Modern systems for large-scale genomics data processing in the cloud

Sergei Yakneen
EMBL
Meyerhofstrasowse 1,
69117 Heidelberg, Germany

09/11/2018

Contents

Chapter 1

Introduction

1.1 Context and Motivation

In the 17 years since the publication of the first draft human genome[84] the fields of genomics and molecular biology have undergone a major shift. The direction of this shift is towards an increasing adoption of computational approaches alongside experimental methods, bringing both of these fields of study into the realm of information science. This transition has been facilitated by two major factors - the advent of next generation sequencing[151], and the development of the Internet and cloud computing[17]. Next generation sequencing has been responsible for bringing down the cost of DNA sequencing to the point where it has become possible to sequence and study entire populations of individuals[66], while the Internet and cloud computing are democratising access to large-scale computational resources such that computation on big datasets, which was previously only accessible to large institutions, is becoming tractable to a growing group of researchers and citizen scientists.

The continued appetite for sequencing of larger and larger cohorts of individuals by the research community is driven by the desire to better understand the evolutionary history of the human species[79], to identify causes and mechanisms of action of rare genetic diseases that affect a very small proportion of the population[14], and to elucidate and potentially target the genetic component of more common diseases such as cancer[179], heart disease[15], or dementia[153] that place a heavy burden on our society. All of these factors together mean that the need for the generation and interpretation of genomic data is growing at an unprecedented scale.

Yet, the analysis of DNA sequencing data to study human genomes remains a largely unsolved problem. The protein coding sequence of the human genome, its *exome*, constitutes roughly 1% of the human DNA and successful studies have carried out exome-based analyses on cohorts at the scale of tens of thousands of individuals[92]. However, the other 99% of the human genome, it's non-coding regions, contain crucial information such as gene regulatory elements[25] that are essential to our full understanding of the mechanisms and processes that are underlying the human genetic landscape. Given current technologies, Whole Genome Sequencing

(WGS) is considerably more expensive and generates data-set sizes at the petabyte (PB) scale that are challenging for even the largest international consortia to tackle [161]. WGS studies at 100,000 participants scale that are planned for the coming years[41] will further increase data-set size and complexity by several orders of magnitude, a challenge that is presently unanswered by the current generation of bioinformatics infrastructures and algorithms.

A bigger and more distant challenge is the development of clinical sequencing and genomics which will truly bring whole-genome sequencing applications to population scale. Currently DNA sequencing has limited adoption within the clinical practice with applications limited to rare Mendelian disorders[91] and certain types of cancers[146] where a small set of genomic loci is interrogated via a gene panel[8] with a set of well-delineated disease sub-types based on these genetic markers. The use of whole-genome sequencing for clinical applications is presently nearly non-existent due to its high cost compared to the clinical utility of its findings, yet the potential for the impact of this approach remains substantial as certain genomic variants such as Structural Variations (SVs) typically have a large effect on an individual's phenotype due to their size[138], but are generally not amenable to interrogation via gene panels.

The magnitude of the opportunity for improvement in the space of DNA sequencing and genomics is thus clear to us - we seek a way to improve the current methods of DNA data analysis such that it becomes tractable and cost-effective to undertake whole-genome sequencing studies within research and clinical contexts at the scale of hundreds of thousands to millions of human genomes.

1.2 Challenges and Problem Statement

Let's examine the key challenges that need to be addressed in order to enable efficient genomic data analysis at the scale that is desired by the research and clinical communities.

Several broad groups of challenges are identified below and further examined throughout this thesis:

Data Set Size - The size of the raw genomic data generated by population-scale studies will be hundreds to thousands of petabytes making it impractical to move and make copies of the data[162].

Data Retention - The cost of generating the data is significantly higher than the cost of storing the data, thus making it impractical to throw away the raw data after initial analysis[125].

Data Formats - The data formats used for storing genomic data are primarily large size character and binary files (FASTA, SAM, BAM, VCF)[105, 32] that have loose specifications and scale poorly to large cohort sizes. File indexing

structures typically support indexing by genomic coordinate only, thus limiting queryability.

Data Fragmentation - The data will be generated at multiple sequencing centres located in different jurisdictions with a wide variety of genomic data handling requirements. Data processing must proceed at multiple locations that respect the requirements of each jurisdiction[124].

Data Type Diversity - Comprehensive characterization of a person's genome that is useful in a clinical setting implies the collection and integrative analysis of many diverse data types - including germline[110] and somatic[62] genomic variants, transcriptomics[177], epigenomics[80], metabolomics[174], and clinical information. Uniform collection, processing, and integration of these data types is required to successfully associate the role of this genomic variation on disease phenotypes[146].

Data Processing Stages - Data processing for genomics analysis proceeds through a sequence of stages from base-calling, to quality-control, to genome alignment, to variant calling, to annotation, to downstream analysis[36]. Each stage typically has non-trivial computational requirements needing several days on a multi-core machine to complete with increased failure risk as a function of data set size. Intermediate results from one stage are often required as input for downstream stages. Fully sequential processing makes inefficient use of the data by redundantly loading and interrogating the data in memory over a series of passes through the sample.

Toolset Fragmentation - Although comprehensive genomic characterisation of each sample is typically of interest to researchers, specific bioinformatics tools only provide solutions to a limited subspace of the overall problem, thus requiring integration of multiple tools that may produce incongruent outputs and compete for resources producing computational bottlenecks.

Having listed these challenges we attempt to restate the problem in simpler terms before providing a high level overview of the types of approaches and solutions that will be developed and considered in detail in the body of this thesis in order to deliver a conceptual and practical framework for the effective management of genomic data at the desired scale.

