



USAID
FROM THE AMERICAN PEOPLE

RISE
Reaching Impact, Saturation,
and Epidemic Control

Reaching Impact, Saturation, and Epidemic Control
(RISE)

**A comprehensive manual for Next-
Generation Sequencing with a focus
on SARS-CoV-2**



October 2022

PREFACE

The COVID-19 pandemic, which has ravaged the world for the last two years continues to be a major threat to public health. In India, an estimated 44 million people have been infected and ~529,000 people have died [as of 25th of October 2022, (WHO COVID-19 Dashboard; <https://covid19.who.int/>)], placing the country second on the global list of highest confirmed cases and third on the list of highest COVID-19 related deaths. Despite the roll-out of vaccines, which currently provides coverage for ~66% of the country's population, COVID-19 infections are still increasing in the country, with thousands of active cases reported daily.

Active SARS-CoV-2 genome surveillance forms a crucial part towards curbing COVID-19 disease. As part of its commitment to strengthening healthcare systems to combatting COVID-19 disease, USAID RISE is developing and enhancing the Next Generation Sequencing (NGS) capacity in India to increase SARS-CoV-2 genome surveillance. 12 model NGS equipped laboratories are in the process of being established across the RISE supported states. The Oxford Nanopore Technologies Limited developed MinION is perfectly suited to the needs in India owing to its low cost, portability, and the limited and variable sequencing volume in the public sector.

This guide is designed to be an easy-to-use manual and resource for laboratories that will be involved in the sequencing of SARS-CoV-2 genomes derived from COVID-19 positive samples in India. The manual is divided into six chapters, providing an overview of the various nucleic acid sequencing techniques and technologies. Since the USAID RISE supported NGS laboratories are equipped with the Oxford Nanopore Technologies Limited MinION device specifically to conduct SARS-CoV-2 genome surveillance, the manual provides comprehensive resources to explain SARS-CoV-2 RNA genome nucleic acid sample handling and processing methodologies using the ONT MinION. Additionally, this document also serves as a “tool-kit” to support onsite training with the inclusion of Standard Operating Procedures (SOPs) and protocols for library preparation and SARS-CoV-2 sequencing, as well as complete workflows on available bioinformatic tools and pipelines for processing and analysing raw sequenced data. Lastly, some international and local quality management SOPs and guidelines for genomic surveillance have been added to the Annexures to assist with NGS implementation for public health.

Please keep in mind that due to the rapidly evolving nature of ONT's nanopore technology, its application in public health laboratories is yet to mature, and as a result, this document will need to be frequently amended. [FIND makes no warranty or guarantee, for any report, data or information provided herein]. Thus, the user of this document bears the sole responsibility of ensuring that their laboratory practices meet the appropriate country-specific standards.

This document has been prepared in partnership with subject matter experts in the fields of genomics, and laboratory sciences, and has been reviewed by the National Institute of Virology - ICMR, Pune, INDIA.

Disclaimer: Images used herein have been either sourced from the public domain, creative commons or are in accordance with fair use terms under the copyright law, and **not for distribution**.

Any trade/product names used in this document is for identification purposes only and does not imply endorsement by the USAID RISE.

The use, duplication or distribution of this document or any part thereof is prohibited without the written permission of the USAID RISE. Unauthorized use may violate copyright laws and result in civil and/or criminal penalties.

EXECUTIVE SUMMARY

The first cases of novel coronavirus disease 2019 (COVID-19) were reported by the World Health Organization (WHO) in December 2019. Since then, there have been almost 625 million reported cases of COVID-19, and more than 6 million deaths due to this disease. COVID-19 is caused by a virus known as SARS-CoV-2. The COVID-19 pandemic is one of the most severe public health catastrophes the world has seen.

In the early stages of the pandemic there were few tools at our disposal to tackle the spread of the virus. The only control measures were non-pharmaceutical interventions (NPIs), such as social distancing, wearing of face masks and, most extreme of all, lockdowns. While such measures can reduce the transmission of COVID-19, they also have detrimental effects on mental health, children's education, and a country's economy.

We now have a wider range of tools at our disposal. There has been some progress in treatment of COVID-19 with the availability of new and repurposed drugs. Importantly, several effective vaccines have been developed, which indicates that there may be light at the end of the pandemic tunnel. However, despite the medical and technological advances that have been made, their rollout globally has been uneven, and there are considerable issues in terms of access, especially to vaccines. Therefore, diagnostics remain a key weapon in the fight against COVID-19.

Sequencing is a process used to decipher and interpret the genetic makeup of a biological organism; Next Generation Sequencing (NGS) are the available high-throughput, rapid, and scalable sequencing technologies used to determine the order of nucleotides present in DNA or RNA sequences of complete genomes or their parts. Applying NGS techniques enable rapid identification of unknown pathogens, discovery of genetic variations, and molecular understanding of disease-causing pathogens, to inform the development and utility of tests, treatments, and vaccines. An increasingly critical application of sequencing is genomic surveillance, which uses sequenced data from outbreak causing pathogens to identify them, and to understand how pathogens are introduced and spread through a population.

NGS-based diagnostic tests for COVID-19 became available from June 2020 and are not widely used. As the COVID-19 pandemic progresses, both knowledge of the disease and virus, and strategies for managing infection rate and reducing transmission, are evolving.

Indian SARS-CoV-2 Genomics Consortium (INSACOG) was established to expand Whole Genome Sequencing of SARS-CoV-2 across the nation, aiding our understanding of how the virus spreads and evolves. The Consortium initially started with a network of ten regional genome sequencing laboratories spread across the country and has now expanded to 54 additional INSACOG Genome

Sequencing Laboratories (IGSLs), that are mapped to most States and UTs to facilitate a smooth processing of all available positive samples.

This toolkit summarizes our current knowledge of NGS technology development, protocols, and best practices. It begins with an overview of the human genome project, followed by an in-depth look at various sequencing platforms, evolution, and their comparisons, along with a complete NGS workflow process. There are chapters on COVID-19 genome surveillance, sequencing protocols using Nanopore technology, and an overview of the available easy to use bioinformatic tools developed specifically for biologists, besides the myriad conventional tools. The final chapter provides details on setting up of an NGS lab and the associated quality control procedures required in the laboratory.

The audience of this toolkit are medical and scientific personnel new to NGS technologies and is focused to SARS-CoV-2 sequencing. This will be a useful resource material during training and will serve as a reference for laboratories involved in SARS-CoV-2 sequencing. It is also a comprehensive resource for all stakeholders involved in the diagnostic laboratory ecosystem, including policymakers, laboratory managers, laboratory technicians, and data analysts.

ACKNOWLEDGMENTS

Reaching Impact, Saturation, and Epidemic Control (RISE) is a 5-year global project funded by the U.S. Agency for International Development (USAID). RISE works with countries to achieve a shared vision of attaining and maintaining epidemic control, with stronger local partners capable of managing and achieving results through sustainable, self-reliant, and resilient health systems by 2024.

We are grateful to the team at FIND, the global alliance for diagnostics, for insights derived from its vast experience in laboratory management. Further, this volume would not have been possible without the support of JHPIEGO, a non-profit organization for international health affiliated with Johns Hopkins University.

CONTRIBUTORS:

FIND, the global alliance for diagnostics

Dr Sanjay Sarin, Vice President, Access

Dr Sarabjit Chadha, Regional Technical Director-India & South-East Asia

Dr Anita Suresh, Deputy Director of Genomics & Sequencing, FIND Geneva

Authors

Dr Oluwaseun Oyewole, External Scientific Writer

Dr Lakshmi Soundararajan, Senior Microbiologist

Mr Nithyanandan T, Senior Technical Associate

Ms Preetishirin Katapur, Deputy Project Manager

Reviewers

Dr Varsha Potdar, Scientist 'E' & Influenza Group Leader, ICMR- National Institute of Virology

Dr Prashant Singh, Senior Advisor Laboratory Strengthening, JHPIEGO

Dr Swapna Uplekar, Senior Scientific Officer, FIND Geneva

Technical team, FIND Geneva

JHPIEGO, John Hopkins University Affiliates (India)

Dr Vineet Kumar, Chief of Party, RISE

Dr Nochiketa Mohanty, Deputy Chief of Party, RISE

Intended Audience(s)

This comprehensive manual is intended for diagnostic laboratory personnel planning to use NGS for SARS-CoV-2 genome surveillance, health policy formulators and decision makers, including members of committees appointed for developing or updating a SARS-CoV-2 sequencing facility, and other relevant stakeholders who influence SARS-CoV-2 sequencing related activities, such as laboratory personnel, laboratory managers, administrators, and program managers working in SARS-CoV-2 diagnostics.

This document was made possible with support from the United States Agency for International Development (USAID) funded RISE program, under the terms of the cooperative agreement 7200AA19CA00003. The contents are the responsibility of the RISE program and do not necessarily reflect the views of USAID or the United States Government.

USAID Mission Activity Manager:

USAID Agreement Officer Representative:

Submitted by:

FIND, in collaboration with: JHPIEGO



icmr | **NIV**
INDIAN COUNCIL OF
MEDICAL RESEARCH
Serving the nation since 1911
NATIONAL INSTITUTE
OF VIROLOGY

आई सी एम आर - राष्ट्रीय विषाणु विज्ञान संस्थान

भारतीय आयुर्विज्ञान अनुसंधान परिषद
स्वास्थ्य अनुसंधान विभाग
स्वास्थ्य एवं परिवार कल्याण मंत्रालय, भारत सरकार

ICMR - NATIONAL INSTITUTE OF VIROLOGY

Indian Council of Medical Research
Department of Health Research
Ministry of Health & Family Welfare, Govt. of India

20 - ए. डा. आंबेडकर मार्ग, पोस्ट बॉक्स संख्या 11, पुणे - 411 001, भारत. 20-A. Dr. Ambedkar Road, Post Box No. 11, Pune 411 001, India.
Tel. : NIV Camp +91-020-26127301, 26006290, Fax : 26122669, 26126643 / NIV Pashan +91-020-26006390 Fax : No. 25871895 / 25870640
E-mail : director.niv@icmr.gov.in Website : www.niv.co.in

Dr Sarabjit S Chadha
Regional Technical Director – (India and S.E. Asia)
Sarabjit.Chadha@finddx.org

17 Oct. 2022

Sub: Regarding comprehensive manual on the NGS principles and protocols with a dedicated chapter on SARS-CoV-2 sequencing

Dear Sir,

With reference to the invitation dated August 9 2022 the Comprehensive manual on the SARS CoV 2 prepared by USAID-RISE was reviewed for technical correctness and provided couple of suggestions to improve the quality of the manual. All the suggestions were well taken and appropriately incorporated in the final version.

The NGS manual found to be resourceful and covers all the important aspects of sequencing. It also gives comprehensive information of evolution of sequencing technique from Sanger sequencing to the third generation Next generation sequencing. The manual is specially designed for new beginners in NGS technology and that to the SARS COV 2 sequencing. This will be truly useful resource material during training and the reference for the laboratories involved in SARS CoV 2 sequencing.

It is also recommended to update the platform specific protocols timely if applicable and any relevant feedback received from the laboratories so that the manual remains in demand in future time.

Varsha Potdar, PhD,
Scientist 'E' & Influenza Group Leader
ICMR- National Institute of Virology
National Influenza centre

विश्व स्वास्थ्य संघटन

उभरते वायरल संक्रमणों का सहयोग केन्द्र
राष्ट्रीय शीतज्वर केन्द्र
पोलिओ, खसरा एवं रुबेला के लिए रेफरल प्रयोगशाला



WORLD HEALTH ORGANIZATION

Collaborating Centre for Emerging Viral Infections
National Influenza Centre
Referral Lab for Polio, Measles and Rubella

LIST OF ABBREVIATIONS

NGS	Next Generation Sequencing
HGP	Human Genome Project
DOE	Department of Energy
NIH	National Institute of Health
HUGO	Human Genome Organization
DNA	Deoxyribonucleic acid
ELSI	Ethical, Legal and Social Issues
NHGRI	National Human Genome Research Institute
bp	base pairs
kb	kilobases
PCR	Polymerase Chain Reaction
PacBio	Pacific Biosciences
SMRT	Single Molecule Real Time
ONT	Oxford Nanopore Technology
dNTPs	deoxyribonucleotide triphosphates
ddNTPs	di-deoxyribonucleotide triphosphates
ABI	Applied Biosystem
SBS	Sequencing by Synthesis
cPAL	Combinatorial Probe Anchor Ligation
DNB	DNA Nanoball
BGI	Beijing Genomics Institute
ZMW	Zero-Mode Waveguide
CLR	Continuous Long Read
CCS	Circular Consensus Sequence
PE	Paired end
SE	Single end
Gb	Gigabyte
RNA	Ribonucleic acid
UV	Ultraviolet

QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
DNA-seq	DNA Sequencing
WGS	Whole Genome Sequencing
WES	Whole Exome Sequencing
TS	Targeted Sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
INDEL	Insertion/Deletion
RNA-Seq	RNA Sequencing
WTS	Whole Transcriptome Sequencing
mRNA	messenger RNA
cDNA	complementary DNA
rRNA	ribosomal RNA
WGBS	Whole Genome Bisulfite Sequencing
IVD	<i>In vitro</i> diagnostics
dsDNA	double stranded DNA
HMW	High Molecular Weight
COVID-19	coronavirus disease 2019
WHO	World Health Organisation
VOI	Variant of Interest
VOC	Variant of Concern
GISAID	Global Initiative on Sharing All Influenza Data
CT	Cycle Threshold
HPC	High Performance Computing
BAM	Binary Alignment Map
VCF	Variant Call Format
NCBI	National Center for Biotechnology Information
GFF	General Feature Format
GTF	Gene Transfer Format

BLAST	Basic Local Alignment Search Tool
APHL	Association of Public Health Laboratories
CDC	Centers for Disease Control and Prevention
QMS	Quality Management System
INSACOG	Indian SARS-CoV-2 Genomic Consortia

LIST OF FIGURES

FIGURE 1. HIERARCHICAL SHOTGUN SEQUENCING	2
FIGURE 2. IDENTIFICATION OF DRUG TARGETS SINCE THE HUMAN GENOME PROJECT.....	4
FIGURE 3. STRUCTURE OF THE HUMAN CHROMOSOME	6
FIGURE 4. BASE PAIRING IN DNA	7
FIGURE 5. TIMELINE OF THE EVOLUTION OF DNA SEQUENCING	8
FIGURE 6. THREE GENERATIONS OF DNA SEQUENCERS	9
FIGURE 7. COST OF SEQUENCING A HUMAN GENOME (2001-2021).....	10
FIGURE 8. SANGER SEQUENCING	12
FIGURE 9. PRINCIPLE OF PYROSEQUENCING	13
FIGURE 10. ION TORRENT SEQUENCING.....	14
FIGURE 11. ILLUMINA SEQUENCING	15
FIGURE 12. DNA NANOBALL SEQUENCING.....	16
FIGURE 13. PACBIO SMRT SEQUENCING	17
FIGURE 14. PACBIO CCS GENERATION.....	18
FIGURE 15. NANOPORE SEQUENCING.....	19
FIGURE 16. PAIRED-END VS SINGLE-END SEQUENCING	26
FIGURE 17. NEXT GENERATION SEQUENCING WORKFLOW	32
FIGURE 18. NGS LIBRARY PREPARATION.....	34
FIGURE 19. FRAGMENT ANALYSIS OF LIBRARIES BY CAPILLARY-BASED ELECTROPHORESIS.....	36
FIGURE 20. WORKFLOWS FOR TARGET ENRICHMENT METHODS.....	38
FIGURE 21. LIBRARY CONSTRUCTION OF RNA-SEQ.....	40
FIGURE 22. METAGENOMICS SEQUENCING.....	41
FIGURE 23. BISULFITE SEQUENCING.....	43
FIGURE 24. ADAPTER LIGATION (ILLUMINA).....	44
FIGURE 25. PACBIO SMRTBELL TEMPLATE PREPARATION	46
FIGURE 26. SARS-CoV-2 AND ITS LIFE CYCLE	50
FIGURE 27. TIMELINE TO EMERGENCE OF SARS-CoV-2 VARIANTS	52
FIGURE 28. WHOLE GENOME SEQUENCING MULTIPLEX TILING APPROACH.....	55
FIGURE 29. ARTIC CLASSIC PROTOCOL FOR SARS-CoV-2 GENOME.....	57
FIGURE 30. MIDNIGHT PROTOCOL FOR SARS-CoV-2 GENOME SEQUENCING	58
FIGURE 31. A NEXT GENERATION SEQUENCING BIOINFORMATICS WORKFLOW	62
FIGURE 32. AN EXAMPLE FASTA FILE.....	64
FIGURE 33. AN EXAMPLE FASTQ FORMAT	65
FIGURE 34. AN EXAMPLE SAM FORMAT	65
FIGURE 35. AN EXAMPLE VCF FORMAT.....	66
FIGURE 36. AN EXAMPLE GFF FORMAT	67
FIGURE 37. A INTERARTIC-BASED WORKFLOW FOR SARS-CoV-2 SEQUENCING ANALYSIS SARS-CoV-2 SEQUENCING ANALYSIS.....	73
FIGURE 38. LAYOUT OF A NGS LABORATORY	78
FIGURE 39. INSACOG NETWORK.....	84

LIST OF TABLES

TABLE 1. COMPARISON OF SANGER SEQUENCING AND NGS	20
TABLE 2. APPLICATIONS OF NGS	21
TABLE 3. COMPARISON OF SHORT READ SEQUENCERS	28
TABLE 4. COMPARISON OF LONG READ SEQUENCING PLATFORMS	29
TABLE 5. SHORT READ VS. LONG READ SEQUENCING	29
TABLE 6. CLASSIFICATION OF SARS-CoV-2 VARIANTS OF CONCERN	53
TABLE 7. EXAMPLES OF SARS-CoV-2 SEQUENCING CONSORTIA	53
TABLE 8. COMPARISON OF ONT INSTRUMENT MODELS	56
TABLE 9. A SELECTION OF BIOINFORMATIC TOOLS FOR QC ANALYSIS	68
TABLE 10. SELECTED BIOINFORMATIC TOOLS AND PIPELINES FOR NGS DATA ANALYSIS	69
TABLE 11. SELECTION OF BIOINFORMATIC TOOLS AND SOFTWARE FOR NGS DATA VISUALIZATION	71
TABLE 12. GUIDANCE FOR ESTABLISHMENT OF PERFORMANCE CHARACTERISTICS IN NGS LABORATORIES	81
TABLE 13. DO'S AND DON'TS FOR NGS WORKFLOWS	89

Table of Contents

PREFACE	ii
EXECUTIVE SUMMARY	iv
ACKNOWLEDGMENTS	vi
LIST OF FIGURES	xii
LIST OF TABLES	xiii
CHAPTER 1: INTRODUCTION TO NEXT GENERATION SEQUENCING	I
1.1. The Human Genome Project (1990-2003)	1
1.2. Evolution of DNA Sequencing	7
1.3. Sequencing Methods.....	11
1.3.1. First Generation Sequencing.....	11
1.3.2. Second Generation Sequencing.....	13
1.3.3. Third Generation Sequencing.....	17
1.4. Sanger Sequencing vs. NGS	20
1.5. Applications of NGS.....	21
1.6. References	22
CHAPTER 2: COMPARISON OF NGS PLATFORMS	26
2.1. Common parameters across NGS platforms	26
2.2. Short Read Sequencing Platforms.....	28
2.3. Long Read Sequencing Platforms.....	29
2.4. Short Read vs. Long Read Sequencing Technologies	29
2.5. References	30
CHAPTER 3: NGS Workflow	32
3.1. Overview	32
3.2. Library Construction: General steps in the workflow	33
3.3. Library Construction for Different NGS Assays	37
3.4. Characteristic Features of NGS Platforms.....	43
3.5. References	47
CHAPTER 4: DISEASE OUTBREAK AND COVID-19 PANDEMIC	50
4.1. An Overview of COVID-19	50
4.2. COVID-19 Surveillance	51
4.3. SARS-CoV-2 Genome Sequencing using Nanopore Technology	55
4.3.1. Nanopore Instrument Models & Throughput Capacity.....	56

4.4.	SARS-CoV-2 Sequencing Workflows	57
4.5.	References	59
CHAPTER 5: BIOINFORMATIC ANALYSIS.....		62
5.1.	Introduction	62
5.2.	File Formats for Sequences, Alignments and Annotation	64
5.3.	NGS Data Quality Control.....	67
5.4.	Tools and Pipelines for NGS Analysis	68
5.5.	NGS Data Visualization & Exploration.....	71
5.6.	SARS-CoV-2 Analysis Workflow.....	72
5.7.	Reference Databases	74
5.8.	IT Infrastructure and Data Management.....	75
5.9.	References	76
CHAPTER 6: SETTING UP AN NGS LAB		78
6.1.	Infrastructure	78
6.2.	Consumables & Equipment.....	80
6.3.	Personnel	80
6.4.	Quality Control and Validation (Process Management).....	81
6.5.	Quality Management System – Supporting Documents.....	83
6.6.	References	85
GLOSSARY OF TERMS.....		86
ANNEXURES.....		88
REFERENCES		130

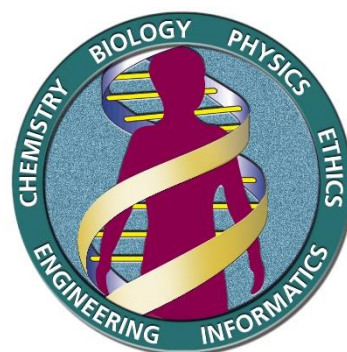
CHAPTER I

INTRODUCTION TO NEXT GENERATION SEQUENCING

CHAPTER I: INTRODUCTION TO NEXT GENERATION SEQUENCING

1.1. The Human Genome Project (1990-2003)

The Human Genome Project (HGP) was a 13-year collaborative, international research effort aimed at completely deciphering the sequence of all genes that constitute the genetic makeup of a human being. The project was coordinated by both the United States' Department of Energy (DOE) and National Institutes of Health (NIH), who [via the United States' Congress] funded many laboratories, research and academic institutions within the country to participate in the project. In addition, the Human Genome Organization (HUGO) organized the project's international collaboration, with notable support from research centres in the United Kingdom, France, Germany, Japan and China. Altogether, the HGP involved more than 2,800 researchers and ~\$2.7 million¹ in funding expenses, making it the first and one of the largest scientific megaprojects ever undertaken (NIH, 2020).



U.S. Department of Energy <<http://www.ornl.gov/hgmis>>

Goals of the HGP

The goals of the HGP were drafted into a series of 5-year plans spanning a total of 15 years. However, the HGP progressed faster than expected, resulting in a revision of the second and final 5-year plans [adapted to 1993-1998 and 1998-2003], which brought the project to completion two years ahead of schedule. A summary of the main goals for the HGP are outlined below:

- Identification of all genes that make up the human DNA
- Determining the sequences of the entire human DNA (~3 billion base pairs [bp])

¹ This amount does not represent the cost for sequencing the first human genome, but rather the total U.S. funding for a broad-spectrum of scientific activities under the HGP such as program management, technology development, genome mapping and bioethics research. The single cost of generating the first human genome sequence is estimated to be between \$500 million and \$1 billion.

- Storage of sequence information in databases
- Development of new and improved tools for data analysis
- Transferral of technologies to the private sector
- Examination of the ethical, legal and social issues of the project

Construction of the Human Genome Sequence (Phase I)

The organization of the human genomic DNA is relatively complex, as all 3 billion bases are tightly wound and packaged into single unit, thread-like structures called chromosomes, with each chromosomal unit consisting of a molecule of DNA wrapped around a core of small proteins called histones (NHGRI, 2022). Therefore, the most challenging aspect of the HGP was determining the

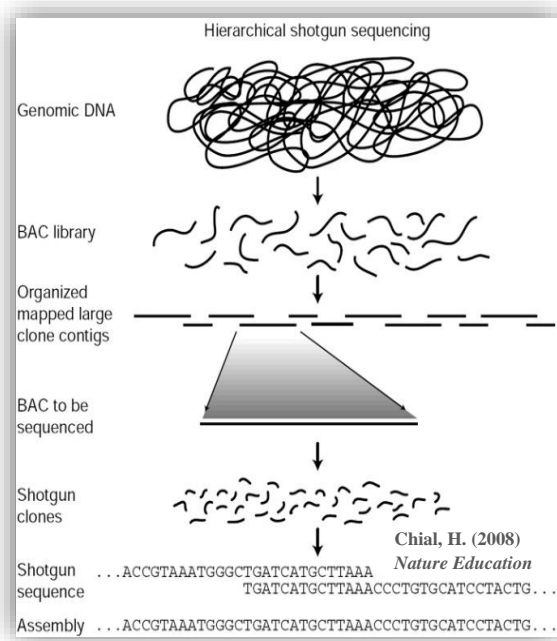


Figure 1. Hierarchical shotgun sequencing

Note. Reprinted from Chial, H. (2008). DNA sequencing technologies key to the Human Genome Project. *Nature Education*, 1(1):219. <https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/> Copyright 2008, Nature Education

order [or “sequence”] of all 3 billion bases that make up the genome (Collins & Fink, 1995). To accomplish this, the researchers first studied and elucidated the genetic makeup of several smaller, model organisms² (mouse, rat, fruit fly, round worm, yeast, and *Escherichia coli*), which then served as comparisons for the human genome. Thus, the first phase (a.k.a. shotgun phase) of the human genome sequence construction commenced using DNA extracted from blood

samples of a representative group of several anonymous volunteers. The DNA of each chromosome was then fragmented into smaller

sizes (~150,000 – 200,000 bp in size), ligated into bacterial artificial chromosome (BAC) vectors and

² Model organisms are non-human species studied by researchers to understand biological processes

cloned. Clones were then mapped to determine their location in the genome, further fragmented, subcloned into BAC vectors to create a shotgun library and sequenced. An ordered set of DNA sequences spanning each BAC clone was computationally generated based on the overlapping shotgun clones and a contiguous sequence was established for all human chromosomes (i.e., 22 autosomes, X, and Y chromosomes) (Chial, 2008). This approach was called the hierarchical shotgun method (**Figure I**).

Finishing the Human Genome Sequence (Phase 2)

Following a 90% completion of the human genome, phase one produced a first draft or rough draft, which was published in the journal *Nature* in February 2001 (Lander et al., 2001). Phase two (a.k.a. finishing phase) commenced immediately to fill in gaps and produce a “finished sequence” with a 99.99% accuracy. In April 2003, a near complete version (~99%) of the genome with a higher accuracy (<1 error per 10,000 bp) was announced and published the following year (International Human Genome Sequencing, 2004). Key findings from both published versions of the genome revealed that:

- the number of protein-coding genes in human beings is much smaller (only ~20,000 – 25,000 genes) than previously hypothesized.
- segmental duplications are much higher in human beings (~5.3%) than in other smaller vertebrate mammals like mouse and rat.
- 50% of the human genome is derived from transposable elements but their activity has declined in the hominid lineage.

With the generation of the “finished sequence”, the HGP was considered “operationally finished” (International Human Genome Sequencing, 2004), having met the minimum requirements for sequence accuracy and completion, along with all projected goals, which were either already met or exceeded expectations ahead of time and under budget (Collins et al., 2003).

Impact of the HGP

One of the greatest benefits of the HGP was the promotion of a cross-disciplinary collaboration within the sciences (between biologists, physicists, engineers, computer scientists and mathematicians), which drove the development and application of sophisticated computational and mathematical approaches to biological data, resulting in the creation of open-source computational software and publicly accessible databases like the NIH's GenBank (Hood & Rowen, 2013). Furthermore, the HGP benefitted both basic biological research and clinical medicine with the provision of a human genome sequence, which has revealed the genetic basis and molecular mechanisms behind a myriad of diseases, improved our understanding of their pathology, and enabled the development of new innovative tools for their diagnosis. Scientists can now easily identify the genes responsible for dozens of inherited diseases including rare genetic disorders e.g., Huntington's Disease, and devise novel therapeutic interventions to manage and/or treat them. Also, the identification of nearly all human genes and proteins via the HGP has greatly expanded the pool of promising new drug candidates with known targets (**Figure 2**), resulting in a boom for the pharmaceutical industry in the last two decades (Gates et al., 2021).

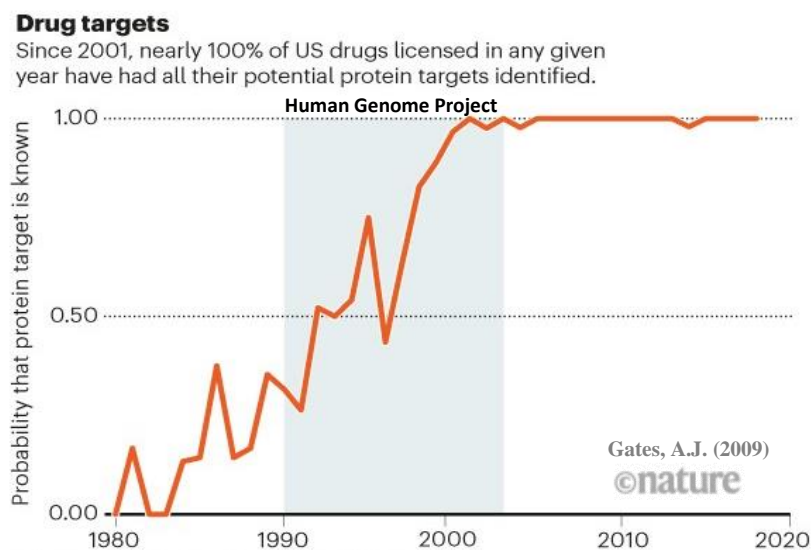


Figure 2. Identification of drug targets since the Human Genome Project

Note. Adapted from Gates, A. J., Gysi, D. M., Kellis, M., & Barabási, A. L. (2021). A wealth of discovery built on the Human Genome Project—by the numbers. *Nature* **590**, 212-215. <https://doi.org/10.1038/d41586-021-00314-6> Copyright 2021, Nature

Ethical, Legal and Social Implications program

As with all advancements in science and technology, the use and application of such technology requires ethical and moral guidance to ensure a transparent and fair process as well as maintain the privacy and confidentiality of its users. For this reason, the HGP allocated 3-5% of its annual budget to address the ethical, legal and social issues (ELSI) related to the availability of genetic information (U.S. Department of Energy, 2003). Some of these issues and the questions they raised include:

- Fairness in the use of genetic information by private and public entities (e.g., schools, insurance companies, courts, employers)
 - *Who should be given access to an individual's genetic information and how will it be used?*
- Privacy and confidentiality of genetic information
 - *Who owns and controls genetic information?*
- Commercialization of products (patents, trademarks, copyrights) and data accessibility
 - *Will patenting DNA sequences restrict their accessibility and use for other development purposes?*
- Psychological impact and stigmatization
 - *How does personal genetic information affect an individual's and society's perception of that individual?*
- Clinical issues
 - *How will genetic tests be assessed and regulated for accuracy, reliability, and utility?*
- Reproductive issues (use of genetic information in reproductive decision-making, reproductive rights, etc.)
 - *Are parents properly counselled about the risks and limitations of (complex) genetic technology?*
- Conceptual and philosophical implications (free will vs. genetic determinism, human responsibility)
 - *Are genes responsible for people's behaviours?*
- Health and environmental issues
 - *Are foods containing genetically modified organisms (GMOs)³ safe for humans and the environment?*

³ GMOs are foods created via genetic engineering.

At present, most of these issues and the questions raised by the ELSI program have already been answered with initiatives put in place to ensure safety, privacy and confidentiality of the patients or individuals involved. Nowadays, informed consent forms are **a must** for most (if not all) clinical and research procedures involving patients and their genetic information. Also, genetic tests are required to meet high-level standards and pass rigorous safety and quality-control measures before they can be marketed or used in clinical medicine.

Completion of the Human Genome Sequence

Although the human genome sequence was considered “essentially finished”, with ~99% sequenced in 2003, this metric only represented euchromatic regions, which make up about 92% of the total genome. The rest of the genome, known as heterochromatic and located primarily in centromeres and telomeres (**Figure 3**), were not sequenced during the project due to limitations of the previous technology (shotgun sequencing) in capturing the highly repetitive properties of this region (International Human Genome Sequencing, 2004). As a result, HGP researchers noted that the “true

completion” of the human genome may require approximately the same number of years it took to complete phase one, due to its dependency on the advancement of sequencing technologies, especially their capacity in assembling repetitive regions. In 2021, the telomere-to-telomere (T2T) consortium announced that the human genome was “finally complete” with all remaining gaps closed and

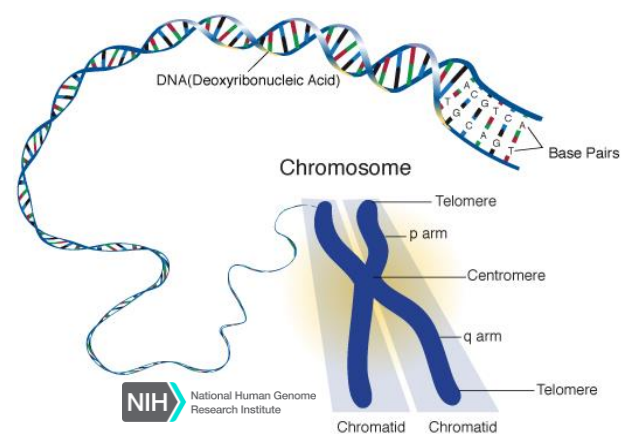


Figure 3. Structure of the Human Chromosome

Note. Reprinted from the National Institute of Health (NIH) <http://genome.gov>

repetitive regions resolved using the power of high throughput sequencing. Thus, the first complete, gapless human genome sequence was officially published in 2022, two decades after the publication of the first sequence draft (Nurk et al., 2022).

Recommended Reading

For more information on the HGP, check out the following links:

- National Human Genome Research Institute (NHGRI) < <https://www.genome.gov/human-genome-project> >
- ODE- Human Genome Project Information Archive < <http://www.ornl.gov/hgmis> >

I.2. Evolution of DNA Sequencing

The genetic information stored in our DNA can be decoded based on the nucleic acid sequence, i.e., the order of the four nucleotide bases that make up the DNA. These bases, namely adenine (A),

thymine (T), cytosine (C) and guanine (G), engage in a specific complementary pairing (A with T and C with G) on opposite strands of the DNA (**Figure 4**), and serve as building blocks

for the translation of the genetic code.

However, the precise order (sequence) of these bases in the DNA was unknown for most part of the early 20th century until a few decades

after Watson and Crick published the 3D, double-helix structure of DNA in 1953. The

first developments in sequencing were pioneered in the late 1970s by Walter Gilbert and Allan Maxam via a chemical cleavage technique (Maxam & Gilbert, 1977), and Frederick Sanger using a chain-termination (dideoxy) method (Sanger et al., 1977).

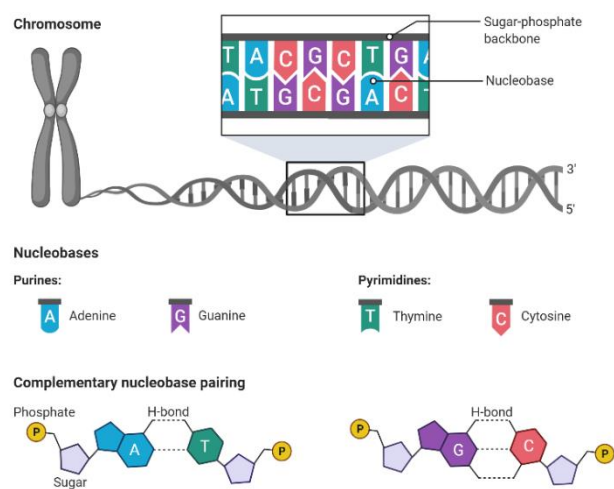
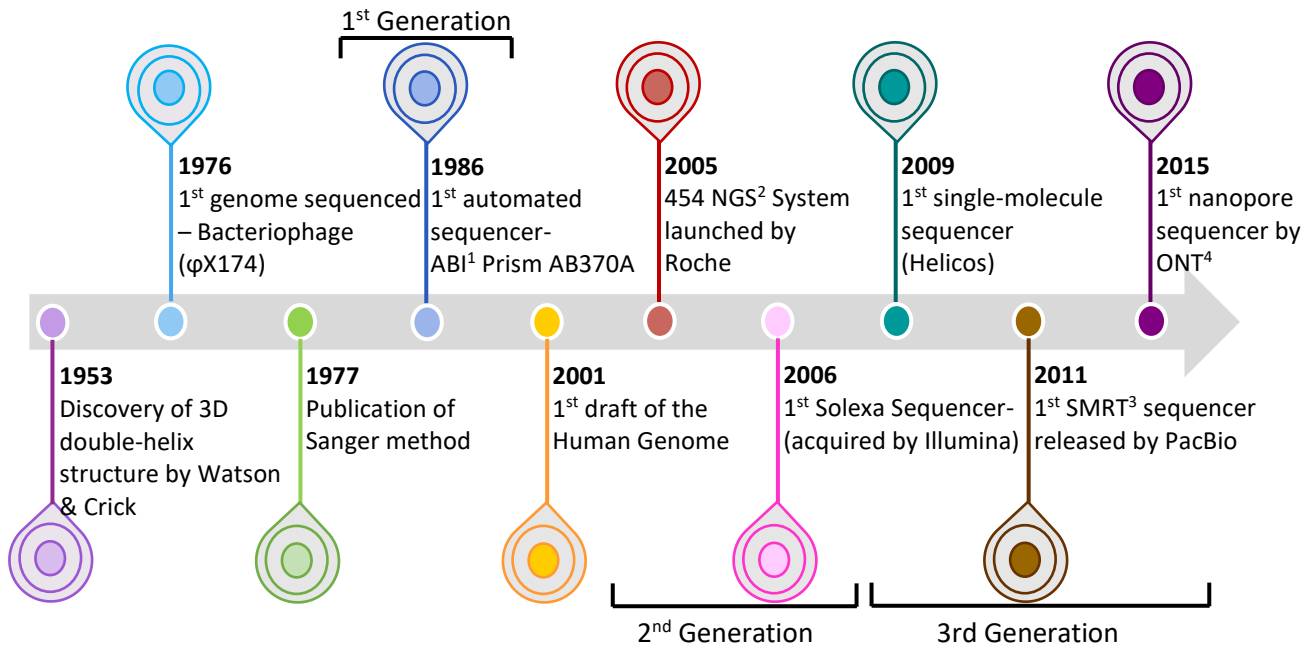


Figure 4. Base Pairing in DNA

Note. Reprinted from "DNA Structure", by BioRender, July 2020, retrieved from <https://app.biorender.com/biorender-templates/> Copyright 2022 by BioRender.



