UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DIPARTIMENTO DI INGEGNERIA E SCIENZE DELL'INFORMAZIONE E MATEMATICA

Dottorato di Ricerca in INFORMATION AND COMMUNICATION TECHNOLOGIES

Curriculum SOFTWARE ENGINEERING AND INTELLIGENT SYSTEMS

XXXVI ciclo

Titolo della tesi

INTEGRATING BIOINFORMATICS AND CLINICAL INSIGHTS: TOWARDS A COMPREHENSIVE FRAMEWORK FOR PRECISION MEDICINE

SSD: INF/01

Dottorando

Bianchi Andrea

Coordinatore del corso

Prof. Cortellessa Vittorio

Tutor

Prof. Di Marco Antinisca

Co-Tutor

Dr. Marzi Francesca

a.a. 2022/2023

I would like to dedicate this thesis to my beloved grandparents, Rocco, Antonietta, Alfio, and Maria, and my two beloved uncles, Zia Mè and Zio Nino, who have invested immeasurable effort to support me in every possible way throughout my university years.
Though they are no longer with us, the spirits and memories of Rocco, Antonietta, Alfio, Maria, Zia Mè, and Zio Nino continue to accompany and guide me on my path.
This work stands as a tribute to their love, sacrifices, and the indelible mark they have left on my heart and my endeavors.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Bianchi Andrea
August 2024

Acknowledgements

And I would like to to express my profound gratitude to Professor Antinisca Di Marco, whom I had the privilege of meeting during my master's degree. Our collaboration began shortly thereafter, marking over eight years of continuous work together. Professor Di Marco has been an unwavering pillar of support; she has offered guidance during challenging times, listened empathetically to my problems, and assisted me with utmost availability whenever I faced difficulties. Her mentorship has been invaluable to my academic and personal growth.

Further, I extend my heartfelt thanks to my wife, Giulia. Since meeting her six years ago, she has profoundly influenced my approach to study and work, encouraging me to strive for excellence and balance in all aspects of life. Her support and love have been indispensable on this journey. My gratitude also goes out to my parents and sisters, whose unconditional love and encouragement have shaped the person I am today. Their belief in me and constant support have been my motivation to pursue my dreams relentlessly. Additionally, I want to acknowledge all the university friends I met during my years of study. Each of you has contributed to this journey in unique ways, offering companionship, shared learning experiences, and invaluable memories that I cherish deeply.

Abstract

The transition towards precision medicine represents a pivotal shift in healthcare, emphasizing the customization of treatments to accommodate the unique genetic, lifestyle, and environmental contexts of individual patients. This evolution is propelled by advancements in Next-Generation Sequencing (NGS) technologies, facilitating an in-depth exploration of the genetic underpinnings of disease and patient response to treatment. The objective of this thesis is to tackle the computational challenges that arise in the integration of extensive genomic data into routine clinical practice, thus bridging the gap between cutting-edge genomic technologies and the realization of personalized patient care.

In pursuit of this goal, the thesis proposes a novel framework that unites bioinformatic analysis with clinical insights, offering a holistic approach to patient treatment and care. The research commences with a thorough investigation of the bioinformatics landscape, identifying and addressing the key challenges within genomic pipelines, with a particular focus on genomics. The research primarily tackled the critical issue of reproducibility in bioinformatics pipelines. Addressing this, the thesis introduced an integrated method to identify various genetic variants, thus creating a detailed genomic profile to highlight variants crucial to patient health. A critical aspect of this research is the integration of these genomic findings with patient clinical reports. The study develops a groundbreaking method to coalesce genomic information with clinical data, aiming to construct a unified framework. This framework is designed to encapsulate the entirety of the bioinformatic analysis—spanning from the identification to the interpretation of genetic variants—and synchronize this information with clinical insights. By doing so, it seeks to provide a seamless and coherent platform that supports the application of genomic discoveries in a clinical context. The innovation lies in the thesis's ability to conceptualize and implement a system that not only processes and analyzes genetic data but also integrates these findings with patient-specific clinical information. This integrated approach facilitates a more nuanced understanding of the patient's condition, enabling healthcare providers to tailor treatments that are truly personalized. Through the development of this framework, the thesis contributes significantly to the fields of bioinformatics and precision medicine, showcasing a model that could potentially enhance the efficacy, safety, and customization of patient care.

Table of contents

List of figures

Introduction

Introduction

Precision Medicine: a paradigm shift

Precision medicine (PM) marks a transformative era in healthcare, shifting from a one-size-fits-all treatment approach to one that is tailored to an individual's unique genetic makeup, lifestyle, and environmental exposure [1]. This change is grounded in the recognition that diseases manifest uniquely in each person, influenced by a complex interplay of genetic, environmental, and lifestyle factors. Precision medicine emphasizes proactive healthcare, focusing on prevention and early diagnosis through personalized genetic and clinical screenings. This allows individuals to proactively assess their risk for certain conditions, marking a significant departure from the traditional reactive healthcare model. By emphasizing early intervention and customized preventive measures, precision medicine advocates for a proactive, anticipatory approach to health management, aligning with the broader goal of maintaining well-being and preventing disease before it occurs. Advances in Next-Generation Sequencing (NGS) technologies have been instrumental in driving this paradigm shift, offering the capability to sequence vast segments of DNA or RNA rapidly and efficiently [2]. This breakthrough allows for an extensive analysis of entire genomes, providing deep insights into genetic variations and their implications for disease, evolution, and personalized medicine. As a result, NGS has democratized genomics, making it a cornerstone of modern health, clinical research and clinical diagnostics, significantly accelerating the pace at which genetic insights combined with medical screening and clinical knowledge can be translated into healthcare solutions.

The evolution towards precision medicine is supported by a multidisciplinary synergy among genomics, bioinformatics, pharmacology, and clinical medicine, with bioinformatics playing a pivotal role. Bioinformatics is central to the precision medicine ecosystem, utilizing advanced algorithms for analyzing complex biological data and turning it into actionable insights [3]. The efficacy of these algorithms is intrinsically linked to the sophistication of the computational infrastructure, which has evolved from basic tools to high-throughput computing systems encompassing powerful servers, cloud computing, and efficient data

storage solutions. This evolution in technology equips researchers to perform expansive genomic analyses and fuse diverse biological information, paving the way for customized treatment strategies. However, this innovation brings significant challenges. The volume and complexity of genomic data require advanced computational tools for effective management, analysis, and interpretation. Integrating this data with clinical insights for actionable healthcare decisions adds another layer of complexity, necessitating robust frameworks that can translate complex genomic information into understandable and usable formats for clinicians.

One of the most pressing issues in this field is ensuring the reproducibility of bioinformatics research, critical for validating findings and their practical application in precision medicine. Additionally, the selection of appropriate tools within a bioinformatics pipeline poses a significant challenge and can substantially impact the outcome of the analysis. The decision-making process for choosing the right tools is complex, given the vast array of options available, each with its strengths, limitations, and specific application contexts. This complexity underscores the necessity for a systematic approach to tool selection, aimed at optimizing the accuracy and efficiency of genomic analyses. Furthermore, once genetic variants are identified through such analyses, directly correlating these variants with patients' symptoms or medical history presents another layer of complexity. The relationship between genetic variants and phenotypic expressions is not always straightforward, making it challenging to draw direct connections without a comprehensive understanding of the underlying biological mechanisms and patient-specific contexts. This difficulty highlights the need for a framework that can encompass a more holistic view of the patient, integrating genetic data with clinical insights to provide a nuanced understanding of disease mechanisms and patient health.

In order to address these challenges, there is a clear need for the development of a framework that not only ensures the accuracy, reproducibility and reliability of bioinformatics research but also facilitates the careful selection of analytical tools and the integration of genetic findings with clinical data. Such a framework would enable a more holistic dimension to patient care, combining the depth of genetic analysis with the breadth of clinical insight to inform more personalized and effective treatment strategies. By bridging the gap between genomic research and clinical practice, this framework would represent a significant step forward in the realization of precision medicine's full potential.

Thesis Objectives

In the pursuit of precision medicine, the imperative for a timely and accurate diagnosis cannot be overstated, as it is integral to delivering effective and personalized patient care. The

current landscape of preventive diagnosis, however, is fraught with challenges that impede its efficacy and efficiency. One of the primary issues is the time-intensive nature of patient examinations, which often require extensive and repetitive tests that strain both patients and healthcare providers [4]. These traditional diagnostic processes are not only laborious but also sometimes fail to capture the comprehensive health profile of the patient due to their fragmented nature. These challenges underscore the urgent need for a holistic framework that not only streamlines the diagnostic process but also enhances the precision of medical assessments. Such a framework should seamlessly amalgamate multifaceted health data, enabling healthcare professionals to access a comprehensive, integrated patient view that supports informed and timely decision-making.

To address these pivotal concerns, this thesis embarks on a mission to fortify the foundations of precision medicine through the lens of bioinformatics. By leveraging the power of computational methods and genomic analytics, this research aims to dissect and understand the complex tapestry of factors that contribute to individual health conditions. This thesis is dedicated to constructing such a framework, using bioinformatics as the cornerstone to enhance the infrastructure of precision medicine. The intent is to develop a unified system that seamlessly integrates and analyzes genetic and clinical data, thereby optimizing the diagnosis and treatment process (Figure 1). This integrated framework aims to remove the conventional, time-consuming diagnostic practices into a streamlined, data-driven process that enhances the accuracy of health assessments and reduces the operational burden on medical practitioners.

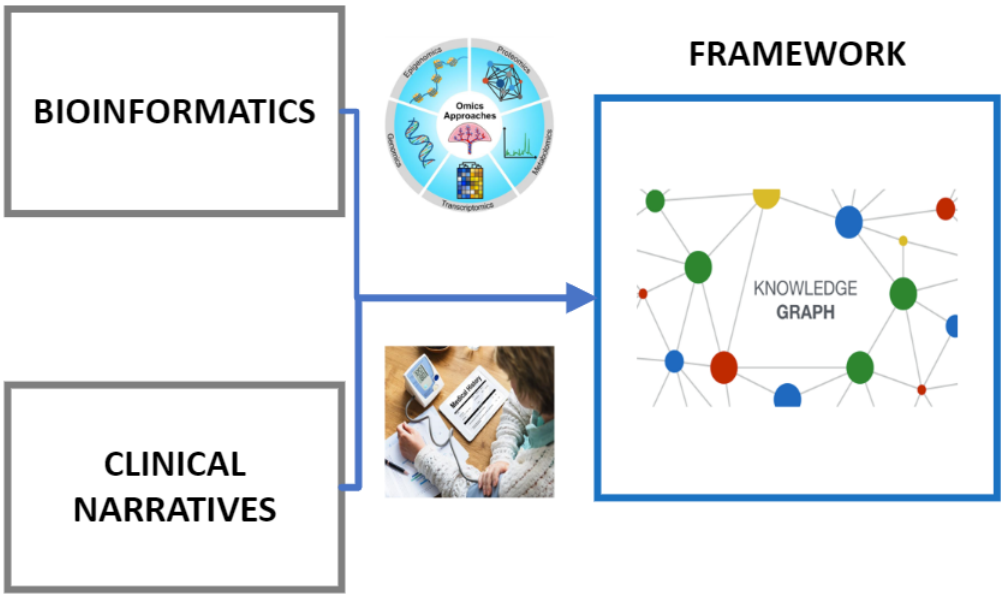


Fig. 1 Framework general overview.

In our quest to enhance precision medicine through bioinformatics, this thesis concentrates on pivotal experiments: RNA-Sequencing (RNA-Seq, from now) and Variant Calling, before merging these bioinformatics findings into an integrated knowledge framework. After a background chapter, the thesis begins with an in-depth exploration of RNA-Seq, highlighting its crucial role in deciphering the complex gene expression patterns in oncology research. This exploration sets the stage for a critical evaluation of existing bioinformatics pipelines, identifying gaps in reproducibility and specificity essential for refining the precision and accuracy of genomic studies. The thesis then transitions to scrutinize variant Calling techniques, crucial for identifying genetic differences that influence disease outcomes, including snvs, indels, and structural variations, thereby providing a comprehensive view of the genomic landscape affecting health. Building upon these analyses, the thesis extends to developing an integrated approach that combines the derived bioinformatics insights with clinical narratives. This culminates in the creation of knowledge graphs, which synergize genomic data and clinical information, thereby facilitating a holistic understanding of patient health. This system not only refines the framework for data integration but also serves as a bridge connecting the theoretical underpinnings of bioinformatics with practical clinical applications. This thesis directly addresses the obstacles that now stand in the way of the seamless integration of genetic data into clinical practice, therefore filling in the gaps that have been discovered. It does this by developing a strong framework. The study's design is guided by a set of focused research questions, each of which has been carefully constructed to address different aspects of these overarching difficulties. This ensures that the study is thoroughly explored and resolved within the context of precision medicine.

- Research Question 1 (**RQ1**): "**What challenges hinder the integration of genomic data into clinical applications?**". The motivation behind this question lies in the current skepticism among medical practitioners regarding the reliability of bioinformatics analysis for clinical decision-making. Despite the wealth of genomic data available, the translation into clinical practice is often hampered by doubts about the 'truth' or accuracy of these analyses. Clinicians and healthcare providers are cautious in utilizing bioinformatics results due to potential issues with data validity, quality, and reproducibility, which are crucial for making accurate patient-related decisions. This gap underscores the necessity of establishing trust in bioinformatics data through rigorous validation and ensuring its relevance and reliability in a clinical context.

  This is addressed by examining the reproducibility of bioinformatics experiments, which stands as a prominent hurdle in the field. The issue is explored in the context of dataset retrievability, acknowledging it as a pivotal concern in the replicability and utility of bioinformatics research. Dataset retrievability pertains to the ease with which

data used in bioinformatics analyses can be accessed and reused by other researchers or clinicians, ensuring that results can be independently verified and trusted. Enhancing the retrievability and transparency of datasets is fundamental to bridging the gap between bioinformatics research and its clinical application, thereby fostering a more reliable and effective integration of genomic data into healthcare.

- Research Question 2 (**RQ2**): "**How can bioinformatics pipelines be optimized for improved biological correctness in genomic analysis?**" The rationale behind this question stems from the observed hesitancy among doctors and medical practitioners to fully trust and integrate bioinformatics analysis into clinical decision-making. The root of this distrust is often the perceived lack of biological and medical reliability in the outcomes of these analyses. Given the critical role that bioinformatics plays in understanding complex genomic data, there is an imperative need to optimize these pipelines to ensure their results are not only scientifically accurate but also clinically relevant and reliable.

  Optimizing bioinformatics pipelines involves a detailed scrutiny of their components to identify and implement the most reliable tools and methods that can improve the biological correctness of the analyses. This means selecting algorithms and tools that have been validated through rigorous scientific research and have proven utility in clinical settings. The aim is to refine these pipelines in such a way that they not only produce scientifically robust data but also generate results that align with biological and medical expectations, thereby enhancing their applicability and trustworthiness in a clinical environment.

- Research Question 3 (**RQ3**): "**What strategies can facilitate a comprehensive analysis of genetic variations affecting individual health?**" This question arises from the critical need in personalized medicine to understand the full spectrum of genetic variations that contribute to individual health profiles. While snvs and indels are well-recognized for their roles in genetic diversity and disease predisposition, structural variants—such as copy number variations, inversions, and translocations—also play a significant part in human biology and disease mechanisms. These larger-scale genomic alterations have been increasingly linked to various conditions, from developmental disorders to cancer, underscoring their importance in comprehensive genomic studies.

  To address this, the thesis proposes an integrated approach that encompasses the concurrent analysis of different types of genetic variants, including snvs, indels, and structural variants. The goal is to create a holistic view of the genomic landscape affecting individual health, recognizing that a complete understanding of genetic contributions

to disease extends beyond the prevalent snvs and indels. This comprehensive strategy aims to streamline the variant detection and analysis process, ensuring that each type of variation is accurately identified and its potential impact on health is thoroughly evaluated.

- Research Question 4 (**RQ4**): "**How can we achieve a seamless integration of genomic and clinical data for a coherent understanding of patient health?**" Building upon the advancements in bioinformatics, we now have access to reliable genomic data outputs from meticulously designed bioinformatics pipelines. The next critical step is to enhance patient care by integrating these genomic insights with detailed clinical narratives, thus providing a holistic view of patient health. The thesis proposes the development of a system adept at not just processing genomic data but also proficient in assimilating clinical texts. By merging genomic outputs with clinical narratives, the system will enable a more nuanced understanding of each patient's unique health condition, thereby improving the accuracy and personalization of medical treatment. This proposed integration not only makes the assimilation of genomic and clinical information seamless but also ensures that the combined data is leveraged comprehensively to enhance patient diagnosis, prognosis, and treatment planning, marking a significant leap forward in the personalized medicine paradigm.

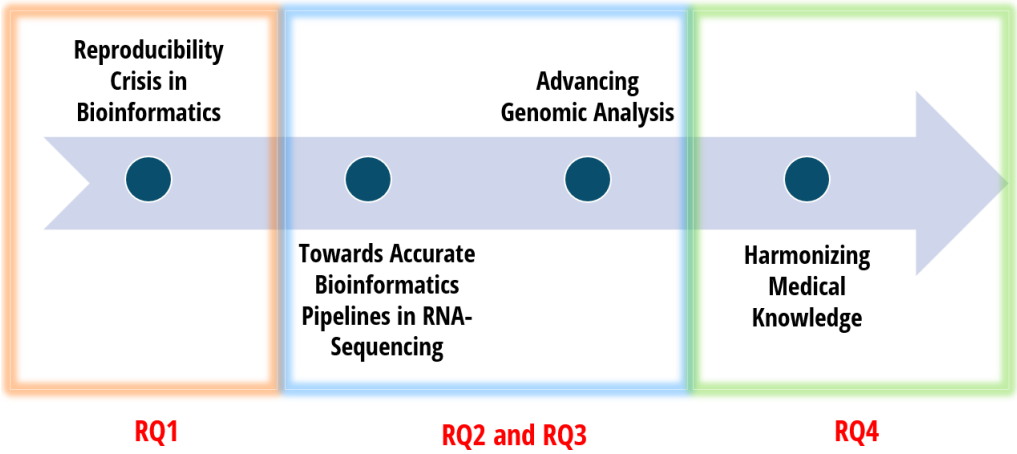


Fig. 2 Association of key research areas with corresponding research questions.

Each objective is carefully aligned with the corresponding research question to ensure that the development of the framework systematically addresses the identified gaps and challenges (Figure 2). By fulfilling these objectives, the thesis contributes a structured approach that enhances the fidelity of bioinformatics experiments, ensures the reproducibility of results, enables a multifaceted analysis of genetic variations, and marries this information with

patient clinical data to provide a holistic picture of patient health. The result is a framework poised to set a new precedent in precision medicine.

Outline

The thesis is organized as in the following:

- Chapter 1 sets the stage by discussing the essence of precision medicine and the transformative role of NGS technologies. It outlines how NGS has revolutionized our approach to genomics, transcriptomics, proteomics, metabolomics, and pharmacogenomics, providing the necessary backdrop for understanding the shift towards data-driven personalized treatment strategies.

- Proceeding to Chapter 2, it addresses the reproducibility crisis prevalent within bioinformatics pipelines, with a particular emphasis on RNA-Sequencing applications in the field of oncology. It engages in a detailed examination to identify critical gaps in reproducibility and specificity that currently impede the field. This chapter emphasizes the urgent requirement for refined methodologies capable of reliably interpreting the expansive datasets produced by NGS technologies. Additionally, it delves into the issue of dataset retrievability, recognizing it as a primary concern for the lack of reproducibility. Drawing on the insights gained from this exploration, the chapter presents a strategy for self-evaluating the dataset retrievability in experiments.

- Chapter 3 delves into the crucial role of bioinformatics pipelines in underpinning the reliability of genomic analyses, with a particular focus on alignment—a critical component within RNA-Seq workflows. This chapter examines the alignment phase by comparing two widely utilized tools, scrutinizing the differences they manifest in downstream analyses and the consequent impact on the results.

- Chapter 4 marks a significant expansion of the thesis. It elaborates on the methodologies developed for a comprehensive analysis of genetic variations, highlighting the importance of an integrated approach in uncovering the complexities of genomic alterations. This chapter is pivotal in bridging the gap between raw genomic data and actionable insights for clinical application, presenting a methodological framework that encompasses the analysis of Single Nucleotide Polymorphisms (snvs), insertions and deletions (indels), and Copy Number Variations (CNVs).

- Chapter 5 presents the thesis's novel contribution to the field: a graph-based system for integrating genomic and clinical data, specifically targeting the challenge of multi-source diagnoses. This chapter discusses the development and application of an entity-relation system that synthesizes disparate data sources into knowledge graphs.

Chapter 1

Background

This chapter provides a comprehensive background necessary to understand the context and significance of the research conducted in this thesis. It delves into the core concepts and technologies that form the foundation of precision medicine, highlighting their roles, advancements, and the challenges they pose in the field of bioinformatics.

1.1 Precision Medicine

The "one size fits all" philosophy of conventional medicine meant that a single medication was prescribed for every patient with a specific illness. However, there are several issues with this method, including the fact that only a portion of these people would respond to the specific medication, a sizable portion would not respond, and a considerable portion would experience side effects. Genetic variants, age, gender, addictions, ethnicity, concurrent drug use, comorbidities, environmental variables, and so forth might all be contributing causes to these inter-individual discrepancies. This resulted in low patient and physician satisfaction, higher expenses, and medication waste due to the mismatch between prescribed treatments and individual patient responses. With increasing ability to not only measure but also to store and share data related to health, in addition to the abundance of available genetic testing tools, the integration of precision medicine with public health will be a very productive union. Electronic health records of a population can be accessed, and along with information from other data, health risks can be gauged to identify subpopulations at higher risk. This allows for preventive modalities to be targeted to these subpopulations, leading to the prevention of chronic diseases, improvement of quality of life and the reduction of healthcare expenses [5]. Precision medicine stands at the forefront of a transformative shift in healthcare, moving from traditional, uniform treatment approaches to highly individualized care. This paradigm shift is rooted in the understanding that treatment efficacy varies significantly

among individuals, influenced by their unique genetic makeup, environmental exposures, and lifestyles (Figure 1.1). The term "precision medicine" gained prominence in the scientific community following its conceptualization to describe the impact of molecular diagnostics on eliminating diagnostic ambiguities. When business strategist Clayton Christensen of Harvard Business School in Boston first used the term "precision medicine" in 2008 to characterise how molecular diagnostics enables doctors to clearly identify the aetiology of a disease without relying solely on intuition, the term became part of the scientific canon [6]. The term didn't become popular until 2011, when a US National Research Council-convened group published a plan to update the taxonomy of diseases by using molecular data, such as causative genetic variations, as opposed to a system of classification based on symptoms [7]. As a result, precision medicine moves beyond personalised medicine's individual focus to concentrate on patient subpopulations. Historically, the concept of tailoring medical care to individual needs is not new, with practices like blood type matching for transfusions dating back over a century. Thus, the concept of precision medicine is not entirely new but has gained significant momentum with advancements in genomics, bioinformatics, and a deeper understanding of the molecular basis of diseases. Central to this approach is the utilization of biomarkers and molecular diagnostics to guide the selection of therapies best suited to an individual's specific condition and genetic profile. This tailoring of healthcare interventions encompasses not only pharmacological treatments but also preventive measures and lifestyle modifications. The advent of genomics and the rapid advancement of Next-Generation Sequencing (NGS) technologies have propelled precision medicine into the clinical forefront, enabling the detailed analysis of genetic variations at an unprecedented scale and cost-efficiency. Since the sequencing of the first human genome, the field has seen a dramatic reduction in both time and cost associated with genomic sequencing, making it a practical tool for diagnosing rare disorders and informing treatment strategies. More than ten years and almost US $3 billion later, the first human genome was sequenced in 2001, and since then, the technology has advanced significantly in speed and affordability. Nowadays, several genomes may be sequenced for about $1,000 each in a few hours [8]. By using this genetic information, precision medicine can create focused medications that are less likely to cause adverse effects and more successful than traditional treatments. The availability of Big Data is what actually distinguishes precision medicine from standard medical treatment. Thanks to the daily advancements in molecular biology and genetic testing, researchers are now able to gather vast amounts of data. This data, when combined with clinical, pharmacological, and socioeconomic information, allows for integrated data sets to be analysed by various computer-based algorithms, allowing for the observation of patterns in the effectiveness of specific treatments and the titration of those treatments to only

the susceptible populations. Precision medicine has found applications across a wide range of medical fields, including oncology, cardiovascular diseases, neurology, and psychiatry [9]. In oncology, for example, molecular profiling of tumors can identify specific genetic mutations that can be targeted by drugs, leading to personalized treatment plans that improve survival rates and quality of life for cancer patients. Similarly, in psychiatry, understanding the genetic factors that influence the response to antidepressants can guide the selection of medication, enhancing treatment efficacy and reducing the trial-and-error approach to finding the right drug [10]. Despite its promise, the implementation of precision medicine faces several challenges. These include the complexity of genetic information, the need for large-scale genomic databases to inform treatment decisions, and the integration of genomic data into clinical practice. Moreover, ethical, legal, and social implications, such as privacy concerns, access to personalized therapies, and the potential for health disparities, must be addressed to fully realize the potential of precision medicine [11].

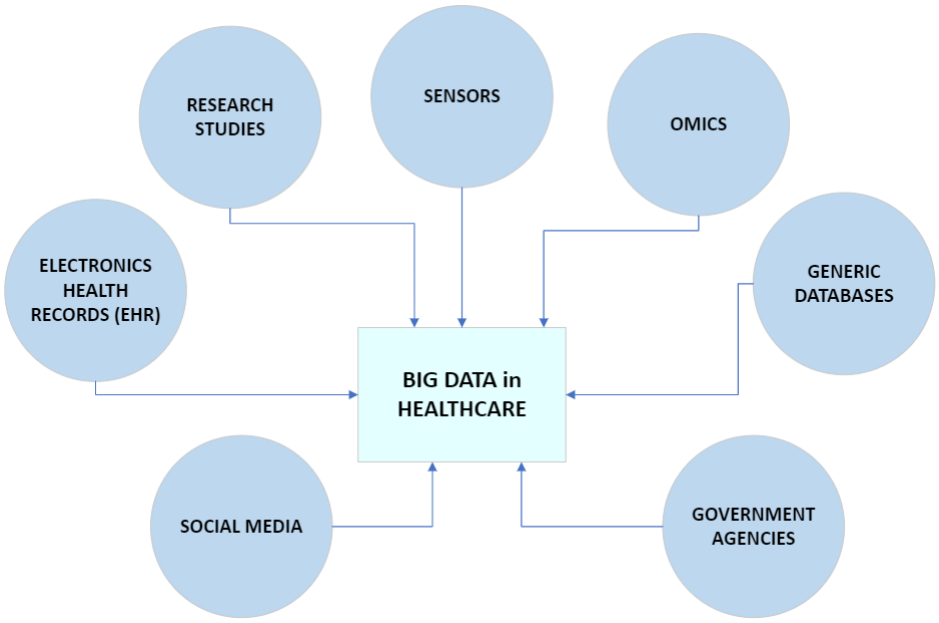


Fig. 1.1 Big data in healthcare.

1.1.1 From Big Data to Precision Medicine

Throughout the last ten years, the sentence "Big Data" has undergone tremendous change, encompassing not just the exponential development in data volume, diversity, and velocity but also our improved ability to analyse and comprehend large datasets. This progress offers the field of biological research both unmatched potential and difficult obstacles. With the advent of powerful hardware resources, we now have the capacity to create, store, and

analyze data at an unprecedented speed and scale. This change is not just quantitative but also qualitative, changing the way we think about medical diagnosis, treatment, and research. Big Data's significance in medicine goes beyond its sheer amount, emphasising instead of its capacity to support exploratory, hypothesis-free study. These studies are by their very nature hypothesis-generating, taking use of our ability to assess several variables at once in order to reveal the deep workings of sophisticated biological systems. This method, which is less dependent on past information, promises to identify linkages and routes that were previously unknown, opening up new possibilities for comprehending illness and creating focused treatments [12].

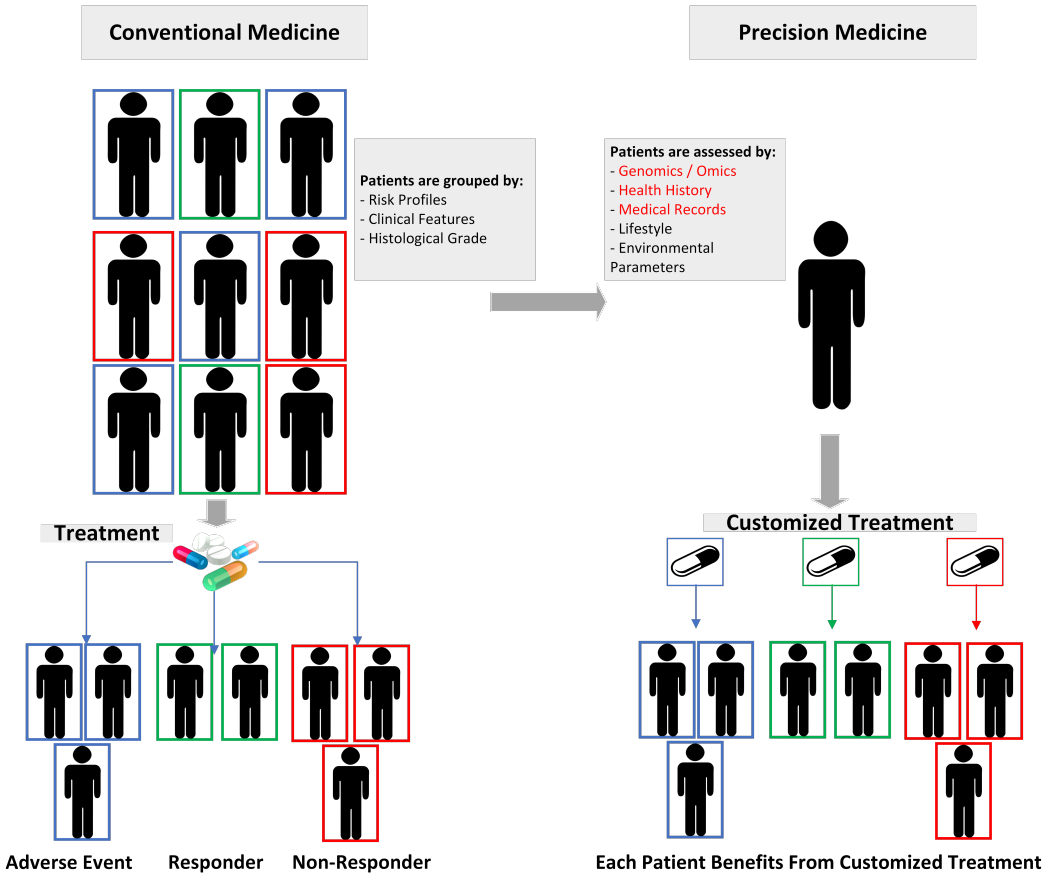


Fig. 1.2 Shift from conventional to precision medicine.

Thus, precision medicine's incorporation of Big Data analytics represents a paradigm change towards a more complex, patient-centered approach to therapy. Precision medicine uses large datasets, such as electronic health records, population health data, and genomic sequences and molecular profiles, to customise therapies to the specific genetic, environmental, and behavioural characteristics of each patient (Figure 1.2). This method not only shows

how Big Data and healthcare are convergent, but it also shows how classic, hypothesis-driven research and new, data-driven theories may work in concert. Developing multinational consortia (for instance, the TEDDY [13] and TrialNet [14] consortia, and others) and engaging the community are two examples of the collaborative approach that is necessary to fully realise the promise of big data. Along with increasing research resources, these programmes also establish new benchmarks for data exchange, analysis, and gathering. Building on progress in pharmacogenetics, pharmacogenomics, and targeted therapy, PM aims at integrating multiple data sources to 'tailor medical treatment to the individual characteristics of each patient.' The development of precision medicine is therefore premised on the collection of large repositories of data including electronic health records (EHRs), standard clinical measurements, genome sequences, environmental data, and lifestyle data collected over time even through mobile devices and apps. Clinical Decision Support Systems (CDSSs) exemplify the application of big data analytics, enhancing patient care by optimizing treatment strategies, ensuring adherence to clinical protocols, and predicting outcomes [15], [16]. These systems, leveraging the wealth of data from EHRs, signify a shift towards an infrastructure that learns and adapts in real-time, embodying the principles of preventative, predictive, and participatory healthcare. The majority of national precision medicine programmes collect various forms of data from hundreds of thousands of people [17]. A multitude of biomaterials and data are made accessible through biorepositories and networks such as the 'UK Biobank' [18] and the U.S. 'All of Us' research cohort [19], aiming to comprise data from at least one million individuals representative of the ethnic diversity of the country, exemplify national precision medicine initiatives that pull disparate data types from hundreds of thousands of citizens. Such initiatives aggregate personal data as a publicly valuable resource, accessible for biomedical research to uncover correlations about biology, lifestyle, environmental exposures, and health outcomes. In a similar vein, other initiatives, such as the Genomics of Drug Sensitivity in Cancer [20] and the Cancer Cell Line Encyclopaedia [21], are creating sizable genomic datasets with the express purpose of examining the relationships between drug sensitivity and genetic indicators in hundreds of cancer cell lines. The evolution of precision medicine into a form of Big Data science has been also accelerated by the ubiquity of sensor-equipped devices (wearable and implantable devices) enabling extensive data collection and the rapid progress in machine learning. These technologies enable continuous monitoring of health parameters, offering a granular view of patient health that informs personalized care strategies. From managing chronic conditions like diabetes to detecting early signs of sepsis, sensing technologies embody the transition towards a more responsive, personalized healthcare system [22]. A defining trait of precision medicine is the engaged role of research participants, who not only contribute their data but

also have the possibility to access this information and receive relevant health findings. This empowerment grants individuals insight into their health-related risks, potentially affecting decisions on insurance and healthcare. Moreover, the rise of social media as a source of big data offers a novel perspective on public health surveillance and patient engagement. Through platforms like Twitter and Facebook, healthcare providers and researchers can monitor public health trends, engage patient communities, and enhance the delivery of health education [23]. This digital dialogue, enriched with insights from environmental and search data, facilitates the early detection of health emergencies and the tailored dissemination of public health messages.

However, there are several obstacles in the way of precision medicine from Big Data. Standardization of data formats, collaborative sharing of data and expertise, and the education of medical researchers in advanced analytics methodologies are among the critical challenges that must be addressed. The ethical, legal, and social implications of data sharing, particularly concerning patient privacy and consent, pose significant barriers to the full realization of Big Data's potential in healthcare. As precision medicine initiatives emphasize data sharing and the public value of Big Data, they confront traditional barriers such as data silos and privacy concerns [24]. The challenge lies in harmonizing the immense variety of data types, ensuring they communicate effectively across formerly isolated platforms. Despite these challenges, the promise of Big Data in advancing precision medicine is undeniable. It offers a path to more effective, personalized treatments, ultimately improving patient outcomes and healthcare efficiency. As we navigate these complexities, the ultimate goal remains clear: to leverage Big Data's vast potential to deliver the right treatments to the right patients at the right time. The dedication to open-source software and the creation of standardised techniques highlight how crucial community-driven initiatives are to tackling the challenges associated with Big Data research [25].

As we have seen, the integration of big data analytics into precision medicine marks a paradigm shift towards more personalized and effective healthcare. This shift is underpinned by our growing ability to gather, analyze, and interpret vast datasets, illuminating the path to tailored treatments. A pivotal component of this data-driven revolution is omics data, with genomics playing a particularly critical role. Precision medicine is based in part on the examination of genomic data, which includes the in-depth analysis of genetic variations. The secrets of understanding human health and illness lie in the intricate interactions between these hereditary variables. In order to provide the foundation for a more sophisticated understanding of the function that genetic data plays in improving precision medicine, the next section will concentrate on the crucial role that omics data, particularly genomics, will play in influencing the direction of medical research and treatment.

1.2 The Omics Revolution and Its Impact on Precision Medicine

The introduction and integration of omics sciences have radically transformed the field of translational medicine over the last decade. These tools, which offer comprehensive insights into the genomic, transcriptomic, proteomic, and metabolic mechanisms underlying health and disease, have revolutionized our approach to studying diseases at the molecular level. A suite of high-throughput experimental technologies, collectively known as "omics," aims to elucidate the complex molecular dynamics of biological systems [26].

The breakthrough in omics research, particularly the publication of the complete human genome, has paved the way for a deeper understanding of the vast array of biological fields, including genomics, transcriptomics, proteomics, metabolomics, and other omics disciplines [27]. The "omics" approach entails a comprehensive evaluation of the sets of molecules, transforming the landscape of biomedical research and enabling a more accurate understanding of the genetic architecture of common diseases.

The thesis will introduce various omics fields, from their foundational role in precision medicine and the direct applicability to understanding disease mechanisms, identifying therapeutic targets, and personalizing patient care. Despite the broader exploration of omics technologies, genomics remains a cornerstone of our investigation, serving as a critical tool for unraveling the molecular basis of diseases and tailoring healthcare to the individual's genetic profile.

1.2.1 Genomics

Genomics, the cornerstone of omics sciences, embarks on the comprehensive study of an organism's entire genome, encompassing both its structure and function. This discipline is instrumental in laying the groundwork for the broader spectrum of clinical omics approaches, standing out due to its fundamental role in unraveling the complexities of genetic information [28]. It serves as a gateway to understanding the intricate details of genetic variations and their profound impact on health and disease. At its core, genomics explores the vast expanse of the human genome, which consists of about three billion DNA base pairs and encodes approximately 20,000 protein-coding genes. A significant focus is placed on the protein-coding regions, or exomes, which represent merely 1–2% of the genome. Despite this small percentage, these regions hold the keys to understanding the genetic underpinnings of various diseases. Beyond the exomes, the remainder of the genome, often misunderstood as 'junk'

DNA, plays essential roles in regulating gene expression and maintaining genomic integrity [29].

Genomic medicine, leveraging advancements such as Genome-Wide Association Studies (GWAS) [30] and Whole Exome Sequencing (WES) [31], has begun to illuminate the genetic variants associated with complex diseases, marking a paradigm shift towards precision medicine. Precision medicine, with genomics at its heart, aims to tailor healthcare to the unique genetic makeup of each individual. It promises a future where diseases can be prevented, diagnosed, and treated with unprecedented accuracy, based on the genetic predispositions and responses of each patient. This approach not only enhances the efficacy of treatments but also minimizes adverse reactions, leading to more targeted and effective healthcare interventions.

1.2.2 Transcriptomics

Transcriptomics is a key subject in molecular science research, providing deep insights into the dynamic nature of gene expression. The field focuses on the transcriptome, which is the full range of RNA transcripts generated by the genome. These transcripts include the well-known messenger RNAs (mRNAs), which act as templates for protein synthesis, as well as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), which are essential for the assembly of proteins. Since it was initially used in the 1990s, the word "transcriptome" has been used to refer to this entire collection of RNA molecules that represents the complete capacity of genome expression inside a particular cell type or tissue [32].

In addition to mRNAs, rRNAs, and tRNAs, the transcriptome encompasses a variety of non-coding RNAs (ncRNAs), such as microRNAs and long ncRNAs, which do not translate into proteins but are increasingly recognized for their regulatory functions. These ncRNAs are instrumental in modulating gene expression and protein activity, highlighting their significance in a myriad of cellular processes. It is astonishing to note that more than 90% of the human genome is transcribed into RNAs, yet only a mere 2% constitutes the coding region. This vast expanse of non-coding sequences generates RNAs that are central to the regulation and complexity of gene expression. By examining the transcriptome, scientists can uncover the intricate regulatory networks that govern biological processes. This exploration not only enhances our grasp of the fundamental mechanisms of life but also opens new avenues for disease prediction and prevention. Transcriptomics, by mapping the expression patterns and regulatory pathways of genes, offers a powerful lens through which we can interpret the complex language of the genome, ultimately guiding us towards novel insights in biology and medicine.