Our problem statement is then as follows:

Human genomic data sets will, in the future, be generated for analysis in various locations throughout the world, at the aggregate rate of multiple petabytes of data per day in the context of disease and clinical practice. The desired outcome of these analyses is the comprehensive characterisation of genomic features and their association with phenotypic variables of interest[182]. The goal of the research community is in capturing the maximum number of samples – N, with high accuracy – A, to increase statistical power of studies[75], while the interest of clinicians is to capture specific individuals with high accuracy – A, and in the shortest possible time – T, in order to inform clinical decision making[175]. Both parties wish to do so at

minimal possible total cost – $C = c_g + c_s + c_a + c_r$, taking into account the cost of data generation, cost of data storage, cost of data analysis, and cost of subsequent data retrieval. Because of the high cost of generating this data each time, the data, once generated, will need to be stored for the foreseeable future. The overwhelming data set size prevents data movement between locations, requiring analysis algorithms to be colocated with the data.

The analysis is hampered by reliance on data formats that have not been designed for operation at such large scale and the necessity to execute a variety of computational algorithms[101, 192, 7, 120] on the data that have been individually developed by different authors within an academic context, using different technologies that compete with each other for computational resources, and at-times produce contradictory results that require human intervention to integrate. The underlying assumption of genomic coordinate-sorted ordering and traversal of the data made by most algorithms limits the modes of reasoning about the dataset to a series of pre-processing steps, followed by another series of coordinate-wise traversals through the data, which impose severe processing time costs, such as the requirement to have generated, seen and sorted all of the data, before an analysis can proceed as well as the inability to stop and interpret analysis results mid-processing.

The optimization problem of maximising N, and A, while minimizing C for research purposes remains unsolved for values of N above 3000 samples when it comes to high-coverage whole genome sequencing, while the problem of maximizing A, and minimising T, and C is presently not solved in the clinical setting for any sample size. It is our proposed solution for tackling these issues that we turn to next.

1.3 Proposed Solution

We assume that N, the number of samples that can be successfully sequenced will depend almost entirely on the total cost C, which itself, among other factors, is determined by the desired accuracy and processing time. We thus focus most of our efforts on the joint optimization of cost, accuracy, and time as necessary conditions for the maximisation of effective sample size N and enablement of whole genome sequencing for clinical practice.

We note that the cost of data generation C is dependent on the sequencing technology used, the underlying chemistry, and the cost of the reagents[123]. Improving these characteristics falls outside the scope of our discussion, and we assume the cost of the data generation component c_g of C to be constant throughout this thesis.

We establish and discuss at length the characteristics of the optimization function in the body of the thesis but here note briefly that analysis accuracy A is evaluated along the usual dimensions of sensitivity and specificity and can generally be improved by generating more data for a given sample up to a theoretical maximum inherent in the sequencing technology used and the nature of the analysis algorithms employed. Generation of more data naturally leads to increased analysis time T and cost C. The time to accomplish the analysis can be reduced by either giving up ac-

curacy (by looking at less data, or using faster but less accurate algorithms[102]), by increasing the level of parallelisation within the computational pipeline i.e. parallelising steps that are currently sequential[85], or by utilising additional computational resources, thereby increasing costs. The various components of cost, in turn, can be optimized by improved data storage and retrieval structures[132] (via multi-level caches and hybrid storage media, for example), by improved-efficiency analysis algorithms, and by reduction of analysis accuracy and increase of analysis time (via cheaper hardware).

It is clear from the discussion above that cost, accuracy, and processing time are not orthogonal concerns i.e. changes in one may lead to changes in the other two. It thus appears that no optimization effort is likely to simultaneously satisfy the requirements of all parties that are interested in large scale genomic analysis, and a successful computational framework for delivering such analyses must allow efficient and dynamic optimization of these parameters to fit the needs of the end user. This is typically not the case for present day genomics frameworks because of the sequential way they look at data[119, 170] i.e. all of the data is generated before it is processed by downstream tools, and accuracy and processing time need to be decided on before launching a set of tools because they step through the genome in coordinate-wise manner.

To address these challenges we develop and describe within this thesis a new computational framework, called Rheos, that is based on the concepts of data streaming, cloud computing, and service orientation to provide a comprehensive toolset for genomic data analysis that can potentially scale to processing of millions of genomes while arming its users with the capability to make timely, responsive, and principled decisions about the tradeoffs between analysis cost, accuracy, and duration.

Three distinct characteristics set Rheos apart from current generation genomic analysis frameworks and each of these allows us tackle some of the issues and challenges described in Section **??**. These are:

- Service Orientation
- Event and Data Streaming
- Random Data Ordering

Service orientation[42] allows us to decompose the overall problem of comprehensively reasoning about genomic data into a set of small loosely-coupled components, each of which is optimized to tackle a particular well-defined subset of the complete set of requirements of the system. Each service has a contract that it makes with its clients, it has an explicit set of inputs that it knows how to process, it has an interface that defines the modes of communication it supports, it has a set of outputs that it produces according to its capabilities, and it has a set of operational characteristics that makes explicit commitments about the service's reliability, speed, etc[133]. This has a number of benefits - a service can be small enough that it optimizes the solution to a particular problem without being subject to the same

competing constraints that larger tools are subject to, which provides opportunities for improved performance and hardware utilization. As long as the service respects its input and output commitments it is free to maintain arbitrary internal representations of the data enabling optimization of data storage and query costs (c_s and c_r). A service can be monitored such that hardware is allocated elastically up and down based on demand to ensure optimal utilization, as well as providing a continued measure of whether the service is meeting its operational reliability requirements to its clients[29]. This is especially useful in contexts where demand for certain calculations is highly variable.