ABI¹: Applied Biosystems; NGS²: Next Generation Sequencing; SMRT³: Single molecule real-time sequencing; ONT⁴: Oxford Nanopore Technologies

Figure 5. Timeline of the evolution of DNA sequencing

Further improvements, automation and commercialization of the Sanger method produced the so-called first generation DNA sequencers (Figure 5), which were widely popular and adopted by the research community from the 1980s to the mid-2000s (Slatko et al., 2018). These new developments (fluorescent labelling, automated capillary electrophoresis) in the Sanger method increased sequencing efficiency and enabled completion of the Human Genome Project in 2003 (Pareek et al., 2011).

Nonetheless, outside of HGP, the cost of Sanger sequencing was still relatively high (\$100,000) and automation rather inefficient (~12,000 bp output per day) for widespread adoption in other large-scale sequencing projects (Adams, 2008). Additionally, the required shotgun approach of fragmentation and vector cloning for DNA fragments > 1,000 bp prior to Sanger sequencing was notably difficult, time-consuming, and labour-intensive, as shown in the HGP. These shortcomings of Sanger sequencing drove the development of high-throughput sequencing, or NGS technologies, which are scalable with a capacity for massive, parallelized sequencing, allowing entire genomes to be sequenced at once. The first class of NGS technologies introduced were the second generation or short-read sequencing

technologies (**Figure 6**), which produce millions of clonally amplified bases as short reads (50-500 bp)

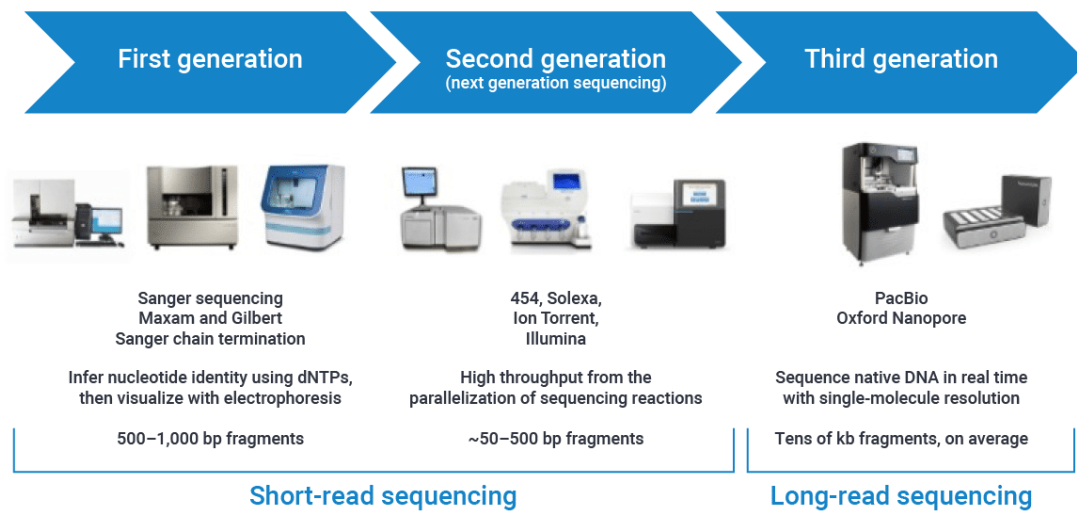


Figure 6. Three Generations of DNA Sequencers

Note. Reprinted from PacBio <https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/> Copyright 2022, PacBio

in single sequence runs using unique chemistries and detection methods (pyrosequencing, reversible-dye terminator, ligation sequencing, proton detection) (Slatko et al., 2018). The first commercialized NGS platform was developed by 454 Life Sciences (later acquired by Roche) in 2005, with the capacity of generating ~25 million bases in a 4-hour run (Margulies et al., 2005). Currently, the leading short-read sequencing technology, Illumina, boasts a production-scale maximum output of 20 billion reads with read lengths of 250 bp read under 2 days.

In the past decade, sequencing advancements have brought forth a new class of NGS technologies that produce longer reads, spanning challenging genomic regions, and providing a more complete picture of structural variants and repetitive regions, which are poorly characterized using short-read sequencers (Mantere et al., 2019). With average read lengths >20 kilobases (kb), these third generation or “long-read” sequencing technologies have revolutionized the sequencing landscape, most notably enabling the final completion of the human genome sequence (Nurk et al., 2022). Compared to short-read sequencers, which clonally amplify DNA fragments, long-read sequencing methods require no prior polymerase chain reaction (PCR) amplification thus avoiding amplification bias issues. This PCR-free process allows the long-read sequencers to target single molecules of DNA in their native state

and sequence them in real-time (Mantere et al., 2019). Nevertheless, a major caveat of long-read sequencers are their intrinsically high error rates (Delahaye & Nicolas, 2021), which restricts their use in routine laboratory practices for genotyping. However, these sequencers are in a constant state of development and already report improved error rates (Delahaye & Nicolas, 2021; Sereika et al., 2022), reaching below 1% (Wenger et al., 2019) rivaling that of the short-read sequencers (Stoler & Nekrutenko, 2021). At present, Pacific Bioscience’s (PacBio) single molecule real-time (SMRT) sequencers and Oxford Nanopore Technologies’ (ONT) sequencers dominate the long-read sequencing market (**Figure 6**).

Impact of NGS on the cost of sequencing

The emergence of NGS technologies and their continued development over the past 15 years have led to substantial reductions in the cost of sequencing, allowing great progress in the field of genomics, spurring an increase in the number of published genomes over the past decade. Although the “cost” of sequencing a genome varies based on a multitude of parameters and nuances (e.g., cost of sequencing instrument, consumables, data analysis and storage, etc.), it is evident that the refinements

and miniaturizations of NGS technologies, propelled by NHGRI’s generous funding scheme (Hayden, 2014), have dramatically driven down costs, outpacing Moore’s law⁴ (**Figure 7**), and enabling the research community to achieve its goal of sequencing individual genomes at a cost of \$1,000 by the year 2014. When

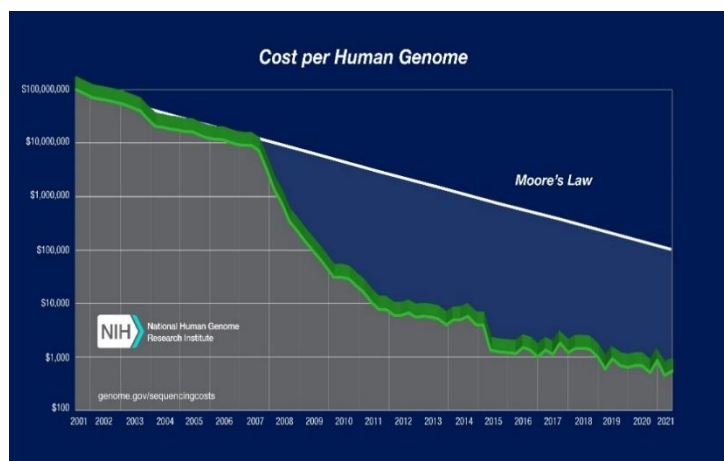


Figure 7. Cost of sequencing a human genome (2001-2021)

<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>

⁴ Moore’s law is based on the observation of Intel co-founder, Gordon Moore, who saw that a doubling in computation power in the semiconductor industry resulted in a halving of the price every two years.

compared to costs during the HGP era in the early 2000s, this translates to a 10,000-fold cost reduction (Pareek et al., 2011).

Genome sequencing costs have further plummeted owing to commercial enterprises, who offer genome sequencing services at competitive prices. As costs continue to decline, sequencing will become even more affordable and accessible for labs of all sizes and resource-limited countries.

Recommended Reading

For more information on the evolution of DNA sequencing, check out the links below:

- Hayden, E. Technology: The \$1,000 genome. *Nature* 507, 294–295 (2014).
<https://doi.org/10.1038/507294a>
- NHGRI: The Cost of Sequencing a Human Genome <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>

1.3. Sequencing Methods

1.3.1. First Generation Sequencing

In the late 1970s, Frederick Sanger and colleagues developed a novel sequencing technique by combining DNA polymerase with a mixture of standard deoxynucleotide triphosphates (dNTPs) and a low ratio of modified di-deoxynucleotide triphosphates (ddNTPs), which lack the 3' hydroxyl (OH) group needed to form a phosphodiester bond between nucleotides on opposite strands during strand elongation. Thus, once a ddNTP is incorporated into a growing strand, it causes an early termination of the polymerase extension activity, and after a series of reactions, multiple DNA fragments of varying lengths are produced (Adams, 2008) (**Figure 8**). These resulting chain-terminated DNA fragments (up to 300 bp) are then size separated using gel electrophoresis (Sanger et al., 1977).

Around the same time Allan Maxam and Walter Gilbert developed a sequencing technique, where terminally labelled DNA fragments were chemically cleaved and the reaction products (up to 100 bp) resolved by gel electrophoresis (Maxam & Gilbert, 1977). Of the two, Sanger's method became widely adopted and later automated (incorporation of fluorescently labelled ddNTPs, switching from slab gels

to capillary electrophoresis and visualization by electropherogram) and commercialized by Applied Biosystems (ABI) (**Figure 8**). Current Sanger sequencing machines can run 96 samples at a time and generate reads ranging from 500 – 1,000 bp in length, which is a significant improvement from the four parallel reaction runs used to sequence a single sample in the original method. Although current NGS methods have replaced Sanger sequencing for large-scale genomic applications, Sanger sequencing remains the benchmark for accuracy (99.99%) and is still actively used in smaller-scale projects and sequence validation studies.

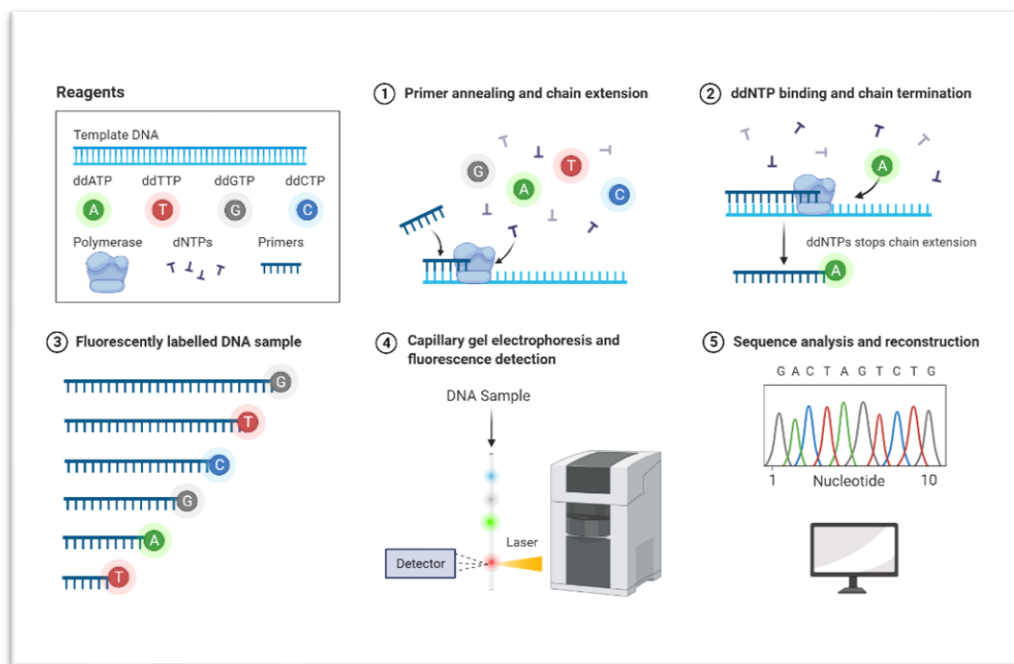


Figure 8. Sanger sequencing

Note. Reprinted from "Sanger Sequencing", by BioRender.com (April 2020). Retrieved from <https://app.biorender.com/biorender-templates> Copyright 2022 by BioRender.

- **Pyrosequencing**

Pyrosequencing is less complex, involves fewer steps, and has a superior limit of detection compared with Sanger sequencing. This method of DNA sequencing differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release and the generation of light on nucleotide incorporation, rather than chain termination with dideoxynucleotides. This method was first demonstrated by Pal Nyren in 1987 that DNA polymerization can be monitored by measuring pyrophosphate production, which can be detected by light.

In the pyrosequencing, dNTPs are sequentially dispensed into the chamber containing the template with the primer and DNA polymerase bound. When the correct complementary bases are injected, the DNA polymerase catalyzes the nucleotide addition, and an inorganic pyrophosphate (PPi) is released during the condensation reaction

(**Figure 9**). Through the sequence of solutions: ATP sulfurylase, luciferase and apyrase, and with the substrates adenosine 5' phosphosulfate (APS) and luciferin, chemiluminescent signals are produced which allows the determination of the sequence of the template (Collen et al, 2013).

Pyrosequencing is ideal for quantitative real-time applications to characterize single nucleotide polymorphisms (SNPs), insertion-deletions (indels), and unknown sequence variants, it can sensitively quantify allele frequencies and DNA methylation levels at both CpG and non-CpG (CpN) sites.

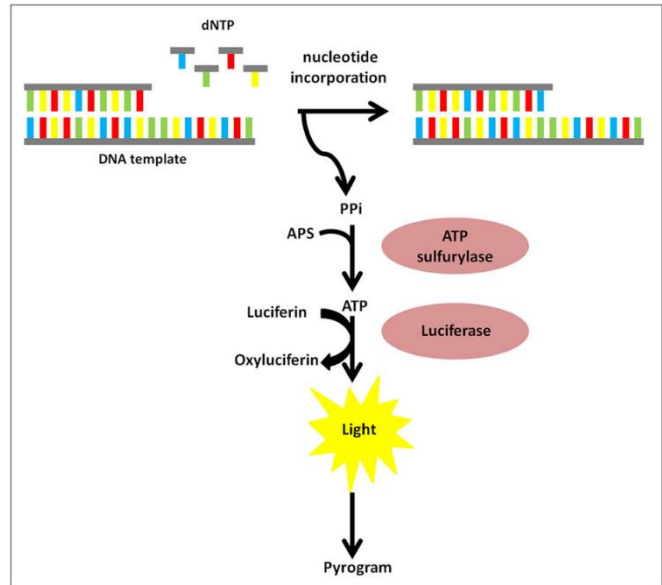


Figure 9. Principle of Pyrosequencing

Note: Reprinted from Magda Rybicka et al (2016). Current molecular methods for the detection of hepatitis B virus quasispecies. Rev Med Virol. 2016 Sep;26(5):369-81. doi: 10.1002/rmv.1897. Epub 2016 Aug 9.

1.3.2. Second Generation Sequencing

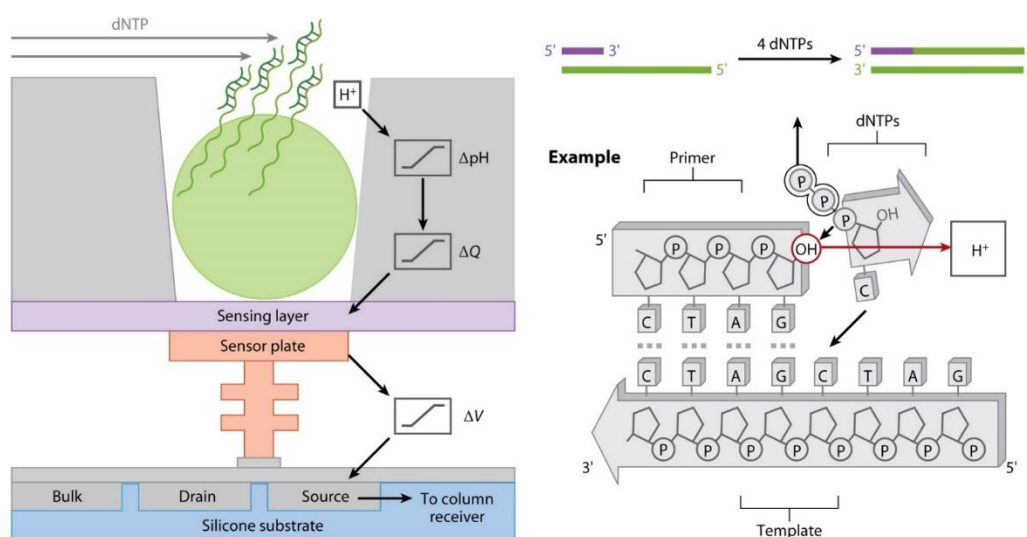
Due to a growing need to sequence larger volumes of DNA faster and at a lower cost, high throughput NGS methods and platforms were developed. The first of these NGS methods, were second generation platforms, which generated shorter reads (50-500 bp) than the Sanger sequencer, but at massively higher throughput (millions to billions of reads per run) using varied chemistries. The technology behind these chemistries are grouped into two major categories: (i) sequencing by synthesis (SBS), a development of Sanger sequencing, excluding the ddNTPs, which utilizes repeated cycles of synthesis, imaging and other methods to incorporate nucleotides into the growing chain of sequenced DNA (Slatko et al., 2018); and (ii) sequencing by hybridization and ligation, which uses

repeated hybridization and washing away of non-hybridized DNA to build larger contiguous sequences (Slatko et al., 2018). Some of the chemistries developed under these two technologies are described below:

Sequencing by synthesis

- **Ion Torrent sequencing**

The Ion Personal Genome Machine (PGM) is an ion-semiconductor sequencer, that was launched in 2010 by Ion Torrent and commercialized by Thermo Fisher Scientific. It is based on the detection of hydrogen ions, which are released as a nucleotide is incorporated into the growing DNA strand. Ion torrent reactions occur in micro chambers containing millions of wells, beneath which an ion-sensitive, complementary metal-oxide semiconductor pH sensor (CMOS) exists (Hu et al., 2021). Sequencing on an Ion torrent machine begins with the clonal amplification of DNA fragments on beads (via emulsion PCR), followed by loading into microwells with sequencing reagents. Incorporation of a single nucleotide triggers the release of a hydrogen ion that changes the pH solution of the microwells, which is quickly detected and recorded by the semiconductor




 Mardis ER. 2013. *Annu. Rev. Anal. Chem.* 6:287–303

Figure 10. Ion Torrent sequencing

Note. Reprinted from Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu Rev Anal Chem*, 6(1), 287-303. <https://doi.org/10.1146/annurev-anchem-062012-092628> Copyright 2022, Annual Reviews

pH sensor in real-time (**Figure 10**), without the need for cameras, light-source or scanner (Slatko et al., 2018).

- **Illumina sequencing**

Considered the leading short-read sequencing technology, Illumina sequencing instruments are based on a technique that uses fluorescently labelled reversible terminators, first developed by Solexa for their Genome Analyzer machine in 2006, which Illumina acquired the following year.

In this approach, DNA fragments are loaded onto a glass flow cell containing oligonucleotides and clonally amplified through a process known as “bridge amplification” (**Figure 11**). During each sequencing cycle, a fluorescently labelled reverse terminator-bound dNTP is incorporated into the growing nucleic acid chain and the resulting fluorescent signal is imaged (**Figure 11**),

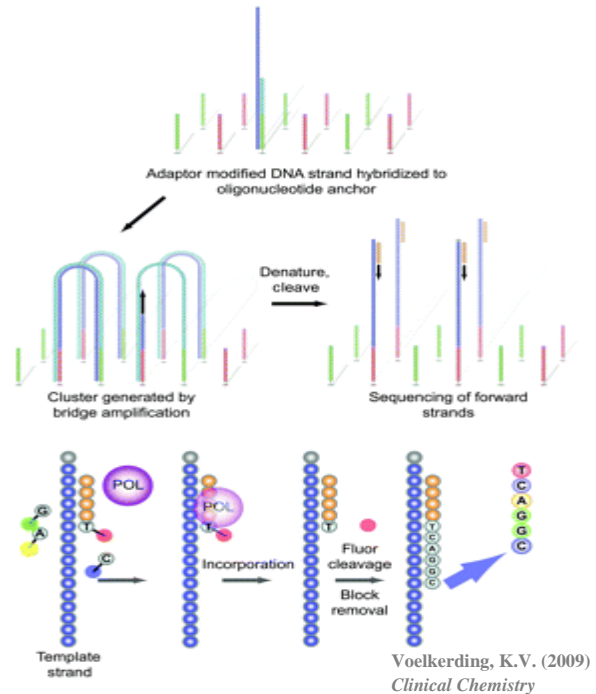


Figure 11. Illumina sequencing

Note. Reprinted from Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4), 641-658. <https://doi.org/10.1373/clinchem.2008.112789> Copyright 2009, Oxford University Press

followed by cleavage of the terminator and fluorescent dye from the incorporated dNTP to allow the next labelled dNTP to be added (Hu et al., 2021). Illumina’s reversible terminator sequencing approach makes it the most accurate NGS technology on the market with an error rate of 0.1%. In addition, sequencing on Illumina platforms occur from both ends of the DNA fragment (paired-end sequencing), which enables the generation of high-quality sequence data, in-depth coverage, and a high number of reads. A major limitation of Illumina sequencing is its read length (150-300bp), which are challenged when sequencing genomes with repetitive regions and structural variations (Goodwin et al., 2016).

Sequencing by hybridization and ligation

- **BGI/Complete Genomics (DNA Nanoball sequencing)**

In 2009, Complete Genomics developed a hybridization and ligation-based sequencing platform that incorporates rolling circle amplification and combinatorial Probe Anchor Ligation (cPAL) technology (Drmanac et al., 2010). Here, the DNA template is first fragmented and circularized, and afterwards clonally amplified into DNA nanoballs (DNBs) by rolling circle amplification, which unlike typical PCR amplification, utilizes the original circularized DNA as a template for each round of amplification, thus minimizing error accumulation and amplification biases. The clonally amplified DNBs are then loaded onto the sequencing (silicon) chip, which has a patterned binding site facilitating the distribution of the DNBs in a manner where only one DNB binds to a binding site, thereby preventing interference between fluorescent signals from multiple DNBs (**Figure 12**). Sequencing commences via cPAL, where an anchor sequence and a fluorescently labelled dNTP probe are ligated to the DNB, and the resulting fluorescent signal is captured and imaged (**Figure 12**). The probe-anchor complex is then removed

to prepare the DNB for a new probe-anchor combination and the cycle is repeated until the template is sequenced. In 2013, Beijing Genomics Institute (BGI) acquired Complete Genomics, and created a modification of the cPAL, called the combinatorial Probe Anchor Synthesis (cPAS) approach, which improves read lengths (Goodwin et al., 2016), though the exact process is not known (Korostin et al., 2020). The major advantage of DNB sequencing is that it is less expensive than Illumina, yet it provides much of the same benefits such as low per base errors for accurate variant base calls, high output and

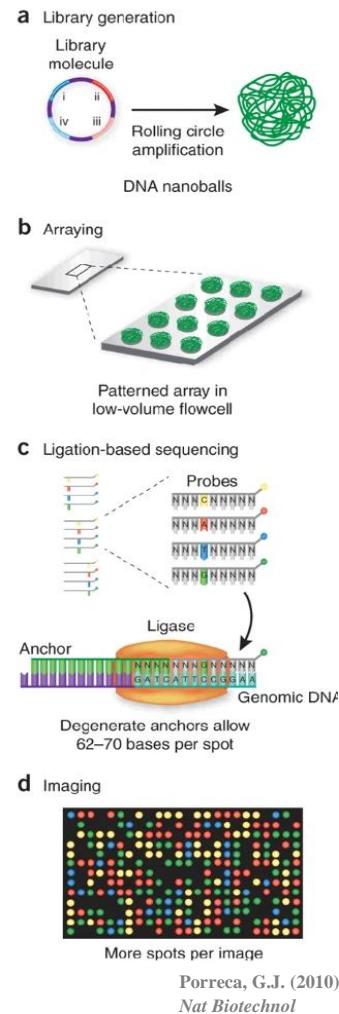


Figure 12. DNA Nanoball sequencing

Note. Adapted from Porreca, G. J. (2010). Genome sequencing on nanoballs. *Nature biotechnology*, 28(1), 43-44. <https://doi.org/10.1038/nbt0110-43>
Copyright 2010, Nature Publishing Group

coverage (Korostin et al., 2020). DNB sequencing platforms are currently marketed by MGI Tech, which is a BGI subsidiary.

1.3.3. Third Generation Sequencing

The third generation of high throughput NGS technology were developed to overcome two main limitations of the second generation sequencers, i.e., short read lengths and PCR bias introduced by clonal amplification. These newer third generation sequencers are characterized as “long-read sequencers” as they typically generate sequences with longer read lengths (>20 kb) directly from native DNA. The two main long-read technologies are PacBio’s SMRT sequencing and ONT’s nanopore sequencing, which operate on very different principles, detailed below.

SMRT sequencing

SMRT sequencing technologies were made commercially available by PacBio in 2011. The principle behind them is based on a single molecule real time sequencing by synthesis approach that occurs on a SMRT cell with millions of small pores, called zero-mode waveguides (ZMWs), where a DNA polymerase molecule and a single strand DNA are immobilized at the bottom (**Figure 13**). During the sequencing reaction, the DNA polymerase binds to the DNA fragment and incorporates fluorescently labelled dNTPs into the growing chain. Once the nucleotide is bound, imaging occurs at the bottom of the well and the fluorophore is released out of the ZMW, allowing the next nucleotide

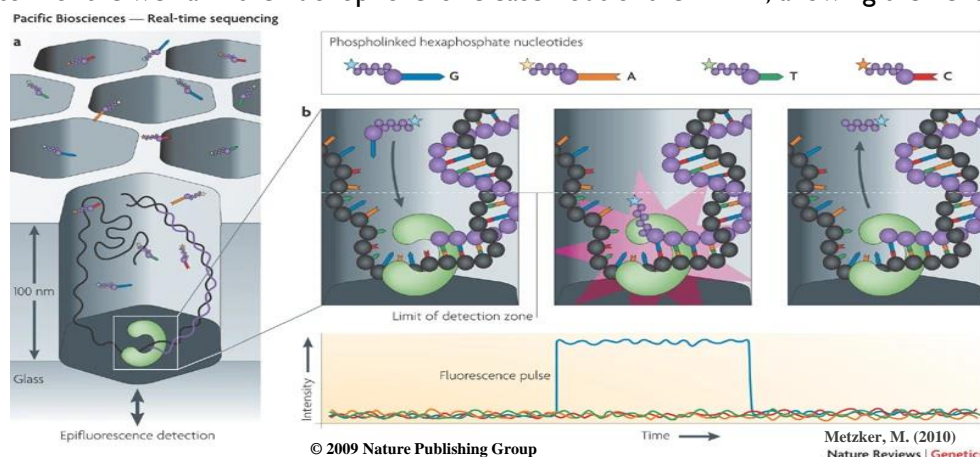


Figure 13. PacBio SMRT Sequencing

Note. Reprinted from Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1), 31-46. <https://doi.org/10.1038/nrg2626> Copyright 2009, Nature Publishing Group.

to be incorporated (Slatko et al., 2018) (**Figure 13**). The sequence obtained from each ZMW is called a continuous long read (CLR), which is further broken down into subreads and then collapsed into a circular consensus sequence (CCS, a.k.a. HIFI) (**Figure 14**).

The previous RSII platform only generated subreads, which are the partitioned format of the CLR reads, produced from a single pass of the polymerase with an error rate of 13-15%. However, the newer Sequel platforms can generate CCS/HIFI reads from the alignment of multiple subreads (**Figure 14**), which cancels out base errors and improves read accuracy to >99% (Ardui et al., 2018). With its highly accurate CCS/HIFI reads and exceptional long read lengths (>20 kb), SMRT sequencing can be used to identify DNA modifications, structural variants as well as highly repetitive regions, which is challenging with short-read sequencers.

Additional advantages include fast sample preparation, lack of PCR bias and a quick turnover rate (sequence runs are finished within a day). A major disadvantage is the high cost associated with the purchase and maintenance of the SMRT sequencers.

Nanopore sequencing

ONT's nanopore sequencing technologies were introduced to the market in 2014 and are based on the measurement of fluctuations in the ionic current, when single-stranded nucleic acids pass through biological nanopores (Amarasinghe et al., 2020). In this case, the nanopore is a α -haemolysin protein isolated from *Staphylococcus aureus*. During sequencing, DNA is bound to a motor enzyme and loaded onto a flow cell containing a membrane with thousands of nanopores (**Figure 15**). The application of an electric current initiates sequencing with the attached motor enzyme, which moves the single DNA strand through the pore. The passage of each nucleotide through the pore results in a characteristic

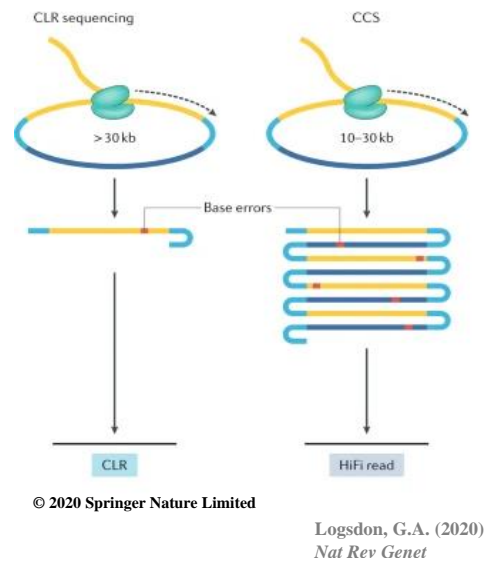


Figure 14. PacBio CCS Generation

Note. Adapted from Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10), 597-614. <https://doi.org/10.1038/s41576-020-0236-x> Copyright 2020, Springer Nature Limited

“squiggle” (disruption) in the ion current, which is detected by the sensors and recorded in real-time (Hu et al., 2021) (**Figure 15**).

Notable features of nanopore sequencing are the read length output and throughput, which can be considerably longer (up to >100 kb) and higher than those from PacBio (Jain et al., 2018). In addition,

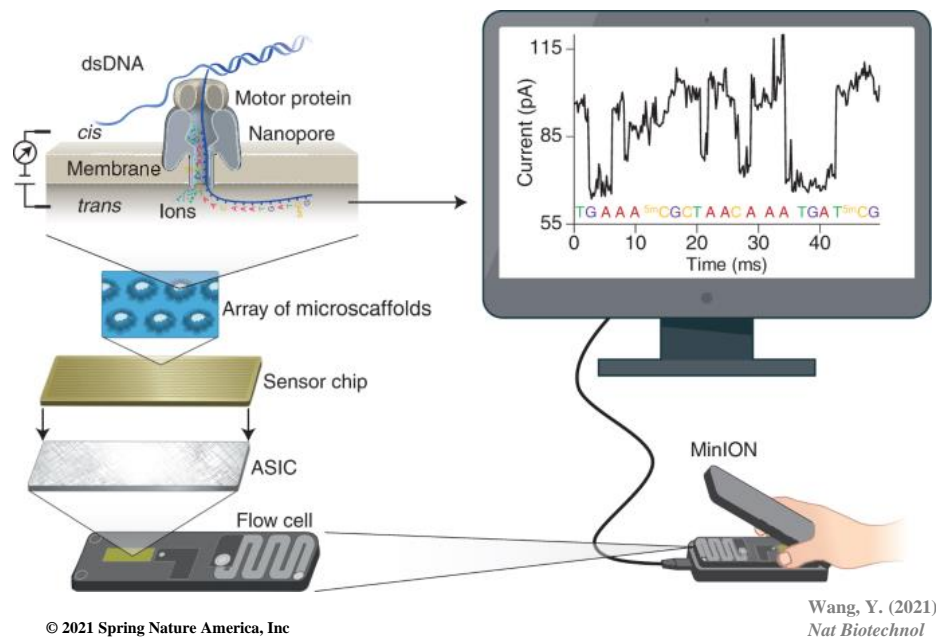


Figure 15. Nanopore sequencing

Note. Reprinted from Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature biotechnology*, 39(11), 1348-1365. <https://doi.org/10.1038/s41587-021-01108-x>
Copyright 2021, Spring Nature America, Inc

the lower cost associated, and the portability of nanopore sequencers offer cost-saving sequencing alternatives to laboratories and settings with limited budgets and space capacity. A major limitation of ONT sequencers are higher error rates (~6%) relative to short read technologies and PacBio CCS, which hinders its application in variant calling (Delahaye & Nicolas, 2021). Although newer chemistries ($\geq R10.4$) show improved accuracies (99%) (Sereika et al., 2022), a verification of these read accuracies in homopolymeric genomes is yet to be performed, as these regions have proven to be the main source of errors for ONT sequencers (Delahaye & Nicolas, 2021), and the previous assessment was performed using bacterial genomes (Sereika et al., 2022), which rarely have long homopolymer stretches (Delahaye & Nicolas, 2021; Sereika et al., 2022). Furthermore, the ONT nanopore and

motor protein are in a continual state of refinement, which may limit opportunities for comparisons between studies, as new chemistry updates are sometimes released within a few months of previous ones. To date, 10 chemistries have been released. These include R6 (June 2014), R7 (July 2014), R7.3 (October 2014), R9 (May 2016), R9.4 (October 2016), R9.5 (May 2017), R10 (March 2019), R10.3 (January 2020), R10.4 (December 2021), R10.4.1 (March 2022) (Wang et al., 2021).

Recommended Reading

For more information on the different sequencing methods, check out the following links:

- Sanger sequencing: <https://www.sigmaaldrich.com/CH/de/technical-documents/protocol/genomics/sequencing/sanger-sequencing>
- Illumina Sequencing <https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html>
- ONT Nanopore Sequencing <https://nanoporetech.com/applications/dna-nanopore-sequencing> ; <https://www.youtube.com/watch?v=RcP85JHLmnl> ; <https://www.youtube.com/watch?v=E9-Rm5AoZGw>
- PacBio SMRT Sequencing <https://www.pacb.com/technology/hifi-sequencing/>; <https://www.pacb.com/technology/hifi-sequencing/how-it-works/>; <https://www.youtube.com/watch?v=NHCJ8PtYCFc>
- Ion Torrent sequencing <https://www.youtube.com/watch?v=zBPKj0mMcDg>
- MGI Tech DNA Nanoball sequencing <https://www.youtube.com/watch?v=CAZwdtORXMw>

I.4. Sanger Sequencing vs. NGS

The benefits and limitations of both Sanger sequencing and NGS have been summarized in **Table I**.

Choosing the right method for your project will depend on the desired scale of sequencing as well as the required downstream application.

Table I. Comparison of Sanger sequencing and NGS

Pros and Cons	Sanger Sequencing	NGS
Pros	○ Fast and cost-effective for low number of targets (1-20 targets)	○ Higher sequencing depth allows higher sensitivity (down to 1%)
	○ Established workflow	○ Higher discovery power for novel variants
	○ Simplified data analysis	○ Higher mutation resolution

Pros and Cons	Sanger Sequencing	NGS
		<ul style="list-style-type: none"> Higher scalability (More data output with the same amount of input DNA)
Cons	<ul style="list-style-type: none"> Low sensitivity (~15-20% detection limit) 	<ul style="list-style-type: none"> Less cost-effective for low number of targets (1-20 targets)
	<ul style="list-style-type: none"> Less cost-effective for high number of targets (>20 targets) 	<ul style="list-style-type: none"> Time consuming for low number of targets (1-20 targets)
	<ul style="list-style-type: none"> Low discovery power for novel variants 	<ul style="list-style-type: none"> More complex workflows
	<ul style="list-style-type: none"> Low scalability (if more data output is desired, sample input must be increased) 	<ul style="list-style-type: none"> More complex data analysis

Note. Adapted from Illumina. <https://emea.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html> Copyright 2022, Illumina.

1.5. Applications of NGS

The power of NGS technologies has led to their application across a wide range of fields including biology, agriculture, clinical diagnostics, forensics, drug discovery, etc. A detailed summary of these applications is shown in **Table 2**.

