TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

FROM SEQUENCING TO KNOWLEDGE:

Design and Implementation of Tools for Genomic and Transcriptomic Data Analysis and Visualization

Dhanaprakash Jambulingam

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1954 | MEDICA – ODONTOLOGICA | TURKU 2026

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

FROM SEQUENCING TO KNOWLEDGE:

Design and Implementation of Tools for Genomic and Transcriptomic Data Analysis and Visualization

Dhanaprakash Jambulingam

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1954 | MEDICA – ODONTOLOGICA | TURKU 2026

University of Turku

Faculty of Medicine
Institute of Biomedicine
Medical Biochemistry and Genetics
Turku Doctoral Programme of Molecular Medicine (TuDMM)

Supervised by

Professor Johanna Schleutker, PhD
Institute of Biomedicine
University of Turku
Turku, Finland

Docent Csilla Sipeky, PhD
Institute of Biomedicine
University of Turku
Turku, Finland

Dr. Vidal Fey, PhD
Institute of Biomedicine
University of Turku
Turku, Finland

Faculty of Medicine and Health Technology
Tampere University
Tampere, Finland

Reviewed by

Docent Esa Pitkänen, PhD
Institute for Molecular Medicine Finland (FIMM), HiLIFE
University of Helsinki
Helsinki, Finland

Faculty of Medicine
University of Helsinki
Helsinki, Finland

Associate Professor Valerio Izzi, PhD
Faculty of Biochemistry and Molecular Medicine
University of Oulu
Oulu, Finland

Opponent

Professor Veli Mäkinen, PhD
Department of Computer Science
University of Helsinki
Helsinki, Finland

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

ISBN 978-952-02-0576-8 (PRINT)
ISBN 978-952-02-0577-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
Painosalama, Turku, Finland 2026

To my family

UNIVERSITY OF TURKU
Faculty of Medicine
Institute of Biomedicine
Medical Biochemistry and Genetics
DHANAPRAKASH JAMBULINGAM: From Sequencing to Knowledge: Design and Implementation of tools for Genomic and Transcriptomic Data Analysis and Visualization
Doctoral Dissertation, 148 pp.
Turku Doctoral Programme of Molecular Medicine (TuDMM)
February 2026

ABSTRACT

Advances in next-generation sequencing (NGS) technologies after the completion of the Human Genome Project have increased the sequencing speed and brought down the cost of sequencing leading to an influx of large scale genomic and transcriptomic studies involving whole-genome sequencing (WGS), whole-exome sequencing (WES) and RNA-sequencing (RNA-seq). Prior to 2005, before the advent of massively parallel sequencing, the primary bottleneck was sequence generation. With widespread adoption of NGS technologies, this has shifted to computational analysis, where storage and computing capacity have become key challenges. Numerous workflows exist for analysing sequencing data but each with its own set of advantages and disadvantages. To address this, I present an integrated bioinformatics framework that unites three modular tools – Kuura, Sampo, and BioCPR – to facilitate comprehensive multi-omics analysis.

Kuura performs end-to-end WES and WGS analysis with no user intervention. Kuura uses a consensus-based variant calling approach where it normalizes and integrates variant calls from multiple callers to enhance reliability and reduce bias. The output of Kuura is an annotated Variant Call Format (VCF) containing high-confidence variant calls. Sampo uses a combinational approach for transcriptomic data where preprocessing and alignment are handled through Nextflow, while differential expression analysis and visualization are performed in R, to improve workflow organisation. BioCPR, implemented in R, provides an interactive platform for analysis of expression data with correlation-based clustering, enabling users to identify co-expression and regulatory networks.

The framework allows seamless integration, with expression matrices from Sampo directly analysed in BioCPR, and variants identified by Kuura cross-referenced with expression networks. Its reproducible design makes it applicable to germline cancer susceptibility and other complex diseases, enabling discovery of disease-relevant genes and pathways.

KEYWORDS: NGS, WGS, WES, RNA-seq, Variant calling, Differential expression, multi-omics

TURUN YLIOPISTO
Lääketieteellinen tiedekunta
Biolääketieteen laitos
Lääketieteellinen biokemia ja genetiikka
DHANAPRAKASH JAMBULINGAM: Sekvensoinnista tietoon: Genomisen ja transkriptomisen datan analysoinnin ja visualisoinnin työkalujen suunnittelu ja toteutus
Väitöskirja, 148 s.
Molekyylilääketieteen tohtoriohjelma (TuDMM)
Helmikuu 2026

TIIVISTELMÄ

Ihmisen genomiprojektin valmistumisen jälkeen uuden sukupolven sekvensointitekniikoiden (NGS) kehitys on lisännyt sekvensoinnin nopeutta ja laskenut kustannuksia, mikä on johtanut laajamittaisten genomi- ja transkriptomitutkimusten kasvuun, joissa hyödynnetään koko genomin sekvensointia (WGS), koko eksomin sekvensointia (WES) ja RNA-sekvensointia (RNA-seq). Ennen vuotta 2005, ennen massiivisesti rinnakkaista sekvensointia, pullonkaulana oli sekvenssidatan tuottaminen, mutta nyt tämä on siirtynyt laskennalliseen analyysiin, jossa tallennus ja laskenta ovat keskeisiä haasteita. Sekvenssidatan analysointiin on olemassa lukuisia työnkulkuja, mutta jokaisella on omat etunsa ja rajoituksensa. Tämän ratkaisemiseksi esitän integroidun bioinformatiikan viitekehyksen, joka yhdistää kolme modulaarista työkalua – Kuura, Sampo ja BioCPR – kattavan moniomiikka-analyysin helpottamiseksi.

Kuura suorittaa kokonaisvaltaisen WES- ja WGS-analyysin ilman käyttäjän väliintuloa. Kuura hyödyntää konsensuspohjaista varianttien tunnistusmenetelmää, jossa se normalisoi ja yhdistää useiden ohjelmien tulokset luotettavuuden parantamiseksi ja harhan vähentämiseksi. Kuuran tuloksena on annotoitu VCF-tiedosto, joka sisältää luotettavasti nimetyt variantit. Sampo on suunniteltu yhdistelmämenetelmäksi transkriptomidatan analysointiin. Analyysi alkaa RNA-seq-raaka-aineistosta ja tuottaa differentiaalisia geenien ilmentymisen matriiseja ja visualisointeja. R-kielellä toteutettu BioCPR tarjoaa interaktiivisen alustan ilmentymisdatan analyysiin korrelaatiopohjaisen klusteroinnin avulla, mahdollistaen yhteisilmentymisen ja säätelyverkostojen tunnistamisen.

Viitekehys mahdollistaa saumattoman integraation, sillä Sampon tuottamat ilmentymismatriisit analysoidaan suoraan BioCPR:ssä ja Kuuran tunnistamat variantit voidaan yhdistää ilmentymisverkostoihin. Sen toistettavuuteen perustuva rakenne tekee siitä soveltuvan perinnöllisen syövän ja muiden monitekijäisten sairauksien tutkimukseen, mahdollistaen sairauteen liittyvien geenien ja reittien tunnistamisen.

AVAINSANAT: NGS, WGS, WES, RNA-seq, Varianttien tunnistus, Geenien ilmentyminen, Moniomiikka

Table of Contents

Abbreviations

A	Adenine
ASD	Autism spectrum disorder
BAM	Binary alignment/map
BCL	Binary Base Call
BP	Base pair
BQSR	Base Quality Score Recalibration
BRCA	BReast CAncer gene
BWA	Burrows-Wheeler Aligner
C	Cytosine
cDNA	Complementary DNA
cgroup	control group
CNN	Convolutional neural network
CNP	Copy number polymorphism
CNV	Copy number variant
CPM	Counts per million
CPU	Central processing unit
DAG	Directed acyclic graph
DD	Developmental delay
ddNTP	Dideoxyribonucleoside triphosphate
DEG	Differentially expressed gene
DGE	Differential gene expression
DNA	Deoxyribonucleic acid
DS	Double stranded
eQTL	Expression quantitative trait locus
ERBB2	erb-b2 receptor tyrosine kinase 2
ESCC	Esophageal squamous cell carcinoma
G	Guanine
GATK	Genome Analysis Toolkit
GIAB	Genome in a bottle
GPU	Graphics processing unit
GTF	Gene transfer format

GUI	Graphical user interface
GWAS	Genome-Wide Association Studies
HGP	Human Genome Project
HOXB13	Homeobox B13
HPC	High-performance computing
HR	Homologous recombination
HTS	High-throughput sequencing
ID	Intellectual disability
indel	Small insertions and deletion
IDT	Integrated DNA Technologies
JI	Jaccard Index
JVM	Java virtual machine
KB	Kilobase
LXC	LinuX Container
lincRNA	Long intergenic non-coding RNA
LINE-1	Long interspersed nuclear element-1
lrWGS	Long-read WGS
MAQ	Mapping and Assembly with Quality
MEM	Maximal exact match
MGE	Mobile genetic element
miRNA	MicroRNA
mRNA	Messenger RNA
nDNA	nuclear DNA
NIST	National Institute of Standards and Technology
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
OS	Operating system
PBS	Portable Batch System
PCA	Principal Component Analysis
PCAWG	Pan-cancer analysis of whole genomes
PRS	Polygenic risk scores
QC	Quality control
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
ROI	Region of interest
RT	Reverse transcription
SAM	Sequence alignment/map
SGE	Sun Grid Engine
SLURM	Simple Linux Utility for Resource Management
snoRNA	Small nucleolar RNA

SNP	Single nucleotide Polymorphism
SNV	Single nucleotide variant
SRA	Sequence Read Archive
srWGS	Short-read WGS
STAR	Spliced Transcripts Alignment to a Reference
SV	Structural variant
SVA	SINE-VNTR-Alu
T	Thymine
TCGA	The Cancer Genome Atlas
TP	True positives
tRNA	Transfer RNA
VCS	Version control systems
VEP	Variant effect predictor
VDSLOD	Variant Quality Score Log-Odds
VQSR	Variant Quality Score Recalibration
VUS	Variants of uncertain significance
W-C	Watson-Crick
WES	Whole exome sequencing
WGS	Whole genome sequencing

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Fey V*, Jambulingam D*, Sara H, Heron S, Sipeky C, Schleutker J. BioCPR – A Tool for Correlation Plots. Data. 2021; 6(9):97. https://doi.org/10.3390/data6090097

II Jambulingam D, Rathinakannan VS, Heron S, Schleutker J, Fey V. Kuura – An automated workflow for analyzing WES and WGS data. PLOS ONE. 2024; 19(1): e0296785. https://doi.org/10.1371/journal.pone.0296785

III Jambulingam D, Heron S, Schleutker J, Fey V. Sampo – A combinatorial approach for identifying differentially expressed genes. Manuscript.

*Joint first authors

1 Introduction

The completion of Human Genome Project (HGP) marked the transition in sequencing technologies from labour intensive low throughput methods to modern high throughput and automated platforms. The HGP relied on Sanger sequencing, which in turn relied on automation using capillary electrophoresis (Heather & Chain, 2016; Sanger et al., 1977). Although this method was considered ground-breaking, it was still slow and expensive (Barba et al., 2013; L. Liu et al., 2012; Venter et al., 2001). The success of HGP paved the way for next generation sequencing technologies that drastically lowered the cost and timeframes associated with sequencing thereby enhancing our ability to investigate the genetic basis of diseases (Akintunde et al., 2025).

However, the abundance of data generated by NGS methods presents substantial computational challenges (Lussier et al., 2013). Efficiently processing raw sequencing data into biological insights requires sophisticated analytical strategies. Developing fully automated, reproducible and scalable analysis pipelines has therefore become essential. Especially workflows that can handle genomic and transcriptomic data analysis play a crucial role in uncovering pathogenic variants as well as uncovering biomarkers and genomic signatures that help to understand pathogenic mechanisms (Bianchi et al., 2016; Karakoyun et al., 2023; Kulkarni & Frommolt, 2017).

In this thesis, I introduce a suite of computational tools that together form an integrated framework for genomic and transcriptomic data analysis: Kuura provides variant calls from DNA sequencing data, Sampo quantifies gene expression differences and provides downstream analysis including functional enrichment analysis, gene ontology annotations and visualizations, and BioCPR visualizes the resulting expression relationships. For example, one could use Kuura to identify candidate pathogenic variants in a cohort, use Sampo to find differentially expressed genes (DEG) in the same samples, and then use BioCPR to examine how those genes co-express under different conditions before conducting computationally intensive analyses such as Genome-Wide Association Studies (GWAS) (Kar et al., 2015). The modular design means that outputs from one pipeline can serve as inputs to another

or to third-party tools. Together, they fill a gap in disease research by providing end-to-end, reproducible analysis pipelines outside of any single disease domain.

This thesis presents a comprehensive suite of tools for NGS data analysis, extending the concept of unified analysis frameworks. The studies' main emphasis has been modularity for enabling straightforward extension and maintenance of individual components, integration of genomic and transcriptomic analysis and reproducibility using Nextflow, Docker, and structured configuration files (Langer et al., 2025; Strozzi et al., 2019).

2 Review of the Literature

2.1 Introduction to sequencing

The genetic blueprint of an organism is stored in its deoxyribonucleic acid (DNA), a double-stranded helical structure consisting of four nucleotide bases – adenine (A), guanine (G), cytosine (C) and thymine (T), whose backbone is made of alternating sugar and phosphate molecules. The DNA nucleotides are classified into two structural groups – purines (A, G) and pyrimidines (C, T), and the helical strands of the DNA are held together by complementary hydrogen bonds between the nucleotides, such that adenine pairs with thymine (A=T) and guanine pairs with cytosine (G≡C), forming Watson-Crick (W-C) base pairs that maintain the double-helical structure, ensuring strength and fidelity in replication and transcription process.

While base pairing in DNA, follows canonical W-C base pairing rules, RNA molecules have greater flexibility due to their single-stranded nature. During the translation process, the incoming transfer RNA (tRNA) anticodon reads the messenger RNA (mRNA) codon. The first two codon nucleotides form strict canonical W-C base pairs with the anticodon, the third codon nucleotide (wobble position) forms non-canonical interactions such as guanine-uracil (demethylated form of thymine found in RNA). This method of pairing is known as Wobble base pairing, and it provides flexibility for single tRNA to recognise multiple codons ensuring degeneracy of the genetic code (F. H. C. Crick, 1966; Hoernes et al., 2018; Saint-Léger & Ribas De Pouplana, 2015).

Sequencing is the method of determining the precise order of nucleotides in DNA or RNA molecules. Sequencing of nucleic acids had its breakthrough in 1977 when Frederick Sanger and his team introduced the dideoxy chain-termination method for sequencing DNA what is now known as Sanger Sequencing (Sanger et al., 1977). Prior to this, sequencing efforts were mostly focused on RNA sequencing due to challenges in handling the double helical structure of the DNA.

2.1.1 Sanger Sequencing

After the discovery of the structure of DNA by Watson and Crick in 1953, multiple attempts to sequence the DNA have been made (Watson & Crick, 1953). In 1965, Robert Holley succeeded in sequencing the first tRNA (specifically alanine transfer RNA) followed by Walter Fiers, who in 1972 determined the first DNA sequence of a complete gene (Holley et al., 1965; Jou et al., 1972). The next breakthrough came in 1977, when Frederick Sanger and his team introduced what is now known as Sanger sequencing. This method uses chain terminating nucleotides that lack the 3'OH group, which deactivates the DNA polymerase as no phospodiester bond can be formed, thereby halting the growing DNA chain at that location (Sanger et al., 1977; Sanger & Coulson, 1975). The ddNTPs used for this method were originally radioactively, and later fluorescence labelled enabling detection using sequencing gels or automated sequencing machines (Sanger et al., 1977; Slatko et al., 2018). This method became the standard for sequencing studies for almost 30 years. In 1987, Applied Biosystems invented the first automatic sequencing machine using capillary electrophoresis, thereby making sequencing faster and more accurate, which ultimately paved the way for the Human Genome Project (1990) (L. Liu et al., 2012). By the completion of the project, the method had matured but restrictions such as high cost and limited sequencing capabilities for studying large genomes made it inefficient (Barba et al., 2013; Venter et al., 2001).

2.1.2 Next generation sequencing

Following the completion of the Human Genome Project, the next significant advancement came in the form of massively parallel sequencing technologies that made it possible to sequence large volumes of DNA fragments in parallel. Next Generation Sequencing (NGS) or high-throughput sequencing is a collective term for sequencing technologies developed after the Sanger sequencing method that were capable of massively parallel analysis: they encompass various platforms that differ in their sequencing mechanisms as well as read length (Esplin et al., 2014; L. Liu et al., 2012). Today there are numerous NGS platforms available on the market, each with its own unique features that makes them suitable for different types of studies (Metzker, 2010).

Figure 1 shows the basic steps involved in sequencing workflows of NGS technologies – sample preparation, DNA extraction, adapter ligation and amplification followed by sequencing. Further developments of sequencing technologies have helped move into third and fourth generation sequencing methods such as long-read sequencing, single-cell sequencing and in-situ sequencing (Ke et al., 2016; McGinn & Gut, 2013; van Dijk et al., 2018). These technologies differ from NGS because of the way they sequence the DNA; newer

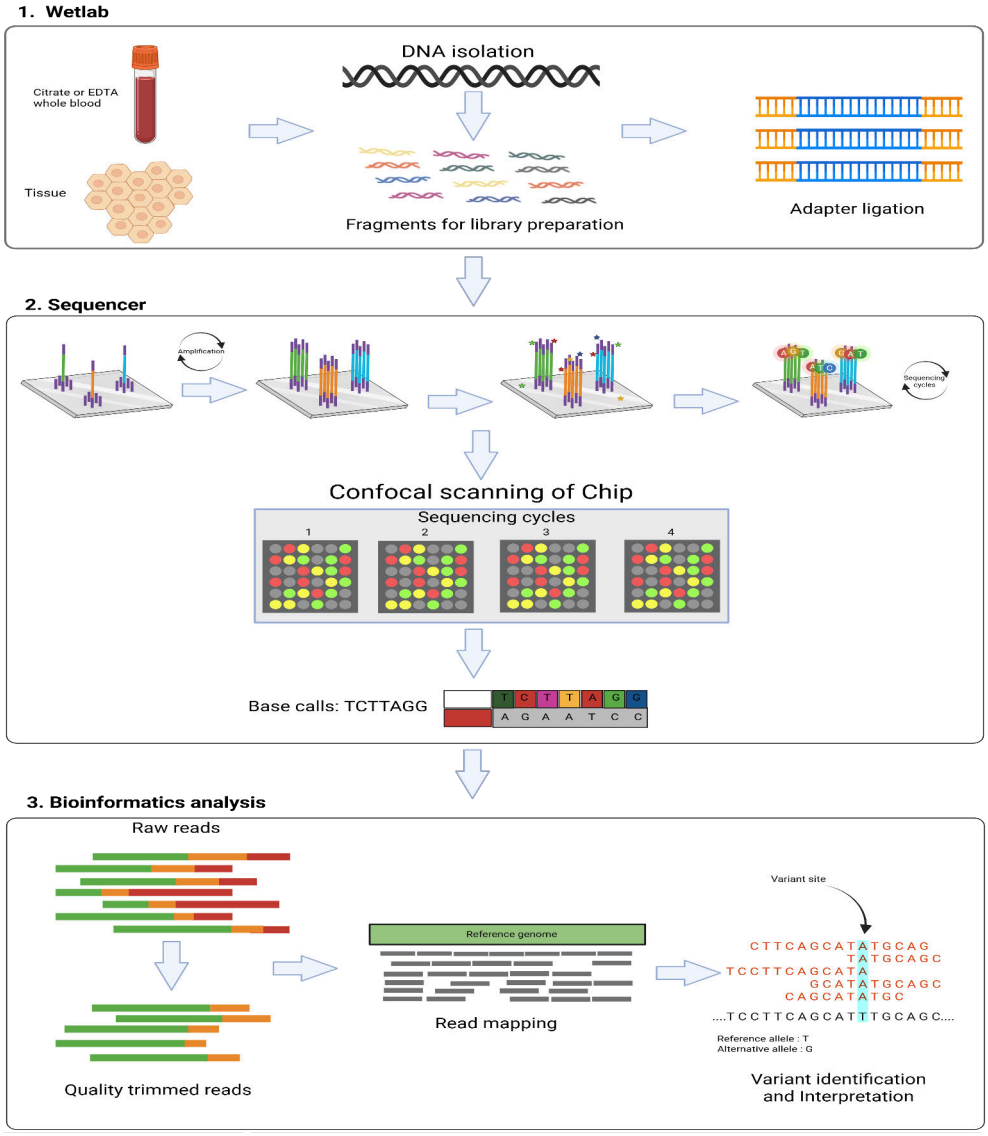


Figure 1. Example of Illumina sequencing workflow for whole genome sequencing. Panel 1. Preparation of sample for sequencing in wet lab. Panel 2. Library preparation, cluster generation and sequencing. Panel 3. Bioinformatics analysis starting from raw reads to identifying variants. Image created in Biorender.com

technologies sequence directly from template DNA while NGS requires the DNA to be amplified during library preparation (Park & Kim, 2016). Despite the recent developments, NGS technology is still the most widely used technique (McGinn & Gut, 2013).

2.2 Genome

The human DNA sequence is over 3 billion base pairs in length wound around histones, positively charged proteins, thereby creating nucleosomes which are further condensed into thread-like structures known as chromosomes, through which the genetic information is passed on to subsequent generations. The DNA sequence of an organism in its entirety including its coding sequences and non-coding sequences is known as that organism's genome, a copy of which is present in the nucleus of every cell of the organism. The length of a genome varies depending on the species. The human genome contains approximately 3 billion base pairs with approximately 20,000 protein coding genes and roughly 198,000 transcripts while the mouse genome contains approximately 2.7 billion base pairs with approximately 22,000 genes and 119,000 transcripts (Breschi et al., 2017; Sharma & Sampath, 2019; Taanman, 1999).