1.2.3 A particular experiment - RNA-Seq Analysis procedure

As depicted in Figure 1.3, a typical RNA-Seq experiment is comprised of three principal (macro)-phases. The initial two phases — Experimental Design and Laboratory Performance — are primarily conducted in a biological laboratory, focusing on the manipulation of wet materials to obtain DNA or RNA sequences. The final phase, data analysis, is a bioinformatics task dedicated to extracting meaningful insights from the sequenced data, employing specialized software tools and methodologies.

The process initiates with the experimental design, wherein decisions are made regarding the type of library to be utilized, the number of replicates to be generated, and the depth of sequencing required. Once the experimental framework is established, the procedure advances to RNA extraction, which is then followed by mRNA enrichment or ribosomal RNA depletion. Subsequently, cDNA synthesis and the preparation of the adapter-ligated sequencing library are undertaken. Sequencing is performed on high-throughput platforms, typically yielding data in the *.fastq* format, although alternative formats may occasionally be employed.

Upon acquisition of the RNA/DNA sequences, computational data analysis is undertaken to ascertain differential gene expression, a focus emphasized in this paper and illustrated in the green boxes in the Figure 1.3. This analysis encompasses the following tasks:

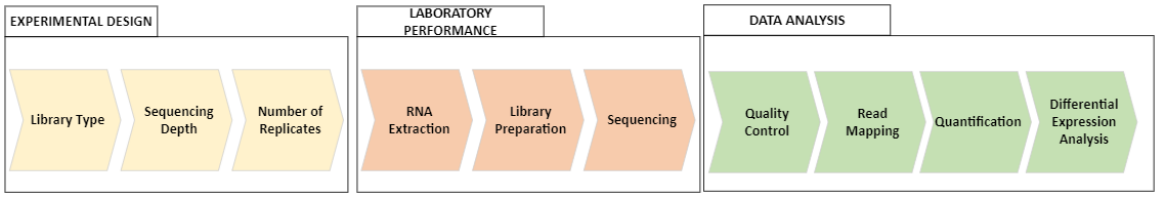


Fig. 1.3 A classical workflow for the RNA-Seq process.

- **Quality Control:** Comprising quality check and trimming steps, this stage is vital for assessing the integrity of the raw data, represented as sequences of reads. These reads are meticulously scanned to identify issues such as low-confidence bases, adapters, duplicates, and base call errors. The necessity for the trimming step is determined based on these assessments— if the sequencing data is deemed suboptimal, improvements are made by excising adapters, low-quality sequences, and duplicates. This step can be omitted if the data quality is already satisfactory. Such measures are imperative to mitigate any potential biases in the final analysis.

- **Read Mapping (or Alignment):** In this stage, reads are mapped to a reference genome or transcriptome, or assembled de novo if a reference is unavailable. The choice of

reference is contingent on factors such as the specific research question, available computational resources, and the nature of the organism under study. This phase involves aligning the sequenced reads to a segment of the reference sequence. It is noteworthy that read mapping is one of the most resource-intensive steps in RNA-Seq data analysis due to the vast number of reads and the large, complex nature of reference genomes or transcriptomes, which require non-contiguous mapping of spliced reads.

- **Quantification:** Following successful mapping, this phase quantifies the number of reads aligned to each gene or transcript (referred to as *features*). Essentially, each read is associated with a specific feature based on its mapping location. The outcomes of this quantification are compiled into an expression matrix (or 'counting matrix'), with each row representing an expression feature (gene or transcript) and each column representing a sample. The matrix entries are typically the actual read counts.

- **Differential Expression Analysis:** The final phase involves identifying genes that exhibit significant expression changes across various experimental conditions. Appropriate statistical models are applied to the normalized counts from the quantification step to extract meaningful information from the RNA-Seq data.

1.3 From Genomics to Variant Calling

The path from genomics to variant calling is critical in the field of precision medicine, underpinning our ability to understand and treat genetic disorders with an unprecedented level of specificity. This process begins with the comprehensive study of an organism's genome, provided by genomics, and transitions to the identification and analysis of variations within that genome, known as variants. Understanding these variations is crucial for diagnosing genetic conditions, predicting disease risk, and tailoring individualized treatment plans. Although each person is unique, our genomes show that, genetically speaking, we are 99.9% alike. All the characteristics that distinguish each of us as unique are caused by the remaining 0.1%. The term "genomic variation" refers to the variations in our genomes, which might range in size and whether or not they affect our health. More precisely, a genetic variant is a difference in the DNA sequence among individuals, groups, or populations. Variants occur when a specific position in the genome differs from the reference sequence used as a standard to compare genetic material. These variations can be as small as a single nucleotide change or can involve larger segments through insertions, deletions, and structural rearrangements.

1.4 Germline vs. Somatic Variants

Understanding the nature and origin of gene variants is crucial in the study of genetics and its application to health and disease [33].

- **Inherited (Hereditary) Variants**: Inherited, or hereditary, variants are passed from parents to their offspring and are present in virtually every cell of the body throughout an individual's life. These are also known as germline variants because they originate in the germ cells—egg or sperm—of the parents. When an egg and sperm cell unite, the fertilized egg contains a mix of DNA from both parents, including any variants they carry. These variants will be present in the cells of the offspring that develop from the fertilized egg. Germline variants are fundamental to the transmission of genetic information and traits from one generation to the next, including predispositions to certain diseases.
- **Non-Inherited (Somatic) Variants**: Non-inherited variants, also called mutations, in contrast, arise during an individual's lifetime and are present only in certain cells, not in every cell of the body. Such variants are typically found in somatic cells (those other than sperm and egg cells) and are therefore referred to as somatic variants. Somatic variants are not heritable and cannot be passed down to the next generation. These changes can result from environmental factors, such as ultraviolet radiation from the sun, or from errors during DNA replication as cells divide.

Most gene variants are benign, contributing to the natural diversity among individuals, including traits like eye color, hair color, and blood type. However, a small fraction of variants can influence the risk of developing certain diseases. Understanding the distinction between germline and somatic variants, along with the concept of mosaicism, is critical for the diagnosis, treatment, and prevention of genetic disorders, as well as for the development of precision medicine strategies tailored to individual genetic profiles.

Types of Variants

As we shift our emphasis to a more in-depth examination inside our thesis, we now provide the kinds of genetic variations that will be essential to our talks to follow. In genomics, each kind of variants has a unique function that advances our knowledge of genetic variety, disease causes, and the possibilities for customised treatment.

- **Single Nucleotide Polymorphisms (SNPs)**: stand as the most ubiquitous type of genetic variation within the human population, each representing a single nucleotide

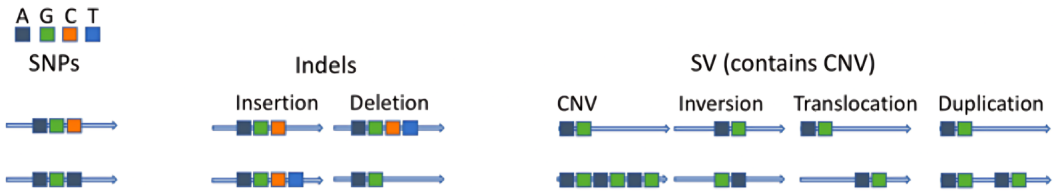


Fig. 1.4 Main DNA Sequence Variations. Single-nucleotide polymorphisms (SNPs) involve the substitution of a single nucleotide. Insertions and deletions (indels) represent the addition or loss of nucleotides. Copy number variations (CNVs), inversions, translocations, and duplications, are larger-scale alterations that can have significant impacts on genetic function and organismal traits (https://www.csc.fi/-/crunching-ngs-data-on-pouta-cloud).

alteration within the genome's vast expanse. These minute changes, involving one of the four nucleotides—adenine (A), thymine (T), cytosine (C), or guanine (G)—are foundational to our understanding of genetic diversity. With an average occurrence of one SNP per 1,000 nucleotides, it's estimated that each human genome harbors around 4 to 5 million SNPs, illuminating the profound variation and complexity of our genetic makeup. The identification and study of SNPs have opened new vistas in genomics, revealing that these genetic markers are not uniformly distributed across the genome but are found both within the coding regions of genes and in the vast stretches of non-coding DNA. The significance of SNPs transcends their mere presence in the genome; they serve as critical markers for locating genes associated with diseases and are instrumental in understanding the genetic underpinnings of various health conditions. Notably, SNPs located within genes or regulatory regions near genes can have a direct impact on health by influencing gene function and expression [34]. The vast majority of SNPs are benign, exerting no observable effect on an individual's health or development. However, a select few have proven crucial in the study of human health, predicting individuals' responses to certain drugs, their susceptibility to environmental factors, and their risk of developing diseases. This has profound implications for pharmacogenomics, where understanding an individual's SNP profile can guide drug therapy, tailoring treatments to reduce adverse effects and enhance efficacy (https://medlineplus.gov/genetics/understanding/genomicresearch/snp/). Moreover, SNPs are invaluable in tracing the inheritance of disease-associated genes within families, offering insights into complex diseases such as heart disease, diabetes, and cancer. Through genome-wide association studies (GWAS) and other genetic research methodologies, genetic variations that contribute to disease susceptibility are identified,

providing a clearer picture of the biological pathways involved and paving the way for novel therapeutic targets.

- **Insertions and Deletions (Indels)**: Indels are variants where nucleotides are inserted into or deleted from the genome. These changes can disrupt the coding sequence of genes, potentially leading to disease. They encompass events less than one kilobase in length (usually less than 50 base pairs) that relate to the insertion and/or deletion of nucleotides into genomic DNA. Moreover, some indels can cause frameshift mutations, which alter the reading frame of a gene and can have significant effects on gene function, potentially leading to severe consequences such as loss of function or gain of harmful function. They are significant in clinical settings, particularly in oncology, as they can activate kinases in cancer, making them targets for kinase inhibitors used in targeted therapies [35].

- **Copy Number Variants (CNVs)**: Structural variants (SVs), including copy number variants, represent a significant and complex type of genetic variation within the genome. SVs encompass a broad range of alterations that affect the structure of the genome, such as inversions, translocations, and more intricate rearrangements that combine multiple forms of changes (https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/cnv). Among these, CNVs are particularly noteworthy as they involve variations in the number of copies of specific regions of the DNA, resulting in either an increase (gain) or decrease (loss) in the copies of those regions. These genomic variations play a crucial role in human diversity, evolution, and disease. CNVs can span anywhere from a few hundred base pairs to millions, covering sizable portions of the genome. Their impact on gene function and expression can be profound, as gains can lead to gene dosage effects, while losses may remove critical genetic information. Inversions involve the reversal of a segment of DNA, while translocations involve the rearrangement of segments between non-homologous chromosomes. The implications of CNVs are wide-ranging; they have been associated with a variety of diseases and conditions, including developmental disorders, neurological conditions, and susceptibility to infectious diseases [36].

1.4.1 The Process of Variant Calling

The accurate identification and interpretation of genetic variants are essential for advancing our understanding of human genetics, disease mechanisms, and the development of targeted therapies. As sequencing technologies and bioinformatics tools continue to evolve, the process of variant calling becomes more refined, enabling more precise and personalized

approaches to medicine. Finding variations in sequence data is a crucial step in the genomics process known as "variant calling." Through this procedure, variations in the reference genome and the sequenced genome, including SNPs, Indels, CNVs, and other genetic changes, may be identified by scientists and medical professionals. Accurately detecting these genetic differences is the main goal of variant calling, as these variations are crucial for comprehending hereditary illnesses, creating personalised medical strategies, and expanding our understanding of human genetics [37].

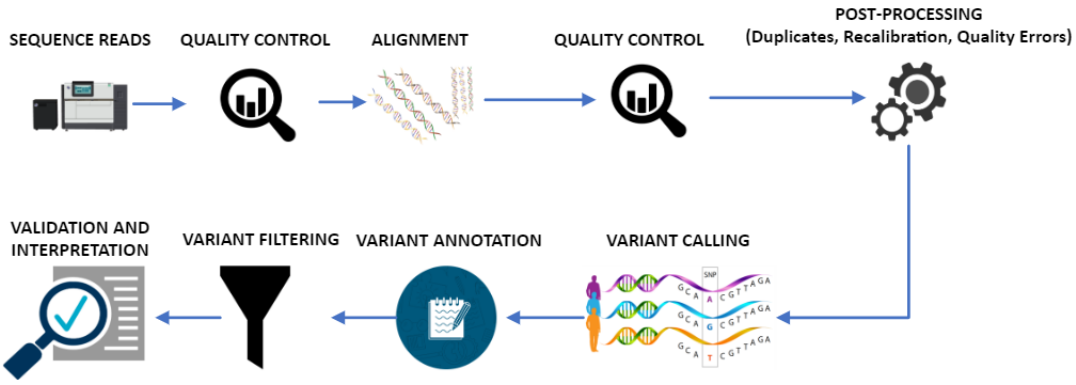


Fig. 1.5 A general Variant Calling Workflow.

The general workflow of variant calling involves several key steps, each contributing to the accurate identification and interpretation of genetic variants [38], [37]. The process typically begins with the collection of DNA samples, followed by sequencing to generate a vast amount of genomic data. This data, often in the form of short DNA sequence reads, is then aligned to a reference genome, a standardized sequence representing the idealized sequence of a species' genome. Alignment is crucial as it locates the sequence reads within the context of the known genome, facilitating the detection of variations. After alignment, the next step involves the actual calling of variants. This process employs algorithms to compare the aligned sequences to the reference, identifying locations where the sequence data differs from the reference. Variants are called based on the evidence provided by the sequence reads, including the quality of the reads, the depth of coverage (the number of times a nucleotide is sequenced), and the agreement among reads about a potential variant.

Following variant calling, the identified variants undergo filtering to remove false positives, which may arise due to sequencing errors, alignment inaccuracies, or low-quality data. Filtering criteria might include the depth of coverage, quality scores, and the consistency of variant evidence across reads. The filtered set of variants is then annotated, a process that provides information about the potential impact of each variant on gene function, protein structure, and, by extension, potential implications for health and disease. Annotation draws

on databases of known genetic variants, functional genomics data, and biomedical literature to interpret the biological significance of called variants.

The final step involves the validation and interpretation of the called variants, often requiring additional experimental or computational approaches to confirm the accuracy and significance of the findings. The result is a list of genetic variants that are likely to be true positives, each with potential biological implications.

1.5 Knowledge Graphs for Integrating Genomics and Clinical Data

Knowledge graphs represent a compelling methodology for the integration and analysis of diverse data sources, standing at the forefront of biomedical research and healthcare innovation. This section delves into the conceptual underpinnings of knowledge graphs, their utility in healthcare, particularly in integrating genomics and clinical data, and exemplifies their applications in advancing precision medicine.

Fig. 1.6 A simple example of KG.

1.5.1 Conceptual Framework of Knowledge Graphs

A Knowledge Graph (KG) is a data structure that interlinks a wide array of information through nodes (entities) and edges (relationships), forming a network that depicts how pieces of information are interconnected, like in the simple example of Figure 1.6. These entities can range from simple concepts or events to complex structured data. The relationships, on the other hand, are not merely connections but carry semantic meaning that defines the nature of the link between the nodes [39]. This structure makes it possible to represent data in a way that is similar to how humans think, which makes it easier to explore and understand complicated datasets in an understandable way [40].

1.5.2 Application of Knowledge Graphs

The foundation of KG lies in semantic web technologies, leveraging ontologies to define and organize the data. Ontologies, in this context, provide a formal vocabulary of terms and the relationships among them, ensuring that the data is not only machine-readable but also machine-understandable [41]. This semantic layer allows for the integration of heterogeneous data sources, ranging from structured databases to unstructured text, under a unified framework. Moreover, in the realm of KGs, emerging technologies like Artificial Intelligence (AI) and machine learning are revolutionizing the way we construct and analyze complex biomedical data. These technologies enhance the capability of knowledge graphs by automating the identification of patterns and relationships within vast datasets, including genomics. AI algorithms can predict new connections and insights, making knowledge graphs more dynamic and insightful [42],[43]. This evolution in technology is critical for advancing precision medicine, as it enables a deeper, more nuanced understanding of the genetic and clinical factors influencing health and disease.

Knowledge Graphs in Precision Medicine

As we delve deeper into the realms of precision medicine and genomics, the integration of diverse biomedical data becomes paramount for advancing personalized healthcare. The complexity of genomic information, combined with the vast array of clinical data, presents a significant challenge for data management, interpretation, and application in clinical settings. This is where KG come into play, offering a dynamic and interconnected framework for capturing, organizing, and querying complex data relationships. In precision medicine, KGs facilitate the tailoring of medical treatment to the individual characteristics of each patient. By integrating genomics data with clinical outcomes, healthcare providers can identify genetic markers associated with disease susceptibility and drug response. This integration enables the prediction of disease risk, the prevention of adverse drug reactions, and the selection of optimal therapeutic strategies for individual patients, thereby enhancing treatment efficacy and patient safety. KGs are instrumental in precision medicine for tailoring medical treatments to the unique characteristics of each patient. By synergizing genomics data with clinical outcomes, KGs enable healthcare providers to identify genetic markers tied to disease susceptibility and drug responses. This holistic approach fosters the prediction of disease risks, the mitigation of adverse drug reactions, and the formulation of optimal therapeutic strategies tailored to individual patients, thereby elevating treatment efficacy and patient safety.

Within the scope of precision medicine, KGs find other critical applications [39]. Indeed, KGs are pivotal in accelerating the discovery of new drugs or repurposing existing ones by elucidating relationships among drugs, genes, diseases, and biological pathways. This methodology unveils novel drug-disease associations and predicts the effectiveness of existing medications for treating new conditions, significantly reducing drug development time and cost while addressing unmet medical needs. KGs also enhance our comprehension of the intricate biological mechanisms underpinning diseases. Integrating genomics data with clinical insights and biomedical literature through KGs reveals patterns and correlations that shed light on disease etiology and progression. This comprehensive perspective is key to identifying potential therapeutic targets and biomarkers for disease diagnosis and prognosis, propelling our understanding of human health and disease forward [44].

1.5.3 Examples of Knowledge Graph Applications in Healthcare

Several pioneering projects and platforms have demonstrated the potential of knowledge graphs in healthcare. The Human Phenotype Ontology (HPO) project utilizes knowledge graphs to associate genetic disorders with phenotypic abnormalities, facilitating the diagnosis of rare genetic diseases [45]. Another example is the use of knowledge graphs in the Cancer Genome Atlas (TCGA) project, which integrates genomic and clinical data to improve our understanding of cancer biology and guide the development of targeted therapies [46]. While KGs offer significant advantages, their implementation is not without challenges. Data quality, interoperability, and the need for advanced analytics capabilities are critical considerations. Moreover, ensuring the privacy and security of sensitive health information within KGs is paramount [47].

Chapter 2

Unveiling the Reproducibility Crisis in Bioinformatics Pipelines

In this chapter, we conduct a comprehensive mapping study of RNA-Seq pipelines in the field of oncology, which serves as a lens to scrutinize the reproducibility crisis in bioinformatics. This investigation allows us to map out the current landscape and identify key reproducibility challenges within bioinformatics pipelines. The findings underline the importance of dataset retrievability as a critical factor impacting reproducibility. To address this issue, we introduce a decision tree approach, designed to evaluate and enhance the retrievability of datasets and experimental settings, pinpointing it as a pivotal concern in the realm of bioinformatics reproducibility.

2.1 Motivation

Research in the fields of biology and medicine has entered a new era with the introduction of Next Generation Sequencing technologies. These technologies, which are distinguished by their greater efficiency and speed, have significantly improved our capacity to examine the molecular details of different cell types that come from a variety of organisms and environmental contexts. A key feature of NGS technologies is their ability to generate large datasets consisting of millions of nucleotide sequences, or *reads*. The advent of new data formats and experimental typologies brought about by this data explosion has made the creation of advanced tools and processes necessary. These improvements play a critical role in both aiding the future processing of these reads and in assembling them into larger, coherent sequences. Gaining thorough understanding of intricate biological processes is the goal.

RNA-Sequencing (*RNA-Seq*), a method that has become essential in the investigation of genomic sequences, is one particularly notable application of NGS. Differential expression analysis, which is entirely based on RNA-Seq, compares gene expression profiles under various experimental settings. These circumstances can include distinct cell populations or the same cells exposed to several drugs. This analytical method is particularly useful in oncological research, since it can provide important insights into the differences in gene expression between malignant and healthy cells as well as the effects of pharmacological therapies. These revelations could then help us better understand cancer biology and guide the creation of specific treatment approaches.

Despite the promise and utility of RNA-Seq, researchers often face challenges in selecting suitable pipelines and tools for the reproducible analysis of RNA-Seq data. The diversity and complexity of available tools and the variability in experimental conditions compound this challenge. This study, therefore, aims to furnish a comprehensive guide for the analysis of RNA-Sequencing experiments, with a particular emphasis on differential expression analysis in oncological contexts. By conducting an extensive review of the prevalent pipelines and their applicability under varying conditions, this chapter seeks to aid researchers in making informed decisions regarding pipeline selection.

In this chapter, we aim to quantify tool usage across various RNA-Seq phases, thereby illuminating the landscape of pipeline selection in cancer research. This analysis not only contributes to the understanding of the most prevalent methodologies but also highlights areas where tool selection is underreported, echoing the concerns raised in [48] about the challenges of reproducibility in RNA-Seq experiments. By providing a detailed examination of pipeline usage, our study seeks to enhance the clarity and reproducibility of RNA-Seq research in the field of oncology.

2.1.1 Related Works

The work presented in [49] provides a broad overview of RNA-Seq applications but does not concentrate on specific aspects, nor does it delve into the most widely used pipelines. While it occasionally references prominent tools used in different stages of RNA-Seq, its scope extends beyond the differential expression analysis, encompassing additional RNA-Seq applications like result visualization and gene fusion discovery. In contrast, [50] narrows its focus to the differential expression phase, outlining the primary tools used and dissecting their respective advantages and limitations. Similarly, [51] offers a detailed critique of these tools, specifically from the perspective of their statistical models and underlying assumptions. [52] provides a comprehensive description of a classical RNA-Seq experiment, spanning from sequencing to pathway enrichment analysis. It generally discusses the tools

employed in various stages and attempts to elucidate methodological differences. This practical approach is further exemplified in [53], which presents a complete pipeline using existing bioinformatics tools, offering an actionable guide for pipeline development. The steps delineated in [54] mirror those in [52], yet the focus is limited to a select set of tools without extensive comparison or description. [55] offers a brief overview of RNA-Sequencing applications such as differential expression analysis and short-read mapping. Although it lists popular tools, it does not extensively cover all phases of the applications, and recent tools are notably absent due to its earlier publication. The study in [56] spans from data preprocessing to alignment and differential expression in RNA-Sequencing. It uniquely compiles a list and descriptions of existing bioinformatics tools and software for each phase, alongside relevant platforms for tool usage and data analysis. However, the list provided does not necessarily reflect the most frequently used pipelines. [57] surveys various RNA-seq applications like alternative splicing and variant detection. It discusses tools used in different phases but stops short of comparing them. A thorough step-by-step description of the RNA-Seq process and its tools is offered in [58], which includes a comprehensive collection of tools, even recent ones. However, it mainly focuses on tool taxonomy rather than their applications. [59] is particularly notable, comparing 278 pipelines composed of various sequencing mapping, quantification, and normalization tools. Despite its comprehensive approach, it does not proceed to the differential analysis phase and lacks specific insights into its actual usage in oncology.

Our research differs significantly from the aforementioned works:

Unlike [49], [50], [55], [57], [58], and [51], our study is not merely descriptive but rather quantitatively focused. We aim to identify and quantify the most utilized pipelines in tumor-related RNA-Sequencing experiments, spanning from sequencing to differential expression analysis. Diverging from [56], [52], [53], and [54], our study does not stay on tool differences. We concentrate exclusively on tumor-related pipelines without asserting the superiority of any particular tool or method. While [59] delves into pipeline concepts and compares a multitude of them, our study extends to tools used in differential expression analysis and examines their prevalence in oncological research, a perspective not covered in [59].

2.2 Automating the Mapping Study

In the fast-paced domain of bioinformatics, especially within clinical genetics, the application of Systematic Mapping Studies (SMS) becomes pivotal. Unlike Systematic Literature Reviews (SLR), which delve into each paper's methodologies and results, SMS aims to effi-

ciently survey scientific literature to extract specific information of interest [60]. This study's goal is to navigate the diverse landscape of RNA-Sequencing (RNA-Seq) analysis tools, specifically targeting studies involving complete genomes or transcriptomes and excluding gene panel research. Our approach, delineated across five structured phases—ranging from the initial definition of research questions and criteria to the comprehensive development of a classification schema—is meticulously visualized in Figure 2.1. This process includes crafting and executing a detailed search strategy, selecting pertinent papers based on stringent criteria, conducting keywording and in-depth full-text analysis, and culminating in a multidimensional visual schema that encapsulates the research landscape, particularly focusing on oncology [61].

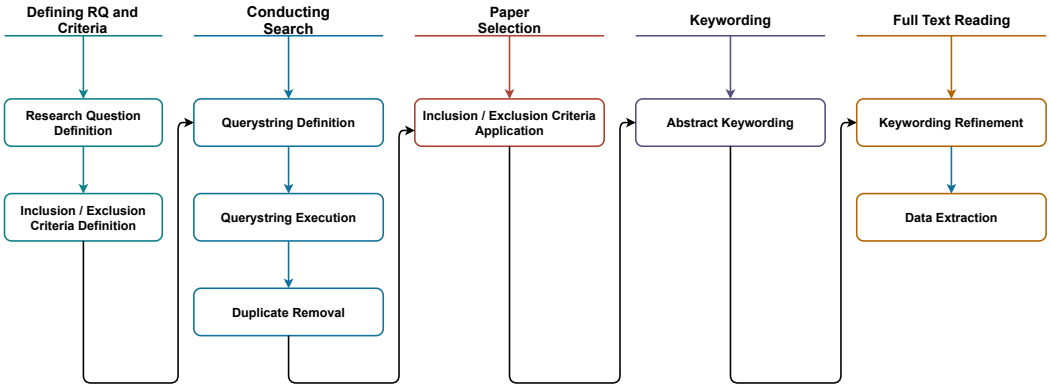


Fig. 2.1 The mapping study process, separated by phases. In the coloured rectangles we identify all the steps. The arrows show the steps followed in each phase.

2.2.1 Defining RQ and Criteria

In this subsection, we initially present the specific research questions (RQs) that are foundational to our systematic mapping study, as detailed further in Section2.2.1. Following this, we define the inclusion and exclusion criteria within Section 2.2.1, which act as the primary filters for the literature screening process. These carefully crafted criteria are intended to guide our selection process, ensuring it remains systematic and closely aligned with the objectives of our investigation.

Research Question Definition

Rooted deeply in a specific area of study with a keen focus on the reproducibility of experiments, our investigation emphasizes key factors including the datasets employed, the types of tumors analyzed, thresholds set for differential expression analysis, the array of tools

used throughout various stages, and the settings applied. This targeted focus underpins our study's primary objective, which is to explore and answer critical research questions that arise within these parameters, aiming to enhance the reliability and replicability of bioinformatics experiments in this domain:

- **RQ1**: *Which pipelines are most prevalent in RNA-Seq data analysis for tumor studies, specifically aimed at differential expression?* This question is pivotal, as understanding which pipelines are favored can shed light on their usability and reliability compared to others, which is vital for the reproducibility of experiments—a cornerstone for ongoing research integrity. We aim to dissect RQ1 further by examining the specific tools used in various phases of analysis and evaluating how thoroughly these tools are described in the literature, ensuring a comprehensive understanding of each pipeline's application and potential for replication.

- **RQ2**: *What cancer types are most frequently studied using NGS technology with a focus on differential expression?* Our goal is to pinpoint the tumor types that are the focus of significant research efforts in this domain. Understanding the prevalence of certain tumor types in research could reveal which datasets are more accessible and widely used, providing insights into current trends and gaps in oncological research leveraging NGS technology.

- **RQ3**: *What are the defining characteristics of datasets used in analyzed experiments?* Given the diverse nature of RNA-Seq analysis, we scrutinize how the characteristics of datasets influence the selection of specific pipelines. This exploration acknowledges the reality that no single pipeline serves as the gold standard for all RNA-Seq studies, highlighting the importance of dataset attributes in guiding pipeline choice.

- **RQ4**: *What are the experimental settings of the analyzed experiments?* This question investigates the specific experimental setups utilized in studies, focusing on aspects crucial for experiment replicability. It emphasizes the alignment and differential expression analysis phases, known for their impact on result consistency. Understanding the choice of reference genome versions and differential expression thresholds is essential, as these parameters significantly influence the identification and interpretation of gene significance within the experimental context.

Inclusion/Exclusion Criteria Definition

We crafted specific inclusion and exclusion criteria focusing on titles, abstracts, and full texts, especially valuing full texts for deeper insights where abstracts fall short. Our criteria

aim to include studies detailing comprehensive pipelines or those utilizing novel or recent tools in their analysis phases. Moreover, works that provide thorough discussions on RNA-Sequencing, offering broad overviews and practical insights, are particularly valued for their potential to contribute significantly to the field.

Inclusion Criteria:

- Papers that illustrate one or more pipelines.
- Studies with a comprehensive focus on the RNA-Sequencing process.
- Inclusion of potentially noteworthy tools.

Our exclusion criteria are meticulously designed to filter out studies not pertinent to our primary objectives, including those not focused on tumor research via RNA-Sequencing, conference materials lacking comprehensive details, and any retracted publications. Furthermore, we specifically exclude research on non-human subjects and studies that only address a single step of the RNA-Sequencing process without offering a holistic experimental view. This ensures our review remains focused on human tumor studies, incorporating comprehensive experiments and innovative methodologies.

Exclusion Criteria:

- Papers focusing only on a portion of the RNA-Sequencing process.
- Non-human data studies.
- Studies not aligned with our objectives.
- Research RNA-Sequencing based but not cancer related.
- Conference posters or abstracts.
- Retracted studies.

The outcome of this phase defines the scope for the publications, guiding the subsequent query formulation.

2.2.2 Conducting the Search

This subsection encapsulates the strategy and execution of our literature search. It begins with the *Query String Definition*, where we outline the construction of a targeted search string from key terms derived from our research aims. This is followed by *Query Execution*,

detailing the process and outcomes of running the search string to gather relevant studies. Together, these steps form a methodical approach to identifying the primary studies that will form the basis of our systematic mapping study.

Query String Definition

The search for primary studies is a pivotal step, particularly as our focus is within the biomedical domain. We chose PubMed Central (PMC) https://www.ncbi.nlm.nih.gov/pmc/ as our primary repository due to its comprehensive archive of full-text articles. To capture recent developments, we limited our search to the past six years, allowing us to identify current trends and evaluate the usage of both new and possibly outdated tools.

We distilled essential search terms from our research questions to craft an effective query string. This string was constructed to encompass terms related to differential expression analysis in tumor studies using RNA-Seq. The resultant query string on PMC is provided below.

Query String Execution

We executed the formulated query string on PMC on November 20, 2023. This comprehensive search was intentionally designed to encompass all relevant papers published over the last eight years, from 2016 to 2023. This specific time frame allowed us to capture a broad and up-to-date perspective of the advancements and trends in this rapidly evolving domain of RNA-Seq (in oncology). The search yielded a significant corpus of 5202 papers, reflecting the extensive research activity in this area during the specified period.

Duplicate Removal

Given that our search was conducted within a single database, we encountered few duplicates. After identifying and removing 8 duplicate papers, the total count of selected papers was adjusted to 5194.

2.2.3 Paper Selection

From the initial dataset, applying our defined criteria led to the selection of 212 publications, as illustrated in Figure 2.2.

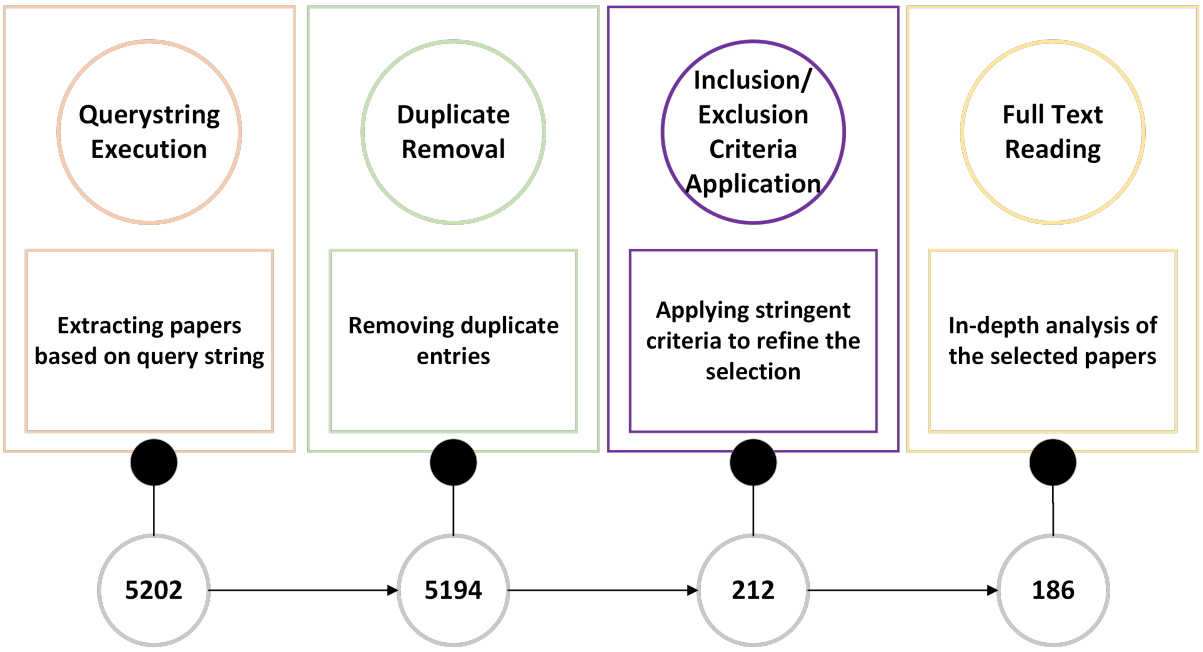


Fig. 2.2 Sequential phases from initial search to final paper selection, with corresponding paper counts.

2.2.4 Keywording

In preparation for extracting information from our selected papers, we initiated a classification scheme through keywording of abstracts. Initially, we employed Wordstat (https://provalisresearch.com/products/content-analysis-software/) software for automation, inputting the abstracts into an Excel file. However, the specialized medical nature of the publications meant the specific details we sought were not adequately extracted. This led to additional manual screening to pinpoint the critical attributes of the publications relevant to our research needs, ensuring a thorough and tailored analysis of the literature for our study's objectives.

2.2.5 Full Text Reading

We conducted a full-text analysis of all 212 selected papers. During this phase, 25 papers were removed as they were secondary studies, lacked detailed experiments, or did not provide the precise data necessary for our analysis. Consequently, we finalized 186 papers for data extraction.

Keyword Refinement

A challenge we encountered was the heterogeneity of the papers' structure, with relevant information often buried within specific sections like "bioinformatics analysis" or "supplementary material." To address this, we manually entered the full text of each publication, including supplementary materials, to extract the required information. This step formed the basis for the subsequent phase of structured information extraction.

Data Extraction

Following our established schema, we extracted data from the selected studies. Our focus on experimental reproducibility and the tools used in various pipelines led us to concentrate on three main aspects: the dataset, the tools in each RNA-Seq phase, and the experimental settings.

2.3 Results

To clarify the plots in upcoming sections, the label *NOSPEC* is used when a specific phase is not mentioned in an experiment, while *NOPROV* indicates phases that were definitively not conducted. The term ND represents a combination of NOSPEC and NOPROV across phases. It's important to note that the total count of papers per stage in some plots, like in Figure 2.4, might not equal the overall selected paper count. This discrepancy arises because documents often utilize multiple tools within a single phase to enhance analysis, and each tool has been accounted for individually.

2.3.1 Which pipelines are most prevalent in RNA-Seq data analysis for tumor studies, specifically aimed at differential expression?

Addressing the question of prevalent pipelines in tumor study analyses aimed at differential expression, our study reveals no uniform standard across the board, echoing findings from [49]. While some tools are favored in certain experimental phases, their selection may hinge more on user familiarity than on proven superiority. Our examination uncovered a preference for specific tools, indicating a level of trust in their performance (as shown in Figure 2.4). Notably, a comprehensive pipeline specification within studies is rare, underscoring a variety in approach. Among those, pipelines integrating *fastqc / trimmomatic / star / featurecounts / DESeq2* and *fastqc / trimmomatic / star / htseq / DESeq2* emerge as the most cited. Despite the frequent omission of quality control and trimming phases, the *tophat2 / cufflinks* pipeline,

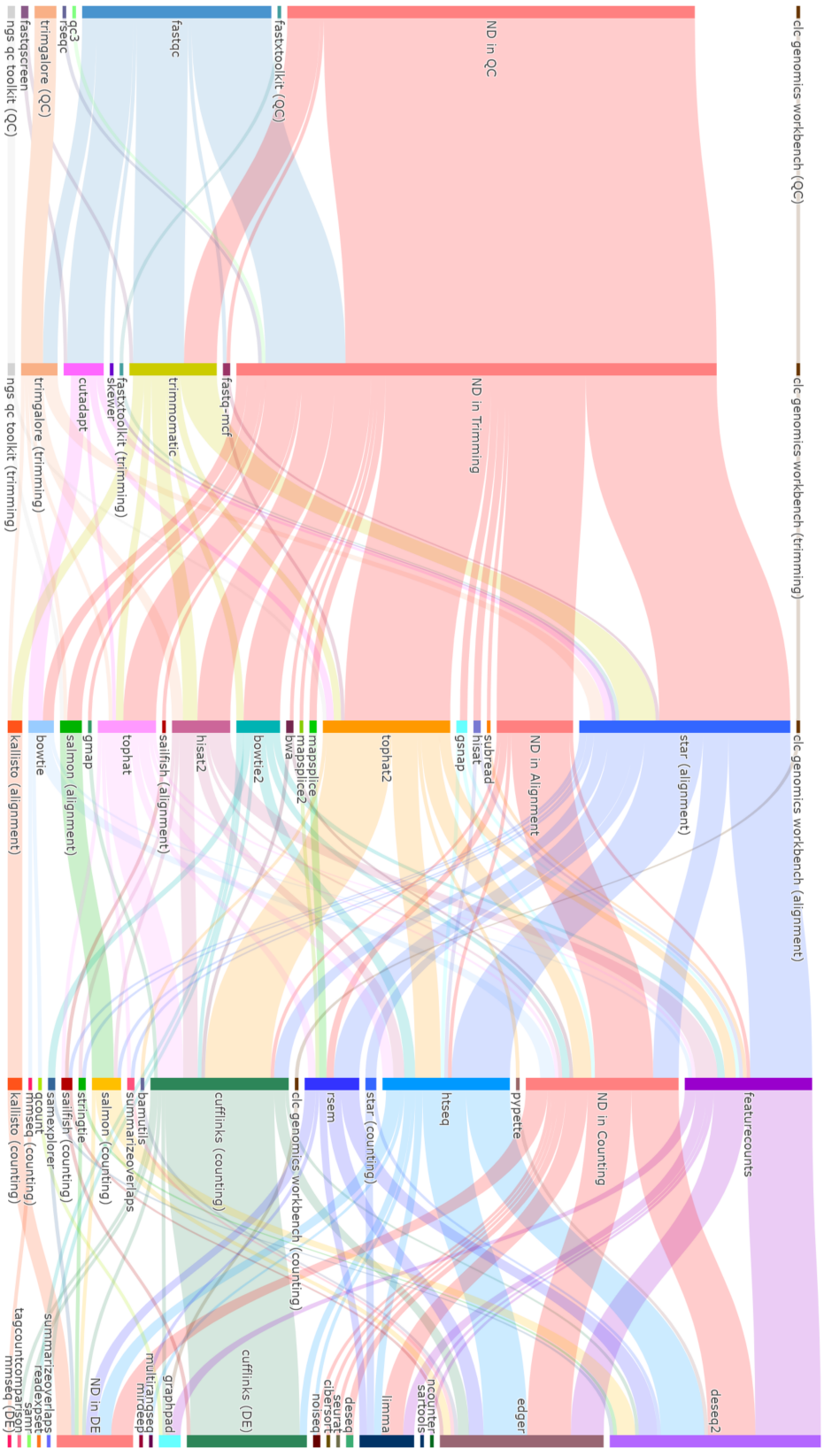


Fig. 2.3 Used tools, separated by phases.

known as the Tuxedo Pipeline, stands out for its widespread adoption (71 selected papers in total). The analysis underscores the diverse methodologies employed, with no single pipeline dominating the field.