Table 2. Applications of NGS

Category	Examples
Biology	
Whole genome sequencing	Study the genomes of pathogenic microorganisms
Targeted gene sequencing (amplicon-based)	Mutation discovery
Targeted gene expression profiling	Quantify mRNAs for gene, examine the pattern of gene expression in cells
Methylation Sequencing	Analyse genome-wide DNA methylation
Metagenomic Sequencing	Study the microbiome (microbial diversity) in environmental and biological samples
miRNA and small RNA analysis	microRNA profiling
DNA-Protein Interaction Analysis	Genome-wide mapping of protein-DNA interactions
Transcriptome sequencing	Discover novel RNA variants and splice sites
Infectious Diseases	
Public health surveillance	Strain characterization, rapid pathogen detection, transmission mapping
Antimicrobial resistance profiling	Identify known resistance-associated mutations for treatment management and new drug resistance genes
Agriculture (Plant Biology)	
Targeted gene expression profiling	Identify novel genes associated with traits or disease
Trait screening	Identify desired traits for selective breeding

Clinical diagnostics	
Genetic testing/screening	Cystic fibrosis, non-invasive prenatal testing (NIPT), human leukocyte antigen (HLA) typing
Cell-Free Sequencing & Liquid Biopsy Analysis	Mutation detection at high resolution
Forensics	
Mitochondrial DNA (mtDNA)	Identify forensic samples
Maternal blood typing	Establish family relationships
Drug discovery	
Whole genome sequencing	Identify novel drug targets
Pharmacogenomics	Identify gene variants associated with drug response

1.6. References

- Adams, J. (2008). DNA sequencing technologies. *Nature Education*, 1(1), 193.
- Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, 21(1), 30. <https://doi.org/10.1186/s13059-020-1935-5>
- Ardui, S., Ameer, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*, 46(5), 2159-2168. <https://doi.org/10.1093/nar/gky066>
- Chial, H. (2008). DNA sequencing technologies key to the Human Genome Project. *Nature Education*, 1(1), 219.
- Colleen T. Harrington, BA; Elaine I. Lin, BA; Matthew T. Olson, MD; James R. Eshleman, MD, PhD. *Arch Pathol Lab Med* (2013) 137 (9): 1296–1303. <https://doi.org/10.5858/arpa.2012-0463-RA>
- Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol Health Res World*, 19(3), 190-195. <https://www.ncbi.nlm.nih.gov/pubmed/31798046>
- Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617), 286-290. <https://doi.org/10.1126/science.1084564>
- Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS One*, 16(10), e0257521. <https://doi.org/10.1371/journal.pone.0257521>
- Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borchering, A. P., Brownley, A., Cedeno, R., Chen, L., . . . Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961), 78-81. <https://doi.org/10.1126/science.1181498>
- Gates, A. J., Gysi, D. M., Kellis, M., & Barabasi, A. L. (2021). A wealth of discovery built on the Human Genome Project - by the numbers. *Nature*, 590(7845), 212-215. <https://doi.org/10.1038/d41586-021-00314-6>
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6), 333-351. <https://doi.org/10.1038/nrg.2016.49>
- Hayden, E. C. (2014). Technology: The \$1,000 genome. *Nature*, 507(7492), 294-295. <https://doi.org/10.1038/507294a>
- Hood, L., & Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. *Genome Med*, 5(9), 79. <https://doi.org/10.1186/gm483>
- Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum Immunol*, 82(11), 801-811. <https://doi.org/10.1016/j.humimm.2021.02.012>

14. International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931-945. <https://doi.org/10.1038/nature03001>
15. Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., . . . Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*, 36(4), 338-345. <https://doi.org/10.1038/nbt.4060>
16. Korostin, D., Kulemin, N., Naumov, V., Belova, V., Kwon, D., & Gorbachev, A. (2020). Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One*, 15(3), e0230301. <https://doi.org/10.1371/journal.pone.0230301>
17. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921. <https://doi.org/10.1038/35057062>
18. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012, 251364. <https://doi.org/10.1155/2012/251364>
19. Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Front Genet*, 10, 426. <https://doi.org/10.3389/fgene.2019.00426>
20. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376-380. <https://doi.org/10.1038/nature03959>
21. Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74(2), 560-564. <https://doi.org/10.1073/pnas.74.2.560>
22. NHGRI. (2022). *Chromosome*. Retrieved 31/07/2022 from <https://www.genome.gov/genetics-glossary/Chromosome#:~:text=Chromosomes%20are%20threadlike%20structures%20made,in%20the%20nucleus%20of%20cells>.
23. NIH. (2020). *The Human Genome Project*. <https://www.genome.gov/human-genome-project>
24. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., . . . Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44-53. <https://doi.org/10.1126/science.abj6987>
25. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J Appl Genet*, 52(4), 413-435. <https://doi.org/10.1007/s13353-011-0057-x>
26. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., & Nyren, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242(1), 84-89. <https://doi.org/10.1006/abio.1996.0432>
27. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12), 5463-5467. <https://doi.org/10.1073/pnas.74.12.5463>
28. Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sorensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*, 19(7), 823-826. <https://doi.org/10.1038/s41592-022-01539-7>

29. Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol*, 122(1), e59. <https://doi.org/10.1002/cpmb.59>
30. Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform*, 3(1), lqab019. <https://doi.org/10.1093/nargab/lqab019>
31. U.S. Department of Energy. (2003). <http://www.ornl.gov/hgmis>
32. Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*, 39(11), 1348-1365. <https://doi.org/10.1038/s41587-021-01108-x>
33. Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Functammasan, A., Kolesnikov, A., Olson, N. D., Topfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, 37(10), 1155-1162. <https://doi.org/10.1038/s41587-019-0217-9>

CHAPTER 2
COMPARISON OF NGS PLATFORMS

CHAPTER 2: COMPARISON OF NGS PLATFORMS

2.1. Common parameters across NGS platforms

Apart from the sequencing technology, NGS platforms can be compared using different parameters.

Read length: Most platforms either support a paired-end (PE) sequencing mode, where both ends of the fragment are sequenced or a single-end (SE) sequencing mode, where only one end is sequenced (**Figure 16**). Read lengths from PE sequencing platforms are usually denoted as “2 x read length”. For example, 2 x 150 bp indicates that both forward and reverse read outputs are ~150 bp in length. On the other hand, the read length from a SE sequencing platform is expressed as “1 x read length” e.g., 1 x 50 bp or simply as the length of the read e.g., 50 bp.

Read output: NGS platforms generate massive amounts of data, which can be up to billions of reads and this data output is measured in bytes. For example, Illumina platforms that can produce ~4 million reads generally require ~1.2 Gigabytes (Gb) of data storage space (Illumina, 2022).



Figure 16. Paired-end vs single-end sequencing

Accuracy: The accuracy of a NGS platform generally refers to the degree of correctness of its base calling (i.e., assignment of bases / nucleotide order of the template during sequencing). The base calling accuracy is measured by the Phred quality score, a.k.a. Q-score, which indicates the probability that a given base is called incorrectly by the sequencer (Illumina, 2011). For example, a Q score of 10 (Q10) indicates a 1 in 10 times probability of an incorrect base call, which corresponds to a base call accuracy of 90%. High quality NGS runs generally have Q scores of 30 indicates a 1 in 1000 times probability of an incorrect base call and above ($\geq Q30$), i.e., a base call accuracy of $\geq 99.9\%$ (Illumina, 2011). However, it is important to note that errors are a normal occurrence during sequencing and can arise during clonal amplification, sequencing cycles and imaging, resulting in ~0.1-1% of all bases being called incorrectly (Fox et al., 2014). Although Q scores are the quality metric used by short read platforms, base calling on long read platforms is performed either through internally developed basecallers

specific to the sequencing chemistry (PacBio) or neural network based basecallers that are trained and validated using machine learning and are available as production software or open source (ONT).

2.2. Short Read Sequencing Platforms

Table 3. Comparison of short read sequencers

	Illumina ^a						Thermofisher Scientific ^b			Qiagen ^c	MGI ^d			
System Platform(s)	iSeq	MiniSeq	MiSeq	NextSeq 550	NextSeq 1000&2000	NovaSeq 6000	GeneStudio S5	Genexus (automated)	Ion PGM Dx	GeneReader [§]	DNBSEQ-T7	DNBSEQ-G400	DNBSEQ-G400 ³	DNBSEQ-G50
Sequencing principle	Sequencing by synthesis									Sequencing by ligation				
Sequencing detection	Fluorescent						Ion			Fluorescent	Fluorescent			
Read length [†]	2 x 75 bp 2 x 150 bp 36 bp 50 bp 75 bp	2 x 75 2x150bp 75 bp 100 bp	2 x 25 bp 2 x 75 bp 2 x 150 bp 2 x 250 bp 2 x 300 bp 36 bp	2 x 75 bp 2 x 150 bp 75 bp	2 x 50 bp 2 x 100 bp 2 x 150 bp 50 bp	2 x 50 bp 2 x 100 bp 2 x 150 bp 2 x 250 bp 35 bp	200 bp 400 bp 600 bp	400 bp	200 bp	100-150 bp ¹	2 x 100 bp 2 x 150 bp 50 bp	2 x 100 bp 2 x 150 bp 2 x 200 bp 50 bp 100 bp 400 bp	2 x 100 bp 50 bp 100 bp	2 x 100 bp 2 x 150 bp 50 bp 100 bp
Output per run	0.1-1.2 Gb	1.7-7.5 Gb	0.3-15 Gb	25-120 Gb	30-360 Gb	65-3000 Gb	0.3-50 Gb	24 Gb	0.6-1 Gb	N/A- 16 mil reads/flowcell ¹	6 Tb	55-1440 Gb	75-720 Gb	10-150 Gb
Sequencing run time	9.5-19 hrs	5-24 hrs	4-56 hrs	11-29 hrs	11-48 hrs	13-44 hrs	2.5-4 hrs ⁴	14-31 hrs ⁴	4.4 hrs ⁴	Up to 45 hrs ²	20-30 hrs	14-109 hrs	9.5-50.5 hrs	9-40 hrs
Accuracy/Quality score* per base	Q30 ≥80%		Q30≥70%	Q30 ≥75%			Q20 ≥99%			Q25 >85% ¹	Q30 >80%	Q30 >70%	Q30 >75%	Q30 >80%
Advantages	<ul style="list-style-type: none"> - Lowest error rates of NGS technologies - Wide application - High throughput (up to billions of reads) 						<ul style="list-style-type: none"> - Short run times - Low cost - No fluorescent labelling or optics 			Integrated automated workflows	<ul style="list-style-type: none"> - Less expensive alternative to Illumina - Wide application - Supports both paired end and single end reads 			
Disadvantages	<ul style="list-style-type: none"> - PCR amplification bias - Long run time - Short read lengths are unable to resolve complex genomic regions - High instrument costs 						<ul style="list-style-type: none"> - Higher errors in homopolymer regions - Lower throughput relative to other short read technologies 			Limited application: only specific clinical targets	<ul style="list-style-type: none"> - Short read lengths are unable to resolve genomic complexities (or "complex genomic regions" for consistency) - Longer workflows - Run times can be long 			

[†]Paired-end (PE) reads are represented as 2 x read length, while single-end (SE) reads are represented as a unit read length (e.g., 50 bp).
^{*}Quality score predicts the probability of an erroneous base call. A quality score of 30 (Q30) indicates an error rate of 1 in 1000, which corresponds to a call accuracy of 99.9%.
[§] Qiagen has discontinued GeneReader development, but existing instrument owners will continue to be supported for an indefinite period.
¹ Information sourced from an online powerpoint presentation (http://media.aiom.it/userfiles/files/doc/AIOM-Servizi/slide/20170512NA_42_Cogne_Cerfeda_Qiagen.pdf) and may not be accurate
² Information sourced from the online Instrument user manual
³ Sequencer version only available in certain countries and the software is configured for a specific reagent (Hot MPS)
⁴ Run time for Genexus represents total turnaround time, i.e., the time from start to finish of the entire workflow from sample and library preparation to sequencing, analysis and reporting whereas run times for GeneStudio and Ion PGM Dx only represents time from start to finish of sequencing run. Turnaround time (sequencing run + analysis time) for GeneStudio systems ranges between 3-21.5 hrs.

References:
^aIllumina. (2022). <https://emea.illumina.com/systems/sequencing-platforms.html> / <https://emea.illumina.com/systems/sequencing-platforms/nextseq-1000-2000/specifications.html> / <https://emea.illumina.com/systems/sequencing-platforms/nextseq/specifications.html> / <https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html> / <https://emea.illumina.com/systems/sequencing-platforms/miniseq/specifications.html> / <https://emea.illumina.com/systems/sequencing-platforms/iseq/specifications.html>
^bThermofisher Scientific. <https://www.thermofisher.com/ch/en/home/brands/ion-torrent.html> / <https://www.thermofisher.com/ch/en/home/clinical/diagnostic-testing/instruments-automation/genetic-analysis-instruments/ion-pgm-dx.html>
^cQiagen. (2022). GeneReader NGS System. <https://www.qiagen.com/us/products/instruments-and-automation/genereader-system/qiagen-genereader-platform/>
^dMGI Tech. (2022). Sequencers. <https://en.mgi-tech.com/products/>

Note. Adapted from Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum Immunol*, 82(11), 801-811. <https://doi.org/10.1016/j.humimm.2021.02.012>
 Copyright 2021, American Society for Histocompatibility and Immunogenetics

2.3. Long Read Sequencing Platforms

Table 4. Comparison of long read sequencing platforms

System Platform(s)	PacBio SMRT ¹			ONT Nanopore ²			
	Sequel	Sequel II	Sequel IIe	Flongle	MinION	GridION	PromethION
Sequencing principle	Single molecule sequencing						
Detection	Fluorescent			Electrical conductivity			
Read length (max)	300 kb			Longest read so far > 4 Mb			
Output per run (max)	75 Gb	600 Gb		2.8 Gb	50 Gb	250 Gb	580 Gb – 14 Tb
Run time (min-max)	Up to 20 hrs	Up to 30 hrs		2 mins - 16 hrs	2 mins - 72 hrs		
Accuracy	CLR: 13% error rate ³ HIFI reads: 99.9% accuracy			Flow cells before R9.4.1: ~12% error rate ³ , Flow cells R9.4.1: ~6% error rate ⁴ ; Flow cells ≥R10.4: 99% accuracy ⁵			
Advantages	<ul style="list-style-type: none"> - Long reads that resolve genomic ambiguities - Fast turnaround time - Stochastic (random) errors in raw reads can be resolved with consensus generation 			<ul style="list-style-type: none"> - Low cost - Portable - Rapid sequencing, - Long and ultra-long reads that span complex, ambiguous genomic regions - Higher throughput 			
Disadvantages	<ul style="list-style-type: none"> - Expensive sequencing equipment - Non-portable (large size) - Difficult installation - Lower throughput relative to ONT 			<ul style="list-style-type: none"> - Higher error rates in raw reads, especially in homopolymers and regions with high GC content⁴ - Systematic errors in raw reads cannot be resolved with high coverage 			
References:							
¹ PacBio. (2022). PacBio Sequel Systems. https://www.pacb.com/technology/hifi-sequencing/sequel-system/							
² Oxford Nanopore Technologies. (2022). Product specifications. https://nanoporetech.com/products/specifications							
³ Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. <i>Nat Rev Genet</i> , 17(6), 333-351. https://doi.org/10.1038/nrg.2016.49							
⁴ Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. <i>PLoS One</i> , 16(10), e0257521. https://doi.org/10.1371/journal.pone.0257521							
⁵ Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sorensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. <i>Nat Methods</i> , 19(7), 823-826. https://doi.org/10.1038/s41592-022-01539-7							

Note. Adapted from Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum Immunol*, 82(11), 801-811. <https://doi.org/10.1016/j.humimm.2021.02.012> Copyright 2021, American Society for Histocompatibility and Immunogenetics

2.4. Short Read vs. Long Read Sequencing Technologies

Table 5. Short read vs. long read sequencing

	Short read sequencing	Long read sequencing
Read lengths	Shorter (restricted lengths): 150-300 bp	Longer (unrestricted lengths): > 10 kb (average)
Sample preparation	Longer and more complex workflows	Simplified and quicker workflows
Sequencing run time	Fixed run times	Rapid turnover with real-time data acquisition
Raw read accuracy	Lower per base error rate	Higher per base error rate (minimized with consensus base-calling)
Throughput	Millions to billions	Hundreds of thousands to millions
Genomic characterizations	<ul style="list-style-type: none"> - Unable to span complex and repetitive genomic regions - Unable to detect base modifications 	<ul style="list-style-type: none"> - Resolves complex and repetitive regions - Able to detect base modifications
Estimated investment (platform costs, consumables, cost per Gb)	Can be cost-intensive	Can be cost saving*
* This is based on the unpublished excel documentation of sequencing costs by Dr. Albert J. Vilella (https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8Xklo3YxIWaZA5vVMuhU1kg41g4xLkXc/htmlview?hl=en_GB)		

2.5. References

1. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl*, 1. <https://doi.org/10.4172/jngsa.1000106>
2. Illumina. (2011). *Quality Scores for Next-Generation Sequencing* https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
3. Illumina. (2022). *Illumina sequencing platforms*. <https://emea.illumina.com/systems/sequencing-platforms.html>

CHAPTER 3
NGS Workflow

CHAPTER 3: NGS Workflow

3.1. Overview

Regardless of the sequencing technology and downstream application, a typical NGS workflow consists of four major steps (**Figure 17**):

1. **Sample preparation:** The first step of an NGS workflow begins with the extraction of nucleic acids (i.e., DNA or ribonucleic acid [RNA]) from biological samples (e.g., blood, saliva, cell cultures, tissues, bone marrow) using an appropriate isolation method or extraction kit that guarantees the best yield, purity and quality needed for subsequent workflow steps. Details of the sample extraction process can be referred from chapter 5 of a comprehensive manual on SARS-CoV-2 diagnostics (https://www.finddx.org/wp-content/uploads/2022/08/20220819_usaid_covid_toolkit_FV_EN.pdf). The yield, purity and quality are critical measures for the success of NGS and can be assessed using ultraviolet (UV) spectrophotometric assays, fluorometric assays, gel-based or microfluidic electrophoresis (Refer **Annexure 3**).

2. **Library preparation:** Here, the isolated nucleic acid is prepared in a format that allows it to be recognized and processed by the sequencer. This generally involves fragmenting the nucleic acid and attaching the resulting fragments to sequences that are compatible with those of the sequencer (**Figure 17**). Depending on the

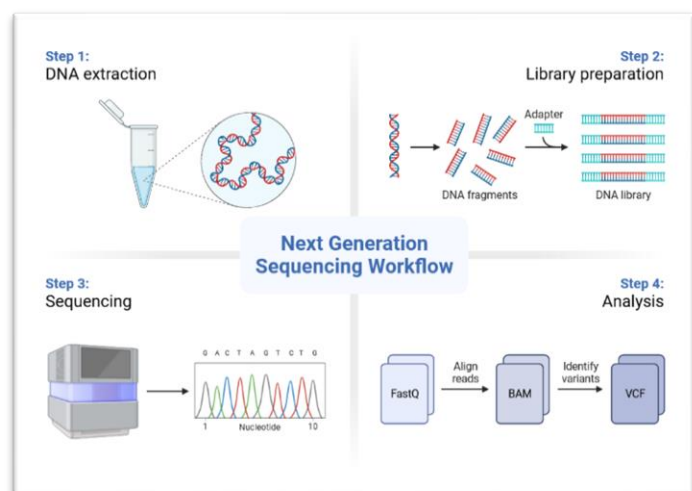


Figure 17. Next generation sequencing workflow

Note. Reprinted from "Next Generation Sequencing Workflow", by BioRender, April 2022. Retrieved from <https://app.biorender.com/biorender-templates> Copyright 2022 by BioRender.

NGS application, additional steps such as a library amplification by PCR or a target enrichment (for specific genes or genomic regions) may be performed. The final prepared library is then

quantified to ensure high quality and uniformity of the nucleic acids before sequencing can begin.

3. **Sequencing:** In this step, the prepared library is loaded onto the sequencer, and reads are produced using one of the many sequencing methods such as sequencing by synthesis, sequencing by ligation or single molecule sequencing.
4. **NGS data analysis:** The final step of the workflow is processing and interpreting the generated sequenced data using different bioinformatic tools and computational methods. In the processing stage, the sequenced data is cleaned by removing low quality reads and trimming off adapter regions. Following this, reads from the different samples that were sequenced together in the same run (multiplexed) are sorted into separate sample files (demultiplexed). In the analysis stage, the cleaned sequence data are examined by mapping to a reference sequence, generating assemblies, annotating genes and identifying variants or new transcripts. Finally, in the interpretation stage, the biological relevance of the sequence is determined e.g., the implications of newly discovered genes, transcripts or genetic variants (Thermo Fisher Scientific, n.d.-b).

3.2. Library Construction: General steps in the workflow

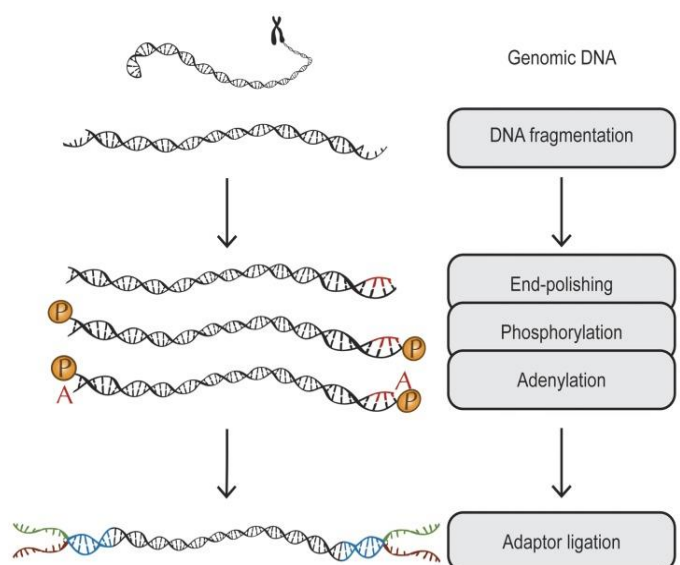
In the second step of the NGS workflow, the template input (DNA or RNA) is prepared into a collection of sequencing ready fragments known as “libraries”. Library preparation involves the core steps of fragmentation and end-repair, adapter ligation and library quantification. NGS libraries can be prepared either through a ligation-based method, where fragmentation, end-repair and adapter ligation occur in separate steps or a tagmentation-based method using bead-linked transposomes, which combines all three steps into one. A summary of the principal and optional steps involved in the construction of NGS libraries is provided below:

1. **Fragmentation:** Here, the long strands of the template DNA (or RNA) are sheared or fragmented into smaller pieces that are optimal to the size range of the NGS platform to be

used (Hu et al., 2021). Fragmentation may be performed via physical/mechanical, enzymatic, or chemical approaches (Head et al., 2014).

- i. **Physical/mechanical:** This is achieved via acoustic devices and sonicators, which employ low to high frequency wavelength energy to shear DNA.
 - ii. **Enzymatic:** Transposases and restriction endonucleases fragment DNA enzymatically by creating nicks in the DNA strands, which breaks them into fragments.
 - iii. **Chemical:** In this method, DNA is heat digested using a divalent metal-cation such as magnesium or zinc (Head et al., 2014).
2. **End-repair:** After fragmentation, the ends of the fragmented DNA pieces are repaired or “polished” by blunt ending to remove protruding ends (overhangs) and the 5’ ends of the blunted fragments are phosphorylated to enable efficient ligation in subsequent steps.

A-tailing (optional): For some libraries e.g., Illumina, single adenine bases are also added to the 3’ ends of the blunted DNA fragments via an A-tailing reaction to prepare the libraries for ligation to adapters with complementary thymine overhangs (**Figure 18**).



Neiman, M.
(2012) *PLoS One*

Figure 18. NGS library preparation

Note. Reprinted from Neiman, M., Sundling, S., Grönberg, H., Hall, P., Czene, K., Lindberg, J., & Klevebring, D. (2012). Library preparation and multiplex capture for massive parallel sequencing applications made efficient and easy. *PLoS One*, 7(11), e48616. <https://doi.org/10.1371/journal.pone.0048616> License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

3. **Adapter ligation:** In this step, short oligonucleotides (~20-40 bp in length) of known sequence, called adapters are covalently annealed to the blunted ends of the DNA fragment (**Figure 18**), allowing them to attach to the flow cell and be recognized by the

sequencing instrument. Adapters also include barcodes or indexes, which allow the identification of individual samples in a pool of samples simultaneously sequenced in a single run (multiplexed sequencing). In some protocols, the DNA fragments flanked by adapters are referred to as “inserts”.

Size selection and purification (optional): Depending on the library protocol, size selection may be performed after fragmentation, adapter ligation or PCR amplification (Thermo Fisher Scientific, n.d.-c). Here, the desired library fragment sizes are selected for sequencing and all other unwanted components (e.g., excess (unbound) adapters and primers, adapter dimers and primer dimers) are removed. The size selection process is essentially a purification step, as it separates the template fragments from all other unwanted components, whose removal is necessary before sequencing, due to their tendency to compete with the library fragments in flow cell binding, resulting in lowered data output and increased sequence noise (Thermo Fisher Scientific, n.d.-c).

Sequencing libraries generally consist of fragments of various sizes. However, uniformity of library fragment sizes is essential for maximizing data output, as longer fragments can result in unsequenced insert sections while shorter fragments can cause the sequencing run to abort, as it runs out of template for base incorporation (Illumina, 2020). Size selection can be performed using a bead-based or gel-electrophoresis method. In the bead-based method, magnetic beads are used to isolate DNA fragments having the desired size whereas with electrophoretic-based methods, the adapter-ligated DNA fragments are run on a gel, which separates them by size. Choosing the best method for your project largely depends on the sample amount, sample throughput and size range of the libraries (Thermo Fisher Scientific, n.d.-b). The bead-based method is more suitable where sample amounts are low and/or sample throughput is high, as high recovery of DNA is ensured and process automation is possible. The gel electrophoresis method works best where large amounts of sample are available and separation of fragments with a narrow size range is required.

Library amplification (optional): Depending on the downstream application and the type of sample input, NGS libraries may require amplification by PCR. Amplification is essential when working with low quantities of sample input, as it increases the library amount, which ensures sufficient coverage during sequencing. PCR amplification also enriches for fragments with adapters ligated on both ends, thereby ensuring that the PCR primers hybridize to the sequences connected to the adapters. However, PCR-based libraries are prone to bias, errors and chimeras, which affect sequence analysis. Therefore, care must be taken to minimise the introduction of such PCR artefacts in the library.

Note. Library amplification is **not** the same as clonal amplification. The former concerns increasing library input quantity before loading on a flow cell whereas the latter deals with amplifying fragment libraries after loading on a flow cell, so that fluorescent signals are strong enough to be detected by the sequencer (Thermo Fisher Scientific, n.d.-b).

4. **Library quantification:** In the final step of the library construction, libraries are quantified and normalized as a quality control (QC) measure to ensure appropriate amounts and equalized concentrations of the fragments are loaded on the flow cell for successful sequencing. Normalization is a critical step before sequencing as uneven library concentrations can either overcluster or undercluster flow cells, resulting in suboptimal data quality. Some commonly used methods for library quantification include microfluidic capillary

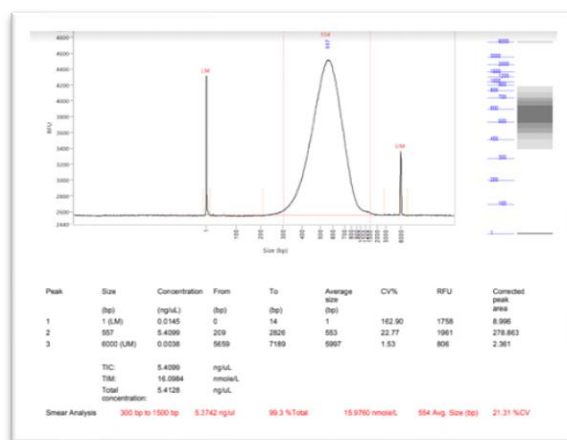


Figure 19. Fragment analysis of libraries by capillary-based electrophoresis

electrophoresis (e.g., Agilent Bioanalyzer), fluorometric assays (e.g., Qubit, PicoGreen) and quantitative PCR (qPCR). The microfluidic capillary electrophoresis measures both the concentration and fragment size of a sequencing library (Figure 19), however, it is typically used

in combination with one of the other two methods, which provide more accurate quantification of the library concentration.

3.3. Library Construction for Different NGS Assays

The library preparation process can differ depending on the NGS assay or application. NGS assays usually target DNA or RNA and can be broadly categorized under DNA sequencing and RNA sequencing. A list of different NGS assays and their specific library preparation requirements is provided below.

DNA sequencing (DNA-seq)

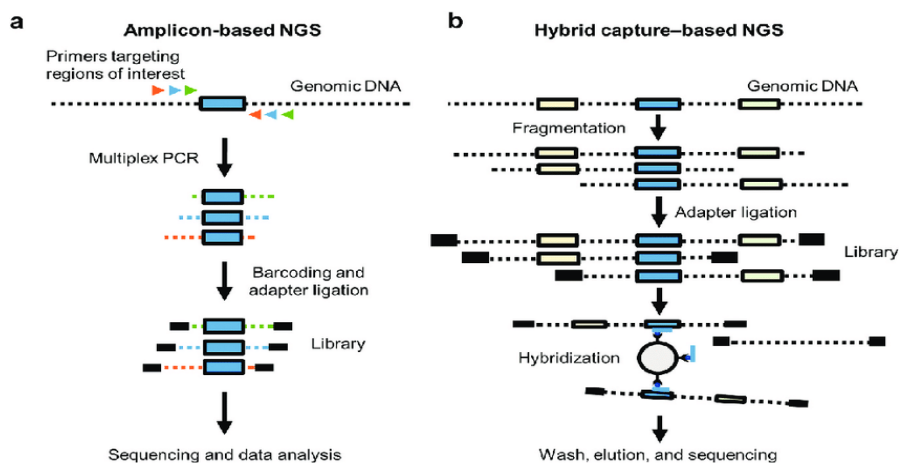
- **Whole genome sequencing (WGS):** This is a sequencing method used to determine the entire (or nearly entire) DNA sequence of an organism, be it human, plant, animal, or microbe.
- **Whole exome sequencing (WES):** In this approach, only the protein-coding regions of the genome (exome) are sequenced. As this region only comprises 2-3% of the human genome, exomes can be sequenced at a greater [depth](#) (150-200x) for lower cost relative to WGS (Bewicke-Copley et al., 2019). As WES targets protein-coding regions, target enrichment is required with this assay.
- **Targeted sequencing (TS):** This type of sequencing focuses on specific genes or coding regions known or suspected to contribute to the pathogenesis of a disease. Targeted gene panels can be custom designed or purchased with predesigned gene content and are widely used for instance in cancer studies. As TS only focuses on specific genomic areas, it requires less sample input and sequences at a significantly deeper coverage depth (200-1000x+), allowing the identification of low-frequency variants at a lower cost than WGS and WES (Bewicke-Copley et al., 2019). In order to efficiently identify regions of interest, target enrichment is utilized during library preparation for targeted sequencing.

Library construction for DNA sequencing

Regardless of the sequencing technology, the library preparation for DNA sequencing assays typically follows a ligation-based approach and can be PCR-free using high molecular weight DNA (500-1000 ng) as starting material or PCR-based with lower input quantities (100-500 ng). For much lower input quantities (1-10 ng), there are specific amplification protocols, which can be combined with a fragmentation-based library construction workflow.

Target enrichment

During TS, samples are enriched for sequences of interest either via an amplicon-based or a hybridization capture-based enrichment process. Amplicon-based enrichment is basically a PCR amplification process, which amplifies regions of interest (up to hundreds of genes) using specially designed primers in a single run prior to library preparation whereas the hybridization approach uses target-specific biotinylated probes or “baits” to capture genomic regions of interest and is usually performed after library preparation (**Figure 20**). The amplicon-based approach is quick, easy and requires less starting material (10-100 ng), though coverage and data quality are affected by primer design, PCR efficiency and amplification bias (Thermo Fisher Scientific, n.d.-c). On the other hand, the hybridization approach generates higher quality data and more uniform coverage, albeit at the expense



Subramanian, J. (2021)
Expert Review of Anticancer

Figure 20. Workflows for target enrichment methods

Note. Reprinted from Subramanian, J., & Tawfik, O. (2021). Detection of MET exon 14 skipping mutations in non-small cell lung cancer: overview and community perspective. *Expert Review of Anticancer Therapy*, 21(8), 877-886.
<https://doi.org/10.1080/14737140.2021.1924683> License: [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/)

of a longer workflow and higher input amounts (>500 ng). The hybridization approach is therefore more suitable for large scale experiments that target a higher number of genes (>50 genes) and require a more sensitive, comprehensive profile of variants such as WES, while the amplicon approach is more appropriate for smaller sets of target regions (<50 genes), which simply require identification of single nucleotide polymorphisms (SNPs) or single nucleotide variants (SNVs) and insertions/deletions (INDELs) (Illumina, n.d.).

Other probes used in hybridization approach: Linked capture probes (Pel et al., 2018) and molecular inversion probes (Niedzicka et al., 2016).

RNA sequencing (RNA-Seq)

- **Whole transcriptome sequencing (WTS)/Total RNA sequencing:** Here, both coding and non-coding RNA are sequenced to provide a “total view” of the [transcriptome](#).
- **Messenger RNA (mRNA) sequencing:** In this sequencing approach, the coding region of the transcriptome is selected (enriched) and sequenced. This method is typically applied in diseases studies to detect [allele](#)-specific expression.
- **Small RNA sequencing:** Here, small non-coding RNAs such as microRNAs (miRNAs) are isolated and sequenced to examine their differential expression.

Library construction for RNA-Seq

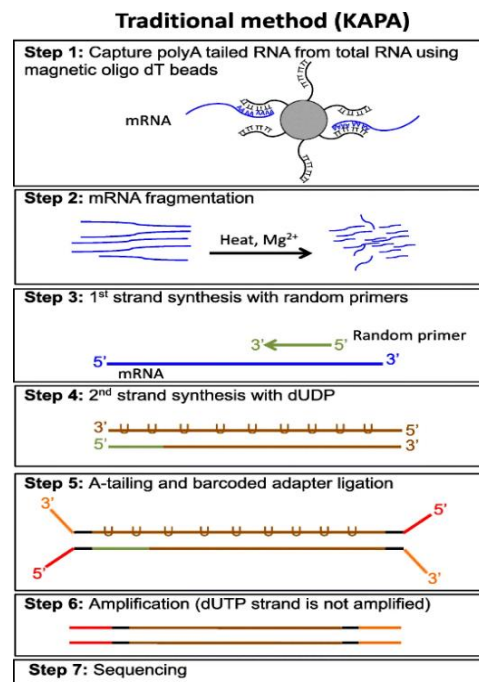
The library construction process for RNA-Seq assays generates complementary DNA (cDNA) via a stranded or non-stranded/standard approach. Strand-specific or stranded RNA libraries preserve the orientation of the transcript by distinguishing the first and second strands of cDNA whereas non-stranded libraries lose this information (Zhao et al., 2015). Depending on the NGS application, a stranded or a non-stranded approach may be required. The protocol for a stranded RNA library is outlined below:

Total RNA isolation: Total RNA is isolated from cells or tissues of interest

1. **mRNA enrichment and rRNA depletion:** For mRNA-Seq assays, the mRNA region is captured from previously purified total RNA using magnetic beads conjugated to oligo(thymine) DNA (oligo(dT)), which bind to the poly(A) tail of the mRNA (**Figure 21**).

For total RNA-Seq assays, ribosomal RNA (rRNA) is depleted to enable the identification of low abundance RNAs.

2. **1st strand synthesis:** As mRNA is single stranded, it is first converted to single stranded (ss)-cDNA molecule via reverse transcription.
3. **2nd strand cDNA synthesis:** The ss-cDNA is then converted to a double stranded (ds)-cDNA and labelled with uracil via a deoxyuridine triphosphate (dUTP) second strand marking method (**Figure 21**).



Ma, F. (2019)
BMC genomics

4. **End repair and A-tailing**
5. **Adapter ligation**
6. **Second-strand digestion:** After library preparation, the second strand with uracils is degraded, ensuring only the first strand is amplified in the next step along with its strand information (**Figure 21**).
7. **PCR amplification and purification**

Figure 21. Library construction of RNA-Seq

Note. Reprinted from Ma, F., Fuqua, B. K., Hasin, Y., Yukhtman, C., Vulpe, C. D., Lusi, A. J., & Pellegrini, M. (2019). A comparison between whole transcript and 3'RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC genomics*, 20(1), 1-12. <https://doi.org/10.1186/s12864-018-5393-3> License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

Other NGS assays

Metagenomics sequencing: This is a method used to sequence all the genomes present in a sample

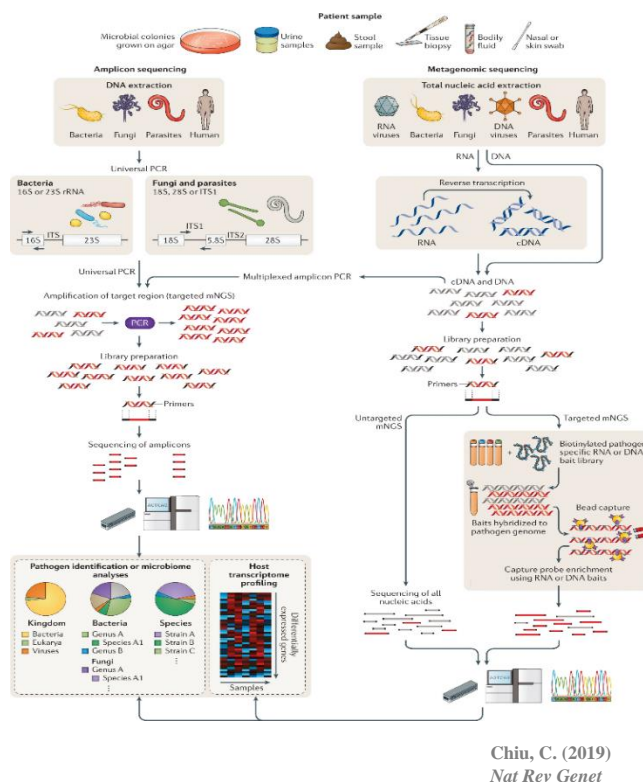


Figure 22. Metagenomics sequencing

Note. Reprinted from Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nature Reviews Genetics*, 20(6), 341-355. <https://doi.org/10.1038/s41576-019-0113-7> Copyright 2019, Springer Nature Limited

in order to gain insights on the diversity and complexity of the sample community. Compared to genomics-based approaches (e.g. WGS), which target DNA from a single organism, metagenomics explores the entire genetic content in a sample, which may originate from multiple organisms and entities, such as bacteria, viruses, viroids, etc. Some applications of metagenomics include estimation of microbial abundance and diversity in various environments, investigation of unculturable microorganisms, functional analysis of genes and gene clusters as well as in clinical investigations of antimicrobial resistance,

human host gene expression and oncology. Metagenomics sequencing can be performed via a targeted or shotgun approach (**Figure 22**). In the targeted approach, conserved genomic regions (e.g., 16S rRNA, 18S rRNA, internal transcribed spacer [ITS]) with sufficient variability for species differentiation are amplified by PCR and sequenced whereas in the shotgun approach, DNA from all cells in the sample is extracted, fragmented, and sequenced.