2.2.1 Whole genome sequencing

Whole genome sequencing (WGS) is a comprehensive sequencing technique to determine the entire DNA sequence of an organism, covering both its coding and non-coding regions as well as structural elements such as telomeres and centromeres.

Since WGS sequences the whole genome, it enables the screening and identification of a wide variety of genetic variations such as single nucleotide variants (SNV), small insertions and deletions (indels), copy number variants (CNVs) and other structural variants (SVs) (Collins et al., 2020; Goodwin et al., 2016; Shendure & Ji, 2008). Targeted sequencing like, for example, Whole Exome Sequencing (WES) meant to sequence protein-coding regions can also reliably identify smaller variants such as SNVs and indels present within these regions (Belkadi et al., 2015; Meienberg et al., 2015).

As powerful as it is as identifying genetic variations, WGS also presents unique challenges, such as computational requirements for storing and analysing such large data, higher costs compared to targeted sequencing and the complexity of interpreting variants of uncertain significance (Bagger et al., 2024; Bhérer et al., 2024; Freeman et al., 2025).

There are two types of WGS technologies available – short-read WGS (srWGS) and long-read WGS (lrWGS) (Bentley et al., 2008; Record & Reilly, 2024). srWGS typically produces reads shorter than 300 bp whereas lrWGS can produce reads ranging from 10 kilobases to several megabases, depending on the technology used (Logsdon et al., 2020; Record & Reilly, 2024; *Sequencing Human Whole Genomes on the NovaSeq 6000 System*, n.d.). Even though, the bioinformatic analysis tools are compatible for both srWGS and lrWGS, srWGS still remains the most widely used approach owing to its high accuracy, cost effectiveness and higher read depth (Choo et al., 2023; Record & Reilly, 2024).

2.2.2 Whole exome sequencing

Whole exome sequencing (WES) is a targeted sequencing technique aimed at determining the nucleotide sequence of protein coding exonic regions, which represents around 1–2% of the total human genome but contain around ~85% of known pathogenic variants (Barbitoff et al., 2020; Choi et al., 2009; López et al., 2025). In contrast to WGS, WES uses hybrid capture techniques to limit sequencing only to the exome (Mertes et al., 2011). As WES targets only the exonic regions, it generates substantially less data than WGS, the smaller data directly reduces costs associated with sequencing, storage and analysis, making WES a cost efficient solution. However, as sequencing costs continue to decline, WGS is becoming more accessible and potentially the preferred future method for studies focussed on the exome (Björn et al., 2018; Ortendahl et al., 2025).

The primary limitation of WES lie in its targeted design, which relies on pre-designed hybridization probes and a reference genome to define exonic regions, restricting analysis to coding regions (Miya et al., 2015; Seaby et al., 2016). In addition, technical bias during library preparation and hybrid capture can cause uneven coverage across the exome and increase read duplication (Björn et al., 2018; Cheng et al., 2024; Ross et al., 2013; Seaby et al., 2016).

Even though WES is a targeted approach, copy number variants (CNVs) can be identified from exome data using algorithms that make use of normalized read-depth signals, allelic imbalances, and statistical frameworks such as hidden Markov models and singular value decomposition. However, their accuracy is limited compared to CNV identification using WGS data (Conroy et al., 2000; Rajagopalan et al., 2020; Yao et al., 2017).WES is widely used in the study of diseases with broader genetic heterogeneity as it is computationally faster and its region of interest (ROI) cover all protein-coding exons (Brancato et al., 2025).

2.3 Transcriptome

In biological organisms, the flow of genetic information from DNA to RNA to protein is described by a foundational concept known as central dogma of molecular biology (Cobb, 2017; F. Crick, 1970). The DNA sequence is transcribed into RNA and then the RNA is translated into amino acid sequences that fold into functional proteins and eventually determine phenotypes at cellular and organism levels. Although the central dogma of molecular biology is still a valid concept that describes the direction of genetic information flow, it is no longer considered to be unidirectional but rather multidirectional (F. H. C. Crick et al., 1961; Haseltine et al., 2024). Notable exceptions are reverse transcription, where an enzyme called reverse transcriptase (RT) uses RNA as a template to synthesize complementary DNA (cDNA), and certain mobile genetic elements (MGE) containing reverse

transcriptase genes, called retrotransposons (Finnegan, 2012; Verwilt et al., 2023; Zabrady et al., 2023).

Similar to the term genome which refers to the entirety of the DNA sequence, the transcriptome refers to the complete set of transcripts in a cell or tissue at a given moment. The transcriptome consists of both coding transcripts such as messenger RNAs (mRNAs) and non-coding transcripts such as long intergenic non-coding RNAs (lincRNAs), small nucleolar RNAs (snoRNAs), microRNAs (miRNAs) and others, which are thought to have a role in gene regulation. There are multiple techniques for transcriptome analysis, amongst them RNA sequencing is one of the most widely used methods (Kukurba & Montgomery, 2015; Lowe et al., 2017; Z. Wang et al., 2009).

2.3.1 RNA sequencing

Ribonucleic acid sequencing (RNA-seq) is a high-throughput sequencing (HTS) approach used to analyse the transcriptome by sequencing cDNA fragments derived from RNA to quantify gene expression between different groups, condition or samples. Unlike DNA, mRNA is a transient molecule and chemically unstable as its backbone is composed of ribose as the sugar molecule instead of more inert deoxyribose found in DNA. In addition, RNA uses uracil base (U) instead of thymine found in DNA.

In comparison to Sanger sequencing and microarray-based methods, RNA-seq offers better coverage of the transcriptome and aids in identifying novel transcripts and alternatively spliced genes (Kukurba & Montgomery, 2015; Z. Wang et al., 2009).

There are several NGS platforms that differ in their underlying chemistries and library preparation protocols for sequencing the transcriptome. Figure 2 shows an overview of the basic RNA-seq workflow of Illumina Inc. The RNA-seq workflow involves RNA extraction, mRNA enrichment or rRNA depletion, RNA fragmentation, reverse transcription to cDNA, adapter ligation, cluster generation and amplification, and sequencing (Deshpande et al., 2023; Shouib et al., 2025; Stark et al., 2019). Like any biological analysis, experimental design needs to be considered when planning an RNA-seq experiment. Some key considerations include – sequencing depth, read length, choice between single or paired end sequencing, number of biological replicates required to ensure statistical significance.

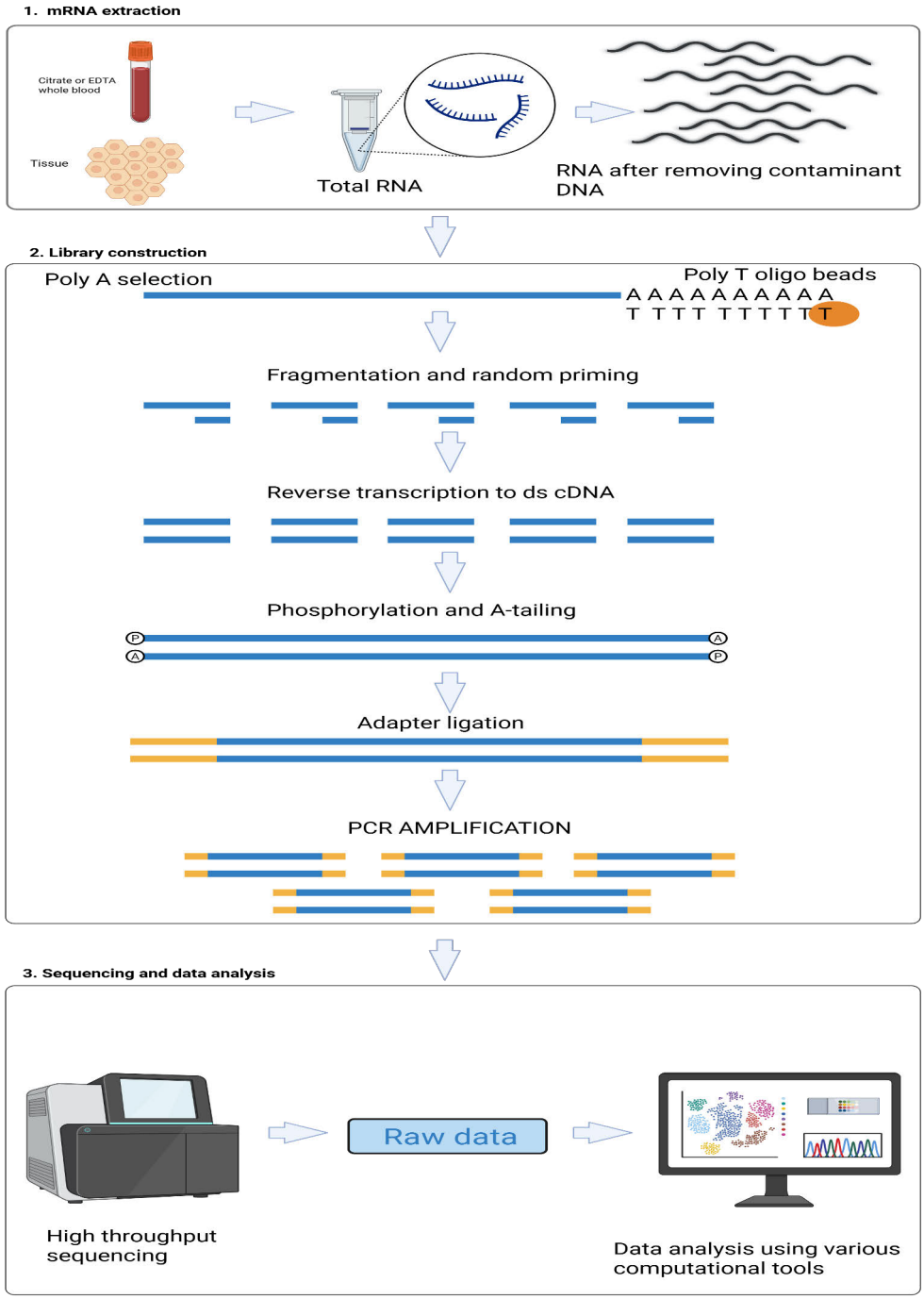


Figure 2. Example of Illumina sequencing workflow for RNA-seq data. Panel 1, Illustrates the process of extracting RNA from biological samples. Panel 2, Details the library preparation process for RNA-seq analysis. Panel 3, Shows the next steps after library preparation, sequencing followed by data analysis. Image created in Biorender.com

2.4 Applications of NGS technologies

The development of NGS technologies has cut down the costs and time associated with sequencing, thereby granting researchers access to the human genome in the context of personalized medicine (Bagger et al., 2024). The modest cost of sequencing has opened doors for NGS technologies to be applied to various situation such as full genome resequencing or targeted sequencing for variant identification, to identify structural rearrangements, including copy number variants (CNV), inversions and translocations, or RNA-seq to study gene expression (Shendure & Ji, 2008). The vast array of uses makes NGS indispensable in sequencing studies and it is rapidly becoming the standard technique for clinical applications such as cancer screening, treatment planning, and Mendelian disease diagnosis (Brlek et al., 2024; Cirulli & Goldstein, 2010). Figure 3 shows a visual representation of the applications of NGS technologies.

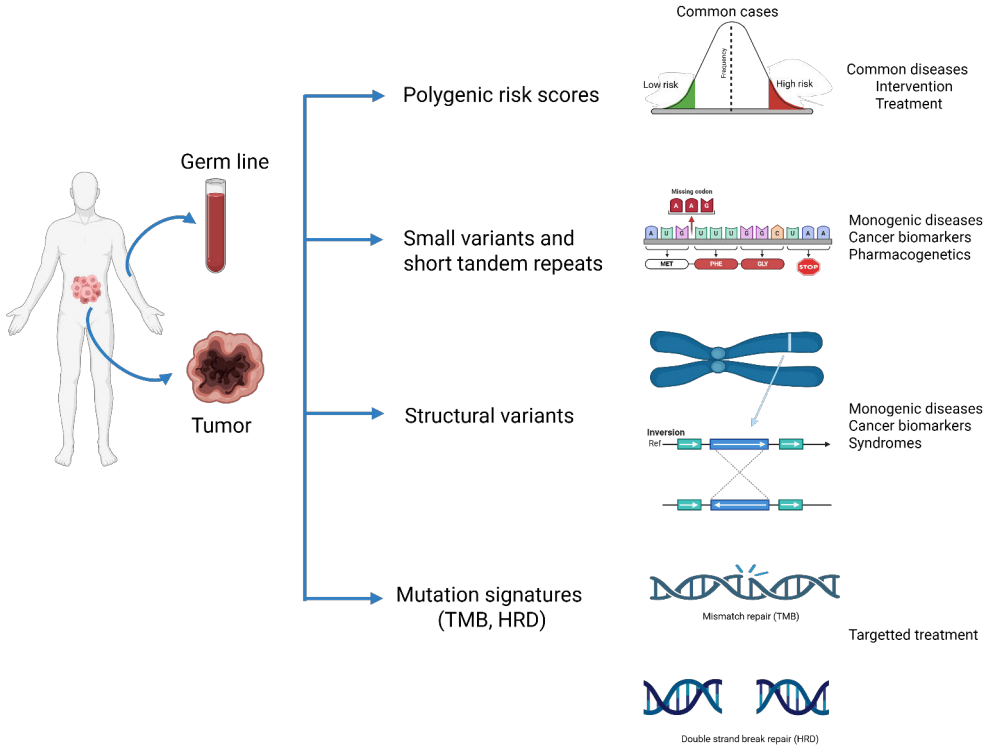


Figure 3. Examples of clinical applications of WGS. The figure shows how WGS combines germline and somatic genomic features to improve cancer risk assessment. Image is modified version of Bagger et al 2024. Image created in Biorender.com

2.4.1 Genetic variations

The National Human Genome Research Institute (NHGRI) defines genetic variation as "DNA sequence differences among individuals or populations" (*Genomic Variation*, n.d.) Differences in the genome sequence arise through mutation and recombination and contribute to the biological diversity observed among individuals and populations (Breschi et al., 2017; Halldorsson et al., 2019; Venter et al., 2001).

A typical human genome on average contains ~3.9 million SNVs and ~1 million indels and together they account for just over 0.13% of the total genome by length. Although SNVs are the most abundant type of genetic variation, SVs have much greater impact on the genome overall. An average genome carries ~31,000 larger SVs, which account for 0.63% of the total genome (Taylor et al., 2024). Variants are classified based on their cellular origin and heritability into germline variants and somatic variants.

Germline variants

Germline variants occur in gametes (sperm cells or oocytes) or their precursor germ cells during gametogenesis. Variants present in a gamete are, replicated during meiosis and are transmitted to the zygote during fertilization. The variants are then propagated through subsequent mitotic cell divisions in the developing embryo and are therefore present in every cell of the resulting offspring, making them heritable (*Definition of Germline Variant – NCI Dictionary of Cancer Terms – NCI*, n.d.; Rahbari et al., 2015; Solís-Moruno et al., 2022; Stratton et al., 2009).

Somatic variants

In contrast to germline variants which are inherited, somatic variants arise as de novo changes in somatic cells after fertilization (Greenman et al., 2007). These post-zygotic variants are not transmitted to offspring; instead, they are propagated through mitotic cell divisions leading to genetic mosaicism, where an individual can possess multiple genetically distinct cell populations (Vijg, 2014). Somatic variants accumulate over time and are a normal characteristic of aging tissue. Some of these variants give cells a growth advantage, called driver variants, allowing cell lineage to expand rapidly and thereby can lead to tumour formation. As the number of somatic variants grow with age, the likelihood of cancer also increases thereby putting especially aged individuals at risk (Solís-Moruno et al., 2022; Stratton et al., 2009; Vijg & Dong, 2020).

In addition to their cellular origin, genetic variations are classified based on their size from those that affect a single base pair to those that result in large chromosomal

rearrangements spanning millions of bases (Mullaney et al., 2010). Genetic variations can be classified into SNVs, indels, SVs and CNVs.

2.4.1.1 Single nucleotide variants

Variations originating from a change in a single nucleotide are known as SNVs. SNVs present in more than 1% of the population are commonly known as single nucleotide polymorphisms (SNP) (Richards et al., 2015; Sachidanandam et al., 2001; Vallejos-Vidal et al., 2020). SNPs are used for example in Genome Wide Association Studies (GWAS) - a research method that tests the genomes of large populations to identify genetic markers, particularly SNPs, that are statistically associated with specific diseases or traits. GWAS traditionally uses genotyping arrays that assay a fixed panel of common variants. Imputation is then used to infer additional untyped variants using haplotype phasing and reference populations. The choice of genotyping platform depends on study goals, but with declining costs WGS could become the method of choice. The summary statistics generated by GWAS also enable the calculation of polygenic risk scores (PRS). PRS are used in estimating an individual's genetic predisposition to a specific disease or trait by aggregating the effects of many common genetic variants across the genome, rather than relying on a single gene (Jiang et al., 2024; Uffelmann et al., 2021; Z. Zhao et al., 2021).

Nucleotide substitutions, a common form of point mutation, is categorized into transitions and transversions based on the nature of nucleotide bases involved. In a transition, a purine is replaced by another purine (adenine ↔ guanine) or a pyrimidine is replaced by another pyrimidine (cytosine ↔ thymine) (C. Guo et al., 2017). Transitions preserve the nucleotide base class and occur more frequently, mostly due to the chemical similarity between bases (Z. Zhang & Gerstein, 2003). In contrast, transversions involve the substitution of a purine for a pyrimidine or vice versa (e.g., adenine → cytosine). A transversion alters the nucleotide base's structural class and occurs less often (C. Guo et al., 2017).

2.4.1.1.1 Coding Regions

SNVs within the coding regions can be either synonymous or non-synonymous, depending on whether they affect the amino acid sequence of the encoded protein. Due to the degeneracy of the genetic code, where multiple codons can code for the same amino acid, some nucleotide changes do not modify the encoded protein. These are known as synonymous SNVs. Even though these changes leave the amino acid sequence intact, they can still affect gene expression through multiple mechanisms including binding of transcription factors, mRNA stability and splicing (Gotea et al., 2015; Livingstone et al., 2017; Z. Zeng et al., 2021; X. Zhang et al., 2017).

In contrast, non-synonymous SNVs do change the amino acid sequence of the protein and are further classified into missense and nonsense mutations. Missense mutations result in the substitution of one amino acid for another, which can affect protein structure, function, and stability (Sahni et al., 2015; Wong et al., 2020). Nonsense mutations introduce a premature termination codon (e.g., TAG, TAA, TGA), truncating the protein product and resulting in a non-functional protein (Wong et al., 2020).

2.4.1.1.2 Non-Coding Regions

SNVs occurring in the intronic and intergenic regions do not directly alter the amino acid sequence of a protein. However, they are not necessarily neutral and can impact gene function. Intronic SNVs can interfere with splice site recognition, triggering alternative splicing events that may yield proteins with altered amino acid sequences. Additionally, noncoding SNVs can influence gene expression by affecting regulatory elements, splicing efficiency, or transcriptional activity. (Dababneh et al., 2025; Douglas & Wood, 2011; Law et al., 2007; H. Lin et al., 2019; Pagani & Baralle, 2004; Scotti & Swanson, 2015).

2.4.1.2 Small insertion and deletion

Indels are small genomic variants that involve the gain or loss of nucleotide sequences and are typically defined as affecting fewer than ~50 bp of sequence. Larger insertion and deletion events (>50 bp) are classified as SVs, which also include duplications, inversions, deletions and chromosomal rearrangements (Chaushevska et al., 2025; Z. Liu et al., 2022).

2.4.1.3 Structural Variants

SVs are a major category of genomic alteration involving rearrangements of DNA segments larger than ~50 base pairs and can range up to millions of bases in size (Alkan et al., 2011; Pande et al., 2025; Redon et al., 2006). These variants are an important source of genetic diversity and can contribute to disease development when they disturb gene function (Shlien & Malkin, 2009). SVs are broadly classified into two types including CNVs, involving gain or loss of DNA segments, and rearrangements, including inversions and translocations, that alter the genome without changing overall copy number (Pande et al., 2025).

2.4.1.3.1 Copy Number Variants

CNVs are a subtype of SVs characterized by a change in the number of copies of a genomic segment, through deletions (loss) or duplications (gain). They range considerably in size from kilobases (kb) to several megabases (Mb), often spanning multiple genes (Mollon et al., 2023). When a CNV has a population frequency greater than 1%, it is termed a copy number polymorphism (CNP) (Richards et al., 2015). As they directly impact the number of copies of a gene, CNVs are associated with a wide spectrum of diseases, including, e.g., a number of neurodevelopmental and psychiatric conditions, such as autism spectrum disorder (ASD), intellectual disability (ID) and developmental delay (DD), and schizophrenia (Girirajan et al., 2011; Mollon et al., 2023; Shlien & Malkin, 2009).

2.4.1.3.2 Rearrangements

Rearrangements are a class of balanced SVs that alter the physical arrangement of genomic segments without changing the total DNA copy number (Karaoğlanoğlu et al., 2020). During meiosis in germ cells, rearrangements can lead to production of gametes with unbalanced genomic material, resulting in genetic disorders in the offspring (Balachandran & Beck, 2020; Xiao et al., 2023), in particular, translocations and inversions.

Translocations

A chromosome translocation is a type of genetic variant where a segment detaches from one chromosome and reattaches to a different, non-homologous chromosome (Canoy et al., 2022). Chromosomal translocations can be classified as balanced or unbalanced. In a balanced translocation, chromosomal segments are rearranged without any net loss or gain of genetic material. Since the event is copy-number neutral, carriers are often phenotypically normal, though they may have reproductive risks. In an unbalanced translocation, on the other hand, genetic material is either duplicated or deleted, which can result in infertility, recurrent spontaneous abortions, neonatal death and ID (Pei et al., 2022; X. Zeng et al., 2023). Chromosomal translocations are also a common mechanism for the formation of fusion genes, particularly in cancer (Mohammad et al., 2024).