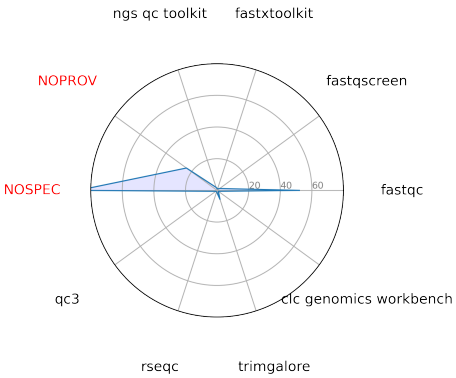
Moreover, our investigation delves into the tool preferences across each stage of RNA-Seq analysis in tumor studies. We adopt a granular approach, examining tools utilized at every process stage, represented through spider plots for clear visualization (as shown in Figure 2.4). This detailed exploration, extending to pathway analysis as a natural continuation of DE analysis, aims to gauge the selected studies' comprehensiveness. Our findings highlight a diverse usage pattern, with no single pipeline dominating. Despite this variety, certain tools emerge as prevalent in specific phases, underscoring patterns that may not be immediately evident. For instance, the initial quality control phase often lacks specificity or execution in many studies, while tools like *fastqc* and *trimmomatic* show prominence in their respective phases. The alignment phase reveals a preference for *star* and *Tophat2*, aligning with literature that underscores their efficacy ([62], [63]). Interestingly, the DE analysis phase shows a detailed specification of tools, with *DESeq2* and *edgeR* leading the usage. To provide a complete picture, we also examine the enrichment analysis step, recognizing its role in understanding the biological impact of differential expression.

To grasp the completeness of pipeline documentation in RNA-Seq analyses for tumor studies, we delve into how each study specifies the experimental phases, contrasting those marked with NOSPEC and NOPROV labels for unspecified or unperformed stages (Figure 2.5). Despite a general lack of specificity across many papers, a notable portion thoroughly details their methodologies, with 28 papers documenting all steps and another 36 covering at least four. This analysis, particularly illuminated in the progression depicted in Figure 2.6, showcases a trend where initial phases are more frequently detailed than later ones, emphasizing a pattern of diminishing specificity as the process advances. This pattern not only highlights areas where methodological transparency wanes but also pinpoints the critical need for complete documentation to enhance experiment applicability. We report in the following the 28 selected papers that specify the tools used in all phases:

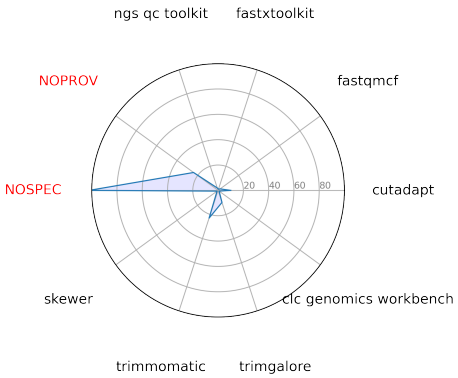
- [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91]

2.3.2 RQ2 What tumor types are most frequently studied using NGS technology with a focus on differential expression?

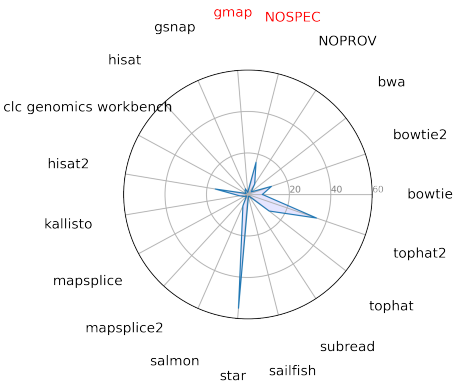
There is a strong tendency among the selected documents to analyze datasets consisting of different sets of tumors through the RNA-Seq process. This is because almost all the



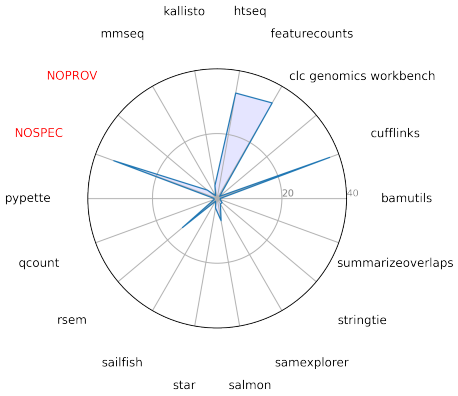
(a) Tools in quality control phase.



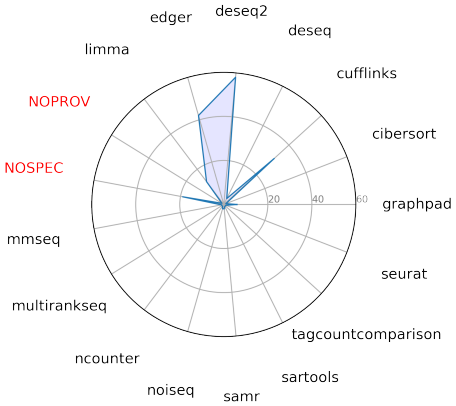
(b) Tools in trimming phase.



(c) Tools in alignment phase.



(d) Tools in counting phase.



(e) Tools in DE analysis phase.

Fig. 2.4 Phase-by-phase pattern.

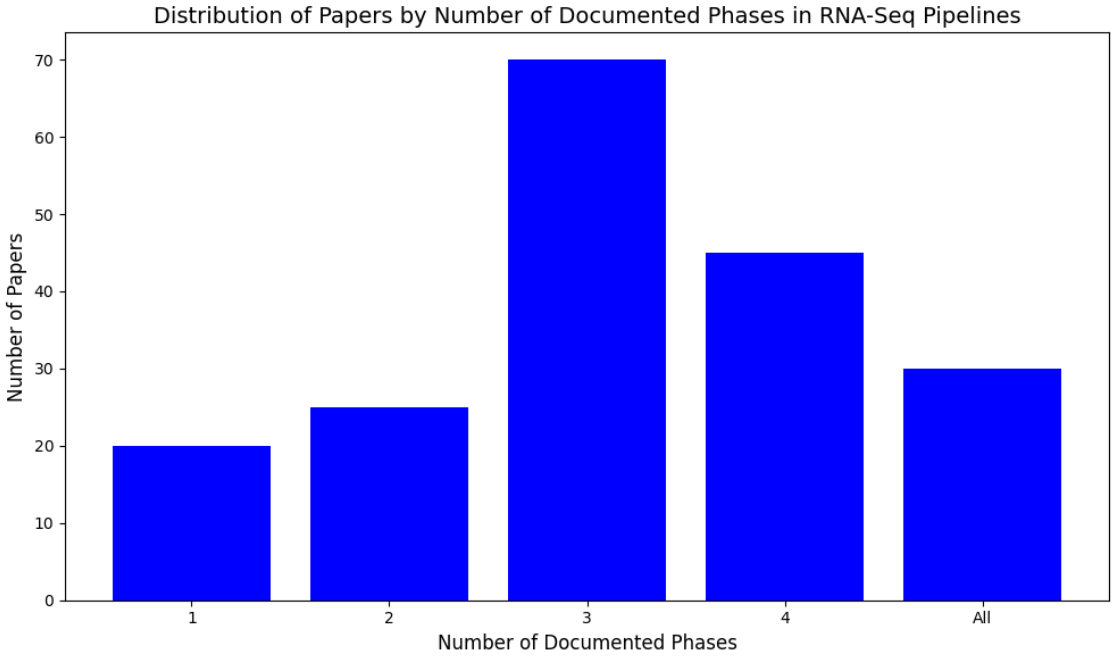


Fig. 2.5 This figure illustrates the distribution of research papers based on the number of RNA-Seq pipeline phases they document. The x-axis represents the number of phases specified, ranging from 1 to all phases, with the bars corresponding to the number of papers that disclose specific tools used for each phase. The y-axis indicates the number of papers. The data is sorted incrementally from 1 phase to all phases, providing a clear view of how comprehensively each paper documents its pipeline.

experiments try to establish the effectiveness of a given treatment on different types of tumors (183 papers), while others are works that establish the effectiveness and the speed of execution of a certain pipeline on datasets (3 papers: [67], [92] and [93]). Figure 2.7 catalogs the tumors investigated within these studies.

2.3.3 RQ3 What are the main characteristics of the used dataset in the analyzed experiments?

To explore the third research question regarding the main characteristics of datasets in analyzed experiments, we integrate responses from subdivided inquiries into a cohesive analysis. This encompasses evaluating dataset sizes as described by runs due to the frequent omission of total patient numbers in the selected works, prioritizing 95 papers that detailed dataset dimensions. Therefore, excluding the papers that did not specify the dataset, because it was not made available or only partially (Figure 2.8).

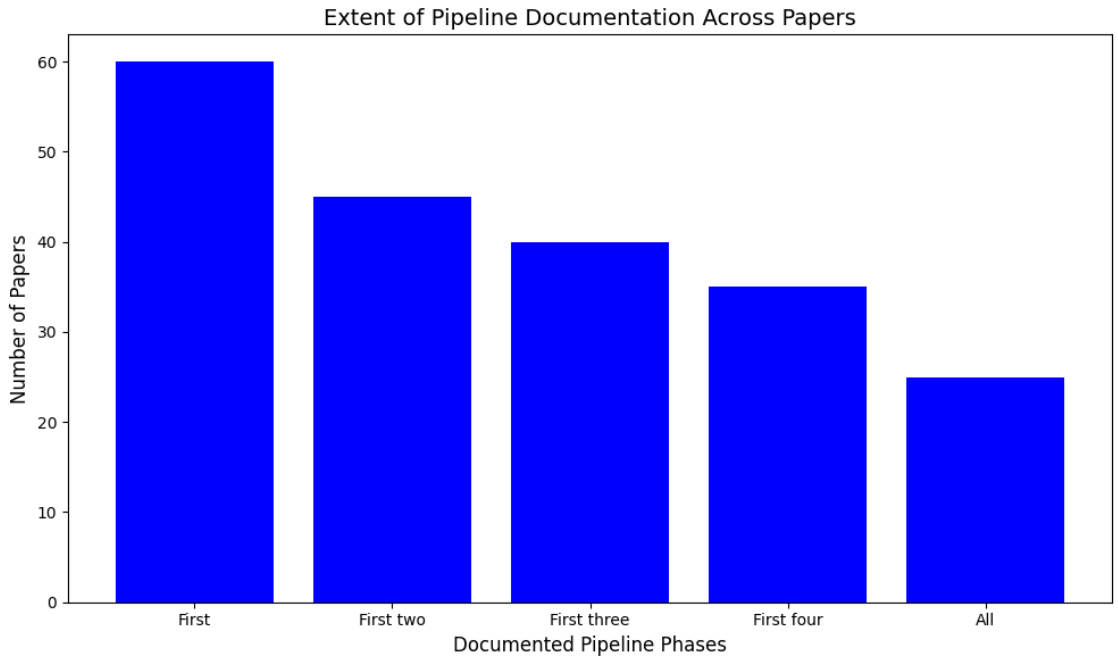


Fig. 2.6 This figure shows the extent of pipeline documentation in each paper, beginning with the Quality Control phase. 'First' means only the initial phase is detailed, with subsequent terms indicating progressively more phases documented. Specifying a phase involves naming the tools used.

The analysis, as depicted in Figure 2.9, highlights that the majority of studies utilize the Illumina platform for sequencing. However, some studies have not disclosed their sequencing platform. Regardless of the platform, the focus of our analysis shifts towards the sequencing design, specifically whether single-end or paired-end reads were used. The figure illustrates a clear preference for paired-end designs, which are crucial for enhancing alignment accuracy during the analysis phase, as supported by [94]. This preference underscores the importance of sequencing design in achieving precise results.

2.3.4 RQ3.3 Are the used dataset in the analyzed experiments publicly available?

First of all we want to say that with "publicly accessible dataset", we mean that in the paper a link is given to it or a special code able to localize it on online publicly accessible data archives (we have usually searched the SRA Archive). With the aim to answer this RQ, the selected papers include three main cases: dataset available, dataset not available or dataset partially available. In the latter case, it means that not all the runs have been uploaded and made available on online repositories. On the other hand, when we are faced with an

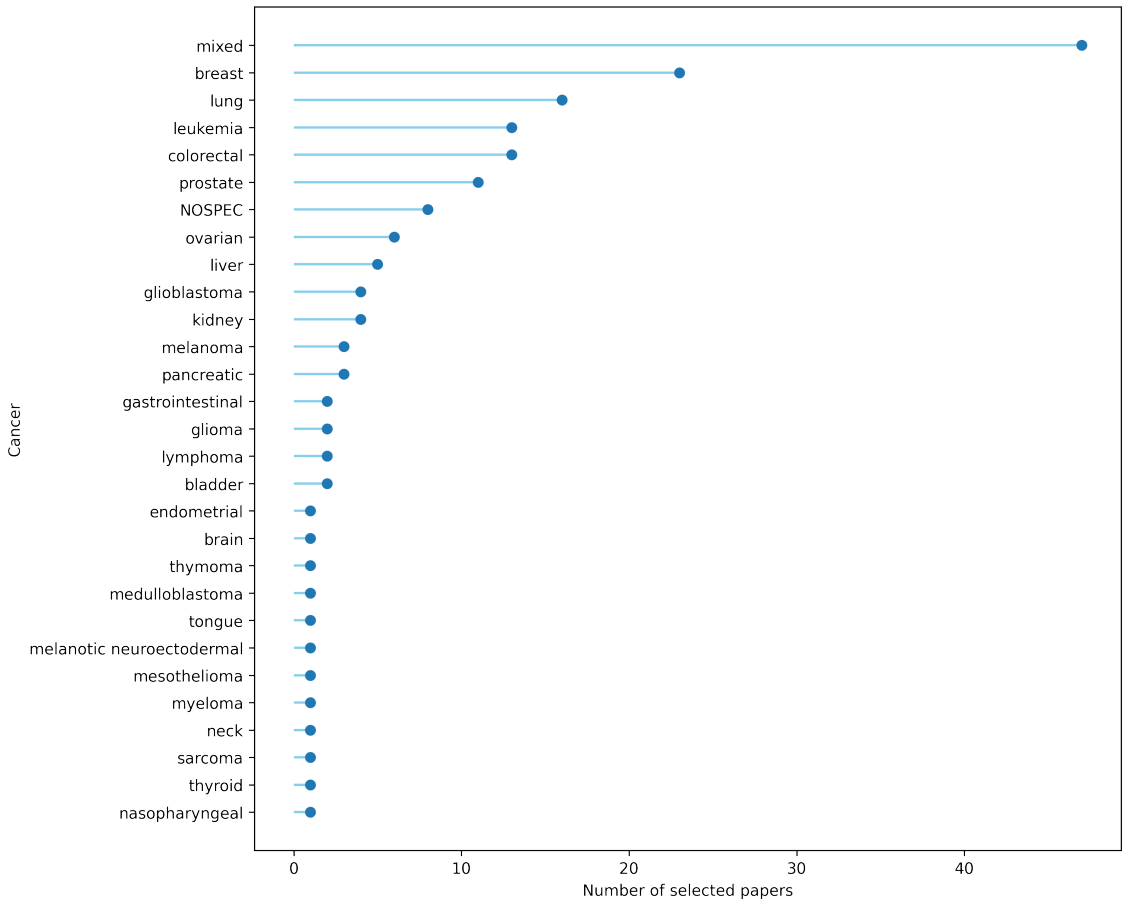


Fig. 2.7 Cancer distribution for experiment. The term "mixed" means a work that has taken into consideration the study of different types of tumor.

unavailable dataset it is because the passage of the dataset in a confidential manner between research groups is required. We note that a large percentage of the total papers make their used datasets accessible (89 papers), compared to those that make the dataset accessible on demand (73 papers). Finally, we have 14 papers where not all the runs are available.

2.3.5 RQ4 What are the experimental settings of the analyzed experiments?

To delve into Research Question 4 regarding the experimental setup of studies, we consolidate our examination of both the reference genome versions utilized and the thresholds applied for differential expression analysis. Our analysis underscores the critical role of genome version selection, noting a division between studies employing older versions like hg19 or GRCh37 and those adopting newer versions such as hg38 or grch38. The usage of the stable version,

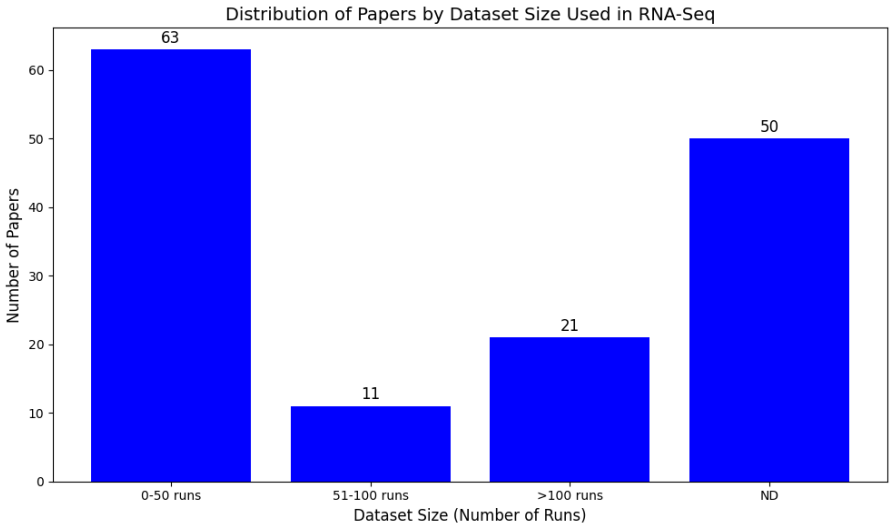


Fig. 2.8 Dataset size. We note that most of the papers do not specify the dataset used (ND label in the figure). Then we have that a part of the selected papers use RNA-Seq on a limited number of runs. (0-50 runs corresponds to 63 papers). Finally we have that 51-100 runs corresponds to 11 papers and >100 runs to 21 papers.

which corresponds to the version hg19 or GRCh37 (60 papers and 21 papers, respectively), prevails. This choice significantly impacts the experiment's replicability and accuracy in variant detection [95]. The genome version that was utilised is not specified in a number of articles (40), which undervalues this information in terms of future repeatability. Finally, a tiny subset of three studies has been chosen to use hs37, a significantly altered variant of the hg19 genome.

Considering the Table 2.1, in examining experimental settings, we assess the commonly used thresholds for two critical measures in differential expression (DE) analysis: padj (indicating statistical significance) and fold change (denoting biological significance). An analysis reveals that 69 studies did not specify either threshold, affecting clarity in DE analysis. Conversely, 33 studies detailed only the padj values without mentioning fold change, and a few (six papers) did the opposite. Notably, the combinations Padj = 0.05 / Fold Change 1 and Padj = 0.05 / Fold Change = 2 are most prevalent, underscoring standard practices in threshold selection for these experiments.

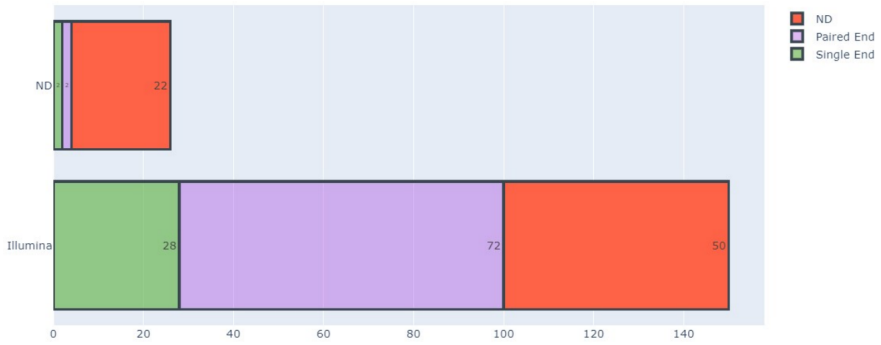


Fig. 2.9 Platform and layout distribution.

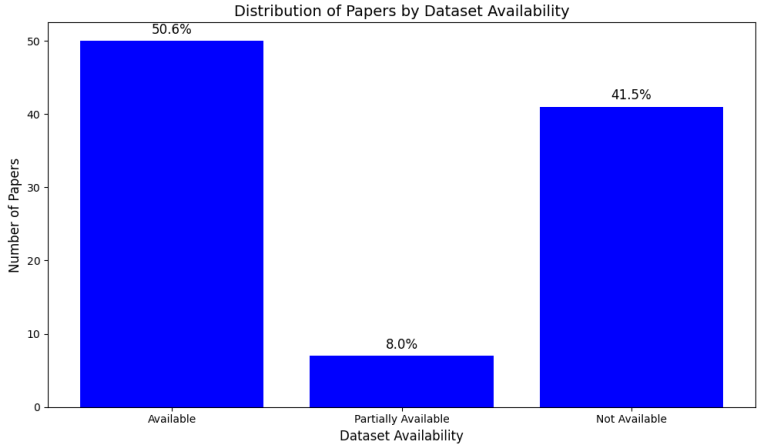


Fig. 2.10 Dataset availability. The possible cases are the following: dataset available (available), dataset not available (not available), dataset partially available (partially available)

2.4 Assessing Reproducibility: A Decision Tree Approach for Dataset Accessibility

Drawing from the principles established in the survey's results on data reproducibility, we built a unified approach that extends our discussion on dataset retrievability. Thus, incorporating the additional details and clarifications provided, we further refine our discussion on enhancing data reproducibility through a structured approach to dataset retrievability. Our methodology, underscored by a comprehensive decision tree, serves as a fundamental guide for researchers in navigating the complexities of making datasets accessible for subsequent research endeavors. This decision tree, illustrated in the referenced figure and available on the Zenodo repository [96], systematically addresses the pivotal aspects of dataset accessibility

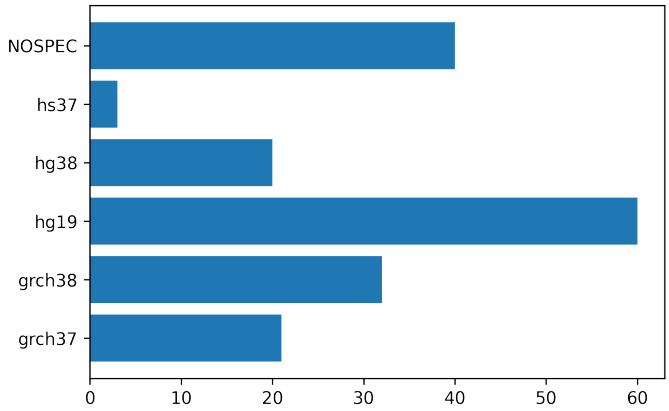


Fig. 2.11 Genome versions in dataset.

and processing methods, laying the groundwork for our commitment to bolstering scientific transparency and reproducibility.

The decision tree is ingeniously designed with eight main interrogatives, structured as internal nodes, leading to specific recommendations encapsulated in the leaf nodes. These recommendations are meticulously tailored to guide researchers in providing the necessary metadata or materials corresponding to each unique dataset scenario. The bifurcation of the tree into two principal domains is predicated on the dataset's shareability: one path explores scenarios where the dataset cannot be shared, delving into alternative access methods or the implications of restricted information due to regulatory constraints. The other pathway investigates scenarios where the dataset is shareable, further distinguishing between raw and preprocessed data forms and elucidating the requisite steps for each, including public accessibility, preprocessing methodologies, and the specifics of data processing tools or scripts.

For datasets that are not shareable, our approach mandates a clear delineation of available alternative access mechanisms or a justification for the reproducibility limitations imposed by regulatory frameworks. This ensures that, even in cases of restricted access, efforts are made to outline the extents of dataset availability and the conditions under which they can be accessed.

Conversely, when datasets are shareable, our schema meticulously distinguishes between raw and preprocessed datasets. It guides researchers through a series of decisions regarding the public availability of these datasets, the accessibility of preprocessing methods, and the detailed instructions for data processing, whether it involves scripts, tools (including their versions and configurations), or other methods. This structured approach not only facilitates

Table 2.1 Association between thresholds

Padj	Fold Change	No of Papers
ND	0.5	1
	1	3
	2	2
	ND	69
0.01	0.5	1
	0.75	1
	1	5
	1.5	4
	1.7	1
	2	2
	ND	6
0.05	0.25	1
	0.5	1
	0.6	2
	1	16
	1.5	7
	2	26
	3	1
	ND	27

the disclosure of dataset access and processing details but also significantly contributes to the replicability and reliability of scientific research.

By employing this decision tree, we not only adhere to the principles unearthed from our survey on data reproducibility but also enhance our discussion on dataset retrievability.Our proposed method employs a decision tree to guide researchers through the nuanced process of making datasets accessible for future research. This decision tree is meticulously crafted to accommodate various scenarios—whether the data can be shared publicly or is restricted, and if it's the raw dataset or a processed form. It offers a clear protocol for researchers to disclose detailed information about dataset access, processing scripts or tools used, including their versions and configuration parameters, thus ensuring the replicability of their work. This decision tree stands as a critical tool, aligning with our survey findings and reinforcing our commitment to advancing scientific transparency and reproducibility.

2.5 Conclusion

In this chapter, we conducted a comprehensive analysis of the reproducibility of RNA-Seq pipelines in tumor studies, with a particular focus on differential expression analysis. A key finding is the lack of uniformity in pipeline documentation, with a significant portion of

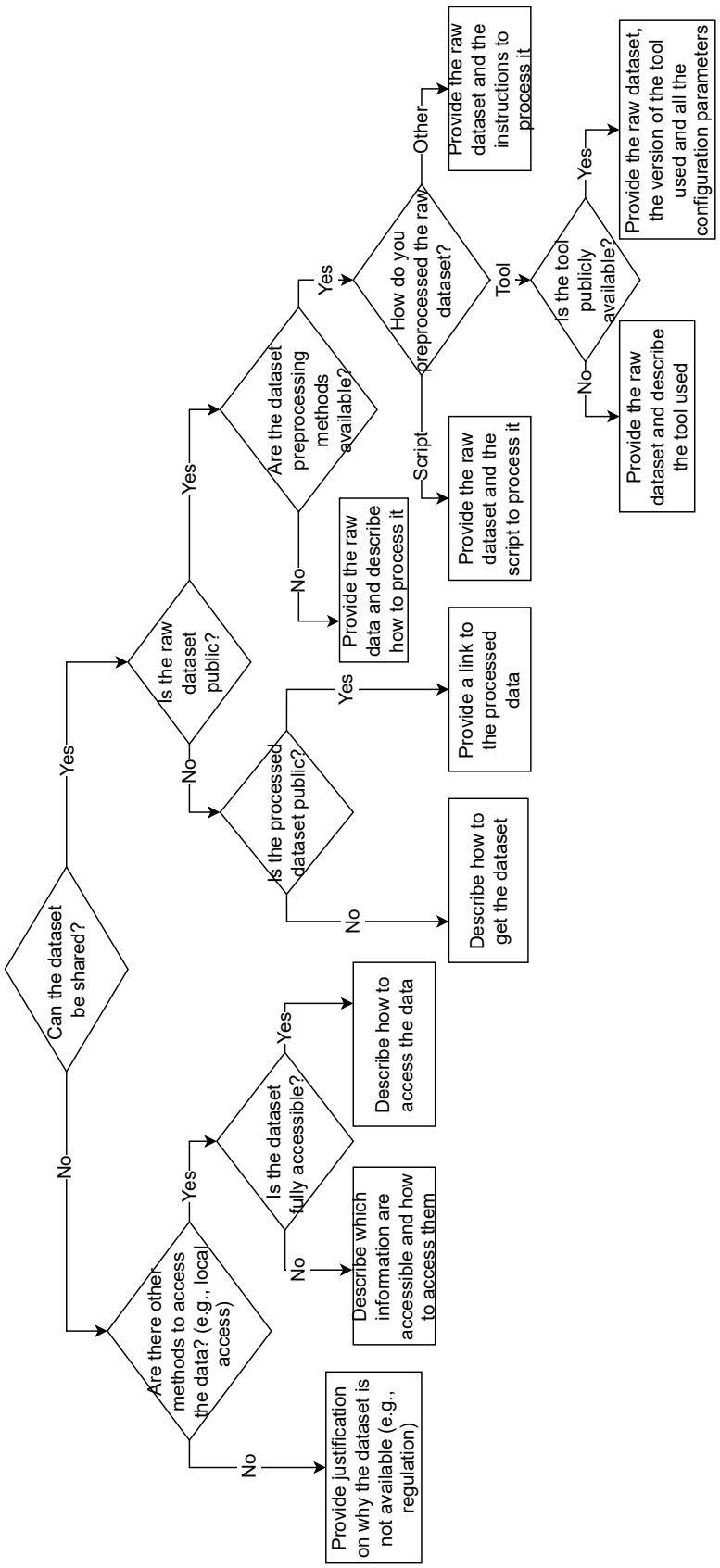


Fig. 2.12 Data Retrievability Tree

studies failing to fully specify the tools and parameters used across all phases. Specifically, only 28 out of the selected studies provided complete documentation of all pipeline phases, representing approximately 15

Furthermore, our analysis revealed that 69 studies did not specify critical thresholds for differential expression analysis, such as padj or fold change, and 40 studies did not disclose the reference genome version used. These omissions are particularly concerning as they directly impact the reproducibility of the results. The choice of genome version, for instance, plays a critical role in variant detection and the overall accuracy of the analysis. Similarly, the specification of padj and fold change thresholds is essential for ensuring that the results can be accurately replicated by other researchers. The variability in reporting practices underscores the necessity for more stringent guidelines within the bioinformatics community. To improve reproducibility, it is imperative that future studies provide comprehensive documentation of all analytical parameters, including the thresholds used for differential expression and the specific genome version employed. This level of transparency is crucial for enabling other researchers to replicate findings under identical conditions, thereby enhancing the reliability of scientific research.

Chapter 3

Towards Accurate Bioinformatics Pipelines

Recognizing reproducibility issues affecting the trust physicians place in bioinformatics results [97] and the imperative need for dependable bioinformatics outcomes, we embarked on a meticulous examination of how the selection of specific tools within these pipelines can influence the results, irrespective of the experiment being conducted.

To ascertain whether the optimization of bioinformatics pipelines could indeed yield more biologically reliable results, we conducted a comparative analysis of the two most commonly used tools identified in our earlier study: *HISAT2* and *STAR*. This comparison aimed to illuminate the significant impact that tool choice has on downstream analyses, particularly focusing on the alignment and quantification phases integral to RNA-Seq processes. Our investigation delved into the performance and output quality of these tools, examining how each one's unique handling of the data affects the subsequent steps in the analysis, especially the identification and interpretation of differentially expressed genes. Through this analysis, we sought to understand how the choice of one tool over another could potentially alter the biological insights gained from the data, thus informing a more informed and strategic selection of bioinformatics tools that underpin reliable research findings.

3.1 Motivation

We set out to explore the intricacies of bioinformatics pipelines in order to achieve reliability and clarity in the analysis of genetic data. Along the way, we encountered an important but sometimes disregarded feature: the uncertainty surrounding the pipelines' first phases. Although quality control, trimming, and alignment are widely acknowledged as essential

procedures in many genomic research, particular information about the instruments and settings used is often missing. This omission affects the dependability and interpretability of outcomes from variant calling to RNA-Seq studies, and it goes beyond simple procedural errors.

In this context, the alignment process captured our interest since it is a critical point in time when the selection of tools has a major impact on the final analytical result. Motivated by an extensive survey we did in the previous chapter revealing STAR and HISAT2 as the two most often utilised alignment tools and from their established use and their demonstrated advantages in terms of performance [98], we found ourselves actively involved in an RNA-Seq investigation [99]. This project, which sought to clarify the possible impact of the medication Ruxolitinib on myelofibrosis (MF), a crippling bone marrow disorder, offered a painful background for our comparative study. While roxolitinib has FDA approval for treating MF symptoms, it is not effective in curing the disease or lowering the amount of mutant cells in the body. This is because some MF cells are resistant to the medication, presumably as a result of extra genes or pathways that support cell survival even in the presence of ruxolitinib's targets, the JAK2/STAT5 pathway. It highlights the urgent need for a more comprehensive genetic understanding. Our previous investigations, as outlined in related work, highlighted the role of proteasomal genes in MF cell survival and suggested that targeting these genes could enhance treatment efficacy.

Unlike [99] but rather driven by these realisations and the need to overcome the difficulties presented by early pipeline phases, we started a thorough analysis of HISAT2 and STAR2. This investigation, which was guided by our RNA-Seq work, was not just a technical exercise but also a crucial step in understanding how alignment tool selection affects computational efficiency and biological interpretability. Through the use of these commonly-used instruments in the particular framework of our RNA-Seq investigation on MF, the goal was to illuminate the wider consequences of preliminary bioinformatics choices.

3.2 Related Works

Precise alignment is pivotal for the success of downstream analyses, yet RNA-seq reads containing splice junctions introduce significant challenges to alignment precision. To surmount these challenges, various software solutions have been devised for aligning sequences to a reference genome. Among these, TopHat2 [100], HISAT2 [101], and STAR2 [102] have emerged as notable examples. HISAT2, building upon the foundational Bowtie2 [103], has replaced TopHat2 as the preferred choice due to its enhanced computational efficiency. Despite their rapid processing capabilities, the selection of an appropriate aligner is crit-

ical for accurate downstream analysis, necessitating a thorough evaluation of each tool's performance.

This investigation aims to scrutinize HISAT2 and STAR2 through comparative analysis, assessing their efficiency and the quality of their output to determine the most suitable tool for precise alignment and subsequent analysis. Both HISAT2 and STAR2, representing the latest advancements in alignment technology, have improved upon the limitations of their predecessors, offering increased accuracy, speed, and memory efficiency. The divergent alignment strategies, index sizes, sensitivity levels, and speed of HISAT2 and STAR2, along with their optimizations for different read types, underscore the necessity of this comparative study. HISAT2, employing a graph FM index (GFM) [101], boasts a compact index size and rapid processing albeit with restricted multi-threading capabilities, whereas STAR2 leverages Spliced Transcripts Alignment [102], offering greater sensitivity and enhanced multi-threading at the expense of speed.

Limited research has been conducted comparing these tools within cancer dataset analyses. Notably, one study [104] demonstrated STAR2's superior uniquely mapped read percentage over HISAT2 across various genome assemblies. Another investigation [105] reported HISAT2's tendency to align fewer reads and its higher propensity for aligning to pseudogenes, impacting alignment accuracy. Building upon these findings, our study extends the comparison to include an assessment of biological relevance in the results produced by HISAT2 and STAR2. By utilizing the hg38 genome assembly, recognized for its comprehensive coverage in genome-wide analyses, we aim to explore differential expression outcomes from each pipeline. Additionally, we will examine the computational efficiency of these pipelines, offering insights into their operational time complexities.

3.3 RNA-Seq General Workflow and Implemented Pipelines

In RNA-Seq analyses, there is not a universally superior pipeline applicable to all scenarios. The choice of methodology often varies based on the objectives of the research and the organisms being sequenced, necessitating a selection from a diverse array of software tools [106],[107]. The generalized workflow for RNA-Sequencing and the specific tools selected for our two unique pipelines are depicted in Figure 3.1.

A key determinant of bioinformatics tools' efficacy in the mapping phase is the speed of alignment [108]. Consequently, we opted to modify the alignment tool across our pipelines while maintaining consistency in the other stages. The distinction between our pipelines

lies solely in their mapping components, chosen for their balanced methodologies and the robustness and efficiency as highlighted in [98], positioning them as leading options currently available. As illustrated in Figure 3.1, an RNA-Seq process encompasses four principal phases: *Quality Control* (detailed in Section 3.3.1), Alignment (explored in Section 3.3.2), Quantification (discussed in Section 3.3.3), and Differential Expression (DE) Analysis (covered in Section 5.5).

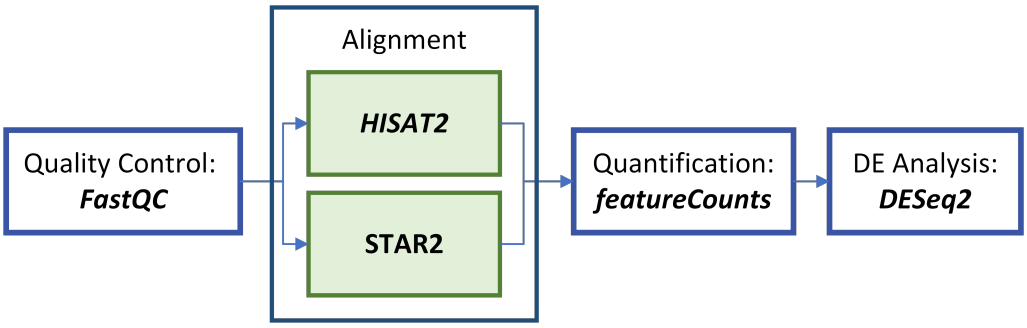


Fig. 3.1 General workflow for RNA-Seq analysis and the implemented pipelines.

3.3.1 Quality Control

The initial phase in the RNA-Seq analysis is Quality Control (QC), a crucial step to ensure the integrity of raw data. This process typically includes adapter trimming to discard sequences not originating from the target organism and the elimination of low-quality reads and bases with ambiguous calls. For data procured from the Illumina platform, FastQC [109] serves as the preferred tool for assessing data quality due to its widespread adoption. Should quality assessments indicate issues, the protocol incorporates additional measures to excise adapter sequences and trim bases of subpar quality, for instance, employing Trimmomatic v0.39 [110].

In our specific analysis, the data quality was deemed adequate, obviating the need for adapter removal or trimming as per our evaluations.