Library construction for metagenomic sequencing

Metagenomic libraries typically require high-quality DNA extract. For some samples, a high DNA quality is ensured after depletion of the host (human or non-microbe) DNA in an enrichment step either via a selective lysis and degradation method or hybridization-capture method before or after

extraction (Shi et al., 2022) (**Figure 22**). Afterwards the DNA is fragmented, adapters are ligated, and the library is quantified.

Methylation or bisulfite sequencing: Bisulfite sequencing is a technique that utilizes bisulfite treatment of DNA prior to sequencing to distinguish between methylated and unmethylated cytosines, thereby allowing the identification of methylation patterns across entire genomes or genomic regions of interest (Li & Tollefsbol, 2011). Cytosine methylation is an epigenetic mechanism that adds a methyl group (CH₃) to the 5' carbon on the pyrimidine ring of the cytosine nucleotide leading to the formation of a 5'-methylcytosine (5mC). Methylated cytosines are known to play a critical role in gene expression, embryonic development, cellular proliferation, differentiation and chromosome stability (Li & Tollefsbol, 2011). Abnormal or aberrant methylation (e.g., hypermethylation in gene regulatory sites such as promoters) can cause genomic instability resulting in diseases such as cancer.

Library construction for bisulfite sequencing

Bisulfite sequencing may be based on a whole genome or targeted/amplicon method. Whole genome bisulfite sequencing (WGBS) can be performed via: (i) a conventional approach, where bisulfite treatment occurs after a ligation-based library preparation; (ii) an alternative Post-Bisulfite Adaptor Tagging (PBAT) approach (**Figure 23**), where bisulfite conversion precedes adapter ligation, thus circumventing the BS-induced DNA degradation that frequently occurs in the conventional approach (Miura et al., 2012); (iii) a tagmentation-based WGBS (T-WGBS), where the bisulfite treatment occurs after tagmentation of DNA with transposome fragments and following gap repair (Wang et al., 2013).

On the other hand, the library preparation process for targeted bisulfite sequencing or bisulfite amplicon sequencing (BSAS) involves bisulfite conversion of the input DNA, followed by PCR amplification using specially designed primers to enrich for regions of interest for DNA methylation analysis (**Figure 23**). Depending on the library preparation kit, PCR amplicons may additionally be constructed into dual-indexed libraries through a simple tagmentation process and finally sequenced (Masser et al., 2015) (**Figure 23**). Alternatively, targeted bisulfite sequencing may be based on a hybridization step, where bisulfite-converted DNA is hybridized to a pre-designed oligonucleotide

capture array to enrich for target regions (Creative Biomart, n.d.). In this workflow, library preparation occurs before bisulfite treatment and hybridization capture (**Figure 23**).

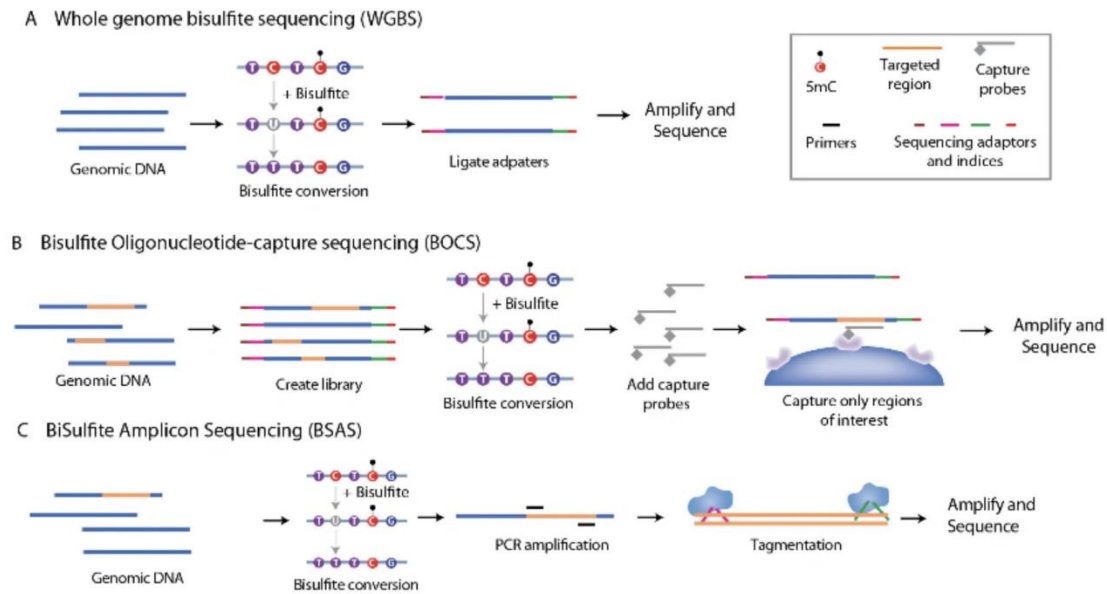


Figure 23. Bisulfite sequencing

Note. Reprinted from Oklahoma Nathan Shock Center on Aging, <https://oklahomanathanshockcenteronaging.org/genomic-sciences-core-services/> Copyright 2019, Oklahoma Nathan Shock Center on Aging

Recommended Reading

For more information on the different NGS assays and their library preparation methods, check out the following links:

- Target enrichment <https://www.youtube.com/watch?v=EKctfcqv93A>
- RNA Sequencing <https://www.youtube.com/watch?v=agipRvXIZv4>
- Stranded vs. non-stranded RNA-Seq <https://web.azenta.com/stranded-vs-non-stranded-rna-seq>; https://www.youtube.com/watch?v=_nKWSXTC9a0
- Metagenomic sequencing https://www.youtube.com/watch?v=RcYXTpNS_XU
- DNA methylation and Bisulfite sequencing <https://www.youtube.com/watch?v=OclazFGQv0g>
- Whole genome bisulfite sequencing <https://www.youtube.com/watch?v=ZYVGSgIf-AI>

3.4. Characteristic Features of NGS Platforms

A summary of the specificities of the two main short read (Illumina and Ion Torrent) and long read (PacBio & ONT) technologies is outlined below.

Illumina: Sequencing on Illumina platforms is based on the SBS approach and occurs via their proprietary reversible terminator technology, which generates millions of highly accurate (<0.1% errors) short reads, making it the technology of choice for applications that require high throughput and sensitive analysis, be it small or large-scale. Among all the available short-read sequencing technologies, Illumina platforms support the widest range of applications from research to clinical diagnostics with two regulated sequencing platforms (MiSeqDx & NextSeq 550Dx) for routine diagnostic use (e.g., in cancer diagnostics, cystic fibrosis testing and in vitro diagnostic [IVD] assay development). Due to its short-read technology, Illumina library construction protocols require a fragmentation step and can be ligation or tagmentation-based. Although some protocols can take up to a whole day to perform, there are a variety of other workflow options, which require only a few hours. In addition, Illumina's library prep workflows support a broad range of input (1-500 ng for DNA and 1-1000 ng for RNA) and various sample types (blood, saliva, dried blood spots or bacterial colonies, formalin fixed paraffin-embedded [FFPE] tissues) with the option of multiplexing up to 384 reactions, which makes it flexible for different types of investigations. Furthermore, Illumina's libraries are specially constructed with an adapter complex of sequences that allow the libraries to bind to flow cells (P5 & P7), primer binding sites to initiate sequencing (Rd1 SP & Rd2 SP) and sample identifiers (Index 1 & Index 2) for multiplexed samples (**Figure 24**). Finally, a particular characteristic of Illumina sequencing is the clonal bridge amplification process, which generates clusters for sequencing.

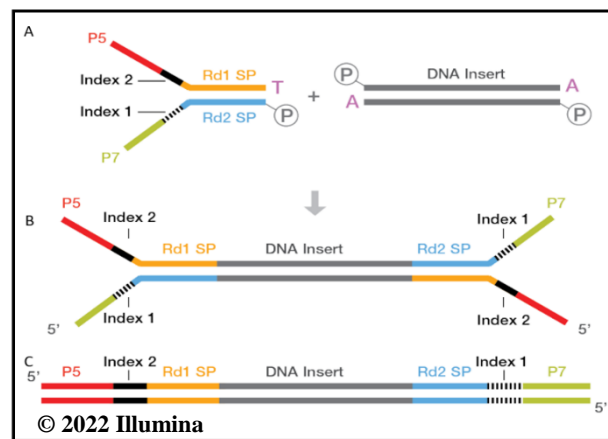


Figure 24. Adapter ligation (Illumina)

Note. Reprinted from Illumina. (2020). How short inserts affect sequencing performance.

<https://emea.support.illumina.com/bulletins/2020/12/how-short-inserts-affect-sequencing-performance.html> Copyright 2022, Illumina.

and various sample types (blood, saliva, dried blood spots or bacterial colonies, formalin fixed paraffin-embedded [FFPE] tissues) with the option of multiplexing up to 384 reactions, which makes it flexible for different types of investigations. Furthermore, Illumina's libraries are specially constructed with an adapter complex of sequences that allow the libraries to bind to flow cells (P5 & P7), primer binding sites to initiate sequencing (Rd1 SP & Rd2 SP) and sample identifiers (Index 1 & Index 2) for multiplexed samples (**Figure 24**). Finally, a particular characteristic of Illumina sequencing is the clonal bridge amplification process, which generates clusters for sequencing.

Ion Torrent: Ion torrent platforms are the only NGS technology that do not utilize scanning, cameras, or light to capture incorporated bases; instead, directly translating bases to digital information on a semiconductor chip via proton detection, making them simpler, faster and more cost effective

than other short read technologies on the market. These platforms produce accurate (<0.1% error) single-end reads that are longer (max 600 bp) than those from other short-read technologies, though at a lower throughput (max 50 Gb) and are therefore suitable for small- to medium-scale applications requiring intermediate read lengths for sensitive analysis at low costs. Ion torrent also has a limited application range and is mainly used for research purposes, with the exception of one platform (Ion PGM Dx), which is used in diagnostic laboratories for IVD and assay development. For applications requiring automation, Ion Torrent Genexus system offers an automated NGS workflow that performs all steps from sample and library preparation to sequencing and analysis with the generation of end data reports. The library construction workflow for Ion torrent is simple, fast and mainly ligation-based, requiring a fragmentation step and end repair without A-tailing. However, depending on the protocol, the amount of input required may be higher (e.g., ≥ 100 ng for RNA) and multiplexing options may be limited (only 48 reactions for RNA-Seq) compared to workflows from other technologies that can facilitate lower input amounts and up to 384-plex reactions (Thermo Fisher Scientific, n.d.-a).

PacBio: PacBio Sequel systems use SMRT technology to sequence single molecules of DNA in real-time, producing long reads with average lengths >10 kb. Characteristic features of these platforms include (i) the use of SMRTcells that contain wells called ZMWs, where the DNA is immobilized; (ii) a SMRTbell library preparation process that creates circularized templates (SMRTbells) by ligating universal hairpin adapters onto dsDNA fragments (**Figure 25**); (iii) a CCS sequencing mode, generates consensus sequences (HIFI reads) with accuracies of >99.9%; and (iv) sequencing runs, which are called “movies”. The PacBio HIFI library construction workflow is straightforward and relatively quick (4.5-6 hrs). SMRTbell libraries generally require a high quantity of input material (300-1000 ng of high molecular weight [HMW] DNA for DNA sequencing and ~300 ng of total RNA for RNA-Seq) and

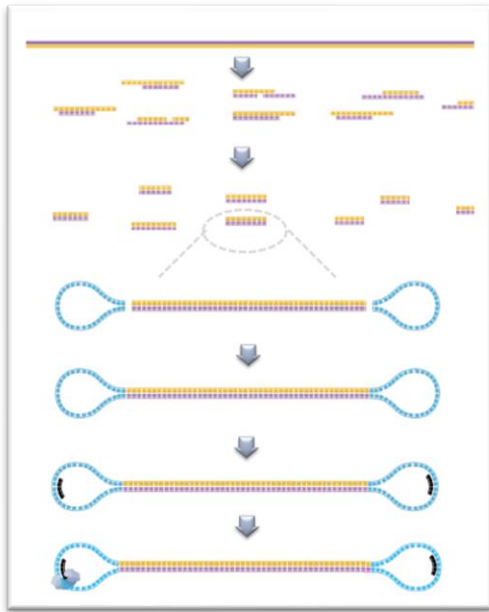


Figure 25. PacBio SMRTbell template preparation

Note. Reprinted from Pacific Biosciences. (2014). Template preparation <https://www.pacb.com/wp-content/uploads/2014/04/TemplatePreparation.pdf> Copyright 2014, Pacific Biosciences of California, Inc.

follow the typical ligation-based library preparation steps with the only difference being the addition of a nuclease treatment step to remove damaged SMRTbell templates (PacBio, 2022). For SMRTbell multiplexed libraries, sample identifiers are known as barcodes (BC) and are attached prior to hairpin adapter ligation. Despite high accuracy reads, PacBio platforms only support research applications and are most relevant where long reads are needed, especially for accurate detection of structural variations, DNA methylations and repetitive regions. Although a high quantity of input is required for the library preparation, a new protocol that supports ultra-low DNA inputs (5-20 ng) is

currently available (PacBio, 2022). PacBio Sequel systems are able to support small to large-scale projects, however due to their large size (>300 kg) and high cost, many investigations requiring SMRT technology are preferably outsourced to commercial companies offering sequencing services at low costs.

ONT Nanopore: ONT platforms utilize nanopore technology to sequence native (single molecule) DNA in real time and can generate both long (10-100 kb) and ultra-long reads (>100 Mb). Characteristic features of these platforms include their portability, low cost, high throughput and use of reusable, detachable flow cells containing ion permeable nanopores adapted from biological pore-forming proteins (α -hemolysin). ONT also provides the unique option of performing sequencing anywhere, i.e., outside of the traditional laboratory environment with two pocket-sized/hand-held, sequencing platforms (MinION) as well as an automated USB-powered device for sample and library preparation (VolTRAX), thus making it suitable for use in resource-limited settings and field-based investigations. Despite supporting a wide range of applications, ONT platforms are limited to the

research field because of their high error rates (~6%) (Delahaye & Nicolas, 2021). However, the newest flow cell chemistries (R10.4.1) report an improvement with higher accuracies of 99% (Sereika et al., 2022). Library preparation for ONT workflow is easy, flexible and quick and can be ligation-based (using a ligation sequencing kit) or tagmentation-based (using a rapid sequencing kit). Libraries are specially prepared by attaching a motor protein sequence adapter complex to strand ends, allowing the motor protein to bind to nanopores in the flow cell and control the movement of DNA/RNA strands through the nanopore. Library protocols generally require ~1000 ng of HMW dsDNA or 200 ng of total RNA with no fragmentation, however, they can be adapted for lower input quantities, with the inclusion of fragmentation and a size selection step to enrich for longer read lengths is necessary prior to adapter ligation. ONT platforms therefore offer a low-cost alternative for research investigations requiring complete sequencing control and accurate long or ultra-long reads from quick and flexible protocols.

3.5. References

1. Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., Korfi, K., & Wang, J. (2019). Applications and analysis of targeted genomic sequencing in cancer studies. *Comput Struct Biotechnol J*, 17, 1348-1359. <https://doi.org/10.1016/j.csbj.2019.10.004>
2. Creative Biomart. (n.d.). *Targeted Bisulfite Sequencing*. Retrieved 25/07/2022 <https://www.creativebiomart.net/epigenetics/services/targeted-bisulfite-sequencing/>
3. Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS One*, 16(10), e0257521. <https://doi.org/10.1371/journal.pone.0257521>
4. Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56(2), 61-64, 66, 68, passim. <https://doi.org/10.2144/000114133>
5. Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum Immunol*, 82(11), 801-811. <https://doi.org/10.1016/j.humimm.2021.02.012>
6. Illumina. (2020). *How inserts affect sequencing performance*. <https://emea.support.illumina.com/bulletins/2020/12/how-short-inserts-affect-sequencing-performance.html>
7. Illumina. (n.d.). *Target Enrichment*. Retrieved 15/07/2022 from <https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/target-enrichment.html>
8. Li, Y., & Tollefsbol, T. O. (2011). DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol*, 791, 11-21. https://doi.org/10.1007/978-1-61779-316-5_2
9. Masser, D. R., Stanford, D. R., & Freeman, W. M. (2015). Targeted DNA methylation analysis by next-generation sequencing. *J Vis Exp*(96). <https://doi.org/10.3791/52488>
10. Miura, F., Enomoto, Y., Dairiki, R., & Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res*, 40(17), e136. <https://doi.org/10.1093/nar/gks454>

11. Niedzicka, M., Fijarczyk, A., Dudek, K., Stuglik, M., & Babik, W. (2016). Molecular Inversion Probes for targeted resequencing in non-model organisms. *Sci Rep*, 6, 24051. <https://doi.org/10.1038/srep24051>
12. PacBio. (2022). *Featured Documentation*. https://www.pacb.com/support/documentation/?fwp_workflow_step=library-preparation&fwp_sort=preserve
13. Pel, J., Leung, A., Choi, W. W. Y., Despotovic, M., Ung, W. L., Shibahara, G., Gelinas, L., & Marziali, A. (2018). Rapid and highly-specific generation of targeted DNA sequencing libraries enabled by linking capture probes with universal primers. *PLoS One*, 13(12), e0208283. <https://doi.org/10.1371/journal.pone.0208283>
14. Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sorensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*, 19(7), 823-826. <https://doi.org/10.1038/s41592-022-01539-7>
15. Shi, Y., Wang, G., Lau, H. C., & Yu, J. (2022). Metagenomic Sequencing for Microbial DNA in Human Samples: Emerging Technological Advances. *Int J Mol Sci*, 23(4). <https://doi.org/10.3390/ijms23042181>
16. Thermo Fisher Scientific. (n.d.-a). *Ion Torrent Next-Generation Sequencing Construct Library*. Retrieved 17/07/2022 from <https://www.thermofisher.com/ch/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-construct-library.html>
17. Thermo Fisher Scientific. (n.d.-b). *Next-Generation Sequencing Illumina Workflow-4 Key Steps*. Retrieved 14/07/2022 from <https://www.thermofisher.com/ch/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/illumina-workflow.html>
18. Thermo Fisher Scientific. (n.d.-c). *Preparation of DNA Sequencing Libraries for Illumina Systems-6 Key Steps in the Workflow*. Retrieved 14/07/2022 from <https://www.thermofisher.com/ch/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/dna-sequencing-preparation-illumina.html>
19. Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bahr, M., Wolf, S., Shendure, J., Eils, R., Plass, C., & Weichenhan, D. (2013). Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc*, 8(10), 2022-2032. <https://doi.org/10.1038/nprot.2013.118>
20. Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., & Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16, 675. <https://doi.org/10.1186/s12864-015-1876-7>

CHAPTER 4
DISEASE OUTBREAK AND COVID-19
PANDEMIC

CHAPTER 4: DISEASE OUTBREAK AND COVID-19 PANDEMIC

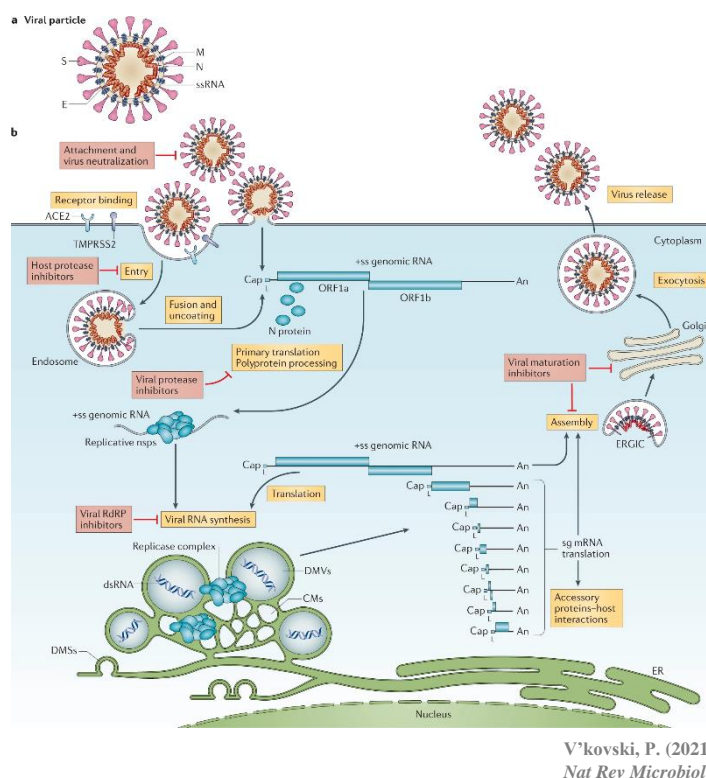
4.1. An Overview of COVID-19

COVID-19 is an infectious disease that was first identified in samples from Wuhan, China, in December of 2019, after which it quickly spread worldwide resulting in a global pandemic. So far (i.e., as of the time of this document's write-up), the COVID-19 pandemic has caused over 625 million infections and more than 6 million deaths (World Health Organization, 2022), with the true death toll being estimated to be between 14.6 and 25.4 million (The Economist, 2022), placing COVID-19 among the seven deadliest pandemics in history (Prabhu & Gergen, 2021).

COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a single-stranded, positive sensed RNA (+ssRNA)

virus of ~ 30 kb composed of four structural proteins: (i) a spike protein (S) that mediates entry into host cells, (ii) a membrane protein (M) that facilitates virion budding, (iii) an envelope protein (E) that facilitates assembly and release of virions, and (iv) a nucleocapsid protein (N) that protects the viral RNA genome and packages it into a ribonucleoprotein complex (Yang & Rao, 2021). The life cycle of the virus within the host consists

of five steps: viral entry, replication and transcription using host cell machinery, assembly and release (**Figure 26**).



V'kovski, P. (2021)
Nat Rev Microbiol

Figure 26. SARS-CoV-2 and its life cycle

Note. Reprinted from V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology*, 19(3), 155-170. <https://doi.org/10.1038/s41579-020-00468-6> Copyright 2020, Springer Nature Limited

The origin of SARS-CoV-2 is still being speculated. However, given a 96% genetic identity with coronaviruses found in bats (Zhou et al., 2020), SARS-CoV-2 is believed to be the result of a zoonotic transfer (“spillover”) from animal reservoirs, either directly or through an intermediate host (Brant et al., 2021). Given the fast rate of SARS-CoV-2 transmission, vaccines were seen as the most effective preventive measure against the virus and indeed, since their roll-out, there has been a substantial reduction in the incidence (number of new cases) and mortality rate of COVID-19 cases, with ~20 million deaths averted in 185 countries during the first year of vaccine implementation (Watson et al., 2022). As of 18 July 2022, ~67% of the world has received a first dose of a COVID-19 vaccine and 12 billion doses have been administered globally (Mathieu et al., 2021). Nevertheless, the emergence of newly mutated viral variants with higher infectivity and/or immune escape ability is concerning, as it threatens to undo the progress achieved by vaccinations through breakthrough infections (infections in fully vaccinated persons). Thus, continued surveillance is required to mitigate future outbreaks by new variants.

4.2. COVID-19 Surveillance

Surveillance reporting for COVID-19 generally requires specific data for monitoring new developments such as the number of confirmed cases⁵, deaths, hospital and intensive care unit (ICU) admissions, tests performed, doses administered, and persons vaccinated per dose. This information is collated and summarized as a visual display on a configurable web-based application called a dashboard. Many countries have deployed their own surveillance dashboards, allowing users to view and compare country-specific progression of COVID-19. However, a global collation of official COVID-19 surveillance data from governments and health ministries worldwide is available via “Our World in Data” (<https://ourworldindata.org/coronavirus>) as well as through the WHO (<https://covid19.who.int/>).

⁵ A confirmed case is defined as the positive identification of SARS-CoV-2 nucleic acid or antigen in a clinical specimen by a laboratory personnel.

Tracking SARS-CoV-2 Variants

For the first 11 months of the COVID-19 pandemic, SARS-CoV-2 was reportedly in a period of evolutionary stasis (equilibrium) (Harvey et al., 2021). However, since late 2020, the virus has notably evolved with the emergence of several variants of distinct genetic lineages (**Figure 27**). These variants mutations most notably in the spike protein, which have been associated with increased infectivity, disease severity and/or immune escape (Harvey et al., 2021). The World Health Organization (WHO) categorizes these emerging variants into two main classes (WHO, 2022):

1. **Variant of Interest (VOI):** VOIs are viral variants that possess genetic markers (or

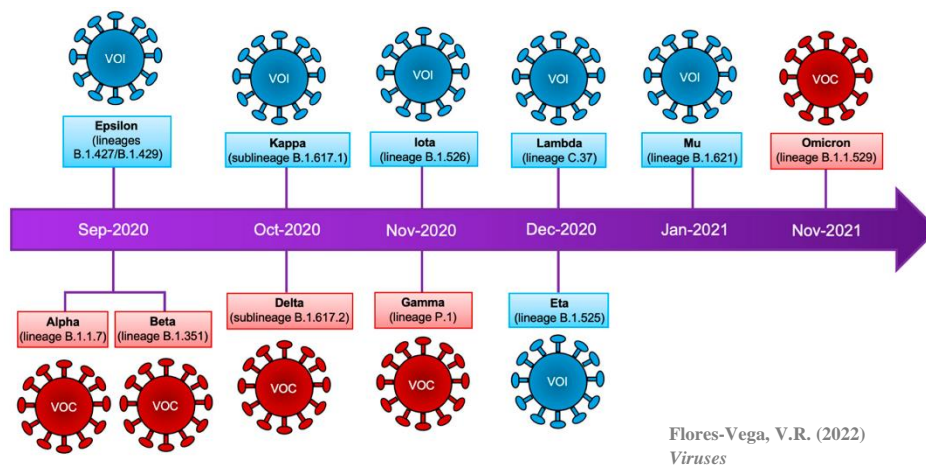


Figure 27. Timeline to emergence of SARS-CoV-2 variants

Note. Reprinted from Flores-Vega, V. R., Monroy-Molina, J. V., Jiménez-Hernández, L. E., Torres, A. G., Santos-Preciado, J. I., & Rosales-Reyes, R. (2022). SARS-CoV-2: Evolution and Emergence of New Viral Variants. *Viruses*, 14(4), 653. <https://doi.org/10.3390/v14040653> License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

mutations), which are predicted to affect transmission, diagnostics, therapeutics, or immune escape and are seen as an emerging risk to global health, having been identified in transmission events, multiple clusters, or countries (WHO, 2022).

2. **Variant of Concern (VOC):** VOCs are seen as highly significant to global public health due to their increased transmissibility, ability to cause a more severe clinical course, failure to be detected by diagnostic assays, escape from natural or vaccine-derived immunity and decreased susceptibility to therapeutics (WHO, 2022).

So far, five VOCs and eight VOIs have been identified (**Figure 27**). These have been subsequently classified into different lineages and clades (**Table 6**). Although no VOIs are currently circulating, the

newest VOC, known as Omicron , which was identified in November 2021 and harbours over 30 mutations in the spike protein compared to the index virus (Callaway, 2021), has since overtaken all other SARS-CoV-2 strains to become the predominant variant circulating globally.

Table 6. Classification of SARS-CoV-2 Variants of Concern

WHO	PANGO (Lineage)	GISAID (Clade)	Nextstrain (Clade)	Country of Identification
Alpha	B.1.17	GRY	20I (V1)	UK
Beta	B.1.1351	GH/501Y.V2	20H (V2)	South Africa
Gamma	P.1	GR/501Y.V3	20J (V3)	Japan/Brazil
Delta	B.1.617.2	G/478K.V1	21A	India
Omicron	B.1.1.529	BR/484A	21K	Botswana/South Africa

WHO, World Health Organization, PANGO, Phylogenetic Assignment of Named Global Outbreak Lineages; GISAID, Global Initiative on Sharing All Influenza Data; UK, United Kingdom; USA, United States of America.

Note. Reprinted from Flores-Vega, V. R., Monroy-Molina, J. V., Jiménez-Hernández, L. E., Torres, A. G., Santos-Preciado, J. I., & Rosales-Reyes, R. (2022). SARS-CoV-2: Evolution and Emergence of New Viral Variants. *Viruses*, 14(4), 653. <https://doi.org/10.3390/v14040653> License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

These variant observations have been largely accomplished by the sharing of SARS-CoV-2 genomic sequences via the Global Initiative on Sharing All Influenza Data (GISAID; <https://www.gisaid.org/>), which to date has received over 12 million genome sequence submissions. By leveraging this information, surveillance strategies have been able to track the emergence, spread and phylogeographical evolution of SARS-CoV-2 variants via Nextstrain (<https://nextstrain.org/ncov/gisaid/global/6m>), CoVariants (<https://covariants.org/>), CoV-Spectrum (<https://cov-spectrum.org/about>) as well as conduct mutation comparisons across lineages via outbreak.info (<https://outbreak.info/situation-reports>).

In addition, many initiatives and consortia have been established to coordinate SARS-CoV-2 sequencing efforts at the national and global levels. These have been summarized in **Table 7**.

Table 7. Examples of SARS-CoV-2 Sequencing Consortia

NAME	Country/Region	URL
Irish Coronavirus sequencing consortium	Ireland	https://www.teagasc.ie/food/research-and-innovation/research-areas/food-bioscience/irish-coronavirus-sequencing-consortium/
German COVID-19 OMICS Initiative (DeCOI)	Germany	https://decoi.eu/
Coronavirus Sequencing in Quebec (CoVSeQ)	Quebec, Canada	https://covseq.ca/
The COVID-19 host genetics initiative	Global	https://www.covid19hg.org/

COG-UK Project Hospital-Onset COVID-19 Infections Study (COG-UK HOCl)	UK	https://www.cogconsortium.uk/studies-publications/national-studies/the-hoci-study/
Mutational Dynamics of SARS-CoV-2 in Austria	Austria	https://www.sarscov2-austria.org/
SPHERES - sequencing for public health emergency response, epidemiology and surveillance consortium	USA	https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html
Genetics Of Mortality In Critical Care – The GenOMICC Study	UK-Based	https://genomicc.org/
Canadian COVID Genomics Network (CanCOGeN)	Canada	https://www.genomecanada.ca/en/cancogen
Africa CDC Institute for Pathogen Genomics	Africa	https://africacdc.org/africa-cdc-institutes/africa-cdc-institute-for-pathogen-genomics/
Netherlands sequencing efforts at RIVM (National Institute for Public Health and the Environment)	Netherlands	https://www.rivm.nl/en/news/update-on-spread-of-uk-coronavirus-variant-voc-20201201-in-netherlands
COVID Network for Genomics Surveillance South Africa (NGS-SA)	South Africa	http://www.krisp.org.za/ngs-sa/ngs-sa-network-for-genomic-surveillance-south-africa/
SeqCOVID	Spain	http://seqcovid.csic.es/
COVID-19 Network Investigations (CONI) Alliance	Thailand	https://coni.team/
COVID-19 Case in Cambodia	Cambodia	https://public.idseq.net/
Public Health Alliance for Genomic Epidemiology (PHA4GE)	Global	https://pha4ge.org/
COVID-19 High Performance Computing (HPC) Consortium	Global	https://covid19-hpc-consortium.org/
ARTIC network	Global	https://artic.network/
Indian SARS-CoV-2 Consortium on Genomics	India	https://pib.gov.in/PressReleaseframePage.aspx?PRID=1707177
Danish Covid-19 Genome Consortium (DCGC)	Denmark	https://www.covid19genomics.dk/home
National Institute of Infectious Diseases	Japan	https://www.niid.go.jp/niid/en/
Swiss SARS-CoV-2 Sequencing consortium	Switzerland	https://bsse.ethz.ch/cevo/cevo-press/2020/05/first-data-for-genomic-surveillance-of-sars-cov-2-in-switzerland-made-available.html

Note. Reprinted from Next generation sequencing for SARS-CoV-2. Foundation for Innovative New Diagnostics (FIND). 2020; Available from: <https://www.finddx.org/covid-19/covid-19-genomic-surveillance/>

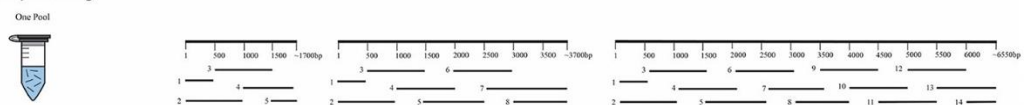
In summary, given the genetic plasticity (ability to mutate) of SARS-CoV-2, which threatens to undercut vaccine efforts, surveillance remains the best strategy for monitoring and tracking new developments related to COVID-19, ensuring better preparedness and prompt response to future outbreaks. Additionally, genomic surveillance of SARS-CoV-2, which has added a new dimension to COVID-19 surveillance, should be implemented in public health surveillance strategies, if possible. However, it should be noted that routine genomic surveillance does not require sequencing of every

single COVID-19 case, but rather representative populations to identify new variants and monitor trends in circulating variants (Centers for Disease Control and Prevention, 2022). Standards for genomic surveillance of SARS-CoV-2 variants are currently available in a WHO guidance document (https://www.who.int/publications/i/item/WHO_2019-nCoV_surveillance_variants).

4.3. SARS-CoV-2 Genome Sequencing using Nanopore Technology

Sequencing of the SARS-CoV-2 genome using ONT technology can be performed using a variety of protocols. However, the most common protocols used are those developed by the ARTIC network for viral surveillance (<https://artic.network/ncov-2019>), which have proven to be reproducible with a high sensitivity for clinical samples (Bull et al., 2020). These protocols mainly follow an amplicon-based approach, where multiplexed PCR tiling of samples is combined with sequencing to maximize coverage (**Figure 28**). Here, tiled amplicons of a specific size are generated from reverse-transcribed cDNA using overlapping primers that have been designed to span the whole viral genome. Afterwards the samples are barcoded using ONT's Native barcoding kit and libraries are prepared using ONT's ligation sequencing kit.

Whole Genome Multiplex Tiling



Taylor, M.K. (2020)
*Frontiers in Cellular and
Infection Microbiology*

Figure 28. Whole genome sequencing multiplex tiling approach

Note. Reprinted from Taylor, M. K., Williams, E. P., Wongsurawat, T., Jenjaroenpun, P., Nookaew, I., & Jonsson, C. B. (2020). Amplicon-based, next-generation sequencing approaches to characterize single nucleotide polymorphisms of Orthohantavirus species. *Frontiers in Cellular and Infection Microbiology*, 10, 565591. <https://doi.org/10.3389/fcimb.2020.565591> License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

Nanopore sequencing using the ARTIC protocols were analytically validated for its adoption in public health viral surveillance using Illumina sequencing as a comparison (Bull et al., 2020). The results of the analytical validation showed that ONT sequencing produces highly accurate consensus sequences with >99% sensitivity and >99% precision and enables variant detection comparative to Illumina at >99% sensitivity and >99% precision using the Nanopolish pipeline. However, the study also revealed the unsuitability of ONT in detecting indel variants and rare SNVs. Nevertheless, the standardized ARTIC

protocols have been utilized by various studies in ONT sequencing (Meredith et al., 2020; Rios et al., 2021; Yakovleva et al., 2021), and in some cases optimized (Freed et al., 2020; Li et al., 2020), for the characterization of SARS-CoV-2 viruses from clinical specimens. ONT can therefore be used for the reliable sequencing of SARS-CoV-2 in surveillance studies.

4.3.1. Nanopore Instrument Models & Throughput Capacity

A comprehensive comparison of all four ONT sequencing platforms is provided in **Table 8** below.

Table 8. Comparison of ONT Instrument Models

	Flongle ¹	MinION	MinION Mk1C ²	GridION	PromethION			
Models					P2 Solo ³	P2 ³	PromethION N 24	PromethION N 48
No of flow cells	1	1	1	5	2	2	24	48
Max no of channels / flow cell	126	512	512	512	2,675	2,675	2,675	2,675
Max run time	16 hrs	72 hrs	72 hrs	72 hrs	72 hrs	72 hrs	72 hrs	72 hrs
Max output	2.8 Gb	50 Gb	50 Gb	250 Gb	580 Gb	580 Tb	7 Tb	14 Tb
Power requirement	By laptop		25 W	800 W	N/A	N/A	2.2 kW	2.2 kW
Weight	20 g	87 g	450 g	11 kg	N/A	N/A	Sequencer: 28 kg Data Acquisition unit: 25 kg	
Connectivity	None	None	Ethernet, microSD, USB	1 Gb Ethernet	N/A	N/A	Dual 10 Gb Fibre or Ethernet	
Storage	None	None	1 TB SSD	4 TB SSD	N/A	N/A	60 TB SSD	
RAM	None	None	8 Gb	64 Gb	N/A	N/A	512 Gb	
Applications	Library QC, amplicons, plasmid, viral and bacterial sequencing	Whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome, small transcriptomes, smaller multiplexing		Larger genomes, whole transcriptomes, larger samples	Larger genome projects, population scale sequencing, highly multiplexed small genomes, whole transcriptomes (cDNA or direct RNA)			
Multiplexing								
WGS - small genomes	Low plex	Low to medium plex			Highly multiplexed			
WGS - large genomes	N/A	Low pass (low coverage)	Yes	Yes	Yes			
Targeted sequencing	Low plex	Low to medium plex			Highly multiplexed			
¹ Flongle flow cells are single-use and not reusable. They can also be used on the MinION and MinION Mk1C devices. ² The MinION Mk1C is an all-in-one device compatible with an integrated software for base-calling and analysis of sequencing performed on MinION and Flongle flow cells. ³ P2 (PromethION 2) devices are compact versions of PromethION 24 & 48, designed for small to medium sized academic and research labs to allow them explore applications enabled by high output nanopore sequencing before committing to the PromethION 24 or 48 sequencers. P2 has a self-contained GPU unit while P2 solo requires a host computer or GridION Nk1 for support. References: Oxford Nanopore Technologies. https://nanoporetech.com/products/specifications								

4.4. SARS-CoV-2 Sequencing Workflows

There are three different workflow options for SARS-CoV-2 sequencing using ONT devices (available in the Nanopore community <https://nanoporetech.com/community>) and depending on sample throughput and/or the type of NGS laboratory, any of these protocols may be used.