Inversions

A chromosome inversion is a type of SV where a segment of a chromosome is excised and reinserted in the reverse orientation without net gain or loss of DNA (Khandekar et al., 2023). Inversions can be further classified as pericentric, when

involving the centromere, or paracentric when occurring only within one chromosome arm (Karamysheva et al., 2022). While many inversions are balanced, they still have pathogenic effects by disrupting genes at breakpoint regions, altering gene regulation through position effects or generate fusion genes, contributing to diseases such as cancer and reproductive disorders (Berdan et al., 2023; Pande et al., 2025).

Non-canonical / complex structural variants

In addition to the above mentioned rearrangements, some SVs arise through complex mechanisms that do not fit in the most common SV categories. Chromothripsis is a type of complex SV where a catastrophic event leads to shattering of the chromosomal region into hundreds of fragments, which are then reassembled in random order, thereby resulting in complex genomic rearrangements (Groot et al., 2023; Hancks, 2018). Retrotranspositions are a type of SVs, where MGEs such as long interspersed nuclear element-1 (LINE-1), SINE-VNTR-Alu (SVA) and other retrotransposons copy themselves via an RNA intermediate followed by insertion into new genomic locations, thereby altering the genome (Bergin et al., 2023; Cordaux & Batzer, 2009; Hancks, 2018).

2.5 Data Analysis

Advances in NGS technologies have reduced costs and time taken for sequencing thereby leading to generation of vast amounts of sequencing data and a necessity for development of novel bioinformatics tools. Modern NGS data need to go through multiple analytical stages before we can obtain biologically meaningful results from them. The need to manage the complex multi-step NGS data analysis led to the development of bioinformatics pipelines and workflow management systems integrating multiple individual tools into reproducible analysis frameworks (Köster & Rahmann, 2012; Larson et al., 2022; Satam et al., 2023).

2.5.1 Workflow management systems

Analyzing sequencing data involves careful chaining of multiple bioinformatics tools, each designed to handle a particular stage in the analysis process. The increasing complexity of the analysis pipelines and the increasing size of datasets make it essential to manage and run analyses on many samples reliably and reproducibly. To address this, modern workflow management systems have been developed that can orchestrate complex pipelines using workload managers, handle

execution across diverse computing environments and at the same time, are portable and reproducible (Vera Alvarez et al., 2021; Wratten et al., 2021).

2.5.1.1 Nextflow

Nextflow is a workflow management system that enables researchers to define and execute complex computational workflows in a scalable, portable and reproducible manner (DI Tommaso et al., 2017). Nextflow uses a dataflow programming model where processes exchange data through asynchronous channels, allowing nextflow to automatically determine task dependencies without the need for pre-computed dependency graphs. Unlike systems where outputs must always be file-based, nextflow can handle varying data and task requirements, including in-memory objects. In addition, nextflow integrates a wide range of execution platforms, including high-performance computing (HPC) and cloud computing environments and software package managers to ensure software dependencies are isolated and reproducible (Agudelo-Romero et al., 2025; Rashid et al., 2024).

2.5.1.2 Snakemake

Snakemake is a workflow management systems that follows a “Make-like” rule based approach, where computation is based on pre-estimation of all computational dependencies (Köster & Rahmann, 2012). Here, the user defines rules on how raw input data is transformed into expected outputs and lets the system compute the dependencies before execution. The dependencies between rules are represented in the form of a directed acyclic graph (DAG), which ensures that each step is run in the correct order and enables parallel execution where possible. Even though the approach is less dynamic than the dataflow model used by other workflow systems, it allows users to define complex multi-step pipelines with clear structure and reproducibility. Snakemake workflows are modular and portable and can be implemented in a wide range of execution platforms and computational environments.

2.5.1.3 Galaxy

Galaxy is a web-based platform that enables accessible and reproducible data analysis through a graphical user interface (GUI), Allowing researchers without programming experience to construct and run workflows. The platform includes a large library of tools and the Galaxy Tool Shed which enables integration of external tools into workflows (Abueg et al., 2024). Even though Galaxy’s public servers provide a powerful resource for small to medium scale analysis, large datasets and

sample sizes may prove challenging. Deploying a local Galaxy server can address resource limitations and provide security for sensitive data. Among GUI-based workflow management systems, Galaxy is widely adopted (Abueg et al., 2024; Giardine et al., 2005).

2.5.2 Execution environment

The execution environment, more commonly referred to as run environment, refers to the complete set of software, libraries and configuration files required to carry out a computational analysis. The run environment needs to be preserved to ensure the reproducibility and portability of bioinformatics workflows (Grüning et al., 2018; Mölder et al., 2025; Qin et al., 2024).This is aided by containerization technologies, which isolate the run environment from the operating system (OS). Containers package software, libraries, and dependencies into lightweight units that can be executed consistently across local machines, HPC clusters, and cloud platforms. Unlike virtual machines, which emulate an entire OS, containers share the host kernel, making them significantly faster and more resource-efficient (Merkel Dirk, 2014).

2.5.2.1 Docker

Docker is an open-source containerization platform that enables researchers to encapsulate software tools together with all necessary dependencies and environmental settings into portable images that can be shared using registries such as DockerHub. The platform uses Linux container technology, specifically LinuX Containers (LXC), control groups (cgroups) and copy-on-write file systems to resolve software dependency and runtime environment inconsistencies (Merkel Dirk, 2014). Since the run environment is packed into an immutable image, docker containers can be executed consistently across different host environments from local desktops to HPC systems, making docker suitable for distributing bioinformatics workflows (Merkel Dirk, 2014).

2.5.2.2 Apptainer/Singularity

Apptainer (formerly Singularity) is an open-source container platform designed specifically for scientific computing in HPC systems. Unlike enterprise focused containers such as Docker, which require root level access to execute workflows, Apptainer's security model enables users to execute containers without elevated security privileges (Mitra-Behura et al., 2021). Like Docker, Apptainer addresses fundamental challenges such as portability and reproducibility by encapsulating the

entire run environment. In addition, it is compatible with docker images, thereby bridging existing ecosystem gaps. Singularity works seamlessly with all HPC resource managers such as Simple Linux Utility for Resource Management (SLURM), Sun Grid Engine (SGE) and Portable Batch System (PBS)/Torque, making it an essential tool to manage run environments (Kurtzer et al., 2017).

2.5.3 Execution platforms

Large scale data analysis such as sequencing analysis requires large scale computational infrastructure. HPC clusters and commercial cloud computing platforms offer two distinct models for scientific computation, each with characteristic strengths and constraints.

2.5.3.1 HPC clusters

HPC clusters are widely used in academic and research institutions, to meet the computational demand of large-scale genomic data analysis. These clusters consist of multiple interconnected servers, called compute nodes, often managed by workload managers such as SLURM, PBS/Torque, or SGE (Elshambakey et al., 2024), which optimize the execution of parallel workloads, as they efficiently allocate resources such as central processing unit (CPUs) cores, graphical processing unit (GPUs) and memory across multiple simultaneous tasks while enforcing limits that prevent them from overloading the system. This makes them well suited for computationally heavy sequence analysis tasks such as read alignment, variant calling, and differential expression analysis (Hofert, 2024; Kurc et al., 2009; Yoo et al., 2003).

2.5.3.2 Cloud computing platforms

Commercial cloud computing platforms provide on-demand access to computing resources which enables efficient storing and processing of data remotely. Unlike traditional on-premise HPC infrastructure, cloud platforms require lower capital making them easier to adopt. However, their pay-as-you-go model can lead to increasing costs in the long run depending on the utilization (Armbrust et al., 2010; Sachdeva et al., 2024).

2.5.4 Version control systems

Version control systems (VCS) are essential to ensure reproducibility and collaboration in research by providing a structured way to track changes, manage

versions and coordinate work. Git is among the most widely used VCS currently available. Unlike older VCS, the distributed architecture of git allows it to track all changes locally improving speed while ensuring that every user has a complete copy of the project repository which also can be restored in case of failure. Git's branching model allows researchers to simultaneously work on independent development of new features without impacting the master project, and a unique staging area acts as an intermediatory point between working directory and final commit allowing users to choose which commits to include. Code repositories such as GitHub and GitLab extend git's functionality with features including issue tracking, pull requests, code review and access control, making it easier to track changes and support collaborative software development (Blischak et al., 2016; Perez-Riverol et al., 2016).

2.5.5 Genome Analysis Toolkit best practices for DNA-seq data

The Genome Analysis Toolkit (GATK) developed by the Broad institute provides widely adopted best practice recommendations for variant discovery from NGS data. The GATK best practices guideline describes a multi-step workflow that transforms raw sequencing reads in the form of fastq files into accurate, analysis ready variant calls.

The workflow begins with reads that have passed quality control (QC), which are then aligned to a reference genome using Burrows-Wheeler Aligner (BWA) - Maximal Exact Match (MEM) (Li & Durbin, 2009; Zanti et al., 2021). This step produces a Sequence Alignment/Map file (SAM) file which contains the genomic position of each read relative to the reference genome.

The aligned reads are then processed to improve data quality, including sorting, marking duplicate reads generated during PCR amplification and base quality score recalibration (BQSR) to correct systematic errors in the quality scores assigned by the sequencing machine. These preprocessing steps are designed to reduce false variant calls arising from technical artifacts.

Variant calling is performed using GATK's HaplotypeCaller, which identifies both SNVs and indels by locally reconstructing indels. Following variant calling, quality filtering is applied to distinguish true variants from noise, this is commonly done using variant quality score recalibration (VQSR) or hard filtering. At present, BWA-GATK workflow is one of the most widely used workflows and is considered the gold standard due to its emphasis on accuracy and efficiency (G. A. Van der Auwera et al., 2013; G. Van der Auwera & O'Connor, 2020).

3 Motivation and aims of the thesis

Advances in next-generation sequencing (NGS) technologies have significantly reduced sequencing time and cost, shifting the primary bottleneck from data generation to computational analysis (Noor et al., 2015; Satam et al., 2023; Torri et al., 2012).

In the context of gene co-expression analysis, existing methods frequently separated computational analysis from visualization, requiring users to preprocess and filter the results prior to visualization. This separation reduced accessibility to researchers with limited programming expertise and at the same time restricted exploratory data analysis. Although GUI based tools for co-expression analysis have been developed, they were often commercial or computationally demanding, motivating the development of an integrated and accessible solution for co-expression analysis and visualization.

A wide range of tools and workflows exist for analysing DNA-seq and RNA-seq data, however their practical use is often restricted by limited flexibility in customization, steep learning curve to get acquainted with the workflow or poor transparency of intermediate analysis steps (DI Tommaso et al., 2017; Leipzig, 2017). At the time of this work, many published pipelines were implemented as rigid ad-hoc solutions that were difficult or impossible to customise and lacked re-entrancy, thereby limiting their broader adoption (Cinaglia & Cannataro, 2025; Leipzig, 2017; Wratten et al., 2021). In response to these limitations, workflow management systems were developed to simplify workflow development and maintenance, while improving modularity, reproducibility, and portability. Although workflow management systems such as Nextflow, Workflow Description Language (WDL) and Snakemake, now address these issues, they were relatively new when this work began. There was a need for a simple, transparent, and customizable workflow, leading to the development of Kuura pipeline for analysis of WES and WGS data.

Similarly, many RNA-seq workflows were limited in scalability and focused mainly on upstream processing, often ending before differential gene expression (DGE) analysis. Downstream analysis and visualization required additional manual steps, reducing reproducibility. This work aimed to address these limitations by

developing a scalable, containerized RNA-seq workflow that integrates downstream analysis and visualization (Corchete et al., 2020; Davis-Turak et al., 2017).

In addition, as the cohort sizes increase, there is an even greater need for modular and reproducible frameworks that lets researchers mix-and-match tools, adjust parameters and inspect intermediate outputs with minimal effort.

To address these needs, the specific aims of the thesis were as follows:

1. Development of tools to facilitate analysis and visualization of gene co-expression.
2. Development of a robust and automated workflow for comprehensive analysis of WES and WGS data.
3. Development of a scalable, containerized RNA-seq workflow for automated and reproducible differential gene expression and further downstream analysis

Publication I covers the development and methodology of BioCPR, an analysis and visualization tool for gene co-expression, **Publication II** covers the development and optimization of Kuura, a variant calling pipeline for analysing WES and WGS data, **Publication III** covers the development of Sampo, an RNA-seq workflow to analyse and visualise DEGs.

4 Materials and Methods

In this section, Roman numerals I, II and III after each header indicate in which article the method was used.

4.1 Datasets

The datasets presented here were used in the publications included in this thesis. The datasets were downloaded from public data repositories and are available without any restrictions.

4.1.1 Dataset used in Publication I

The aim of this publication was to develop an interactive and user-friendly tool to facilitate the analysis of gene-gene correlations and identify co-expression networks from differential gene expression data. The tool has been used in our group for quick analysis and visualization of expression data, for publishing it was opted to use a publicly available dataset from The Cancer Genome Atlas (TCGA) – The Molecular Taxonomy of Primary Prostate Cancer (Abeshouse et al., 2015). The dataset was obtained using the cBioPortal (Cerami et al., 2012) with study id "prad_tcga_pub".

The dataset originates from a study where the authors conducted molecular analysis of 333 primary prostate carcinomas. The gene expression dataset was obtained via the cBioPortal and the dataset was converted into a format compatible with BioCPR – a matrix containing gene symbols (rows) by samples (columns) in a tab-delimited format with all non-numeric columns removed. In the processed input matrix, each row corresponds to a unique gene symbol, and each column represents an individual sample. All non-numeric columns such as a pvalue, log2FoldChange and other metadata are excluded to enable direct parsing of the input by BioCPR. This preprocessing step eliminates the need for additional user side formatting, thereby preserving data integrity and reproducibility of subsequent analyses.

4.1.2 Dataset used in Publication II

The aim of this publication was to develop a fully automated end-to-end analysis pipeline for analysing WES and WGS sequencing data. The datasets are widely used for benchmarking variant calling tools and are considered the gold standard for this purpose. The datasets are publicly available and were downloaded from the EBI FTP server. The characteristics and metadata of the datasets are detailed in Table 1,

Table 1. Information on the datasets used for validating publication II.

Sample ID	NIST ID	SRA accession number	Ancestry	Project
NA12878	HG001	SRR14724473	CEPH/UTAH	HapMap
NA24385	HG002	SRR14724472	Ashkenazi Jewish	Personal genome project
NA24631	HG005	SRR14724469	Han Chinese ancestry	

4.1.3 Dataset used in Publication III

The aim of this publication was to develop a scalable, containerized RNA-seq pipeline that runs end-to-end beginning with fastq files to identifying DEGs and finally concluding with enrichment analysis and visualization, thereby providing comprehensive sets of results with minimal intervention. A publicly available dataset from the sequence read archive (SRA) – SRP193095 was used to demonstrate the working of the Sampo pipeline (You et al., 2019).

The dataset comprises 23 paired esophageal squamous cell carcinoma (ESCC) and adjacent normal tissue samples and was designed to identify long non-coding RNAs (lncRNAs) that drive ESCC progression.

4.2 Methodology and Analysis Tools

Sequencing data is initially generated in the Binary Base Call (BCL) format. To make it compatible with external data analysis tools, the data is converted to a text-based FASTQ file format. The FASTQ format serves as the raw input for most workflows and consists of four newline-separated fields per sequence, as illustrated in Figure 4 (Cock et al., 2010).

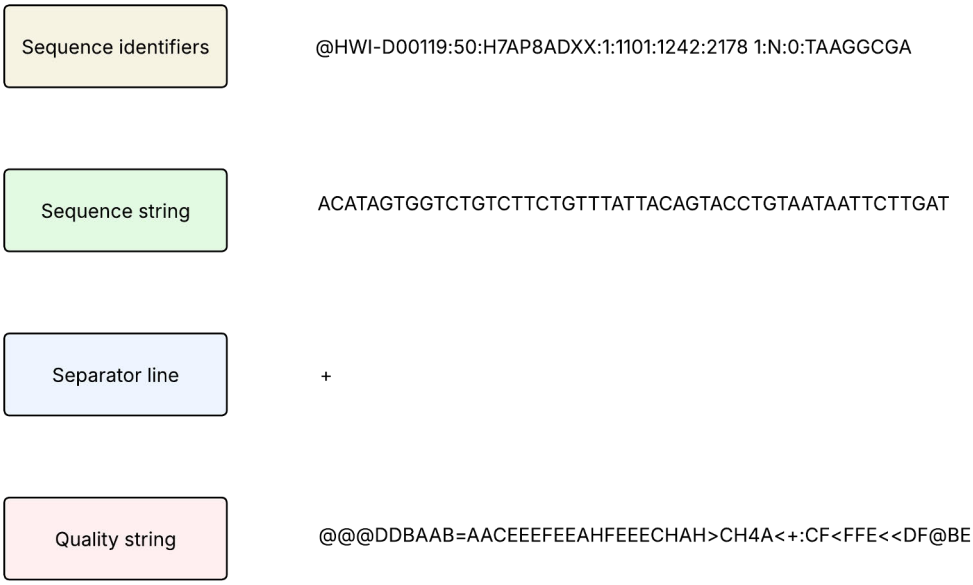


Figure 4. Composition of a FASTQ file. A FASTQ file consists of 4 lines: 1. sequence identifier starting with @, 2. nucleotide sequence, 3. separator line, 4. ASCII encoded quality scores where each quality score corresponds to the base at the same position in the sequence.

4.2.1 Preprocessing (II, III)

4.2.1.1 Quality control (II, III)

Preprocessing sequencing data is an essential step in standard NGS data analysis. It includes assessing the quality of raw reads that are in FASTQ format, followed by removing low quality bases and adapter sequences, and checking the cleaned data before proceeding with further analysis. Preprocessing is done to ensure that sequencing errors and contaminants do not bias alignment or variant calling (Andrews et al., 2012; S. Chen et al., 2018).

The initial step in any NGS data analysis is generating summary statistics to obtain an overview of the raw sequencing data. This usually involves assessing the quality of reads stored in FASTQ format and obtain per-base and per-read quality profiles (S. Chen et al., 2018; Cock et al., 2010). Over the years, multiple tools have been developed to address this issue and generate a detailed summary of the FASTQ data such as FASTQC (Andrews et al., 2012), MultiQC (Ewels et al., 2016), PRINSEQ (Schmieder & Edwards, 2011), QUACK (Thrash et al., 2018) and others. Amongst the available tools FASTQC is by far the most widely used due to its quick overview of the data as well as user friendly outputs and is used as the default tool for generating quality control summary in this work.

4.2.1.2 Read trimming (II, III)

After generating a quality summary and identifying low-quality reads and adapter content, the reads are trimmed to remove these sections and ensure that the read alignment along the genome is accurate, thereby reducing further downstream biases.

Multiple tools are available for read preprocessing, with each offering distinct advantages depending on dataset characteristics and research objectives (Yang et al., 2019). Commonly used tools for preprocessing include cutadapt (Martin, 2011), trimmomatic (Bolger et al., 2014), fastp (S. Chen et al., 2018), SOAPnuke(Y. Chen et al., 2018) or SeqPurge (Sturm et al., 2016).

Amongst commonly used tools, cutadapt and fastp were selected in this work as they cover two most common preprocessing scenarios in NGS data. Cutadapt was chosen for cases where information on adapter sequences are available as it allows exact trimming of those sequences along with user defined quality and length thresholds. Fastp was chosen as the default alternative when adapter sequences are not available as it can perform dynamic adapter detection along with quality filtering and generate post QC quality reports.

4.2.1.3 Post-trimming quality control (II, III)

Post trimming, it is good practice to run FastQC again to verify that low-quality bases and adapters have been successfully removed and the read length distribution and per-base quality scores are within acceptable ranges. While trimming typically improves overall read quality, over-trimming can result in loss of data, so it is important to balance stringency with retention (Williams et al., 2016).

4.2.2 Alignment and Base Quality Score Recalibration (II, III)

4.2.2.1 Read Alignment (II, III)

Following preprocessing, the initial step in processing sequencing data is to align the cleaned reads to a reference genome also known as read alignment or read mapping. The reads are aligned to a known reference genome, in this case the human reference genome assembly GRCh38, to determine the precise location of the reads within the genome. Alignment involves matching sequencing reads to the reference genome, accounting for possible sequencing errors, insertions, deletions, and polymorphisms (Hatem et al., 2013).

Alignment algorithms typically employ either hash-based or index-based strategies. Hash-based aligners, such as Mapping and Assembly with Quality (MAQ), construct a hash table to index the reads and then search the reference genome against the hash table for matches (Li et al., 2008). Index-based aligners, like BWA (Li & Durbin, 2010) and Bowtie2 (Langmead & Salzberg, 2012), create a compact and searchable index of the reference genome using FM-index and Burrows-Wheeler Transform to enable efficient alignment.

In the Kuura pipeline, DNA sequencing reads are aligned to the reference genome using BWA-MEM (Li, 2013), which generates SAM files. These SAM files are then converted to compressed BAM format using SAMtools (Li et al., 2009). The aligned reads in BAM format are sorted by their genomic coordinates to organize them according to their position in the reference genome.

In case of RNA-seq analysis, the alignment accuracy of reads is further complicated by the presence of splicing events. Mature mRNA transcripts are produced through splicing, where introns are removed and exons are joined together. During RNA-seq, this results in sequencing reads that may span splice junctions (Wang et al., 2008). Traditional aligners not designed to handle spliced reads may fail to align these reads correctly, necessitating the use of splice-aware aligners (Trapnell et al., 2010).

Splice-aware aligners such as Spliced Transcripts Alignment to a Reference (STAR) (Dobin et al., 2013) and Hierarchical indexing for spliced alignment of transcripts 2 (HiSAT2) (Kim et al., 2019) can accurately map reads spanning exon-exon junctions by accounting for the splicing patterns, thereby facilitating reliable downstream analysis.

In the Sampo pipeline, sequencing reads are aligned to the reference genome using STAR aligner. The aligner builds a genome index using the reference genome and gene annotation file (GTF), which helps in recognizing known gene structures and splice sites. After generating the genome index, reads are then aligned to the reference genome. Unlike DNA aligners, STAR can accurately map reads that span exon-exon junctions by splitting reads at intron boundaries.