The issue of inter-service communication is of major importance because of the large size of the data-set and the potential for various difficult-to-debug run-time race and error conditions inherent in a distributed system[56]. Currently, most bioinformatics tools do not communicate with each other directly via an API, instead they use popular file formats such as SAM/BAM/CRAM[105, 54], and VCF[32], as well as a myriad of more esoteric file formats not only as a storage medium but also as a means of communicating information between each other. This paradigm hurts the ultimate scalability of the entire system because of the necessity to write data to disk and possibly move it over the network in order to enable communication across tools. Furthermore, a file-based information exchange mechanism forces a coarse-grained, sample-level, communication between components that wish to avoid tight coupling between each other, even though most of the reasoning about genomic data occurs at locus, or small locus-neighbourhood, levels[39].

Rheos adopts a data and event stream approach to accomplish scalable fine-grained communication between services[126]. This approach allows each service to listen to and produce data at the level of granularity that it needs to make decisions, and that its downstream dependencies are interested in (for instance at read, locus, or breakpoint levels). When primary data is ingested into the Rheos system (from a sequencer, or a data repository) the data stream can start to be analyzed immediately[67], unlike file-based systems that need to wait for the entire sample to transmit before beginning. This approach can potentially enable real-time analysis given sufficient allocation of computational resources[4]. Since the raw data is extremely large, it is advantageous to move this data between machines, and between disk and RAM as little as possible, thus instead of passing the raw data around the network various services pass around events of interest about the raw data amongst themselves[43]. When a particular service needs the raw data (rather than the corresponding events) for its decision-making it can be shipped this data as necessary, or it can be instructed to run on the host that has already cached this data in memory. Data streaming allows for extreme scalability, but a key challenge when dealing with data streams is that one is no longer guaranteed to ever be able to see "all of the data" for a particular sample, at least in any meaningful amount of time[55]. Because genomic algorithms frequently make use of various summary statistics accumulated over the data-set[134, 94], not being able reason over all the data at once means that approximations for these summary statistics are required. Rheos uses approximations calculated within time windows over the data stream[34, 10] and we consider their properties in detail in the body of this thesis.

A key assumption made by nearly all algorithms in the genomics space that

participate in variant calling and reason over sequence reads is that the reads are coordinate-sorted with respect to the reference genome to which they are aligned[105, 58, 20, 142]. The algorithms then proceed by traversing the genome in coordinate-wise fashion from the beginning of chromosome 1 to the end of chromosome Y interrogating each locus in turn by examining the set of reads that overlap that locus (a read pileup)[105]. Getting the reads into a state that is usable by these algorithms then requires, at a minimum, that all the reads for a given sample have been generated, have gone through QC[183], have been aligned[102], have been investigated for PCR duplicates[170], and have been sorted[170]. Each of these steps can take hours or even days to complete, especially on high coverage whole genome samples. We take a different approach with Rheos by relaxing the requirement for the reads to have been sorted before any variant calling can take place, and instead develop a set of variant calling algorithms that do not assume any particular order within the data that they observe. This allows Rheos to make use of sequence data as soon as it comes off the sequencing machine, thereby dramatically reducing the total time T required to process genomic data compared to the current generation of algorithms. Rheos accomplishes this by employing the service- and stream-based approaches discussed above to process each read on-the-fly as it moves through the system. The read is first assessed for quality, then aligned to a reference genome by the alignment service. This service emits an event with a coordinate that corresponds to the alignment. Variant calling services listen to this event stream and incorporate the evidence for genomic variation supplied by this read into their models of the genomic features that exist at that particular locus for that sample via a statistical framework based on an iterated application of Bayes' rule[189, 12].

Because current generation tools can see all of the data for a particular locus at once they can incorporate all of the evidence supplied by this data in a minimal number of calculations, corresponding to each particular algorithm[105, 58]. Rheos, on the other hand, to incorporate the same amount of evidence will need to perform a larger number of calculations in a redundant manner, incorporating the data as it is observed. This cost is compensated for, however, by the fact that Rheos can immediately incorporate new data about a particular locus when it becomes available without the need to have accumulated all of the data for all of the loci, generating significant time savings. Furthermore, because data arrives in no particular order the set of variant calls produced by Rheos at any given point in time represent a comprehensive characterisation of the sample as if the sample was sequenced at an average coverage consistent with the amount of data that has been observed so far. Observing more data is equivalent to raising the average coverage uniformly throughout the genome, thereby improving call accuracy[6]. This provides us with a framework to actively and dynamically trade off call-set accuracy A for processing time T and cost C as actual data is being observed thereby enabling novel applications whereby sequencing is abandoned early when issues such as sample-swap[46], or contamination[19] are detected. In addition, when sufficient accuracy is reached based on observed data at a particular locus, the framework may choose to stop looking at further data, whereas current generation approaches necessitate committing to a particular sequencing depth a-priori. Furthermore, because current methods iterate through the data in a coordinate-wise manner, their partial results are not really usable until the entire data-set has been traversed (as they represent only

a particular region of the genome), whereas Rheos call-sets represent progressive elaboration of a complete genomic characterisation and are thus usable at any level of accuracy that is fit for the purposes of the underlying analysis. We develop the details of the statistical framework used by Rheos and compare its theoretical and real performance to current generation frameworks in the body of this thesis.

We conceived of Rheos as a modern bioinformatics framework that aims to enable the large scale genomics studies of the future[41, 81, 115] in both research and clinical contexts by providing a toolset that allows for interpretation and comprehensive characterisation of high coverage human whole genome samples at the scale of millions of samples. In order to meet the diverse requirements of its users the framework allows users to make informed and dynamic tradeoffs along the optimization dimensions of of cost, accuracy, and time. Rheos unique abilities rest upon three characteristics that set it apart from current generation tools, these are: service orientation, data streaming, and random data ordering. Taken together these characteristics enable Rheos to perform at unprecedented levels of scale while retaining call-set accuracy and reducing per-sample processing time. We dedicate the main body of this thesis to the development of the theoretical framework underlying Rheos, exploring its characteristics, benefits and tradeoffs, discussing its implementation, and evaluating and comparing Rheos' performance to the current best practices in genomics algorithms on real data.