The ARTIC Classic protocol is based on the ARTIC protocol for nCov-2019 developed by Josh Quick (<https://artic.network/ncov-2019>). The workflow includes amplification of the SARS-CoV-2 genome in ~400 bp overlapping segments followed by a post-amplification normalization and quantification step (**Figure 29**). In generating the tiled amplicons, the multiplex PCR is carried out in two separate reactions using two different sets of primers (110 primers in pool 1 and 108 primers in pool 2) (primer set available on

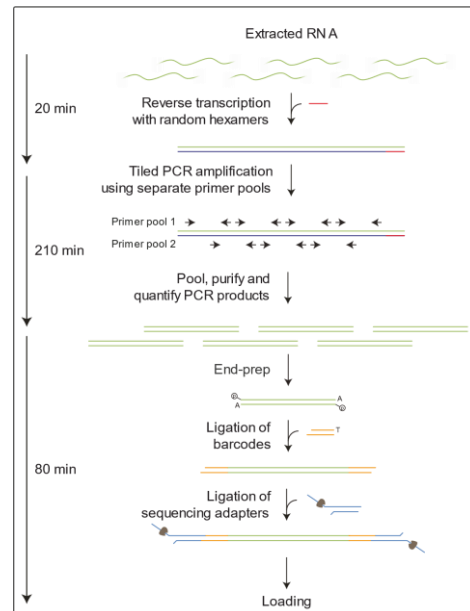


Figure 29. ARTIC Classic protocol for SARS-CoV-2 genome

Note. Reprinted from Oxford Nanopore Technologies. (2022). PCR tiling of SARS-CoV-2 virus- classic protocol (SQK-LSK109 with EXP-NBD196) https://community.nanoporetech.com/docs/prepare/sars_cov_2

<https://github.com/artic-network/artic->

[ncov2019/tree/master/primer_schemes/nCoV-](https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-)

[2019/V3](https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3)), which are then pooled together for

subsequent steps. Afterwards, amplicons are barcoded using the native barcoding kit, which allows for multiplexing of up to 96 samples per library, and sequencing libraries are prepared using the ligation kit. This protocol has been optimized for maximum coverage and is ideal for routine sequencing of smaller batches of samples with varying viral loads (i.e., different cycle threshold [CT] values quantified using a real-time RT-PCR). It is also best suited for labs with previous experience using ONT sequencing.

The ARTIC Eco protocol is based on the LoCost protocol (https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bp216n26rgqe/v3?version_warning=no), which is an adapted version of the ARTIC classic protocol, also developed by Josh Quick with the aim of reducing hands-

on time during library preparation and lowering sequencing costs by removing the post-amplification normalization step and using less reagents. This removal of the normalization and quantification steps post-amplification reduces hands-on-time by ~1-2 hrs depending on sample throughput while the use of less reaction volumes saves on reagent costs. This protocol also uses sequencing auxiliary vials and short fragment buffers in combination with the native barcoding kit for library preparation, which further reduces costs per sample compared to the classic protocol. However, this method can result in a dropout of samples with low viral loads or high CT values. Therefore, it is recommended to batch samples based on CT values to minimize such occurrences when using this protocol. The Eco protocol is ideal for getting started quickly with SARS-CoV-2 genome sequencing and is recommended for labs with medium level sample throughput.

The Midnight protocol is a new development that generates ~1200 bp amplicons in a tiled fashion across the SARS-CoV-2 genome using primers designed by Freed *et al.*, (Freed *et al.*, 2020) using the Primal scheme method to find primer panels (Quick *et al.*, 2017). Here, the number of primers used in the two PCR reaction pools are smaller than those used in the standard ARTIC protocol, with only 30 primers

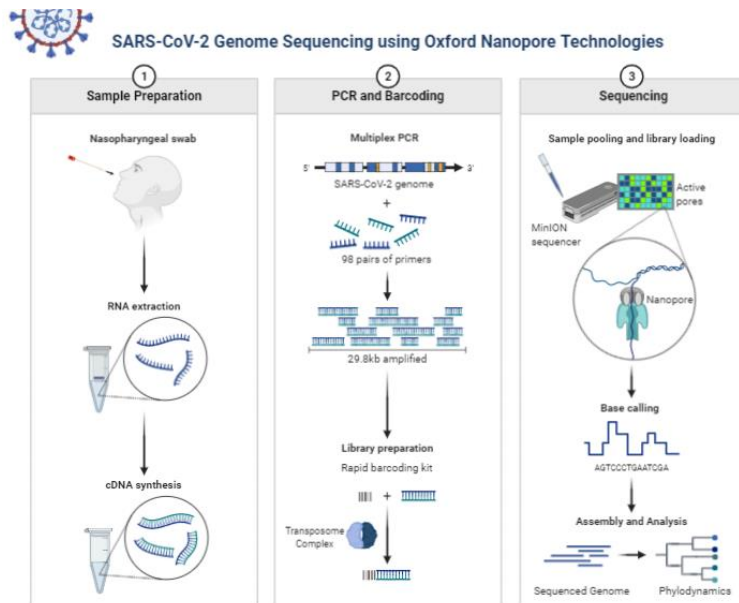


Figure 30. Midnight protocol for SARS-CoV-2 genome sequencing

Note. Reprinted from "SARS-CoV-2 Genome Sequencing using Oxford Nanopore Technologies", by BioRender, August 2020, retrieved from <https://app.biorender.com/biorender-templates> Copyright 2022 by BioRender.

used in pool 1 and 28 primers in pool 2. Details on these primer set are available in the published online protocol (<https://www.protocols.io/view/sars-cov2-genome-sequencing-protocol-1200bp-amplic-rm7vz8q64vx1/v6>).

The library preparation step is faster here than other workflows due to the simplicity of the rapid barcoding kit (**Figure 30**). Given its faster turnaround, reduced hands-on-time and simplicity, this protocol can be automated, especially where sample throughput is high, resulting in the lowest cost per sample for SARS-CoV-2 genome sequencing.

4.5. References

1. Brant, A. C., Tian, W., Majerciak, V., Yang, W., & Zheng, Z. M. (2021). SARS-CoV-2: from its discovery to genome structure, transcription, and replication. *Cell Biosci*, *11*(1), 136. <https://doi.org/10.1186/s13578-021-00643-z>
2. Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., Naing, Z., Yeang, M., Verich, A., Gamaarachchi, H., Kim, K. W., Luciani, F., Stelzer-Braid, S., Eden, J. S., Rawlinson, W. D., van Hal, S. J., & Deveson, I. W. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun*, *11*(1), 6272. <https://doi.org/10.1038/s41467-020-20075-6>
3. Callaway, E. (2021). Beyond Omicron: what's next for COVID's viral evolution. *Nature*, *600*(7888), 204-207. <https://doi.org/10.1038/d41586-021-03619-8>
4. Centers for Disease Control and Prevention. (2022). *What is Genomic Surveillance?* <https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html>
5. Freed, N. E., Vlkova, M., Faisal, M. B., & Silander, O. K. (2020). Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol Methods Protoc*, *5*(1), bpaa014. <https://doi.org/10.1093/biomethods/bpaa014>
6. Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Consortium, C.-G. U., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*, *19*(7), 409-424. <https://doi.org/10.1038/s41579-021-00573-0>
7. Li, J., Wang, H., Mao, L., Yu, H., Yu, X., Sun, Z., Qian, X., Cheng, S., Chen, S., Chen, J., Pan, J., Shi, J., & Wang, X. (2020). Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. *Sci Rep*, *10*(1), 17492. <https://doi.org/10.1038/s41598-020-74656-y>
8. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodes-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nat Hum Behav*, *5*(7), 947-953. <https://doi.org/10.1038/s41562-021-01122-8>
9. Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., Curran, M. D., Parmar, S., Caller, L. G., Caddy, S. L., Khokhar, F. A., Yakovleva, A., Hall, G., Feltwell, T., Forrest, S., Sridhar, S., Weekes, M. P., Baker, S., Brown, N., . . . Goodfellow, I. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis*, *20*(11), 1263-1272. [https://doi.org/10.1016/S1473-3099\(20\)30562-4](https://doi.org/10.1016/S1473-3099(20)30562-4)
10. Prabhu, M., & Gergen, J. (2021). *History's Seven Deadliest Plagues*. <https://www.gavi.org/vaccineswork/historys-seven-deadliest-plagues>
11. Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S. C., Black, A., Bedford, T., Carroll, M. W., Nunes, M., . . . Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc*, *12*(6), 1261-1276. <https://doi.org/10.1038/nprot.2017.066>

12. Rios, G., Lacoux, C., Leclercq, V., Diamant, A., Lebrigand, K., Lazuka, A., Soyeux, E., Lacroix, S., Fassy, J., Couesnon, A., Thiery, R., Mari, B., Pradier, C., Waldmann, R., & Barbry, P. (2021). Monitoring SARS-CoV-2 variants alterations in Nice neighborhoods by wastewater nanopore sequencing. *Lancet Reg Health Eur*, 10, 100202. <https://doi.org/10.1016/j.lanpe.2021.100202>
13. The Economist. (2022). The pandemic's true death toll. *The Economist*. <https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates>
14. Watson, O. J., Barnsley, G., Toor, J., Hogan, A. B., Winskill, P., & Ghani, A. C. (2022). Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis*. [https://doi.org/10.1016/S1473-3099\(22\)00320-6](https://doi.org/10.1016/S1473-3099(22)00320-6)
15. WHO. (2022). *Tracking SARS-CoV-2 variants*. Retrieved 30/07/2022 from <https://www.who.int/activities/tracking-SARS-CoV-2-variants>
16. World Health Organization. (2022). *WHO Coronavirus (COVID-19) Dashboard* <https://covid19.who.int/>
17. Yakovleva, A., Kovalenko, G., Redlinger, M., Liulchuk, M. G., Bortz, E., Zadorozhna, V. I., Scherbinska, A. M., Wertheim, J. O., Goodfellow, I., Meredith, L., & Vasylyeva, T. I. (2021). Tracking SARS-COV-2 Variants Using Nanopore Sequencing in Ukraine in Summer 2021. *Res Sq*. <https://doi.org/10.21203/rs.3.rs-1044446/v1>
18. Yang, H., & Rao, Z. (2021). Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nat Rev Microbiol*, 19(11), 685-700. <https://doi.org/10.1038/s41579-021-00630-8>
19. Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. D., Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., . . . Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270-273. <https://doi.org/10.1038/s41586-020-2012-7>

CHAPTER 5
BIOINFORMATIC ANALYSIS

CHAPTER 5: BIOINFORMATIC ANALYSIS

5.1. Introduction

Due to the massive amount and complexity of data generated by NGS platforms, computational tools and skills are necessary to process, analyse and interpret NGS data. In recent years, the field of bioinformatics - a discipline that develops and applies advanced computational tools for the analysis and interpretation of high-dimensional biological data (Oliver et al., 2015) – has seen considerable development, with the establishment of new, open-source tools and integrated analytic pipelines. These tools have become less complex and easier to use, allowing non-bioinformaticians to perform their own analysis with little to no support from highly trained bioinformaticians. The initial steps in learning how to perform one's own NGS bioinformatic analysis are understanding a typical NGS bioinformatics workflow and understanding the question(s) to be answered by sequencing.

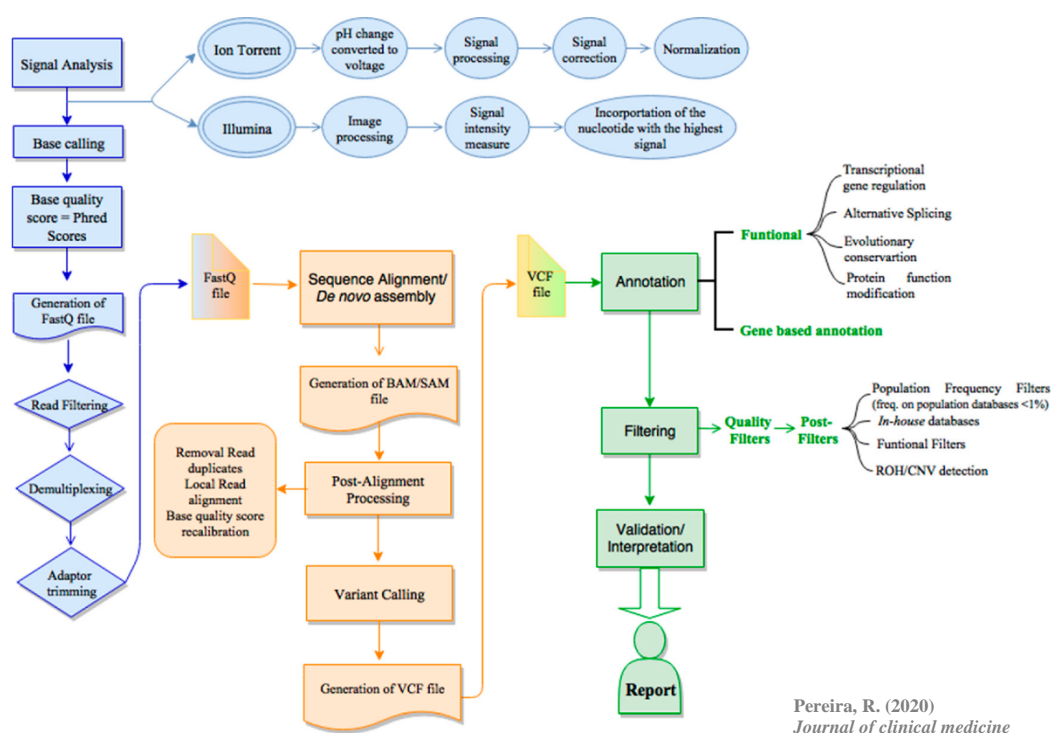


Figure 31. A next generation sequencing bioinformatics workflow

Primary analysis is depicted in blue, secondary analysis in orange and tertiary analysis in green

Note. Reprinted from Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of clinical medicine*, 9(1), 132. <https://doi.org/10.3390/jcm9010132> License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

Regardless of the NGS platform, a bioinformatic pipeline or workflow for NGS analysis can be divided into primary, secondary or tertiary analysis (**Figure 31**). During primary analysis, the raw sequencing

data is detected and converted into nucleotides and reads. This involves base calling and quality base score assignment, ending with a quality control clean-up of the generated raw reads (e.g., adapter trimming, demultiplexing, application of filters). In secondary analyses, the high quality filtered reads are aligned against a reference sequence (or *de novo* assembled) and variants are detected (variant calling). Finally, during tertiary analysis, the NGS information generated from the prior step is interpreted by associating it with sample-specific genomic profiles (Oliver et al., 2015). For example, in the case of detected variants, these may be annotated (i.e., assigned a functional or genetic information), filtered, visualized, and then linked to observed disease phenotypes.

It is important to note that all bioinformatic workflows differ depending on the research question or type of investigation. A microbiology lab investigating different bacterial strains may simply be interested in generating and comparing assembly profiles for each strain, thus foregoing the complicated variant analytic step depicted in **Figure 31**. On the other hand, a cancer research lab investigating different mutations and their associated human phenotypes (traits) will require more detailed variant analysis. Therefore, when choosing established bioinformatic tools and pipelines, it is important to consider their relevance to the research question or investigation. Such relevance attributes may include:

- **Target species and/or quality of data:** Some tools and pipelines are specially designed for certain organisms, NGS platforms or read types while others have a more flexible application (Brandies & Hogg, 2021). Reading associated published papers or the README file accompanying the tool will help in defining this relevance.
- **Available computing resources and time restrictions:** Some labs may have an integrated informatic infrastructure, such as a high performance computing (HPC) cluster that provides sufficient memory and computing cores for NGS analysis while others may be required to perform the analysis on personal computers, which have limited RAM or cores (Brandies & Hogg, 2021). Discussing with the IT personnel at your institution will be helpful in figuring out your needs.

- **Availability of tools:** Some tools require a license while others are freely available. Speaking with colleagues or asking questions on online forums (e.g., BioStars, Nanopore community) may assist in your decision-making (Brandies & Hogg, 2021).

5.2. File Formats for Sequences, Alignments and Annotation

Data generated from NGS platforms are stored in various standardized formats according to their application. FASTA/FASTQ formats are used for storing nucleotide or amino acid (protein) sequences. SAM/BAM/VCF formats are used for storing information from alignments of nucleotides or amino acids while GFF/GTF formats are used for storing annotation information.

Sequence Formats

- **FASTA:** This is the de facto standard for sequence data (Wong et al., 2019). The first line in the fasta file is referred to as the header and begins with a '>' (greater than) symbol followed by a description of the sequence, which sometimes includes a unique identifier known as the sequence identifier (**Figure 32**). After the header line is the sequence of the nucleic acid or protein in one-letter code. Fasta files can contain one sequence or multiple sequences. The latter is referred to as a multi-fasta file.

```
>OW444422.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2 2019-
nCoV_Muc-IMB1_3.Vero-E6 in hamster H.310 genome assembly, complete genome: monopartite
CCCAGGTAACAAACCAACCAACTTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACCTTAAAATCTGTG
TGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACA
GGACACGAGTAACCTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGC
ACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGCCCTGGTTTCAACGAGAAA
```

Figure 32. An example fasta file

- **FASTQ:** The fastq format extends the fasta format by adding corresponding quality scores for each nucleotide. Fastq files have four lines per sequence. **Line 1** begins with a '@' character and is followed by a sequence identifier and an optional description. **Line 2** is the sequence. **Line 3** begins with a '+' character and is optionally followed again by the same sequence identifier and description. **Line 4** encodes the quality values for the sequence, which corresponds to a quality score (**Figure 33**).

'##' and a data field with eight mandatory columns ((CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO) (**Figure 35**).

Note. Reprinted from Dave Tang. https://davetang.github.io/learning_vcf_file/

On SNP databases such as that of the National Center for Biotechnology Information (NCBI),

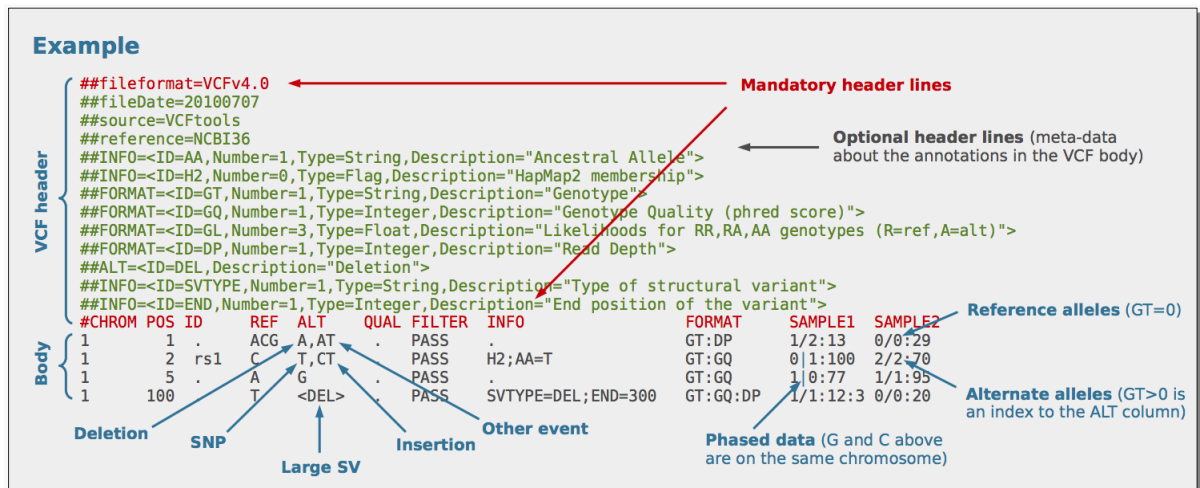


Figure 35. An example VCF format

variant changes are denoted using a '>' symbol. For example, an adenine (A) at a genomic position 364, which normally has a thymine (T) will be denoted as 'g.364T>A'. For transcripts, the 'g' is replaced with a 'c', e.g., c.364T>A, while proteins use a 'p'. This information is usually placed behind the sequence identifier, after a ':' (colon) e.g., NC_000147.1:g.364T>A. More information on the SNP nomenclature of NCBI databases can be found here: https://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs_attributes.html

Annotation Formats

- **GFF:** The General Feature Format (GFF) is a text-based file format used to describe genes and other features of DNA, RNA and protein sequences. GFF is the standard for genome annotations and can be used for any feature type (transcripts, exons, introns, promoter, etc.). The structure of the GFF consists of one line per feature and 9 fields/columns (**Figure 36**). The newest version of GFF (GFF3) is an improvement over the previous version (GFF2), which

has since been deprecated due to its limitation of not being to handle the three-level hierarchy of gene → transcript → exon (Institute for Systems Genomics, n.d.).



Figure 36. An example GFF format

Note. Reprinted from Bonatelli, M. (2021). Bacterial Taxonomy with whole-genome data. https://ucdavis-bioinformatics-training.github.io/2021-ASM-genome-assembly/markdown_docs/Bacterial-taxonomy

- **GTF:** The Gene Transfer Format (GTF) has the same format and structure as the GFF for the first 7 fields but differs in the content and format of the 9th field. GTF is also primarily used for genes/transcripts (Institute for Systems Genomics, n.d.).

5.3. NGS Data Quality Control

Quality control of NGS data is critical at the end of each stage of the bioinformatics workflow. **In the pre-processing stage**, quality control is performed by removing adapters, demultiplexing reads and applying filters to remove low-quality reads (e.g., reads below the required read length or reads with lower base quality scores). **In the secondary analysis step**, quality control is also performed after mapping reads to a reference genome by interrogating unmapped reads to identify PCR artefacts such as duplicated reads or potential contaminants, which may have been introduced during library preparation. Unmapped reads can be blasted using the Basic Local Alignment Search Tool (BLAST; <https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to find identical sequences, which may allow one to infer the origin or source of a sequence contamination. **In microbiology labs, the most frequent source of contamination are bacteria present on the skin surface, which can be easily introduced into samples, if care is not taken. It is therefore imperative that all lab personnel follow the guidelines for good laboratory practice to minimize the introduction of such contaminants during wet lab preparation. Some of the tools used in quality control are provided in Table 9.**

Note. Tools vs. Pipelines: In bioinformatics, it is quite common to see the terms, “tool” and “pipeline”, used interchangeably, however, they have slightly different interpretations. While a “tool” mainly uses algorithms for a single type of analysis, a “pipeline” combines statistical models, computational tools and/or algorithms in a series of steps to process raw sequencing data and interpret it. In essence, a tool is constructed from algorithms for a single process whereas a pipeline is an integration of various algorithms and tools for multiple processes.

Table 9. A selection of bioinformatic tools for QC Analysis

NGS Technology	Bioinformatic QC Tool	QC Process	Available via
Illumina			
	FastQC	Generates single QC reports	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	MultiQC	Generates QC report for multiple samples at once	https://multiqc.info/
	FastUniq	Duplicate removal	https://sourceforge.net/projects/fastuniq/
	Picard	Marks duplicates for removal by 3 rd party software e.g., Samtools	https://broadinstitute.github.io/picard/
ONT			
	NanoFilt	Filters and trims reads	https://github.com/wdecoster/nanofilt
	Porechop	Trims Adapters, demultiplexes barcodes	https://github.com/rrwick/Porechop
	Filtlong	Filters long reads by quality	https://github.com/rrwick/Filtlong
	MiniScrub	Removes low quality ONT reads	https://bitbucket.org/berkeleylab/jgi-miniscrub/src/master/
PacBio			
	Lima	Trims Adapters, removes primers, demultiplexes barcodes	https://lima.how/
	zwmfilter	Filters PacBio BAM data on ZMW IDs	https://github.com/PacificBiosciences/zwmfilter/
	pbmarkdup	Mark duplicate reads from an amplified library	https://github.com/PacificBiosciences/pbmarkdup/

5.4. Tools and Pipelines for NGS Analysis

There are two main types of analysis that can be performed with quality processed NGS data.

- **Genome Assembly:** This is a process of aligning and merging sequenced reads to reconstruct the original sequence. There are two types of assembly processes: **Reference-**

based alignment is a process where reads are assembled by mapping to a representative genomic example of the sequence that needs to be assembled, otherwise known as a reference genome whereas **de novo assembly** is the process of assembling a genome from scratch without using a reference genome. A type of de novo assembly approach called “**hybrid assembly**” leverages the accuracy of short reads to polish error-prone, long reads during the assembly process, thereby generating highly accurate, complete assemblies. **Polishing** is the process of improving the base accuracy of assembled sequences.

- **Variation analysis:** This is the identification of variants such as SNPs/SNVs, INDELs, SVs in reads that have been mapped to a reference sequence or genome. Variants may reflect a variation in a species genetic sequence with no phenotypic effect or represent true genetic mutations that produce observable phenotypes e.g., eye colour or in worse cases, diseases. Variants of the latter are usually annotated and further studied to confirm their function.

A non-exhaustive list of tools and pipelines that are commonly used for NGS data analysis are provided in **Table 10**.

Table 10. Selected bioinformatic tools and pipelines for NGS Data Analysis

	Tool/Pipeline	Applicable technologies	Organism-specific?	Available via
Aligners				
	Minimap2	Short and long reads	No	https://github.com/lh3/minimap2
	BWA	Short to moderate reads (up to 1Mb)	No	http://bio-bwa.sourceforge.net/
	Bowtie2	Ideal for short reads. Can align longer reads, but slower	No	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
	MUMmer	Short and long reads	No	https://github.com/mummer4/mummer
	Mafft	Short and long reads	No	https://mafft.cbrc.jp/alignment/software/
	NanoPipe	ONT reads	No	https://github.com/IOB-Muenster/nanopipe2
Assemblers				
	SPAdes	Short reads, long reads, hybrid assembly	No	https://github.com/ablab/spades
	Unicycler*	Short reads, long reads, hybrid assembly	Yes, Bacteria	https://github.com/rrwick/Unicycler
	Tricycler	Long reads	Yes, Bacteria	https://github.com/rrwick/Tricycler
	Flye	Long reads	No	https://github.com/fenderglass/Flye

	Raven*	Long uncorrected reads	No	https://github.com/lbcb-sci/raven
	Miniasm*	Long reads	No	https://github.com/lh3/miniasm
	Haploflow	ONT reads	Yes, Viruses	https://github.com/hzi-bifo/Haploflow
	Shasta	Long reads (optimized for ONT but can be used for PacBio)	No	https://github.com/chanzuckerberg/s_hasta
	Canu*	Long reads	No	https://github.com/marbl/canu
	NextDenovo*	ONT and PacBio CLR reads	No	https://github.com/Nextomics/NextDenovo
	Wtdbg2/Redbean*	Long reads	No	https://github.com/ruanjue/wtdbg2
	MaSuRCA	Short or long reads	No	https://github.com/alekseyzimin/masurca
	Wengan	Hybrid assembly	No	https://github.com/adigenova/wengan
	Haslr	Hybrid assembly	No	https://github.com/vpc-ccg/haslr
	WeFaceNano	ONT reads	Yes, Bacterial plasmids	https://github.com/ErasmusMC-Bioinformatics/WeFaceNano
	microPIPE	ONT & Illumina reads	Yes, Bacteria	https://github.com/BeatsonLab-MicrobialGenomics/micropipe
Polishing				
	Medaka	ONT reads	No	https://github.com/nanoporetech/medaka
	Pilon	Optimized for Illumina reads	No	https://github.com/broadinstitute/pilon/wiki
	NextPolish	Optimized for long read assemblies	No	https://github.com/Nextomics/NextPolish
	Racon	Short or long reads	No	https://github.com/lbcb-sci/racon
	Polypolish	Short reads	Optimized for Bacterial genomes	https://github.com/rrwick/Polypolish/wiki
	Hypo	Short or long reads	No	https://github.com/kensung-lab/hypo
	ntEdit	Short or long reads	No	https://github.com/bcgsc/ntEdit
	Minipolish (This is basically Racon in graph format)	Short or long reads	No	https://github.com/rrwick/Minipolish
	Nanopolish	ONT reads	No	https://github.com/jts/nanopolish
Variant callers				
	GATK	Short reads	No	https://gatk.broadinstitute.org/hc/en-us
	Varscan2	Requires pileup file from samtools (https://www.htslib.org/docs/samtools.html)	No	https://dkoboldt.github.io/varscan/
	Medaka	ONT reads	No	https://github.com/nanoporetech/medaka
	Pilon	Optimized for Illumina reads	No	https://github.com/broadinstitute/pilon/wiki
	Minorseq	PacBio	No	https://github.com/PacificBiosciences/minorseq/
	Nanocaller	ONT reads	No	https://github.com/WGLab/NanoCaller
	Nanopolish	ONT reads	No	https://github.com/jts/nanopolish
*Some tools were developed for older chemistries of long read technologies, which had higher error rates (~13-15% errors). As newer chemistries of current ONT and PacBio sequencers have improved accuracies (>99%), these tools may not be applicable for them. Users are therefore recommended to always read the full description of a tool/pipeline and verify their applicability to updated versions of a NGS technology.				

Note. The tools and pipelines outlined in **Table 10** are applicable to genomic-based studies and do not cover transcriptomic or proteomic analysis. However, tools and pipelines for the latter may be found in the resources outlined under the “Recommended Reading” section below.

5.5. NGS Data Visualization & Exploration

There are many tools available for visualizing NGS data that has been processed and analyzed. Some require a license while others are open source. Commercially available software have the advantage of supporting an automated, streamlined analysis and visualization of NGS data, albeit at a license cost, which may be a worthy investment for non-bioinformatically trained personnel. On the other hand, open-source tools/pipelines are free and flexible for different applications but may require some customization and bioinformatics skills (scripting) for high-throughput analysis and visualization. Some examples of commercially available software and applicable open-source tools for NGS visualization have been summarized in **Table 11** below.

Table 11. Selection of bioinformatic tools and software for NGS data visualization

NGS Bioinformatic tool/pipeline	Comment
Commercial software	
Geneious prime	<ul style="list-style-type: none"> ○ 30-day free trial ○ Full NGS data analysis workflow (pre-processing, alignment, assembly, variant calling) ○ Limited options for long read analysis ○ Exportable publication-ready reports
QIAGEN CLC Genomic Workbench	<ul style="list-style-type: none"> ○ 14-day free trial ○ Complete workflows for genomics, transcriptomics, epigenomics and metagenomics ○ Supports both short and long read technologies ○ Exportable publication-ready reports
Softgenetics NextGENe	<ul style="list-style-type: none"> ○ 30-day free trial ○ Variant analysis, alignment and assembly ○ Mainly for Illumina and Ion Torrent platforms
DNASTAR Lasergene	<ul style="list-style-type: none"> ○ 14-day free trial ○ Supports multiple workflow applications, available in differently priced packages (Lasergene Molecular Biology, Lasergene genomics, Lasergene protein, DNASTAR Lasergene) ○ Supports analysis both short and long read technologies
Open source	
Integrative Genomic Viewer (IGV)	<ul style="list-style-type: none"> ○ Easy to use graphical user interface (GUI) ○ No programming skills required ○ Requires good quality reference sequence ○ Allows high-quality visualization of variants
Circos	<ul style="list-style-type: none"> ○ Steep learning curve (knowledge of perl is helpful) ○ Generates beautiful circular visualization images of high publication quality ○ Suitable for showing transmission dynamics in surveillance and outbreak studies

R Bioconductor	<ul style="list-style-type: none"> ○ Multiple software options available for every skill level (beginner to advanced) ○ Requires basic knowledge of R programming
cBioPortal	<ul style="list-style-type: none"> ○ Visualization of cancer genomics datasets ○ Easy-to-use web interface
NanoGalaxy	<ul style="list-style-type: none"> ○ Complete pipeline for pre-processing, analysis and visualization ○ Supports ONT reads ○ Web-based interface
NanoPipe	<ul style="list-style-type: none"> ○ Visualizes alignment and polymorphisms ○ Supports only ONT reads ○ Web-based interface
Alvis	<ul style="list-style-type: none"> ○ Visualizes alignments and detects chimeras ○ Supports only ONT reads ○ Command-line interface (requires some knowledge of Linux)
Artemis	<ul style="list-style-type: none"> ○ Visualization of sequences and annotation ○ Straightforward GUI ○ Can optionally be used on the command-line
QUAST	<ul style="list-style-type: none"> ○ Outputs quality assessment reports for genome assemblies ○ Command-line and web interface ○ Supports Illumina, ONT and PacBio reads
Icarus	<ul style="list-style-type: none"> ○ Visualizes draft genome assemblies ○ Supports Illumina, PacBio and ONT reads ○ Command-line interface
BLAST Ring Image Generator (BRIG)	<ul style="list-style-type: none"> ○ Easy-to-use GUI ○ Requires multiple finetuning steps to generate beautiful images ○ Is not actively updated (Last update was in 2018 and bugs have been identified since then)

Recommended Reading

For more information on Bioinformatic tools/pipelines for NGS data processing, analysis and visualization, check out the following resources:

- PacBio tools (distributed via Bioconda) <https://github.com/PacificBiosciences/pbbioconda>
- Nanopore resource center [https://nanoporetech.com/resource-centre?tags\[value\]\[0\]=tools](https://nanoporetech.com/resource-centre?tags[value][0]=tools)
- Bioinformatics for Researchers in Life Sciences: Tools and Learning Resources <https://publications.iadb.org/en/bioinformatics-researchers-life-sciences-tools-and-learning-resources>
- IGV Tutorial https://www.igv.org/workshops/NCIApril2017/IGV_SlideDeck.pdf
- Artemis- DNA Plotter Tutorial <https://home.cc.umanitoba.ca/~psgendb/tutorials/artemis/dnaplotter/dnaplotter.html>

5.6. SARS-CoV-2 Analysis Workflow

Since the publication of the first SARS-CoV-2 genomic sequence, numerous bioinformatic tools and pipelines have been established to aid in the analysis of SARS-CoV-2 NGS data. Some of these include the Genome Detective Virus Tool (Cleemput et al., 2020), CosmosID-HUB COVID-19 (<https://www.cosmosid.com/cosmosid-hub-covid-19/>), GalaxyProject SARS-CoV-2 workflows (<https://galaxyproject.org/projects/covid19/>), Qiagen CoV-2 Insights Service (Qiagen, n.d.), V-Pipe (Posada-Céspedes et al., 2021), HaVoC (Truong Nguyen et al., 2021), etc. In addition, the ARTIC

network for viral surveillance, who were responsible for developing standardized protocols for SARS-CoV-2 sequencing using ONT sequencing devices, have developed an interactive web application (InterARTIC) for the analysis of viral WGS data generated from nanopore sequencing (Ferguson et al., 2021). According to the developers, this application does not require any bioinformatics experience or third-party software installation and can be easily used to analyse nanopore sequencing data and assemble complete viral genomes from individual patient isolates (Ferguson et al., 2021). InterARTIC can be downloaded via github (<https://github.com/Psy-Fer/interARTIC/>), installed and executed on a standard laptop/desktop PC or alternatively on a GridION/PromethION. A video tutorial to assist in setting up the InterARTIC is available on youtube: <https://youtu.be/RCArn-xOkHg>

The InterARTIC workflow has already been summarized by Ferguson et al (Ferguson et al., 2021). In brief, the workflow starts with a fastq input generated from the ONT Guppy basecaller and performs

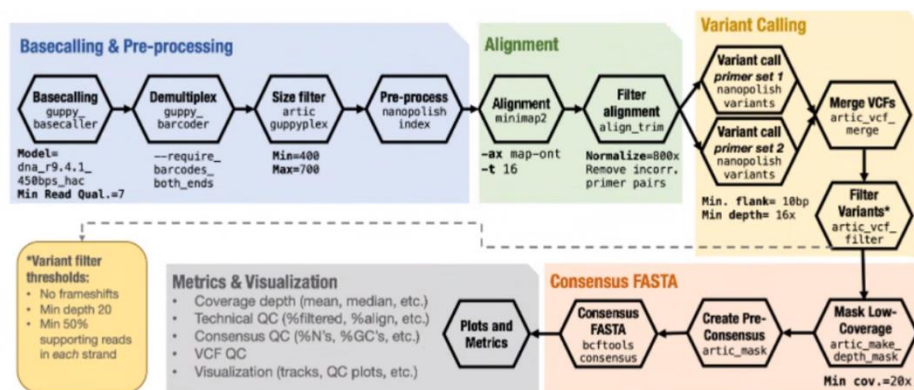


Figure 37. A InterARTIC-based workflow for SARS-CoV-2 sequencing analysis SARS-CoV-2 sequencing analysis

an optional demultiplexing step using the porechop tool. Next, the sequencing reads are aligned to a viral reference genome using Minimap2, primers are trimmed, and genetic variants are identified relative to the reference genome either via Nanopolish, Medaka or both. Afterwards, low quality variants are filtered out and low-quality regions are masked from the reference genome. The final variant candidates are then incorporated into the masked reference genome to produce a consensus genome sequence for the sequenced viral isolate. This consensus genome sequence can then be used in other downstream applications such as lineage classification (via Pangolin) and phylogeographic analysis via Nextstrain and afterwards uploaded on a public repository (e.g., GISAID).

Alternatively, users with more bioinformatic experience, can create their own workflow for SARS-CoV-2 analysis, as shown in **Figure 37**.