4.2.2.2 Duplicated reads and Base Quality Score Recalibration (II)

Following read alignment, duplicated reads need to be removed from the BAM file. Duplicated reads are groups of reads that map to the same genomic location. Such duplicates often arise from technical artifacts, such as PCR amplification during library preparation or sequencing artifacts such as optical duplicates. In contrast, multiple reads mapping to the same location may represent independent biological fragments in high coverage regions, and they cannot be reliably distinguished from technical duplicates based on alignment alone (Bansal, 2017). As a result, all such

reads are flagged using Picard's MarkDuplicates tool to prevent bias and improve accuracy during variant detection. In each set of duplicates, the read with highest quality score is kept as a representative and all remaining reads are marked as duplicates.

Sequencing machines are prone to make mistakes when estimating the accuracy of each base call. Like duplicated reads, it is necessary to address the discrepancies in quality scores as they are often over-estimated. The systemic errors in quality scores are corrected using Base Quality Score Recalibration (BQSR). The BQSR process uses machine learning algorithms to model errors and adjust quality scores based on covariates like sequence context, base position within the read and respective read groups (Depristo et al., 2011). It is important to note that the BQSR process does not correct the base calls themselves but rather adjust the associated quality scores. The process is divided into two stages, where in the first stage GATK BaseRecalibrator uses known variant datasets, such as dbSNP, to exclude positions containing true variant to avoid recalibrating those as errors and creates a recalibration file. In the second stage, GATK ApplyBQSR uses the recalibration file to adjust each base's quality score ensuring that only systematic biases are corrected (G. A. Van der Auwera et al., 2013).

4.2.3 Variant Calling (II)

Variant calling is the computational process of examining each genomic position to determine where the nucleotide sequences differ from the reference genome it was aligned to (Olson et al., 2023).

To improve the confidence of the germline variant calls, Kuura uses a consensus-based approach by integrating five distinct variant callers. By doing so, this approach reduces the biases from individual variant calling algorithms and increases the reliability of the variant calls (O'Rawe et al., 2013). The variant callers used are GATK HaplotypeCaller, DeeptVariant, FreeBayes, Strelka2 and VarScan2.

4.2.3.1 GATK HaplotypeCaller (II)

Several variant callers have been developed to accurately identify variants calls from sequencing data amongst them, GATK HaplotypeCaller is one of the widely used (Lin et al., 2022). HaplotypeCaller can call both SNVs and indels simultaneously via local de-novo assembly of haplotypes in an active region (Pirooznia et al., 2014; G. A. Van der Auwera et al., 2013). It uses Bayesian statistics to determine genotype likelihoods.

Moving forward in the analysis, we need to filter out probable artifacts and examine the accuracy of the variant calls. To this end we used GATK's Variant

Quality Score Recalibration (VQSR), a machine learning algorithm that uses a Gaussian mixture model to classify variants based on how their annotation values cluster when trained with a set of high-confidence variant calls (Pirooznia et al., 2014; G. A. Van der Auwera et al., 2013). The VQSR process involves two main steps as briefed below.

- *Model Building*

GATK VariantRecalibrator uses known, highly validated variant resources (e.g., from HapMap or 1000 Genomes Project) to select a subset of variants from our variant calls that are deemed true positives. This true-positive subset is used as a training set where its annotations such as base quality, read depth or mapping quality are used to train a recalibration model to differentiate between true positives and artifacts (false positives). The model then assigns a scoring metric called Variant Quality Score Log-Odds (VQSLOD) to the INFO field of each variant.

- *Variant Scoring*

GATK ApplyVQSR uses a filtering threshold based on the recalibration table to mark which variants passed and which failed in the output VCF files.

4.2.3.2 DeepVariant (II)

Standard variant callers use a combination of different statistical models to identify and score variants, while having to account for factors such as base quality, mapping bias, and local sequence context. DeepVariant was developed to improve the available statistical models with a single deep learning model. It uses a deep convolutional neural network (CNN) that represents aligned reads as multi-channel images (read pileups) to learn the relationship between sequencing artifacts and true genotype calls directly from training data, enabling it to achieve higher accuracy (Poplin et al., 2018). DeepVariant performs variant calling in three stages.

- Data encoding

Mapped data is scanned for sites that differ from the reference genome. At sites identified as potential variants, the read and reference data is transformed to multi-channel tensors, known as pileup images, for the neural network to analyse. Standard images are made of 3 channels - red, green and blue, whereas DeepVariant's tensors are encoded as six layers - 'read base', 'base quality', 'mapping quality', 'strand', 'read supports variant' and 'base differs from ref'.

- Deep Learning Inference

Each tensor is evaluated via a pretrained CNN that has been trained using truth-set genomes to distinguish statistical patterns between artifacts and true positives. For

each tensor evaluated, the CNN generates three distinct genotype likelihoods representing homozygous reference (0/0), heterozygous (0/1), and homozygous variant (1/1).

- Refinement

The actual variant calling happens in the post-processing stage, where the genotype likelihoods generated by the neural network are converted to variant calls to produce a VCF file, complete with all necessary quality score and genomic annotations.

4.2.3.3 FreeBayes (II)

FreeBayes is a haplotype-based variant caller that uses a Bayesian model to detect SNVs, indels and complex variants from mapped sequencing reads. Unlike traditional statistical methods that evaluate each genomic site independently, Freebayes performs local probabilistic reassembly of short-read sequences and uses this reassembly to infer a set of possible haplotypes. Variants are then called from these inferred haplotypes rather than from their precise alignments (Garrison & Marth, 2012).

4.2.3.4 Strelka2 (II)

Strelka2 is a small-variant caller designed for both germline and somatic variant calling. It implements a two-tiered haplotype modelling strategy where it uses a fast alignment-based method in simple genomic regions and a local assembly-based method in complex genomic regions.

Strelka2 can self-calibrate to the data it is processing by using a statistical mixture-model-based approach to dynamically estimate error rates for both SNVs and indels directly from sequencing data, thereby accounting for technical variability and reducing false positive calls (Kim et al., 2018).

4.2.3.5 VarScan2 (II)

VarScan2 is a platform-independent heuristic-based variant detector for detecting both somatic and germline variants. It is a position-based approach that works directly on alignment pileups where it applies user-defined parameters for read depth, allele frequency and base quality scores to classify true variants from sequencing artifacts. In addition to germline variants, it can be used for somatic variant analysis, as well, where it uses Fisher's exact test to contrast allele frequencies in matched tumour-normal pairs to classify variants into germline, somatic or loss of heterozygosity (LOH). The heuristic-based approach used in VarScan2 can limit sensitivity for very low-frequency variants as compared to other, more computationally intensive methods available (Koboldt et al., 2012).

4.2.4 Variant Annotation (II)

To understand the biological and clinical effects of variants, it is necessary to assign functional and biological context to them, which is known as variant annotation. High-confidence variant calls like those identified consistently across multiple variant callers are subjected to functional annotation using Ensembl's Variant Effect Predictor (VEP) (McLaren et al., 2016). Ensembl VEP predicts the functional consequences of variants (e.g., missense, nonsense, splice-site) by integrating data from multiple genomic databases, including GENCODE for gene models, ClinVar for clinical significance, and gnomAD for population frequency. Considering the size of the dataset used for annotation, Ensembl VEP is run in offline mode with a local cache of annotation data for the GRCh38 human genome version.

4.2.5 Read Quantification and Cleaning (III)

To quantify the number of reads mapping to each gene in RNA-seq data, featurecounts (Liao et al., 2014) tool is used. This assigns sequencing reads to genomic features (e.g., exons, genes) based on alignment coordinates and produces a count matrix that reflects gene expression levels across samples.

The resulting count data is cleaned to remove any anomalies or inconsistencies, ensuring that the data is suitable for further differential expression analysis. This is done by removing the initial header line containing the program version and commands, followed by filtering the contents to retain only gene identifiers and raw read counts for the corresponding sample.

4.2.6 Differential Expression Analysis and Visualization (diffwrap) (III)

The final stage involves identifying genes that are differentially expressed between different biological conditions. Sampo pipeline uses the diffwrap R package to identify DEGs, which provides a streamlined interface to the edgeR and limma R packages thereby producing DEGs results and numerous visualizations.

4.2.7 Correlation Analysis (I)

An open-source R/Shiny application that facilitates interactive analysis and visualization of DEGs through dynamic heatmaps in article I. The tool is built in R with the Shiny framework to create a web-based interface for easy access. The tool used Pearson correlation coefficients under the assumption that log-transformed expression data follows a normal distribution. Statistical significance

of correlations is determined using two-tailed Student's t-tests, implemented through R's pt() function and results are visualized via the coreheat package.

4.3 Artificial intelligence (thesis)

OpenAI’s ChatGPT, based on the GPT-4 language model, was used to improve the grammar, spelling, and overall readability of this thesis

5 Results

5.1 Core functionalities and applications of BioCPR (I)

BioCPR is an open-source application built in the R language and uses the R Shiny library to create a portable web application that can be run locally or hosted on a server. The tool was primarily developed to ease the analysis and visualization of gene expression data and identify co-expressing genes. The tool takes differential gene expression data in the form of a matrix, where the columns correspond to different samples and the rows correspond to gene symbols.

The tool displays the content of the expression dataset for inspection and enables users to perform optimal preprocessing steps, such as standardizing gene symbols (e.g., converting Ensembl IDs to gene symbols) and pre-filtering genes. The expression dataset can be filtered in two mutually exclusive ways.

I By number of genes, where the user specifies the number of genes to be included in the plot and based on that top N most variable genes are selected.

II By gene symbol, where the user can define one or more genes of interest to be included in the analysis by selecting them.

This functionality ensures that the user has an opportunity to explore relationship between genes within their dataset by focusing on genes with high variance or examining the interaction between a gene of interest with its closely correlated partners.

Genes with higher variance across samples are prioritized in co-expression analysis as they provide more informative signal for correlation estimates, thereby increasing sensitivity. In contrast, low variance genes tend to be dominated by technical noise, making it difficult to distinguish true biological co-expression from technical variability (Zhang et al., 2021; Zhao et al., 2025). Visualization-based examination of co-expression patterns has been widely used for interpreting high-dimensional transcriptomics data, specifically for identifying gene modules, hub genes, and condition-specific networks. In line with prior studies, this method

enables focused, hypothesis-generating analysis rather than definitive novel interaction discovery (Morabito et al., 2023; Ospina et al., 2025).

By adjusting the option "Number of genes surrounding selection", after the correlation matrix is generated, the tool selects the gene of interest and includes only the gene plus a user defined number of closely correlated neighbours (genes surrounding our gene of interest in a clustered correlation matrix). The interactive filtering approach allows users to focus on a particular set of genes of interest without manually editing the input dataset.

Proceeding further to visualization, BioCPR uses Pearson's correlation for the calculation of correlation coefficients from gene expression data which are then visualized as heatmaps. To ensure statistical validity of the observed associations, p-values are calculated for each gene-pair correlation by transforming the Pearson correlation coefficient (r) into t-statistic using the formula.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Where n represents the number of samples with complete observations for the given gene pair. Two-tailed p-values are obtained from the Student's t-distribution with $n-2$ degrees of freedom using the pt function from the R stats package.

The resulting p-values are shown as visual annotations in the heatmap to help interpret dense regions. Here, a p-value of < 0.05 is assigned *, a p-value of < 0.01 is assigned ** and finally a p-value of $< 0.001$is assigned ***, to emphasize their significance. This helps to visualize the significantly correlated gene pairs with relative ease and distinguish them from noise in dense datasets.

Given the exploratory nature of correlation heatmaps and the large number of pairwise tests performed in co-expression analysis, multiple-testing correction is not applied by default. Accordingly, the significance annotations are intended to guide heuristic interpretation rather than to support confirmatory statistical inference. Users are encouraged to apply appropriate multiple-testing correction or independent validation in downstream analyses when drawing biological conclusions. In addition, correlation heatmaps can be further customised and the options provided for customization are listed in Table 2.

5.2 Interpreting the heatmap (I)

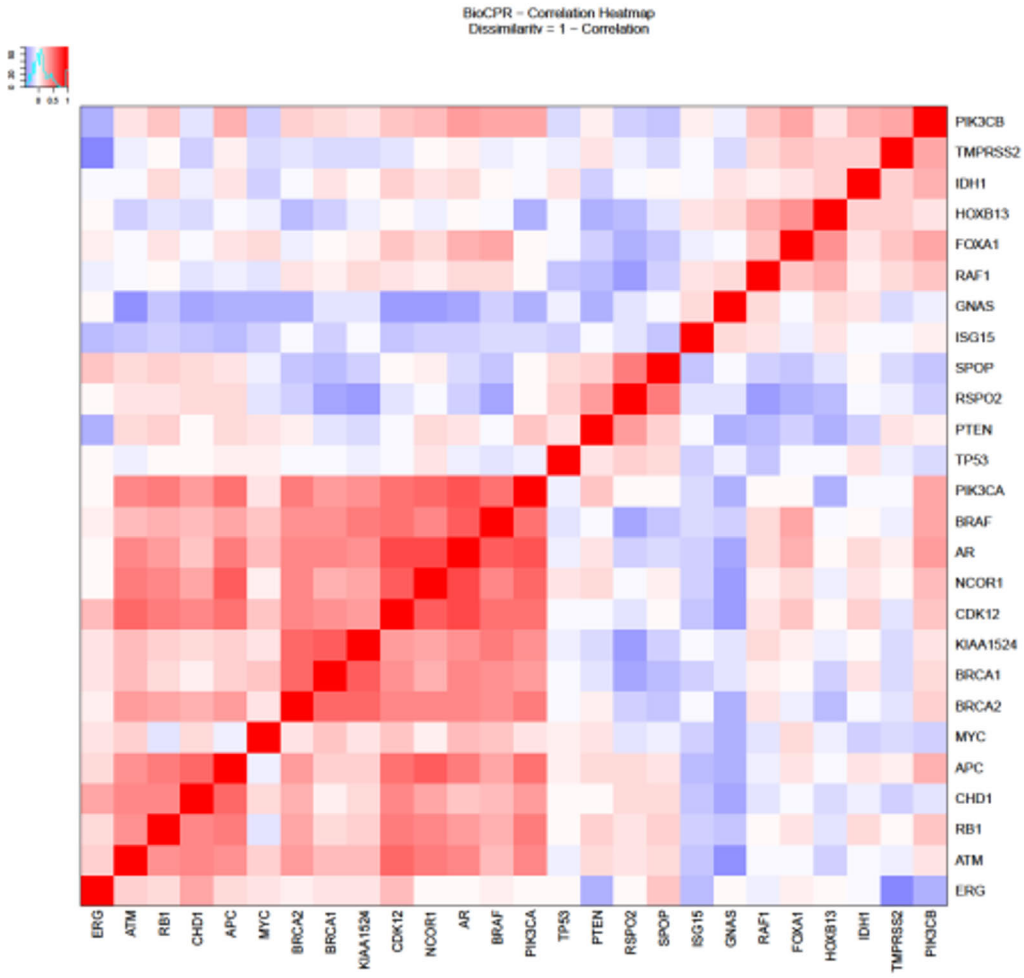


Figure 5. Correlation heatmap created from the subset of PRAD TCGA dataset. The heatmap highlights positive correlations in deep red and negative correlations in deep blue. The histogram on the top right corner of the image shows the distribution of correlation coefficients.

The tool was validated using the publicly available dataset – The Molecular Taxonomy of Primary Prostate Cancer (Abeshouse et al., 2015) to assess its ability to analyse, cluster and visualize gene co-expression patterns. Using this dataset, BioCPR generated an interactive heatmap that displayed structured blocks of positively and negatively correlated genes.

Table 2. Functionalities available for visualization and dynamic filtering of correlation plots in BioCPR. The table summarizes available tools for interactive exploration of correlation plots, including visualization options and filters that allow users to refine results based on selected criteria.

Functionalities	Effect
Add Significance Stars	Heatmap cells are annotated with *, ** or *** to indicate their p-value thresholds (p < 0.05/0.01/0.001)
Highlight Selected Genes	If the user had selected a list of genes for pre-filtering, they are shown with an outline on the heatmap
Image size	Used to change the size of the generated plot
Gene Label Size	Adjusts row/column label font size so long gene names remain visible
Enter Plot Title	Custom title for the heatmap upon export
Download as PDF	Render heatmap as a PDF
Show advanced controls	View/hide advanced controls section
Cluster and Filter Correlations	Enable/disable hierarchical clustering
Allowed Fraction of NAs per Row (%)	Tolerance for missing values per row (gene)
Filter by Correlation Value	Enable/disable threshold-based filtering
Correlation Filter threshold	Correlation value threshold to retain/remove rows
Correlation Filter Margin per Row (%)	Fraction of rows where a specified percentage of their correlation coefficients are higher than the threshold
Filter by Cutting the Dendrogram	Enable/disable tree cutting
Threshold for Tree Cutting	Tree cutting threshold
Number of Genes on a Tree Branch to be Considered a Cluster	Number of genes in tree branch to be considered a cluster.

The correlation coefficients were calculated from the expression matrix and subsequently clustered using a hierarchical clustering pipeline implemented in BioCPR. The correlation matrix is filtered for missing values, transformed into a dissimilarity matrix using the formula $(1 - r)$ and then processed using agglomerative hierarchical clustering with the complete linkage method via the $hclust$ function. Once the correlation matrix is clustered, the reordered correlation matrix was visualized as an interactive heatmap as shown in Figure 5 and is accompanied by a numeric correlation matrix.

Each cell of the heatmap represents the direction and strength of the correlation matrix, here deep red colour indicates a strong positive correlation, and deep blue colour indicates a negative correlation. A colour key and histogram are provided as

a small inset on the top left margin to summarize the distribution of correlation coefficients.

We visualized strong positive correlation between DNA repair genes such as BRCA1-BRCA2 and negative correlations between HOXB13 and genes such as BRCA1/2. These patterns are consistent with known co-expression patterns reported in previous gene expression and cancer co-expression analyses. For example, Miller et al. (2021) establish a computational framework for using gene co-expression correlations to predict gene function, while Custódio et al. (2022) provide clinical evidence for the coordinated regulation of BRCA1/2 and other homologous recombination (HR) genes in ovarian tumours. Finally, Mitsiades et al. (2024) identify a distinct genomic mechanism in which co-amplification of ERBB2/HOXB13 occurs alongside the interstitial loss of BRCA1 (Custódio et al., 2022; Miller & Bishop, 2021; Mitsiades et al., 2024).

5.3 Kuura workflow and output (II)

Kuura is implemented as complete WES/WGS analysis pipeline, containerised using docker to ensure reproducibility and orchestrated with nextflow for full automation and scalability. The pipeline comprises of four main stages: 1. quality control and read trimming, (2) alignment and BQSR, (3) variant calling, and (4) variant consensus and annotation. An overview of the pipeline structure is shown in Figure 6.

The key feature of the Kuura pipeline is the implementation of multi-caller consensus strategy for variant identification. In the final stage, the VCF outputs from five different variant callers are intersected to generate a high-confidence variant set, while variants supported by single callers are retained separately as low-confidence variant sets for optional downstream review. Consensus or ensemble approaches that integrate variant calls from multiple callers have previously been shown to improve the accuracy of variant detection compared to individual tools, primarily by reducing false-positive calls while maintaining sensitivity in both WES and WGS datasets (Chiara et al., 2018; Ewels et al., 2020; Garcia et al., 2020; Zhao et al., 2020). High-confidence variants are then functionally annotated using the Ensembl VEP and MultiQC is used to aggregate QC metrics and pipeline logs into a consolidated report. The consensus-based approach is intended to improve the reliability and precision of variant calls (Ewels et al., 2020; Krishnan et al., 2021; Sandmann et al., 2017).

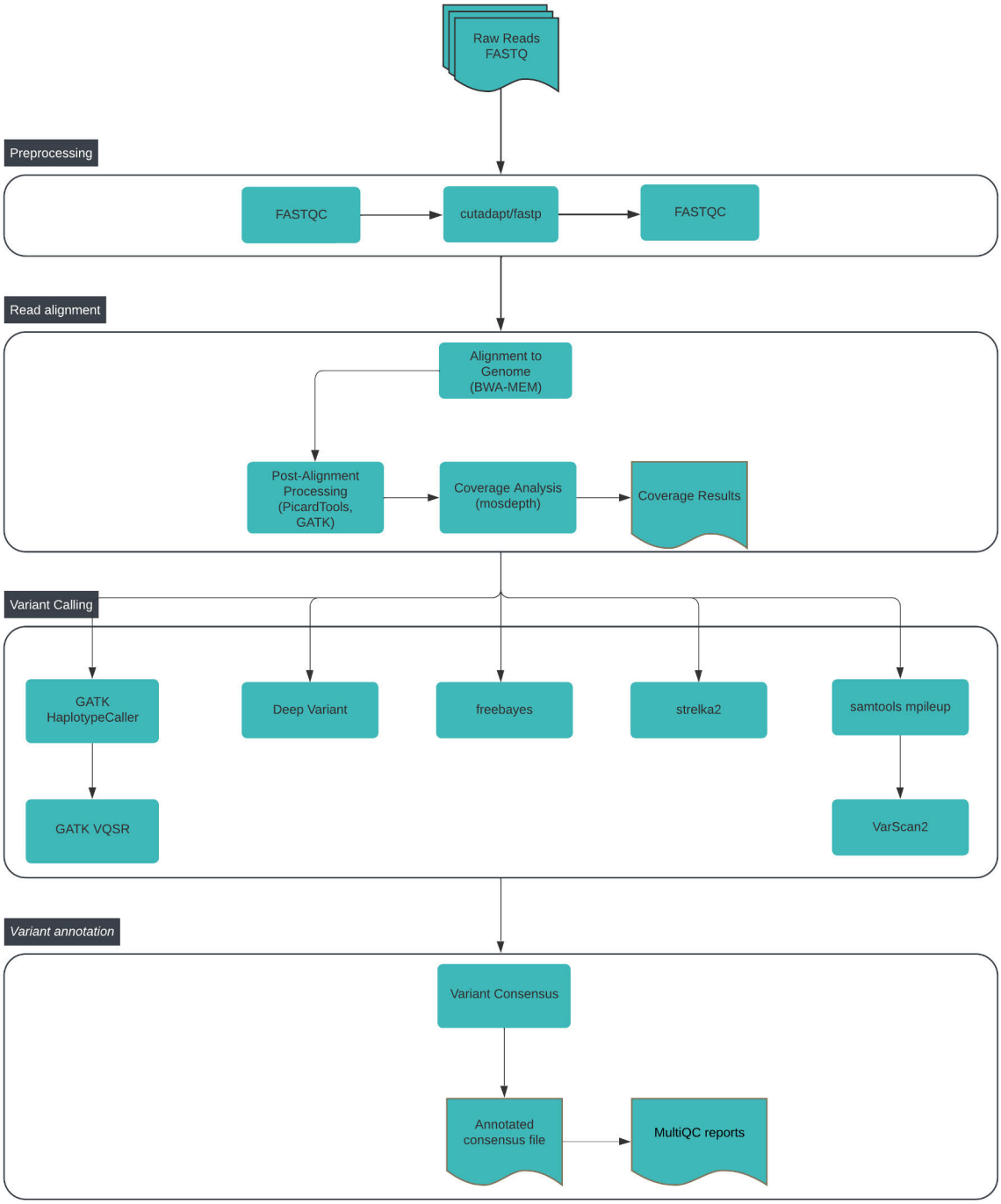


Figure 6. Overall architecture of the Kuura – WES and WGS analysis pipeline. The flowchart illustrates the four stages of the Kuura pipeline and the processes that takes place in each stage.