1.4 Thesis Outline

This thesis focuses on the development of the conceptual framework behind the Rheos platform, the theoretical properties of the various algorithms employed by Rheos, and the implementation and experimental validation of the framework on real data. Chapter 2 provides an introduction to the fields of genomics, including cancer genomics and clinical genomics, a survey of the main tools and algorithms that are commonly used in genomics is provided including the details of the underlying statistical models and operational characteristics. The chapter concludes with a look at workflow frameworks that tie individual algorithms together into computational pipelines. Chapter 5 sets up and describes the conceptual framework underlying Rheos based on the approaches of Service Orientation, Data Streaming, and Random Data Ordering, mentioned above. We describe the overall architecture as well as the model behind individual services that comprise Rheos and investigate the theoretical properties of the algorithms that underlie Rheos-based genomic analysis. Chapter **??** describes the actual implementation of the Rheos framework's components and investigates their operational characteristics. Chapter **??** is dedicated to the experimental evaluation of Rheos in comparison to other extant frameworks and algorithms using real genomic data. We conclude this work in Chapter 6 with a discussion of the results and an examination of the future direction of Rheos development.

Chapter 2

Background and Related Work

2.1 Genomics

The field of genomics is closely related to, yet distinct, from the field of genetics, which itself stems from the work of such seminal figures as Charles Darwin[33] and Gregor Mendel[64]. While genetics largely focuses on the study of single (or relatively small numbers of) genes - the *genotype*, and how genetic variation and mutation affect the physical traits of a given cell or organism - its *phenotype*, genomics focuses on larger scale events and mechanisms that tend to act on the entirety of an organism's genome, shaping its architecture and ultimately affecting its survival.

2.1.1 History of Genomics

Each living cell is a bio-chemical machine that carries out a number of complex behaviours such as interactions with the surrounding environment, motility, metabolism, and reproduction, that are necessary for its survival and proliferation, based on a genetic program that is encoded within the cell's DNA. The DNA is nominally subdivided into functionally distinct areas known as *genes*. The cell utilizes the program within each gene by first *transcribing* the DNA into an intermediary information-carrier molecule called RNA, and then *translating* this RNA into molecules called *proteins* that are utilized by the cell to carry out the majority of its functions. Understanding and interpretation of the underlying genetic program thus underpins our ability to comprehend the entirety of the different behaviours that each cell undertakes.

The success of this undertaking is contingent, first and foremost, on our ability to effectively read off the information encoded in the DNA, an activity known as *sequencing*. We are able to sequence DNA thanks to the pioneering work of researchers Rosalind Franklin[53], James Watson, and Francis Crick[178] who first elucidated the physical structure of DNA, then followed by the work of Fred Sanger[149, 148] who devised the first effective DNA sequencing method. The sequencing method

allows us to transform information that is physically encoded on the DNA molecule via a sequence of four distinct types of *basepairs* - Adenine, Cytosine, Guanine, and Thymine into a string stored on a computer using a four-letter alphabet - A,C,T, and G, thus turning DNA interpretation into a digital information processing problem.

While the entire length of the DNA of an organism ranges from several hundred thousand basepairs for simple organisms like viruses and bacteria, to about 3,000,000,000 basepairs for a human, to over 150,000,000,000 for certain plants[135] the limitations of Sanger DNA sequencing technology are such that the sequencing machine can only produce DNA fragment strings, known as *reads* that are 800 - 1,000 basepairs long[149]. Reconstituting the original complete DNA sequence from partial overlaps between reads is thus a costly, time consuming, and computationally intensive problem known as *de-novo assembly*[191]. Once one such full sequence (known as a *reference* sequnce) is assembled however, sequencing other individuals of the same (or closely related) species becomes a significantly easier undertaking. Rather than assembling the sequence *de-novo* one can search for a position on the reference sequence that provides the best matching *alignment* between the reference and each read obtained for the specimen under study. This technique is known as *genomic alignment*[101] or *mapping* and yields for each fragment a coordinate that represents where on the reference sequence the fragment maps to. Furthermore, because the DNA of any two organisms of the same species is largely identical, with differences occurring at about 0.1% of all sites (although this depends on DNA mutation rate)[127] researchers are able to significantly reduce the amount of information that is required to fully represent the genome of a specimen by retaining only the information that describes the sites where that specimen is different from the reference sequence for that species.

A general approach has thus emerged, where each new species of interest undergoes a relatively costly *de-novo* assembly process for the first genome, which then becomes the reference genome for that species. The sequencing of further individuals of that species utilizes, relatively cheaper, *alignment* and identification of *variants* (sites where the individual differs from the reference) to investigate the effect these variants may have on different phenotypes of interest such as disease susceptibility and survival[112].

Although genomicists study many different types of organisms the study of human genomes garners by far the most attention and research funding[**needcitation**] due to the natural desire of humans to better understand ourselves and influence, where possible, genetic factors impacting human longevity and health. Subsequent to the development of DNA sequencing methods by Fred Sanger one of the most audacious and crucial projects for the development of genomics as a branch of science has been The Human Genome Project[84] - an international effort to sequence and *de-novo* assemble the first complete human genome consisting of chromosomes 1-22, X, and Y (as well as mitochondrial DNA) and totalling approximately 3 billion basepairs. The project ran for over 10 years, completing in 2001, and cost more than $3 billion USD. Although the main project effort was completed using the Sanger sequencing method, a competing version of the human genome was simultaneously published by a commercial company led by JC Venter[172], using a new sequencing method called shotgun sequencing[173], a method that formed the basis for a new revolution

in sequencing technology, now termed Next Generation Sequencing[151].