Recommended Reading

For more information on Bioinformatic tools/pipelines for SARS-CoV-2 analysis, check out the following resources:

- CDC repository of SARS-CoV-2 sequencing resources https://github.com/CDCgov/SARS-CoV-2_Sequencing
- German Network for Bioinformatics Infrastructure (de.NBI) <https://www.denbi.de/covid-19/coronavirus-tools>

5.7. Reference Databases

Due to the large amount of data produced by NGS platforms, online sequence databases have been established to enable the storage of and access to sequences (DNA, RNA or protein) of various organisms, stemming from multiple sources, thereby promoting international collaboration and ensuring scientific reproducibility. Examples of nucleotide sequence databases include the Genbank of the NCBI (<https://www.ncbi.nlm.nih.gov/genbank/>), the European Nucleotide Archive (ENA) of European Molecular Biological Laboratory and Bioinformatics Institute (EMBL-EBI) (<https://www.ebi.ac.uk/ena/browser/home>) and the DNA Data Bank of Japan (DDBJ) (<https://www.ddbj.nig.ac.jp/index-e.html>).

Although nucleotide databases are publicly accessible, the data submitted on these platforms are not curated (subjected to quality assurance checks) before release. Thus, sequences may be unreliable for certain comparative studies such as variant analysis, as users cannot easily verify sources of sequence bias (e.g., sample quality, PCR artefacts, sequencing technology). A solution to this has been the establishment of reference sequence databases. A reference sequence database is an open-access, curated collection of non-redundant sequences representing genomes, transcripts, and proteins for selected microbes, viruses, organelles and eukaryotic organisms (O'Leary et al., 2016). As it is subject to quality assurance checks before release, reference sequences provide a stable and consistent system for reporting gene-specific data, clinical variation and cross-species comparisons (O'Leary et al., 2016). NCBI's Reference Sequence Database, RefSeq (<https://www.ncbi.nlm.nih.gov/refseq/>), is one of the

most widely used and cited databases for high-quality reference sequences in biomedical research (Pruitt et al., 2005).

For sequenced SARS-CoV-2 sequences, GISAID is the most commonly used database. GISAID is a public-private partnership that was initially established to promote the international collection, sharing and analysis of influenza virus sequences (<https://gisaid.org/>). However, since the start of the COVID-19 pandemic, it now serves an additional purpose for the global sharing of SARS-CoV-2 genome sequence data, which are encouraged to be uploaded on the database, so they can be used to monitor the progression of the disease worldwide.

5.8. IT Infrastructure and Data Management

Deciding on the IT infrastructure for running your bioinformatic pipelines or tools will depend on the size of your data/computing requirements, species of interest, NGS technology, prior experience, timeline and budget (Brandies & Hogg, 2021). Estimating the computing requirements for your lab will allow you to have an approximate idea of the IT infrastructure that is most suitable for your needs. Although many institutions have a local HPC or access to a national/international HPC infrastructure for bioinformatic analysis and storage, these are currently being replaced by the more scalable cloud computing resources, which offer a number of advantages over the traditional shared HPC resources such as (i) unlimited scalability and ease of reproducibility, (ii) lack of queuing system for running jobs, allowing faster analysis and processing of scripts, (iii) flexibility in tailoring computing resources for each bioinformatic tool or pipeline and (iv) complete control over one's computing environment (Brandies & Hogg, 2021). Despite these benefits, commercial cloud computing resources have associated costs and a steep learning curve. However, simplified and cost-effective cloud solutions such as RONIN (<https://ronin.cloud/>) are available for researchers as well as free cloud computing services like Galaxy (<https://usegalaxy.org/>), ecocloud (<https://ecocloud.org.au/>), nectar (<https://ardc.edu.au/services/nectar-research-cloud/>) and cyverse (<https://cyverse.org/>) (Brandies & Hogg, 2021).

Regarding data storage, external hard drives offer the simplest solution for small-scale studies. However, these have the potential to be physically damaged or the possibility of viruses transferred with the data resulting in data loss. Laboratories with a well set up IT Infrastructure may offer well-secured local servers for storage of NGS data. Nonetheless, cloud services offer the most secure and long-term option for data storage. Amazon web services (<https://aws.amazon.com/health/genomics/>) currently offers multiple genomic tools and services for genomic-based applications.

5.9. References

1. Brandies, P. A., & Hogg, C. J. (2021). Ten simple rules for getting started with command-line bioinformatics. *PLoS Comput Biol*, 17(2), e1008645. <https://doi.org/10.1371/journal.pcbi.1008645>
2. Cleemput, S., Dumon, W., Fonseca, V., Abdool Karim, W., Giovanetti, M., Alcantara, L. C., Deforche, K., & de Oliveira, T. (2020). Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*, 36(11), 3552-3555. <https://doi.org/10.1093/bioinformatics/btaa145>
3. Ferguson, J. M., Gamaarachchi, H., Nguyen, T., Gollon, A., Tong, S., Aquilina-Reid, C., Bowen-James, R., & Deveson, I. W. (2021). InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btab846>
4. Institute for Systems Genomics, C. B. C. (n.d.). *File Formats Tutorial*. Retrieved 21/07/2022 from <https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/>
5. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1), D733-745. <https://doi.org/10.1093/nar/gkv1189>
6. Oliver, G. R., Hart, S. N., & Klee, E. W. (2015). Bioinformatics for clinical next generation sequencing. *Clin Chem*, 61(1), 124-135. <https://doi.org/10.1373/clinchem.2014.224360>
7. Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J., & Beerenwinkel, N. (2021). V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btab015>
8. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue), D501-504. <https://doi.org/10.1093/nar/gki025>
9. Qiagen. (n.d.). *Analysis of SARS-CoV-2 data*. Retrieved 31/07/2022 from <https://digitalinsights.qiagen.com/resources/science/sars-cov-2-resources/>
10. Truong Nguyen, P. T., Plyusnin, I., Sironen, T., Vapalahti, O., Kant, R., & Smura, T. (2021). HAVoC, a bioinformatic pipeline for reference-based consensus assembly and lineage assignment for SARS-CoV-2 sequences. *BMC Bioinformatics*, 22(1), 373. <https://doi.org/10.1186/s12859-021-04294-2>
11. Wong, K.-C., Zhang, J., Yan, S., Li, X., Lin, Q., Kwong, S., & Liang, C. (2019). DNA sequencing technologies: sequencing data protocols and bioinformatics tools. *ACM Computing Surveys (CSUR)*, 52(5), 1-30.

CHAPTER 6
SETTING UP AN NGS LAB

CHAPTER 6: SETTING UP AN NGS LAB

6.1. Infrastructure

Setting up an NGS laboratory brings with it a unique set of challenges, which differ depending on the setting. NGS platforms generally require hours to days for massive parallel processing of samples, therefore an optimal laboratory infrastructure is critical to the success of their routine use. Thus, in setting up an NGS laboratory, the following factors should be considered when determining which infrastructure (i.e., rooms) to host a NGS instrument.

- **Separation of pre-PCR and post-PCR work areas:** In a NGS laboratory, pre-PCR and post-PCR activities should always be conducted in physically separate areas to prevent the risk of contamination by amplicons spread through aerosols in the laboratory environment. Ideally, two separate pre-PCR rooms should be used: one for PCR master mix preparation and the other for sample preparation (nucleic extraction)/template addition (**Figure 38**). The third room, i.e., the post-PCR room can be used for PCR amplification, post-amplification, sequencing and analysis (**Figure 38**). However, this can also be split into two rooms, depending on space availability with a partition in the Room 2 to include master mix preparation room preparation,

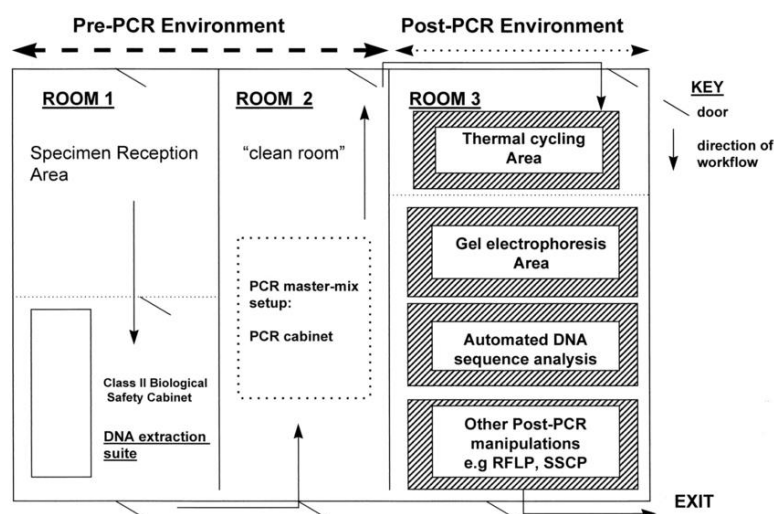


Figure 38. Layout of a NGS laboratory

Note. Reprinted from Millar, B. C., Xu, J., & Moore, J. E. (2002). Risk assessment models and contamination management: implications for broad-range ribosomal DNA PCR as a diagnostic tool in medical bacteriology. *Journal of Clinical Microbiology*, 40(5), 1575-1580. Copyright 2002, American Society for Microbiology.

Each room should have its own separate set of consumables (pipettes, pipette tips, lab reagents, lab coats, gloves) and equipment (vortex, mini centrifuge), which are clearly labelled. Gloves and lab coats should be changed when moving between rooms, and reagents and equipment should not be moved between rooms. Each room should have one fridge and freezer for storing reagents and working stocks. The pre-PCR room(s) should be additionally equipped with laminar flow cabinets for extraction and PCR-based workflows. Lastly, workflows should proceed unidirectionally from the pre-PCR room to the post-PCR room.

- **Protection from high temperatures:** Excessive heat can affect the performance of NGS platforms (Association of Public Health Laboratories, 2016). Therefore, the room hosting these instruments should have adequate cooling controls to prevent NGS platforms from becoming overheated. The Instrument should also not be placed in direct sunlight, though, if unavoidable, it is recommended to use a screen or shading during operation.
- **Protection from humidity:** Manufacturers recommend sequencers be placed in an area with 20-60% humidity for optimal performance (Association of Public Health Laboratories, 2016). Laboratories located in extreme conditions (e.g., high altitude and low humid areas) may require further optimization to account for humidity differences.
- **Vibration free space/zone:** Sequencers are notably sensitive to vibrations, which can affect their performance. Sources of vibrations include nearby centrifuges or instruments on the same surface/bench, opening and closing of room doors, unstable table legs. Air conditioning units, which are commonly used in countries with tropical climates, also produce noise and vibration that can be transferred to the building structure. Depending on the laboratory infrastructure, it may be worth investing in some anti vibration pads for Heating, Ventilation and Air Conditioning (HVAC) to minimize/control these effects.
- **Low traffic zone:** The NGS Instrument should not be placed in a high traffic area (e.g., a room that serves too many purposes) to prevent accidental bumping by lab personnel.

- **Uninterrupted power supply:** It is especially important that the electricity supply for a NGS platform is secure and surge-free (Yu, 2014). The electricity should be supplied by an uninterrupted power source which is ideally backed up by a generator/battery back-up, since sequencing can last a couple of days.
- **Informatics network:** NGS platforms generate massive amounts of data, which require sophisticated network connections for their storage, management, and analysis (Yu, 2014). It is therefore imperative that the platform is placed in an area that will allow for easy and efficient transfer of data from site of generation to the analysis and storage servers. Ideally, this transfer is facilitated using a desktop computer with sufficient memory that is directly connected to the NGS platform and designated solely for its use.

6.2. Consumables & Equipment

Several consumables and equipment support the sample preparation and sequencing process and are therefore necessary when performing NGS. These have been summarized in **Annexure 3**.

6.3. Personnel

In setting up an NGS lab, the time and cost required to effectively train laboratory personnel must also be considered. However, this will vary from laboratory to laboratory and with reference to time, will largely depend on the molecular background of the individual, which can take as little as a few weeks to a couple of months. Training a new staff member usually starts with the trainee observing an experienced staff member until they are comfortable with the protocol, after which the trainee can perform the procedure under observance from the trainer/experienced staff member. **According to the Association of Public Health Laboratories (APHL), the minimal number of times for a procedure to be performed under observation and then independently for training is three times** (Association of Public Health Laboratories, 2016). Some trainees may also be given “blinded” samples that contain known strains to further verify the accuracy of the trainee’s skills. Please refer to the [Annexures](#) for guidance templates used in training laboratory personnel in nanopore sequencing.

6.4. Quality Control and Validation (Process Management)

NGS labs require established QC steps to inform decisions on whether to continue or halt a workflow for a given sample. It is important to know when to stop the testing of a sample as early as possible during the workflow process (e.g., due to poor quality/quantity of extracted DNA or poor quality of the library), as this saves time and costs, rather than completing the workflow and obtaining sub-optimal results (Hutchins et al., 2019). Laboratories implementing NGS methods therefore need to establish quality control checkpoints for both the sample preparation (wet lab) and data analysis/bioinformatics (dry lab) aspects of the workflow to ensure that only samples and/or sequence data that meet the laboratory-established minimum quality standards (NABL 112 and ISO 15189) can move forward in the workflow process.

In addition to QC measures, NGS laboratories also need to perform a method validation, where the performance characteristics of the NGS approach are validated and documented before routine application to clinical samples (Yu, 2014). These performance characteristics have been summarized in **Table 12**. It is however important to note that performance specifications are dependent on the NGS assay and consequently, only those that apply to the NGS assay need to be established. For example, qualitative assays that identify the presence or absence of a microbe do not need to establish the reference range or reportable range.

Table 12. Guidance for establishment of performance characteristics in NGS laboratories

Performance characteristic	Definition as applied to NGS	Guidance for implementation in a NGS lab
Accuracy	Closeness of agreement between the sequencing result and the reference sequence (or true value of the reference material)	Reference sequences can be derived from reference materials, i.e., samples with well characterized biologically reference organisms, synthetic DNA or datasets.
Precision	Degree to which repeated measurements give the same results, repeatedly (within-run precision) and reproducibly (between-run precision)	Precision can be measured based on: <ol style="list-style-type: none"> 1. Repeatability (within-run precision): degree to which the same result is achieved in sequencing the same sample multiple times under the same conditions (e.g., sequencing samples in replicates during the same run) 2. Reproducibility (between-run precision): degree to which the same result is obtained for a sample when sequencing is performed by different users or on different instruments (e.g., sequencing

		the same samples on different runs or at different sites/locations)
Analytic sensitivity	Likelihood that the NGS will detect a target (e.g., variant or targeted region), if present.	<p>The true positive rate is the best measurement for sensitivity and is calculated by dividing the number of true positives by the sum of true positives and false negatives: $TP / [TP + FN]$</p> <p>The limit of detection (LOD) is also associated with the analytic sensitivity and can be used to detect the presence of low-level variants or sequences.</p> <p>For NGS labs, the LOD can be defined as the minimum amount of input material that allows for a consistently positive identification of all replicates for a defined sequence target.</p> <p>Recommendation for microbial identification: using serial dilutions of a known pathogen in a clinically relevant matrix, (e.g., spiked viral particles in blood/plasma) to establish the minimum coverage needed to detect the pathogen</p>
Analytic specificity	Likelihood that the NGS will not detect a target (e.g., variant or targeted region), if not present	<p>The false positive rate is the best measurement for specificity and is calculated by dividing the number of true negatives by the sum of true negatives and false positives: $TN / [TN + FP]$</p> <p>Recommendation for microbial identification and variant detection: the false positive rate should be evaluated at various read depths.</p>
Reportable range	Region(s) of the sequenced genome for which the sequence of an acceptable quality can be derived by the laboratory	This parameter describes genomic regions (e.g., genes or targeted regions) that are sequenced and included in analysis or from which information is drawn for comparison
Reference range (mainly specific for human genomes)	Reportable sequence variants or targeted regions that the NGS method can detect and occurs in a reference population	This parameter describes the type of sequence variants (e.g., structural variants , insertions, deletions, etc.) that occur at a genomic position in a reference population

Note. Adapted from Gargis, A. S., Kalman, L., & Lubin, I. M. (2016). Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. *Journal of clinical microbiology*, 54(12), 2857–2865. <https://doi.org/10.1128/JCM.00949-16>
Copyright 2016, American Society for Microbiology

The validation process is divided into three phases: (i) a test development phase, where testing is performed multiple times (in iterative cycles) until all NGS assay conditions and bioinformatic pipeline settings are optimized with a final SOP for the entire workflow established; (ii) a validation phase, where the NGS performance specifications are established using a number of diverse sample types and conditions (e.g., different users) to demonstrate the platform’s ability to accurately identify the intended target (e.g., pathogen identification); and (iii) a quality management phase, where appropriate QC procedures are put in place to monitor routine test performance, independent assessments are performed periodically and changes to the NGS assay are re-validated (Gargis et al., 2016). Quality

assurance procedures may also include confirmatory testing with a gold standard to minimize errors or exclude the possibility of contamination. For example, given higher error rates of nanopore sequencing, sequenced targeted regions on these platforms can be validated by cross examination of results from Sanger or Illumina sequencing platforms, which have near-zero error rates.

Note. Due to the continuously changing sequencing chemistries, reagents, kits, software upgrades and other modifications of NGS platforms, ongoing validation will be necessary to re-establish performance specifications or demonstrate unchanged characteristics (Yu, 2014).

6.5. Quality Management System – Supporting Documents

In 2015, the Center for Disease Control and Prevention (CDC) formed a NGS Quality Workgroup to address the challenges of developing and implementing quality NGS methods within CDC laboratories (Hutchins et al., 2019). This led to the development of generalized solutions and practical guidance under the quality management system (QMS) framework to assist laboratory implementation and maintenance of quality practices for NGS workflows. These SOPs, procedures and other documents are provided as templates and can be easily adapted and revised by NGS laboratories for their individual needs. For the purpose of this training manual, only those templates specific to nanopore sequencing have been included in the [Annexures](#) of this document. Templates corresponding to other types of NGS are available on the CDC website (<https://www.cdc.gov/labquality/qms-tools-and-resources.html>).

INSACOG

In January 2021, the Indian SARS-CoV-2 Genomic Consortia (INSACOG) was joint initiative of the Union Health Ministry of Health and Department of Biotechnology (DBT) (under the Ministry of Science and Technology) with the Council for Scientific & Industrial Research (CSIR) and Indian Council of Medical Research (ICMR) to expand the whole-genome sequencing of SARS-CoV-2. This was aimed to assess the SARS COV 2 evolution in India and its correlation with clinical and epidemiological data and preparedness for public health

interventions. The structure of INSACOG is HUB and SPOKE model. HUB laboratories serve as mentors and are responsible for training, handholding and QC check for the RGLs / Satellite and regional laboratories. The Indian biological data centre (IBDC) is a one-point data collection and analysis centre. All the laboratories share raw data to IBDC and the clade and lineage information to Integrated Health Information Platform (IHIP) under Integrated Disease surveillance Program along with metadata to be used for public health interventions. Any additional laboratory once a member of INSACOG, will get the IHIP login for epi data and lineage information and IBDC login for submission of sequenced data (raw fastq files).

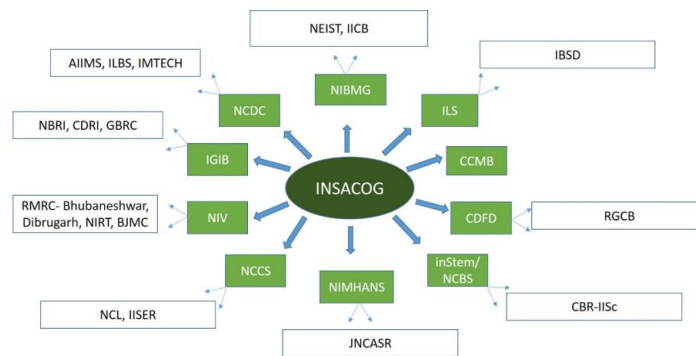


Figure 39. INSACOG Network

Note. Reprinted from <https://dbtindia.gov.in/sites/default/files/INSACOG.pdf>

Currently, the network has 54 laboratories. For information on how to join the INSACOG network as well as their established SOPs and guidelines, please see the [Annexures](#).

Recommended Reading

For more information on quality and validation practices in a NGS Lab, check out the links below:

- NGS Quality Initiative <https://www.cdc.gov/labquality/ngs-quality-initiative.html>
- QMS Tools and Resources <https://www.cdc.gov/labquality/qms-tools-and-resources.html>
- King, J., Harder, T., Beer, M. *et al.* Rapid multiplex MinION nanopore sequencing workflow for Influenza A viruses. *BMC Infect Dis* **20**, 648 (2020). <https://doi.org/10.1186/s12879-020-05367-y>

- Tyler, A.D., Mataseje, L., Urfano, C.J. et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci Rep* 8, 10931 (2018). <https://doi.org/10.1038/s41598-018-29334-5>
- WHO. Genomic Sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health <https://www.who.int/publications/i/item/9789240018440>
- WHO. Guidance for surveillance of SARS-CoV-2 variants: Interim guidance. https://www.who.int/publications/i/item/WHO_2019-nCoV_surveillance_variants
- National Center for Disease Control (NCDC). COVID19 Guidelines <https://ncdc.gov.in/index1.php?lang=1&level=1&sublinkid=703&lid=550>

6.6. References

1. Association of Public Health Laboratories. (2016). Next Generation Sequencing Implementation Guide. In.
2. Gargis, A. S., Kalman, L., & Lubin, I. M. (2016). Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. *J Clin Microbiol*, 54(12), 2857-2865. <https://doi.org/10.1128/JCM.00949-16>
3. Hutchins, R. J., Phan, K. L., Saboor, A., Miller, J. D., Muehlenbachs, A., & Workgroup, C. N. Q. (2019). Practical Guidance to Implementing Quality Management Systems in Public Health Laboratories Performing Next-Generation Sequencing: Personnel, Equipment, and Process Management (Phase 1). *J Clin Microbiol*, 57(8). <https://doi.org/10.1128/JCM.00261-19>
4. Yu, B. (2014). Setting up next-generation sequencing in the medical laboratory. *Methods Mol Biol*, 1168, 195-206. https://doi.org/10.1007/978-1-4939-0847-9_11

GLOSSARY OF TERMS

Genome - A genome refers to the complete set of genetic information (DNA or RNA in some viruses) in an organism

DNA - Deoxyribonucleic acid is a molecular structure consisting of two chains that carry the genetic instructions for an organism's functioning and development.

Base pair – Base pair (bp) refers to the number of nucleotides in one of the strands of DNA. Other higher units of measurements include the kilobase (kb or kbp), which equates to 10^3 bp; the Megabase (Mb or Mbp), which equates to 10^6 bp; or the Gigabase (Gb), which equates to 10^9 bp.

Gene - A gene is a sequence of DNA that encodes a function. In genetics, they are known as the basic unit of inheritance and are the traits or features that are passed from parents to offspring.

Sequence - A sequence refers to the specific order of nucleotide bases in a DNA segment or whole genome.

DNA Database - A DNA database is an organized collection of data (sequence information) that is stored and accessed electronically.

Mapping - Read mapping is the process of aligning reads to a known genome sequence (reference genome).

Euchromatin – Euchromatic regions refer to the less condensed, gene-rich regions of the chromosome, which can be transcribed.

Heterochromatin – Heterochromatic regions are highly condensed, gene-poor regions of the chromosome, which cannot be transcribed. Examples of heterochromatic sites are the centromeres and telomeres, which are made up of large, repetitive DNA elements.

Genetic code - The genetic code is defined as the three-letter combination of nucleotides (codons), which directs the translation of DNA into amino acids.

Coverage (or depth) – The sequencing coverage or depth (used interchangeably) refers to the number of times a base is sequenced allowing it to align or “map” to a reference genome. It is usually expressed with an “X”. For example, a coverage of 30X means that an average of 30 reads uniquely mapped to the reference genome.

Read - A read is defined as a sequenced piece or segment of DNA.

Structural variation - Structural variations (SV) are large (50 bp -1 kb) genomic alterations in an organism's chromosome. They are classified as either deletions, duplications, insertions, inversions, copy number variations (CNVs) or translocations. SVs are usually associated with diseases like cancer.

Multiplexing - This is a process of pooling DNA fragments from different samples and sequencing them at the same time in a single run. Sample multiplexing is cost-efficient way of sequencing targeted regions or smaller genomes

Demultiplexing - This is the process of sorting NGS reads from different samples that were sequenced together into their respective samples of origin using barcode information.

Homopolymer - A homopolymeric region in genomics refers to a stretch of repetitive regions in a genomic sequence e.g.TGTGAAACCCC

Assay - In diagnostics, an assay is a qualitative or quantitative test, used to determine the presence or the amount of a substance in a clinical sample

Probe - Probes or baits are oligonucleotides used to retrieve specific RNA or genomic fragments of interest for sequencing

Epigenetics - This is the study of heritable changes in gene expression that are caused by factors other than an individual's DNA sequence.

Lineage - A lineage refers to the relationship between ancestors and descendants and usually refers to a direct line of descent from an ancestor.

Clade - A clade consists of a group of organisms and their most recent common ancestor.

HPC – High performance computing uses parallel data processing to improve computing performance and perform complex calculations at high speeds. This may involve aggregating the computing power of supercomputer and computing clusters to deliver more performance than a typical standalone personal computer.

Core - A core is an individual processor that processes computational tasks within the central processing unit (CPU). A multi-core processor has two or more cores.

Open-source A software is considered open source if its source code is made freely available to the public, allowing anyone to view, modify and enhance it. Open-source software are developed in a decentralized way to promote open collaboration in the community.

Database - In genomics, a database is a organized collection of sequence data that is stored and set up for easy access electronically.

Gold Standard - In diagnostics, a gold standard refers to the best available diagnostic test or methodology, which newer tests can be benchmarked against.

Tagmentation - On-bead tagmentation library prep uses bead-linked transposomes for a more uniform tagmentation reaction compared to in-solution tagmentation reactions. Bead-linked transposome chemistry integrates DNA extraction, fragmentation, library preparation, and library normalization steps. This reduces the number of workflow steps, requiring low sample input and reducing both hands-on and turnaround time.

ANNEXURES

ANNEXURE 1- Generalized Do's and Don'ts of NGS Workflows

ANNEXURE 2- List of Consumables and Equipment for an NGS Lab

ANNEXURE 3- Equipment's for library preparation QC and amplification

ANNEXURE 4- MinION Employee Training SOP

ANNEXURE 5- MinION QC Workflows Guidance

ANNEXURE 6- MinION Rapid QC Workflows Guidance

ANNEXURE 7- Application Form for Joining INSACOG Network

ANNEXURE 8- INSACOG Data Submission Guidelines

ANNEXURE 9- INSACOG Updated SOPs and Guidelines

ANNEXURE I

Generalized Do's and Don'ts for NGS Workflows

Errors during NGS workflows are unavoidable. However, keeping the following measures in mind may help minimize their occurrence, improve sequencing efficiency, and maximize output.

Table 13. Do's and Don'ts for NGS Workflows

Step	Do...	Do not...
All workflow steps		
	...wear gloves at all times and avoid touching contaminated surfaces with gloved hands	...use the same gloves in different laboratory spaces or rooms
	...sterilise workbench and tools with ethanol and bleach before and after each experiment	...use old dilutions of ethanol and bleach for decontamination
	...return enzymes or polymerases to the freezer immediately after use to maintain stability	...use the same pipette tips between samples, rows or columns on a plate
	...mix and spin down reagents before opening	...store NGS reagents and samples in the same place
	... perform a pilot run with “mock” (known) samples to test protocols and entire workflow	...use unoptimized or unvalidated protocols and reagents
Sample preparation		
	...process input samples according to recommended instructions	...overload the purification system
	...place eluted nucleic acids on ice for immediate use or freeze for long term storage	...leave eluted DNA/RNA at room temperature
	...use sensitive/recommended quantification assays for isolated nucleic acids	...use low quality DNA/RNA
Library preparation		
	...minimize PCR cycling	...increase cycling over what is recommended
	...include internal controls and/or replicates	...process too many samples at once
	...multiplex samples to reduce costs while maintaining sufficient coverage, where possible	...waste reagents on abrupt and ill-planned sequencing experiments
Sequencing & Analysis		
	...monitor the sequencing run metrics	...overload flow cells
	...use reference sequences for comparisons	...assign biological relevance to low quality sequence data
	...compress data and keep a storage backup	...use external memory sticks on the main data storage network

ANNEXURE 2

Full list of Consumables and Equipment needed for the SARS-CoV-2 Sequencing

Equipment/Consumables required for processing <24 samples:

Room No 1: Master Mix Preparation

1. 8 well PCR Strips
2. 1.5- or 2.0-ml micro centrifuge tubes
3. 0.5 - 10 μ l or 2 - 20 μ l Pipette (single channel) & filter tips
4. 20 – 200 μ l Pipette (single channel) & filter tips
5. -20°C PCR cooler & racks (96 well & 1.5- or 2.0-ml boxes) (transferrable between rooms) *
6. Ice Bucket (transferrable between rooms) *

Room No 2: Library Preparation (*Template Addition Room if available can be utilized for this protocol*)

1. 8 well PCR Strips
2. 1.5- or 2.0-ml micro centrifuge tubes
3. 0.5 ml thin walled, clear PCR tubes
4. 15- or 50-ml centrifuge tubes (for aliquoting the qubit buffers)
5. 1.5- or 2.0 ml magnetic Stand
6. 1.5- or 2.0-ml tube spinner/minifuge/centrifuge
7. 8 well PCR strip spinner
8. Qubit Instrument
9. Thermal cycler (PCR machine) or RT-PCR instrument
10. 0.5 - 10 μ l or 2 - 20 μ l Pipette (single channel) & filter tips
11. 20 – 200 μ l Pipette (single channel) & filter tips
12. 100 – 1000 μ l Pipette (single channel) & filter tips
13. 96 well Plate adhesive seal or aluminum foil
14. Scalpel blade or scissor
15. Vortex mixer
16. Digital Timer
17. -20°C PCR cooler & racks (96 well & 1.5- or 2.0-ml boxes) (transferrable between rooms) *
18. Ice Bucket (transferrable between rooms) *

Room No 3: Sequencing Room

1. 1.5- or 2.0-ml micro centrifuge tubes
2. 20 – 200 μ l Pipette (single channel) & filter tips
3. 100 – 1000 μ l Pipette (single channel) & filter tips
4. 1.5- or 2.0-ml tube spinner/minifuge/centrifugation
5. Vortex mixer
6. Ice Bucket (transferrable between rooms) *

Equipment/Consumables required for processing >24 samples:

Room No 1: Master Mix Preparation

1. 8 well PCR strips
2. 96 well PCR Plate

3. 1.5- or 2.0-ml micro centrifuge tubes
4. 20 – 200 µl Pipette (single channel) & filter tips
5. 0.5 – 10 µl Pipette (multi-channel) & filter tips
6. 8 well PCR strip spinner
7. Ice Bucket (transferrable between rooms) *
8. -20°C PCR cooler & racks (96 well & 1.5- or 2.0-ml boxes) (transferrable between rooms) *

Room No 2: Library Preparation (Template Addition Room if available can be utilized for this protocol)

1. 1.5- or 2.0-ml micro centrifuge tubes
2. 15- or 50-ml Centrifuge tubes
3. 0.5 ml thin walled, clear PCR tubes
4. 0.5 - 10 µl or 2 - 20 µl Pipette (single channel) & filter tips
5. 0.5 – 10 µl pipette (multi-channel) & filter tips
6. 10 – 100 µl pipette (multi-channel) & filter tips
7. 20 – 200 µl Pipette (single channel) & Filter tips
8. 100 – 1000 µl Pipette (single channel) & filter tips
9. 1.5- or 2.0-ml Magnetic Stand
10. 1.5- or 2.0-ml tube spinner/minifuge/centrifugation
11. Qubit Instrument
12. Thermal cycler (PCR machine) or RT-PCR instrument
13. 96 well plate spinner or plate centrifuge
14. Vortex mixer
15. -20°C PCR cooler & racks (96 well & 1.5- or 2.0-ml boxes) (transferrable between rooms) *
16. 96 well Plate adhesive seal or aluminum foil
17. scalpel blade or scissor
18. Ice Bucket (transferrable between rooms) *
19. Digital Timer or stopwatch

Room No 3: Sequencing Room

1. 1.5- or 2.0-ml micro centrifuge tubes
2. 20 – 200 µl Pipette (single channel) & filter tips
3. 100 – 1000 µl Pipette (single channel) & filter tips
4. 1.5- or 2.0-ml tube spinner/minifuge/centrifuge
5. Vortex mixer
6. Ice Bucket (transferrable between rooms) *

Reagents required

Room No 1: Master Mix Preparation

1. LunaScript RT SuperMix Reagent
2. Nuclease free water
3. SARS CoV2 Primer Vials A & B
4. Q5 Hot Start HF 2X Master Mix

Room No 2: Library Preparation (Template Addition Room if available can be utilized for this protocol)

1. Barcode Plate
2. Nuclease free water
3. Ampure/SPRI Beads
4. 80% Ethanol (freshly prepared)
5. Elution Buffer
6. Rapid Adapter F vial
7. Qubit dsDNA HS reagents

Room No 3: Sequencing Room

1. Sequencing buffer II
2. Loading Beads II or loading Solution
3. Flush Buffer
4. Flush Tether
5. Flongle or Flow cell

* **Note.** As previously discussed under [section 6.1](#), equipment should ideally not be transferred between rooms. However, in the case of cooler racks, these can be sterilized and cleaned for contamination before using in different rooms. In addition, a single ice bucket can be taken into different rooms, as long as care is taken to ensure the ice bucket is not placed on any surface, the ice from the bucket is immediately transferred to a container present in the room and the ice bucket is quickly taken out of the room afterwards.

ANNEXURE 3

Equipment's for library preparation QC and amplification

Equipment for quantifying nucleic acid

Each step of the NGS process, starting from DNA isolation to library preparation, requires an accurate measurement of the amount of input DNA. Several options are available for both accurate and sensitive quantification of DNA. Among these, the most used instruments include:

- **Fluorometer:** These use fluorescent dyes that selectively bind to DNA, RNA, or protein, thereby enabling fast and sensitive quantification, even at low concentrations. As the dyes only bind to target molecules, fluorometers can provide accurate measurements even in the presence of contaminants. The Qubit fluorometer is one of the most widely used fluorometric instruments in NGS applications.
- **Spectrophotometer:** Spectrophotometric instruments determine the concentration of compounds by using the UV-absorbance method, which measures the natural absorbance of light at 260 nm (for DNA and RNA) or 280 nm (for proteins). However, unlike fluorometers, the lack of specificity at 260 nm means that a measurement at this wavelength could correspond to multiple types of nucleic acids (double stranded DNA [dsDNA], single stranded DNA [ssDNA] and RNA), which can confound NGS analysis. On the other hand, compared to fluorometers, spectrophotometric methods can also detect the presence of a wide variety of sample contaminants based on their UV-absorption patterns. The Nanodrop and QIAxpert are frequently used spectrophotometers in biological applications.
- **Real-time PCR:** The real-time PCR or quantitative PCR (qPCR) is a PCR-based technique that combines the amplification of a target DNA and the quantification of the DNA concentration in a single reaction. Detection of the amplified target DNA sequence (PCR product) is achieved using fluorescent-based chemistries, which correlate the PCR product concentration to the fluorescence intensity.

Equipment for measuring nucleic acid quality

Sequencing protocols require both high-quality and sufficient amounts of DNA to ensure successful sequencing runs. Assessment of the quality and size of the input DNA is an important and necessary quality control step that determines whether NGS workflows can commence to the next step. Some instruments that support these measurements are:

- **Automated microfluidic capillary electrophoresis stations:** These devices replace the traditional, gel electrophoresis and deliver high-resolution quality measurements (i.e., size, concentration, and integrity) of nucleic acids and proteins using integrated, streamlined workflows. Stations can process up to 96 samples at the same cost in a relatively short time, thus providing scalable throughput. Examples include the Agilent Bioanalyzer and TapeStation systems, and the QIAxcel.

Thermocyclers

Thermocyclers are essential for amplifying DNA prior to and/or during library preparation. However, it may be necessary to optimize protocols when using certain thermocyclers or purchase new thermocyclers as recommended by the NGS protocol.

