David: *I've always thought that a lot of universities and clinical settings would benefit from a mobile sequencing unit housed in a large vehicle with short and long-read technologies. This could rock up with advance notice on a Monday, and researchers could provide samples and have data by the Friday, and even some analysis to boot. It could also deal with local pathogen outbreaks etc., and travel to unusual locations abroad.*

FLG: Is there something on the horizon that could change the sequencing world?

David: *A sequencer the size of a microchip that can sequence DNA without library prep (or at least minimal) and be used in many applications. I have this vision somewhere in the future whereby a mini real-time sequencer can be implanted in people to continually detect circulating tumours in the blood stream and virtually irradicate terminal cancer due to extremely early detection.*

"I'VE ALWAYS THOUGHT THAT A LOT OF UNIVERSITIES AND CLINICAL SETTINGS WOULD BENEFIT FROM A MOBILE SEQUENCING UNIT HOUSED IN A LARGE VEHICLE WITH SHORT AND LONG-READ TECHNOLOGIES."

Chapter 9 references

1. Babu, M. & Snyder, M. **Multi-omics profiling for health**. *Molecular & Cellular Proteomics* **22**, 100561 (2023).
2. Battle, A. *et al.* **Impact of regulatory variation from RNA to protein**. *Science* **347**, 664-667 (2015).
3. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. **Methods and applications for single-cell and spatial multi-omics.** *Nature Reviews Genetics*, 1-22 (2023).
4. Baysoy, A., Bai, Z., Satija, R. & Fan, R. **The technological landscape and applications of single-cell multi-omics.** *Nature Reviews Molecular Cell Biology*, 1-19 (2023).
5. Li, X. **Harnessing the potential of spatial multiomics: a timely opportunity.** *Signal Transduction and Targeted Therapy* **8**, 234 (2023).
6. Zhang, D. *et al.* **Spatial epigenome–transcriptome co-profiling of mammalian tissues.** *Nature* **616**, 113-122 (2023).
7. Liu, Y. *et al.* **High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue**. *Cell* **183**, 1665-1681.e18 (2020).
8. Liu, Y. *et al.* **High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq.** *Nature Biotechnology* (2023).
9. Roychowdhury, R. *et al.* **Multi-omics pipeline and omics-integration approach to decipher plant's abiotic stress tolerance responses.** *Genes* **14**, 1281 (2023).
10. Argelaguet, R., Cuomo, A.S.E., Stegle, O. & Marioni, J.C. **Computational principles and challenges in single-cell data integration.** *Nature Biotechnology* **39**, 1202-1215 (2021).
11. Hao, Y. *et al.* **Dictionary learning for integrative, multimodal and scalable single-cell analysis**. *Nature Biotechnology* (2023).
12. Ghazanfar, S., Guibentif, C. & Marioni, J.C. **Stabilized mosaic single-cell data integration using unshared features.** *Nature Biotechnology* (2023).
13. Cao, Z.-J. & Gao, G. **Multi-omics single-cell data integration and regulatory inference with graph-linked embedding.** *Nature Biotechnology* **40**, 1458-1466 (2022).
14. Sinitcyn, P. *et al.* **Global detection of human variants and isoforms by deep proteome sequencing**. *Nature Biotechnology* **41**, 1776-1786 (2023).
15. Donovan, M.K.R. *et al.* **Functionally distinct BMP1 isoforms show an opposite pattern of abundance in plasma from non-small cell lung cancer subjects and controls.** *PLOS ONE* **18**, e0282821 (2023).
16. Álvez, M.B. *et al.* **Next generation pan-cancer blood proteome profiling using proximity extension assay.** *Nature Communications* **14**, 4308 (2023).
17. Brinkerhoff, H., Kang, A.S.W., Liu, J., Aksimentiev, A. & Dekker, C. **Multiple rereads of single proteins at single–amino acid resolution using nanopores.** *Science* **374**, 1509-1513 (2021).
18. Yan, S. *et al.* **Single molecule ratcheting motion of peptides in a Mycobacterium smegmatis Porin A (MspA) nanopore.** *Nano letters* **21**, 6703-6710 (2021).
19. Chen, Z. *et al.* **Controlled movement of ssDNA conjugated peptide through Mycobacterium smegmatis porin A (MspA) nanopore by a helicase motor for peptide sequencing application.** *Chemical science* **12**, 15750-15756 (2021).
20. Ying, Y.-L. *et al.* **Nanopore-based technologies beyond DNA sequencing.** *Nature Nanotechnology* **17**, 1136-1146 (2022).
21. Timp, W. & Timp, G. **Beyond mass spectrometry, the next step in proteomics.** *Science Advances* **6**, eaax8978 (2020).
22. Motone, K. *et al.* **Multi-pass, single-molecule nanopore reading of long protein strands with single-amino acid sensitivity.** *bioRxiv*, 2023.10.19.563182 (2023).
23. Xue, L. *et al.* **Solid-state nanopore sensors.** *Nature Reviews Materials* **5**, 931-951 (2020).
24. Storm, A.J., Chen, J.H., Ling, X.S., Zandbergen, H.W. & Dekker, C. **Fabrication of solid-state nanopores with single-nanometre precision.** *Nature Materials* **2**, 537-540 (2003).
25. Gilboa, T., Zrehen, A., Girsault, A. & Meller, A. **Optically-monitored nanopore fabrication using a focused laser beam.** *Scientific Reports* **8**, 9765 (2018).
26. Fragasso, A., Schmid, S. & Dekker, C. **Comparing current noise in biological and solid-state nanopores**. *ACS Nano* **14**, 1338-1349 (2020).
27. Lindsay, S. **The promises and challenges of solid-state sequencing.** *Nat Nanotechnol* **11**, 109-11 (2016).
28. Liu, H., Zhou, Q., Wang, W., Fang, F. & Zhang, J. **Solid-state nanopore array: manufacturing and applications**. *Small* **19**, 2205680 (2023).
29. Sosinsky, A. *et al.* **Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme.** *Nature Medicine* (2024).

Front Line Genomics

frontlinegenomics.com

info@frontlinegenomics.com
@FLGenomics
@frontlinegenomics

Festival of Genomics: festivalofgenomics.com