2.1.2 Next Generation Sequencing

The Next Generation Sequencing methodology[114] relies on fragmenting the DNA of a subject into millions of fragments that are between 100-500 basepairs (bp) in length, then sequencing all of the short fragments and aligning all the reads to the reference with the aid of a relatively fast algorithm[102]. Because NGS sequencing methods are prone to certain errors and biases[38], it is necessary to sequence enough DNA fragments to overlap (or cover) every location in the genome several times (typically 10-30), in order to build a statistical model that will be able to determine the underlying sequence, known as *genotyping*[128], with a high degree of confidence. Thus, at present, a single sequenced DNA sample will typically contain 1 billion reads with a file size of 150GB when compressed.

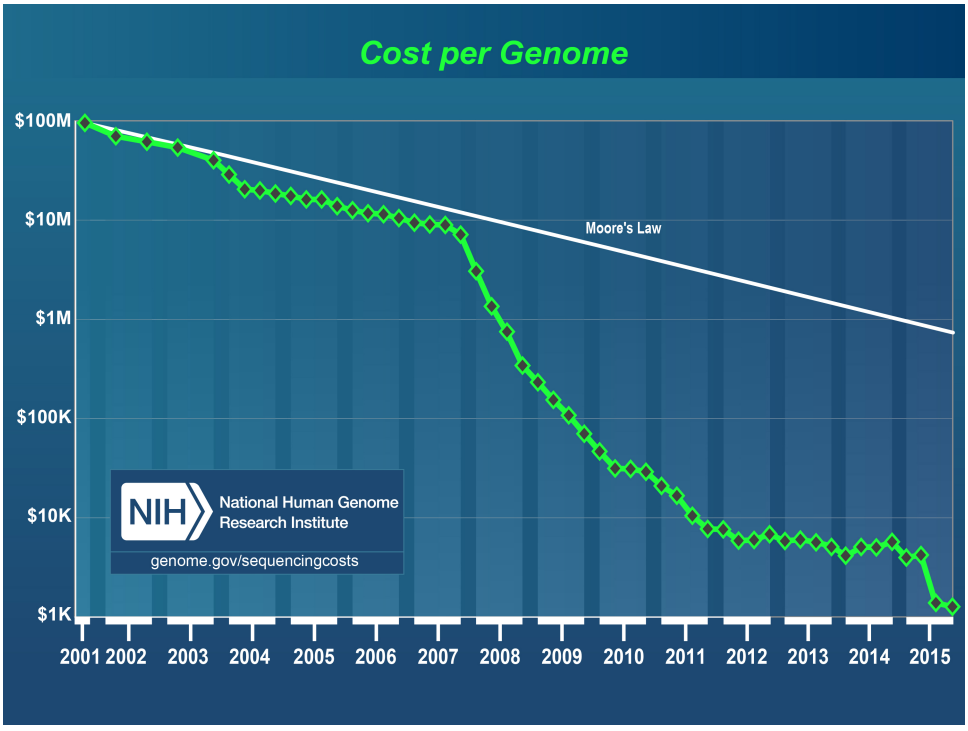


Figure 2.1: Cost of DNA sequencing[167]

Figure 2.1 shows the change in the cost of DNA sequencing over the course of the past 15 years. The precipitous drop in sequencing cost observed since 2008 coincides with wide adoption of NGS methodologies. This drop in price has made tractable a new set of large scale genomics sequencing projects that aim to characterize human genetic diversity at population scale, projects such as the 1000 Genomes Project[24], and the large scale sequencing of the Icelandic population[66].

2.1.3 Genomics Studies

2.1.4 Cancer Genomics

Cancer is a genetic disease that has an extremely high burden on the human population. In 2012, the global incidence of new cases worldwide has been estimated as 14.1 million, and deaths at 8.2 million[168]. The economic cost of cancer to the European Union has been estimated at 126 billion euro in 2009[109], and in the US $124.5 billion USD in 2010[185]. Because of the genetic nature of the disease studying genomes of cancer patients helps uncover the mechanisms behind the development and evolution of cancer[164].

Cancerous tumours arise from a single cell which over time accumulates a series of somatic mutations that cause it to exhibit properties such as: increased mutation rate, increased proliferation, anchorage independent growth, and resisting cell death[68] . Only certain mutations, however, contribute to the development of cancer, while others are benign. Cancer genomics studies aim to identify and characterize those mutations that are cancer drivers and play a role in the formation or progression of tumours[164].

Studying cancer genomes is more complex and expensive than studying the genomes of healthy individuals because each patient requires that two DNA samples are collected - that of the normal tissue, and that of the tumour. This is necessary to identify those mutations that are somatic - i.e. only occur in the tumour cell population[144].