ANNEXURE 4

MinION Employee Training SOP

CDC		APHL ASSOCIATION OF PUBLIC HEALTH LABORATORIES	
Insert Laboratory Specific Name Here			
MinION Employee Training SOP			
1.0 Purpose			
This procedure outlines the steps for training personnel to acquire the skills and knowledge necessary to run the Oxford Nanopore MinION sequencer from initial sample quality control to the review of sequencing run quality metrics.			
2.0 Scope			
This document applies to all staff that operate the Oxford Nanopore MinION next generation sequencer and supervisors that oversee these operations. Training on the Oxford Nanopore MinION sequencer is a process that includes building a base of sequencing knowledge, observing the trainer perform the sequencing procedures, performing sequencing procedures under direct trainer supervision, and individually executing the sequencing procedures.			
3.0 Related Documents			
Title	Document Control Number		
MinION Employee Training Form			
MinION Trainer Designation Form			
<i>"Lab-developed Risk Assessment/Mitigation Steps"</i>			
4.0 Responsibilities			
Position	Responsibility		
All laboratory staff	<ul style="list-style-type: none">• Complete all necessary training requirements		
Team Lead	<ul style="list-style-type: none">• Determine the training needs for the laboratory team• Ensure all staff are trained and evaluated according to this procedure• Designate the trainer by completing the MinION Trainer Designation Form• Create training plans, review training materials, and assign trainers as needed		
Trainers	<ul style="list-style-type: none">• Develop training materials• Train staff as directed by the Team Lead• Document training activities		
Branch Chief	<ul style="list-style-type: none">• Ensure applicable laboratory staff are accountable for completing all training and evaluation requirements described in this procedure• Review and approve this procedure		
Quality Manager	<ul style="list-style-type: none">• Review training documentation		
5.0 Training Information Resources			
Document #:	Revision #:	Effective Date:	Page 1 of 4

- 5.1 *Reference your laboratory SOP or the MinION SOP your laboratory uses here.*
- 5.2 *Reference your laboratory-developed risk assessment/mitigation document here; this may be specific to the MinION or to the specific nucleic acid source.*
- 5.3 Biosafety in Microbiological and Biomedical Laboratories, 5th Edition, HHS Publication Number (CDC) 21-1112
- 5.4 MinION Support Training Videos (select the videos relevant to your lab processes; add other videos as appropriate)
 - a. [MinION: A Portable, Real-Time DNA/RNA Sequencing Device](#)
 - b. [Flongle: For Rapid Nanopore Sequencing of Smaller Samples](#)
 - c. [Loading an Oxford Nanopore Flow Cell](#)
 - d. [RNA Sequencing with Nanopore Technology](#)
- 5.5 Required Reading (*select documents relevant to your lab processes; add other documents as appropriate*)
 - a. MinION Flow Cell Check Protocol
 - b. MinION Rapid Sequencing Protocol
 - c. MinION 1D Genomic DNA by Ligation Protocol
 - d. Oxford Nanopore Community Discussion Board

6.0 Equipment/Materials

- 6.1 Oxford Nanopore MinION Sequencer
- 6.2 Library preparation and sequencing reagents

7.0 Safety Precautions

- 7.1 All BSL-2 practices, safety equipment, and facility design must comply with the requirements listed in the most current version of Biosafety in Microbiology and Biomedical Laboratories.
- 7.2 Appropriate PPE must be worn at all times when working in the laboratory, including laboratory coat, gloves, and safety glasses (if splashes are anticipated)

8.0 Procedure

- 8.1 The trainee will build a basic understanding of MinION next generation sequencing (NGS) technology by:
 - a. Reviewing the MinION support training videos (Section 5.4)
 - b. Complete the required readings (Section 5.5)
- 8.2 The trainer will perform all steps within the sequencing SOP in the laboratory while the trainee observes.
 - a. The trainer will verbally walk the trainee through the entire sequencing process from the beginning to end using the operational protocol as a training guide (Section 5.5)
 - b. This 1:1 review will cover initial sample quality control, preparing sample libraries, preparing the sequencing instrument, running the sequencing instrument, clean-up, and review of sequencing run quality control metrics.

Document #:	Revision #:	Effective Date:	Page 2 of 4
-------------	-------------	-----------------	--------------------

- 8.3** The trainee will perform all steps within the sequencing SOP under direct and full observation of the trainer.
- a. The trainer will quiz the trainee on multiple aspects of the protocol, including the questions below *(the laboratory should populate this section with questions relevant to their procedure)*:
 - i. Describe how the MinION identifies signals from the library fragment during sequencing.
 - ii. What can be done if the library does not absorb by capillary action into the SpotON priming port?
 - iii. Why do you open the flow cell priming port and then add 200 µl?
 - iv. Why is it important to have greater than 800 active pores?
 - v. What could cause a large number of unavailable/inactive pores after loading the flow cell?
 - b. The trainer will review the trainee's quality control data as described in the sequencing protocol to assess the competency of the trainee.
- 8.4** Once the trainee successfully performs a sequencing run under the observation of the trainer, the trainee will perform an unaccompanied sequencing run.
- a. The trainer will review the trainee's quality control data and run data to assess the competency of the trainee.
- 8.5** Is it the responsibility of the primary user to ensure that preventative maintenance is scheduled and executed.
- a. The trainee will observe proper user performed preventive maintenance.
 - b. The trainee will perform user performed preventive maintenance.
 - c. The trainer will assess the trainee's ability to properly maintain the instrument according to established maintenance procedures.

9.0 Continued Learning

- 9.1 Trainers and primary users should regularly attend Oxford Nanopore MinION webinars, read primary literature, and review new product releases.
- 9.2 It is expected that trainers will try new protocols in the laboratory and teach new skills to primary users on a semiannual basis.

10.0 References

11.0 Appendices

12.0 Revision History

Rev #	DCR #	Change Summary	Date

13.0 Approval

Approved By: _____ Date: _____
 Author

 Print Name and Title

Document #:	Revision #:	Effective Date:	Page 3 of 4
-------------	-------------	-----------------	--------------------



Approved By: _____ Date: _____
Technical Reviewer

Print Name and Title

Approved By: _____ Date: _____
Quality Manager / Designee

Print Name and Title

Document #:	Revision #:	Effective Date:	Page 4 of 4
-------------	-------------	-----------------	-------------



Annexure 5

MinION QC Workflows Guidance

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION 1D QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 1 of 6
-------------	-------------	-----------------	-------------

1.0 Purpose

- 1.1 This document provides quality control (QC) guidance for nucleic acid sequencing using the Oxford Nanopore MinION technology. The guidance takes into account specific QC checkpoints between laboratory processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that are informative for downstream bioinformatics processes.
- 1.2 The quality of nucleic acid extraction and manipulation, fractionations and size selection, and library preparations affects fragment size uniformity and library diversity, which is important for achieving complete and even coverage of the total nucleic acid to be sequenced. Gaps resulting from poor sample preparation cannot be corrected downstream by error correction methods employed by some sequencing technologies. In addition, quality scores do not reflect errors introduced during sample preparation, as the sequencing signal will appear clean and error-free. The maximal achievable accuracy of most sequencing platforms is limited by the sample accuracy.

2.0 NGS QC Checkpoints

The following sections correspond to the process steps prior to sequencing, as outlined in Figure 1.

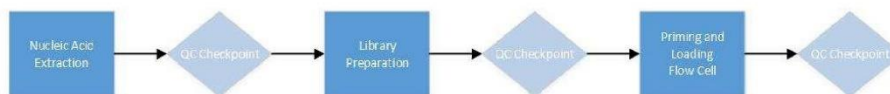


Figure 1: NGS QC Checkpoints for MinION 1D Workflows

2.1 Nucleic Acid Extraction

- a. High quality nucleic acid purification is essential for obtaining accurate NGS data. The extraction method depends greatly on the sample type and matrices involved. See Appendix A for extraction methods recommended by Nanopore.
Note: Proteinase K has been known to cause pore degradation. It is recommended to use an extraction method that does not use proteinase K.

2.2 Post Extraction Nucleic Acid QC Checkpoint

It is important to check input DNA for quality before beginning library preparation. Low molecular weight, incorrectly qualified and/or contaminated DNA (e.g. salt, EDTA, protein, organic solvents) can have a significant impact on downstream processes and ultimately, your sequencing run.

a. Criteria for Input DNA

- i. Purity as measured using Nanodrop – OD 260/280 of ~1.8 and OD 260/230 of 2.0-2.2. A 260/280 which is higher than ~1.8 indicates the presence of RNA. A 260/280 which is lower than ~1.8 can indicate the presence of protein or phenol. Establish the precise acceptable 260/280 range for your test during development and validation.
- ii. Average fragment size >30kb. Fragment size may be measured using several methods (e.g., pulse-field, low percentage agarose gel analysis, blue pippin). This quality checkpoint

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION 1D QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 2 of 6
-------------	-------------	-----------------	-------------

is important during the development and validation of the test. Labs may elect to omit this quality check after validation if the test has proven robust and stable.

- iii. Input mass, as measured by Qubit – 1 µg or 1.5 µg if carrying out a DNA repair step. In order to maximize sequencing yield, it is important that the nanopores are kept filled with DNA to minimize the time they are idle between strands. For further optimization of fragment length to improve throughput, see table 1 in section 2.6.
- b. Use the configuration test cell to confirm the MinION is communicating with the computer.
 - i. The configuration protocol has been successfully completed when the message “Customer configuration run has completed” is displayed in the notifications panel. If configuration reports that it has failed, reinsert the flow cell and trouble shoot per manufacturer’s instructions. Upon successful configuration, the MinION and MinkNOW systems are ready for platform QC of the flow cell (see Section 2.7).

2.3 DNA Fragmentation (optional): DNA fragmentation is an optional step for when experiments require specific fragment sizes.

2.4 Fragmentation QC Checkpoint

- a. Determine the fragment size, quantity, and quality using the Agilent Bioanalyzer or similar instrument. Confirm the fragment size is within the expected range. If the results yield smaller fragments, this is indicative of substantial shearing/degradation of the input material and is likely to reduce the quality of the library preparation and the read length distribution

2.5 Library Preparation: Perform library preparation according to the selected protocol. It is recommended that the repaired/end-prepped DNA sample is subjected to clean-up with AMPure XP beads. This clean-up can be omitted for simplicity and to reduce library preparation time. However, it has been observed that omission of this clean-up can: reduce subsequent adapter ligation efficiency, increase the prevalence of chimeric reads, and lead to an increase in pores being unavailable for sequencing.

2.6 Library Preparation QC Checkpoint

- a. In order to maximize sequencing yield, it is important that the nanopores are kept filled with DNA to minimize the time they are idle between strands. The less material goes into the flow cell, the fewer “threadable ends” will be present to be captured by the pores. Therefore, the pores will be searching for molecules for longer, and if the pores are not always sequencing, throughput could be compromised.
- b. Note: During development and optimization of a method it is advisable to check the fragment size and final DNA input concentration of the library before proceeding to priming and loading the library. The below table may be used to inform optimization experiments.

Mass of extracted nucleic acid	No. of moles if library fragment length = 2kb	No. of moles if library fragment length = 8kb	No. of moles if library fragment length = 50 kb
10 µg	7.7 pmol	1.9 pmol	308 fmol
5 µg	3.9 pmol	963 fmol	154 fmol
3.5 µg	2.7 pmol	674 fmol	108 fmol
2 µg	1.5 pmol	385 fmol	62 fmol

2

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION 1D QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 3 of 6
-------------	-------------	-----------------	-------------

Mass of extracted nucleic acid	No. of moles if library fragment length = 2kb	No. of moles if library fragment length = 8kb	No. of moles if library fragment length = 50 kb
1.5 µg	1.2 pmol	289 fmol	46 fmol
1 µg	770 fmol	193 fmol	31 fmol
500 ng	385 fmol	96 fmol	15 fmol
400 ng	308 fmol	77 fmol	12 fmol
200 ng	154 fmol	39 fmol	6.2 fmol
100 ng	77 fmol	19 fmol	3.1 fmol
30 ng	23 fmol	5.8 fmol	0.9 fmol
10 ng	7.7 fmol	1.9 fmol	0.3 fmol
10 pg	0.0077 fmol	0.009 fmol	0.0003 fmol

Table 1: Fragment Length

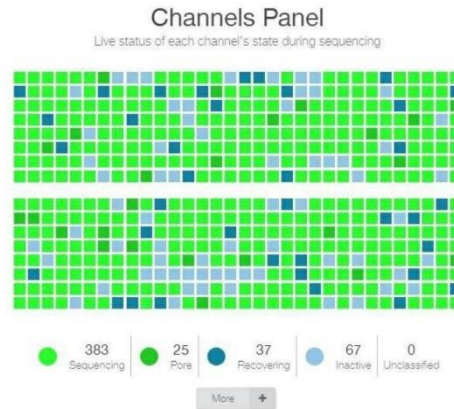
- i. In order to keep the pores full, the current R9.4.1 and R9.5.1 pores require about **5-50 fmol of good quality library put into the flow cell**.
 - ii. Quantify 1 µL of adapter ligated DNA using a Qubit fluorometer. Expected recovery is 430 ng.
- 2.7 Pre-loading QC Checkpoint:** As the MinKNOW script progresses, check the following:
- a. Number of active pores should be 800 or greater
 - b. Heatsink temperatures: (34°C)
- 2.8 Post Loading QC Checkpoint:** The library is loaded dropwise. Ensure each drop flows into the port before adding the next. Be sure to pipette slowly when adding priming mix to priming port to ensure the membrane stays intact. For further details on loading the Oxford Nanopore MinION flow cell click [here](#).
- 2.9 Post Loading QC Checkpoint**
- a. Number of active pores should be above 800
 - b. Development of the read histogram: Confirm the histogram reflects expected read lengths for the experimental design being used.
 - c. Pore occupancy: Monitor the pore occupancy for the first 30 minutes of a sequencing experiment. If you are not observing the expected percentage of pores in stand sequencing, stop the run, wash the flow cell and store it for use in another run. A good library will be indicated by a higher proportion of light green channels in Sequencing state (neon green) than are in Pore state (green). The combination of Sequencing and Pore channels indicate the number of active pores at any point in time. A low proportion of sequencing channels will reduce the throughput of the run.
 - i. **Recovering** (dark blue) indicates channels that may become available for sequencing again. A high proportion of this may indicate additional clean up steps are required during your library preparation.
 - ii. **Inactive** (light blue) indicates channels that are no longer available for sequencing. A high proportion of these as soon as the run begins may indicate an osmotic imbalance.

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION 1D QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 4 of 6
-------------	-------------	-----------------	-------------

- iii. **Unclassified** are channels that have not yet been assigning one of the above classifications.



- d. **Good quality library:** A good quality library will result in most of the pores being in the "Sequencing" state (neon green), and very few in "Pore" (green), "Recovering" (dark blue), or "Inactive" (light blue). A library that results in a Duty Time graph like the example below is likely to give a good sequencing throughput. The graph populates over time, and can be used as a way to assess the quality of your sequencing experiment, and make an early decision whether to continue with the experiment or to stop the run.



The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION 1D QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 5 of 6
-------------	-------------	-----------------	-------------

- e. Base Calling Report: Confirm the local basecalling is being recorded in the base calling report and is within expected range (*insert laboratory specific range here*).

3.0 Appendices

3.1 Appendix A – NGS MinION Extraction Methods

4.0 Revision History

Rev #	DCR #	Change Summary	Date

5.0 Approval

Approved By: _____ Date: _____
 Author

 Print Name and Title

Approved By: _____ Date: _____
 Supervisor

 Print Name and Title

Approved By: _____ Date: _____
 Quality Manager

 Print Name:

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION 1D QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 6 of 6
-------------	-------------	-----------------	-------------

Appendix A – Extraction Methods

Extraction Method	
Gram-negative bacterial DNA	Click here for protocol
Gram-positive bacterial DNA	Click here for protocol
Yeast DNA	Click here for protocol
Yeast RNA	Click here for protocol

Annexure 6

MinION Rapid QC Workflows Guidance

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION Rapid QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 1 of 5
-------------	-------------	-----------------	-------------

1.0 Purpose

- 1.1 This document provides quality control (QC) guidance for nucleic acid sequencing using the Oxford Nanopore MinION Rapid Sequencing technology. The guidance takes into account specific QC checkpoints between laboratory processes to ensure each step is completed correctly, with high confidence, and to generate quality data metrics that are informative for downstream bioinformatics processes.
- 1.2 The quality of nucleic acid extraction and manipulation, fractionations and size selection, and library preparations affects fragment size uniformity and library diversity, which is important for achieving complete and even coverage of the total nucleic acid to be sequenced. Gaps resulting from poor sample preparation cannot be corrected downstream by error correction methods employed by some sequencing technologies. In addition, quality scores do not reflect errors introduced during sample preparation, as the sequencing signal will appear clean and error-free. The maximal achievable accuracy of most sequencing platforms is limited by the sample accuracy.

2.0 NGS QC Checkpoints

The following sections correspond to the process steps prior to sequencing, as outlined in Figure 1.

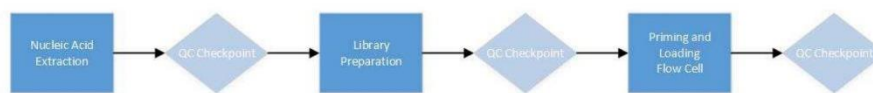


Figure 1: NGS QC Checkpoints for MinION Rapid Sequencing Workflows

2.1 Nucleic Acid Extraction

High quality nucleic acid purification is essential for obtaining accurate NGS data. The extraction method depends greatly on the sample type and matrices involved. See Appendix A for extraction methods recommended by Nanopore.

Note: Proteinase K has been known to cause pore degradation. It is recommended to use an extraction method that does not use proteinase K.

2.2 Post Extraction Nucleic Acid QC Checkpoint

It is important to check input DNA for quality before beginning library preparation. Low molecular weight, incorrectly qualified and/or contaminated DNA (e.g. salt, EDTA, protein, organic solvents) can have a significant impact on downstream processes and ultimately, your sequencing run.

a. Criteria for Input DNA

- i. Purity as measured using Nanodrop – OD 260/280 of ~1.8 and OD 260/230 of 2.0-2.2. A 260/280 which is higher than ~1.8 indicates the presence of RNA. A 260/280 which is lower than ~1.8 can indicate the presence of protein or phenol. Establish the precise acceptable 260/280 range for your test during development and validation.
- ii. Average fragment size >30kb. Fragment size may be measured using several methods (e.g., pulse-field, low percentage agarose gel analysis, blue pippin). This quality checkpoint

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION Rapid QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 2 of 5
-------------	-------------	-----------------	-------------

is important during the development and validation of the test. Labs may elect to omit this quality check after validation if the test has proven robust and stable.

- iii. Input mass, as measured by Qubit - ~400 ng. In order to maximize sequencing yield, it is important that the nanopores are kept filled with DNA to minimize the time they are idle between strands. For further optimization of fragment length to improve throughput, see table 1 in section 2.3.
- b. Use the configuration test cell to confirm the MinION is communicating with the computer.
 - i. The configuration protocol has been successfully completed when the message "Customer configuration run has completed" is displayed in the notifications panel. If configuration reports that it has failed, reinsert the flow cell and trouble shoot per manufacturer's instructions. Upon successful configuration, the MinION and MinKNOW systems are ready for platform QC of the flow cell (see section 2.4).

2.3 Library Preparation

There are many library preparation kits available specific to the kind of sequencing and downstream application. The library preparation does not have a separate checkpoint during routine testing. Proceed to Pre-loading QC checkpoint.

2.4 Library Preparation QC Checkpoint

Note: During development and optimization of a method it is advisable to check the fragment size and final DNA input concentration of the library before proceeding to priming and loading the library. The below table may be used to inform optimization experiments.

Mass of extracted nucleic acid	No. of moles if library fragment length = 2kb	No. of moles if library fragment length = 8kb	No. of moles if library fragment length = 50 kb
10 µg	7.7 pmol	1.9 pmol	308 fmol
5 µg	3.9 pmol	963 fmol	154 fmol
3.5 µg	2.7 pmol	674 fmol	108 fmol
2 µg	1.5 pmol	385 fmol	62 fmol
1.5 µg	1.2 pmol	289 fmol	46 fmol
1 µg	770 fmol	193 fmol	31 fmol
500 ng	385 fmol	96 fmol	15 fmol
400 ng	308 fmol	77 fmol	12 fmol
200 ng	154 fmol	39 fmol	6.2 fmol
100 ng	77 fmol	19 fmol	3.1 fmol
30 ng	23 fmol	5.8 fmol	0.9 fmol
10 ng	7.7 fmol	1.9 fmol	0.3 fmol
10 pg	0.0077 fmol	0.009 fmol	0.0003 fmol

Table 1: Fragment Length

2.5 Pre-loading QC Checkpoint: As the MinKNOW script progresses, check the following:

- a. Number of active pores should be 800 or greater
- b. Heatsink temperatures: (34°C)

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

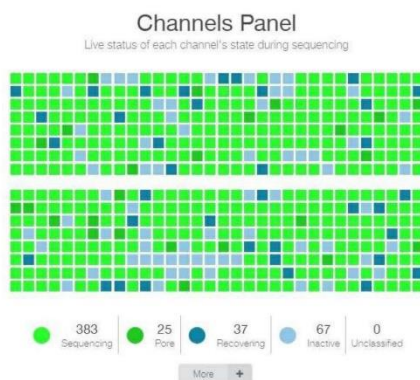
NGS MinION Rapid QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 3 of 5
-------------	-------------	-----------------	-------------

2.6 Priming and Loading: Add priming mix very slowly to ensure the membrane and protein pores are not damaged. The library is loaded dropwise. Ensure each drop flows into the port before adding the next. For further details on loading the Oxford Nanopore MinION flow cell click [here](#).

2.7 Post Loading QC Checkpoint

- a. Number of active pores should be above 800
- b. Development of the read histogram: Confirm the histogram reflects expected read lengths for the experimental design being used.
- c. Pore occupancy: Monitor the pore occupancy for the first 30 minutes of a sequencing experiment. If you are not observing the expected percentage of pores in strand sequencing, stop the run, wash the flow cell and store it for use in another run. A good library will be indicated by a higher proportion of light green channels in Sequencing state (neon green) than are in Pore state (green). The combination of Sequencing and Pore channels indicate the number of active pores at any point in time. A low proportion of sequencing channels will reduce the throughput of the run.
 - i. **Recovering** (dark blue) indicates channels that may become available for sequencing again. A high proportion of this may indicate additional clean up steps are required during your library preparation.
 - ii. **Inactive** (light blue) indicates channels that are no longer available for sequencing. A high proportion of these as soon as the run begins may indicate an osmotic imbalance.
 - iii. **Unclassified** are channels that have not yet been assigning one of the above classifications.

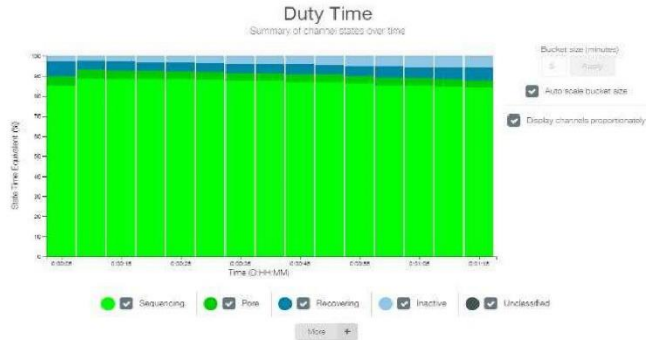


- d. Good quality library: A good quality library will result in most of the pores being in the "Sequencing" state (neon green), and very few in "Pore" (green), "Recovering" (dark blue), or "Inactive" (light blue). A library that results a Duty Time graph like the example below is likely to give a good sequencing throughput. The graph populates over time, and can be used as a way to assess the quality of your sequencing experiment, and make an early decision whether to continue with the experiment or to stop the run.

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION Rapid QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 4 of 5
-------------	-------------	-----------------	-------------



- e. Base Calling Report: Confirm the local basecalling is being recorded in the base calling report and is within expected range (*insert laboratory specific range here*).

3.0 Appendices

Appendix A- NGS MinION Extraction Methods

4.0 Revision History

Rev #	DCR #	Change Summary	Date

5.0 Approval

Approved By: _____ Date: _____
Author

Print Name and Title

Approved By: _____ Date: _____
Supervisor

Print Name and Title

Approved By: _____ Date: _____
Quality Manager

Print Name

The NGS Quality Workgroup developed these documents and tools for use by next-generation sequencing laboratories. These documents and tools were developed based upon best available information, reviewed, edited, and approved by the participants in the group listed above. Prior to implementing these processes in your lab, review the date the document was finalized (included in the file name) and take any necessary actions to ensure the information remains applicable. These documents and tools are not controlled files; you are encouraged to modify the format (e.g. header/footer, sections) as needed to meet the document control requirements of the quality management system within your laboratory.

NGS MinION Rapid QC Workflows Guidance

Document #:	Revision #:	Effective Date:	Page 5 of 5
-------------	-------------	-----------------	-------------

Appendix A – Extraction Methods

Extraction Method	
Gram-negative bacterial DNA	Click here for protocol
Gram-positive bacterial DNA	Click here for protocol
Yeast DNA	Click here for protocol
Yeast RNA	Click here for protocol

Annexure 7

Application Form for Joining INSACOG Network

SOP for Joining INSACOG Network

The Indian SARS-CoV-2 Genomic Consortia (INSACOG) established for genomic surveillance in the country was setup in Jan 2021 with initial participation of 10 National research laboratories belonging to the Department of Biotechnology, Indian Council of Medical Research, Council of Scientific and Industrial Research and Ministry of Health and Family Welfare, Govt. of India. At a later stage 33 more research laboratories have been included in the network for enhancing sequencing effort in the country. These 43 laboratories operate on a **hub and spoke model** where the 10 IGSLs (INSACOG Genome Sequencing Laboratories) provide handholding for the new laboratories and act as hubs. Subsequently some more sequencing laboratories were included as a part of the sequencing network. There is a felt need for increasing sequencing efforts further for identification of variants of concern (VoC) and variants of Interest (VoI) in the Indian samples so as to advice the policy makers for effective management of COVID-19.

Based on several rounds of discussion in the INSACOG-SCAG it was felt necessary to clearly define roles and responsibility of the sequencing laboratories involved in genomic surveillance efforts. This was essential as the identified/ involved sequencing laboratories have varied levels of infrastructure, sequencing platforms, bioinformatics pipelines and human resources for carrying out sequencing in a timely manner.

The SCAG, based on several rounds of discussion recommends placing the Genome Sequencing Laboratories in three groups:

A. Group1: INSACOG Genome Sequencing Laboratories (Hub Laboratories):

- Since the inception of sequencing surveillance program as well as INSACOG network, 10 National Research Laboratories of the Department of Biotechnology (5 laboratories), Council of Scientific and Industrial Research (2 laboratories), ICMR (1 Laboratory), MoH&FW (2 Laboratories) were identified as INSACOG Genome Sequencing Laboratories (IGSLs).
- These laboratories have significant experience and expertise in sequencing SARS-COV-2 genomes, have adequate and appropriate sequencing platforms, expertise in bioinformatic analysis and interpretations.
- These laboratories receive samples from sentinel sites designated by NCDC and those operating through state surveillance network. Subsequently, as some of these laboratories are also testing facilities, sequencing of retrospective samples was also carried out.
- These laboratories, therefore constitute the hub of the INSACOG network, and are provided with login credentials for submitting analysed sequencing data on the IHIP-Integrated Health Information Platform portal developed and managed by NCDC. The IGSLs are also required to submit the FASTA & FASTAQ files of the analysed sequences to centralised repositories of the two of the IGSLs (NIBMG, Kalyani and IGIB, New Delhi).
- These laboratories have also developed facilities for storage and retrieval of samples, as and when needed. The financial support for carrying out sequencing surveillance are met from the support received from the respective agencies through EMR project or through institutional support.

B. Group 2: Laboratories nominated by Central Govt. Ministries/ Department as well as State Government nominated Laboratories:

- With the surge in COVID-19 positivity rate, it was felt necessary to include some more laboratories in the genome surveillance projects.
- INSACOG network was then expanded to include 33 more sequencing laboratories belonging to Central Govt/ State Govt./ Not for Profit Organisation.
- These laboratories were identified based on the availability of sequencing platforms, previous engagement with COVID -19 activities/ sequencing of COVID-19 samples, competence in bioinformatic analysis.
- These laboratories were assigned to one of the HUB laboratories for mentoring and hand holding. The analysed sequence information from these laboratories is generally shared with the HUB laboratories, which in turn, submit the sequence information to IHIP portal as well as the IGIB and NIBMG hub.
- In a few cases, taking into account the number of sequences of required quality generated, some of these laboratories were provided direct IHIP access, based on the recommendation of the HUB labs.
- The expenses for carrying out sequencing work for these laboratories are met from the respective agencies/ departments/ state governments.

Guidelines and procedure for the involvement of State Govt. nominated laboratories will be as follows:

- The interested State Government Nominated Laboratories will be required to provide the following information to INSACOG (as per the specific format enclosed).
- Once this information is received, the technical committee constituted by the INSACOG-SCAG will undertake assessment by interacting with the concerned laboratory. The technical committee will also interact and assign a hub laboratory (IGSL) for the nominated laboratory.
- The INSACOG SCAG, based on the assessment report will recommend to DBT the selected sequencing laboratories as well as the IGSLs to which they will be attached.
- The laboratories should be able to sequence at least 50 samples per week.
- In case the laboratories don't have necessary infrastructure, they can resubmit Expression of Interest to the INSACOG as and when their facilities are in place.
- The state government laboratories may obtain the samples from the sentinel sites of the respective states for which NCDC could provide user IDs and Password for accessing the sample details, on recommendation of the hub IGSLs.
- In case the state government laboratories don't have BSL-2 facilities then respective hub IGSLs can provide isolated RNA for sequencing.
- These laboratories should provide the sequence information to the IGSLs for sequence analysis and further submission to the national portal. This is essential to avoid any misinterpretation of the data generated. But this could be revisited if IGSLs to which the state government laboratories are associated feels that uploading of the data can independently be done by the respective laboratories.

C. Group 3: Laboratories belonging to Private Sector Commercial Labs and other entities:

- In order to enhance the scale of sequencing and for a sustainable effort, the inclusion of interested private-sector laboratories and companies is being considered.
- Due consideration of all important issues including Ethical, Economical, Data quality and security IPR etc. (as mentioned in the guidelines) should be taken into consideration by the INSACOG SCAG while deciding on inclusion of private sector laboratories for sequencing efforts. Sequencing result should not be disclosed in public or media or to any third party.
- These laboratories should be required to be aligned with one of the HUB laboratories for IHP entries, quality assurance and data interpretation.
- These laboratories will be required to clearly indicate source of funding for undertaking sequencing activities.
- These laboratories should specify the turnaround time of sequencing depending on minimum and maximum number of samples they could sequence.

Guidelines and procedure for the involvement of private sector laboratories and companies will be as follows:

- Private sector laboratories to provide an expression of interest to sequence samples routed through an INSACOG IGSL to INSACOG secretariat (in the mail ID: manager.insacog@nibmg.ac.in). Sequencing costs to be met by the private laboratory. Also charges shall not be levied from the patients. Private lab has to also ensure that financial resources to be used for the purpose are in compliance of Govt. of India norms and all due processes and clearances etc. are obtained in advance.
- Samples will be routed and reported via an IGSL (hub laboratory) of INSACOG.
- Private laboratories will agree to be assessed for sequencing capacities and infrastructure. The technical committee constituted by the ISACOG-SCAG will undertake assessment by interacting with the concerned laboratory.
- Based on the information provided, the INSACOG -SCAG will recommend to DBT for the inclusion or otherwise, of the private laboratory as a sequencing laboratory.
- The details of the accredited private laboratories will be communicated to the all the INSACOG- IGSLs and also posted on the DBT INSACOG Portal.
- For data privacy and protection, labs have to adhere to Biotech - PRIDE (Biotech - Promotion of Research and Innovation through Data Exchange) guidelines. Overall, data protection as per Govt policy must be ensured in the whole process.
- Sequencing SOPs and reporting to the IGSL will follow the norms for the INSACOG network.
- IGSLs will provide isolated RNA to the private laboratories for construction of c-DNA libraries and sequencing. These samples will be coded and no information on SRF IDs and other details will be provided.
- Private laboratories will provide sequence information as FASTQ and Consensus FASTA files to the IGSLs for submission to the national portal.
- Sequencing information will be confidential and not shared with any third party, media or the public.
- The private laboratories will preserve the RNA samples for a specific period to be decided by IGSLs in consultation with INSACOG-SCAG.
- The HUB IGSLs to which the laboratories are assigned will perform a monthly QC

check of sequencing, by re-sequencing of a subset of the RNA sequenced at the private laboratory.

- Association with INSACOG shall not be used for advertising purposes in any manner.
- INSACOG reserves the right to withdraw recognition of the lab at any point of time.

FORMAT FOR APPLICATION

Expression of Interest by State Govt./ Central Govt. nominated laboratories and Not for Profit NGOs for Joining the INSACOG Network

1. Name and Complete Address of the Laboratory
2. Details of the Lab. In charge with Telephone No. and Email Id
3. IGSL partner laboratory/ies _ Letter[s] of support to be included.
4. Available sequencing platform with model no./ year of procurement and capacity .
5. Weekly and monthly capacity of sequencing viral genomes
6. Trained laboratory personnel available for sequencing and bioinformatics analysis
7. Access controlled freezers for storage of RNA
8. Approximate turnaround time (TAT) for reporting sequencing information.
9. Ability to provide both FASTQ and consensus FASTA files for each sample sequenced.
10. Previous experience of whole genome sequencing service.
11. If isolating RNA (based on the point in the previous page?) IBSC approval of the BSL2+ capability in the laboratory

Signature

Name

Date

FORMAT FOR APPLICATION For laboratory belonging to Private Sector

EXPRESSION OF INTEREST (on company letterhead) for Joining the INSACOG Network

1. Name and Complete Address of the Laboratory
2. Details of the Lab. In charge with Telephone No. and Email Id
3. IGSL partner laboratory/ies _ Letter[s] of support to be included.
4. Available sequencing platform with model no./ year of procurement and capacity.
5. Weekly and monthly capacity of sequencing viral genomes
6. Weekly/monthly commitment to sequence for INSACOG
7. Trained laboratory personnel available for sequencing and bioinformatics analysis
8. Access controlled freezers for storage of RNA
9. Approximate turnaround time (TAT) for reporting sequencing information.
10. Ability to provide both FASTQ and consensus FASTA files for each sample sequenced.
11. Previous experience of whole genome sequencing service.
12. Source of financial resources.

Signature

Date

Annexure 8

INSACOG Data Submission Guidelines



INSACOG data submission guidelines at IBDC

Version 1.2

IBDC has developed a separate track for the submission of Covid-19 sequencing data at the **Indian Nucleotide Data Archive - Controlled Access (INDA-CA)**.

STEP1: User registration

In order to submit COVID19 data sets, the user must register at INDA-CA(<https://inda.rcb.ac.in/indasecure>).

1. Click on the 'Register' button on the top panel on the INDA-CA homepage.
2. Provide all the relevant details.
3. All new user accounts need to be activated after verification by IBDC staff. An email would be sent confirming the account activation at the registered email.

STEP 2: Select data submission track

1. Visit IBDC website (www.ibdc.rcb.res.in) and click on '**Submit Data**' tab or directly go to **INDA-CA** home page (skip to step 3).
2. Select '**INDA-CA**' on the next page.
3. **Select 'Template-based submission (INSACOG)' track within the box .**

Important points:

IBDC provides a special track for submission of the INSACOG COVID 19 sequencing data:

- a. **Template-based submission:** Metadata is provided in an 'MS Excel' file and sequence data is uploaded via FTP.
4. Use login credentials provided by IBDC staff to login via any of the above submission tracks.

STEP 3: Submit data using the 'Template-based submission' track

1. Log in to the '**Template Based submission**' track using the login credentials received from IBDC.
2. Download the **template 'MS Excel'** by clicking on the '**MetaData INSACOG template**' link provided on the dashboard (Left panel). Fill in the metadata for multiple samples in the template file.
3. **Register 'Study':** Register a study by providing relevant details.
4. **Upload Metadata:** Click the link "**Upload metadata**" on the left panel of the dashboard.
 - a. Select an appropriate 'Study' to which the data has to be submitted.
 - b. Subsequently, select the appropriate 'metadata file'.
 - c. Click on the '**Upload and Save button**'.

If fields of the metadata file are as per format a success message will be displayed on the page. If any discrepancy is found an appropriate message will be displayed and the user has to correct the file and re-upload.

Important points:

- The metadata template downloaded should be filled in according to the headers.
- Some fields are mandatory which have to be filled and it is advised **not to edit the first two lines** of the metadata template.

1

- Following fields are mandatory: '**Virus name**', '**Collection Date**', '**State**', '**Sequencing Technology**', '**Sample id given by the sample provider**', '**Filename1**', '**md5sum1**' and '**Strategy**'.
 - **Virus name** should be unique, duplicate entries will not be taken.
 - Please enter the date in a uniform format of "**dd/mm/yyyy**" or "**dd-mm-yyyy**" in **Collection date and Last vaccinated fields** .
 - **Filename1** (Raw data filename) should be given with the extension for example "**Filename.fastq.gz**".
 - Please do not enter text or any other characters in date fields.
 - Please check for special character/tab/space before the dates in Collection date or in Last vaccinated column.
- The filename given in the metadata details **should be exactly the same** as the file which has to be uploaded in the next step.
- A valid **Md5sum** value for each file has to be given.
- **Strategy** field should be given either as **single** or **paired**. Values apart from this will not be considered.

5. **Raw Data Upload:** Once the metadata has been uploaded successfully user can use the FTP details after clicking on the '**Raw Data upload**' tab on the left panel on the dashboard. It is recommended to create a new folder in the **FTP** directory and upload the **fastq files** in the created folder. **The filename should be the same as given in the metadata file uploaded.**

6. Map Uploaded files

- In this step user has to map the uploaded 'raw data files' with the 'metadata'.
- User has to select the folder of the uploaded files and click on the button "**Click here to Map All Uploaded Files**". All the files with matching metadata will be mapped, a count of mapped and not mapped files will be shown on the page. IBDC accession will be provided when the files are mapped correctly for each uploaded sample.

Important points:

- The files will be mapped only if the filename given in the metadata matches the uploaded filenames.
- INDA-CA will verify the 'md5sum' value provided in the meta-data file with that of the data file. If they don't match user will be notified via email and the correct file has to be re-uploaded in the same folder.

Support: For any queries please write to us at 'indasupport@ibdc.rcb.res.in'

Genomic Surveillance for SARS-CoV-2
In India
Indian SARS-CoV-2 Genomics Consortium
(INSACOG)
(Updated guidelines and SOPs)
(dated 15.07.2021)



1. Background

Globally, the SARS-CoV-2 virus has posed the biggest public health challenge of the century. However, India, has largely been able to keep the mortality low through effective diagnosis, appropriate treatment measures and contact tracing. In order to fully understand the spread and evolution of the SARS CoV-2 virus, and to tackle its future spread sequencing and analysing the genomic data of SARS CoV-2 is required. Any changes to the genetic code, or mutations in the SARS-CoV-2 virus, can be observed in the samples through genomic sequencing studies. Global experience has revealed the importance of genomic variants playing a key role in transmission and subsequent surges. The study of accumulation of mutations in the viral genomes enables us to compare virus samples and viral lineages in order to understand if local outbreaks are caused by transmission of single or multiple viral lineages. Analysis of SARS-CoV-2 genome sequences also allows us to study the evolution of the virus and assess whether these mutations influence transmission, clinical outcomes, severity, and their role in developing public health intervention measures and vaccines.