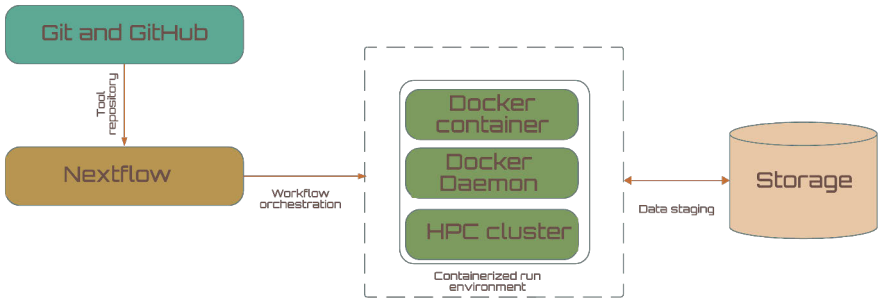


Figure 7. Overall architecture of the workflow for Kuura and Sampo. The architecture remains the same while the tools and methodology used for the analysis differ.

The pipeline was designed to be reproducible and easily deployable across different computational environments. An overview of the pipeline architecture is shown in Figure 7. The pipeline starts with raw reads in the form of FASTQ files and runs the whole analysis producing one annotated high-confidence VCF file per sample, together with a summary report describing the overall analysis.

All software dependencies required for the analysis are encapsulated within docker images to ensure consistent execution, while workflow orchestration is handled by nextflow, which manages job scheduling, docker container deployment, and scalability on HPC systems (DI Tommaso et al., 2017). As illustrated in Figure 7, the pipeline is deployed from a version-controlled GitHub repository and executed within Docker containers on an HPC environment via Docker daemons with outputs redirected to user-defined storage volumes. This architecture makes the pipeline easily deployable in different analysis environments and ensures reproducibility. Table 3 provides a detailed list of computational resources used by Kuura pipeline.

The Kuura pipeline was validated with gold standard datasets NA12878, NA24385, NA24631, and the complete precision and recall metrics for each individual variant caller are reported in Table 4. Metrics were computed using hap.py against GIAB v4.2.1 truth sets, restricted to the IDT exome capture regions defined by the evaluation BED file. Consequently, recall represents the proportion of benchmark variants correctly identified within the captured and callable regions, rather than the total variant count across the entire genome.

Consistent with previous benchmarking studies, DeepVariant achieved high precision as well as recall of around ~0.999 and ~0.990 respectively, while GATK HaplotypeCaller also showed strong performance, with precision of approximately ~0.996 and recall of ~0.958 (Poplin et al., 2018). FreeBayes demonstrated comparatively lower precision of ~0.945 but high recall of ~0.989, VarScan2 achieved balanced precision and recall of ~0.982 and ~0.972 despite identifying fewer total SNPs than other variant callers, whereas Strelka2 exhibited lower overall performance, with a precision of ~0.916 and recall of ~0.210.

Table 3. Summary of the tools and corresponding docker containers used at each stage of the analysis. Table recreated from Publication II (own work).

Stage	Tools used	Container
Quality control	FASTQC	utuprcagenetics/dnapipe:0.1
	Cutadapt/fastp	
	FASTQC	
Genome alignment and quality score recalibration	BWA-MEM	
	GATK MarkDuplicates	broadinstitute/gatk
	SAMtools	utuprcagenetics/dnapipe:0.1
	GATK BaseRecalibrator	broadinstitute/gatk
	GATK ApplyBQSR	
	Mosdepth	quay.io/biocontainers/mosdepth:0.2.4--he527e40_0
	bedtools	utuprcagenetics/dnapipe:0.1
Variant calling & variant recalibration	GATK HaplotypeCaller	broadinstitute/gatk
	GATK VariantRecalibrator	
	GATK ApplyVQSR	
	DeepVariant	google/deepvariant:1.4.0
	Strelka2	utuprcagenetics/dnapipe:0.1
	Freebayes	
	VarScan2	
Variant consensus & annotation	BCFtools	
	VEP	ensemblorg/ensembl-vep
Summary	MultiQC	utuprcagenetics/dnapipe:0.1

Table 4. Total number of SNPs identified, number of true positive SNPs, precision and recall values for different variant callers evaluated against gold-standard datasets (NA12878, NA24385 and NA24631). Precision and recall were calculated using hap.py against GIAB v4.2.1 truth sets, restricted to IDT exome capture region.

Variant Caller	NA12878				NA24385				NA24631			
	# of Variants	TRUTH TP	Precision	Recall	# of Variants	TRUTH TP	Precision	Recall	# of Variants	TRUTH TP	Precision	Recall
GATK Haplotype-caller	160713	23539	0.996	0.958	155973	24257	0.995	0.954	150884	23895	0.995	0.952
Deep Variant	261796	24310	0.999	0.990	252227	25107	0.999	0.988	247151	24795	0.999	0.988
Freebayes	353434	24289	0.945	0.989	342915	25083	0.938	0.987	339658	24787	0.948	0.987
Strelka2	333574	5159	0.916	0.210	317965	5453	0.908	0.215	307473	5755	0.908	0.229
VarScan2	95416	23886	0.982	0.972	94501	24674	0.978	0.971	92490	24413	0.983	0.972

Different variant callers identified overlapping but non-identical sets of variants, primarily due to the difference in their underlying algorithms and assumptions (O'Rawe et al., 2013). To account for this variability, Kuura applies a consensus-based strategy in which variants that are consistently identified by all five variant callers are prioritised as a high-confidence set. Similar multi-caller consensus strategies have shown to reduce false-positive calls and improve the reliability of germline variant detection (Cantarel et al., 2014; Chiara et al., 2018; Guo et al., 2015; Kanzi et al., 2020).

5.4 Sampo workflow and output (III)

Sampo is implemented as an end-to-end RNA-seq data analysis pipeline that starts with raw sequence reads in FASTQ format and ends with performing differential expression analysis, optionally including functional enrichment and visualization. The pipeline is developed using nextflow to ensure reproducibility, scalability and interoperability across computing environments (DI Tommaso et al., 2017). Core statistical analysis and visualization are performed using diffwrap, an in-house R package that wraps established methods from the edgeR and limma frameworks for differential expression and enrichment analysis (Ritchie et al., 2015; Robinson et al., 2009).

RNA-seq workflows vary widely in their architecture, tool selection, and analytical goals. As such, benchmarking across pipelines is inherently limited and often fails to reflect the nuances and flexibility that a specific pipeline offers. Therefore, we validated the correctness and completeness by re-analysing a publicly available RNA-seq dataset from the SRA - study SRP193095, consisting of 23 paired esophageal squamous cell carcinoma (ESCC) and adjacent normal tissue samples.

To enable direct comparison between the results generated by the Sampo pipeline and previously published analysis, gene expression values from the two independently processed dataset were harmonized. A custom batch correction was applied using the diffwrap package, which fits a generalized linear model to generate corrected pseudo-counts. To ensure numerical stability and variance stabilization, negative pseudo-counts were set to a minimum of 0.1 before computing normalized log-transformed counts per million (logCPM). Variance stabilization is used to reduce mean-variance dependency and to improve interpretability in further downstream analyses, particularly for lowly expressed genes (Anders & Huber, 2010; Love et al., 2014).

These variance stabilized pseudo counts are used for exploratory data analysis. We used principal component analysis (PCA) to identify major sources of variation and to assess consistency between the datasets, as PCA helps visualizing both technical as well as biological variability (Conesa et al., 2016).

As shown in Figure 8, the first principal component (PC1), accounted for 58.3% of the total variance and separated samples by data source, while the second principal component (PC2) explaining 21.5% of the variance reflects biological conditions. The mirrored clustering observed along PC1 indicates that analytical workflows represent a major source of variation; however, the consistent separation of groups along PC2 suggests that both pipelines preserve a comparable biological signal.

In addition to PCA, consistency between gene-level results produced by the Sampo pipeline and the results of the GEO study was quantified using the Jaccard index (JI). The JI measures similarity between two datasets as the ratio of their intersection (number of genes shared by both analyses) to their union (total number of unique genes identified across both) and is used to assess gene overlap (Real & Vargas, 1996). It is calculated as

$$J(GEO, SAMPO) = \frac{|P_{GEO} \cap P_{Sampo}|}{|P_{GEO} \cup P_{Sampo}|}$$

where, P_{GEO}represents genes identified as expressed in the GEO analysis and P_{Sampo} represents genes identified using the Sampo pipeline. The resulting JI values (Table 5) indicate a high degree of overlap between the two gene sets, supporting consistency between the analyses.

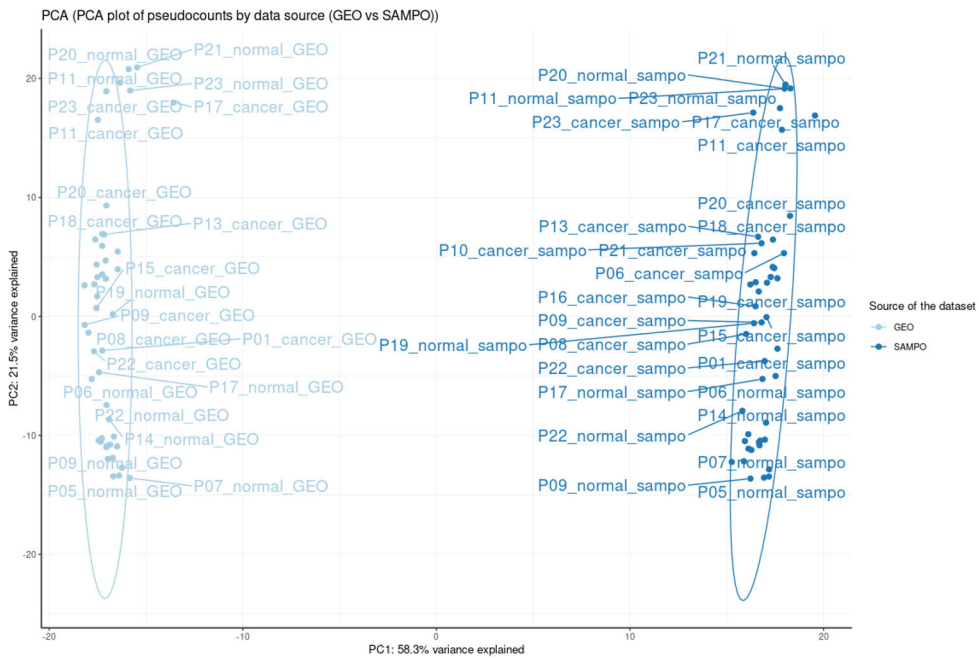


Figure 8. PCA of variance stabilized pseudo-counts from GEO and SAMPO analyses. PCA was performed on normalized logCPM values. The mirrored clustering pattern indicates the underlying biological signal is preserved despite pipeline specific technical variation.

A schematic overview of the Sampo pipeline is shown in Figure 9. The pipeline starts from raw FASTQ data and performs QC, read alignment, read quantification and DGE analysis in a modular manner. Each module of the pipeline is independently configurable, and outputs structured results into clearly defined directories under a common results root folder specified in the config files. The intermediate results such as trimmed reads, quality reports, BAM files, and count matrices are retained to support auditing and additional downstream analysis. The final output comprises DGE results, functional enrichment analysis and a comprehensive set of visualizations, including QC plots (multidimensional scaling (MDS) plots, PCA biplots, two- and three-dimensional PCA scatter plots, and hierarchical clustering dendrograms), as well as DEG plots (MA plots, volcano plots, p-value and false discovery rate (FDR) histograms and heatmaps, and also gene and sample correlation matrices).

Table 5. Jaccard similarity between SAMPO and GEO pipeline results for matched tumour and normal samples.

Sample	Cancer	Normal
P01	0.961	0.906
P02	0.949	0.941
P03	0.948	0.890
P04	0.958	0.917
P05	0.959	0.930
P06	0.951	0.961
P07	0.947	0.902
P08	0.955	0.944
P09	0.958	0.921
P10	0.947	0.941
P11	0.963	0.939
P12	0.959	0.945
P13	0.955	0.939
P14	0.965	0.924
P15	0.962	0.942
P16	0.963	0.933
P17	0.801	0.960
P18	0.948	0.938
P19	0.950	0.958
P20	0.958	0.949
P21	0.942	0.941
P22	0.952	0.946
P23	0.963	0.937

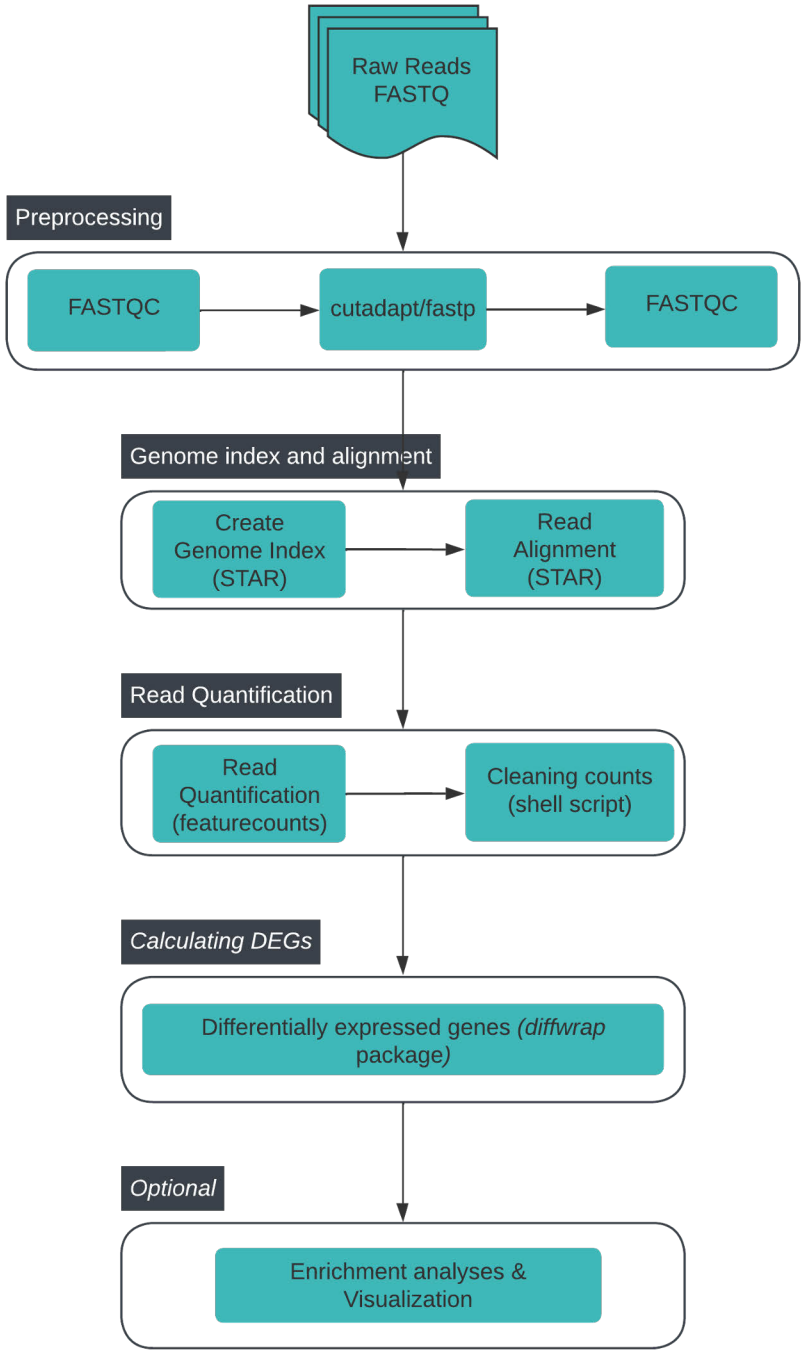


Figure 9. Overall architecture of the Sampo RNA-seq data analysis pipeline.

6 Discussion

The rise of NGS technologies has made it accessible to have a whole genome or exome sequenced along with their matching transcriptome. These data help study a wide range of biological features and conditions, including DGE, allele specific expression, fusion genes, differentially spliced genes, intron retention, SNVs, indels, CNVs, and SVs (Bell & Beck, 2009; Conesa et al., 2016). Collectively, studying these biological features support integrated analyses that link genetic variation to functional effect (Aguet et al., 2020; Fu et al., 2025; Stark et al., 2019).

Increase in sequencing throughput has come with a bottleneck in terms of computational analysis. The analysis of NGS data is a multi-step approach with numerous tools available at each stage and substantial variation in recommended practices (Conesa et al., 2016). The results of the analysis are influenced by factors such as experimental design, library preparation, read length, coverage and statistical analysis choices. Newer tools and pipelines are benchmarked against existing ones but benchmarking alone does not as address broader problems related to reproducibility, scalability and modularity (Ewels et al., 2020; Tello et al., 2019).

The work presented in this thesis focuses on the development of three interoperable tools: Kuura, Sampo and BioCPR, which emphasize modular design, reproducible execution and cross-modality integration. The tools presented here do not introduce new variant calling algorithms or statistical models. Instead, they are designed to provide practical, end-to-end analysis workflows that are transparent, customizable, and compatible across genomic and transcriptomic data types. The tools complement each other's output and at the same time, is independently deployable to carry out their own analysis. Collectively, the tool kits cover three broader areas of sequencing studies - SNV and indel identification from WES or WGS data, DEG identification from RNA-seq data and finally a tool to analyse and visualize DEG data. A flowchart depicting the overview of the analysis tools and their functional integration is shown in Figure 10.

In study I, BioCPR was developed as an analysis and visualization tool for differential gene expression data with focus on exploratory identification of co-expression patterns that can shed light on new gene-gene interactions and networks. The tool provides an interactive GUI implemented in R using the shiny framework,

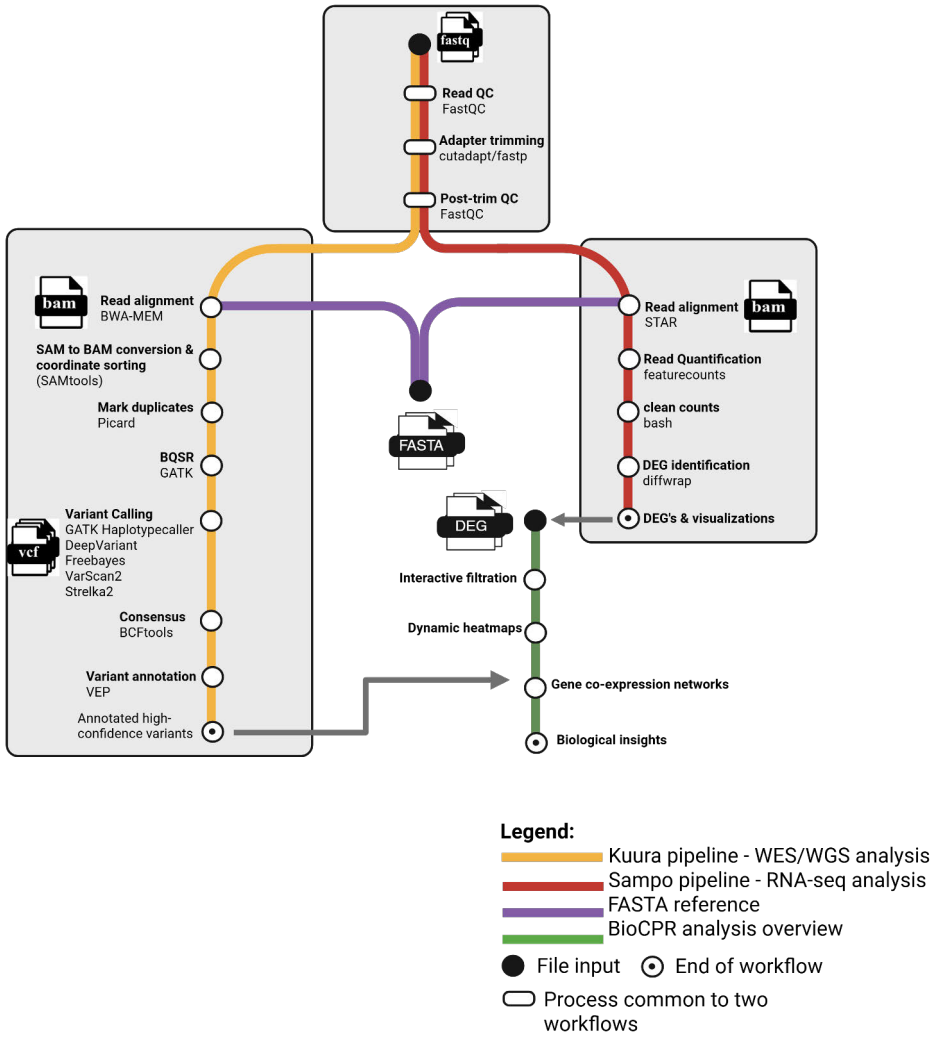


Figure 10. Flowchart showing the works carried out during this thesis. Kuura, Sampo and BioCPR, and how they complement each other for better understanding of sequencing data.

enabling users to explore correlation and clustering analysis without requiring extensive programming experience. A wide range of established R packages exist that support gene expression analysis, many relying on scripted workflows generating static visualizations. BioCPR complements these approaches by enabling dynamic filtering and interactive exploration of expression relationships, enabling hypothesis generation and interpretation rather than replacing established statistical methods.