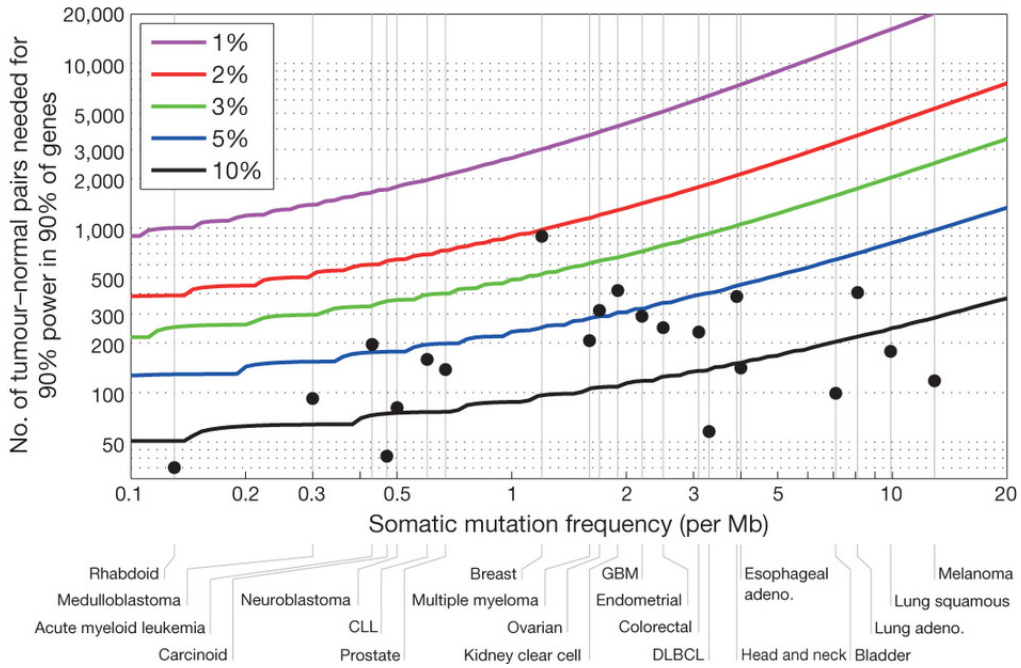


Figure 2.2: Sequencing sample size required by mutation rate[89].

Although there is a large number of identified mutations that are implicated in

cancer (2,002,811 SNV, 10,534 gene fusions, 61,299 genome rearrangements, 695,504 CNV segments in COSMIC v70; August 2014)[51], each mutation has a low chance of being present in any given tumour. Figure 2.2 demonstrates the sample size required to have 90% statistical power to identify 90% of the variants that occur with a set frequency in tumours with varying background mutation rates. Thus, identifying 90% of the mutations occurring with a frequency of at most 1% in Lung Adenocarcinoma requires a sample size of at least 10,000 patients. The necessity to sequence large cohorts of patients in order to be able to comprehensively detect cancer related genomic variants has led to the creation of several large scale cancer sequencing studies.

2.1.5 Clinical Genomics

2.2 Computational Methods for Next Generation Sequencing

Because the size of a typical genome is millions to billions of basepairs long, and current DNA sequencing technology frequently generates errors during the sequencing process, requiring multiple samples of each genomic location to be generated, the amount of data required to be examined in order to characterize even a single sample is well beyond the capabilities of any human. Thus, a multitude of computational approaches are required in order to make the task tractable for individual samples as well as cohorts, and entire populations.

The task of comprehensive characterization of genomic data for an individual is typically decomposed into a series of computational steps, each with its own data representation, and typically developed by a separate research group, which are then assembled into computational pipelines and executed by workflow engines on diverse computing environments. Our goal in this section is to enumerate and describe the individual steps and to provide a survey of the key computational tools and data formats that presently form the set of best practices in this rapidly evolving branch of science. Since Rheos is designed to improve upon these best practices we identify in each section the key mathematical and algorithmic ideas that underpin each approach in order to adapt and translate them into the Rheos framework.

The data that is used in virtually all modern genomics studies is generated on a next generation DNA sequencing machine. Several types of sequencers have been developed but the most frequently used ones are made by Illumina. The raw data produced by such a sequencer is a set of image files, where the color of each pixel represents the corresponding nucleotide base in a DNA strand that is being sequenced in each micro-well of a flowcell, representing the sample of interest. The succession of images produced by each cycle of sequencing then results in a set of reads, a collection of randomly ordered DNA fragments that are further analyzed by downstream tools. The first challenge in generating these reads is the accurate interpretation of pixel colors and mapping them to the corresponding nucleotide

bases, known as base-calling. Because all of the currently available DNA sequencing methodologies are imperfect at reading the underlying DNA sequence a number of errors is introduced into the process at various stages and special QA software is required in order to detect and assess the location and severity of the errors. A typical output of the QA process is a filtered set of reads where the lowest quality reads have been filtered out and each base within each read is assigned a quality score which represents the best current estimate of the probability that the base has been called incorrectly. The currently most frequently used file format for storing DNA sequence reads along with their read qualities is a text file known as fastq.

Depending on whether the organism under study has previously been sequenced there may already exist a reference sequence for it i.e. a file that for each genomic location describes the most frequently occurring nucleotide for that species at that location. Humans, and many other species of organisms already have reference sequences available. If the reference sequence for the organism under study is available then the next processing step involves searching for the position in the reference sequence that best matches each read that has been generated for the sample under study in the previous step. The coordinate of the best match is then assumed to be the location in the genome where that particular read has originated from. This process is known as genome alignment and it is very resource intensive for species with large genomes such as humans (3 billion bases) because a typical sequencing effort will generate at least 1 billion reads for a single sample, and each read needs to be mapped to the reference genome. This problem is made more difficult by the fact that an organism's genome typically has a large proportion of repeated sequence fragments and thus the generated reads do not uniquely align to a single location on the reference. A list of matching positions is generated instead, where each match needs to be scored and the highest scoring match is assumed to be the true origin of the read. Many alignment algorithms exist but the most accurate and fast ones use a two step process of indexing, implemented via hash tables or prefix/suffix tries, to generate a short list of promising match locations, followed by a more exact local alignment that uses dynamic programming to generate a best match. The alignment process is further complicated by the presence of sequencing errors, various genomic variants, and disease state such as cancer, all of which generate significant (and sometimes drastic) differences between the obtained reads and the reference genome, thus necessitating inexact matching approaches. The best algorithms that are currently available have a typical runtime of 24-48 hours on a modern 8-core machine. The most widely adopted standard for storing the alignment data on disk is the SAM[105] (and its binary and indexed counterpart BAM) format developed in the context of the 1000 Genomes Project. In addition to the sequence data and base qualities that are already available in fastq, the SAM format adds a reference coordinate to each read, an overall mapping quality for the read, and whether each position in the read matches the reference sequence, along with other useful metadata.