3.3.2 Alignment

When comparing high-throughput sequencing data, in terms of individual reads, to a reference genome or transcriptome, alignment is a crucial step in many bioinformatics investigations. In bioinformatics, alignment is the act of putting DNA, RNA, or protein sequences in a certain order to find homologous sections that could point to structural, functional, or evolutionary linkages. Alignment particularly refers to the mapping of individual short reads produced by

high-throughput sequencing methods to a reference transcriptome or genome in the context of RNA-Seq data processing. This crucial step enables researchers to pinpoint the exact position of each read within the genome, which makes it possible to accurately quantify the levels of gene expression and identify different genomic structures like splice junctions, exons, and introns. To make the alignment process more efficient, most alignment tools create an index of the reference genome or transcriptome. Because genomes and transcriptomes are large, an index helps the alignment tool quickly find where each read might match. The index works by organizing the sequence into a searchable structure, much like a table of contents or an address book, that allows the tool to skip directly to relevant sections instead of searching through the entire genome. This significantly speeds up the process, as the tool can quickly narrow down the locations to check, making alignment much faster and less computationally demanding. The output of the alignment process is typically a Binary Alignment Map (BAM) file, which contains information on how each read aligns to the reference genome. These BAM files are then used as input for subsequent steps, such as quantification and differential expression analysis, making the alignment step foundational to the overall RNA-Seq analysis pipeline. We looked at the two tools -HISAT2 and STAR2- to accomplish the alignment stage.

Alignment Tool #1: HISAT2

HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts) [111] stands out in the realm of RNA-Seq mapping due to its innovative hierarchical indexing approach, which not only enhances its speed but also elevates its efficiency beyond many competing tools. This method not only reduces the index's disk space requirement but also accelerates index construction. Optimized for handling both single-end and paired-end reads, HISAT2 boasts an alignment process that can be up to twice as quick as traditional methods, thanks to its refined strategy [111]. A key feature of HISAT2 is its adeptness at aligning sequences across regions marked by complex splicing, attributed to its hierarchical indexing. This capability ensures high sensitivity, making HISAT2 indispensable for extensive RNA sequencing projects. Despite these advantages, the computational demand of HISAT2's indexing strategy is higher compared to some alternatives, which might limit its use on less powerful systems. Nonetheless, HISAT2's design to construct a thorough index initially allows for faster subsequent alignments with reduced memory needs, highlighting its suitability for large-scale studies where rapid and efficient data processing is paramount.

Alignment Tool #2: STAR2

STAR2 (Spliced Transcripts Alignment to a Reference) [102] ranks among the leading RNA-Seq mapping applications, renowned for its splicing-aware alignment technique. This method significantly enhances sensitivity, enabling a higher number of reads to be precisely aligned to the reference genome than with alternative approaches. Featuring advanced multi-threading support, STAR2 efficiently leverages multicore processors, albeit at the cost of requiring larger index sizes compared to HISAT2. Distinguishing itself from aligners like HISAT2, which uses hierarchical indexing, STAR2 implements a unique strategy for spliced alignment, aptly named the *unique spliced alignment strategy*. This technique intricately matches RNA-Seq reads to the reference genome in scenarios involving spliced exons, through an index that identifies spliced reads and aligns them accordingly. The creation of this index involves delineating splice junctions and anchor points from the reference genome, facilitating the accurate identification of spliced exons within the reads, even amidst complex splicing patterns.

Despite its proficiency in handling large-scale RNA sequencing projects, STAR2's intensive computational demands may render it less suitable for less powerful computing environments. Unlike HISAT2, STAR2 requires users to construct the genome index from the ground up for every new reference genome, a process that can be both time-consuming and resource-intensive, particularly for larger genomes. This necessity for manual index construction could detract from its efficiency and the conservation of resources when pipelines are reused, contrasting with other software that may offer pre-built indexes for quicker setup and lower computational overhead.

3.3.3 Quantification

After completing the mapping phase, the next task involves tallying the reads linked to specific features of interest (in this case, genes) to facilitate differential expression analysis. This analysis contrasts gene expression across various experimental setups. To accomplish this, we employed *featureCounts* [112], processing all BAM files generated during the alignment stage from both pipelines simultaneously.

3.3.4 DE Analysis

DE analysis represents a pivotal phase in RNA-seq data evaluation, aiming to identify genes that vary in expression across different experimental conditions. For this purpose, the *DESeq2* [113] package was specifically employed to analyze the quantification data. Before conduct-

ing DE analysis, we normalized the count data to ensure that comparisons between samples are accurate, taking into account differences in sequencing depth and RNA composition. *DESeq2* package employs the median of ratios method for the normalization. This method starts by creating a pseudo-reference sample for each gene, calculated as the geometric mean of the raw counts across all samples. This pseudo-reference serves as a baseline to compare gene expression across different samples. For each gene in every sample, DESeq2 then computes the ratio of the raw count to this geometric mean, which reflects how much each gene's count deviates from the reference, taking into account variations in sequencing depth and RNA composition between samples. To adjust for these differences, DESeq2 calculates a normalization factor, known as a size factor, for each sample. This size factor is determined by taking the median of all the ratios for a given sample, ensuring that the majority of genes, which are typically not differentially expressed, have comparable expression levels after normalization. Using this median of ratios method, DESeq2 effectively normalizes the data, making it robust to differences in sequencing depth and RNA composition and allowing for accurate identification of genes that are differentially expressed across different conditions.

After normalizing the quantification data from RNA-seq, DE analysis was conducted to pinpoint genes that were either upregulated or downregulated. In this study, the R programming language was harnessed to execute DE analysis on data derived from the HISAT2 and STAR2 pipelines.

3.4 Experimental Settings

In this section we present the experimental settings we used in our experiments. In particular, in Section 3.4.1 we describe the used dataset, in Section 4.3.4 we describe the hardware we used. Finally, in Section 3.4.3 we describe the software configuration.

3.4.1 Dataset Description

Our objective was to scrutinize the mRNA expression in CD34+ hematopoietic stem cells harvested from the peripheral blood of myelofibrosis patients. Initially, the dataset comprised samples from five individuals. However, quality control measures revealed Escherichia coli contamination in the samples from one patient. To maintain the integrity of our analysis, we excluded the contaminated samples, focusing on 32 fastq files derived from four patients, with each offering two samples—one treated with the Ruxolitinib drug and the other untreated, with each patient providing two replicates. The data, stored in *fastq* format, averaged 3 GB per file when uncompressed, totaling 96 GB for the study. Sequencing of these eight

samples employed a paired-end approach to enhance data duplication accuracy, utilizing the Illumina HiSeq 2500 system. This sequencing effort yielded between 60 to 90 million bases per sample, with a majority surpassing 70 million reads. Access to the raw sequencing data from this research is restricted to on-site requests, in compliance with patient confidentiality protocols.

3.4.2 Hardware Configuration

The research was carried out on Caliban, a computing cluster with multiple nodes, where all calculations were executed on a single node equipped with 48 CPUs at 2.16 GHz each, enabling efficient parallel computation and swift data handling. This node features 141.48 GB of RAM and 1.5 TB of local disk space, sufficient for storing all computational data and results.

3.4.3 Software Configuration

The environment for bioinformatics tools and their configurations that we selected is Anaconda version 3. Specifically, we developed bespoke bash scripts using Anaconda's package management (Conda) to automate the cluster's RNA-Seq operation. Table 3.1 presents the commands utilised for launching the chosen tools together with their corresponding settings. We list the tools that were utilised, together with the *command*, *arguments* that were relevant to the experiments, and the workflow *phase* that each tool implemented on the table's rows.

For both pipelines, the GRCh38 genome from Ensembl [115] served as the foundational reference for aligning reads. Accompanying the genome, a Gene Transfer Format (GTF) file, detailing crucial information about splice sites and exons, was pivotal in constructing the indexes necessary for alignment. Recognizing the dual demands of time and computational resources required for indexing, efforts to streamline this process leveraged a multi-core processing environment, utilizing 40 threads for both the development of the index and the alignment operations themselves.

To maintain integrity and ensure a balanced evaluation of the two pipelines, indexes were generated anew rather than relying on pre-existing versions. This approach, dictated by the essential role of index construction in alignment accuracy, aimed to eliminate potential biases and uphold the comparison's validity. It's important to highlight that STAR2, in particular, necessitates the manual creation of indexes for each reference genome, as it lacks pre-built indexes. This requirement underscores the need for meticulous preparation and resource allocation in the setup phase of alignment tasks, ensuring both pipelines operate from a common ground for a fair and insightful analysis.

Table 3.1 Commands used for each tool in the bioinformatics pipelines and the related configurations

Tool	Phase	Command	Arguments
FASTQc	Quality Control	fastqc	-t 40 -o output_dir input_file
HISAT2	Alignment (Indexing)	hisat2-build	– p 40 – ss splice_sites.txt – exon exons.txt
STAR2	Alignment (Indexing)	star	– runThreadN 40 – runMode genomeGenerate – sjdbGTFfile Homo_sapiens.GRCh38.97.gtf – genomeDir GRCh38 – genomeFastaFiles GRCh38.dna.primary.fa
STAR2	Alignment	star	– runThreadN 40 – genomeDir GRCh38 – sjdbGTFfile Homo_sapiens.GRCh38.97.gtf - readFilesIn Sample_R1.fastq Sample_R2.fastq - outSAMtype BAM SortedByCoordinate - outSAMunmapped Within - outSAMattributes Standard - quantMode GeneCounts - outFileNamePrefix AlignmentSample1 - twopassMode Basic
featureCounts	Quantification	featureCounts	-T 40 -p -t exon -g gene_name -a Homo_sapiens.GRCh38.97.gtf -o countmatrix.txt S1.bam ... Sn.bam
DESeq2	DE Analysis	R Script (custom)	Official Code: [114]

During the DE analysis, we aimed to determine over- or down-expressed genes using two different thresholds to establish biological and statistical relevance. The *padj* threshold represents the level of statistical significance of the differential gene expression and corresponds to an adjusted p-value, which was set to a value of 0.05. The adjustment for multiple testing was performed using the Benjamini-Hochberg method, which controls the false discovery rate (FDR). In simple terms, the Benjamini-Hochberg method sorts all the p-values and adjusts them to account for the fact that many tests are being performed simultaneously, reducing the likelihood of false positives and ensuring that the reported significant results are reliable.

The *log fold change* threshold, or log2 fold change, reflects the magnitude of the biological differences between experimental conditions. For our analysis, this threshold was set to 1 in absolute value, meaning it did not play a role in determining which genes were considered differentially expressed. Essentially, all genes with significant adjusted p-values were included regardless of the fold change magnitude, allowing for a comprehensive analysis of differential expression without imposing a minimum effect size.

These thresholds and the multiple testing correction method were applied consistently across both pipelines to ensure that the results were comparable and biologically relevant.

Code Availability

In this study, the pipelines were implemented as bash scripts and executed on a Linux Cluster. The associated code is accessible at the provided reference [116] and is governed by the Creative Commons Attribution 4.0 International license.

3.5 Results

This section compares the execution times and biological outcomes of the various bioinformatics pipelines through a thorough review. To comprehend the overhead brought on by the alignment tools, the computation time study concentrates on the three primary pipeline steps: index construction, alignment, and quantification. Conversely, the evaluation of biological results seeks to compare the pipelines with respect to the biological relevance of the Differential Expression step (in Section3.5) and the Alignment and Quantification steps (in Section 3.5) on one side. Our goal is to present a detailed analysis of the two pipelines, emphasising their advantages and disadvantages.

Computing time

In the experiment, we configured the tools of the two pipelines with consistent configuration parameters (e.g., the same number of threads, the considered mapping regions), as reported in Section 3.4.

Table 3.2 Average computing time of index creation, alignment, and quantification steps with standard deviations.

Pipeline Name	Index Creation (minutes)	Alignment (minutes)	Quantification (minutes)
HISAT2	54.0 ± 3.2	10.19 ± 0.75	3.34 ± 0.20
STAR2	28.0 ± 2.5	10.43 ± 0.68	2.18 ± 0.15

The data presented in Table 3.2 indicate that HISAT2 demands more time to create an index but aligns reads at a pace comparable to STAR2, albeit with a marginal lag in quantification speed. This variance in the quantification phase could potentially stem from the distinct alignment methodologies employed by HISAT2 and STAR2. Given that both software and hardware configurations were consistent across tests, the differences in processing durations can likely be ascribed to: *i)* the inherent disparities in how each algorithm processes data and *ii)* the variation in their approaches to multithreading and optimization strategies.

The data presented in Table 3.2 show that while HISAT2 requires more time for index creation than STAR2, the difference in alignment and quantification times between the two tools is minimal. The additional 26 minutes for index creation in HISAT2 is relatively insignificant in the context of an entire scientific project. If index creation time is a critical factor in a project's timeline, HISAT2's longer setup time might be a significant consideration. Conversely, for projects where the slight delay in quantification does not impact overall deadlines or where computational resources are limited, HISAT2's performance could be deemed acceptable.

Biological Relevance of Alignment and Quantification

In evaluating the alignment efficiency, an essential aspect to compare is the overall and uniquely mapped reads as outlined in Table 3.3. The comparative analysis revealed that both HISAT2 and STAR2 exhibit similar overall alignment rates, with HISAT2 aligning 98.03% and STAR2 slightly higher at 98.78% of the reads. Despite the minimal variance in overall alignment, a more notable difference emerges in the proportion of uniquely mapped reads. Specifically, HISAT2 managed to uniquely align 80.47% of reads, whereas STAR2 achieved a slightly higher rate of 81.66%, translating to a difference of over 300,000 uniquely mapped

reads between the two, with HISAT2 aligning 24,309,436 reads and STAR2 24,667,395 reads.

This disparity in uniquely mapped reads, although seemingly modest, holds significant implications for subsequent analytical steps and the interpretability of biological outcomes. It is crucial to acknowledge HISAT2's marginally higher incidence of multi-mapped reads (17.12% for HISAT2 vs. 15.9% for STAR2), which poses challenges to precise mapping and, by extension, can influence downstream analyses like differential expression analysis. The phenomenon of multi-mapping, if not addressed, might obscure true gene expression patterns, yet it also presents an opportunity for HISAT2 to potentially uncover a broader array of differentially expressed genes compared to STAR2. This aspect underscores the importance of considering both the quantity and quality of mapped reads in choosing an alignment tool, as it directly affects the reliability of downstream data interpretation and the discovery of biologically significant insights [117].

Table 3.3 Percentage and number of mapped reads obtained after the alignment.

Pipeline Name	Overall Alignment Rate (%)	Uniquely Mapped Read Rate (%)	No. of (Uniquely) Mapped Reads (in millions)
HISAT2	98.03 ± 0.15	80.47 ± 0.22	24.309 ± 0.084
STAR2	98.78 ± 0.13	81.66 ± 0.18	24.667 ± 0.079

Building upon the noted discrepancies in uniquely mapped reads and the potential for multi-mapping to affect the discernment of genuine gene expression profiles, we expanded our analysis to consider the alignment of reads against pseudogene regions. Pseudogenes are DNA sequences that resemble functional genes but are generally nonfunctional due to mutations, deletions, or combinations [118]. Despite their noncoding status, pseudogenes can be transcribed, complicating the alignment process and potentially masquerading as functional gene expression in RNA-Seq data.

In our expanded analysis, represented in the bar chart in the Figure 3.2, we scrutinized the alignment patterns of HISAT2 and STAR to these pseudogene loci. Notably, HISAT2 aligns a higher percentage of reads to pseudogenes than STAR in both conditions, with a pronounced increase under the treated condition. This suggests that HISAT2 may have lower specificity in differentiating between genes and pseudogenes, particularly when experimental conditions are altered, such as by drug treatment. STAR's performance is consistent across both conditions, indicating a more stable alignment behavior irrespective of the treatment, which may be preferable in studies where pseudogene expression could confound the results. The significant rise in pseudogene alignment by HISAT2 in the treated samples could point

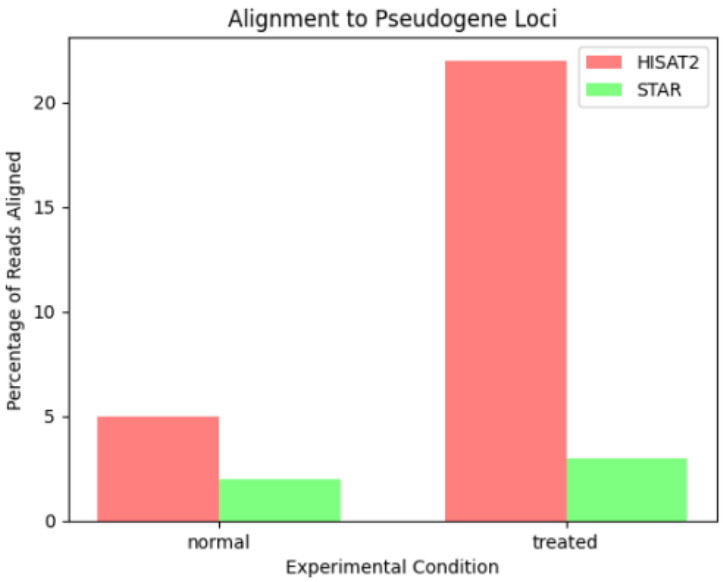


Fig. 3.2 Comparative alignment of reads to pseudogene loci using HISAT2 and STAR under normal and treated experimental conditions.

to treatment-related changes in the transcriptome, perhaps indicative of the expression of pseudogenes or changes in RNA splicing or processing.

These observations underline the importance of selecting an appropriate alignment tool based on the nature of the experiment, especially considering the potential biological implications. For instance, in studies where the accurate differentiation between genes and pseudogenes is critical, the consistent low alignment to pseudogenes by STAR might make it the tool of choice. Conversely, the increase in pseudogene alignment with HISAT2 might necessitate additional data filtering steps to ensure the reliability of downstream analyses, such as identifying differentially expressed genes. This careful consideration is essential to ensure that the results of RNA-Seq data analysis are both accurate and biologically meaningful.

Biological Relevance of Differential Expression

In the displayed Venn diagrams of Figure 3.3, we have illustrated the number of differentially expressed genes detected by the HISAT2 and STAR2 pipelines, with a comprehensive breakdown of total gene counts as well as those that are over-expressed and down-expressed.

Our analysis revealed that HISAT2 flagged a larger number of genes (197) as differentially expressed than STAR2 (147), all within the defined thresholds of statistical and biological significance (Section 3.4). Closer scrutiny of the Venn diagrams shows a substantial overlap, where 138 genes were commonly identified by both tools, suggesting a core set of differen-

tially expressed genes recognized irrespective of the alignment tool used. It is interesting to note that almost all the genes detected by STAR2 were also picked up by HISAT2, with the latter additionally identifying 59 genes (35 over-expressed and 24 down-expressed) not recognized by STAR2. Conversely, STAR2 uniquely identified only 9 genes.

This disparity in gene detection could imply that HISAT2, potentially due to its alignment strategy that encompasses more regions corresponding to pseudogenes, might have a broader detection scope. Pseudogenes, the genomic remnants that resemble functional genes but are typically noncoding, can introduce complexities in alignment and gene identification processes. Since they share sequences with functional genes, aligners might misinterpret pseudogene-derived reads as originating from functional genes, leading to potential discrepancies in gene expression analysis.

This observation suggests that while HISAT2 may cast a wider net in detecting gene expression changes, it also raises the possibility of including reads from pseudogenes, which may not be biologically relevant in the context of functional gene expression. On the flip side, the more conservative count of STAR2 might reflect a more targeted approach, potentially missing some true positives but also reducing the noise from pseudogenes.

The implications of these findings are twofold. First, they highlight the necessity for careful consideration when selecting an alignment tool based on the study's focus—whether it is to capture as broad a spectrum of gene expression changes as possible or to ensure a more stringent, potentially more accurate, set of differentially expressed genes. Second, they underline the importance of understanding the inherent biases and operational differences between alignment tools, which can have a direct impact on the downstream analysis and biological interpretations drawn from RNA-Seq data [105].

To unravel the biological significance of the data, bioinformatics typically zooms in on the most differentially expressed genes, as they often hold the key to understanding the physiological changes under study (Figure 3.4). In this analysis, the focus was placed on the top 30 differentially expressed genes, irrespective of their being upregulated or downregulated. It was found that 27 out of these 30 genes showed consistent expression across both HISAT2 and STAR2 pipelines.

Such consistency between the pipelines suggests that, despite the presence of some false positives within the broader gene expression data, the most critical genes from a biological standpoint tend to be reliably detected by both methods. Hence, these top 30 genes, reflecting the core of differentially expressed genes, can serve as a reliable basis for further detailed examination and interpretation. This convergence of results from both pipelines enhances confidence in the biological relevance of the identified genes, providing a solid foundation for subsequent research and validation efforts.

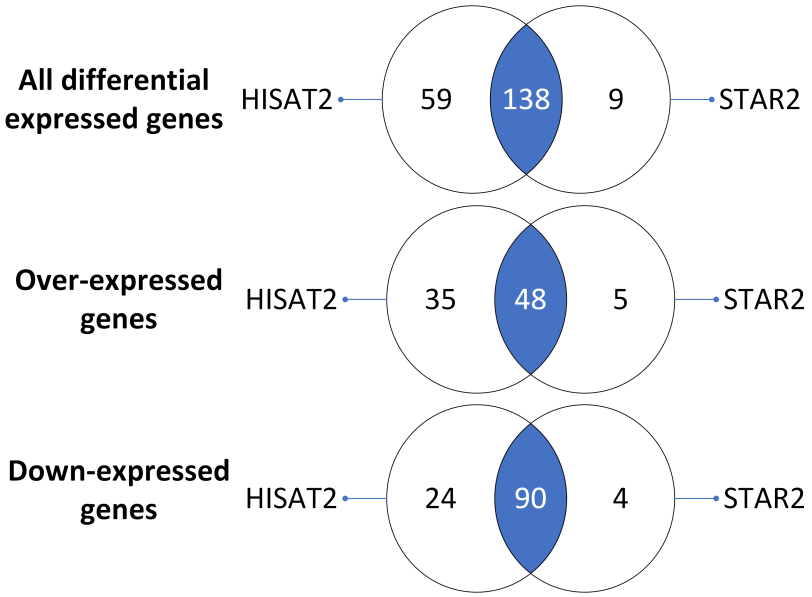


Fig. 3.3 Comparison of gene expression results across pipelines and gene sets. The first Venn diagram (from the top) shows the overlap and differences in overall differentially expressed genes between two pipelines. The second diagram focuses on over-expressed genes, while the third diagram compares down-expressed genes.

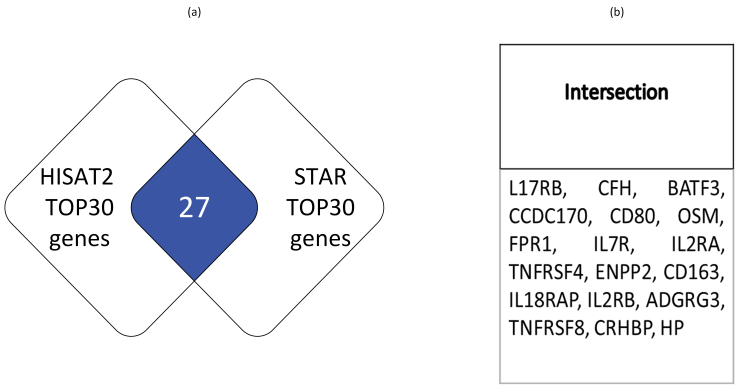


Fig. 3.4 On the left (Figure a), the comparison of gene expression results across pipelines and gene sets in terms of most important differentially expressed genes (top 30). On the right (Figure b), the focus is on the intersection and the expressed genes that were found down-regulated

To illustrate a case in point regarding the physiological relevance of our gene expression data: the top 30 differentially expressed genes identified in our study have shown alignment with clinical observations. This group predominantly comprises genes implicated in the pathophysiology of myelofibrosis, which appeared to be repressed upon treatment with the therapeutic drug, leading to a reduction in inflammation—a hallmark of this condition (Figure 3.4). Remarkably, each gene within this subset was found to be down-regulated, signaling the drug's efficacy in modulating gene expression. The fact that these down-regulated genes are all associated with the inflammatory processes specific to myelofibrosis corroborates the notion that the drug intervention is exerting a precise and targeted influence on the biological pathways underpinning the disease. This correlation between the gene expression results and the expected therapeutic outcomes lends credence to the drug's role in manipulating key genes within the disease pathway, affirming the therapeutic strategy employed in this research.

3.5.1 Limitations

Despite the promising results obtained in this study, there are certain limitations to our approach that should be taken into consideration. In this section, we discuss these limitations and their potential impact on the interpretation of our findings.

- The main limitation of this study is that it analyzes only a single dataset. To draw more generalizable conclusions about the effectiveness and reliability of different mapping tools, it would be necessary to analyze multiple datasets from diverse sources. Such a broader analysis could provide a more comprehensive assessment of the tools' performance across various experimental conditions and biological contexts. However, even with this single dataset, we have begun to observe patterns in how different tools perform, suggesting initial trends that could be further validated with additional data. While expanding the scope to include multiple datasets is beyond this thesis, it is important to recognize that conclusions based on a single dataset may not fully reflect the variability and robustness of the tools in diverse settings.

- The dataset employed here is proprietary, provided by our collaborators, and not available for public distribution. We acknowledge the limitation this poses for the reproducibility of our findings. Additionally, sourcing public datasets for myelofibrosis is inherently difficult given the rarity of the condition.

- The scope of our analysis included differential gene expression using DESeq2, but it did not extend to pathway or functional enrichment analysis. Such analyses are

crucial for confirming the efficacy of the analytical pipelines and for uncovering new biological processes involved in the onset and progression of myelofibrosis.

- As a means of improving the robustness of our results, we aim to integrate other methods for differential expression analysis in future work. Although we used the DESeq2 method in this study, the integration of other related methods will enable us to evaluate potential false positives or false negatives that may arise from using only one method. Such a comprehensive evaluation will provide a more accurate and reliable comparison of the different cleaning methods.

These considerations serve as a reminder of the necessity for comprehensive and transparent research practices and the continuous refinement of methods to achieve the most reliable and informative results.

3.6 Discussion

Through the lens of our comparative study between HISAT2 and STAR2, it is evident that the choice of alignment tool can distinctly influence the outcome in terms of computational efficiency, alignment accuracy, and the identification of differentially expressed genes. Employing robust statistical thresholds allowed us to identify a significant cohort of genes exhibiting differential expression, many of which were responsive to the myelofibrosis treatment drug. Notably, genes down-regulated by the drug coincided with those associated with inflammation—a key feature of the disease—implying that the drug's mechanism is effectively targeting the underlying biological processes of myelofibrosis.
Drawing from the collective findings, several insights emerge:

- STAR2 demonstrated superior alignment accuracy in comparison to HISAT2, suggesting its preference for projects where precise differential expression analysis is paramount. Nonetheless, HISAT2's broader gene discovery potential makes it suitable for exploratory studies aimed at uncovering novel genes.

- Contrary to initial assumptions, STAR2 exhibited improved execution times despite the absence of pre-built indices, suggesting that its alignment process is well-optimized for speed.

- The disparity in unique read mapping, with STAR2 aligning substantially more reads, resulted in a paradox of fewer identified genes. This invites further exploration to unravel how the differential expression patterns of these uniquely mapped reads contribute to the overall gene expression landscape.

- Intriguingly, the genes identified exclusively by either HISAT2 or STAR2 may open new investigative avenues into the pathology of myelofibrosis and the pharmacodynamics of Ruxolitinib, potentially leading to novel insights that have yet to be explored.

Moreover, the role of pseudogenes in the analysis emerged as a pivotal factor. HISAT2's higher alignment to pseudogenes, especially under treatment conditions, poses additional considerations for data interpretation. Pseudogenes, which mimic functional gene sequences but are not typically expressed, can complicate the alignment process, potentially inflating gene counts or misrepresenting gene expression levels. One effective approach to mitigate this issue would be to provide a GTF file containing only protein-coding genes during the alignment step. This would reduce the alignment of reads to pseudogenes, thus enhancing the specificity of the analysis and ensuring more accurate gene expression results. The precision of STAR2 in this regard suggests a more selective approach, possibly minimizing the influence of these genomic elements.

Additionally, an important consideration in bioinformatics research is computational reproducibility. To ensure the reproducibility of our findings, future studies should consider using virtualization or containerization technologies, such as Docker or Singularity. These tools provide a standardized environment for executing bioinformatics pipelines, ensuring that all dependencies and software versions are consistently maintained across different computational setups. This approach would not only facilitate reproducibility but also enhance the transparency and robustness of bioinformatics research.

In light of these observations, it becomes clear that the selection of an alignment tool should not only be dictated by performance metrics but also by the specific requirements and goals of the research at hand. The ramifications of this study highlight the ongoing need for comprehensive evaluations of bioinformatics tools, ensuring that their applications in research are both informed and intentional, to foster advancements in our understanding and treatment of complex diseases like myelofibrosis. As we conclude this chapter on the differential expression analysis, it's important to recognize the continuity in genomic research methodologies as we transition to the next stage of our study. We now shift our focus to variant analysis, which is an expansion of the foundational bioinformatics processes. These steps, while preserved across different types of genomic analyses, play a particularly pivotal role in variant calling pipelines, allowing us to delve deeper into the genomic alterations that may underpin diseases like myelofibrosis.

All the work, including the results and analysis discussed in this section, has been detailed and published in our article by [119], further enriching the discourse on the impact of alignment tool selection in genomic research.

Chapter 4

Identifying Potential Cancer-Associated Variants through Integrated Genomic Analysis

Following our exploration of pipeline optimization, this chapter shifts focus to the critical task of identifying genetic variants associated with cancer. Instead of developing new bioinformatics tools, our aim is to leverage existing methodologies to conduct a comprehensive analysis of genomic variations—specifically SNVs, Indels and CNVs—that may have a potential role in cancer predisposition and pathogenesis.

Understanding the genetic underpinnings of cancer requires an integrated approach that considers both point mutations and structural variations across the genome. Here, we present a detailed analysis using public WES datasets from familial breast cancer cases that do not carry the common *BRCA1/2* mutations. Our approach synthesizes multiple established tools into a cohesive workflow to maximize the detection of relevant variants that could inform on novel genetic factors contributing to cancer. By applying this integrative strategy, we aim to highlight potential variants that warrant further investigation in cancer research, thereby contributing to the broader understanding of cancer genetics and advancing knowledge in the field of genomics.

4.1 Motivation

Following the comprehensive analysis of bioinformatics pipelines in previous chapters, our focus naturally shifted towards variant calling as a crucial step in genomic research. The knowledge and experience gained from those chapters provided a solid foundation for devel-

oping a more integrated approach to identifying genetic variations. This chapter highlights the evolution of our methodology, building upon the insights and practical knowledge obtained from previous studies.

The motivation for this work stems from the need to refine and enhance variant detection techniques, specifically for identifying SNPs, Indels, and CNVs. By integrating these various types of genetic alterations within a unified framework, we aim to improve the accuracy and reproducibility of genomic analysis. This integrated approach not only addresses the complexities of variant calling but also opens new possibilities for discovering potential cancer-associated variants, advancing our understanding of genomic contributions to disease.

4.2 Related Works

In the realm of contemporary genomics, the ability to thoroughly assess genetic variations stands as a cornerstone in unraveling the complex genetic underpinnings of diseases, notably cancer [120]. The study of SNV, Indels and CNVs is paramount for elucidating disease mechanisms and enhancing diagnostic precision [121]. The call for an integrated analysis of these genetic alterations has been highlighted in recent research [122], [123], [124], pointing out a pivotal gap: the tendency to analyze SNVs/Indels and CNVs in isolation rather than within a unified framework [125]. This segmentation hampers the construction of a holistic view of genomic alterations. Additionally, the scarcity of publicly available codes exacerbates the challenge for researchers, particularly for those with limited bioinformatics expertise.

The use of bespoke pipelines for SNP and Indel detection is prevalent across various laboratories [126] or as seen in https://www.alamyahealth.com/next-era-whole-genome-sequencing/. Yet, these custom solutions often remain under lock and key, shrouded in proprietary secrecy that bars the broader scientific community from scrutinizing or enhancing these tools [127]. This opacity not only stifles methodological validation and replication but also curtails collaborative advancements in the field. The literature reflects ongoing struggles to standardize variant analysis, with researchers frequently resorting to custom or established platforms like Galaxy (https://usegalaxy.org/) that may fall short of meeting precise research demands. This delineates a crucial gap in genomic research, underscoring the necessity for an approachable, integrated system.

Addressing this void, our work introduces an integrated pipeline that amalgamates the analysis of SNVs, Indels, and CNVs into a coherent, single framework. This holistic strategy not only broadens the analytical scope but also champions the reproducibility of findings—a critical yet often neglected facet in current research paradigms.

Our investigation utilizes public databases to re-examine NGS data, with a particular emphasis on Whole Exome Sequencing (WES) datasets from familial/hereditary breast cancer (BC) cases devoid of *BRCA1/2* mutations (non-BRCA). This focused inquiry enables us to validate our integrated approach in a precise and controlled context. Given that a substantial portion of familial BC instances eludes genetic explanation beyond the most commonly implicated *BRCA1/2* genes [128], and despite the identification of other genes with lower penetrance [129], a significant majority of these cases remain genetically unresolved. Consequently, there is an imperative need to uncover new predisposing factors that could elucidate the genetic susceptibility in non-BRCA familial BC cases.

4.3 Materials and Methods

4.3.1 Methodological Approach

Our integrated technique was designed to simultaneously identify CNVs, Indels, and SNVs in WES data. This comprehensive strategy includes a pipeline for identifying germline SNVs and Indels as well as a parallel process for identifying CNVs. Although the two pipelines have different end goals, they both have common basic phases that guarantee consistent data processing. Even while our pipelines' tools are essential for defining the current internal processes, they may be easily modified or replaced with only slight configuration adjustments. This flexibility also applies to the references that are utilised, which may be easily changed to suit the requirements of different projects. These references include genomes, targeted parts, and known sites.

4.3.2 Dataset

In developing a comprehensive and robust approach for the analysis of WES data, it is imperative to validate the proposed methodologies within a controlled setting. This controlled environment facilitates a precise evaluation of the approach, ensuring its reliability and effectiveness in uncovering new insights into familial BC susceptibility. To this end, we utilized previously sequenced data as the foundation for our experimental setting.

The datasets employed encompass two WES datasets: **PRJEB3235** (36 items) [130] and **PRJEB31704** (7 items) [131]. The dataset corresponding to id *PRJEB3235* **is freely accessible from the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/srasra/?term=PRJEB3235)** and provides sequencing data for eleven BC cases: 07S240, DAD1, family F2887 (F2887-13 and -24), family F3311 (F3311-5 and -43), I-1408, family RUL036

(RUL036-2 and -7), family RUL153 (RUL153-2 and -3) and seven HapMap controls. Regarding the dataset with id **PRJEB31704**, we employed 7 samples: family F11 (BC patients F11S01 and F11S02 and their informative relatives F11S03 and F11S04), family F12 (BC patients F12S01, F12S02 and their informative relative F12S03) [131]. Both the datasets correspond to samples sequenced on the Illumina platform. They were chosen for their potential since they contain novel genetic variants associated with BC susceptibility, as they included data from families of patients who were negative for *BRCA1/2* mutations.

Data Availability

To ensure the reproducibility and transparency of our work, the datasets underpinning our analysis are readily available for scrutiny. The primary data can be found on Zenodo (https://doi.org/10.5281/zenodo.10078336) and GitHub (https://github.com/anbianchi/IntegratedSNVINDELSandCNV/), providing open access to the resources crucial for our study. Additionally, these datasets were derived from sources in the public domain: BioProject **PRJEB3235** and **PRJEB31704**, further details of which can be explored through the National Center for Biotechnology Information (NCBI) BioProject repository.

4.3.3 Software

The core of our analytical methodology is anchored in the utilization of Snakemake [132], a workflow management system lauded for its promotion of durable, scalable, and replicable data analyses. Snakemake serves as the backbone of our strategy, orchestrating a cohesive suite of bioinformatics tools that span the entirety of our analysis process.

At the heart of our integrated approach lies a selection of universally applied tools, critical for ensuring the uniformity and reliability of our pipeline. For quality control, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) provides an initial assessment of the raw sequencing data. Read trimming is adeptly handled by Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), ensuring that only high-quality sequences proceed to downstream analyses. Samtools (http://www.htslib.org/) offers indispensable functionalities for indexing and generating statistics on BAM files, while Picard (https://broadinstitute.github.io/picard/) is our go-to for sorting reads and removing duplicates. The Genome Analysis Toolkit (GATK) [133] plays a pivotal role in recalibrating base qualities, further refining the quality of our dataset.

Upon the completion of base realignment, our workflow bifurcates, adapting its toolset to the specific demands of each analysis branch. For the detection of SNVs and Indels, the GATK's germline short variant discovery pipeline (https://tinyurl.com/ypcx7fnx) is employed,

Scope	Tool	Version	Parameters
Preprocessing	Trimmomatic	0.39	MAXINFO:40:0.9 MINLEN:36
	Trimmomatic PE	0.39	MAXINFO:40:0.9 MINLEN:36
	BWA-MEM	0.7.17	Default
	Samtools Flagstat	2.6.0	Default
	Mark Duplicates	3.0.0	remove_duplicates: true create_index: true validation_stringency: silent
	Picard Sortsam	1.17	sort_order:coordinate extra:create_index true
	Gatk Base Recalibrator	4.4.0.0	intervals: 100bp_exon.bed known-sites: Mills_and_1000G.ind.hg19
Snv/Indels	GATK HaplotypeCaller	4.4.0.0	-ERC GVCF
	GATK GenotypeGVCFs	4.4.0.0	Default
	GATK VQSR	4.4.0.0	-mode SNP, -mode INDEL
	GATK VariantFiltration	4.4.0.0	-
	Annovar	2020-06-07	-build hg19
CNV	Gatk Apply BQSR	4.4.0.0	extra = –intervals 100bp_exon.bed
	ExomeDepth	1.1.16	100bp_exon.bed
	cn.mops	1.44.0	100bp_exon.bed
	AnnotSV	3.2.3	Default

Table 4.1 List of tools, version number and parameters employed in our pipeline.

leveraging its comprehensive resources for short variant calling. In contrast, CNV detection harnesses the capabilities of ExomeDepth [134] and cn.mops [135], both renowned for their efficacy in CNV analysis. Annotation tasks for SNV/Indels and CNVs are respectively undertaken by Annovar (version 2020-06-08) [136] and AnnotSV (version 3.2.3) [137], enriching our variants with essential contextual information.

Each tool selected for our pipeline was chosen not only for its individual performance but also for how it complements the workflow as a whole. This strategic assembly of tools, detailed in **Table 4.1**, ensures a comprehensive and nuanced approach to variant analysis, striving for the highest standards of data integrity and analytical depth. Through this meticulous integration of diverse bioinformatics resources, we aim to advance the precision and understanding of genomic analyses, particularly in the realm of variant detection.

4.3.4 Hardware

The research was conducted on Caliban, a cluster environment provided by the DISIM Department of the University of L'Aquila and comprising multiple nodes. Specifically, the experiments were executed on a system running CentOS Linux release 7.4.1708 (Core) and powered by an Intel(R) Xeon(R) CPU E5-2698 v4, operating at 2.20GHz, with an available RAM of 141GB.

4.3.5 Resources

In this subsection, we detail the resources utilized in our integrated approach. These resources are crucial for ensuring the efficacy and reliability of our analysis.

Reference Genome: The reference genome used is hg19, taken from UCSC Genome Browser (https://genome.ucsc.edu/). In particular, the *hg19.fa* file we used is in (https://tinyurl.com/4rjnnetn).

Targeted Regions (Bed Files): We used a bed file format to precisely define the regions of interest for our tools. We collected the input by employing exon locations with a 100 bp flanking area, as reported in [138], given the likelihood of detecting high-quality off-target variations from WES. The Table Browser at UCSC (https://genome.ucsc.edu/cgi-bin/hgTables) was used to get these websites.

Known Sites: For pivotal steps such as base recalibration and variant filtering within our analysis, we leveraged known sites, a crucial resource for enhancing the accuracy and reliability of our results. These reference sites were meticulously selected from the comprehensive GATK resource bundle, accessible via (https://tinyurl.com/3f5n4cy7). This incorporation significantly aids in refining the precision of our variant calls, minimizing the likelihood of false positives and enhancing the overall quality of our analysis.