Study II, addresses the limitations associated with a single variant caller for germline variant detection. There are multiple variant callers and pipelines to analyse raw sequencing data, each one using its own statistical testing for inferring variants

and its quality. However, no pipeline or tool is uniformly optimal for all kinds of genomic contexts, as differences in algorithms, statistical assumptions, and filtering strategies result in significantly different call sets which are only partially overlapping. For example, DeepVariant performs well in highly mapped regions but loses sensitivity in homopolymer tracts, whereas Freebayes and VarScan2, can identify low-allele-frequency events, yet are plagued with higher false positives in regions with extensive strand bias (Koboldt, 2020; Poplin et al., 2018; Sandmann et al., 2017). As a result, ensemble- and consensus-based strategies have been widely explored. Large-scale efforts such as the Genome in a Bottle (GIAB) Consortium and the Pan-Cancer Analysis of Whole Genomes (PCAWG) project have demonstrated that integrating calls from multiple methods can improve robustness and reduce false positives, particularly in difficult genomic regions (Campbell et al., 2020; Zook et al., 2014).

The Kuura pipeline builds directly on these principles by implementing a multi-caller consensus framework using five widely adopted variant callers - GATK HaplotypeCaller, DeepVariant, FreeBayes, Strelka2, and VarScan2. The whole pipeline is implemented using nextflow to create a scalable, automated workflow, while the analysis environment is isolated in a docker container to ensure seamless deployment and reproducibility across computing environments. Variant calls consistently identified by all five variant callers are regarded as high confidence calls and are annotated using Ensembl VEP. The variants identified by single callers are retained as low confidence calls for future reference. This method is useful in areas where precision is more important, like in a clinical setting where consistent and reliable methods are needed for identifying variants (Garrison & Marth, 2012; S. Kim et al., 2018; Koboldt et al., 2012; McKenna et al., 2010; McLaren et al., 2016; Poplin et al., 2018; G. A. Van der Auwera et al., 2013). Compared to other DNA-seq pipelines, Kuura is not designed to be feature-intensive; it contains only essential features and limits configuration complexity to achieve an end-to-end workflow that produces comprehensive yet reliable, ready-to-use outputs for downstream analysis. For experienced users, individual components or the entire workflow can be replaced or upgraded, making it easier to integrate new tools or updated best practices without restructuring the pipeline.

The Kuura analysis pipeline in study II focussed on DNA-sequencing, Sampo tackles the next challenge: How best to analyse and identify DEGs from RNA-seq. The Sampo pipeline is implemented as a modular RNA-seq pipeline that integrates standardized pre-processing with downstream statistical analysis via the diffwrap R package. Raw data processing and read quantification are handled within the nextflow framework, after which normalization, differential expression testing, enrichment analysis and visualization are performed in R. Differential expression outputs from Sampo are directly compatible with BioCPR, enabling downstream co-expression analysis and visualization without additional data transformation.

Collectively, Kuura, Sampo, and BioCPR constitute an integrated analysis ecosystem that supports coordinated genomic and transcriptomic analysis. For matched samples, high-confidence variants identified by Kuura can be examined alongside with the gene expression changes identified by Sampo, enabling integrative analyses such as expression quantitative trait locus (eQTL) testing and pathway-level interpretation (Aguet et al., 2020; Eisen et al., 1998; Ongen et al., 2016; Shabalin, 2012).

6.1 Study scope

The analysis framework defined here forms a complete analysis ecosystem. The closest match to this would be the nf-core analysis pipelines, a community effort to develop and maintain a set of analysis pipelines built using nextflow.

BioCPR is an analysis and visualization tool for DGE data. There are alternatives that have partial overlap with BioCPR functionalities, however at the time of development, there were no widely adopted, ready to use visualization tool using the shiny framework. For users proficient in programming the same results can be obtained by using existing R libraries.

The closest equivalent to Kuura and Sampo would be community-driven pipelines such as nf-core/sarek, nf-core/rnaseq, and nf-core/differentialabundance, which emphasize strict standardization, extensive feature coverage, and collective governance (Ewels et al., 2020; Langer et al., 2025). In contrast, Kuura and Sampo offer greater ease of customization for end users. Similarly, platforms such as CSC's Chipster and Galaxy provide user-friendly graphical environments that eliminates the need of command-line complexity and programming requirements, making sophisticated analyses accessible to a broader range of users (Abueg et al., 2024; Giardine et al., 2005).

Overall, the tools presented in this thesis should be considered complementary to the existing solutions rather than replacements. Their primary contribution lies in achieving an ecosystem that remains accessible and adaptable to diverse research needs.

6.2 Users and competencies

The ecosystem is primarily designed for computational biologists and bioinformaticians who are able to work with Linux environments, understand workflow management (Nextflow), and containerized execution (Docker or Apptainer).

BioCPR reduces the barrier for biologists needing interactive exploration as it provides an accessible Shiny interface for exploring co-expression patterns

interactively. For researchers with minimal command-line experience, alternatives such as Galaxy or Chipster (For RNA-seq data) offer customisable analysis options.

6.3 Computational environments

Kuura and Sampo pipelines execute efficiently on HPC clusters. Since both use nextflow-based pipeline deployment and containerization with docker, they can be easily integrated with workload management tools like SLURM and scaled for large cohorts. Local execution is feasible for small datasets but would be severely limited by storage and computing requirements.

6.4 Study limitations

In study I, BioCPR was developed and applied as an interactive visualization tool to analyse and explore co-expression networks in DGE data. While correlation-based methods are widely used and useful in identifying co-regulatory gene networks, they are also subject to influence by confounding factors such as unaccounted batch effects that can be confused with biological variability (Chen et al., 2011; Leek et al., 2010). Additionally, Pearson correlation assumes linear relationships between variables and is most suited for normally distributed data. It is also particularly sensitive to outliers, which can disproportionately influence the correlation estimate.

The output of BioCPR depends on the upstream input data from RNA-seq pipelines (such as Sampo), and any biases or limitations in the upstream process affect the validity of co-expression results and subsequent biological interpretation.

Study II focused on the development of the Kuura pipeline and its application in variant detection and prioritization from WES and WGS data. The pipeline uses a consensus-based approach that integrates results from five variant callers to produce a high-confidence variant set. This approach improves precision and reduces false positives but also risks excluding true variants detected only by a subset of variant callers. In addition, the detection of variants is dependent on factors such as sequencing depth and uniform coverage, to ensure reliable variant identification. Variant annotations depend on multiple external curated databases and gaps in annotation databases could also lead to a potential variant being overlooked.

In study III, the Sampo pipeline was developed for end-to-end RNA-seq analysis, from raw reads to the identification of DEGs. Sampo integrates established tools for quality control, alignment, quantification and DEG identification, yet the statistical power of the analysis is highly influenced by factors such as sample size, replicates and sequencing depth. Small sample sizes reduce the reliability of variance estimates, thereby increasing the number of false negatives. Since, Sampo relies on reference genome annotation before quantifying reads, there are possibilities of

missing novel transcripts or unannotated splicing events (Frankish et al., 2019; Schurch et al., 2016).

Collectively, these limitations underscore the need to account for both technical and biological factors not only during result interpretation, but also during experiment planning. In addition to the abovementioned limitations, the bioinformatics tools and reference databases used in the study receive constant updates and evolve over time. Hence, it is crucial to maintain and update the analysis pipelines if they are to be relevant over time.

6.5 Future prospects

Advances in sequencing technologies and analysis tools are continuously reshaping the field of genomics and transcriptomics. In particular, third and fourth generation sequencing technologies are transforming both variant discovery and transcriptome profiling by enabling full-length isoform reconstruction, improved structural variant detection, and direct measurement of epigenetic modifications (Moustakli et al., 2025; Santucci et al., 2024). As datasets continue to grow in size and complexity, so does the need for scalable, adaptable and reproducible analysis workflows (Wratten et al., 2021). The workflows developed in this thesis are modular in nature, making them adaptable for newer emerging technologies. However, their long-term usability depends on continuous updates to include advances in data formats, analytical algorithms and tools.

BioCPR can benefit from including pre-processing directly built into the app, so that users can directly upload the output from DGE analysis without any additional preprocessing, thereby reducing the need to format the input data.

A key area for development in the Kuura pipeline could be in variant annotation and prioritization and integration of long-read-specific aligners. Since Kuura already integrates multiple variant callers and annotation through Ensembl VEP, there could be a strategy to prioritize annotated variants into different tiers using approaches similar to the one described by Rantapero et al (2020) (Rantapero et al., 2020), which would enable Kuura to stratify variants into clinically and biologically meaningful categories. Tier 1 could include pathogenic or likely pathogenic variants supported by ClinVar or curated resources, Tier 2 variants of uncertain significance (VUS) with predicted functional impact, and Tier 3 benign variants, further filtered by high allele frequency or lack of predicted relevance. This systematic prioritization could make downstream validation more effective.

The Sampo pipeline could be expanded to integrate alternative splicing detection and, for matched samples, combining variant information from the Kuura pipeline and DGE results from Sampo to perform expression quantitative trait loci (eQTL) analysis, thereby directly mapping genetic variants to transcriptional changes (Aguet et al., 2020).

7 Conclusions

This thesis presents the development and application of three modular bioinformatics pipelines designed for comprehensive analysis and visualization of genomic and transcriptomic data.

BioCPR provides an interactive platform for analysing and visualization of gene co-expression networks from DGE data. Kuura enables automated variant discovery and annotation from WES and WGS data. Sampo automates RNA-seq analysis from raw data to identifying DEGs and visualization. Together, these tools form an integrated framework capable of both independent analysis of sequencing data as well as multi-omics integration, while their modular structure ensures adaptability to rapidly evolving analytical approaches.

Acknowledgements

This work was carried out at the Genetic cancer predisposition research group, Institute of Biomedicine, University of Turku. I would like to thank Professor Sari Mäkelä for providing excellent research facilities.

I am deeply grateful to my supervisors, Professor Johanna Schleutker, Docent Csilla Sipeky and Dr. Vidal Fey, for their guidance and support throughout this work. Johanna, thank you for giving me the opportunity to pursue my PhD in your group, for fostering a supportive and inspiring working environment, and for trusting me to work independently while consistently offering guidance. Csilla, thank you for your unwavering support and motivation throughout this work, for encouraging me during challenging times, and for pushing me to become the better version of myself. Vidal, thank you for always being there to guide me, helping me build a strong foundation in bioinformatics and for giving me confidence in my work. It has been a privilege to have you as my supervisors.

I would also like to express my deepest gratitude to Docent Esa Pitkänen and Associate Professor Valerio Izzi, for taking the time to review my thesis and for their constructive comments, which have greatly improved this work. I also want to thank Professor Veli Mäkinen for accepting the role of my esteemed opponent, I truly appreciate the opportunity to discuss my research with you.

I would also like to thank my thesis committee members, Professor Merja Heinäniemi, Docent Maria Sundvall and late Docent Johanna Tuomela, for their valuable advice and practical suggestions during this thesis work.

I am greatly obliged to the Turku Doctoral Programme of Molecular Medicine (TuDMM), its former and current directors, Professor Kati Elima and Professor Noora Kotaja, and its coordinators Dr. Eeva Valve and Dr. Verna Louhivuori. I am grateful to TuDMM for financial support during the final phase of my work, for funding conference travel and for organizing many scientific and social events. I also acknowledge financial support from Turku University Foundation, Yrjö Jahnsson Foundation and the Finnish Cancer Society.

I would also like to extend my thanks to Chief Academic Officer Outi Irjala and interim Chief Academic Officer Anni Halonen for their support and guidance during this work.

I would like to express my gratitude to Heli Törmänen, Kristiina Nuutila, Päivi Aalto, Anja Similä, Tiina Haarala for their invaluable administrative support and guidance.

I warmly thank all my co-authors, Henri Sara, Samuel Heron and Venkat Subramaniam Rathinakannan, for their valuable contribution in original publications included in this thesis.

I had the privilege of working with talented colleagues, and I thank everyone for making my time in Johanna's group memorable. Thank you for your friendship and discussions on science as well as everyday life, it is not often that one comes across such friendly and talented colleagues. I am grateful to all the past and present members of Schleutker lab, Aliisa Takala, Amanda Tursi, Dr. Christoffer Löf, Docent Csilla Sipeky, Dr. Elina Kaikkonen, Dr. Gudrun Wahlström, Jukka Karhu, Mohammed Ennejmy, Nasrin Sultana, Dr. Neha Goel, Dr. Olli Metsälä, Dr. Samuel Heron, Tuuli Levänen, Venkat Subramaniam Rathinakannan, Dr. Vidal Fey and Viivi Kurkilahti.

I would also like to express my gratitude to Professor Matti Nykter, Dr. Tommi Rantapero, and Dr. Ebrahim Afyounian for their support and guidance when needed.

I also thank my colleagues on the 5th floor of Medisiina D for creating a welcoming and supportive atmosphere, especially Priyadharshini Parimelazhagan Santhi, Saiganesh Sriraman and Syeda Afshan.

Life outside work would not have been not as colourful without my friends. Thank you, Ponnuswamy, Simo, Pihla, Jouko, Tiina, Katariina, Christian, Vimal, Prakirth, Xu, Natasha, Alvaro, Liisa, Pasi, Samuli, Santeri, Hanna, Gopinath, Carlton, Razowan, Nirmal, Dhayakumar, Senthil and Hariharan. Thank you all for the great company and for sharing so many memorable moments.

Thank you, Imran and Venkat, for being great friends and for always having my back. Life in Turku would have been boring without you two. Even though we live in different places now, looking back on our time together always makes me smile.

Through all the ups and downs of the PhD, dancing was a break I always looked forward to. Jessica, Pierre, Karin, Giancarlo, Otto, Anna, Saara, Salla and Joonas, thank you for the fun times and laughs on the dance floor.

Ashik, Sakeena, Prabhakaran, Guru, Venkey, Arun, Sriram, Dinesh, Mohan, Saravanan, Prabhavathy and Vaishnavi, we may not talk as frequently as we used to, but whenever we do, it feels like going back in time. You have helped shape who I am today, and I have countless memories to look back on and cherish. Thank you all.

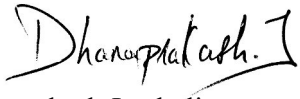
I would like to thank my parents (late Mr. Jambulingam Thangavelu and Mrs. Sundari Parthasarathy) and siblings (Sathish, Dinesh and Sugesh) for their love and support throughout my life. They were always there whenever I needed them, this PhD would not have been possible without their love and support. I would also like

to thank my aunts, uncles, grandparents, cousins, in-laws and all my relatives for their love and for including me in their thoughts.

Finally, and most importantly, my wife Martina, you always believed in me and have been a pillar of my strength throughout this journey. You were so patient and encouraging, I am not sure I would have managed to get here without your support. Our dear son Iniyan, your smiles, laughter and hugs make me forget all my worries and is a reminder of what matters most. I am deeply grateful to have you both in my life.

I am fortunate to have wonderful friends, with whom I have made unforgettable memories. Thank you all for your support.

Turku, February 2026



Dhanaprakash Jambulingam

References

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., Auman, J. T., Balasundaram, M., Balu, S., Barbieri, C. E., Bauer, T., Benz, C. C., Bergeron, A., Beroukhim, R., Berrios, M., … Zmuda, E. (2015). The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, *163*(4), 1011–1025. https://doi.org/10.1016/J.CELL.2015.10.025/ASSET/0061A3ED-8888-44A7-A066-2251A102BA70/MAIN.ASSETS/GR6.JPG

Abueg, L. A. L., Afgan, E., Allart, O., Awan, A. H., Bacon, W. A., Baker, D., Bassetti, M., Batut, B., Bernt, M., Blankenberg, D., Bombarely, A., Bretaudeau, A., Bromhead, C. J., Burke, M. L., Capon, P. K., Čech, M., Chavero-Díez, M., Chilton, J. M., Collins, T. J., … Zoabi, R. (2024). The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*, *52*(W1), W83. https://doi.org/10.1093/NAR/GKAE410

Agudelo-Romero, P., Conradie, T., Caparros-Martin, J. A., Martino, D. J., Kicic, A., Stick, S. M., Hakkaart, C., & Sharma, A. (2025). Advancing bioinformatics capacity through Nextflow and nf-core: lessons from an early-to mid-career researchers–focused program at The Kids Research Institute Australia. *Frontiers in Bioinformatics*, *5*, 1610015. https://doi.org/10.3389/FBINF.2025.1610015/TEXT

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Flynn, E. D., Parsana, P., Fresard, L., Gamazon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., … Volpi, S. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/SCIENCE.AAZ1776/SUPPL_FILE/AAZ1776_TABLESS10-S16.XLSX

Akintunde, O., Tucker, T., & Carabetta, V. J. (2025). The Evolution of Next-Generation Sequencing Technologies. *Methods in Molecular Biology*, *2866*, 3–29. https://doi.org/10.1007/978-1-0716-4192-7_1,

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics 2011 12:5*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology 2010 11:10*, *11*(10), R106-. https://doi.org/10.1186/GB-2010-11-10-R106

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012). *FastQC: a quality control tool for high throughput sequence data*. Babraham Institute.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, *53*(4), 50–58. https://doi.org/10.1145/1721654.1721672

Bagger, F. O., Borgwardt, L., Jespersen, A. S., Hansen, A. R., Bertelsen, B., Kodama, M., & Nielsen, F. C. (2024). Whole genome sequencing in clinical practice. *BMC Medical Genomics 2024 17:1*, *17*(1), 1–16. https://doi.org/10.1186/S12920-024-01795-W

Balachandran, P., & Beck, C. R. (2020). Structural variant identification and characterization. *Chromosome Research : An International Journal on the Molecular, Supramolecular and*

Evolutionary Aspects of Chromosome Biology, *28*(1), 31. https://doi.org/10.1007/S10577-019-09623-Z

Bansal, V. (2017). A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC Bioinformatics*, *18*(3), 113–123. https://doi.org/10.1186/S12859-017-1471-9/FIGURES/5

Barba, M., Czosnek, H., & Hadidi, A. (2013). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, *6*(1), 106–136. https://doi.org/10.3390/V6010106,

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., Kostareva, A. A., Glotov, O. S., & Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific Reports 2020 10:1*, *10*(1), 2057-. https://doi.org/10.1038/s41598-020-59026-y

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J. L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5473–5478. https://doi.org/10.1073/PNAS.1418631112/SUPPL_FILE/PNAS.1418631112.SD02.XLSX

Bell, C. G., & Beck, S. (2009). Advances in the identification and analysis of allele-specific expression. *Genome Medicine*, *1*(5), 1–5. https://doi.org/10.1186/GM56/FIGURES/1

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/NATURE07517;KWRD=SCIENCE

Berdan, E. L., Barton, N. H., Butlin, R., Charlesworth, B., Faria, R., Fragata, I., Gilbert, K. J., Jay, P., Kapun, M., Lotterhos, K. E., Mérot, C., Durmaz Mitchell, E., Pascual, M., Peichel, C. L., Rafajlović, M., Westram, A. M., Schaeffer, S. W., Johannesson, K., & Flatt, T. (2023). How chromosomal inversions reorient the evolutionary process. *Journal of Evolutionary Biology*, *36*(12), 1761–1782. https://doi.org/10.1111/JEB.14242

Bergin, C. J., Silva, A. M. da, Benoit, Y. D., Bergin, C. J., Silva, A. M. da, & Benoit, Y. D. (2023). Where to Draw the LINE—Are Retrotransposable Elements Here to Stay? *Cancers 2023, Vol. 15, 15*(16). https://doi.org/10.3390/CANCERS15164119

Bhérer, C., Eveleigh, R., Trajanoska, K., St-Cyr, J., Paccard, A., Nadukkalam Ravindran, P., Caron, E., Bader Asbah, N., McClelland, P., Wei, C., Baumgartner, I., Schindewolf, M., Döring, Y., Perley, D., Lefebvre, F., Lepage, P., Bourgey, M., Bourque, G., Ragoussis, J., … Taliun, D. (2024). A cost-effective sequencing method for genetic studies combining high-depth whole exome and low-depth whole genome. *Npj Genomic Medicine 2024 9:1*, *9*(1), 8-. https://doi.org/10.1038/s41525-024-00390-3

Bianchi, V., Ceol, A., Ogier, A. G. E., De Pretis, S., Galeota, E., Kishore, K., Bora, P., Croci, O., Campaner, S., Amati, B., Morelli, M. J., & Pelizzola, M. (2016). Integrated systems for NGS data management and analysis: Open issues and available solutions. *Frontiers in Genetics*, *7*(MAY), 188599. https://doi.org/10.3389/FGENE.2016.00075/BIBTEX

Björn, N., Pradhananga, S., … B. S.-J. of N., & 2018, undefined. (2018). Comparison of variant calls from whole genome and whole exome sequencing data using matched samples. *Diva-Portal.OrgN Björn, S Pradhananga, B Sigurgeirsson, J Lundberg, H Green, P SahlénJournal of Next Generation Sequencing & Applications, 2018•diva-Portal.Org*, *5*, 1. https://doi.org/10.4172/2469-9853.1000154