When a reference sequence does not exist, or when it is undesirable to use one, genome alignment tools are inapplicable and a different approach, called de-novo assembly, is used. Under this approach each read is broken into smaller subsequences called k-mers (of length k), these k-mers are then used to build a graph structure

called a de Bruijn graph. Unique paths through the graph represent possible arrangements of reads that correspond to the underlying sequence and the highest scoring path is chosen as the true sequence. Using the de-novo assembly approach has some advantages over alignment-based methods because it models the structure of the organism's genome directly as it is observed rather than in relation to a reference. This is because no reference is perfect, but instead each reference has its own set of errors that were introduced in its construction. Furthermore, genomic structural variants, which represent large (hundreds to millions of basepairs long) sequences that may be deleted, duplicated, or inverted within a given genome challenge alignment software because of the alignment errors that they introduce and require sophisticated algorithms to later detect, whereas in the de-novo assembly approach these variants are directly modelled as they occur in the underlying sequence and are thus easier to identify. De-novo assembly has its own set of challenges however related to difficulties dealing with repetitive sequences that are found within the genome, as well as the extremely high resource requirements of de-novo assembly algorithms, especially when it comes to memory. The de Bruijn graph is typically built in memory and can be multiple terabytes in size, thus requiring computers with extremely high memory to process. Since, even when using in-memory graph construction the runtime for a single sample is typically several days, it is impractical to move the graph representation to disk without dramatically increasing the algorithm runtime to the point where its duration becomes unreasonable. In practice whole genome de-novo assembly is currently rarely used for processing human genomic data because of the challenges described above. Instead, modern algorithms supplement read alignment with local assembly of particular genomic regions of interest in order to reap some of the benefits offered by assembly-based methods without incurring all of the costs.

Once the reads have been aligned they are typically sorted by genomic coordinate so that all of the reads that overlap a given coordinate can be examined together at once. This is an expensive sortation step that does not lend itself well to parallelization and takes several hours to complete per sample. Subsequent to the sortation step is another round of data QA which aims to throw out low quality reads that poorly align to the reference. Care must be taken however, because these low quality reads may not only signal underlying data or sequencing issues like sample contamination, or lane-swap, but may also signal the presence of structural variants or integration of retrovirus DNA into the host under study, both of which are of high interest to properly identify. Thus, it is common to split the sample into reads of high quality that are further assessed with one set of algorithms and a set of reads that map with low quality, or fail to map at all, to be assessed with a different set of algorithms.

At this point the data is ready to begin the process of variant calling, that is, identifying the genomic features of the sample that are different from the reference sequence for that organism (i.e. mutations). It is important to distinguish germline variant calling from somatic variant calling at this time. In germline variant calling we are trying to identify the set of variants that have been passed to the individuals under study from their parents and are thus present in every cell of the organism forming the underlying genetic background of that individual where some variants

may be neutral to the organism's survival, some may be beneficial, and some may be deleterious. Comprehensively identifying and classifying these is of significant research and clinical interest as they confer susceptibility or resistance to certain disease vectors as well as potential medical remedies and may act as biomarkers to predict disease prognosis or response to treatment within the groups of patients that harbour them.

Somatic mutations are those that each individual cell accumulates over its lifetime and they are of especial interest in the context of cancer where a certain set of mutations accumulated in a particular sequence and over a period of time disrupt the normal cell lifecycle and result in the formation of a malignant tumour. In this context researchers typically sequence both healthy cells (such as those drawn from the patient's blood) and cancerous cells. Mutations are identified in both and the difference between these sets of mutations is then stipulated to be the set of somatic mutations present within that tumour. Just like in the germline case, not all of the somatic mutations contribute to the formation of the cancer and the appropriate identification and classification of those mutations that do (so-called cancer drivers) is an important question of significant clinical and research importance which we consider further below. From a technical standpoint calling somatic variants is significantly more complex than calling germline variants because healthy cells generally conform to the underlying genetic characteristics of the organism, such as the number of chromosomes and ploidy (23 chromosomes, diploid, for humans), whereas in the cancer cells these characteristics can be severely disrupted with entire chromosomes missing or present in amplified copy number, requiring different and more complex statistical models to accurately identify. An additional complexity that is unique to somatic variant calling is the concept of sub-clonal mutations. These are mutations that have been acquired only by some of the cells within a tumour. Since sequencing samples data from a large number of cells within a tumour the reads from which are all pooled together, only a comparatively low number of reads will contain information about sub-clonal mutations, thus making them more difficult to detect, even though such mutations may have a significant impact on the tumour phenotype and thus would be very important to properly identify.

We typically think of three classes of genomic variants that are identified by different methods and oftentimes by separate tools. The simplest to accurately detect, and most frequently occurring are Single Nucleotide Polymorphisms (SNPs), in the germline case, and Single Nucleotide Variants (SNVs), in the somatic case. These are single basepair substitutions where the germline genome differs from the reference sequence by a single letter (for SNPs), or the somatic genome differs from the germline genome by a single letter (for SNVs). SNPs are quite common in humans and occur at the rate of approximately 1 per 1,000 bases on average, or, equivalently, 3 million per individual. Somatic SNVs have a widely varying incidence rate depending on the type of cancer involved with typical rates between (INSERT RATES HERE). For humans, which are diploid (i.e. have two copies of each of the chromosomes, except for the sex chromosomes X and Y), we classify SNPs and SNVs as being either heterozygous (with one reference allele and one variant allele) or homozygous (with both alleles being variant). Methods to detect and accurately genotype SNPs and SNVs typically rely on counting the reads that overlap a given

genomic position and evaluating a statistical model that contrasts the probability of the site being reference versus the probability of the site being variant in the face of potential sequencing errors which are expressed as base quality scores and mapping quality scores (as previously described). The models employed for somatic SNV detection and genotyping are significantly more complex than the models for germline variant detection because of the possibility of sub-clonal mutations (as previously described) as well as regions of amplified copy number (i.e. regions where the organism is no longer diploid but can have any number of additional copies of a chromosomal region, or an entire chromosome). More advanced methods output not only lists of variant sites for a sample but calculate a distribution of genotype likelihoods, i.e. all the possible genotypes at a given variant site along with their relative probabilities so that these can be integrated into the models of downstream statistical analyses in a comprehensive manner.