4.3.6 Implementation

The integrated method for SNV, Indels and CNV detection we used is depicted in **Figure 4.1**. It is organized into three primary segments: *Preprocessing*, *SNV and Indels identification* and *CNV detection*. They are detailed in the following sections.

4.3.7 Preprocessing

Our integrated approach has its roots in many important preprocessing processes that provide a common groundwork for the subsequent specialised studies that target the identification of SNVs, Indels, and CNVs. These phases are primarily concerned with organising and fine-tuning the data so that later on, accurate variation detection may be achieved.

Raw Data Quality Control: the approach begins with a quality control check on the raw WES data using FastQC (https://tinyurl.com/333dk7kp).

Read Trimming: Following the quality control, the reads undergo trimming to eliminate low-quality bases and adapter sequences, employing Trimmomatic with parameters set to MAXINFO:40:0.9 and MINLEN:36. These parameters, specifically chosen based

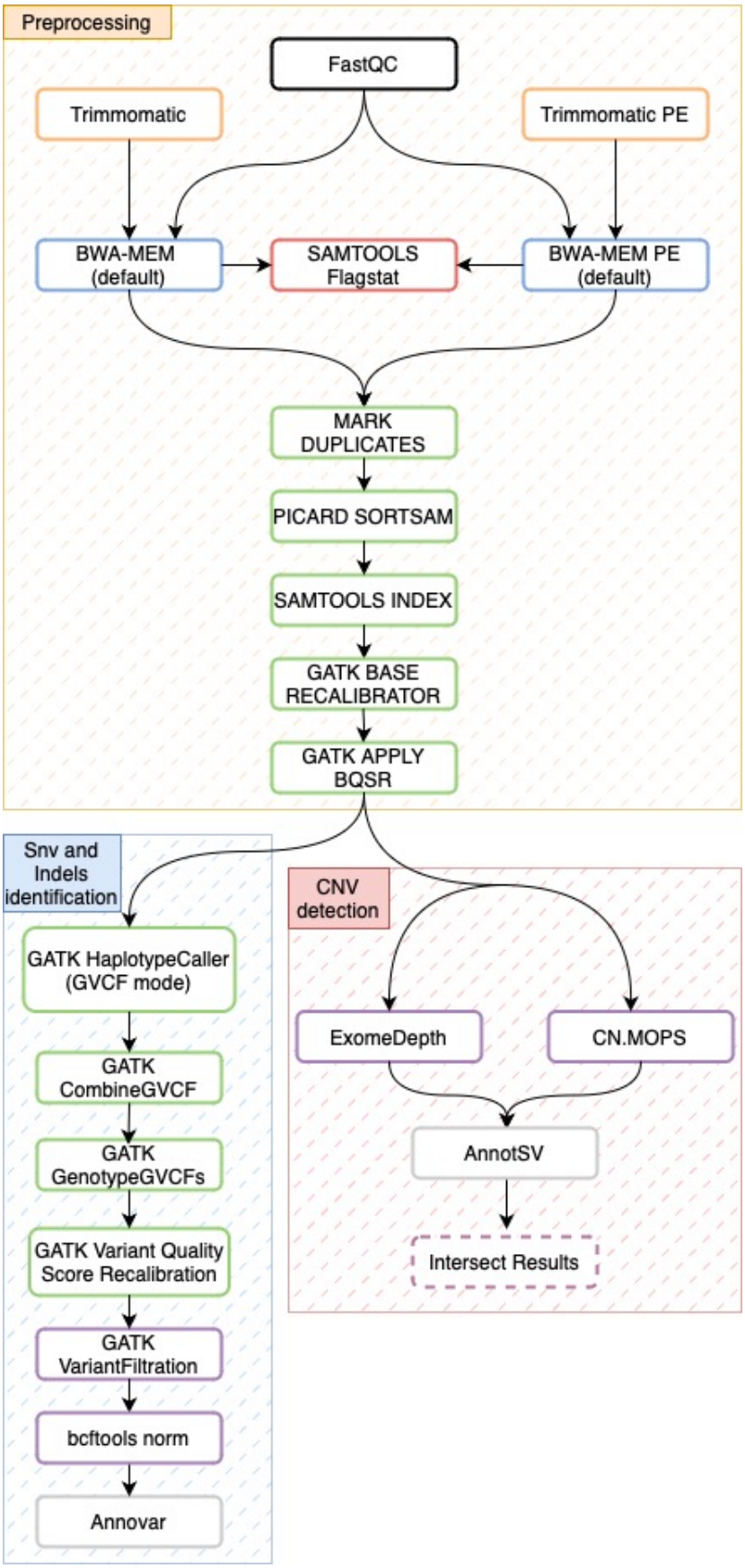


Fig. 4.1 Proposed method for SNV/indels and CNV detection.

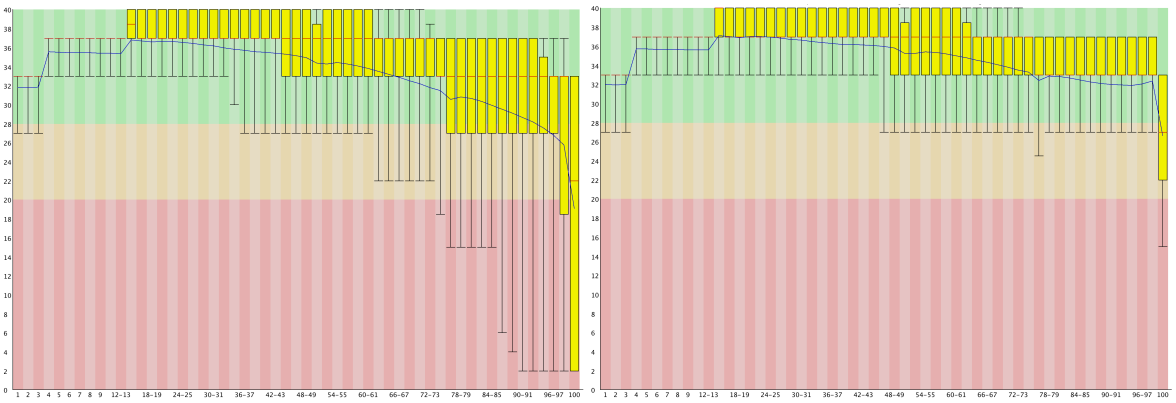


Fig. 4.2 Pre-trimming (left) and post-trimming (right) base read quality of a representative sample. Values in red denote poor quality. After trimming, the read quality is greatly increased.

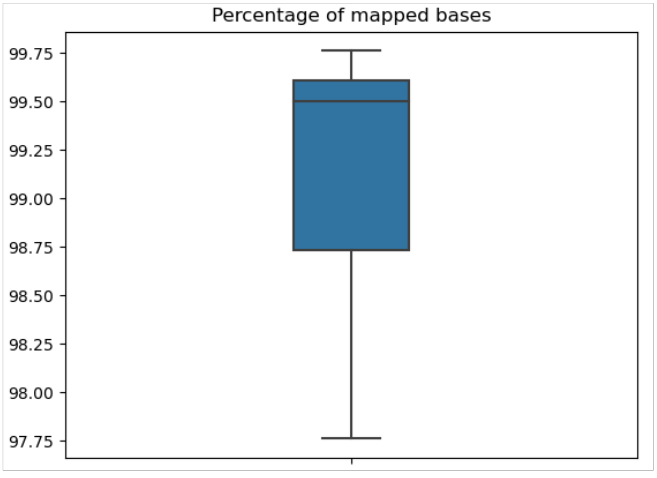


Fig. 4.3 Boxplot of the percentage of mapped bases for all samples of the two datasets.

on the nature of our data, but adaptable also to other data, enhance the overall read quality (**Figure 4.2**).

Read Alignment: The high-quality trimmed reads are then aligned to the reference genome using the BWA-mem tool with default settings (https://bio-bwa.sourceforge.net/bwa.shtml). Accurate mapping of reads to their respective genomic locations is pivotal at this juncture (**Figure 4.3**).

Duplicate Removal and sorting: Duplicate reads are identified and removed with MarkDuplicates. This process curtails redundancy and ensures accurate coverage calculations at specific genomic loci. Then, the aligned reads are sorted by genomic coordinates using Picard.

Alignment Quality Assessment: The alignment quality is assessed using Samtools Flagstat to generate statistics on the mapped files. A high percentage of properly mapped bases attests the reliability of the alignment process.

Base Quality Score Recalibration: Lastly, a base quality score recalibration is executed using the BaseRecalibrator tool from GATK to amend any biases or errors in the quality scores assigned by the sequencer. Known sites from the GATK bundle are utilized to provide additional contextual information for the recalibration process.

4.3.8 SNV and Indels identification pipeline

Within this crucial phase of our analytical pipeline, as illustrated in Figure 4.1 (located on the left side, at the bottom), we meticulously follow the GATK4 best practices for the identification of germline single SNVs and Indels. The GATK framework for short variant discovery offers a robust and streamlined methodology for the precise detection of SNPs and Indels within high-throughput sequencing datasets. Initiating with rigorous quality control measures and the alignment of sequencing reads against a reference genome, the process is designed to ensure the highest data integrity. Subsequent preprocessing steps, including duplicate marking and base quality score recalibration, are critical for enhancing the overall quality of the dataset. The core of variant discovery utilizes the HaplotypeCaller in GVCF mode, a strategy that ensures a thorough capture of variant evidence for individual samples, thereby laying the groundwork for accurate joint genotyping across the research cohort. This step amalgamates individual GVCFs into a singular, meticulously curated variant call set, which is then refined through Variant Quality Score Recalibration (VQSR) and/or stringent hard filtering techniques to further enhance the precision of variant calls.

The culmination of this process is the annotation of variants to elucidate their biological significance, providing valuable insights into the potential impact of each variant. This methodical approach, detailed in **Table 4.1** along with the versions and specific parameters

of the tools used, is pivotal in achieving a high fidelity in variant discovery, particularly in the context of genetic research.

Our application of this strategy focuses on exome sequences, exploring inherited genetic variants within a cohort. Utilizing the HaplotypeCaller in GVCF mode allows for the nuanced identification of SNVs and Indels. Following the HaplotypeCaller step, joint genotyping is performed using GenotypeGVCFs, culminating in a comprehensive list of variants. Our approach incorporates both VQSR and hard filtering to meticulously refine the variant calls, ensuring a dual-layered filter that significantly bolsters the trustworthiness of our findings. Ultimately, variant annotation via Annovar provides a deeper layer of context, offering a richer understanding of the genetic landscape uncovered in our study. This dual-pronged filtering strategy, coupled with detailed variant annotation, underscores our commitment to not only identifying genetic variants but also understanding their broader implications within the realm of human genetics.

To guarantee more targeted variations, we employed Genoox's Franklin's tools (https://franklin.genoox.com/clinical-db/home), which are based on the following criteria: The following categories of variants were identified: i) those found in splicing sites and exonic regions; ii) rare heterozygous variants with minor allele frequency (MAF) below 0.01 in the databases of ExAC, GnomAD, and 1000 genomes; iii) start-loss, stop-gain, stop-loss, and frameshift variants; and) missense variants flagged as deleterious by at least half (6 out of 11) of the in silico prediction tools that were taken into consideration (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, LR, VEST3, and CADD). A read depth over 10 and a quality per depth above 10 are also considered to reduce the false call rate [139].

The American College of Medical Genetics and Genomics-Association for Molecular Pathology (ACGM) guidelines [140] classify variants as either pathogenic, likely/possibly pathogenic, likely/possibly benign or uncertainly significant. These variants are identified through the Franklin ACMG annotation. In the variant and gene prioritisation phase, which aims to identify candidate BC risk loci, we take into consideration: i) variants characterised by evidence of pathogenicity or conflicting interpretation of pathogenicity in the Clinvar database, ii) variants in known/candidate cancer predisposition genes [141]. Additionally, a thorough review of the published data on filtered variations is carried out. Kindly take notice that the application of Genoox's programme was a supplementary step to our main process, which was to rank the changes found by annotation of VCF files.

4.3.9 CNV pipeline

The following section delves into the CNV analysis within our study, as detailed in Figure 4.1 (located on the right side, at the bottom). This component of our methodology, including the specific tools utilized and their respective versions, is systematically outlined in Table 4.1. For the task of CNV identification, we elected to integrate two highly regarded tools, each bringing a unique strength to our analysis.

ExomeDepth ([134]), our first choice, is renowned for its adeptness in identifying CNVs, with a particular forte in uncovering rare variants [142]. It utilizes sophisticated statistical models in conjunction with read-depth information to accurately predict CNV events across the genome. On the other hand, cn.mops ([135]) serves as our second tool, distinguished by its specialized approach to CNV detection that focuses on read depth variability. Unlike simpler read depth methods, cn.mops employs a statistical model that accounts for both the expected and observed read depths, making it particularly effective at identifying CNVs in regions with variable read coverage. The statistical model used by cn.mops works by comparing the read depth at each genomic location to what is typically expected based on other samples. In simple terms, cn.mops creates a "model" of how many reads (short DNA sequences) should be present in a typical genome at each spot. It then compares the actual read counts from the sample being analyzed to this expected model. If the read counts are significantly higher or lower than expected, cn.mops flags these regions as potential CNVs. This capability to handle variability in read depth allows cn.mops to detect CNVs with higher sensitivity and specificity, especially in complex genomic regions where other tools might struggle.

Acknowledging the criticality of integrating multiple CNV detection tools to circumvent the biases that can often skew CNV analysis, our strategy is to harness the strengths of both ExomeDepth and cn.mops. This dual-tool approach is informed by a comprehensive review of the field, aligning with best practices that highlight the necessity of employing diverse methodologies to achieve a more balanced and accurate CNV detection [143]. By incorporating these two prominent and well-validated tools, our pipeline not only benefits from the individual merits of each but also provides a robust framework for CNV analysis, ensuring that our findings are both reliable and reflective of the complex nature of genomic variations.

During the preprocessing phase of our analytical framework, the generated mapped files (with a .bam extension) are seamlessly integrated into the CNV detection workflow, serving as inputs for both ExomeDepth and cn.mops. These sophisticated tools are further equipped to process bed files, which supply exon coordinates delineating the targeted genomic areas under investigation. Such coordinates are instrumental in guiding the analysis towards

specific regions where CNVs may reside, with variations in read counts spotlighting these areas of potential interest.

The CNV detection process meticulously evaluates these discrepancies to identify and catalog CNVs, employing a ranking system that considers both the confidence in and the relevance of each CNV call. This prioritization ensures that CNVs deemed most significant and with the highest reliability are selected for in-depth analysis. The culmination of this stage is the compilation of CNV findings into an accessible format, detailing vital information such as genomic locations and the nature of the CNVs identified, whether they are duplications or deletions.

To enrich the CNV data with comprehensive gene-related insights, the resulting output files (in .bed format) from both ExomeDepth and cn.mops undergo further enrichment through AnnotSV. This annotation step is critical for associating the detected CNVs with specific genes and understanding their potential impact. AnnotSV's contribution to our pipeline enhances the interpretability of the CNV data, providing a more detailed understanding of the genomic landscape affected by these variations. This level of detail is paramount for advancing our comprehension of CNV involvement in genetic conditions, ensuring that our findings are not only accurate but also deeply informative.

We eliminated changes deemed frequent from the analysis by filtering out allele frequencies less than 0.01 in order to provide extremely dependable results. We also evaluated metrics (reads predicted versus observed) between 0.4 and 0.7 for deletions and larger than 1.3 for duplications, which were computed by ExomeDepth for the detected deletion and duplication events. Additionally, manual checks were made for consistency of filtered CNVs, frequency within the analysed BC cases, and quality factors. Take note that following the ranking intersection, only the CNVs found by both tools are kept. We included a prioritisation stage for prospective genes of interest, mostly based on ACMG categorization and the selection of known/candidate cancer susceptibility genes, mirroring the discovery process of SNVs and short Indels [141].

4.3.10 Validation Analysis

Our study delves into a rigorous validation process for both SNPs and Indels, as well as CNVs, to affirm the accuracy and reliability of our variant detection methodologies. This process is essential for ensuring that our findings can be confidently applied to genetic research, particularly in understanding complex diseases. In particular, for the comprehensive validation of our pipeline, we utilized the *NA12878* sample from the Genome in a Bottle Consortium (GIAB), as suggested in [144]

4.3.11 Validation of SNPs and Indels

Methodology

The initial step in our validation process involved downloading the raw sequences and the truth sets of variant calls for the NA12878 sample from the GIAB's FTP official site (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/).

Following data acquisition, the validation of SNV and Indels was meticulously performed using the *hap.py* tool developed by Illumina, as suggested in [145]. This tool is specifically engineered to adeptly handle complex variant types, enabling a distinct evaluation of SNPs and Indels. This capability is crucial for our analysis, as it allows for the separation of metrics for each variant type, thereby providing a granular view of the pipeline's performance. To quantify the performance of our pipeline, we calculated several metrics, including precision, recall, and the F1 score. The F1 score, which is the harmonic mean of precision and recall, provides a balanced measure that considers both false positives and false negatives. It is particularly useful when the class distribution is imbalanced, as is often the case in genomic data where true variant calls can be vastly outnumbered by false positives. For the evaluation, we compared our pipeline's filtered call set against the real truth set provided in the FTP site, specifically named `project.NIST.hc.snps.indels.vcf`. This truth set is recognized as the gold standard for variant calling in the NA12878 sample, comprising high-confidence SNPs and Indels identified through rigorous consensus methodology among leading genomics research institutions.

Results

The detailed performance of our pipeline in detecting SNPs and Indels is summarized in Table 4.2. These results underscore the strengths of our pipeline in SNP detection, marked by high precision and a robust F1 score. However, the relative challenge in Indel detection, especially in achieving a higher recall, indicates an area for further research and development. Our ongoing efforts aim to enhance the algorithm's sensitivity to Indels without compromising the high precision observed in SNP detection.

Variant Type	Recall	Precision	F1 Score
SNPs	85.25%	97.86%	91.12%
Indels	71.13%	81.09%	75.78%

Table 4.2 Validation results for SNPs and Indels.

4.3.12 Validation of CNVs

Methodology

For the benchmarking of CNVs, we meticulously adopted a reference SV baseline callset for NA12878, courtesy of the Mt. Sinai School of Medicine. This callset, derived from approximately 44x coverage of PacBio data, encompasses a merged SV VCF file (high-confidence reference callset), which is crucial for a comprehensive CNV analysis. The data, including the merged SV VCF, is publicly accessible at the GIAB repository under NA12878 PacBio data (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/), providing a valuable resource for high-confidence variant calls. To compare our pipeline's CNV calls against this high-quality reference set, we employed Truvari, a tool specifically designed for the evaluation and comparison of genomic variants. Truvari facilitates the assessment of concordance between our detected CNVs and the reference SV callset, enabling a detailed analysis of our pipeline's performance [146]. However, before leveraging Truvari for comparison, it was necessary to adapt the VCF file obtained from the official site to ensure compatibility and accurate analysis. This adaptation involved modifying the VCF to align with the format expected by the tool, ensuring that the comparison accurately reflects the true performance of our CNV detection methodology.

Results

The validation of CNVs yielded the values in Table 4.3. These results, while indicating a solid foundation for CNV detection, also highlight specific areas for improvement and refinement within our methodology. A recall rate of 60.53% suggests that, although the pipeline is capable of identifying a majority of the true positive CNVs present within the dataset, a significant proportion remains undetected. This gap may be attributed to the inherent complexities associated with CNV detection. The challenges posed by these complexities are further compounded by variations in sequencing data quality and the limits of current detection algorithms' sensitivity. The precision rate of 72.72% demonstrates the pipeline's effectiveness in accurately distinguishing true CNVs from false positives. This high level of precision is indicative of the robustness of the algorithms employed, ensuring that the CNVs identified by the pipeline are indeed present in the genome. The F1 score, which is a harmonic mean of precision and recall, stands at 66.07%. This metric reflects the pipeline's overall efficiency in CNV detection, balancing the need to maximize true positive detections while minimizing false discoveries. While the F1 score confirms that the pipeline performs reasonably well in this complex area of genomic analysis, it also underscores the potential for further optimization.

Given these outcomes, it is clear that enhancing the pipeline's ability to detect CNVs, particularly by improving recall, is a pivotal area for future development. Strategies to achieve this may include refining existing detection algorithms, integrating additional CNV-specific quality metrics, and leveraging advanced computational techniques to better interpret complex genomic regions [147].

Variant Type	Recall	Precision	F1 Score
CNV	60.53%	72.72%	66.07%

Table 4.3 Validation results for CNVs.

4.4 Results

Indels and SNV

Post-calling and filtering, 221 variants were identified with 191 being unique (**Supplementary Table 1**). These variants were categorized as per the ACGM classification and the distribution is illustrated in **Figure 4.4**. To validate, we cross-verified with variants from the original publications [131] [130]. In particular, we were able to confirm the presence of the two most relevant variants previously described: *CHEK2* c.1100delC, p.Thr367fs in family RUL153 and *FANCM* c.5791C>T, p.Arg1931* in sample DAD1) [130] (**Supplementary Table 1**).

Several interesting variants, unmentioned in original publications, were also detected (**Table 4.4**). Notably, we found missense variants in the *RBBP8* (sample F3311_5) and *LEPR* genes (sample 07S240), previously described in high-risk, BRCA-negative BC cases [151], [150]. Frameshift mutations in tumor-related genes such as *DLEC1* (c.5068_5071dupAACA, p.Ser1691fs, sample F2887_24) and *AIM2* (c.1027delA, p.Thr343fs, samples DAD1/RUL153_3) were also found. In addition, we detected variants of potential interest in known BC-related genes: these included the variants c.572T>A (p.Ile191Asn, sample F2887_24) and c.8560C>T (p.Arg2854Cys, sample F3311_5) in the *ATM* gene and c.401C>T (p.Thr134Ile, sample I1408) in the *RECQL* gene. We identified numerous Variant of Uncertain Significance (VUS), largely deemed possibly pathogenic by the ACMG classification (**Figure 4.4**). While not directly linked to disease, they underscore the samples' genetic complexity and offer potential research avenues.

CNV

For CNVs identification, we considered only coherent results from ExomeDepth and cn.mops analysis. The merging operation we implemented reduced the number of CNVs from 102.359

Gene	Transcript	Nucleotide Change	AA Change	Effect	ClinVar Classification	Franklin ACMG classification	Sample	Notes
AIM2	NM_004833.3	c.1027delA	p.Thr343fs	Frameshift	N.A.	UC	DAD1/RUL153_3	Gene involved in cell cycle regulation, suppression of tumor proliferation [148]
ATM	NM_000051.3	c.8560C>T	p.Arg2854Cys	Missense	Conf. int. of Pathog.	UC, Pos. Path.	F3311_5	Known BC-related gene [141]
ATM	NM_000051	c.572T>A	p.Ile191Asn	Missense	UC	UC	F2887_24	Known BC-related gene [141]
DLEC1	NM_007335	c.5068_5071dupA	p.Ser1691fs	Frameshift	N.A.	UC, Pos. Path.	F2887_24	Tumor suppressor gene involved in DNA damage response [141]
EWSR1	NM_005243.3	c.1843C>T	p.Arg615*	Stop Gain	N.A.	Likely Pathogenic	F2887_13	Gene linked with BRCA1/2 pathway and associated with non-medullary thyroid cancer susceptibility [149]
LEPR	NM_002303.5	c.1835G>A	p.Arg612His	Missense	Conf. int. of Pathog.	UC, Pos. Path.	07S240	Same variant identified in high-risk, non-BRCA BC case [150]
PDGFRA	NM_006206	c.2411G>A	p.Arg804Gl	Missense	UC	UC	I1408	Gene associated with prostate cancer and sarcoma predisposition [141]
RBBP8	NM_002894	c.298C>T	p.Arg100Tr	Missense	Conf. int. of Pathog.	Likely Pathogenic	F3311_5	Same variant identified in high-risk, early onset, non-BRCA BC case [151]
RECQL	NM_002907.4	c.401C>T	p.Thr134Ile	Missense	Conf. int. of Pathog.	UC	I1408	Known BC-related gene[141]
SEC23B	NM_006363.6	c.2035G>T	p.Glu679*	Stop Gain	N.A.	Likely Pathogenic	F3311_5	Gene associated with cowden syndrome and sporadic thyroid cancer [152]
TP53AIP	NM_022112	c.63dupG	p.Gln22fs	Frameshift	N.A.	UC, Pos. Path.	RUL36_7/RUL153	Gene associated with melanoma susceptibility [141]
TSC2	NM_000548	c.748A>G	p.Lys250Glu	Missense	UC	UC	F2887_13	Gene associated with colorectal and gland cancers predisposition [141]

Table 4.4 List of the most interesting pathogenic, likely/possibly pathogenic (Pos. Path.) and uncertain significance (UC) variants identified in this study. N.A.: not available. Conf. int. of Pathog.: Conflicting interpretation of pathogenicity.

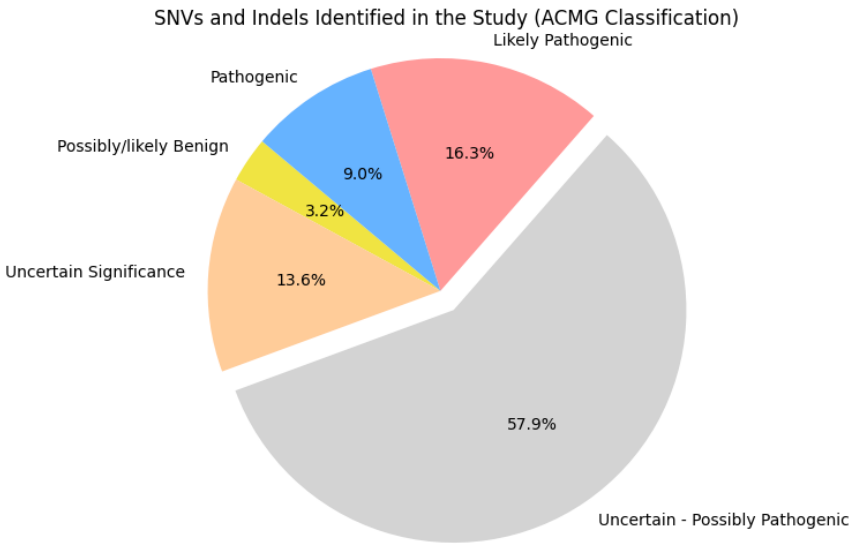


Fig. 4.4 SNVs and indels identified in this study after variant filtering, based on ACMG classification.

to only 983. We then considered parameters, including allele frequency in the population and within the analyzed BC cases, metrics calculated by ExomeDepth and consistency of filtered CNVs, further reducing the number of CNVs.

We detected a total of 69 CNVs affecting 103 genes in BC patients (**Supplementary Table 2**). The ExomeDepth algorithm had an average Bayes factor of 19.56 and read count ratios of 0.56 for deletions and 1.36 for duplications. CNVs were unevenly distributed across chromosomes (chr), with more duplications in chr 1 and 12, and deletions in chr X (**Figure 4.5**). Notably, most X-related CNVs were due to a large region (48316827_49144655) deletion in sample F11S02. We prioritized CNVs based on ACMG annotations and genes linked to cancer predisposition. This analysis revealed the 29496809-29509783 pathogenic deletion on chr 17, overlapping with *NF1* gene and the 86368073-86408032 deletion on chr 9, overlapping with *GKAP1*, both in sample F11S01. Moreover, sample F12S01 showed CNVs overlapping with known cancer-related genes, including *KRAS*, *RBBP8*, *ARNT2*, *ESR1*, *SYCP1*, and *XPO1*.

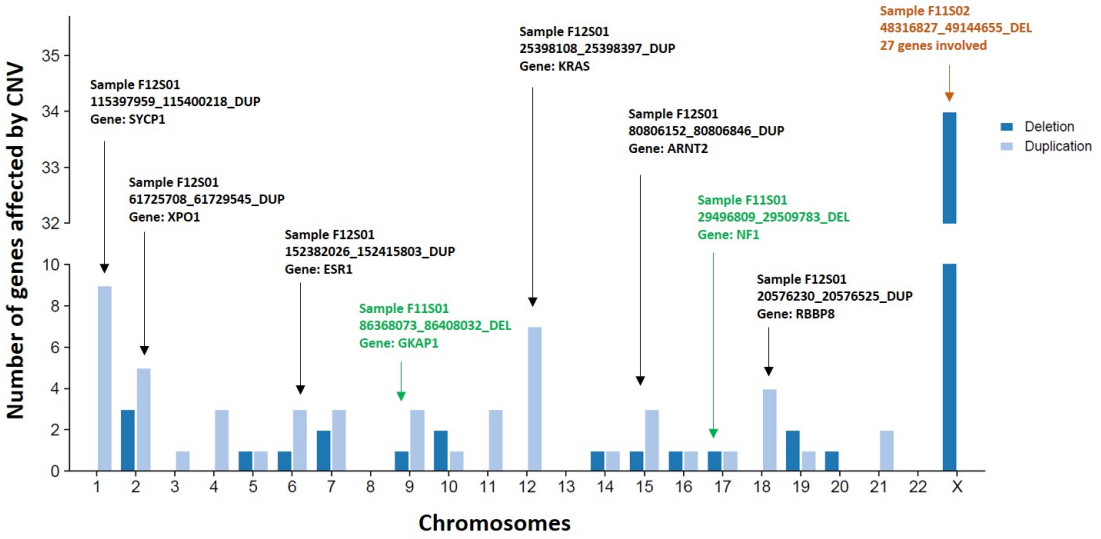


Fig. 4.5 Chromosomal distribution of CNVs. Most relevant CNVs detected in samples F11S01 (green), F11S02 (orange) and F12S02 (black) are shown. The CNVs shown in the figure include those already validated and related to cancer, emphasizing their potential significance in the context of disease.

4.5 Limitations

Our exploration into the genomic intricacies of breast cancer susceptibility through variant calling analysis has unfolded significant insights. However, it is imperative to acknowledge certain limitations that shape the context and interpretation of our findings.

Use of WES Data for CNV Calling: Although WES is highly effective for identifying SNVs and Indels within coding regions, it is less optimal for CNV detection. Even if is out of the cope of the thesis, WGS has been proven to be more reliable for this purpose as it provides uniform coverage across the entire genome, including non-coding regions where CNVs often occur. The use of WES may lead to incomplete or less accurate CNV detection, particularly in regions outside of the targeted exons.

Validation Constraints: A notable constraint in our study is the absence of direct validation for newly identified variants using Sanger sequencing and Multiplex Ligation-dependent Probe Amplification (MLPA). Recognizing this gap, we strategically benchmarked our findings against the NA12878 reference sample from the GIAB, setting a high standard for our NGS methodology's accuracy assessment. This approach, while not a direct substitute for traditional validation methods, draws confidence from studies demonstrating high concordance rates between NGS variants and those confirmed by

established methods [153]. Such findings bolster the reliability of NGS data, suggesting that in contexts of high-quality sequencing, the necessity for confirmatory analyses may be reconsidered.

Technical Challenges in CNV Detection: Detecting CNVs presents a formidable technical challenge within our study. Despite the integration of sophisticated tools like ExomeDepth and cn.mops, the task of accurately identifying CNVs remains fraught with difficulties. These complexities are particularly pronounced in genomic regions characterized by intricate architectural features. Variable sequencing quality and coverage further exacerbate these challenges, influencing the overall sensitivity and specificity of CNV detection. Such technical hurdles underscore the need for continued innovation and refinement in CNV analysis methodologies to enhance accuracy and overcome the inherent obstacles presented by the diverse nature of genomic data.

Adaptation to Other Diseases: While our methodology has proven to be versatile and effective for the breast cancer (BC) subgroup initially targeted, it's important to note a limitation in its direct application to other complex polygenic diseases without adjustments. Each disease context—be it cardiovascular, neurological, or diabetes—may necessitate specific settings, thresholds, and considerations unique to its genetic landscape. Our study primarily focused on optimizing the pipeline for BC, leveraging settings and thresholds tailored to its genomic characteristics. This focus, while yielding significant insights for BC susceptibility, implies that the methodology, as currently configured, may not automatically extend to other diseases with optimal effectiveness. Adapting the pipeline to suit the genomic intricacies of other conditions would require recalibration of parameters and potentially the integration of disease-specific variant databases. This limitation emphasizes the need for bespoke adjustments to our approach when extending its application beyond BC, ensuring that the methodology remains both robust and relevant across different genomic research domains.

4.6 Discussion

The primary goal of this chapter, in keeping with the study's road map, was to explore the field of variant calling analysis. Our goal was to create a comprehensive, reproducible method that effectively catches SNVs, Indels, and CNVs among other genomic variations.

In precision medicine, the identification of novel putative genes or variants implicated in BC susceptibility would have a significant impact on clinical practice [128]. WES analysis has

Fig. 4.6 Visualization of the rare germline variant observed in the final BAM file of the patient under investigation. Generated using JIGV (https://github.com/brentp/jigv).

proven to be a suitable procedure for detecting disease-causing variants and discovering new target genes [154]. Notably, several interesting and promising new putative genes/variants, such as *RCC1 and SERPINA3*, are emerging in BC predisposition by using this method [155, 156]. However, few studies about WES analysis in non-BRCA patients, mainly in a limited number of families, are available [157]. Our study's main contribution is the creation of a specialised method, included in a replicable and adaptable tool we made available to the scientific community, created for processing WES datasets to comprehensively detect SNVs, Indels and CNVs.

We developed a study, using the proposed method, we re-analyzed two WES datasets (PRJEB3235 project [130] and PRJEB31704 project [131]) to look for germline alterations in non-BRCA patients potentially predisposing to familial BC, with the aim of detecting SNVs, Indels and CNVs.

Interestingly, putatively relevant variants, undetected in the original published studies, emerged in the present work: these occurred in genes recently associated with BC predisposition and pathogenesis, including *LEPR* and *RBBP8* [150], [151], or genes involved in cancer-related pathways, including *AIM2, EWSR1, DLEC1, SEC23B and TP53AIP1*, [152], [141]. With this regard, the c.298C>T (p.Arg100Trp) variant in the *RBBP8* gene, which is involved in the homologous recombination DNA repair mechanism, was recently identified in a high-risk, early onset BC case negative for mutations in *BRCA1/2* genes [151]. Similarly, the c.1835G>A (p.Arg612His) variant in the *LEPR* gene, which encodes for the leptin receptor involved in the regulation of lipid metabolism, was identified in a high-risk-familial, BRCA-negative BC patient [150].

CNV detection accuracy varies with the bioinformatics tools and settings utilized. For optimal results, it is advised to merge algorithms from different methods [158]. In our approach, we integrated ExomeDepth and cn.mops. ExomeDepth is considered one of

the most balanced tools for sensitivity and specificity [159], supporting its use in routine targeted NGS diagnostic services for Mendelian diseases [160]. Similarly, cn.MOPS shows the best performance when the size of targeted CNVs is between 100 kb and 10Mb, but it is also a suitable choice for unknown research, as its accuracy is globally satisfactory [158]. Based on a prioritization scale, the most interesting CNV detected in this study was a deletion on chr 17, which includes the known BC susceptibility gene *NF1* [141]. The same sample (F11S01) also showed a deletion on chr 9, overlapping with *GKAP1*, a gene recently suggested to be a candidate susceptibility factor in esophageal squamous cell carcinoma [161]. The deletion of a large genomic region on the X chromosome was also observed in the sample F11S02. This chromosome carries a significant number of oncogenes and tumor suppressor genes, the genetic alteration or dysregulation of which, both at germline and somatic level, has been associated with the development and progression of different cancer types, including BC [162]. Finally, sample F12S01 revealed duplication CNVs overlapping several known cancer-related genes, not only with an oncogenic role such as *ESR1* and *KRAS*, but also identified as tumor suppressors, including the aforementioned *RBBP8* and *ARNT2*. Deletion CNVs result in haploinsufficiency, while duplications can cause triplosensitivity, gene fusion, or disruption. Though most duplications are adjacent to the original locus, inversions or intragenic variations can disrupt genes. Predicting the genetic consequences of such alterations requires specific breakpoint-level analysis [163]. Key CNVs were found in samples without significant SNVs-indels, indicating that in these familial BC cases, susceptibility may arise from these alterations over nucleotide-based variants.

Overall, our methodology has been meticulously crafted with adaptability, allowing for straightforward adjustments to cater to various genomes or resources utilized during research. Importantly, it is capable of adeptly handling both single end-reads and paired-end reads, demonstrating its flexibility in accommodating different data configurations. Employing stringent quality control parameters at various junctures of the pipeline, our tool minimizes the risk of false positives and negatives. This process allowed us to confirm the presence of all the more prominent variants in genes known to be involved in BC susceptibility, such as *CHECK2* and *FANCM*, previously described [130], as well as highlighting the good performance of our method, as demonstrated by the analysis of the reference sample NA12878.

For punctiform variants, our pipeline was validated using the NA12878 reference sample from the Genome in a GIAB, a gold standard in variant calling. Using the *hap.py* tool developed by Illumina, we benchmarked our variant calls against this high-confidence reference set. Our pipeline achieved high recall and precision rates, with an F1 score of

91.12% for SNPs and 75.78% for Indels, which are competitive with other leading variant calling tools in the literature. This high level of accuracy confirms that our method is effective for detecting punctiform variants with both sensitivity and specificity. Regarding CNV detection, we integrated two robust tools, ExomeDepth and cn.mops, and evaluated their performance on WES data. The integration of these tools was specifically chosen to enhance the reliability of CNV calls by leveraging the strengths of both methods. In our study, this dual-tool approach resulted in precision and recall metrics that were higher than those reported for other CNV detection methods using WES data. For example, by only retaining CNVs identified by both tools, we achieved an F1 score of 66.07%, which is a strong result given the inherent challenges of CNV detection in WES data.

While we did not perform a direct comparison with every existing tool, our results indicate that the pipeline offers a high degree of accuracy and reliability across different types of variants. In particular, the accuracy metrics here obtained, make the performance of our pipeline comparable to others already described in the literature for SNV-indels detection [164]. Furthermore, our concerted efforts to refine the pipeline's sensitivity to Indels and enhance CNV detection underscore our commitment to advancing genomic analysis [164, 147].

Of note, CNV callers tends to be more challenging both because: i) these variants are more difficult to accurately detect using short read sequencing data, which makes structural variants calling more error-prone than small variants calling, ii) the precise breakpoints for CNVs are not always well defined, which makes comparison between call-sets more complex. Therefore, even the best performing structural variants callers with whole-genome sequencing data achieve F-1 scores of 0.80-0.90 [165]. However, to date, WES is the most widely used NGS approach in clinical diagnostics and academic research [166], so there is a need to find accurate solutions and strategies for detecting CNVs also from WES data. Here, using two well-known CNV detection tools (ExomeDepth and cn.mops), taking into account quality parameters and retaining only the CNVs identified by both tools, we obtained performance metrics higher to others already described in literature based on WES data [167].

This pipeline, enclosed within the Snakemake workflow management system, thus offers a turnkey solution for researchers, particularly in the realm of BC genetics. By consolidating the analytical process into one unified system, we significantly reduce the time and expertise required to configure and run genomic analyses. This approach allows researchers to swiftly apply our pipeline to their datasets, enabling a more focused investigation of genetic underpinnings in diseases without the burden of technical complexities often associated with genomic data analysis. The openness and accessibility of our pipeline contrast sharply with

the closed nature of many in-house tools [127], providing a valuable resource for the wider research community to engage in collaborative improvements and benchmarking efforts.