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A Quick Introduction to Version Control with Git and GitHub. *PLOS Computational Biology*, *12*(1), e1004668. https://doi.org/10.1371/JOURNAL.PCBI.1004668

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Brancato, D., Treccarichi, S., Bruno, F., Coniglio, E., Vinci, M., Saccone, S., Calì, F., Federico, C., Brancato, D., Treccarichi, S., Bruno, F., Coniglio, E., Vinci, M., Saccone, S., Calì, F., & Federico, C. (2025). NGS Approaches in Clinical Diagnostics: From Workflow to Disease-Specific Applications. *International Journal of Molecular Sciences 2025, Vol. 26, 26*(19). https://doi.org/10.3390/IJMS26199597

Breschi, A., Gingeras, T. R., & Guigó, R. (2017). Comparative transcriptomics in human and mouse. *Nature Reviews Genetics*, *18*(7), 425–440. https://doi.org/10.1038/NRG.2017.19,

Brlek, P., Bulić, L., Bračić, M., Projić, P., Škaro, V., Shah, N., Shah, P., & Primorac, D. (2024). Implementing Whole Genome Sequencing (WGS) in Clinical Practice: Advantages, Challenges, and Future Perspectives. *Cells*, *13*(6). https://doi.org/10.3390/CELLS13060504,

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature 2020 578:7793*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Canoy, R. J., Shmakova, A., Karpukhina, A., Shepelev, M., Germini, D., & Vassetzky, Y. (2022). Factors That Affect the Formation of Chromosomal Translocations in Cells. *Cancers*, *14*(20), 5110. https://doi.org/10.3390/CANCERS14205110

Cantarel, B. L., Weaver, D., McNeill, N., Zhang, J., Mackey, A. J., & Reese, J. (2014). BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-104

Cerami, E., Gao, J., Gogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Chaushevska, M., Alapont-Celaya, K., Schack, A. K., Krych, L., Garrido Navas, M. C., Krithara, A., & Madjarov, G. (2025). Get ready for short tandem repeats analysis using long reads-the challenges and the state of the art. *Frontiers in Genetics*, *16*, 1610026. https://doi.org/10.3389/FGENE.2025.1610026/FULL

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE*, *6*(2), e17238. https://doi.org/10.1371/JOURNAL.PONE.0017238

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i890. https://doi.org/10.1093/BIOINFORMATICS/BTY560

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., & Chen, Q. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, *7*(1), 1–6. https://doi.org/10.1093/GIGASCIENCE/GIX120

Cheng, X., Goktas, M. T., Williamson, L. M., Krzywinski, M., Mulder, D. T., Swanson, L., Slind, J., Sihvonen, J., Chow, C. R., Carr, A., Bosdet, I., Tucker, T., Young, S., Moore, R., Mungall, K. L., Yip, S., & Jones, S. J. M. (2024). Enhancing clinical genomic accuracy with panelGC: a novel metric and tool for quantifying and monitoring GC biases in hybridization capture panel sequencing. *Briefings in Bioinformatics*, *25*(5). https://doi.org/10.1093/BIB/BBAE442

Chiara, M., Gioiosa, S., Chillemi, G., D'Antonio, M., Flati, T., Picardi, E., Zambelli, F., Horner, D. S., Pesole, G., & Castrignanò, T. (2018). CoVaCS: a consensus variant calling system. *BMC Genomics*, *19*(1), 120. https://doi.org/10.1186/S12864-018-4508-1

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., & Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National*

Academy of Sciences of the United States of America, *106*(45), 19096–19101. https://doi.org/10.1073/PNAS.0910672106;WEBSITE:WEBSITE:PNAS-SITE;ISSUE:ISSUE:DOI

Choo, Z. N., Behr, J. M., Deshpande, A., Hadi, K., Yao, X., Tian, H., Takai, K., Zakusilo, G., Rosiene, J., Da Cruz Paula, A., Weigelt, B., Setton, J., Riaz, N., Powell, S. N., Busam, K., Shoushtari, A. N., Ariyan, C., Reis-Filho, J., de Lange, T., & Imieliński, M. (2023). Most large structural variants in cancer genomes can be detected without long reads. *Nature Genetics*, *55*(12), 2139–2148. https://doi.org/10.1038/S41588-023-01540-6;SUBJMETA=114,208,212,2302,2785,631;KWRD=COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,GENOME+ASSEMBLY+ALGORITHMS,GENOMICS

Cinaglia, P., & Cannataro, M. (2025). fDESI: An open-source web application for visual bioinformatics pipeline design. *Neurocomputing*, *647*, 130582. https://doi.org/10.1016/J.NEUCOM.2025.130582

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics 2010 11:6*, *11*(6), 415–425. https://doi.org/10.1038/nrg2779

Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology*, *15*(9), e2003243. https://doi.org/10.1371/JOURNAL.PBIO.2003243

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/NAR/GKP1137

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature 2020 581:7809*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 1–19. https://doi.org/10.1186/S13059-016-0881-8/TABLES/1

Conroy, J. M., Kolda, T. G., O'leary, D. P., & O'leary, T. J. (2000). Chromosome Identification Using Hidden Markov Models: Comparison with Neural Networks, Singular Value Decomposition, Principal Components Analysis, and Fisher Discriminant Analysis. *Laboratory Investigation*, *80*(11), 1629–1641. https://doi.org/10.1038/LABINVEST.3780173

Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., & Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports*, *10*(1), 1–15. https://doi.org/10.1038/S41598-020-76881-X;SUBJMETA=114,199,208,631,67,68;KWRD=CANCER+GENETICS,COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,GENE+EXPRESSION

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews. Genetics*, *10*(10), 691. https://doi.org/10.1038/NRG2640

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature 1970 227:5258*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

Crick, F. H. C. (1966). Codon—anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, *19*(2), 548–555. https://doi.org/10.1016/S0022-2836(66)80022-0

Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General Nature of the Genetic Code for Proteins. *Nature 1961 192:4809*, *192*(4809), 1227–1232. https://doi.org/10.1038/1921227a0

Custódio, N., Savisaar, R., Carvalho, C., Bak-Gordon, P., Ribeiro, M. I., Tavares, J., Nunes, P. B., Peixoto, A., Pinto, C., Escudeiro, C., Teixeira, M. R., & Carmo-Fonseca, M. (2022). Expression Profiling in Ovarian Cancer Reveals Coordinated Regulation of BRCA1/2 and Homologous

Recombination Genes. *Biomedicines*, *10*(2). https://doi.org/10.3390/BIOMEDICINES10020199/S1

Dababneh, S. F., Babini, H., Jiménez-Sábado, V., Teves, S. S., Kim, K. H., & Tibbits, G. F. (2025). Dissecting cardiovascular disease-associated noncoding genetic variants using human iPSC models. *Stem Cell Reports*, *20*(4), 102467. https://doi.org/10.1016/J.STEMCR.2025.102467

Davis-Turak, J., Courtney, S. M., Hazard, E. S., Glen, W. B., da Silveira, W. A., Wesselman, T., Harbin, L. P., Wolf, B. J., Chung, D., & Hardiman, G. (2017). Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Review of Molecular Diagnostics*, *17*(3), 225. https://doi.org/10.1080/14737159.2017.1282822

Definition of germline variant - NCI Dictionary of Cancer Terms - NCI. (n.d.). Retrieved September 1, 2025, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/germline-variant

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. https://doi.org/10.1038/NG.806;SUBJMETA=144,208,2254,2489,514,631;KWRD=DISEASE+GENETICS,NEXT-GENERATION+SEQUENCING

Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P. Łabaj, P., & Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, *14*, 997383. https://doi.org/10.3389/FGENE.2023.997383/FULL

DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/NBT.3820;SUBJMETA

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635

Douglas, A. G. L., & Wood, M. J. A. (2011). RNA splicing: disease and therapy. *Briefings in Functional Genomics*, *10*(3), 151–164. https://doi.org/10.1093/BFGP/ELR020

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868. https://doi.org/10.1073/PNAS.95.25.14863

Elshambakey, M., Maiyza, A. I., Kashkoush, M. S., Fathy, G. M., & Hassan, H. A. (2024). The Egyptian national HPC grid (EN-HPCG): open-source Slurm implementation from cluster to grid approach. *The Journal of Supercomputing 2024 80:12*, *80*(12), 16795–16823. https://doi.org/10.1007/S11227-024-06041-9

Esplin, E. D., Oei, L., & Snyder, M. P. (2014). Personalized sequencing and the future of medicine: Discovery, diagnosis and defeat of disease. *Pharmacogenomics*, *15*(14), 1771–1790. https://doi.org/10.2217/PGS.14.117,

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/S41587-020-0439-X;SUBJMETA=114,631,648,706;KWRD=COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,SCIENTIFIC+COMMUNITY

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354

Finnegan, D. J. (2012). Retrotransposons. *Current Biology*, *22*(11), R432–R437. https://doi.org/10.1016/j.cub.2012.04.025

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/NAR/GKY955

Freeman, K., Taylor, D., Dinnes, J., Clark, C. C. A., Kander, I., Scandrett, K., Chockalingam, S., Dracup, N., Court, R., Butt, F., Visintin, C., Bonham, J. R., Elliman, D., Shortland, G., Mackie, A., Miedzybrodzka, Z. H., Morgan, S. M., Boardman, F. K., Takwoingi, Y., … Taylor-Phillips, S. (2025). Challenges in evaluating whole genome sequencing for newborn screening: series of systematic reviews and roadmap for evidence generation for policy advisers. *BMJ Medicine*, *4*(1), 1726. https://doi.org/10.1136/BMJMED-2025-001726

Fu, J., Zanotelli, V. R. T., Howald, C., Chammartin, N., Kolpakov, I., Xenarios, I., Froese, D. S., Wollscheid, B., Pedrioli, P. G. A., & Goetze, S. (2025). A Multi-Omics Framework for Decoding Disease Mechanisms: Insights From Methylmalonic Aciduria. *Molecular & Cellular Proteomics : MCP*, *24*(7), 100998. https://doi.org/10.1016/J.MCPRO.2025.100998

Garcia, M., Juhos, S., Larsson, M., Olason, P. I., Martin, M., Eisfeldt, J., DiLorenzo, S., Sandgren, J., Díaz De Ståhl, T., Ewels, P., Wirta, V., Nistér, M., Käller, M., & Nystedt, B. (2020). Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research 2020 9:63*, *9*, 63. https://doi.org/10.12688/f1000research.16665.2

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv*. https://arxiv.org/abs/1207.3907

Genomic Variation. (n.d.). Retrieved August 29, 2025, from https://www.genome.gov/genetics-glossary/Genomic-Variation

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., & Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, *15*(10), 1451–1455. https://doi.org/10.1101/GR.4086505

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human Copy Number Variation and Complex Genetic Disease. *Annual Review of Genetics*, *45*, 203. https://doi.org/10.1146/ANNUREV-GENET-102209-163544

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gotea, V., Gartner, J. J., Qutob, N., Elnitski, L., & Samuels, Y. (2015). The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment Cell & Melanoma Research*, *28*(6), 673. https://doi.org/10.1111/PCMR.12413

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., … Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, *446*(7132), 153. https://doi.org/10.1038/NATURE05610

Groot, D. de, Spanjaard, A., Hogenbirk, M. A., Jacobs, H., Groot, D. de, Spanjaard, A., Hogenbirk, M. A., & Jacobs, H. (2023). Chromosomal Rearrangements and Chromothripsis: The Alternative End Generation Model. *International Journal of Molecular Sciences 2023, Vol. 24, 24*(1). https://doi.org/10.3390/IJMS24010794

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A., & Taylor, J. (2018). Practical Computational Reproducibility in the Life Sciences. *Cell Systems*, *6*(6), 631. https://doi.org/10.1016/J.CELS.2018.03.014

Guo, C., McDowell, I. C., Nodzenski, M., Scholtens, D. M., Allen, A. S., Lowe, W. L., & Reddy, T. E. (2017). Transversions have larger regulatory effects than transitions. *BMC Genomics*, *18*(1), 1. https://doi.org/10.1186/S12864-017-3785-4/FIGURES/4

Guo, Y., Ding, X., Shen, Y., Lyon, G. J., & Wang, K. (2015). SeqMule: Automated pipeline for analysis of human exome/genome sequencing data. *Scientific Reports*, *5*(1), 1–10. https://doi.org/10.1038/SREP14283;SUBJMETA

Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., Gunnarsson, B., Oddsson, A., Halldorsson, G. H., Zink, F., Gudjonsson, S. A., Frigge, M. L., Thorleifsson, G., Sigurdsson, A., Stacey, S. N., Sulem, P., Masson, G., Helgason, A., Gudbjartsson, D. F., … Stefansson, K. (2019). Human genetics: Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, *363*(6425). https://doi.org/10.1126/SCIENCE.AAU1043;PAGE:STRING:ARTICLE/CHAPTER

Hancks, D. C. (2018). A Role for Retrotransposons in Chromothripsis. *Methods in Molecular Biology (Clifton, N.J.)*, *1769*, 169. https://doi.org/10.1007/978-1-4939-7780-2_11

Haseltine, W. A., Patarca, R., Haseltine, W. A., & Patarca, R. (2024). The RNA Revolution in the Central Molecular Biology Dogma Evolution. *International Journal of Molecular Sciences 2024, Vol. 25, 25*(23). https://doi.org/10.3390/IJMS252312695

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(1), 1–25. https://doi.org/10.1186/1471-2105-14-184/TABLES/4

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1. https://doi.org/10.1016/J.YGENO.2015.11.003

Hoernes, T. P., Faserl, K., Juen, M. A., Kremser, J., Gasser, C., Fuchs, E., Shi, X., Siewert, A., Lindner, H., Kreutz, C., Micura, R., Joseph, S., Höbartner, C., Westhof, E., Hüttenhofer, A., & Erlacher, M. D. (2018). Translation of non-standard codon nucleotides reveals minimal requirements for codon-anticodon interactions. *Nature Communications 2018 9:1*, *9*(1), 4865-. https://doi.org/10.1038/s41467-018-07321-8

Hofert, M. (2024). High Performance Computing Cluster Setup: A Tutorial. *Journal of Data Science*, *0*(0), 1–22. https://doi.org/10.6339/24-JDS1159

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, *147*(3664), 1462–1465. https://doi.org/10.1126/SCIENCE.147.3664.1462,

Jiang, W., Chen, L., Girgenti, M. J., & Zhao, H. (2024). Tuning parameters for polygenic risk score methods using GWAS summary statistics from training data. *Nature Communications 2024 15:1*, *15*(1), 24-. https://doi.org/10.1038/s41467-023-44009-0

Jou, W. M., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, *237*(5350), 82–88. https://doi.org/10.1038/237082A0;KWRD=SCIENCE

Kanzi, A. M., San, J. E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., & de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Frontiers in Genetics*, *11*, 544162. https://doi.org/10.3389/FGENE.2020.544162/FULL

Kar, S. P., Tyrer, J. P., Li, Q., Lawrenson, K., Aben, K. K. H., Anton-Culver, H., Antonenkova, N., Chenevix-Trench, G., Baker, H., Bandera, E. V., Bean, Y. T., Beckmann, M. W., Berchuck, A., Bisogna, M., Bjørge, L., Bogdanova, N., Brinton, L., Brooks-Wilson, A., Butzow, R., … Pharoah, P. D. P. (2015). Network-based integration of GWAS and gene expression identifies a HOX-centric network associated with serous ovarian cancer risk. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *24*(10), 1574. https://doi.org/10.1158/1055-9965.EPI-14-1270

Karakoyun, H. K., Sayar, C., & Yararbaş, K. (2023). Challenges in clinical interpretation of next-generation sequencing data: Advantages and Pitfalls. *Results in Engineering*, *20*, 101421. https://doi.org/10.1016/J.RINENG.2023.101421

Karamysheva, T. V., Gayner, T. A., Elisaphenko, E. A., Trifonov, V. A., Zakirova, E. G., Orishchenko, K. E., Prokhorovich, M. A., Lopatkina, M. E., Skryabin, N. A., Lebedev, I. N., Rubtsov, N. B.,

Karamysheva, T. V., Gayner, T. A., Elisaphenko, E. A., Trifonov, V. A., Zakirova, E. G., Orishchenko, K. E., Prokhorovich, M. A., Lopatkina, M. E., … Rubtsov, N. B. (2022). The Precise Breakpoint Mapping in Paracentric Inversion 10q22.2q23.3 by Comprehensive Cytogenomic Analysis, Multicolor Banding, and Single-Copy Chromosome Sequencing. *Biomedicines 2022, Vol. 10, 10*(12). https://doi.org/10.3390/BIOMEDICINES10123255

Karaoğlanoğlu, F., Ricketts, C., Ebren, E., Rasekh, M. E., Hajirasouliha, I., & Alkan, C. (2020). VALOR2: characterization of large-scale structural variants using linked-reads. *Genome Biology*, *21*(1), 72. https://doi.org/10.1186/S13059-020-01975-8

Ke, R., Mignardi, M., Hauling, T., & Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Human Mutation*, *37*(12), 1363–1367. https://doi.org/10.1002/HUMU.23051,

Khandekar, A., Vangara, R., Barnes, M., Díaz-Gay, M., Abbasi, A., Bergstrom, E. N., Steele, C. D., Pillay, N., & Alexandrov, L. B. (2023). Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.02.03.527015

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/S41587-019-0201-4;SUBJMETA=114,2056,208,2785,308,457,631,692;KWRD=GENETICS+RESEARCH,GENOME+INFORMATICS,POPULATION+GENETICS

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., & Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594. https://doi.org/10.1038/s41592-018-0051-x

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 91. https://doi.org/10.1186/S13073-020-00791-W

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/GR.129684.111

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/BIOINFORMATICS/BTS480

Krishnan, V., Utiramerur, S., Ng, Z., Datta, S., Snyder, M. P., & Ashley, E. A. (2021). Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. *BMC Bioinformatics 2021 22:1*, *22*(1), 85-. https://doi.org/10.1186/S12859-020-03934-3

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951. https://doi.org/10.1101/PDB.TOP084970

Kulkarni, P., & Frommolt, P. (2017). Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Computational and Structural Biotechnology Journal*, *15*, 471. https://doi.org/10.1016/J.CSBJ.2017.10.001

Kurc, T., Hastings, S., Kumar, V., Langella, S., Sharma, A., Pan, T., Oster, S., Ervin, D., Permar, J., Narayanan, S., Gil, Y., Deelman, E., Hall, M., & Saltz, J. (2009). HPC and Grid Computing for Integrative Biomedical Research. *The International Journal of High Performance Computing Applications*, *23*(3), 252–264. https://doi.org/10.1177/1094342009106192

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, *12*(5), e0177459. https://doi.org/10.1371/JOURNAL.PONE.0177459

Langer, B. E., Amaral, A., Baudement, M.-O., Bonath, F., Charles, M., Chitneedi, P. K., Clark, E. L., Di Tommaso, P., Djebali, S., Ewels, P. A., Eynard, S., Fellows Yates, J. A., Fischer, D., Floden, E. W., Foissac, S., Gabernet, G., Garcia, M. U., Gillard, G., Gundappa, M. K., … community, the

nf-core. (2025). Empowering bioinformatics communities with Nextflow and nf-core. *Genome Biology*, *26*(1), 1–15. https://doi.org/10.1186/S13059-025-03673-9/FIGURES/2

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/NMETH.1923;SUBJMETA=1647,208,212,48,514,631;KWRD=BIOINFORMATICS,GENOMICS,SEQUENCING

Larson, N. B., Oberg, A. L., Adjei, A. A., & Wang, L. (2022). A clinician's guide to bioinformatics for next-generation sequencing. *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer*, *18*(2), 143. https://doi.org/10.1016/J.JTHO.2022.11.006

Law, A. J., Kleinman, J. E., Weinberger, D. R., & Weickert, C. S. (2007). Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. *Human Molecular Genetics*, *16*(2), 129–141. https://doi.org/10.1093/HMG/DDL449

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733–739. https://doi.org/10.1038/NRG2825;SUBJMETA=1513,1647,48,631;KWRD=BIOINFORMATICS,GENETIC+TECHNIQUES

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, *18*(3), 530–536. https://doi.org/10.1093/BIB/BBW020

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://arxiv.org/pdf/1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589. https://doi.org/10.1093/BIOINFORMATICS/BTP698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/GR.078212.108

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656

Lin, H., Hargreaves, K. A., Li, R., Reiter, J. L., Wang, Y., Mort, M., Cooper, D. N., Zhou, Y., Zhang, C., Eadon, M. T., Dolan, M. E., Ipe, J., Skaar, T. C., & Liu, Y. (2019). RegSNPs-intron: A computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biology*, *20*(1), 1–16. https://doi.org/10.1186/S13059-019-1847-4/FIGURES/5

Lin, Y. L., Chang, P. C., Hsu, C., Hung, M. Z., Chien, Y. H., Hwu, W. L., Lai, F. P., & Lee, N. C. (2022). Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports*, *12*(1), 1–6. https://doi.org/10.1038/S41598-022-05833-4;SUBJMETA=114,208,4017,631,692;KWRD=COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,GENETICS,MOLECULAR+MEDICINE

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. *Journal of Biomedicine and Biotechnology*, *2012*, 251364. https://doi.org/10.1155/2012/251364

Liu, Z., Roberts, R., Mercer, T. R., Xu, J., Sedlazeck, F. J., & Tong, W. (2022). Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biology 2022 23:1*, *23*(1), 68-. https://doi.org/10.1186/S13059-022-02636-8

Livingstone, M., Folkman, L., Yang, Y., Zhang, P., Mort, M., Cooper, D. N., Liu, Y., Stantic, B., & Zhou, Y. (2017). Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Human Mutation*, *38*(10), 1336–1347. https://doi.org/10.1002/HUMU.23283