Indian SARS-CoV-2 Genomics Consortium (INSACOG) was established to expand Whole Genome Sequencing of SARS-CoV-2 across the nation, aiding our understanding of how the virus spreads and evolves. The Consortium had initially started with a network of ten regional genome sequencing laboratories spread across the country and has now been expanded to currently include 28 INSACOG Genome Sequencing Laboratories (IGSLs) that are mapped to the States and UTs for facilitating smooth flow of samples. The viral genome sequencing data generated by these IGSLs is analysed by the respective centres and sent to the National Centre for Disease Control (NCDC), Delhi for collation and integration. The Central Surveillance Unit (CSU) under Integrated Disease Surveillance Programme (IDSP) at the National Centre for Disease Control (NCDC) regularly collects data in a decentralized manner from various States/districts. This helps in correlating the data from the genome sequencing laboratories with the field data trends and study the linkages (if any) between the genomic variants and epidemiological trends based on COVID data generated by State and District Surveillance Units of IDSP. These correlations also enhance the understanding of unusual events like Vaccine breakthrough, suspected reinfections, super spreader events, outbreaks etc. The ultimate goal of this activity is to strengthen public health interventions across the country and breaking the chain of transmission. Linking this data with the IDSP epidemiological data and data on patient's symptoms will allow us to better understand the viral infection dynamics, morbidity and mortality trends. Further, the data can be linked with host genomics, immunology, clinical outcomes and risk factors for a more comprehensive outlook.

Over the past few months many variants have been detected through the Whole Genome Sequencing activities undertaken by INSACOG. The information so obtained has been regularly shared with States/UTs to strengthen their public health response to the pandemic. A few of the

variants thus detected have contributed to surges across various regions and other variants are being monitored by INSACOG for their potential role in disease transmission dynamics. For this whole exercise to become more meaningful, it is necessary that timely clinical data and adequate number of RTPCR positive samples for genome sequencing are provided by the States/UTs to facilitate better analysis of the transmissibility and virulence of these variants.

2. Stakeholders

1. State Governments - IDSP-SSUs and DSUs situated in States and Districts
2. Sentinel Sites –Identified RT-PCR labs, secondary & tertiary care hospitals of States and UTs.
3. INSACOG Genome Sequencing Labs
4. Ministry of Health and Family Welfare - NCDC, ICMR
5. Ministry of Science and Technology – DBT/DST/CSIR

3. Objectives of the Indian SARS-CoV -2 Genomics Consortium (INSACOG)

The overall aim of the **Indian SARS-CoV-2 Genomics Consortium** is to monitor the genomic variations in the SARS-CoV-2 on a regular basis through a multi-laboratory network. The mandate of INSACOG has evolved with time and the focus has shifted from primarily tracking variants among international passengers to early detection of variants that may emerge within the Country. In the present scenario, genomic surveillance of SARS-CoV-2 has the following objectives:

- Early detection of genomic variants of public health implication through sentinel surveillance
- To determine the genomic variants in unusual events/trends (Vaccine breakthrough, super-spreader events, high mortality/morbidity trend areas etc.)
- To correlate the genome surveillance data with epidemiological data
- To suggest public health actions based on the analysis of genomic and epidemiological surveillance data

4. Genomic Surveillance Strategy:

The current genomic surveillance is based on a three-pronged strategy that focusses on:

- 4.1. International Travellers**
- 4.2. Regular on-going surveillance in the community (through Sentinel Sites) &**
- 4.3. Event based surveillance in special case scenarios**
- 4.4. Other general considerations**

The salient features of the three-pronged strategy are:

4.1 International Travellers:

- i. The aim is to detect the entry of new variants/mutants in India from other countries.
- ii. The details of the Countries, Points of Entry and Variants to be included in this approach shall be decided on the basis of the available information and the guidance released by the WHO under the provisions of the International Health Regulations (IHR 2005) from time to time.
- iii. Guidance for Epidemiological Surveillance and Response in the context of new variant of SARS-CoV-2 virus detected in other Countries shall be made available on the MoHFW website as and when the need arises.
- iv. This shall involve, inter alia, screening & testing of international travellers followed by sequencing a fixed percentage of the positive cases thus detected.

4.2 Regular on-going surveillance in the community through Sentinel Sites:

- i. This is based on the WHO document “Operational Considerations to expedite Genomic Sequencing Component of GISRS surveillance of SARS CoV-2” that has been adapted taking into consideration the specific requirements of the country and the present sequencing capacity of the laboratories in the country.
- ii. The States and UTs are required to identify adequate number of Sentinel Sites ensuring representation across geographical territory (coverage of at-least 80% of the districts to be ensured) as well as the clinical spectrum of cases reported from the State/UT.
- iii. The sentinel sites can either be the RT-PCR labs or secondary & tertiary care hospitals managing COVID cases.
- iv. The sentinel surveillance methodology requires that ideally each sentinel site sends at-least 15 samples every 15 days to the identified IGSL as per standard specimen collection procedure depicted in Annexure-1.
- v. However, this methodology may be customised by the States/UTs to suit local conditions. States are encouraged to refer samples in numbers that reflect the field situation and are in-sync with the sequencing capacity of the designated IGSL mapped to the State/UT.

4.3 Event based surveillance in special case scenarios:

- i. The IDSP network captures unusual COVID-19 events such as suspected vaccine breakthrough, super-spreader events, clusters of cases with high mortality and/or morbidity and reports to the concerned DSUs for verification of these events
- ii. Upon verification, a detailed epidemiological investigation including instituting genomic analysis of the samples collected from such events is to be carried out by the concerned SSU/DSU in consultation with the CSU
- iii. The number of samples to be sent for sequencing in such events shall be determined by the investigating team and can be 100% for events such as vaccine breakthrough and suspected reinfections.
- iv. The Central Surveillance Unit – IDSP will provide technical assistance to the States/UTs in this regard.

4.4 Other General Considerations of the Strategy:

- i. The relevant case details of any sample detected with the new variant which is found to be significant from public health perspective will be communicated directly to NCDC (Director), NCDC being the Nodal Unit along with the members of INSACOG SCAG.
- ii. Data privacy is to be ensured with respect to personal identifiers as well as findings that may have national or international public health implications
- iii. Submission of genomic data to international databases e.g. GISAID etc shall not be undertaken before submitting the summary findings to NCDC.
- iv. The database of all the samples sent for genome sequencing will be maintained in the special surveillance module of IHIP-INSACOG portal. Further, clinical and outcome details of all such samples must also be stored, updated and shared with CSU-IDSP, NCDC for establishment of clinic-epidemiological correlation. Individual SSUs/DSUs must also proactively review this data.
- v. The States will be intimated in case of identification of any new variant of public health concern after discussion with the technical experts for further epidemiological analysis and planning response strategies.
- vi. All the genomic sequencing data will be maintained in National database at three sites:
 - a. NIBMG, Kalyani,
 - b. IGIB, New Delhi and
 - c. NIV, Pune.
- vii. The central database shall be accessible to all the contributing IGSLs.
- viii. The genome sequence data will be shared with international organisations only after it has been shared with NCDC.

5. Organisational structure:

- A. **Centre level:** A Nodal Unit has been created at NCDC, New Delhi with officers from Division of Bio-technology, Epidemiology and Central Surveillance Unit. This unit will act as a pivot and coordinate with the respective State/district surveillance units and plan the transportation of samples to the designated IGSL. Samples can also be transported by Sentinel Sites directly to sequencing labs. This unit at NCDC, New Delhi will also act as the Nodal National Hub for all Regional Hubs as detailed below.
- B. **Regional level:** It is proposed that the ten identified IGSL will serve as the regional hub laboratories for genome sequencing of the relevant region. These ten IGSLs are supported with 18 Satellite labs for genome sequencing activities (Annexure 2). These satellite labs will be sharing the genome sequencing data with respective hub IGSLs which in turn will update the same on IHIP-INSACOG portal. The Country will be divided into six regions for clearly defining the sample collection/transportation flow, as below:

Regional Hub	List of Hub Labs	State/UT(s)*
East and North East	1. DBT- National Institute of Biomedical Genomics (NIBMG), Kalyani (near Kolkata) Estimated sequencing capacity – 5000 per month	Andaman & Nicobar Islands, West Bengal, Bihar, Jharkhand, Assam, Tripura, Meghalaya, Manipur, Arunachal Pradesh, Sikkim, Nagaland, Mizoram Odisha, Chhattisgarh, Sikkim
	2. DBT-Institute of Life Sciences, (ILS) Bhubaneshwar Estimated sequencing capacity –1200 per month	
West	3. ICMR-National Institute of Virology (NIV), 4. DBT-National Centre for Cell Science, Pune Estimated sequencing capacity –1200 per month	Goa, Maharashtra, Gujarat, western part of MP, UT of Dadar and Nagar Haveli and Daman and Diu
South	5. CSIR-Centre for Cellular and Molecular Biology (CCMB) and 6. DBT-Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad Estimated sequencing capacity – 5000 per month at CCMB) and 1200 at CDFD	Andhra Pradesh, Telangana, Goa (northern part of Karnataka)
	7. DBT InSTEM/NCBS, Bengaluru Estimated sequencing capacity –1200 per month 8. NIMHANS, National Institute of Mental Health and Neuro Sciences Hospital (NIMHANS), Hosur Road, Bangalore	
Central	9. CSIR-Institute of Genomics and Integrative Biology (IGIB), Delhi Estimated sequencing capacity – 10,000 per month	Rajasthan, Punjab, Haryana and western part of UP. Kerala samples will be sequenced at IGIB
North & Central	10. NCDC, Delhi - Division of Bio-technology, Epidemiology and Central Surveillance Unit Estimated sequencing capacity – 3,000 per month	Eastern part of MP, Uttarakhand, Chandigarh, Delhi, Haryana, Himachal Pd., Ladakh, J&K & Punjab
*: States/UTs may plan and indicate the practical feasibility to CSU. CSU will further fine tune the regional linkages with IGSLs.		
NCDC based on a continuous and dynamic assessment exercise would readjust the mapping of states, parts of states to IGSLs.		

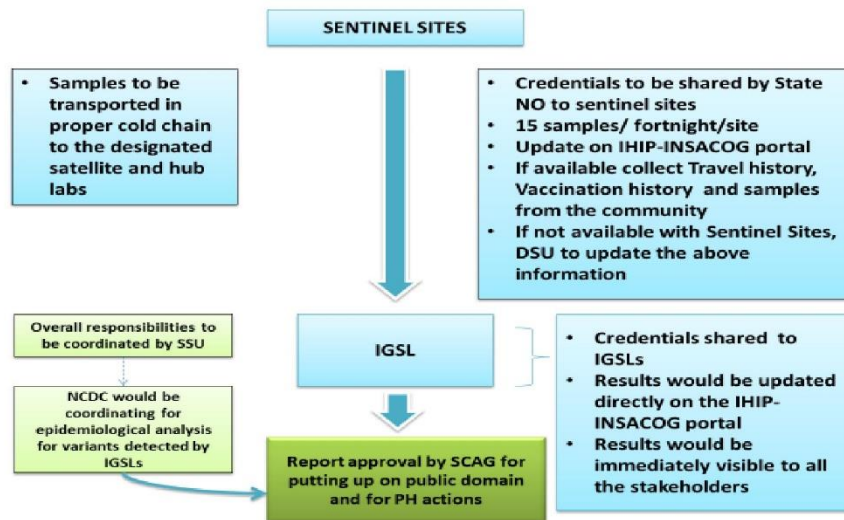
6. Flow of Information:

6.1 All the Sentinel Sites and IGSLs have been provided username and passwords for the INSACOG WGS Surveillance module of IDSP-IHIP.

6.2 The Sentinel Sites will send adequate number of samples for genome sequencing to the mapped IGSLs (10 Hub IGSLs and 18 Satellite IGSLs) – list placed at Annexure 2, and will update the details of the samples on the INSACOG WGS Surveillance module of IDSP-IHIP including the name of IGSL to which samples are being sent.

6.3 The IGSL will share the genome sequencing results (annotated data) with NCDC, Delhi for further analysis and compilation (taking help from NIBMG and IGIB sequencing analysis team).

6.4 The analysis reports will be sent periodically to the MoHFW and also shared with the relevant IGSL and CSU/SSU for necessary action.



7. Key steps in Implementation of the genomic surveillance strategy:

7.1 Each State and UT is to designate a WGS nodal officer who will be the focal point for coordination of WGS related activities between the State, Sentinel Sites, IGSLs and NCDC. The State nodal officer is responsible for overall coordination from sending of the samples to the designated IGSLs/satellite labs along with updating of data for vaccination status of the patient, travel history, and outcome on IHIP WGS portal.

7.2 The nodal officer is also responsible for identification of adequate number of sentinel sites (Minimum of 10 for larger States) in consultation with their respective State Govt./UT Administration. These sentinel sites must adequately represent geographical and clinical spectrum of the cases in the respective State/UTs.

7.3 A nodal officer for each sentinel site needs to be designated. The user credentials for these sentinel sites to be shared by the State Nodal Officer. The list of sentinel sites along with the details of the nodal officer for each site has to be communicated to IDSP, NCDC.

7.4 Adequate number of samples (At least 15 per fortnight) from each of the sentinel site should flow towards the designated IGSL. (Relaxation in this criterion of 15 per fortnight may be considered when the daily new cases reported from the Sentinel Site are less than 1) In such a case, all RTPCR positive samples in the given fortnight may be sent for WGS.

7.5 The data of the samples sent to IGSLs to be entered in the IHIP-INSACOG portal only using the credentials. The user guide for data entry is placed at Annexure – 3.

7.6 The sentinel sites or District Surveillance Units have to enter the travel history, vaccination history and outcome details of the cases sent for WGS, and review this data proactively.

7.7 IGSLs will have to enter the results of the WGS directly on to the IHIP-INSACOG portal using their credentials which is provided by the State nodal officer. The details of results updation on the portal are depicted in Annexure-4.

7.8 In case of any variant detected by the IGSLs and which may be of concern, such as having immune escape capacity or lowering immune response in the host etc., a technical discussion with the SCAG group members of INSACOG would determine for the actions to be taken at the field level. CSU-NCDC will communicate these findings to the concerned States/UTs.

8. Rapid Response Strategy:

8.1 A **Rapid Response Team (RRT)** will be formed in each State/UT by the Health Department. The team will comprise of a clinician, a microbiologist and a member from Medical College (preferably from Community Medicine Department). As soon as any mutation is detected and conveyed to the State/UT, the RRT will be deployed by the State/UT to the site, where it will investigate the mutant based on following aspects:

- i. Contact tracing of case in which the mutation is detected

- ii. Epidemiological aspects of the mutant detected with respect to number of cases, deaths in the community etc.
- iii. Clinical spectrum of the positive case to detect any change in the severity or mortality.
- iv. Samples of all members in the family in which a variant has been detected and their contacts need to be collected and sent for WGS to the mapped IGSL. This activity should be overseen by the Rapid Response Team.
- v. Take necessary containment measures in the area in conjunction with the district administration
- vi. Provide daily status reports to NCDC

8.2 Review by States/UTs

Regular reviews of the whole genome sequencing activity are to be held by Additional Chief /Principal/Secretary (Health)/MD (NHM) for the following activities:

- o Identification of adequate numbers of sentinel sites for the State/UT, ensuring geographical, demographic, and clinical spectrum representativeness
- o Ensuring that adequate number of samples are sent for WGS to the designated laboratories
- o Reviewing correlation of mutants/VOCs with the epidemiological and clinical data available with the State/UTs

9. Standard Operating Procedure for various Stakeholders:

9.1 Sentinel Sites:

- i. The sentinel sites are required to send 15 samples per fortnight to the tagged IGSLs
- ii. The sentinel sites can send samples directly to the satellite lab linked to their identified Hub lab (Details annexed)
- iii. The sample referral form (Metadata) of every sample sent for WGS has to be filled up by the Sentinel Site online on the IHIP portal (Username, passwords have been provided to all the sentinel sites)
- iv. The results would be made available to the sentinel sites only through the online referral & reporting system on IHIP
- v. The sample referral form is required to be filled up as completely as possible (including the sections on clinical severity, vaccination, travel history and outcome, if available) before sending to the IGSLs

9.2 Hub IGSLs:

- i. All the Hub IGSLs have been provided credentials for reporting the WGS results on the IHIP portal
- ii. The Hub IGSLs are expected to provide technical support to the satellite labs enabling them to function independently in the future
- iii. The Hub IGSLs are also required to enter the results of the satellite laboratories till the time the INSACOG recognizes them as independent IGSL (akin to Hub IGSL) for WGS & data entry

- iv. The sample wise results updated by Hub IGSLs shall be immediately visible to the respective sentinel site as well as the concerned IDSP SSU/DSU
- v. All the IGSLs (Hub & Satellite) have also been provided login credentials (as sentinel sites) for their own sample referral for internal WGS (3-5% of internally tested RT-PCR positive samples)

9.3 Satellite IGSLs:

- i. Each satellite IGSL is linked to a designated Hub laboratory
- ii. The sentinel sites can send samples directly to the satellite lab
- iii. The satellite labs are required to share the sequencing results with the Hub IGSLs for updating the same on portal
- iv. The satellite labs have to ensure that all relevant information pertaining to WGS is shared with the concerned Hub IGSL
- v. The satellite labs shall continuously strive to update their technical skills as per the guidance from the Hub laboratories / INSACOG.

9.4 IDSP DSUs:

- i. District Surveillance Officer (DSO) is responsible for coordination of sending samples from the sentinel sites to the designated IGSLs or satellite labs.
- ii. DSO is responsible for updating of the data with respect to vaccination status, travel history and outcome of the patients for whom the samples have been sent for WGS.
- iii. Provide continuous support and constructive feedback in a timely manner and also update the records of samples sent by sentinel sites including clinical and outcome data of each sample sent for WGS..

Annexure 1

Specimen Collection, Packaging and Transport Guidelines for SARS-CoV-2 positive samples for genome sequencing

To be used for genome sequencing of RT-PCR positive samples by the laboratory personnel from Government or private health authorities/ hospitals.

Purpose: Specimen packaging and transport of clinical specimens to IGSL for genome sequencing.

Sample collection:

- a) From the identified Sentinel Sites as per the strategy
- b) From Non-Sentinel Sites (Labs/Hospitals) in case of unusual events such as vaccine breakthrough, super-spreader events, high mortality/morbidity trend areas etc.

Data sheet:

The Sentinel Sites will submit the sample referral form on the INSACOG WGS Surveillance module of IDSP-IHIP portal.

Roles and Responsibilities: The Sentinel Site will collect, package & transport SARSCoV-2 positive samples.

Only those samples which are positive for SARS-CoV-2 by RT PCR preferably with a Ct value of 25 or less should be packaged & transported.

After carrying out the RT-PCR test the remaining samples (within 72 hours of collection, stored at 2-8°), which are RT-PCR positive (Ct value <25), will be transported in VTM with cool pack (4-8 degree) or in ice.

Alternatively, remaining RNA samples may be stored and aliquoted in the 1.5 ml microcentrifuge tubes followed by proper labelling and sealing with the parafilm (stored at -70 degree Celsius). RNA placed together in plastic/ cardboard cryo-box and packed in the thermocol box with dry ice should be shipped to the respective IGSL for sequencing.

Samples should be packaged and transported with all biosafety precautions and should be accompanied with line list and details of samples including the Ct values of all the target genes detected in standard triple packaging.

The packaging consists of three layers as follows.

1. **Primary receptacle:** A labelled primary watertight, leak-proof receptacle containing the specimen. The receptacle is wrapped in enough absorbent material to absorb all fluid in case of breakage.
2. **Secondary receptacle:** A second durable, watertight, leak-proof receptacle to enclose and protect the primary receptacle(s). Several wrapped primary receptacles may be placed in one secondary receptacle. Sufficient additional absorbent material must be used to cushion multiple primary receptacles.
3. **Outer shipping package:** The secondary receptacle is placed in an outer shipping package which protects it and its content from outside influences such as physical damage and water while in transit

Personal protective equipment (apron, hand gloves, face shield, N95 Masks etc.) need to be used and all biosafety precautions should be followed while carrying out sample packaging and transport.

Annexure 2**List of INSACOG Genome Sequencing Laboratories (IGSL) – Hub and Satellite Labs**

Sr.No.	IGSL	State	Type of Lab (Hub/Satellite)	Name of Hub Lab
1	NIBMG, Kalyani	West Bengal	Hub	N/A
2	ILS, Bhubaneswar	Odisha	Hub	N/A
3	IGIB, New Delhi	Delhi	Hub	N/A
4	NCDC, New Delhi	Delhi	Hub	N/A
5	CCMB, Hyderabad	Telangana	Hub	N/A
6	NCCS, Pune	Maharashtra	Hub	N/A
7	NIV, Pune	Maharashtra	Hub	N/A
8	CDFD, Hyderabad	Telangana	Hub	N/A
9	NIMHANS, Bengaluru	Karnataka	Hub	N/A
10	InStem/NCBS, Bengaluru	Karnataka	Hub	N/A
11	CBR-IISc, Bengaluru	Karnataka	Satellite	InStem/NCBS, Bengaluru
12	NCL, Pune	Maharashtra	Satellite	CCMB, Hyderabad
13	JNCASR, Bengaluru	Karnataka	Satellite	NIMHANS, Bengaluru
14	RMRC, Bhubaneswar	Odisha	Satellite	NIV, Pune
15	GBRC, Gandhinagar	Gujarat	Satellite	IGIB, New Delhi
16	IICB, Kolkata	West Bengal	Satellite	NIBMG, Kalyani
17	ILBS, New Delhi	Delhi	Satellite	NCDC, New Delhi
18	IBSD, Imphal	Manipur	Satellite	ILS, Bhubaneswar
19	NEIST, Jorhat	Assam	Satellite	NIBMG, Kalyani
20	CDRI, Lucknow	Uttar Pradesh	Satellite	IGIB, New Delhi
21	NBRI, Lucknow	Uttar Pradesh	Satellite	IGIB, New Delhi
22	IMTECH, Chandigarh	Chandigarh	Satellite	NCDC, New Delhi
23	RGCB, Thiruvananthapuram	Kerala	Satellite	CDFD, Hyderabad
24	NIRT, Chennai	Tamil Nadu	Satellite	NIV, Pune
25	RMRC, Dibrugarh	Assam	Satellite	NIV, Pune
26	AIIMS, New Delhi	Delhi	Satellite	NCDC, New Delhi
27	BJGMC, Pune	Maharashtra	Satellite	NCCS, Pune
28	IISER, Pune	Maharashtra	Satellite	NCCS, Pune

REFERENCES

1. Adams, J. (2008). DNA sequencing technologies. *Nature Education*, 1(1), 193.
2. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, 21(1), 30. <https://doi.org/10.1186/s13059-020-1935-5>
3. Ardui, S., Ameer, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*, 46(5), 2159-2168. <https://doi.org/10.1093/nar/gky066>
4. Association of Public Health Laboratories. (2016). Next Generation Sequencing Implementation Guide. In.
5. Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., Korfi, K., & Wang, J. (2019). Applications and analysis of targeted genomic sequencing in cancer studies. *Comput Struct Biotechnol J*, 17, 1348-1359. <https://doi.org/10.1016/j.csbj.2019.10.004>
6. Brandies, P. A., & Hogg, C. J. (2021). Ten simple rules for getting started with command-line bioinformatics. *PLoS Comput Biol*, 17(2), e1008645. <https://doi.org/10.1371/journal.pcbi.1008645>
7. Brant, A. C., Tian, W., Majerciak, V., Yang, W., & Zheng, Z. M. (2021). SARS-CoV-2: from its discovery to genome structure, transcription, and replication. *Cell Biosci*, 11(1), 136. <https://doi.org/10.1186/s13578-021-00643-z>
8. Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., Naing, Z., Yeang, M., Verich, A., Gamaarachchi, H., Kim, K. W., Luciani, F., Stelzer-Braid, S., Eden, J. S., Rawlinson, W. D., van Hal, S. J., & Deveson, I. W. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun*, 11(1), 6272. <https://doi.org/10.1038/s41467-020-20075-6>
9. Callaway, E. (2021). Beyond Omicron: what's next for COVID's viral evolution. *Nature*, 600(7888), 204-207. <https://doi.org/10.1038/d41586-021-03619-8>
10. Centers for Disease Control and Prevention. (2022). *What is Genomic Surveillance?* <https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html>
11. Chial, H. (2008). DNA sequencing technologies key to the Human Genome Project. *Nature Education*, 1(1), 219.
12. Cleemput, S., Dumon, W., Fonseca, V., Abdool Karim, W., Giovanetti, M., Alcantara, L. C., Deforche, K., & de Oliveira, T. (2020). Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*, 36(11), 3552-3555. <https://doi.org/10.1093/bioinformatics/btaa145>
13. Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol Health Res World*, 19(3), 190-195. <https://www.ncbi.nlm.nih.gov/pubmed/31798046>
14. Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617), 286-290. <https://doi.org/10.1126/science.1084564>
15. Creative Biomart. (n.d.). *Targeted Bisulfite Sequencing*. Retrieved 25/07/2022 from <https://www.creativebiomart.net/epigenetics/services/targeted-bisulfite-sequencing/>
16. Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS One*, 16(10), e0257521. <https://doi.org/10.1371/journal.pone.0257521>
17. Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borchering, A. P., Brownley, A., Cedenio, R., Chen, L., . . . Reid, C. A. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961), 78-81. <https://doi.org/10.1126/science.1181498>
18. Ferguson, J. M., Gamaarachchi, H., Nguyen, T., Gollon, A., Tong, S., Aquilina-Reid, C., Bowen-James, R., & Deveson, I. W. (2021). InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btab846>

19. Freed, N. E., Vlkova, M., Faisal, M. B., & Silander, O. K. (2020). Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol Methods Protoc*, 5(1), bpaa014. <https://doi.org/10.1093/biomethods/bpaa014>
20. Gargis, A. S., Kalman, L., & Lubin, I. M. (2016). Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. *J Clin Microbiol*, 54(12), 2857-2865. <https://doi.org/10.1128/JCM.00949-16>
21. Gates, A. J., Gysi, D. M., Kellis, M., & Barabasi, A. L. (2021). A wealth of discovery built on the Human Genome Project - by the numbers. *Nature*, 590(7845), 212-215. <https://doi.org/10.1038/d41586-021-00314-6>
22. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6), 333-351. <https://doi.org/10.1038/nrg.2016.49>
23. Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Consortium, C.-G. U., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*, 19(7), 409-424. <https://doi.org/10.1038/s41579-021-00573-0>
24. Hayden, E. C. (2014). Technology: The \$1,000 genome. *Nature*, 507(7492), 294-295. <https://doi.org/10.1038/507294a>
25. Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56(2), 61-64, 66, 68, passim. <https://doi.org/10.2144/000114133>
26. Hood, L., & Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. *Genome Med*, 5(9), 79. <https://doi.org/10.1186/gm483>
27. Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum Immunol*, 82(11), 801-811. <https://doi.org/10.1016/j.humimm.2021.02.012>
28. Hutchins, R. J., Phan, K. L., Saboor, A., Miller, J. D., Muehlenbachs, A., & Workgroup, C. N. Q. (2019). Practical Guidance to Implementing Quality Management Systems in Public Health Laboratories Performing Next-Generation Sequencing: Personnel, Equipment, and Process Management (Phase 1). *J Clin Microbiol*, 57(8). <https://doi.org/10.1128/JCM.00261-19>
29. Illumina. (2020). *How inserts affect sequencing performance*. <https://emea.support.illumina.com/bulletins/2020/12/how-short-inserts-affect-sequencing-performance.html>
30. Illumina. (n.d.). *Target Enrichment*. Retrieved 15/07/2022 from <https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/target-enrichment.html>
31. Institute for Systems Genomics, C. B. C. (n.d.). *File Formats Tutorial*. Retrieved 21/07/2022 from <https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/>
32. International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931-945. <https://doi.org/10.1038/nature03001>
33. Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., . . . Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*, 36(4), 338-345. <https://doi.org/10.1038/nbt.4060>
34. Korostin, D., Kulemin, N., Naumov, V., Belova, V., Kwon, D., & Gorbachev, A. (2020). Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One*, 15(3), e0230301. <https://doi.org/10.1371/journal.pone.0230301>
35. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann,

- L., Lehoczky, J., LeVine, R., McEwan, P., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921. <https://doi.org/10.1038/35057062>
36. Li, J., Wang, H., Mao, L., Yu, H., Yu, X., Sun, Z., Qian, X., Cheng, S., Chen, S., Chen, J., Pan, J., Shi, J., & Wang, X. (2020). Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. *Sci Rep*, 10(1), 17492. <https://doi.org/10.1038/s41598-020-74656-y>
 37. Li, Y., & Tollefsbol, T. O. (2011). DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol*, 791, 11-21. https://doi.org/10.1007/978-1-61779-316-5_2
 38. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012, 251364. <https://doi.org/10.1155/2012/251364>
 39. Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Front Genet*, 10, 426. <https://doi.org/10.3389/fgene.2019.00426>
 40. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376-380. <https://doi.org/10.1038/nature03959>
 41. Masser, D. R., Stanford, D. R., & Freeman, W. M. (2015). Targeted DNA methylation analysis by next-generation sequencing. *J Vis Exp*(96). <https://doi.org/10.3791/52488>
 42. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodes-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nat Hum Behav*, 5(7), 947-953. <https://doi.org/10.1038/s41562-021-01122-8>
 43. Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74(2), 560-564. <https://doi.org/10.1073/pnas.74.2.560>
 44. Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., Curran, M. D., Parmar, S., Caller, L. G., Caddy, S. L., Khokhar, F. A., Yakovleva, A., Hall, G., Feltwell, T., Forrest, S., Sridhar, S., Weekes, M. P., Baker, S., Brown, N., . . . Goodfellow, I. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis*, 20(11), 1263-1272. [https://doi.org/10.1016/S1473-3099\(20\)30562-4](https://doi.org/10.1016/S1473-3099(20)30562-4)
 45. Miura, F., Enomoto, Y., Dairiki, R., & Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res*, 40(17), e136. <https://doi.org/10.1093/nar/gks454>
 46. NHGRI. (2022). *Chromosome*. Retrieved 31/07/2022 from <https://www.genome.gov/genetics-glossary/Chromosome#:~:text=Chromosomes%20are%20threadlike%20structures%20made,in%20the%20nucleus%20of%20cells>.
 47. Niedzicka, M., Fijarczyk, A., Dudek, K., Stuglik, M., & Babik, W. (2016). Molecular Inversion Probes for targeted resequencing in non-model organisms. *Sci Rep*, 6, 24051. <https://doi.org/10.1038/srep24051>
 48. NIH. (2020). *The Human Genome Project*. <https://www.genome.gov/human-genome-project>
 49. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., . . . Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44-53. <https://doi.org/10.1126/science.abj6987>
 50. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and

- functional annotation. *Nucleic Acids Res*, 44(D1), D733-745. <https://doi.org/10.1093/nar/gkv1189>
51. Oliver, G. R., Hart, S. N., & Klee, E. W. (2015). Bioinformatics for clinical next generation sequencing. *Clin Chem*, 61(1), 124-135. <https://doi.org/10.1373/clinchem.2014.224360>
 52. PacBio. (2022). *Featured Documentation*. https://www.pacb.com/support/documentation/?fwp_workflow_step=library-preparation&fwp_sort=preserve
 53. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J Appl Genet*, 52(4), 413-435. <https://doi.org/10.1007/s13353-011-0057-x>
 54. Pel, J., Leung, A., Choi, W. W. Y., Despotovic, M., Ung, W. L., Shibahara, G., Gelinas, L., & Marziali, A. (2018). Rapid and highly-specific generation of targeted DNA sequencing libraries enabled by linking capture probes with universal primers. *PLoS One*, 13(12), e0208283. <https://doi.org/10.1371/journal.pone.0208283>
 55. Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J., & Beerenwinkel, N. (2021). V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btab015>
 56. Prabhu, M., & Gergen, J. (2021). *History's Seven Deadliest Plagues*. <https://www.gavi.org/vaccineswork/historys-seven-deadliest-plagues>
 57. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue), D501-504. <https://doi.org/10.1093/nar/gki025>
 58. Qiagen. (n.d.). *Analysis of SARS-CoV-2 data*. Retrieved 31/07/2022 from <https://digitalinsights.qiagen.com/resources/science/sars-cov-2-resources/>
 59. Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., de Jesus, J. G., Giovanetti, M., Hill, S. C., Black, A., Bedford, T., Carroll, M. W., Nunes, M., . . . Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc*, 12(6), 1261-1276. <https://doi.org/10.1038/nprot.2017.066>
 60. Rios, G., Lacoux, C., Leclercq, V., Diamant, A., Lebrigand, K., Lazuka, A., Soyeux, E., Lacroix, S., Fassy, J., Couesnon, A., Thiery, R., Mari, B., Pradier, C., Waldmann, R., & Barbry, P. (2021). Monitoring SARS-CoV-2 variants alterations in Nice neighborhoods by wastewater nanopore sequencing. *Lancet Reg Health Eur*, 10, 100202. <https://doi.org/10.1016/j.lanepe.2021.100202>
 61. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., & Nyren, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242(1), 84-89. <https://doi.org/10.1006/abio.1996.0432>
 62. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12), 5463-5467. <https://doi.org/10.1073/pnas.74.12.5463>
 63. Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sorensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*, 19(7), 823-826. <https://doi.org/10.1038/s41592-022-01539-7>
 64. Shi, Y., Wang, G., Lau, H. C., & Yu, J. (2022). Metagenomic Sequencing for Microbial DNA in Human Samples: Emerging Technological Advances. *Int J Mol Sci*, 23(4). <https://doi.org/10.3390/ijms23042181>
 65. Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol*, 122(1), e59. <https://doi.org/10.1002/cpmb.59>
 66. Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform*, 3(1), lqab019. <https://doi.org/10.1093/nargab/lqab019>
 67. The Economist. (2022). The pandemic's true death toll. *The Economist*. <https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates>

68. Thermo Fisher Scientific. (n.d.-a). *Ion Torrent Next-Generation Sequencing Construct Library*. Retrieved 17/07/2022 from <https://www.thermofisher.com/ch/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-construct-library.html>
69. Thermo Fisher Scientific. (n.d.-b). *Next-Generation Sequencing Illumina Workflow-4 Key Steps*. Retrieved 14/07/2022 from <https://www.thermofisher.com/ch/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/illumina-workflow.html>
70. Thermo Fisher Scientific. (n.d.-c). *Preparation of DNA Sequencing Libraries for Illumina Systems-6 Key Steps in the Workflow*. Retrieved 14/07/2022 from <https://www.thermofisher.com/ch/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/dna-sequencing-preparation-illumina.html>
71. Truong Nguyen, P. T., Plyusnin, I., Sironen, T., Vapalahti, O., Kant, R., & Smura, T. (2021). HAVoC, a bioinformatic pipeline for reference-based consensus assembly and lineage assignment for SARS-CoV-2 sequences. *BMC Bioinformatics*, 22(1), 373. <https://doi.org/10.1186/s12859-021-04294-2>
72. U.S. Department of Energy. (2003). <http://www.ornl.gov/hgmis>
73. Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bahr, M., Wolf, S., Shendure, J., Eils, R., Plass, C., & Weichenhan, D. (2013). Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc*, 8(10), 2022-2032. <https://doi.org/10.1038/nprot.2013.118>
74. Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*, 39(11), 1348-1365. <https://doi.org/10.1038/s41587-021-01108-x>
75. Watson, O. J., Barnsley, G., Toor, J., Hogan, A. B., Winskill, P., & Ghani, A. C. (2022). Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis*. [https://doi.org/10.1016/S1473-3099\(22\)00320-6](https://doi.org/10.1016/S1473-3099(22)00320-6)
76. Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Topfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, 37(10), 1155-1162. <https://doi.org/10.1038/s41587-019-0217-9>
77. WHO. (2022). *Tracking SARS-CoV-2 variants*. Retrieved 30/07/2022 from <https://www.who.int/activities/tracking-SARS-CoV-2-variants>
78. Wong, K.-C., Zhang, J., Yan, S., Li, X., Lin, Q., Kwong, S., & Liang, C. (2019). DNA sequencing technologies: sequencing data protocols and bioinformatics tools. *ACM Computing Surveys (CSUR)*, 52(5), 1-30.
79. World Health Organization. (2022). *WHO Coronavirus (COVID-19) Dashboard* <https://covid19.who.int/>
80. Yakovleva, A., Kovalenko, G., Redlinger, M., Liulchuk, M. G., Bortz, E., Zadorozhna, V. I., Scherbinska, A. M., Wertheim, J. O., Goodfellow, I., Meredith, L., & Vasylyeva, T. I. (2021). Tracking SARS-COV-2 Variants Using Nanopore Sequencing in Ukraine in Summer 2021. *Res Sq*. <https://doi.org/10.21203/rs.3.rs-1044446/v1>
81. Yang, H., & Rao, Z. (2021). Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nat Rev Microbiol*, 19(11), 685-700. <https://doi.org/10.1038/s41579-021-00630-8>
82. Yu, B. (2014). Setting up next-generation sequencing in the medical laboratory. *Methods Mol Biol*, 1168, 195-206. https://doi.org/10.1007/978-1-4939-0847-9_11
83. Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., & Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16, 675. <https://doi.org/10.1186/s12864-015-1876-7>

84. Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. D., Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., . . . Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270-273. <https://doi.org/10.1038/s41586-020-2012-7>