As we transition to the next chapter, we pivot towards integrating clinical text with knowledge graphs, expanding our analytical horizon. This integration, which will also encompass the management of genomic variants, represents a novel frontier in our exploration. It embodies our continuous journey towards a holistic understanding of genetic underpinnings in diseases, particularly in the context of familial BC. By marrying genomic data with clinical insights through knowledge graphs, we aim to unlock deeper, more actionable insights that can further refine precision medicine's promise.

A limitation of our study is the inability to directly validate new variants with Sanger sequencing and MLPA (Multiplex Ligation-dependent Probe Amplification). To address this, we experimentally benchmarked our data against the NA12878 reference material provided by the GIAB, serving as a standard for assessing the accuracy of our NGS approach. Despite the validation constraints, a study validating 1109 NGS variants from 825 clinical exomes reported 100% concordance for SNV and Indel variants and 95.65% for CNVs with traditional methods. This suggests that, especially with high-quality NGS data, confirmatory analysis might not always be essential [153]. Of note, while our initial focus was on a specific BC subgroup, we underline that the methodology we set-up has broader applications and it can certainly be considered a suitable tool for advancing our understanding of other high-impact complex polygenic diseases (e.g.cardiovascular, neurological, diabetes).

Additionally, we initiated another experiment aimed at variant calling in a rare patient. Utilizing the SNP and Indels pipeline integrated into our approach, we employed Single-sample calling instead of Joint Genotyping, as our focus was on a single sample. Collaborating closely with geneticists from our university, we delved into exploring variants in this rare case, particularly emphasizing Snps and InDels. Notably, among the variants identified, one has captured our attention: *TGFBR1(NM_004612.4):c.1335_1338del (p.Cys446AsnfsTer4)*. This variant, validated in the laboratory, holds significance within the genomic landscape. For clarity, the variant of interest is situated in the region "9:101910011-101910014" on the hg19 genome assembly. To provide a visual representation of the variant, we have included Figure 4.6, depicting the plot of the variant as observed in the final BAM file corresponding to the patient under investigation. However, as this endeavor is currently in progress, we are unable to disclose detailed results at this time.

The article has been accepted by NAR Genomics and Bioinformatics | Oxford Academic (https://academic.oup.com/nargab) and available at [168]. Details of the methodologies and technical implementation are available at the provided https://github.com/anbianchi/IntegratedSNVINDELSandCNV.

Chapter 5

Harmonizing Medical Knowledge: Graph-Based Integration of Genomic and Clinical Data

In the final chapter, we target a set of biological and medical entities within a dedicated multi-source diagnosis scenario, integrating bioinformatics data and clinical narratives. This expanded framework is based on knowledge graphs, facilitating a unified view of patient health. This approach enhances the precision of diagnostics and treatment strategies, leveraging comprehensive data integration to support advanced personalized medicine.

5.1 Motivation

At the heart of transforming healthcare, integrating genomics with clinical narratives emerges as a cornerstone for revolutionizing disease prevention and management. The forthcoming chapter pivots to highlight the indispensable role of genomic information within clinical texts, underpinning its crucial impact on disease prevention and precise diagnosis. Emphasizing KGs as our tool of choice, we delve into extracting actionable insights from clinical texts enriched with genetic data. A dedicated scenario, "Multi-source diagnosis," showcases this integration's power, pulling together disparate clinical diagnoses from various hospitals or research centers. This approach not only underscores the importance of genomics in clinical decision-making but also sets the stage for personalized treatment strategies, heralding a new dawn in precision medicine where every patient's treatment is as unique as their genetic blueprint.

In modern healthcare systems, a patient often consults with multiple specialists across different institutions, leading to multiple diagnostic records. These records, though rich in information, can often be fragmented and inconsistent [169]. Especially today, the emergence and identification of numerous rare diseases have underscored the critical importance of specialized diagnostic efforts, particularly in the realm of genetic testing [170], [171]. As a result, for chronic or complex illnesses, a single individual may have many diagnoses, sometimes different and spanning different time periods and institutions.

While this multitude of data sources should, in theory, provide a comprehensive view of a patient's health, it often results in the opposite: a fragmented, and occasionally contradictory puzzle of information [172]. For a medical professional, piecing this puzzle together can be difficult but also time-consuming. If certain medical writings follow standardised forms, such using the Human Phenotype Ontology (HPO) terminology [45] or certain coding schemes, the situation becomes more complicated when one looks at medical texts that don't follow any set format. This overwhelming and fragmented landscape of patient data can lead to gaps in understanding, potentially causing misdiagnoses, redundant testing, and even treatment errors [173], [174]. There is a clear and pressing need for an efficient, intuitive, and unified system that can consolidate this plethora of information into an accessible and insightful format.

Knowledge graph [175] is a systematic way to connect information and data points to knowledge. These graphs may effortlessly combine intricate patient data in the context of medical diagnostics, making them an appropriate solution for managing discussed challenges [176], [177].

This chapter introduces an approach to tackle the problem of multi-source diagnostic data integration. Our system is focused on entities that are pivotal in understanding and managing genetic information and rare diseases: Genes, Diseases, Chemicals, Species, Variants, and Cell Types.

By leveraging Named Entity Recognition (NER), Entity Normalization and Relationship Extraction (RE) techniques on raw medical texts [178], we propose the generation of knowledge graphs for each diagnosis coming from different diagnostic sources. The culmination of this process is the merging of these individual graphs into a unified knowledge graph, offering a panoramic view of a patient's medical history. Notably, our method doesn't just amalgamate; it also highlights unique entities and relationships, ensuring that no nuance or detail is lost in the integration, enabling doctors to capture a holistic view of a patient's health profile from different diagnostic perspectives.

The primary contributions of this chapter are: (*RQI*) a method to generate individual knowledge graphs from raw medical texts, (*RQII*) a mechanism to merge these individual

graphs while highlighting unique entities, and (*RQIII*) a visualisation tool to assist medical professionals in understanding a patient's comprehensive medical history.

5.2 Related Works

Different approaches and goals have been seen in the field of building knowledge graphs from medical and biological texts. In order to provide a more comprehensive representation of medical situations, some research projects aim to augment textual data with multiple notations that include genetics, proteomics, symptoms, and more [44][179]. Others are focused on developing knowledge graphs that are specialised to particular illness types and provide in-depth insights into their complex dynamics [180]. Moreover, some initiatives, such as [181] and [182], aim to generate knowledge graphs straight from spoken dialogues or utterances recorded in-context clinical encounters. In [183], the authors developed MedKaaS Tools, where existing clinical datasets are integrated and translated into insights intended to augment human reasoning and accelerate scientific research. However, the gap of systems that compare and integrate knowledge graphs produced from different diagnoses remains a challenge in the field.

Kulkarni et al. [181] proposed a method to construct a medical knowledge graph directly from clinical conversations between doctors and patients, which also employed a mathematical approach for tuple validation and leveraged the knowledge graph for disease prediction. Unlike this work, our approach specifically targets the integration of diagnostic data from various healthcare centers, providing a unified visualization that emphasizes patient's whole medical journey rather than predictive analysis from singular clinical conversations.

PrimeKG [44] serves as a multimodal knowledge graph for precision medicine, integrating data from 20 resources to offer insights across ten biological scales, from protein perturbations to therapeutic drug actions. [179] introduces the Clinical Knowledge Graph (CKG), an expansive platform designed to integrate diverse biomedical data, including proteomics, to facilitate precision medicine. CKG, encompassing over 16 million nodes and 220 million relationships, aims to represent experimental data, public databases, and literature while implementing advanced statistical and machine learning tools to enhance proteomics workflows. In [180], the research underscores the role of computational methods in distilling knowledge from the vast datasets produced in scientific investigations, emphasizing the critical case of Alzheimer's disease. Differently from [44], [180] and [179], our research is tailored towards unifying diagnostic data from multiple healthcare centres, providing a comprehensive visual picture of a patient's medical trajectory.

Finally, the research in [182], emphasizes the challenges clinicians face in accessing accurate and updated medical information due to the vast and evolving landscape of medical literature. They introduce the Focused Clinical Search Service (HGFCSS), powered by the Elsevier Healthcare Knowledge Graph, designed to interpret focused clinical queries and fetch relevant content from diverse medical literature. [184] investigates an automated methodology for constructing high-quality knowledge bases linking diseases and symptoms directly from electronic medical records. Using three probabilistic models, they created knowledge graphs, which were then validated against Google's manually-built knowledge graph and expert physician opinions. Their results demonstrate the feasibility of automatically constructing high-quality health knowledge graphs from medical records with significant precision. In contrast, our work emphasizes merging diagnostic data from diverse healthcare centres, aiming to visualize a patient's cumulative medical journey rather than linking diseases and symptoms. Finally, [185] introduces QAnalysis, a tool that allows doctors to pop in questions in plain language and get back answers in the form of charts and tables. It does this by breaking down the input, linking it to concepts in a knowledge graph, and then turning those concepts into queries that can be run on Neo4j. On the flip side, our project is more about piecing together diagnostic data from different places into one unified knowledge graph. So, while QAnalysis is answering doctors' questions, we are showing a complete visual story of a patient's medical journey.

5.3 Medical Knowledge Harmonization

Our approach diverges from existing methodologies by crafting a cohesive knowledge graph from scattered medical histories, offering a comprehensive perspective through the amalgamation of diverse diagnostic information. It adeptly navigates through complex medical texts to map out pertinent biomedical entities and their interrelations, thereby crafting an intuitive visual narrative of a patient's health journey. This synthesized graph allows for the rapid identification of significant health patterns and anomalies, markedly improving the precision of diagnoses and the formulation of treatment strategies [39].

The system addresses the challenge posed by disparate diagnostic reports, which may offer conflicting insights due to being produced at various times and by different healthcare providers, potentially missing crucial health information [186], [187]. It overcomes the risk of overlooking vital historical health information by integrating these varying perspectives into a single, unified knowledge graph. This graph not only highlights shared findings across reports using uniform visual cues but also brings attention to unique diagnostic insights from individual reports. Consequently, it furnishes healthcare professionals with a

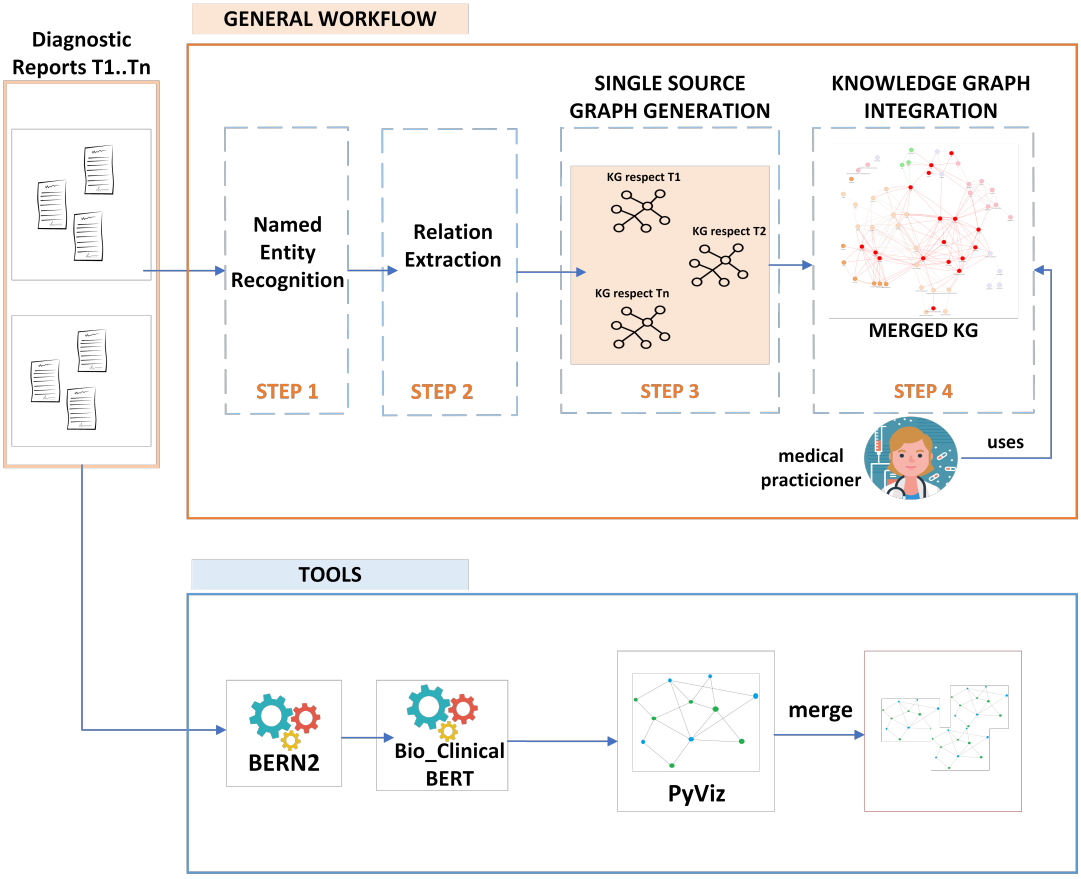


Fig. 5.1 High-Level Workflow for the system.

complete overview of a patient's health, supporting more accurate clinical decisions, avoiding unnecessary procedures, and fostering optimal patient outcomes [188], [44].

Thus, our approach aims to provide a comprehensive understanding of a patient's medical history by integrating disparate diagnostic texts into a single knowledge graph. A sequence of methodical procedures is used to accomplish this transition, as Figure 5.1 illustrates. The diagram illustrates our process in terms of four essential macro-steps: The four processes are as follows: 1) Named Entity Recognition (NER), which identifies relevant entities from the raw texts; 2) Relationship Extraction (RE), which extracts relevant relationships between identified entities; 3) Single Source Graph Generation, which creates separate knowledge graphs for each patient's diagnostic source; and 4) Knowledge Graph Integration, which combines these separate graphs into a single, comprehensive knowledge graph. In the sections that follow, each process step will be covered in detail.

5.3.1 Input Source Determination and Preprocessing

The system initially ingests multiple diagnostic raw texts (T1, T2,..., Tn in Figure 5.1), originating from various healthcare settings and encompassing distinct chronological events in a patient's medical journey. The system is versatile, allowing operation in two distinct modes. The **Data Ingestion Mode** requires a structured dataset as input and proceeds to generate and integrate knowledge graphs based on the data available therein. In the **Manual Mode**, users can utilize the system to process raw diagnostic texts manually. In this mode, users upload individual diagnostic reports into the *diagnostic_reports* folder. The system then processes these texts, extracting and integrating the knowledge graph in a similar manner to the Dataset Mode. Manual mode is particularly useful for ad-hoc analysis of specific diagnostic reports.

5.3.2 Entity Recognition and Normalization

To find medically relevant items, NER is applied to every diagnostic text (T1, T2,... Tn). This stage ensures accurate entity extraction by using NER approaches specifically designed for medical and biological literature. Our approach uses the cutting-edge biomedical tool BERN2 ([189]) for this crucial work. BERN2 can recognise and normalise nine distinct entities: Gene, Disease, Chemical, Species, Mutation, Cell Line, Cell Type, DNA, and RNA. In order to ensure correct entity extraction by navigating through the complex and domain-specific language of medical and biological literature, BERN2 employs unique methodologies for multi-task NER. After identifying the entities, BERN2 proceeds to normalise these entities using specific techniques that improve the accuracy and consistency of the entities found in the diagnostic texts.

5.3.3 Relation Extraction

Following the recognition and normalization of entities within the diagnostic texts, our system embarks on the RE phase. This critical stage is dedicated to uncovering the intricate web of connections among the identified biomedical entities. To achieve this, we harness the sophisticated capabilities of Bio_ClinicalBERT [190], a model meticulously crafted for clinical text analysis. Bio_ClinicalBERT represents a fusion of BioBERT's extensive pretraining on biomedical corpora with subsequent specialization on the MIMIC-III dataset, which comprises a vast array of electronic health records from intensive care unit patients. This model's training regimen, encompassing a diverse collection of clinical notes, equips it with a nuanced understanding of clinical discourse [191].

Although Bio_ClinicalBERT was not initially conceived for directly identifying relationships between entities, its embeddings are imbued with rich biomedical and clinical context. This depth of contextual understanding allows us to employ a heuristic method for inferring potential relationships among the entities detected. By leveraging the contextual embeddings provided by Bio_ClinicalBERT, we can illuminate possible connections that might not be immediately apparent, circumventing the need for a model explicitly trained on RE tasks.

Our experiments are further aligned with the model's domain expertise by incorporating data from the MIMIC database. This strategic choice ensures that our approach benefits from the inherent knowledge and clinical acumen embedded within Bio_ClinicalBERT. Thus, we establish a coherent experimental framework that is both informed by and reflective of the complex realities of clinical data analysis, leveraging state-of-the-art NLP tools to enhance our understanding of patient health narratives.

5.3.4 Knowledge Graph Generation

After extracting entities and their respective relationships, the system leverages on these to construct individual knowledge graphs for each diagnostic text, utilizing entities as nodes and their relationships as edges to graphically illustrate the information embedded within each text. The visualizations are facilitated through the use of PyViz [1] and Networkx [2], ensuring a clear and interactive graphical representation of the data. Following the generation of these individual knowledge graphs, the system goes to the integration phase, wherein it amalgamates these multiple graphs into a unified knowledge graph. This consolidated graph stands as a coherent synthesis of information, amalgamating insights from all diagnostic sources and providing a comprehensive visual depiction of a patient's entire medical history. The visually integrated knowledge graph also highlights common entities and relationships with consistent colours. Unique entities or relationships, which are specific to a particular diagnostic source and not present in others, are highlighted using distinct colors.

5.4 Experimental Settings

Now, we delve into the specifics of how our research was conducted, ensuring transparency and reproducibility. Section 5.4.1 subsection provides insights into the computational infrastructure utilized for our research, detailing the specifications of the machines that supported our work. Then, we discuss the dataset in subsection 5.4.2 section, where we detail its origin,

[1] For more information, visit: https://pypi.org/project/pyvis/0.3.1/

[2] For more information, visit: https://networkx.org/documentation/stable/release/release_2.5.html

the preprocessing and transformation steps undertaken, and the final shape of the data used for our experiments. Then, in section 5.4.3, we outline the considered software tools and libraries in the study.

5.4.1 Hardware Configuration

The research was conducted on Caliban, a cluster environment provided by University Of L'Aquila comprising multiple nodes, utilizing 40 processing units for parallel execution under the "mpi" parallel environment. Specifically, the experiments were executed on a system running CentOS Linux release 7.4.1708 (Core) and powered by an Intel(R) Xeon(R) CPU E5-2698 v4, operating at 2.20GHz, with an available RAM of 141GB.

It is important to note that while the aforementioned hardware configuration was utilized in our research, the experiments can feasibly be conducted on various hardware setups. The tools and methods implemented in our research are not bound to a specific platform or hardware configuration, thereby enhancing the adaptability and reproducibility of our work across varied environments. However, alterations in the hardware setup could influence the computational time for the experiments, despite not affecting the overall outcomes or insights derived from the research.

5.4.2 Dataset

To ensure adherence to ethical standards, our research prioritizes data privacy and security, particularly using data from the *MIMIC-IV-Note: Deidentified free-text clinical notes, version 2.2* dataset, available at PhysioNet (https://physionet.org/content/mimic-iv-note/2.2/). The MIMIC-IV database ensures all patient information is de-identified in alignment with HIPAA's Safe Harbor provisions, guaranteeing anonymity and privacy. It is a collection of deidentified free-text clinical notes for patients included in the MIMIC-IV clinical database [191]. MIMIC-IV-Note contains *331,794* de-identified discharge summaries from *145,915* patients and *2,321,355* de-identified radiology reports for *237,427* patients, all of whom were admitted to the hospital and emergency department at the Beth Israel Deaconess Medical Center in Boston, MA, USA. While the dataset also contains radiology reports, our research strategically narrows its focus exclusively on the discharge summaries to explore the historical medical journey of patients, focusing on fixed entities. Consequently, our analyses and experiments are confined to the file *'discharge.csv'*, found within the *'notes'* folder of the dataset. The *'discharge'* table, and accordingly the *'discharge.csv'* file, encompasses discharge summaries that are inherently crucial to hospitalization data. Our particular emphasis on discharge summaries is driven by our interest in specific entities and the imperative to

construct a coherent and insightful knowledge graph. To create a comprehensive dataset conducive to our research on patient diagnoses, we did meticulous preprocessing on the original dataset. Our objective was to derive a dataset that encapsulated all the diagnoses obtained from each patient's medical journey. This was achieved through the following steps:

- **Filtering Relevant Notes:** from the vast collection of clinical notes in the database, we focused on the *"discharge"* notes, as they provide comprehensive summaries of patients' hospitalizations, encompassing diagnosis, treatment, and medical history.
- **Extraction of *History of Present Illness*:** this section is crucial in the medical context and was therefore crucial for extraction and analysis in our study. It provides a thorough narrative that discusses the patient's condition in a specific instance of time and during a particular hospitalization. Using regular expression-based parsing, we extracted the section of the notes corresponding to the *'History of Present Illness'*, which offers insights into the patient's condition upon admission.
- **Transformation for Multiple Hospitalizations:** recognizing that some patients might have multiple hospitalization records, we restructured the dataset to capture each unique hospitalization event for a patient as a separate column. This transformation allowed us to track and analyze the progression of a patient's medical condition over multiple hospital visits.
- **Filtering Patients with Multiple Diagnostics:** to ensure the richness and comprehensiveness of our dataset, we retained only those patients who had multiple diagnostic events or entities in the original dataset. This filtering step ensured that our dataset was focused on patients with potentially complex medical histories or rare situations.

The resulting dataset encompasses 59,051 unique patients, with distinct hospitalization event and associated *'History of Present Illness'*. The preprocessing of the original dataset was executed in approximately 18 minutes. With the described preprocessing we moved from *145,915* patients in the original dataset to *59,051* patients considered in the evaluation. The final dataset cannot be distributed due to licence restriction on the original one.

5.4.3 Software Configuration

Standard Python libraries, like *re* (for regular expressions), *os* (for operating system-related tasks), *requests* (for HTTP requests), and *hashlib* (for hashing operations) set the base environment.

Table 5.1 Versions of the software and tools utilized in the research.

Software/Tool	Version & Notes
Python	3.6.13
Conda	22.11.1
BERN2	API using Requests version 2.27.1
Bio_ClinicalBERT	loaded using Transformers version 4.9.2
PyVis	0.3.1
NetworkX	2.5.1

For the **NER step** we used **(BERN2 [189])**, an advanced biomedical entity recognition service. Through the *load_bern2_model* function, diagnostic reports are processed to retrieve named entities. The entities are then parsed and structured to be used in subsequent steps. BERN2 is particularly adept at identifying medical entities from raw text, ensuring that every critical piece of information is captured.

For the **RE step** we selected *(Bio_ClinicalBERT [190])*, a variant of BERT that's specialized for clinical and biological texts. This model's embeddings are pivotal in our approach to relation extraction. For each pair of entities in a report, embeddings are generated. Then, the cosine similarity between entity pairs determines if a relation exists, creating it if the similarity surpasses a predetermined threshold fixed to *0,85*.

For the **Knowledge Graph Generation** we considered **(PyVis (https://pypi.org/project/pyvis/0.3.1/)** and Networkx (https://networkx.org/documentation/stable/release/release_2.5.html). The entities and relations derived from the aforementioned steps are organized into individual knowledge graphs using *networkx*. The integrated knowledge graph is visually represented using the *pyvis.network* module for an interactive experience. These graphs not only consolidate information but also offer a visually engaging representation. This ensures that medical professionals can quickly grasp the interconnectedness of various diagnoses. For detailed information regarding the versions of the tools and software used in this study, please refer to Table 5.1.

5.5 Results

This section delves into two experiments designed to evaluate the efficacy of our system in light of the three initial primary research questions: (RQI) which probes the system's capability to autonomously generate individual knowledge graphs from raw medical texts;

(RQII) which examines the system's adeptness in merging various knowledge graphs while accentuating unique entities; and (RQIII) which assesses the utility of the visualization tool in helping medical professionals to decipher a patient's medical history.

The first experiment (detailed in Seciton 5.5.1) presents where the tool analyses the preprocessed dataset of 59,051 patients. This hypothetical situation might be similar to situations in which healthcare systems try to automatically create and preserve knowledge graphs for a large number of patients to aid in future consultations and plan creation. Here, we examine issues including tool scalability, automation effectiveness, and the capability to handle large amounts of data.

The second experiment (detailed in Section 5.6.1) explores instead the situation in which a healthcare provider gives diagnostic text in an effort to obtain a comprehensive picture of a patient's medical history. We examine the tool's effectiveness in deriving a knowledge graph from manually provided data, identifying items, and constructing linkages. This study sheds light on the tool's usefulness in scenarios when on-the-spot analysis is critical, such as during patient consultations or professional team meetings.

It is important to note that, only one example of each experiment will be presented and examined in the following sections due to space limitations. Nevertheless, we have included a subset of pre-analyzed cases in the repository (all the details are in the readme.md of the github repository), which readers can investigate for further insights and analyses in order to enable a deeper exploration and to make it easier to grasp the variety of scenarios our system can accommodate (https://github.com/anbianchi/knowledge_frombio).

5.5.1 Experiment 1: Automated Knowledge Graph Generation in Data Ingestion Mode

Example Case: Patient #10001876. Number of associated medical reports: 2.
Medical Report 1: Ms. ___ presented for evaluation of urinary complaints and after review of records and cystoscopy was diagnosed with a stage III cystocele and stage I vaginal prolapse, both of which were symptomatic. She also had severe vaginal atrophy despite being on Vagifem. Treatment options were reviewed for prolapse including no treatment, pessary, and surgery. She elected for surgical repair. All risks and benefits were reviewed with the patient and consent forms were signed.
Knowledge Graph for Medical Report 1. In Figure 5.2, we report the Individual Knowledge Graph generated by the system for Medical Report 1. Here only 3 interrelated entities have been extracted.

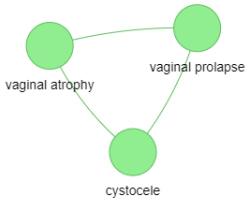


Fig. 5.2 Knowledge graph resulting from Medical Report 1 of the Experiment 1

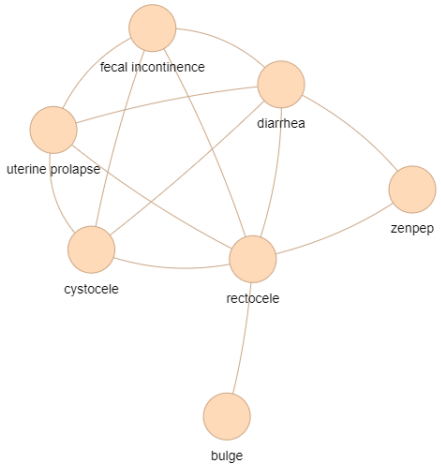


Fig. 5.3 Knowledge graph resulting from Medical Report 2 of Experiment 1

Medical Report 2: She is a ___ patient who presents with ___ rectocele after having a sacral colpopexy and supracervical hysterectomy in ___ for uterine prolapse and cystocele. At that time, she had no rectocele at all. She has symptoms of bulge and pressure in the vagina that has gotten worse over the past few months. She also complains of feeling of incomplete emptying. She states that after she goes to the bathroom, she could go back and urinate some more. She had some frequency, urgency symptoms, which had resolved postoperatively. She also has resolved diarrhea after being started on Zenpep. She is followed by Dr. ___ and her fecal incontinence has resolved as well as resolved diarrhea."

Knowledge Graph for Medical Report 2. In Figure 5.3 we report the Individual Knowledge Graph generated by the system for Medical Report 2

Merged Knowledge Graph of Experiment 1. Figure 5.4 reports the merged knowledge graph for the Experiment 1. In this graph, the common entity *cystocele*, existing in both reports, serves as a crucial point of connection, providing insight into an ongoing or recurring medical condition. This mutual entity not only establishes a link between the separate hospitalizations but also potentially indicates a persistent or chronic condition that warrants consistent monitoring and management. Focusing on such commonly occurring entities

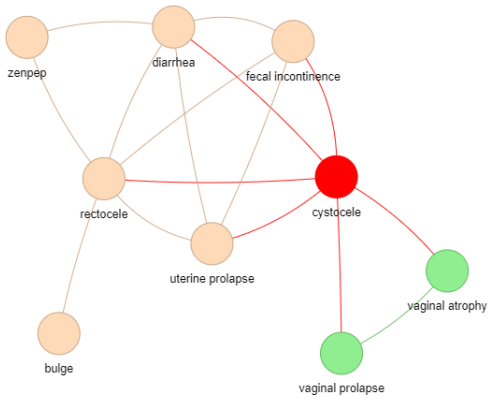


Fig. 5.4 Knowledge graph representing the merging of Medical Reports 1 and 2 for Experiment 1.

is crucial because it helps in tracking the progression or recurrence of specific medical conditions. It aids healthcare professionals in understanding the persistence of certain ailments, which can be vital in shaping therapeutic strategies, ensuring they are aligned with the patient's comprehensive medical history and current health status.

On the other hand, entities that are not common across reports are equally vital to focus upon. These unique entities provide a window into the varied medical conditions and interventions experienced by the patient over time.

For example, the entity *'vaginal prolapse'* appearing in the first report, and entities like *'rectocele'* and *'fecal incontinence'* from the second report, illuminate distinct medical episodes that the patient has navigated through. Also, cystocele, vaginal prolapse, and faecal incontinence are three interrelated pelvic floor disorders in which, respectively, the bulging of the bladder into the vagina, the descent of pelvic organs, and the inability to control bowel movements can all be caused by the weakening or dysfunction of the pelvic supportive structures [192], [193] and [194].

Duration and Technical Details: In this experiment, the generation of knowledge graphs for each diagnostic report and the subsequent creation of the merged graph took approximately 20 seconds in this example case involving two reports. However, it is imperative to note that this time metric can be variable, particularly when dealing with patients with a larger number of reports.

5.5.2 Experiment 2: Generating a Merged Knowledge Graph in Manual Mode

Example Case: Patient #10001876. Number of associated medical reports: 2.

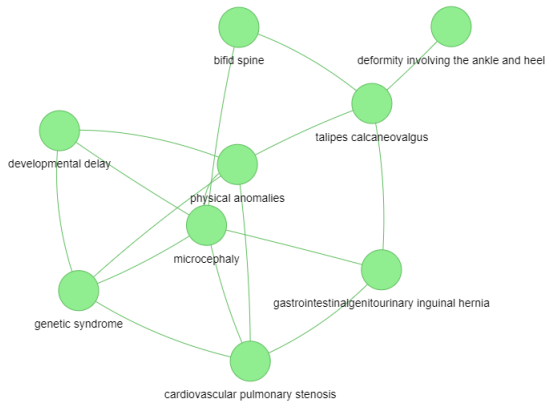


Fig. 5.5 Knowledge graph of Medical Report 1 of Experiment 2

Medical Report 1: Age at last evaluation: ___ year and ___ months. Anthropometric Data: At Birth: Weight: ___ grams, Length: 46.5 cm, Occipitofrontal circumference (OFC): 32 cm. At ___ Year: OFC: 43 cm (P<3), indicative of Microcephaly. Clinical Findings: Neurodevelopmental: Moderate developmental delay noted, Stereotypic movements observed, Microcephaly identified with OFC below the 3rd percentile. Cardiovascular: Pulmonary stenosis diagnosed. Musculoskeletal: Hidden bifid spine detected, Talipes calcaneovalgus (a deformity involving the ankle and heel) present. Gastrointestinal/Genitourinary: Inguinal hernia diagnosed. Genetic Considerations: The combination of microcephaly, developmental delay, and other physical anomalies may suggest a possible genetic syndrome. Genetic testing, including chromosomal microarray and/or whole exome sequencing, may be indicated to identify any underlying genetic etiologies.

Medical Report 2: Age at last evaluation: ___ year and ___ months. Anthropometric Data: Neurodevelopmental: Moderate developmental delay noted, Stereotypic movements observed, Microcephaly identified with OFC below the 3rd percentile. Cardiovascular: Pulmonary stenosis diagnosed. Musculoskeletal: Hidden bifid spine detected, Talipes calcaneovalgus present. Gastrointestinal/Genitourinary: Inguinal hernia diagnosed. Genetic Testing Results: Gene: PAICS (NM_001079525.1), Mutation Identified: c.1165G>C ; p.Gly389Arg. Genetic Considerations: The identified variant in the PAICS gene may potentially explain some of the clinical findings observed in this patient. The PAICS gene encodes a bifunctional enzyme involved in de novo purine biosynthesis. Mutations in this gene could potentially affect cellular proliferation and neurological function, although the exact clinical significance of the identified variant (c.1165G>C ; p.Gly389Arg) needs further evaluation.

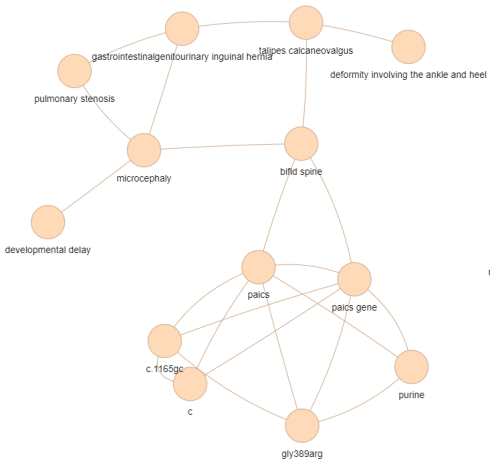


Fig. 5.6 Knowledge graph of Medical Report 2 of Experiment 2

Merged Knowledge Graph of Experiment 2. Figure 5.7 reports the merged knowledge graph for the Experiment 2. In Figure 5.5 and Figure 5.6 we report the Individual Knowledge Graph generated by the system for Medical Report 1 and 2.

In this experiment, an extensive evaluation of a pediatric patient's symptoms necessitated a genetic investigation. Following the genetic testing, a medical variant in the PAICS gene was discovered, due to the linking to particular medical conditions [195], [196] and [197], shedding light on the observed clinical manifestations. This identification enabled a deeper exploration of the patient's condition through knowledge graphs. More evaluations are needed. The merged graph, post-identification of the genetic variant, offered a visual representation of the relationships among symptoms and genetic interactions, facilitating a more integrated understanding. This experiment highlights the pivotal role of genetic testing in advancing diagnostic accuracy and unraveling the complexities of clinical presentations.
Duration and Technical Details: In this experiment, the generation of knowledge graphs for each diagnostic report and the subsequent creation of the merged graph took approximately 16 seconds.

5.5.3 Computational Time and Space Usage

Generating knowledge graphs for each patient in our extensive dataset is computationally intensive. Our method speeds up graph creation for individual patients, but larger datasets still require more time. The time varies with report complexity and the number of entities and relationships processed. We also used the BERN2 API for Named Entity Recognition (NER), limited to 300 requests per 100 seconds per user. To comply with this limit, we added systematic 3-second pauses in our processing pipeline, extending the overall time for our

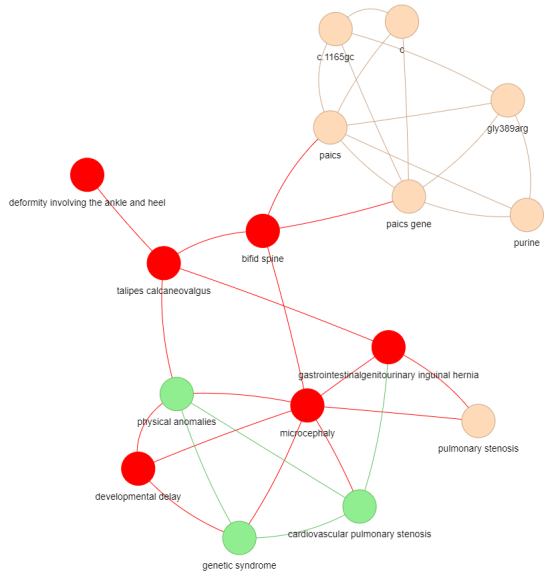


Fig. 5.7 Knowledge graph representing the merging of Medical Reports 1 and 2 for Experiment 2.

Table 5.2 Computational times.

Experiment	**Time Required**
Single patient (between 2-6 reports)	20-50 seconds
Entire dataset (59,051 patients)	12-14 days

59,051-patient dataset. The specific timings are summarized in Table 5.2.

Beyond computational time, storage space is a key factor for the knowledge graphs generated in our experiments. The amount of space each individual graph and their merged versions occupy varies with the number of entities, relationships, and attributes, influenced by the diagnostic reports' complexity and detail. Table 5.3 details the space usage, presenting average, minimum, and maximum figures per patient for single and merged graphs. Overall, individual graphs across all patients and reports require 6.39 GB, and the merged graphs collectively need 3.9 GB of storage space.

5.6 Enhancing Semantic Precision in Relation Extraction

To address the initial system design's notable limitation of lacking semantic labeling in entity relationships, we have refined our relation extraction process by incorporating an advanced feature. This enhancement involves integrating SemRep alongside Bio_ClinicalBERT, signif-

Table 5.3 Space usage.

Graph Type	**Memory average**	**Memory - range**	**Memory (all patients)**
Single Knowledge Graph	24 KB	[6 KB - 42 KB]	6.39 GB
Merged Knowledge Graph	56 KB	[8 KB - 110 KB]	3.9 GB

icantly enriching the semantic depth of the relationships captured in our knowledge graphs and providing a more nuanced understanding of the connections between medical entities.

Bio_ClinicalBERT lays the groundwork for our relation extraction, identifying potential connections between entities within clinical narratives. Despite its effectiveness in contextual understanding within the biomedical domain, it sometimes falls short in explicitly inferring the semantic nature of these relationships. This is where SemRep [198] comes into play . It complements Bio_ClinicalBERT's capabilities by providing precise semantic roles to these relationships, especially in instances where Bio_ClinicalBERT identifies a connection but does not categorize it. This dual approach ensures that each relationship in our knowledge graphs is not only identified but also semantically labeled, capturing the complex interactions depicted in clinical narratives (see at Figure). The practical impact of integrating SemRep is profound. It transforms our knowledge graphs from mere visual representations of entity connections into detailed maps that illustrate the intricate dynamics of a patient's medical history. With the ability to discern and label semantic predications, healthcare professionals are equipped with a more powerful tool for interpreting clinical data, which can significantly influence clinical outcomes and decision-making processes. The integration of SemRep marks a pivotal advancement toward achieving semantically rich and informative knowledge graphs.

5.6.1 Illustrative Case Study: Leveraging Semantic Labels for Enhanced Medical Insight

In this subsection, we explore the transformative potential of incorporating semantic labels into the relation extraction process within our enhanced knowledge graph generation system. Through a specific example, we demonstrate how this capability significantly aids in maintaining diagnostic continuity across multiple medical reports for the same patient.
Example Case: Number of associated medical reports: 2.