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;SUBJMETA=208,212,514,631,649,726;KWRD=GENETIC+VARIATION,GENOMICS,SEQUENCING

López, F. V., Ashton, J. J., Cheng, G., & Ennis, S. (2025). A systematic analysis of contemporary whole exome sequencing capture kits to optimise high-coverage capture of CCDS regions. *NAR Genomics and Bioinformatics*, *7*(3), 115. https://doi.org/10.1093/NARGAB/LQAF115

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), e1005457. https://doi.org/10.1371/JOURNAL.PCBI.1005457

Lussier, Y. A., Li, H., & Maienschein-Cline, M. (2013). Conquering computational challenges of omics data and post-ENCODE paradigms. *Genome Biology*, *14*(8), 1–3. https://doi.org/10.1186/GB-2013-14-8-310/METRICS

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

McGinn, S., & Gut, I. G. (2013). DNA sequencing – spanning the generations. *New Biotechnology*, *30*(4), 366–372. https://doi.org/10.1016/J.NBT.2012.11.012

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/GR.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/S13059-016-0974-4/TABLES/8

Meienberg, J., Zerjavic, K., Keller, I., Okoniewski, M., Patrignani, A., Ludin, K., Xu, Z., Steinmann, B., Carrel, T., Röthlisberger, B., Schlapbach, R., Bruggmann, R., & Matyas, G. (2015). New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Research*, *43*(11), e76–e76. https://doi.org/10.1093/NAR/GKV216

Merkel Dirk. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux Journal*. https://doi.org/10.5555/2600239.2600241

Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J. M. L. M., van der Zaag, P. J., Franke, A., Nilsson, M., Lehrach, H., & Brookes, A. J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics*, *10*(6), 374–386. https://doi.org/10.1093/BFGP/ELR033,

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/NRG2626;KWRD=BIOMEDICINE

Miller, H. E., & Bishop, A. J. R. (2021). Correlation AnalyzeR: functional predictions from gene co-expression correlations. *BMC Bioinformatics 2021 22:1*, *22*(1), 206-. https://doi.org/10.1186/S12859-021-04130-7

Mitra-Behura, S., Fiolka, R. P., & Daetwyler, S. (2021). Singularity Containers Improve Reproducibility and Ease of Use in Computational Image Analysis Workflows. *Frontiers in Bioinformatics*, *1*, 757291. https://doi.org/10.3389/FBINF.2021.757291/FULL

Mitsiades, I. R., Onozato, M., Iafrate, A. J., Hicks, D., Gülhan, D. C., Sgroi, D. C., & Rheinbay, E. (2024). ERBB2/HOXB13 co-amplification with interstitial loss of BRCA1 defines a unique subset of breast cancers. *Breast Cancer Research 2024 26:1*, *26*(1), 185-. https://doi.org/10.1186/S13058-024-01943-1

Miya, F., Kato, M., Shiohama, T., Okamoto, N., Saitoh, S., Yamasaki, M., Shigemizu, D., Abe, T., Morizono, T., Boroevich, K. A., Kosaki, K., Kanemura, Y., & Tsunoda, T. (2015). A combination of targeted enrichment methodologies for whole-exome sequencing reveals novel pathogenic mutations. *Scientific Reports 2015 5:1*, *5*(1), 9331-. https://doi.org/10.1038/srep09331

Mohammad, T., Zolotovskaia, M. A., Suntsova, M. V., & Buzdin, A. A. (2024). Cancer fusion transcripts with human non-coding RNAs. *Frontiers in Oncology*, *14*, 1415801. https://doi.org/10.3389/FONC.2024.1415801

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., van Dyken, P. C., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Vieira, F. G., Meesters, C., Lee, S., Twardziok, S. O., Kanitz, A., VanCampen, J., Malladi, V., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2025). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/F1000RESEARCH.29032.3

Mollon, J., Almasy, L., Jacquemont, S., & Glahn, D. C. (2023). The contribution of copy number variants to psychiatric symptoms and cognitive ability. *Molecular Psychiatry*, *28*(4), 1480. https://doi.org/10.1038/S41380-023-01978-4

Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., & Swarup, V. (2023). hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods*, *3*(6), 100498. https://doi.org/10.1016/j.crmeth.2023.100498

Moustakli, E., Christopoulos, P., Potiris, A., Zikopoulos, A., Mavrogianni, D., Karampas, G., Kathopoulis, N., Anagnostaki, I., Domali, E., Tzallas, A. T., Drakakis, P., & Stavros, S. (2025). Long-Read Sequencing and Structural Variant Detection: Unlocking the Hidden Genome in Rare Genetic Disorders. *Diagnostics 2025, Vol. 15, 15*(14). https://doi.org/10.3390/diagnostics15141803

Mullaney, J. M., Mills, R. E., Stephen Pittard, W., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131. https://doi.org/10.1093/HMG/DDQ400

Noor, A. M., Holmberg, L., Gillett, C., & Grigoriadis, A. (2015). Big Data: the challenge for small research groups in the era of cancer genomics. *British Journal of Cancer 2015 113:10*, *113*(10), 1405–1412. https://doi.org/10.1038/bjc.2015.341

Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., & Zook, J. M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics 2023 24:7*, *24*(7), 464–483. https://doi.org/10.1038/s41576-023-00590-0

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics (Oxford, England)*, *32*(10), 1479–1485. https://doi.org/10.1093/BIOINFORMATICS/BTV722

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., & Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Medicine*, *5*(3), 1–18. https://doi.org/10.1186/GM432/FIGURES/5

Ortendahl, J. D., Cuyun Carter, G., Thakkar, S. G., Bognar, K., Hall, D. W., Abdou, Y., & Gandara, D. (2025). Costs and Benefits of Whole-Exome, Whole-Transcriptome Sequencing in Patients With Advanced Non–Small Cell Lung Cancer. *JCO Precision Oncology*, *9*(9), e2400640. https://doi.org/10.1200/PO-24-00640

Ospina, O. E., Manjarres-Betancur, R., Gonzalez-Calderon, G., Soupir, A. C., Smalley, I., Tsai, K. Y., Markowitz, J., Khaled, M. L., Vallebuona, E., Berglund, A. E., Eschrich, S. A., Yu, X., & Fridley, B. L. (2025). spatialGE Is a User-Friendly Web Application That Facilitates Spatial

Transcriptomics Data Analysis. *Cancer Research*, *85*(5), 848–858. https://doi.org/10.1158/0008-5472.CAN-24-2346

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: Identifying the splicing spoilers. *Nature Reviews Genetics*, *5*(5), 389–396. https://doi.org/10.1038/NRG1327;KWRD=BIOMEDICINE

Pande, S., Dawood, M., Grochowski, C. M., Pande, S., Dawood, M., & Grochowski, C. M. (2025). Structural Variants: Mechanisms, Mapping, and Interpretation in Human Genetics. *Genes 2025, Vol. 16, 16*(8). https://doi.org/10.3390/GENES16080905

Park, S. T., & Kim, J. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *International Neurourology Journal*, *20*(Suppl 2), S76-83. https://doi.org/10.5213/INJ.1632742.371

Pei, Z., Deng, K., Lei, C., Du, D., Yu, G., Sun, X., Xu, C., & Zhang, S. (2022). Identifying Balanced Chromosomal Translocations in Human Embryos by Oxford Nanopore Sequencing and Breakpoints Region Analysis. *Frontiers in Genetics*, *12*, 810900. https://doi.org/10.3389/FGENE.2021.810900/FULL

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., & Vizcaíno, J. A. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLOS Computational Biology*, *12*(7), e1004947. https://doi.org/10.1371/JOURNAL.PCBI.1004947

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14. https://doi.org/10.1186/1479-7364-8-14/FIGURES/4

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & Depristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Qin, Y., Maggio, A., Hawkins, D., Beaudry, L., Kim, A., Pan, D., Gong, T., Fu, Y., Yang, H., & Deng, Y. (2024). Whole-genome bisulfite sequencing data analysis learning module on Google Cloud Platform. *Briefings in Bioinformatics*, *25*(Suppl 1), bbae236. https://doi.org/10.1093/BIB/BBAE236

Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., Stratton, M. R., & Hurles, M. E. (2015). Timing, rates and spectra of human germline mutation. *Nature Genetics*, *48*(2), 126. https://doi.org/10.1038/NG.3469

Rajagopalan, R., Murrell, J. R., Luo, M., & Conlin, L. K. (2020). A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Medicine*, *12*(1), 14. https://doi.org/10.1186/S13073-020-0712-0

Rantapero, T., Wahlfors, T., Kähler, A., Hultman, C., Lindberg, J., Tammela, T. L. J., Nykter, M., Schleutker, J., & Wiklund, F. (2020). Inherited DNA repair gene mutations in men with lethal prostate cancer. *Genes*, *11*(3). https://doi.org/10.3390/GENES11030314,

Rashid, U., Wu, C., Shiller, J., Smith, K., Crowhurst, R., Davy, M., Chen, T. H., Carvajal, I., Bailey, S., Thomson, S., & Deng, C. H. (2024). AssemblyQC: a Nextflow pipeline for reproducible reporting of assembly quality. *Bioinformatics*, *40*(8). https://doi.org/10.1093/BIOINFORMATICS/BTAE477

Real, R., & Vargas, J. M. (1996). The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology*, *45*(3), 380–385. https://doi.org/10.1093/SYSBIO/45.3.380

Record, C. J., & Reilly, M. M. (2024). Lessons and pitfalls of whole genome sequencing. *Practical Neurology*, *24*(4), 263–274. https://doi.org/10.1136/PN-2023-004083

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global

variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/NATURE05329;KWRD=SCIENCE

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/GIM.2015.30

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/NAR/GKV007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139. https://doi.org/10.1093/BIOINFORMATICS/BTP616

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology 2013 14:5*, *14*(5), R51-. https://doi.org/10.1186/GB-2013-14-5-R51

Sachdeva, S., Bhatia, S., Al Harrasi, A., Shah, Y. A., Anwer, K., Philip, A. K., Shah, S. F. A., Khan, A., & Ahsan Halim, S. (2024). Unraveling the role of cloud computing in health care system and biomedical sciences. *Heliyon*, *10*(7), e29044. https://doi.org/10.1016/J.HELIYON.2024.E29044

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., … Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), 928–933. https://doi.org/10.1038/35057149;KWRD=SCIENCE

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y. Y., Yachie, N., … Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, *161*(3), 647–660. https://doi.org/10.1016/J.CELL.2015.04.013/ATTACHMENT/439C3758-C0E6-450F-B181-141ED74D83F1/MMC6.XLSX

Saint-Léger, A., & Ribas De Pouplana, L. (2015). The importance of codon–anticodon interactions in translation elongation. *Biochimie*, *114*, 72–79. https://doi.org/10.1016/J.BIOCHI.2015.04.013

Sandmann, S., De Graaf, A. O., Karimi, M., Van Der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2017). Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/SREP43169;TECHMETA

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/PNAS.74.12.5463

Santucci, K., Cheng, Y., Xu, S. M., & Janitz, M. (2024). Enhancing novel isoform discovery: leveraging nanopore long-read sequencing and machine learning approaches. *Briefings in Functional Genomics*, *23*(6), 683–694. https://doi.org/10.1093/bfgp/elae031

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, *12*(7), 997. https://doi.org/10.3390/BIOLOGY12070997

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/BIOINFORMATICS/BTR026

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*(6), 839–851. https://doi.org/10.1261/RNA.053959.115

Scotti, M. M., & Swanson, M. S. (2015). RNA mis-splicing in disease. *Nature Reviews Genetics 2015 17:1*, *17*(1), 19–32. https://doi.org/10.1038/nrg.2015.3

Seaby, E. G., Pengelly, R. J., & Ennis, S. (2016). Exome sequencing explained: a practical guide to its clinical application. *Briefings in Functional Genomics*, *15*(5), 374–384. https://doi.org/10.1093/BFGP/ELV054

Sequencing Human Whole Genomes on the NovaSeq 6000 System. (n.d.). Retrieved May 22, 2025, from https://support.illumina.com/support-content/WGS-on-NovaSeq.html

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–1358. https://doi.org/10.1093/BIOINFORMATICS/BTS163

Sharma, P., & Sampath, H. (2019). Mitochondrial DNA Integrity: Role in Health and Disease. *Cells*, *8*(2), 100. https://doi.org/10.3390/CELLS8020100

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/NBT1486;KWRD=LIFE+SCIENCES

Shlien, A., & Malkin, D. (2009). Copy number variations and cancer. *Genome Medicine*, *1*(6), 62. https://doi.org/10.1186/GM62

Shouib, R., Eitzen, G., & Steenbergen, R. (2025). A Guide to Basic RNA Sequencing Data Processing and Transcriptomic Analysis. *Bio-Protocol*, *15*(9), e5295. https://doi.org/10.21769/BIOPROTOC.5295

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/CPMB.59

Solís-Moruno, M., Batlle-Masó, L., Bonet, N., Aróstegui, J. I., & Casals, F. (2022). Somatic genetic variation in healthy tissue and non-cancer diseases. *European Journal of Human Genetics*, *31*(1), 48. https://doi.org/10.1038/S41431-022-01213-8

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724. https://doi.org/10.1038/NATURE07943;KWRD=SCIENCE

Strozzi, F., Janssen, R., Wurmus, R., Crusoe, M. R., Githinji, G., Di Tommaso, P., Belhachemi, D., Möller, S., Smant, G., de Ligt, J., & Prins, P. (2019). Scalable Workflows and Reproducible Data Analysis for Genomics. *Methods in Molecular Biology (Clifton, N.J.)*, *1910*, 723. https://doi.org/10.1007/978-1-4939-9074-0_24

Sturm, M., Schroeder, C., & Bauer, P. (2016). SeqPurge: Highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinformatics*, *17*(1), 1–7. https://doi.org/10.1186/S12859-016-1069-7/TABLES/6

Taanman, J. W. (1999). The mitochondrial genome: structure, transcription, translation and replication. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1410*(2), 103–123. https://doi.org/10.1016/S0005-2728(98)00161-3

Taylor, D. J., Eizenga, J. M., Li, Q., Das, A., Jenike, K. M., Kenny, E. E., Miga, K. H., Monlong, J., McCoy, R. C., Paten, B., & Schatz, M. C. (2024). Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. *Annual Review of Genomics and Human Genetics*, *25*(1), 77–104. https://doi.org/10.1146/ANNUREV-GENOM-021623-081639

Tello, D., Gil, J., Loaiza, C. D., Riascos, J. J., Cardozo, N., & Duitama, J. (2019). NGSEP3: accurate variant calling across species and sequencing protocols. *Bioinformatics*, *35*(22), 4716. https://doi.org/10.1093/BIOINFORMATICS/BTZ275

Thrash, A., Arick, M., & Peterson, D. G. (2018). Quack: A quality assurance tool for high throughput sequence data. *Analytical Biochemistry*, *548*, 38–43. https://doi.org/10.1016/J.AB.2018.01.028

Torri, F., Dinov, I. D., Zamanyan, A., Hobel, S., Genco, A., Petrosyan, P., Clark, A. P., Liu, Z., Eggert, P., Pierce, J., Knowles, J. A., Ames, J., Kesselman, C., Toga, A. W., Potkin, S. G., Vawter, M. P., & Macciardi, F. (2012). Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows. *Genes*, *3*(3), 545. https://doi.org/10.3390/GENES3030545

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology 2010 28:5*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers 2021 1:1*, *1*(1), 59-. https://doi.org/10.1038/s43586-021-00056-9

Vallejos-Vidal, E., Reyes-Cerpa, S., Rivas-Pardo, J. A., Maisey, K., Yáñez, J. M., Valenzuela, H., Cea, P. A., Castro-Fernandez, V., Tort, L., Sandino, A. M., Imarai, M., & Reyes-López, F. E. (2020). Single-Nucleotide Polymorphisms (SNP) Mining and Their Effect on the Tridimensional Protein Structure Prediction in a Set of Immunity-Related Expressed Sequence Tags (EST) in Atlantic Salmon (Salmo salar). *Frontiers in Genetics*, *10*, 1406. https://doi.org/10.3389/FGENE.2019.01406/FULL

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.BI1110S43

Van der Auwera, G., & O'Connor, B. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. In *Genomics in the Cloud*. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/SCIENCE.1058040/SUPPL_FILE/1058040S3-6_MED.GIF

Vera Alvarez, R., Pongor, L., Mariño-Ramírez, L., & Landsman, D. (2021). PM4NGS, a project management framework for next-generation sequencing data analysis. *GigaScience*, *10*(1), 1–9. https://doi.org/10.1093/GIGASCIENCE/GIAA141

Verwilt, J., Mestdagh, P., & Vandesompele, J. (2023). Artifacts and biases of the reverse transcription reaction in RNA sequencing. *RNA*, *29*(7), 889. https://doi.org/10.1261/RNA.079623.123

Vijg, J. (2014). Somatic mutations, genome mosaicism, cancer and aging. *Current Opinion in Genetics & Development*, *26*, 141. https://doi.org/10.1016/J.GDE.2014.04.002

Vijg, J., & Dong, X. (2020). Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell*, *182*(1), 12. https://doi.org/10.1016/J.CELL.2020.06.024

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/NATURE07509;KWRD=SCIENCE

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57. https://doi.org/10.1038/NRG2484

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737A0;KWRD=SCIENCE

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, *17*(1), 1–13. https://doi.org/10.1186/S12859-016-0956-2/TABLES/2

Wong, E. T. C., So, V., Guron, M., Kuechler, E. R., Malhis, N., Bui, J. M., & Gsponer, J. (2020). Protein–Protein Interactions Mediated by Intrinsically Disordered Protein Regions Are Enriched in Missense Mutations. *Biomolecules*, *10*(8), 1097. https://doi.org/10.3390/BIOM10081097

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, *18*(10), 1161–1168. https://doi.org/10.1038/S41592-021-01254-9

Xiao, Y., Cheng, D., Luo, K., Li, M., Tan, Y., Lin, G., & Hu, L. (2023). Evaluation of genetic risk of apparently balanced chromosomal rearrangement carriers by breakpoint characterization. *Journal of Assisted Reproduction and Genetics*, *41*(1), 147. https://doi.org/10.1007/S10815-023-02986-7

Yang, S. F., Lu, C. W., Yao, C. Te, & Hung, C. M. (2019). To trim or not to trim: Effects of read trimming on the de novo genome assembly of awidespread east asian passerine, the rufous-capped babbler (Cyanoderma ruficeps Blyth). *Genes*, *10*(10). https://doi.org/10.3390/GENES10100737,

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., Wang, J., & Shen, Y. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*, *10*(1). https://doi.org/10.1186/S13039-017-0333-5

Yoo, A. B., Jette, M. A., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2862*, 44–60. https://doi.org/10.1007/10968987_3

You, B. H., Yoon, J. H., Kang, H., Lee, E. K., Lee, S. K., & Nam, J. W. (2019). HERES, a lncRNA that regulates canonical and noncanonical Wnt signaling pathways via interaction with EZH2. *Proceedings of the National Academy of Sciences*, *116*(49), 24620–24629. https://doi.org/10.1073/pnas.1912126116

Zabrady, M., Zabrady, K., Li, A. W. H., & Doherty, A. J. (2023). Reverse transcriptases prime DNA synthesis. *Nucleic Acids Research*, *51*(14), 7125–7142. https://doi.org/10.1093/NAR/GKAD478

Zanti, M., Michailidou, K., Loizidou, M. A., Machattou, C., Pirpa, P., Christodoulou, K., Spyrou, G. M., Kyriacou, K., & Hadjisavvas, A. (2021). Performance evaluation of pipelines for mapping, variant calling and interval padding, for the analysis of NGS germline panels. *BMC Bioinformatics 2021 22:1*, *22*(1), 218-. https://doi.org/10.1186/S12859-021-04144-1

Zeng, X., Lin, D., Liang, D., Huang, J., Yi, J., Lin, D., & Zhang, Z. (2023). Gene sequencing and result analysis of balanced translocation carriers by third-generation gene sequencing technology. *Scientific Reports 2023 13:1*, *13*(1), 7004-. https://doi.org/10.1038/s41598-022-20356-8

Zeng, Z., Aptekmann, A. A., & Bromberg, Y. (2021). Decoding the effects of synonymous variants. *Nucleic Acids Research*, *49*(22), 12673–12691. https://doi.org/10.1093/NAR/GKAB1159

Zhang, X., Li, M., Lin, H., Rao, X., Feng, W., Yang, Y., Mort, M., Cooper, D. N., Wang, Y., Wang, Y., Wells, C., Zhou, Y., & Liu, Y. (2017). regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution. *Human Genetics*, *136*(9), 1279–1289. https://doi.org/10.1007/S00439-017-1783-X/FIGURES/3

Zhang, Y., Cuerdo, J., Halushka, M. K., & Mccall, M. N. (2021). The effect of tissue composition on gene co-expression. *Briefings in Bioinformatics*, *22*(1), 127–139. https://doi.org/10.1093/BIB/BBZ135

Zhang, Z., & Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, *31*(18), 5338. https://doi.org/10.1093/NAR/GKG745

Zhao, R., Lu, J., Li, Y., Zhou, W., Zhao, N., & Ji, H. (2025). A systematic evaluation of highly variable gene selection methods for single-cell RNA-sequencing. *Genome Biology*, *26*(1), 424. https://doi.org/10.1186/S13059-025-03887-X

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., & Hovig, E. (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific Reports 2020 10:1*, *10*(1), 20222-. https://doi.org/10.1038/s41598-020-77218-4

Zhao, Z., Yi, Y., Song, J., Wu, Y., Zhong, X., Lin, Y., Hohman, T. J., Fletcher, J., & Lu, Q. (2021). PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. *Genome Biology*, *22*(1), 257. https://doi.org/10.1186/S13059-021-02479-9

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology 2014 32:3*, *32*(3), 246–251. https://doi.org/10.1038/nbt.2835

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2026

ISBN 978-952-02-0576-8 (PRINT)
ISBN 978-952-02-0577-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)