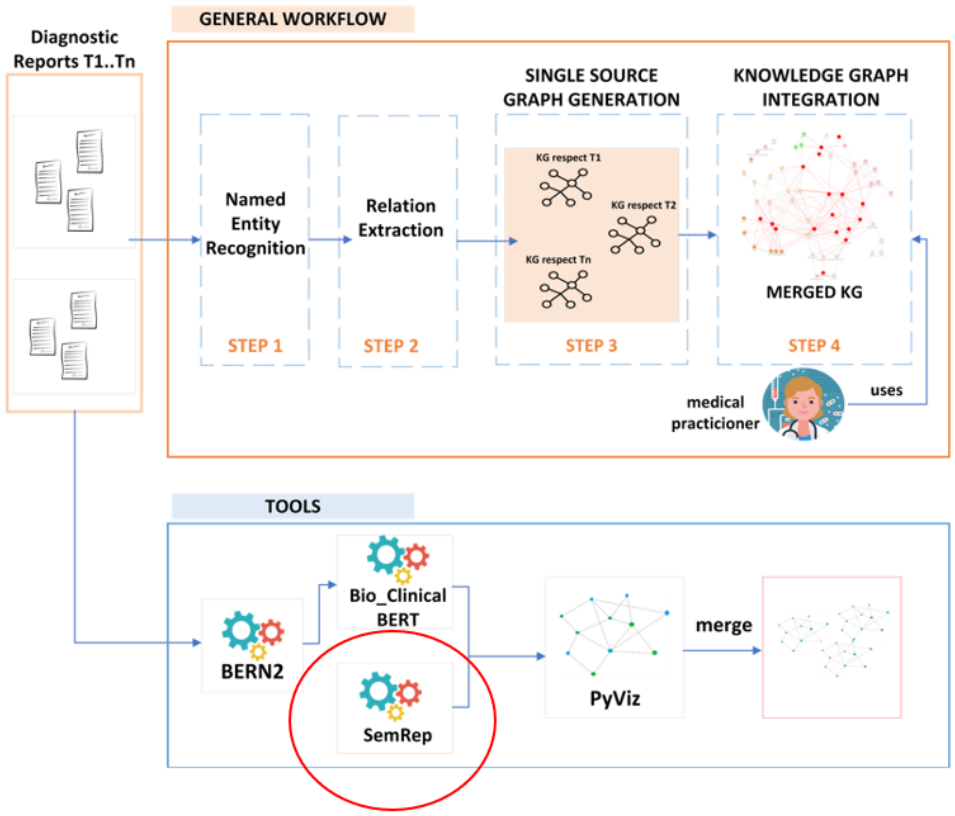


Fig. 5.8 SemRep Integration, in red.

Consider a patient with two associated medical reports, highlighting a scenario where crucial genetic testing information could be overlooked due to a lack of visibility across separate consultations.

Medical Report 1: Age at last evaluation: ___ year and ___ months. Anthropometric Data: Neurodevelopmental: Moderate developmental delay noted, Stereotypic movements observed, Microcephaly identified with OFC below the 3rd percentile. Clinical Findings: Cardiovascular: Pulmonary stenosis diagnosed. Musculoskeletal: Hidden bifid spine detected, Talipes calcaneovalgus present. Gastrointestinal/Genitourinary: Inguinal hernia diagnosed. Genetic Considerations: Genetic Testing Results: Gene: PAICS (NM_001079525.1), Mutation Identified: c.1165G>C ; p.Gly389Arg. The identified variant in the PAICS gene may potentially explain some of the clinical findings observed in this patient. The PAICS gene encodes a bifunctional enzyme involved in de novo purine biosynthesis. Mutations in this gene could potentially affect cellular proliferation and neurological function, although the exact clinical significance of the identified variant (c.1165G>C ; p.Gly389Arg) needs further evaluation.

Medical Report 2: Age at last evaluation: ___ year and ___ months. Anthropometric Data: At Birth: Weight: ___ grams, Length: 46.5 cm, Occipitofrontal circumference (OFC): 32 cm. At ___ Year: OFC: 43 cm (P<3), indicative of Microcephaly. Clinical Findings: Neurodevelopmental: Moderate developmental delay noted, Stereotypic movements observed, Microcephaly identified with OFC below the 3rd percentile. Cardiovascular: Pulmonary stenosis diagnosed. Musculoskeletal: Hidden bifid spine detected, Talipes calcaneovalgus (a deformity involving the ankle and heel) present. Gastrointestinal/Genitourinary: Inguinal hernia diagnosed. Genetic Considerations: The combination of microcephaly, developmental delay, and other physical anomalies may suggest a possible genetic syndrome. Genetic testing, including chromosomal microarray and/or whole exome sequencing, may be indicated to identify any underlying genetic etiologies.

In this case, a patient's first medical report included a genetic test identifying a significant variant in the PAICS gene, crucial for explaining various clinical symptoms (Figure 5.9). However, this vital information was missing from a later report, risking the neglect of important genetic findings and the potential for redundant testing (Figure 5.10).

Our system's semantic labeling effectively highlights entities and their interrelations from different patient visits, such as the conduct and results of genetic tests. These details are accentuated in the knowledge graphs and maintained through the merging process, providing healthcare professionals with a complete view of a patient's medical history, including diagnostics and outcomes not mentioned in recent reports. Semantic labels in the knowledge graphs visually represent the patient's medical history, providing insights into the sequence of diagnostics and treatments. This is especially critical in complex cases where understanding the history of medical decisions is essential for patient care (Figure 5.11).

This example underscores the significance of semantic labeling in deepening knowledge graphs' utility for medical data integration. By integrating and visually emphasizing semantic relations, the system ensures that all historical medical data contributes to a comprehensive patient health overview. This method enhances the retrieval of medical information and highlights the importance of each data point in creating a holistic patient profile, showcasing the system's essential role in improving healthcare delivery.

5.7 Discussion

Addressing the complexity of healthcare information, our system autonomously creates and combines knowledge graphs from raw medical texts, navigating this crucial and challenging domain.

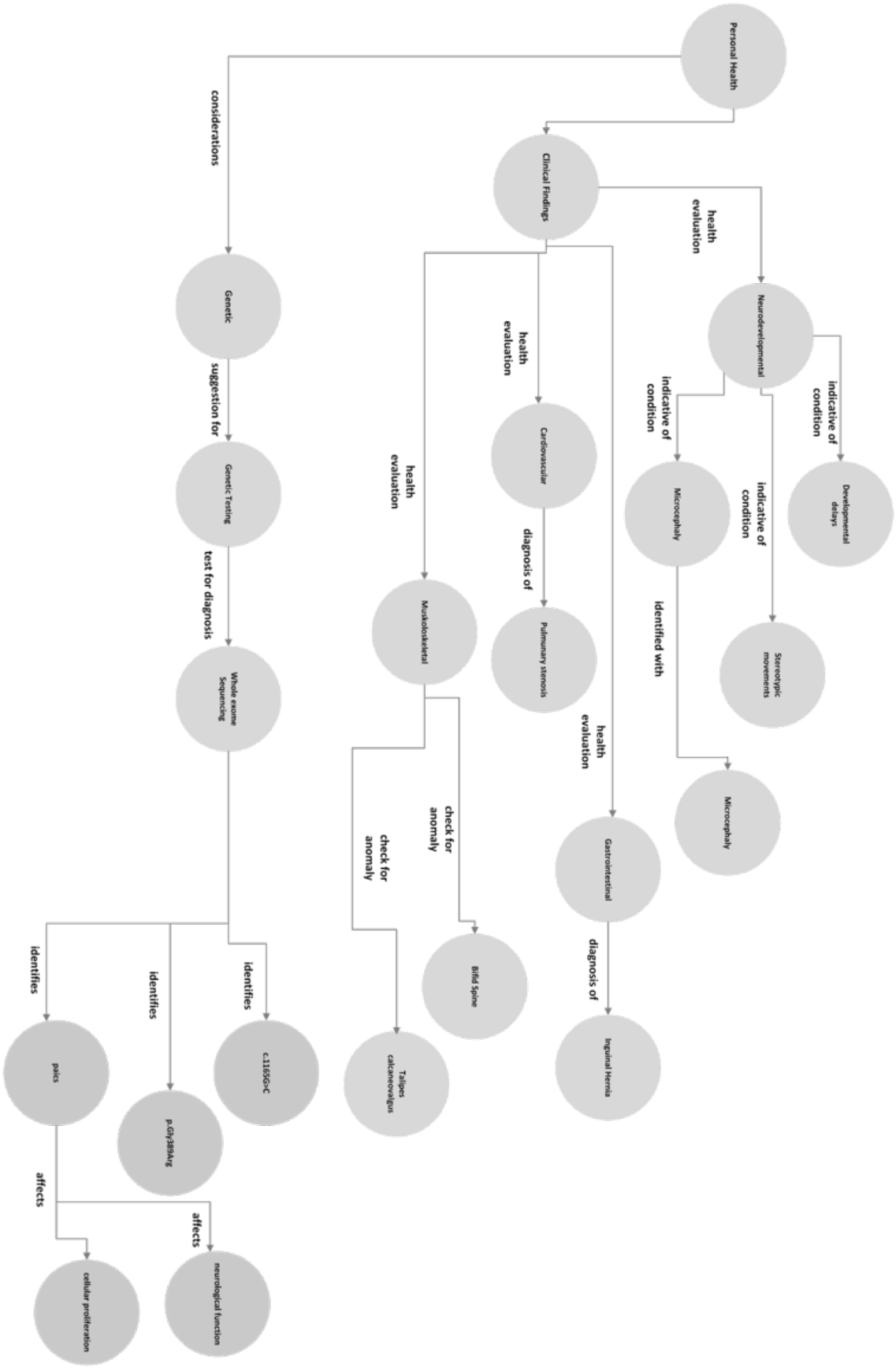


Fig. 5.9 Used tools, separated by phases.

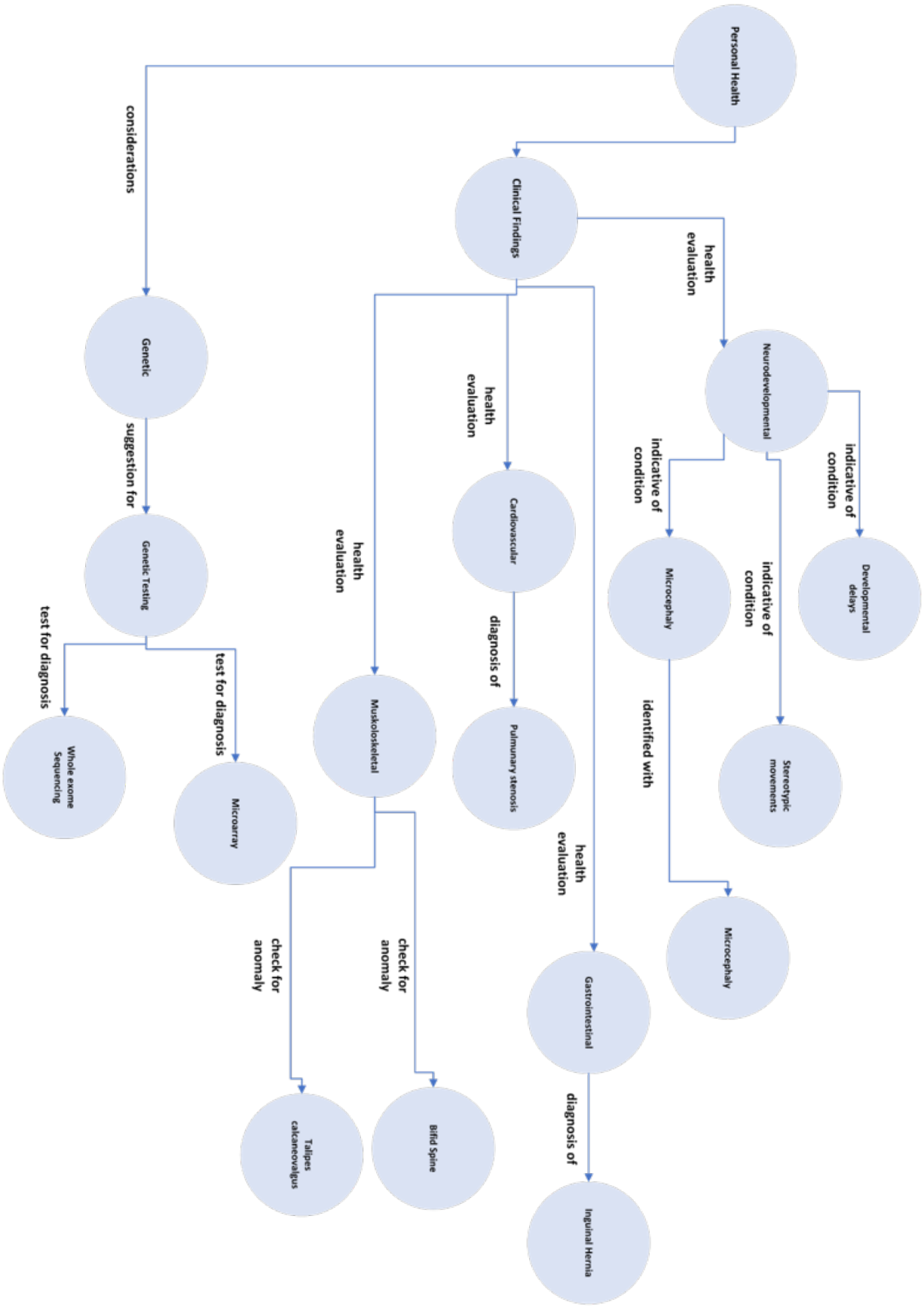


Fig. 5.10 Used tools, separated by phases.

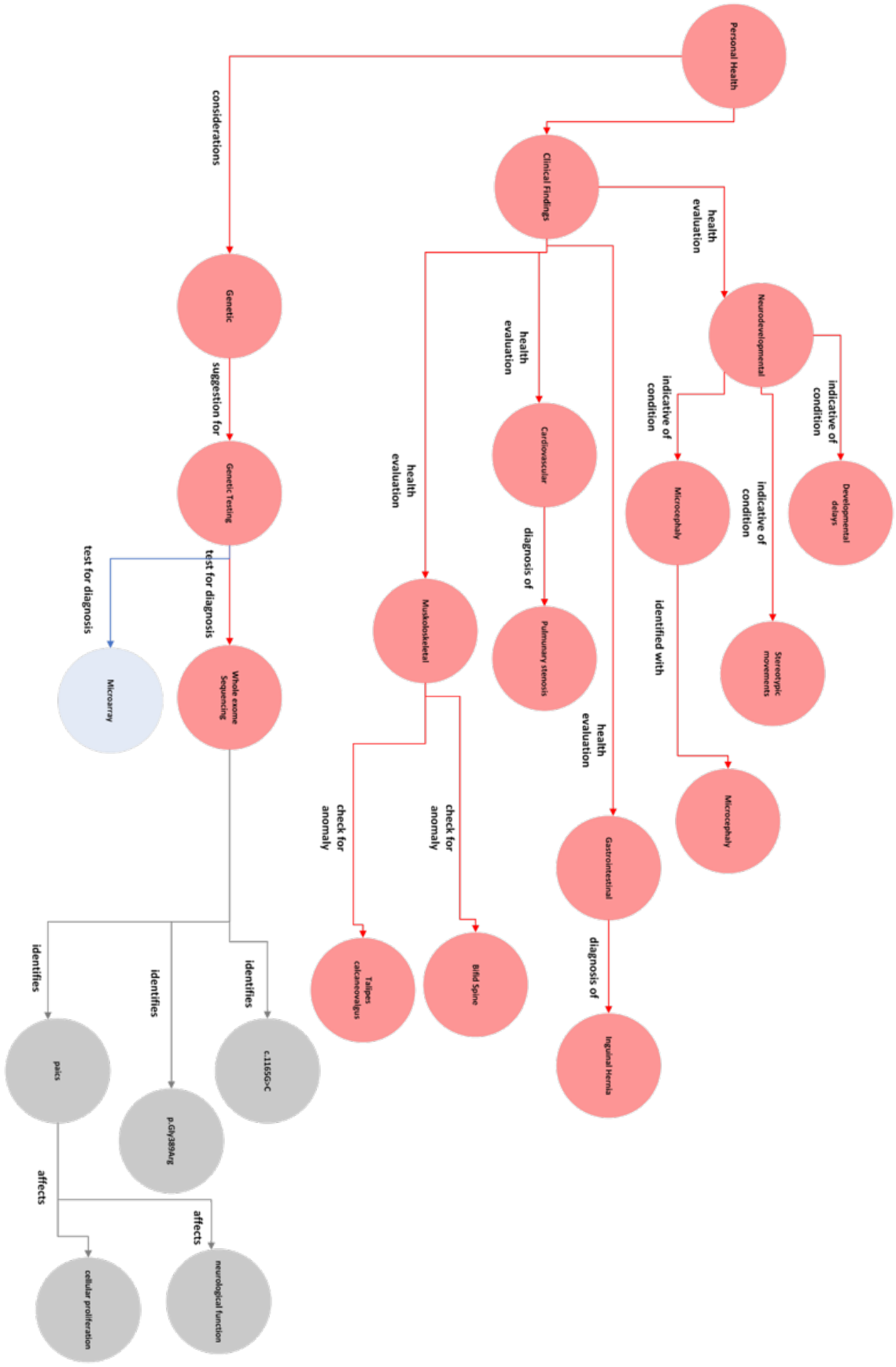


Fig. 5.11 Used tools, separated by phases.

Given the enormous variety of medical and biological entities present in healthcare, it was practical for us to narrow our primary attention to a small number of biomedical entities. This emphasis was seen in the studies, which showed the system's skill at locating, extracting, and connecting these chosen elements to create knowledge graphs that depict a clear and insightful narrative of a patient's medical journey. Focusing on a particular group of entities at this point allowed for deeper and more accurate knowledge as well as opened the door for methodical extension and inclusion of a wider variety of entities in the system's subsequent iterations.

The experiments demonstrated the system's capability to accurately and coherently navigate medical texts, generating individual and merged knowledge graphs that highlight key entities and recurring illnesses, essential for understanding a patient's medical history and refining therapeutic strategies. The visualization tool emerged as a vital asset, offering medical professionals an intelligible visual narrative of a patient's medical journey, enhancing understanding and diagnostic ability. In our study, although 59,051 knowledge graphs were generated, only a small subset of these was reported back to clinicians for evaluation. Specifically, knowledge graphs for 15 patients were provided to healthcare professionals. This limited selection was due to the complexity of reviewing the graphs and the significant effort required from clinicians to examine them thoroughly. For these 15 patients, we also conducted a validation process using their complete medical histories, ensuring that the knowledge graphs accurately reflected all relevant medical information. Furthermore, we double-checked the accuracy of the links and relationships generated within the knowledge graphs by cross-referencing them with existing literature. This meticulous approach aimed to confirm the correctness and reliability of the information presented, enhancing the graphs' utility in clinical practice.

The feedback from clinicians on these validated cases was insightful, demonstrating the potential benefits of knowledge graphs for providing a comprehensive overview of patient histories, particularly in complex cases. However, clinicians also pointed out that the process of analyzing and interpreting the graphs can be time-consuming, indicating a need for further refinement to make the tool more efficient and user-friendly in a clinical setting.

Going forward, we plan to optimize the system based on this initial feedback to improve its usability and reduce the workload for clinicians. Future efforts will focus on improving the clarity and relevance of the information in knowledge graphs, with the aim of making them a more practical and effective tool to support clinical decision making.

5.7.1 Bridging Health Gaps: Societal Benefits of Comprehensive Medical Views

Both patients and professionals frequently struggle with the complexity of managing and comprehending medical information in the wide world of healthcare. The high volume of medical records can be intimidating for practitioners, making it difficult to extract valuable insights from dispersed data. Moreover, during brief appointments, some patients may find it difficult to remember and describe every medical exam, symptom, or medicine they have ever experienced. For instance, it could be particularly difficult for older people or people who are naturally reticent to retell every aspect of their medical history. They could unintentionally leave out anything that is important for their present care, such as past diagnoses, treatments, or symptoms. Our system addresses these challenges by integrating multiple diagnostic reports into a unified visual representation, even if it focuses on a restricted set of biomedical entities. This ensures that every patient, irrespective of their background or communicative abilities, benefits from a comprehensive record that encapsulates their entire health journey. Such a holistic view not only empowers clinicians with a complete understanding of a patient's health but also alleviates the need for patients to remember every detail.

5.7.2 Cost Efficiency

The proposed approach enhances cost efficiency in healthcare by reducing the likelihood of redundant or unnecessary medical examinations. By providing a comprehensive and unified view of a patient's health history, the system enables healthcare providers to make more informed decisions, thereby minimizing the need for repeated tests or procedures.

Patients with complex medical histories, such as those suffering from rare diseases, stand to gain significantly from this approach. The integration of diverse diagnostic information into a single knowledge graph ensures that all relevant medical data is readily accessible, promoting timely and accurate care. Furthermore, in cases where diagnosing a condition proves challenging, physicians can utilize the system to compare the patient's knowledge graph with other merged graphs from similar cases. This capability facilitates a more rapid and accurate diagnosis by leveraging a broader spectrum of existing data, thus streamlining the diagnostic process and reducing associated costs.

5.7.3 Global Scalability

The system exhibited a high degree of scalability, successfully managing a large dataset encompassing 59,051 patients. This scalability is crucial for real-world applications where

the volume of medical data is continuously growing. Also, the system is designed to be modular (please see at the code repository), meaning individual components (like entity extraction or relation prediction) can be updated or replaced without disrupting the entire workflow. This allows the system to easily adapt to new technological advancements.

5.7.4 Limitations and Threats to Validity

Despite the results obtained in this study, there are certain limitations to our approach that should be taken into consideration. Here, we discuss these limitations and their potential impact on the interpretation of our findings.

Input Accuracy: One of the foundational premises of our system is the reliance on accurate and relevant input. It's necessary that users (namely, doctors) provide diagnostic texts pertaining to the same patient. The system is designed to compare and integrate these texts, and any discrepancy in the input, such as including texts from unrelated patients, can lead to misleading results.

Natural Language Dependency: Our current implementation is tailored for the English language. This is largely because we utilize pre-trained tools, which are predominantly trained on English medical and biomedical terminologies. While the system demonstrates efficacy with English texts, its applicability could be limited in regions with different native languages. Expanding the system's capability to cater to diverse languages remains a future target.

Negation Recognition: Our current system, while effective in extracting and linking entities from medical texts, has limitations when it comes to accurately interpreting negated symptoms or medical conditions. For example, phrases like "high glucose: no" or "no signs of infection" are common in clinical narratives, indicating the absence of a condition rather than its presence. Currently, our system does not have a robust mechanism to differentiate between positive and negative mentions of symptoms. This limitation can lead to incorrect assumptions or misinterpretations within the generated knowledge graphs, where a symptom or condition might be inaccurately recorded as present due to the system's inability to recognize negation. To address this issue, we acknowledge the need for negation detection algorithms, that can accurately identify and process negated statements. Incorporating such methods would allow the system to more precisely capture the context of medical data, ensuring that the knowledge graphs reflect a more accurate representation of a patient's medical history.

Bypassing Coreference Resolution: The anonymization of our dataset, which obfuscates names, dates, and other specifics to protect patient confidentiality, impeded accurate coreference resolution, a step that typically ensures pronouns and abbreviations are correctly

mapped to their entities in NLP pipelines. Due to the resulting unreliable outputs from coreference models, we opted to proceed directly to NER without performing coreference resolution.

5.7.5 Future Directions

As our system continues to evolve, one of our primary goals is to ensure its accessibility and usability worldwide. To achieve this, we are actively considering the incorporation of multilingual models, which would enable the system to process and understand medical reports in various languages, catering to a global audience.

Moreover, a promising frontier for our system lies in leveraging the intricate patterns within the knowledge graphs. Our vision is to utilize dedicated pattern recognition techniques that systematically analyze these graphs, pinpointing recurring sequences or clusters of entities and relations that could be indicative of specific medical conditions or trajectories [199]. This kind of structured pattern mining could be instrumental in identifying and potentially predicting diagnostic paths, thereby aiding healthcare professionals in making informed, proactive decisions about patient care [177].

For instance, by analyzing a vast number of knowledge graphs and tracing back the diagnostic journeys of patients with a particular condition, we might discern that certain entity relationships frequently precede the diagnosis of that condition [200]. With this insight, our system could recommend preemptive diagnostic exams when it detects the onset of similar patterns in a patient's medical data. Important and concert benefits can be provided by the proposed system in case of rare diseases: if specific patterns of such diseases are discovered, using graph comparison will speed up the right diagnosis [199].

For the sake of completeness, this work has been accepted to the AIHealth 2024 Conference, held between March 19-24, 2024 (https://www.thinkmind.org/index.php?view=article&articleid=aihealth_2024_1_70_80033). All the methodologies along with the technical implementation can be found at https://github.com/anbianchi/knowledge_frombio/

Conclusions

This thesis represents a comprehensive effort to develop a framework capable of integrating genomic data with clinical information, thereby advancing the field of precision medicine. It builds upon the premise that personalized healthcare is contingent upon our ability to accurately analyze and apply genetic insights in a clinical setting. Throughout this work, we have endeavored to construct and refine bioinformatics tools and methodologies that not only address the inherent challenges of genomic data but also harness its full potential to inform patient care. The development of this framework was initiated with acknowledgment of the existing limitations within bioinformatics, specifically the reproducibility of data and the retrieval of datasets. Through targeted research and development, we introduced a self-evaluating system for reproducibility, enhancing the reliability of bioinformatics experiments. This contribution is poised to improve the standardization across studies, facilitating a more trustworthy and replicable scientific inquiry.

Initially, we dedicated our efforts to RNA-Seq, emphasizing the criticality of selecting the appropriate tools for accurate gene expression analysis. This phase of the research highlighted the significant impact tool choice has on the outcomes and the reproducibility challenges inherent in bioinformatics workflows. By meticulously evaluating RNA-Seq processes, we exposed the limitations in current practices, underscoring the necessity for enhanced precision and reliability in genomic studies. Transitioning from RNA-Seq, the thesis then ventured into the realm of Variant Calling, a method pivotal for detecting genetic variations that could influence health and disease outcomes. Here, we didn't just analyze genetic variations in isolation; instead, we developed an integrated approach that combines the nuanced insights from Variant Calling with the broader contextual data of patient health. This holistic strategy aimed to provide a more comprehensive understanding of how genetic alterations affect individual health profiles, thereby enriching the potential for tailored therapeutic interventions.

The cornerstone of this thesis was the synergistic integration of bioinformatics findings with clinical narratives, culminating in the creation of knowledge graphs. These graphs serve

not just as a novel tool but as a bridge connecting theoretical bioinformatics research with practical clinical applications, offering a more nuanced understanding of patient health.

The harmonization of genomic data with clinical narratives has emerged as a keystone in the realization of precision medicine. The framework developed herein successfully tackles this complex integration, setting a precedent for the field. Through the novel application of entity-relation models and the generation of knowledge graphs, we have provided a way to visualize and analyze the confluence of genetic and clinical data, empowering healthcare professionals to make more informed decisions. The findings and developments presented in this thesis pave the way for a new era in healthcare. The enhanced bioinformatics framework herein is a testimony to the transformative impact of precise genetic analysis in clinical contexts. As we look to the future, this work opens multiple avenues for further research, especially in the realm of advancing computational techniques, exploring the ethical dimensions of data usage, and extending the framework to encompass the ever-growing datasets in genomics. The path ahead is rife with opportunities for continued innovation and implementation. The next steps will involve rigorous field testing of the framework, validation against diverse clinical scenarios, and adaptation to the ever-evolving landscape of genomic research. The immediate focus has been on honing a multi-source diagnostic framework, which integrates bioinformatics data and clinical narratives to construct a comprehensive health profile for individual patients. This approach has laid the groundwork for a nuanced understanding of patient-specific health dynamics, crucial for advancing personalized medicine.

As we look to the future, the ambition is to scale this framework to encompass comparative analyses across entire populations. This expansion aims to unearth patterns and correlations that may not be discernible at the individual level but emerge when viewing the broader population landscape. By comparing genomic and clinical data across diverse groups, the objective is to identify commonalities and differences in health outcomes, genetic predispositions, and treatment responses. It is with optimism that we foresee the integration of our contributions into standard healthcare practices, marking a significant milestone in the journey towards truly personalized medicine.

References

[1] Lloyd Minor. *Discovering precision health: predict, prevent, and cure to advance health and well-being*. John Wiley & Sons, 2020.

[2] Margaret Morash, Hannah Mitchell, Himisha Beltran, Olivier Elemento, and Jyotishman Pathak. The role of next-generation sequencing in precision medicine: a review of outcomes in oncology. *Journal of personalized medicine*, 8(3):30, 2018.

[3] M Pereira, F Malta, M Freire, and P Couto. Application of next-generation sequencing in the era of precision medicine. *Applications of RNA-Seq and Omics Strategies: From Microorganisms to Human Health. London, England, UK: InTechOpen*, pages 293–318, 2017.

[4] S Momčilović, C Cantacessi, V Arsić-Arsenijević, D Otranto, and S Tasić-Otašević. Rapid diagnosis of parasitic diseases: current scenario and future needs. *Clinical Microbiology and Infection*, 25(3):290–309, 2019.

[5] Nardeep Naithani, Sharmila Sinha, Pratibha Misra, Biju Vasudevan, and Rajesh Sahu. Precision medicine: Concept and tools. *medical journal armed forces india*, 77(3):249–257, 2021.

[6] Clayton M Christensen, Jerome H Grossman, and Jason Hwang. The innovator's prescription. 2009.

[7] Alla Katsnelson. Momentum grows to make'personalized'medicine more'precise'. *Nature medicine*, 19(3):249, 2013.

[8] Sing Yu Moorcraft, David Gonzalez, and Brian A Walker. Understanding next generation sequencing in oncology: A guide for oncologists. *Critical reviews in oncology/hematology*, 96(3):463–474, 2015.

[9] James Love-Koh, Alison Peel, Juan Carlos Rejon-Parrilla, Kate Ennis, Rosemary Lovett, Andrea Manca, Anastasia Chalkidou, Hannah Wood, and Matthew Taylor. The future of precision medicine: potential impacts for health technology assessment. *Pharmacoeconomics*, 36:1439–1451, 2018.

[10] Zheng-Guo Wang, Liang Zhang, and Wen-Jun Zhao. Definition and application of precision medicine. *Chinese Journal of Traumatology*, 19(05):249–250, 2016.

[11] Kyle B Brothers and Mark A Rothstein. Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Personalized medicine*, 12(1):43–51, 2015.

[12] Wallace J Hopp, Jun Li, and Guihua Wang. Big data and the precision medicine revolution. *Production and Operations Management*, 27(9):1647–1664, 2018.

[13] TEDDY Study Group. The environmental determinants of diabetes in the young (teddy) study: study design. *Pediatric diabetes*, 8(5):286–298, 2007.

[14] Jeffrey L Mahon, Jay M Sosenko, Lisa Rafkin-Mervis, Heidi Krause-Steinrauf, John M Lachin, Clinton Thompson, Polly J Bingley, Ezio Bonifacio, Jerry P Palmer, George S Eisenbarth, et al. The trialnet natural history study of the development of type 1 diabetes: objectives, design, and initial results. *Pediatric diabetes*, 10(2):97–104, 2009.

[15] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

[16] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine*, pages 795–840. Springer, 2021.

[17] Geoffrey S Ginsburg and Kathryn A Phillips. Precision medicine: from science to value. *Health affairs*, 37(5):694–701, 2018.

[18] Rory Collins. What makes uk biobank special? *The Lancet*, 379(9822):1173–1174, 2012.

[19] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[20] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

[21] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[22] Daniel Richard Leff and Guang-Zhong Yang. Big data for precision medicine. *Engineering*, 1(3):277–279, 2015.

[23] R Rajani, DS Berman, and A Rozanski. Social networks—are they good for your health? the era of facebook and twitter. *QJM: An International Journal of Medicine*, 104(9):819–820, 2011.

[24] David J Duffy. Problems, challenges and promises: perspectives on precision medicine. *Briefings in bioinformatics*, 17(3):494–504, 2016.

[25] Tim Hulsen, Saumya S Jamuar, Alan R Moody, Jason H Karnes, Orsolya Varga, Stine Hedensted, Roberto Spreafico, David A Hafler, and Eoin F McKinney. From big data to precision medicine. *Frontiers in medicine*, 6:34, 2019.

[26] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302, 2018.

[27] Mandana Hasanzad, Negar Sarhangi, Sima Ehsani Chimeh, Nayereh Ayati, Monireh Afzali, Fatemeh Khatami, Shekoufeh Nikfar, and Hamid Reza Aghaei Meybodi. Precision medicine journey through omics approach. *Journal of Diabetes & Metabolic Disorders*, 21(1):881–888, 2022.

[28] Ali Khodadadian, Somaye Darzi, Saeed Haghi-Daredeh, Farzaneh Sadat Eshaghi, Emad Babakhanzadeh, Seyed Hamidreza Mirabutalebi, and Majid Nazari. Genomics and transcriptomics: the powerful technologies in precision medicine. *International Journal of General Medicine*, pages 627–640, 2020.

[29] AJ Marian. Sequencing your genome: what does it mean? *Methodist DeBakey cardiovascular journal*, 10(1):3, 2014.

[30] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.

[31] Daniel C Koboldt, Karyn Meltz Steinberg, David E Larson, Richard K Wilson, and Elaine R Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, 2013.

[32] Victor E Velculescu, Stephen L Madden, Lin Zhang, Alex E Lash, Jian Yu, Carlo Rago, Anita Lal, Clarence J Wang, Gary A Beaudry, Kristin M Ciriello, et al. Analysis of human transcriptomes. *Nature genetics*, 23(4):387–388, 1999.

[33] Brielle Miles and Prasanna Tadi. Genetics, somatic mutation. 2020.

[34] H Richard Johnston, Bronya JB Keats, and Stephanie L Sherman. Population genetics. In *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics*, pages 359–373. Elsevier, 2019.

[35] Jennifer K Sehn. Insertions and deletions (indels). In *Clinical genomics*, pages 129–150. Elsevier, 2015.

[36] EE Eichler. Copy number variation and human disease. *Nat Educ*, 1(3):1, 2008.

[37] Daniel C Koboldt. Best practices for variant calling in clinical sequencing. *Genome Medicine*, 12(1):1–13, 2020.

[38] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20(1):1–14, 2019.

[39] Bilal Abu-Salih, Muhammad Al-Qurishi, Mohammed Alweshah, Mohammad Al-Smadi, Reem Alfayez, and Heba Saadeh. Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. *Journal of Big Data*, 10(1):81, 2023.

[40] Ling Tian, Xue Zhou, Yan-Ping Wu, Wang-Tao Zhou, Jin-Hao Zhang, and Tian-Shu Zhang. Knowledge graph and knowledge reasoning: A systematic review. *Journal of Electronic Science and Technology*, 20(2):100159, 2022.

[41] Piero Andrea Bonatti, Stefan Decker, Axel Polleres, and Valentina Presutti. Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In *Dagstuhl reports*, volume 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[42] Simon Schramm, Christoph Wehner, and Ute Schmid. Comprehensible artificial intelligence on knowledge graphs: A survey. *Journal of Web Semantics*, 79:100806, 2023.

[43] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, pages 1–32, 2023.

[44] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[45] Peter N Robinson and Stefan Mundlos. The human phenotype ontology. *Clinical genetics*, 77(6):525–534, 2010.

[46] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[47] Banan Jamil Awrahman, Chia Aziz Fatah, and Mzhda Yasin Hamaamin. A review of the role and challenges of big data in healthcare informatics and analytics. *Computational intelligence and neuroscience*, 2022, 2022.

[48] Joël Simoneau, Simon Dumontier, Ryan Gosselin, and Michelle S Scott. Current RNA-seq methodology reporting limits reproducibility. *Briefings in Bioinformatics*, 22(1):140–145, 12 2019.

[49] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Szcześniak, Daniel Gaffney, Laura Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17, 01 2016.

[50] Dexiang Gao, Jihye Kim, Hyunmin Kim, Tzu Phang, Heather Selby, and Aik Choon Tan. A survey of statistical software for analysing rna-seq data. *Human genomics*, 5:56–60, 10 2010.

[51] Wei Vivian Li and Jingyi Jessica Li. Modeling and analysis of rna-seq data: a review from a statistical perspective, 2018.

[52] Shanrong Zhao, Baohong Zhang, Ying Zhang, William Gordon, Sarah Du, Theresa Paradis, Michael Vincent, and David von Schack. Bioinformatics for rna-seq data analysis. In Ibrokhim Y. Abdurakhmonov, editor, *Bioinformatics*, chapter 6. IntechOpen, Rijeka, 2016.

[53] Fei Ji and Ruslan I. Sadreyev. Rna-seq: Basic bioinformatics analysis. *Current Protocols in Molecular Biology*, 124(1):e68, 2018.

[54] Alicia Oshlack, Mark Robinson, and Matthew Young. From rna-seq reads to differential expression results. *Genome biology*, 11:220, 12 2010.

[55] Geng Chen, Chun Wang, and Tieliu Shi. Overview of available methods for diverse rna-seq data analyses. *Science China. Life sciences*, 54:1121–8, 12 2011.

[56] Pallavi Gaur and Anoop Chaturvedi. *A Survey of Bioinformatics-Based Tools in RNA-Sequencing (RNA-Seq) Data Analysis*. 04 2017.

[57] Yixing Han, Shouguo Guo, Kathrin Muegge, Wei Zhang, and Bing Zhou. Advanced applications of rna sequencing and challenges. *Bioinformatics and Biology Insights*, 9:29, 11 2015.

[58] Hussain Chowdhury, Dhruba K Bhattacharyya, and Jugal Kalita. Differential expression analysis of rna-seq reads: Overview, taxonomy and tools. *IEEE/ACM transactions on computational biology and bioinformatics*, PP, 10 2018.

[59] Li Tong, Po-Yen Wu, John Phan, Hamid Hassazadeh, Weida Tong, and May Wang. Impact of rna-seq data analysis algorithms on gene expression estimation and downstream prediction. *Scientific Reports*, 10, 12 2020.

[60] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, 17, 06 2008.

[61] Rute Pereira, Jorge Oliveira, and Mário Sousa. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of clinical medicine*, 9(1):132, 2020.

[62] Giacomo Baruzzo, Katharina Hayer, Eun Kim, Barbara Camillo, Garret FitzGerald, and Gregory Grant. Simulation-based comprehensive benchmarking of rna-seq aligners. *Nature Methods*, 14, 12 2016.

[63] Pär Engström, Tamara Steijger, Botond Sipos, Gregory Grant, Andre Kahles, Tyler Alioto, Jonas Behr, Paul Bertone, Regina Bohnert, Davide Campagna, Carrie Davis, Alexander Dobin, Thomas Gingeras, Nick Goldman, Roderic Guigó, Jennifer Harrow, Tim Hubbard, Geraldine Jean, Peter Kosarev, and Georg Zeller. Systematic evaluation of spliced alignment programs for rna-seq data. *Nature methods*, 10, 11 2013.

[64] Alba Rodriguez-Meira, Gemma Buck, Sally-Ann Clark, Benjamin J. Povinelli, Veronica Alcolea, Eleni Louka, Simon McGowan, Angela Hamblin, Nikolaos Sousos, Nikolaos Barkas, Alice Giustacchini, Bethan Psaila, Sten Eirik W. Jacobsen, Supat Thongjuea, and Adam J. Mead. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular Cell*, 73(6):1292–1305.e8, March 2019.

[65] Bojan Losic, Amanda J. Craig, Carlos Villacorta-Martin, Sebastiao N. Martins-Filho, Nicholas Akers, Xintong Chen, Mehmet E. Ahsen, Johann von Felden, Ismail Labgaa, Delia D'Avola, Kimaada Allette, Sergio A. Lira, Glaucia C. Furtado, Teresa Garcia-Lezana, Paula Restrepo, Ashley Stueck, Stephen C. Ward, Maria I. Fiel, Spiros P. Hiotis, Ganesh Gunasekaran, Daniela Sia, Eric E. Schadt, Robert Sebra, Myron Schwartz, Josep M. Llovet, Swan Thung, Gustavo Stolovitzky, and Augusto Villanueva. Intratumoral heterogeneity and clonal evolution in liver cancer. *Nature Communications*, 11(1):291, January 2020. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer;Gastroenterology;Immunology;Oncology Subject_term_id: cancer;gastroenterology;immunology;oncology.

[66] N. Alcala, N. Leblay, A. a. G. Gabriel, L. Mangiante, D. Hervas, T. Giffon, A. S. Sertier, A. Ferrari, J. Derks, A. Ghantous, T. M. Delhomme, A. Chabrier, C. Cuenin, B. Abedi-Ardekani, A. Boland, R. Olaso, V. Meyer, J. Altmuller, F. Le Calvez-Kelm, G. Durand, C. Voegele, S. Boyault, L. Moonen, N. Lemaitre, P. Lorimier, A. C. Toffart, A. Soltermann, J. H. Clement, J. Saenger, J. K. Field, M. Brevet, C. Blanc-Fournier, F. Galateau-Salle, N. Le Stang, P. A. Russell, G. Wright, G. Sozzi, U. Pastorino, S. Lacomme, J. M. Vignaud, V. Hofman, P. Hofman, O. T. Brustugun, M. Lund-Iversen, V. Thomas de Montpreville, L. A. Muscarella, P. Graziano, H. Popper, J. Stojsic, J. F. Deleuze, Z. Herceg, A. Viari, P. Nuernberg, G. Pelosi, A. M. C. Dingemans, M. Milione, L. Roz, L. Brcic, M. Volante, M. G. Papotti, C. Caux, J. Sandoval, H. Hernandez-Vargas, E. Brambilla, E. J. M. Speel, N. Girard, S. Lantuejoul, J. D. McKay, M. Foll, and L. Fernandez-Cuesta. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nature Communications*, 10(1):3407, August 2019.

[67] Jakub Hynst, Karla Plevova, Lenka Radova, Vojtech Bystry, Karol Pal, and Sarka Pospisilova. Bioinformatic pipelines for whole transcriptome sequencing data exploitation in leukemia patients with complex structural variants. *PeerJ*, 7:e7071, 2019.

[68] Markus W. Löffler, Christopher Mohr, Leon Bichmann, Lena Katharina Freudenmann, Mathias Walzer, Christopher M. Schroeder, Nico Trautwein, Franz J. Hilke, Raphael S. Zinser, Lena Mühlenbruch, Daniel J. Kowalewski, Heiko Schuster, Marc Sturm, Jakob Matthes, Olaf Riess, Stefan Czemmel, Sven Nahnsen, Ingmar Königsrainer, Karolin Thiel, Silvio Nadalin, Stefan Beckert, Hans Bösmüller, Falko Fend, Ana Velic, Boris Maček, Sebastian P. Haen, Luigi Buonaguro, Oliver Kohlbacher, Stefan Stevanović, Alfred Königsrainer, and Hans-Georg Rammensee. Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Medicine*, 11:28, April 2019.

[69] Alexey Stupnikov, Paul G. O'Reilly, Caitriona E. McInerney, Aideen C. Roddy, Philip D. Dunne, Alan Gilmore, Hayley P. Ellis, Tom Flannery, Estelle Healy, Stuart A. McIntosh, Kienan Savage, Kathreena M. Kurian, Frank Emmert-Streib, Kevin M. Prise, Manuel Salto-Tellez, and Darragh G. McArt. Impact of Variable RNA-Sequencing Depth on Gene Expression Signatures and Target Compound Robustness: Case Study Examining Brain Tumor (Glioma) Disease Progression. *JCO precision oncology*, 2, September 2018.

[70] Toshima Z. Parris, Elisabeth Werner Rönnerman, Hanna Engqvist, Jana Biermann, Katarina Truvé, Szilárd Nemes, Eva Forssell-Aronsson, Giovanni Solinas, Anikó Kovács, Per Karlsson, and Khalil Helou. Genome-wide multi-omics profiling of the 8p11-p12 amplicon in breast carcinoma. *Oncotarget*, 9(35):24140–24154, May 2018.

[71] Melissa Quintero, Douglas Adamoski, Larissa Menezes Dos Reis, Carolline Fernanda Rodrigues Ascenção, Krishina Ratna Sousa de Oliveira, Kaliandra de Almeida Gonçalves, Marília Meira Dias, Marcelo Falsarella Carazzolle, and Sandra Martha Gomes Dias. Guanylate-binding protein-1 is a potential new therapeutic target for triple-negative breast cancer. *BMC cancer*, 17(1):727, November 2017.

[72] Tristan Gallenne, Kenneth N. Ross, Nils L. Visser, null Salony, Christophe J. Desmet, Ben S. Wittner, Lodewyk F. A. Wessels, Sridhar Ramaswamy, and Daniel S. Peeper. Systematic functional perturbations uncover a prognostic genetic network driving human breast cancer. *Oncotarget*, 8(13):20572–20587, March 2017.

[73] Tiziana de Cristofaro, Tina Di Palma, Amata Amy Soriano, Antonella Monticelli, Ornella Affinito, Sergio Cocozza, and Mariastella Zannini. Candidate genes and pathways downstream of PAX8 involved in ovarian high-grade serous carcinoma. *Oncotarget*, 7(27):41929–41947, July 2016.

[74] Lily Boo, Wan Yong Ho, Norlaily Mohd Ali, Swee Keong Yeap, Huynh Ky, Kok Gan Chan, Wai Fong Yin, Dilan Amila Satharasinghe, Woan Charn Liew, Sheau Wei Tan, Han Kiat Ong, and Soon Keng Cheong. MiRNA Transcriptome Profiling of Spheroid-Enriched Cells with Cancer Stem Cell Properties in Human Breast MCF-7 Cell Line. *International Journal of Biological Sciences*, 12(4):427–445, 2016.

[75] Jong-Young Lee, Seok Joong Yun, Pildu Jeong, Xuan-Mei Piao, Ye-Hwan Kim, Jihye Kim, Sathiyamoorthy Subramaniyam, Young Joon Byun, Ho Won Kang, Sung Phil Seo, Jayoung Kim, Jung Min Kim, Eun Sang Yoo, Isaac Y. Kim, Sung-Kwon Moon, Yung Hyun Choi, and Wun-Jae Kim. Identification of differentially expressed miRNAs and miRNA-targeted genes in bladder cancer. *Oncotarget*, 9(45):27656–27666, June 2018.

[76] Jialiang Yang, Jacob Hagen, Kalyani V. Guntur, Kimaada Allette, Sarah Schuyler, Jyoti Ranjan, Francesca Petralia, Stephane Gesta, Robert Sebra, Milind Mahajan, Bin Zhang, Jun Zhu, Sander Houten, Andrew Kasarskis, Vivek K. Vishnudas, Viatcheslav R. Akmaev, Rangaprasad Sarangarajan, Niven R. Narain, Eric E. Schadt, Carmen A. Argmann, and Zhidong Tu. A next generation sequencing based approach to identify extracellular vesicle mediated mRNA transfers between cells. *BMC genomics*, 18(1):987, December 2017.

[77] Alina Mieczkowska, Adriana Schumacher, Natalia Filipowicz, Anna Wardowska, Maciej Zieliński, Piotr Madanecki, Ewa Nowicka, Paulina Langa, Milena Deptuła, Jacek Zieliński, Karolina Kondej, Alicja Renkielska, Patrick G. Buckley, David K. Crossman, Michael R. Crowley, Artur Czupryn, Piotr Mucha, Paweł Sachadyn, Łukasz Janus, Piotr Skowron, Sylwia Rodziewicz-Motowidło, Mirosława Cichorek, Michał Pikuła, and Arkadiusz Piotrowski. Immunophenotyping and transcriptional profiling of in vitro cultured human adipose tissue derived stem cells. *Scientific Reports*, 8(1):11339, July 2018.

[78] Hye-Yeong Jo, Youngsun Lee, Hongryul Ahn, Hyeong-Jun Han, Ara Kwon, Bo-Young Kim, Hye-Yeong Ha, Sang Cheol Kim, Jung-Hyun Kim, Yong-Ou Kim, Sun Kim, Soo Kyung Koo, and Mi-Hyun Park. Functional in vivo and in vitro effects of 20q11.21 genetic aberrations on hPSC differentiation. *Scientific Reports*, 10(1):18582, October 2020.

[79] Zeda Zhang, Chuanli Zhou, Xiaoling Li, Spencer D. Barnes, Su Deng, Elizabeth Hoover, Chi-Chao Chen, Young Sun Lee, Yanxiao Zhang, Choushi Wang, Lauren A. Metang, Chao Wu, Carla Rodriguez Tirado, Nickolas A. Johnson, John Wongvipat, Kristina Navrazhina, Zhen Cao, Danielle Choi, Chun-Hao Huang, Eliot Linton, Xiaoping Chen, Yupu Liang, Christopher E. Mason, Elisa de Stanchina, Wassim Abida, Amaia Lujambio, Sheng Li, Scott W. Lowe, Joshua T. Mendell, Venkat S. Malladi, Charles L. Sawyers, and Ping Mu. Loss of CHD1 Promotes Heterogeneous Mechanisms of Resistance to AR-Targeted Therapy via Chromatin Dysregulation. *Cancer Cell*, 37(4):584–598.e11, April 2020.

[80] Dimitrios S. Kanakoglou, Theodora-Dafni Michalettou, Christina Vasileiou, Evangelos Gioukakis, Dorothea Maneta, Konstantinos V. Kyriakidis, Alexandros G. Georgakilas, and Ioannis Michalopoulos. Effects of High-Dose Ionizing Radiation in Human Gene Expression: A Meta-Analysis. *International Journal of Molecular Sciences*, 21(6):E1938, March 2020.

[81] Rajinder Gupta, Yannick Schrooders, Duncan Hauser, Marcel van Herwijnen, Wiebke Albrecht, Bas Ter Braak, Tim Brecklinghaus, Jose V. Castell, Leroy Elenschneider, Sylvia Escher, Patrick Guye, Jan G. Hengstler, Ahmed Ghallab, Tanja Hansen, Marcel Leist, Richard Maclennan, Wolfgang Moritz, Laia Tolosa, Tine Tricot, Catherine Verfaillie, Paul Walker, Bob van de Water, Jos Kleinjans, and Florian Caiment. Comparing in vitro human liver models to in vivo human liver using RNA-Seq. *Archives of Toxicology*, 95(2):573–589, February 2021.

[82] Georgios I. Laliotis, Evangelia Chavdoula, Maria D. Paraskevopoulou, Abdul Kaba, Alessandro La Ferlita, Satishkumar Singh, Vollter Anastas, Keith A. Nair, Arturo Orlacchio, Vasiliki Taraslia, Ioannis Vlachos, Marina Capece, Artemis Hatzigeorgiou, Dario Palmieri, Christos Tsatsanis, Salvatore Alaimo, Lalit Sehgal, David P. Carbone, Vincenzo Coppola, and Philip N. Tsichlis. AKT3-mediated IWS1 phosphorylation promotes the proliferation of EGFR-mutant lung adenocarcinomas through cell cycle-regulated U2AF2 RNA splicing. *Nature Communications*, 12(1):4624, July 2021.

[83] Tiira Johansson, Dawit A. Yohannes, Satu Koskela, Jukka Partanen, and Päivi Saavalainen. HLA RNA Sequencing With Unique Molecular Identifiers Reveals High Allele-Specific Variability in mRNA Expression. *Frontiers in Immunology*, 12:629059, 2021.

[84] Shwu-Yuan Wu, Chien-Fei Lee, Hsien-Tsung Lai, Cheng-Tai Yu, Ji-Eun Lee, Hao Zuo, Sophia Y. Tsai, Ming-Jer Tsai, Kai Ge, Yihong Wan, and Cheng-Ming Chiang. Opposing Functions of BRD4 Isoforms in Breast Cancer. *Molecular cell*, 78(6):1114–1132.e10, June 2020.

[85] Ariane L. Moore, Aashil A. Batavia, Jack Kuipers, Jochen Singer, Elodie Burcklen, Peter Schraml, Christian Beisel, Holger Moch, and Niko Beerenwinkel. Spatial

Distribution of Private Gene Mutations in Clear Cell Renal Cell Carcinoma. *Cancers*, 13(9):2163, April 2021.

[86] Ehsan Ghorani, James L. Reading, Jake Y. Henry, Marc Robert de Massy, Rachel Rosenthal, Virginia Turati, Kroopa Joshi, Andrew J. S. Furness, Assma Ben Aissa, Sunil Kumar Saini, Sofie Ramskov, Andrew Georgiou, Mariana Werner Sunderland, Yien Ning Sophia Wong, Maria Vila De Mucha, William Day, Felipe Galvez-Cancino, Pablo D. Becker, Imran Uddin, Mazlina Ismail, Tahel Ronel, Annemarie Woolston, Mariam Jamal-Hanjani, Selvaraju Veeriah, Nicolai J. Birkbak, Gareth A. Wilson, Kevin Litchfield, Lucia Conde, José Afonso Guerra-Assunção, Kevin Blighe, Dhruva Biswas, Roberto Salgado, Tom Lund, Maise Al Bakir, David A. Moore, Crispin T. Hiley, Sherene Loi, Yuxin Sun, Yinyin Yuan, Khalid AbdulJabbar, Samra Turajilic, Javier Herrero, Tariq Enver, Sine R. Hadrup, Allan Hackshaw, Karl S. Peggs, Nicholas McGranahan, Benny Chain, Charles Swanton, and Sergio A. Quezada. The T cell differentiation landscape is shaped by tumour mutations in lung cancer. *Nature Cancer*, 1(5):546–561, May 2020.

[87] Zhifei Luo and Peggy J. Farnham. Genome-wide analysis of HOXC4 and HOXC6 regulated genes and binding sites in prostate cancer cells. *PloS One*, 15(2):e0228590, 2020.

[88] Diana Sousa, Rune Matthiesen, Raquel T. Lima, and M. Helena Vasconcelos. Deep Sequencing Analysis Reveals Distinctive Non-Coding RNAs When Comparing Tumor Multidrug-Resistant Cells and Extracellular Vesicles with Drug-Sensitive Counterparts. *Cancers*, 12(1):E200, January 2020.

[89] Nastaran Masoudi-Khoram, Parviz Abdolmaleki, Nazanin Hosseinkhan, Alireza Nikoofar, Seyed Javad Mowla, Hamideh Monfared, and Gustavo Baldassarre. Differential miRNAs expression pattern of irradiated breast cancer cell lines is correlated with radiation sensitivity. *Scientific Reports*, 10(1):9054, June 2020.

[90] Michal Sima, Kristyna Vrbova, Tana Zavodna, Katerina Honkova, Irena Chvojkova, Antonin Ambroz, Jiri Klema, Andrea Rossnerova, Katerina Polakova, Tomas Malina, Jan Belza, Jan Topinka, and Pavel Rossner. The Differential Effect of Carbon Dots on Gene Expression and DNA Methylation of Human Embryonic Lung Fibroblasts as a Function of Surface Charge and Dose. *International Journal of Molecular Sciences*, 21(13):4763, July 2020.

[91] Louisa Nelson, Anthony Tighe, Anya Golder, Samantha Littler, Bjorn Bakker, Daniela Moralli, Syed Murtuza Baker, Ian J. Donaldson, Diana C. J. Spierings, René Wardenaar, Bethanie Neale, George J. Burghel, Brett Winter-Roach, Richard Edmondson, Andrew R. Clamp, Gordon C. Jayson, Sudha Desai, Catherine M. Green, Andy Hayes, Floris Foijer, Robert D. Morgan, and Stephen S. Taylor. A living biobank of ovarian cancer ex vivo models reveals profound mitotic heterogeneity. *Nature Communications*, 11(1):822, February 2020.

[92] Fei Ji and Ruslan I. Sadreyev. RNA-seq: Basic Bioinformatics Analysis. *Current Protocols in Molecular Biology*, 124(1):e68, October 2018.

[93] Isaac D. Raplee, Alexei V. Evsikov, and Caralina Marín de Evsikova. Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression

Quantification Tools for Clinical Breast Cancer Research. *Journal of Personalized Medicine*, 9(2):E18, April 2019.

[94] Manuel Garber, Manfred Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using rna-seq. *Nature methods*, 8:469–77, 06 2011.

[95] He Li, Moez Dawood, Michael M. Khayat, Jesse R. Farek, Shalini N. Jhangiani, Ziad M. Khan, Tadahiro Mitani, Zeynep Coban-Akdemir, James R. Lupski, Eric Venner, Jennifer E. Posey, Aniko Sabo, and Richard A. Gibbs. Exome variant discrepancies due to reference-genome differences. *The American Journal of Human Genetics*, 108(7):1239–1250, 2021.

[96] Andrea Bianchi, Giordano d'Aloisio, Antinisca Di Marco, and Francesca Marzi. Data reproducibility tree, February 2023.

[97] Pelin Icer Baykal, Paweł Piotr Łabaj, Florian Markowetz, Lynn M Schriml, Daniel J Stekhoven, Serghei Mangul, and Niko Beerenwinkel. Genomic reproducibility in the bioinformatics era. *Genome Biology*, 25, 2024.

[98] Giacomo Baruzzo, Katharina Hayer, Eun Kim, Barbara Camillo, Garret FitzGerald, and Gregory Grant. Simulation-based comprehensive benchmarking of rna-seq aligners. *Nature Methods*, 14, 2016.

[99] Simone Claudiani, Clinton C Mason, Dragana Milojkovic, Andrea Bianchi, Cristina Pellegrini, Antinisca Di Marco, Carme R Fiol, Mark Robinson, Kanagaraju Ponnusamy, Katya Mokretar, et al. Carfilzomib enhances the suppressive effect of ruxolitinib in myelofibrosis. *Cancers*, 13(19):4863, 2021.

[100] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven Salzberg. Tophat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14, 2013.

[101] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915, 2019.

[102] Alexander Dobin, Carrie Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics (Oxford, England)*, 29, 2012.

[103] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[104] S Akila Parvathy Dharshini, Y-H Taguchi, and M Michael Gromiha. Identifying suitable tools for variant detection and differential gene expression using rna-seq data. *Genomics*, 112(3):2166—2172, 2020.

[105] Isaac D. Raplee, Alexei V. Evsikov, and Caralina Marín de Evsikova. Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *Journal of Personalized Medicine*, 9(2), 2019.

[106] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Szcześniak, Daniel Gaffney, Laura Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17, 2016.

[107] Pallavi Gaur and Anoop Chaturvedi. *A Survey of Bioinformatics-Based Tools in RNA-Sequencing (RNA-Seq) Data Analysis*, pages 223–248. 2017.

[108] Paul McGettigan. Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, 17, 2013.

[109] Andrews. Fastqc: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

[110] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[111] Daehwan Kim, Ben Langmead, and Steven Salzberg. Hisat: A fast spliced aligner with low memory requirements. *Nature methods*, 12, 2015.

[112] Yang Liao, Gordon Smyth, and Wei Shi. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30, 2013.

[113] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. 2014.

[114] Andrea Bianchi. Deseq2 script on myelofibrosis dataset (hisat2 and star2 pipelines), March 2023.

[115] Ewan Birney, T Daniel Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, James Cuff, Val Curwen, Tim Cutts, et al. An overview of ensembl. *Genome research*, 14(5):925–928, 2004.

[116] Andrea Bianchi. Hisat2 - star2 pipelines on myelofibrosis, May 2023.

[117] Gabrielle Deschamps-Francoeur, Joël Simoneau, and Michelle S Scott. Handling multi-mapped reads in rna-seq. *Computational and structural biotechnology journal*, 18:1569–1576, 2020.

[118] Evgeniy S Balakirev and Francisco J Ayala. Pseudogenes: are they "junk" or functional dna? *Annual review of genetics*, 37(1):123–151, 2003.

[119] Andrea Bianchi, Antinisca Di Marco, and Cristina Pellegrini. Comparing hisat and star-based pipelines for rna-seq data analysis: a real experience. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 218–224. IEEE, 2023.

[120] Michael F Berger and Elaine R Mardis. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.*, 15(6):353–365, June 2018.

[121] Armand Valsesia, Aurélien Macé, Sébastien Jacquemont, Jacques S Beckmann, and Zoltán Kutalik. The growing importance of cnvs: new insights for detection and clinical interpretation. *Frontiers in genetics*, 4:92, 2013.

[122] Rolph Pfundt, Marisol Del Rosario, Lisenka ELM Vissers, Michael P Kwint, Irene M Janssen, Nicole De Leeuw, Helger G Yntema, Marcel R Nelen, Dorien Lugtenberg, Erik-Jan Kamsteeg, et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine*, 19(6):667–675, 2017.

[123] Qingwei Qi, Yulin Jiang, Xiya Zhou, Hua Meng, Na Hao, Jiazhen Chang, Junjie Bai, Chunli Wang, Mingming Wang, Jiangshan Guo, et al. Simultaneous detection of cnvs and snvs improves the diagnostic yield of fetuses with ultrasound anomalies and normal karyotypes. *Genes*, 11(12):1397, 2020.

[124] Bo Yuan, Lei Wang, Pengfei Liu, Chad Shaw, Hongzheng Dai, Lance Cooper, Wenmiao Zhu, Stephanie A Anderson, Linyan Meng, Xia Wang, et al. Cnvs cause autosomal recessive genetic diseases with or without involvement of snv/indels. *Genetics in Medicine*, 22(10):1633–1641, 2020.

[125] Andre E. Minoche, Ben Lundie, Greg B. Peters, Thomas Ohnesorg, Mark Pinese, David M. Thomas, Andreas Zankl, Tony Roscioli, Nicole Schonrock, Sarah Kummerfeld, Leslie Burnett, Marcel E. Dinger, and Mark J. Cowley. Clinsv: clinical grade structural and copy number variant detection from whole genome sequencing data. *Genome Medicine*, 13(1):32, Feb 2021.

[126] James Reid, Sandra Kachhia, Paul Dougall, John Shovelton, Duarte Molha, Christina Taylor, Jagath Kasturiarachchi, Jolyon Holdstock, Venu Pullabhatla, Laura Parkes, et al. A next generation sequencing solution to detect copy number variants, single nucleotide variants and loss of heterozygosity in intellectual disability and developmental delay samples. *Mosaic*, 5(5/5):100, 2020.

[127] Guney Bademci, Joseph Foster, Nejat Mahdieh, Mortaza Bonyadi, Duygu Duman, F Basak Cengiz, Ibis Menendez, Oscar Diaz-Horta, Atefeh Shirkavand, Sirous Zeinali, et al. Comprehensive analysis via exome sequencing uncovers genetic etiology in autosomal recessive nonsyndromic deafness in a large multiethnic cohort. *Genetics in Medicine*, 18(4):364–371, 2016.

[128] Breast Cancer Association Consortium, Leila Dorling, Sara Carvalho, Jamie Allen, Anna González-Neira, Craig Luccarini, Cecilia Wahlström, Karen A Pooley, Michael T Parsons, Cristina Fortuno, Qin Wang, Manjeet K Bolla, Joe Dennis, and Renske et al. Keeman. Breast cancer risk genes - association analysis in more than 113,000 women. *N. Engl. J. Med.*, 384(5):428–439, February 2021.

[129] Michael G Keeney, Fergus J Couch, Daniel W Visscher, and Noralane M Lindor. Non-brca familial breast cancer: review of reported pathology and molecular findings. *Pathology*, 49(4):363–370, 2017.

[130] Francisco Javier Gracia-Aznarez, Victoria Fernandez, Guillermo Pita, Paolo Peterlongo, Orlando Dominguez, Miguel de la Hoya, Mercedes Duran, Ana Osorio, Leticia

Moreno, Anna Gonzalez-Neira, et al. Whole exome sequencing suggests much of non-brca1/brca2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PloS one*, 8(2):e55681, 2013.

[131] Stavros Glentis, Alexandros C Dimopoulos, Konstantinos Rouskas, George Ntritsos, Evangelos Evangelou, Steven A Narod, Anne-Marie Mes-Masson, William D Foulkes, Barbara Rivera, Patricia N Tonin, et al. Exome sequencing in brca1-and brca2-negative greek families identifies mdm1 and nbeal1 as candidate risk genes for hereditary breast cancer. *Frontiers in genetics*, 10:1005, 2019.

[132] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 08 2012.

[133] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[134] Vincent Plagnol, James Curtis, Michael Epstein, Kin Y Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W Wood, Sophie Hambleton, Siobhan O Burns, Adrian J Thrasher, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754, 2012.

[135] Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arne Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. cn. mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research*, 40(9):e69–e69, 2012.

[136] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[137] Véronique Geoffroy, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, and Jean Muller. Annotsv: an integrated tool for structural variations annotation. *Bioinformatics*, 34(20):3572–3574, 2018.

[138] Yan Guo, Jirong Long, Jing He, Chung-I Li, Qiuyin Cai, Xiao-Ou Shu, Wei Zheng, and Chun Li. Exome sequencing generates high quality data in non-target regions. *BMC genomics*, 13(1):1–10, 2012.

[139] Sanja Mehandziska, Aleksandra Stajkovska, Margarita Stavrevska, Kristina Jakovleva, Marija Janevska, Rodney Rosalia, Ivan Kungulovski, Zan Mitrev, and Goran Kungulovski. Workflow for the implementation of precision genomics in healthcare. *Frontiers in Genetics*, 11:619, 2020.

[140] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. 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):405–423, 2015.

[141] Melissa Rotunno, Rolando Barajas, Mindy Clyne, Elise Hoover, Naoko I Simonds, Tram Kim Lam, Leah E Mechanic, Alisa M Goldstein, and Elizabeth M Gillanders. A systematic literature review of whole exome and genome sequencing population studies of genetic susceptibility to cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 29(8):1519–1534, 2020.

[142] Ramakrishnan Rajagopalan, Jill R Murrell, Minjie Luo, and Laura K Conlin. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome medicine*, 12(1):1–11, 2020.

[143] Veronika Gordeeva, Elena Sharova, Konstantin Babalyan, Rinat Sultanov, Vadim M Govorun, and Georgij Arapidi. Benchmarking germline cnv calling tools from exome sequencing data. *Scientific Reports*, 11(1):14416, 2021.

[144] Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, 3(1):1–26, 2016.

[145] Rajini R Haraksingh, Alexej Abyzov, and Alexander Eckehart Urban. Comprehensive performance comparison of high-resolution array platforms for genome-wide copy number variation (cnv) analysis in humans. *BMC genomics*, 18(1):1–14, 2017.

[146] Adam C English, Vipin K Menon, Richard A Gibbs, Ginger A Metcalf, and Fritz J Sedlazeck. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biology*, 23(1):271, 2022.

[147] Berk Mandiracioglu, Furkan Ozden, Gun Kaynar, Mehmet Alper Yilmaz, Can Alkan, and A Ercument Cicek. Ecole: Learning to call copy number variants on whole exome sequencing data. *Nature Communications*, 15(1):132, 2024.

[148] Zhi-Yong Liu, Jian Yi, and Feng-En Liu. The molecular mechanism of breast cancer cell apoptosis induction by absent in melanoma (aim2). *International journal of clinical and experimental medicine*, 8(9):14750, 2015.

[149] Aayushi Srivastava, Sara Giangiobbe, Diamanto Skopelitou, Beiping Miao, Nagarajan Paramasivam, Chiara Diquigiovanni, Elena Bonora, Kari Hemminki, Asta Försti, and Obul Reddy Bandapalli. Whole genome sequencing prioritizes chek2, ewsr1, and tiam1 as possible predisposition genes for familial non-medullary thyroid cancer. *Frontiers in endocrinology*, 12:600682, 2021.

[150] Fatima Aloraifi, Trudi McDevitt, Rui Martiniano, Jonah McGreevy, Russell McLaughlin, Chris M Egan, Nuala Cody, Marie Meany, Elaine Kenny, Andrew J Green, et al. Detection of novel germline mutations for breast cancer in non-brca 1/2 families. *FEBS J.*, 282(17):3424–3437, 2015.

[151] Reihaneh Zarrizi, Martin R Higgs, Karolin Voßgröne, Maria Rossing, Birgitte Bertelsen, Muthiah Bose, Arne Nedergaard Kousholt, Heike Rösner, Bent Ejlertsen, Grant S Stewart, et al. Germline rbbp8 variants associated with early-onset breast cancer compromise replication fork stability. *The Journal of clinical investigation*, 130(8):4069–4080, 2020.

[152] Lamis Yehia, Farshad Niazi, Ying Ni, Joanne Ngeow, Madhav Sankunny, Zhigang Liu, Wei Wei, Jessica L Mester, Ruth A Keri, Bin Zhang, et al. Germline heterozygous variants in sec23b are associated with cowden syndrome and enriched in apparently sporadic thyroid cancer. *The American Journal of Human Genetics*, 97(5):661–676, 2015.

[153] A Arteche-López, A Ávila-Fernández, R Romero, R Riveiro-Álvarez, MA López-Martínez, A Giménez-Pardo, C Vélez-Monsalve, J Gallego-Merlo, I García-Vara, Berta Almoguera, et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 ngs variants in 825 clinical exomes. *Scientific reports*, 11(1):5697, 2021.

[154] Adam Kiezun, Kiran Garimella, Ron Do, Nathan O. Stitziel, Benjamin M. Neale, Paul J. McLaren, Namrata Gupta, Pamela Sklar, Patrick F. Sullivan, Jennifer L. Moran, Christina M. Hultman, Paul Lichtenstein, Patrik Magnusson, Thomas Lehner, Yin Yao Shugart, Alkes L. Price, Paul I. W. de Bakker, Shaun M. Purcell, and Shamil R. Sunyaev. Exome sequencing and the genetic basis of complex traits. *Nature Genetics*, 44(6):623–630, Jun 2012.

[155] Aouatef Riahi, Hoda Radmanesh, Peter Schürmann, Natalia Bogdanova, Robert Geffers, Rym Meddeb, Maher Kharrat, and Thilo Dörk. Exome sequencing and case–control analyses identify rcc1 as a candidate breast cancer susceptibility gene. *International Journal of Cancer*, 142(12):2512–2517, 2018.

[156] Susanna Koivuluoma, Anna Tervasmäki, Saila Kauppila, Robert Winqvist, Timo Kumpula, Outi Kuismin, Jukka Moilanen, and Katri Pylkäs. Exome sequencing identifies a recurrent variant in serpina3 associating with hereditary susceptibility to breast cancer. *European J. of Cancer*, 143:46–51, 2021.

[157] Veronica Zelli, Chiara Compagnoni, Katia Cannita, Roberta Capelli, Carlo Capalbo, Mauro Di Vito Nolfi, Edoardo Alesse, Francesca Zazzeroni, and Alessandra Tessitore. Applications of next generation sequencing to the analysis of familial breast/ovarian cancer. *High-throughput*, 9(1):1, 2020.

[158] Lanling Zhao, Han Liu, Xiguo Yuan, Kun Gao, and Junbo Duan. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC bioinformatics*, 21(1):1–10, 2020.

[159] José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro, and Bernat Gel. Benchmark of tools for cnv detection from ngs panel data in a genetic diagnostics context. *BioRxiv*, page 850958, 2019.

[160] Jamie M Ellingford, Christopher Campbell, Stephanie Barton, Sanjeev Bhaskar, Saurabh Gupta, Rachel L Taylor, Panagiotis I Sergouniotis, Bradley Horn, Janine A Lamb, Michel Michaelides, et al. Validation of copy number variation analysis for next-generation sequencing diagnostics. *European J. of Human Genetics*, 25(6):719–724, 2017.

[161] Iikki Donner, Riku Katainen, Tomas Tanskanen, Eevi Kaasinen, Mervi Aavikko, Kristian Ovaska, Miia Artama, Eero Pukkala, and Lauri A Aaltonen. Candidate

susceptibility variants for esophageal squamous cell carcinoma. *Genes, Chromosomes and Cancer*, 56(6):453–459, 2017.

[162] Charoula Achilla, Theodosios Papavramidis, Lefteris Angelis, and Anthoula Chatzikyriakidou. The implication of x-linked genetic polymorphisms in susceptibility and sexual dimorphism of cancer. *Anticancer Research*, 42(5):2261–2276, 2022.

[163] Scott Newman, Karen E Hermetz, Brooke Weckselblatt, and M Katharine Rudd. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am. J. Hum. Genet.*, 96(2):208–220, February 2015.

[164] Omar Abdelwahab, François Belzile, and Davoud Torkamaneh. Performance analysis of conventional and ai-based variant callers using short and long reads. *BMC Bioinformatics*, 24(1):472, Dec 2023.

[165] Daniel C. Koboldt. Best practices for variant calling in clinical sequencing. *Genome Medicine*, 12(1):91, Oct 2020.

[166] Lanling Zhao, Han Liu, Xiguo Yuan, Kun Gao, and Junbo Duan. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics*, 21(1):97, Mar 2020.

[167] Veronika Gordeeva, Elena Sharova, Konstantin Babalyan, Rinat Sultanov, Vadim M. Govorun, and Georgij Arapidi. Benchmarking germline cnv calling tools from exome sequencing data. *Scientific Reports*, 11(1):14416, Jul 2021.

[168] Andrea Bianchi, Veronica Zelli, Andrea D'Angelo, Alessandro Di Matteo, Giulia Scoccia, Katia Cannita, Antigone S Dimas, Stavros Glentis, Francesca Zazzeroni, Edoardo Alesse, et al. A method to comprehensively identify germline snvs, indels and cnvs from whole exome sequencing data of brca1/2 negative breast cancer patients. *NAR Genomics and Bioinformatics*, 6(2):lqae033, 2024.

[169] Fabienne C Bourgeois, Karen L Olson, and Kenneth D Mandl. Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Archives of internal medicine*, 170(22):1989–1995, 2010.

[170] Katherine Hempstead, Derek DeLia, Joel C Cantor, Tuan Nguyen, and Jeffrey Brenner. The fragmentation of hospital use among a cohort of high utilizers: implications for emerging care coordination strategies for patients with multiple chronic conditions. *Medical care*, pages S67–S74, 2014.

[171] Luis G Carvajal-Carmona. Challenges in the identification and use of rare disease-associated predisposition variants. *Current opinion in genetics & development*, 20(3):277–281, 2010.

[172] Wei-Qi Wei, Cynthia L Leibson, Jeanine E Ransom, Abel N Kho, Pedro J Caraballo, High Seng Chai, Barbara P Yawn, Jennifer A Pacheco, and Christopher G Chute. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association*, 19(2):219–224, 2012.

[173] Dong Dong, Roger Yat-Nork Chung, Rufina HW Chan, Shiwei Gong, and Richard Huan Xu. Why is misdiagnosis more likely among some people with rare diseases than others? insights from a population-based cross-sectional study in china. *Orphanet journal of rare diseases*, 15(1):1–12, 2020.

[174] Bastien Rance, Michelle Snyder, Janine Lewis, and Olivier Bodenreider. Leveraging terminological resources for mapping between rare disease information sources. *Studies in health technology and informatics*, 192:529, 2013.

[175] Sanju Tiwari, Fatima N Al-Aswadi, and Devottam Gaurav. Recent trends in knowledge graphs: theory and practice. *Soft Computing*, 25:8337–8355, 2021.

[176] Lino Murali, G Gopakumar, Daleesha M Viswanathan, and Prema Nedungadi. Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. *Journal of Biomedical Informatics*, page 104403, 2023.

[177] Katrin Hänsel, Sarah N Dudgeon, Kei-Hoi Cheung, Thomas JS Durant, and Wade L Schulz. From data to wisdom: Biomedical knowledge graphs for real-world data insights. *Journal of Medical Systems*, 47(1):65, 2023.

[178] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673, 2020.

[179] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, and Matthias Mann. Clinical knowledge graph integrates proteomics data into clinical decision-making. *bioRxiv*, pages 2020–05, 2020.

[180] Anderson Rossanez, Julio Cesar Dos Reis, Ricardo da Silva Torres, and Hélène de Ribaupierre. Kgen: a knowledge graph generator from biomedical scientific literature. *BMC medical informatics and decision making*, 20(4):1–24, 2020.

[181] Rugwedi Kulkarni and Yashodhara Haribhakta. Building the knowledge graph from medical conversational text data and its applications. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 1508–1513. IEEE, 2022.

[182] Maulik R Kamdar, Will Dowling, Michael Carroll, Cailey Fitzgerald, Sujit Pal, Steve Ross, Katie Scranton, Dru Henke, and Mevan Samarasinghe. A healthcare knowledge graph-based approach to enable focused clinical search. In *ISWC (Posters/Demos/Industry)*, 2021.

[183] Xiaowei Xu, Xuwen Wang, Meng Wu, Hetong Ma, Liu Shen, and Jiao Li. Development of an interactive medical knowledge graph based tool set. *Procedia Computer Science*, 221:578–584, 2023.

[184] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):5994, 2017.

[185] Tong Ruan, Yueqi Huang, Xuli Liu, Yuhang Xia, and Ju Gao. Qanalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. *BMC medical informatics and decision making*, 19:1–13, 2019.

[186] M Schmidli. Outcome of patients with acute coronary syndrome in hospitals of different sizes. a report from the amis plus registry. *Swiss medical weekly*, 140(2122):314–322, 2010.

[187] Tanya Hewitt, Samia Chreim, and Alan Forster. Incident reporting systems: a comparative study of two hospital divisions. *Archives of Public Health*, 74(1):1–19, 2016.

[188] Peipei Ping, Karol Watson, Jiawei Han, and Alex Bui. Individualized knowledge graph: a viable informatics path to precision medicine. *Circulation research*, 120(7):1078–1080, 2017.

[189] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, 7:73729–73740, 2019.

[190] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[191] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[192] Christine Aboseif and Paul Liu. Pelvic organ prolapse. 2020.

[193] Christl Reisenauer, Andreas Kirschniak, Ulrich Drews, and Diethelm Wallwiener. Anatomical conditions for pelvic floor reconstruction with polypropylene implant and its application for the treatment of vaginal prolapse. *European Journal of Obstetrics & Gynecology and reproductive Biology*, 131(2):214–225, 2007.

[194] Paul Abrams, Karl-Erik Andersson, Apostolos Apostolidis, Lori Birder, Donna Bliss, Linda Brubaker, Linda Cardozo, David Castro-Diaz, PR O'connell, Alan Cottenden, et al. 6th international consultation on incontinence. recommendations of the international scientific committee: evaluation and treatment of urinary incontinence, pelvic organ prolapse and faecal incontinence. *Neurourology and urodynamics*, 37(7):2271–2272, 2018.

[195] Nan Huang, Chang Xu, Liang Deng, Xue Li, Zhixuan Bian, Yue Zhang, Shuping Long, Yan Chen, Ni Zhen, Guohui Li, et al. Paics contributes to gastric carcinogenesis and participates in dna damage response by interacting with histone deacetylase 1/2. *Cell Death & Disease*, 11(7):507, 2020.

[196] Minjun Meng, Yanling Chen, Jianbo Jia, Lianghui Li, and Sumei Yang. Knockdown of paics inhibits malignant proliferation of human breast cancer cell lines. *Biological research*, 51, 2018.

[197] Anna Pelet, Vaclava Skopova, Ulrike Steuerwald, Veronika Baresova, Mohammed Zarhrate, Jean-Marc Plaza, Ales Hnizda, Matyas Krijt, Olga Souckova, Flemming Wibrand, et al. Paics deficiency, a new defect of de novo purine synthesis resulting in multiple congenital anomalies and fatal outcome. *Human Molecular Genetics*, 28(22):3805–3814, 2019.

[198] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. Broad-coverage biomedical relation extraction with semrep. *BMC bioinformatics*, 21:1–28, 2020.

[199] Xiaohui Tao, Thuan Pham, Ji Zhang, Jianming Yong, Wee Pheng Goh, Wenping Zhang, and Yi Cai. Mining health knowledge graph for health risk prediction. *World Wide Web*, 23:2341–2362, 2020.

[200] Huaqiong Wang, Xiaoyu Miao, and Pan Yang. Design and implementation of personal health record systems based on knowledge graph. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 133–136. IEEE, 2018.