Indels represent sequence insertions and deletions that are anywhere from 1 base-pair (bp) to about 50 basepairs long. There is no strict upper bound on the length of an indel and individual tools typically decide on their own cutoffs for length although pretty much all tools place their cutoff at a length that is smaller than the typical read length (150 - 500 bp presently). Indel callers typically look for several mismatched bases in a row between the reference and the sample under study and classify the entire length of the mismatched sequence as an insertion or deletion correspondingly. Other indel callers borrow some of the methodology from structural variant callers which are similar to indels, only typically bigger in size, and are potentially more complex.

Structural variants (SVs) are more large scale genomic rearrangements that occur in both germline and somatic genomes and can have a very drastic effect on the organism's phenotype because they can affect a large number of genes at once, resulting in the loss of function of particular important genes, or the creation of gene fusions where, because of a rearrangement, one gene comes under the programmatic control of another gene thereby disrupting important cellular processes. The most common types of structural variants include insertions, deletions, segmental duplications, inversions, and translocations. Simpler structural variants sometimes combine to produce more complex events that are especially difficult to detect properly. The methods for calling and genotyping of structural variants typically rely on looking at the reads that are deemed low quality for the SNP calling process. These are reads that fail to map to the reference genome, split-reads, which are reads where one part of the read maps to one location on the reference and another part of the read maps to another location, and divergently mapped reads (sequencing is frequently done on read pairs where two ends of a DNA fragment of standard size are sequenced in the opposite directions generating a pair of reads with a standard distance, called insert size, inbetween them), with a shorter or longer than expected insert size. SV callers break down these reads into smaller fragments (k-mers) and attempt to map these k-mers to the reference sequence. The goal is to determine the location of breakpoints, which are positions on the sample genome where a DNA strand break is thought to have occurred as part of the genomic rearrangement that has taken place. Once a list of breakpoints is obtained the algorithm attempts to reconstruct the most likely event sequence that these breakpoints could have arisen

from, pairing up adjacent breakpoints that are the result of a sequence deletion, for example. Thus, each pair of breakpoints typically gives rise to a single SV call in the final output of the caller. SV calling is a complex and error-prone process that generates double-digit false-positive and false-negative rate, especially in the somatic case, where patient genomes can undergo drastic rearrangements as a result of cancer-related processes such as chromothripsis and are thus extremely difficult to resolve with accuracy.

Once variants (SNPs/SNVs, Indels, and SVs) have been comprehensively called, a filtering step is necessary because callers are typically initially tuned for highest sensitivity in order to detect the most variants, thus admitting an increased number of false positive calls. Additionally, because calling of SNPs and SVs typically occurs separately by different tools there can be significant call-set overlap where the SNP caller sees a region as a group of SNPs, whereas the SV caller will see it as a single breakpoint. These overlaps need to be resolved in order to avoid redundant calls. A number of filtering approaches exist, some of which rely on heuristics such as strand bias, or read support to filter out low quality variants. Other filtering approaches rely on curated variant databases or machine learning methods in order to reduce the number of false positive calls. One popular filtering approach involves ensemble calling where several different variant calling methods are used on the same dataset and a variant is excluded unless it is called by multiple tools. These methods are typically able to reduce the false positive rate of the call-set by 5-10% while only nominally affecting the false negative rate.

When a filtered high quality call-set has been prepared it is of interest to determine which of the variants are likely to have an effect on the organism's phenotype and which variants are likely to have no consequence. This is accomplished via variant annotation. The annotation process consults a database of known genes and other genomic elements (promoters, enhancers, etc.) to determine the likely consequence of each variant based on the type of mutation that it represents i.e. a synonymous mutation (that doesn't change the underlying amino acid) is likely to have no phenotypic effect, whereas a stop gain mutation inside the coding region of a known gene may indicate a potential loss of function of that gene and may thus have a considerable effect on the observed phenotype. When annotating somatic mutations it is important to consider known cancer genes and delineate whether mutations are "passengers" or "drivers" depending on whether they are thought to be driving the carcinogenesis process by constitutively activating a cancer gene or deactivating a tumour suppressor, or they are simply acquired as part of the genomic instability that is induced by carcinogenesis. An outcome of the variant annotation process then, is a list of somatic or germline variants accompanied by a designation of the known genomic features that they fall in, along with an assigned functional impact. This is typically the last step of an NGS analysis pipeline after which the variant call-set is considered completed and can be used for any number of downstream analyses depending on the particular research question or clinical application being considered. For instance, the variants may be used as input into a Genome Wide Association Study (GWAS), a Quantitative Trait Locus (QTL) analysis, a rare variant association study, or as input into the computation of a clinical biomarker.

2.2.1 File Formats

DNA sequencing studies generate large amounts of data - a single whole-genome sample sequenced at 30x coverage on a modern Illumina sequencer generates roughly 10^9 reads which are strings of length 150 characters and take 100 GB of space on disk when compressed with gzip. Thus, even a moderately-sized study of several thousand individuals needs to grapple with the efficient management of hundreds of terabytes of data. Because of the size of these datasets data storage, access, and exchange formats play a major role in determining the speed, cost, and efficiency with which large-scale analyses can be undertaken. The field of genomics has developed in bursts associated with the major international projects that have been undertaken in the past 30 years, including the Human Genome Project[84], the HapMap Project[26], and the 1000 Genomes Project[24]. It is the latter project that has given rise to most of the currently adopted file format standards in use today, including FASTA/FASTQ for raw sequence data, SAM/BAM for reference-mapped sequence data, and VCF for representing genomic variants. Because these file formats have become the primary information exchange medium in the field of genomics they have a large influence on software tools and data access patterns that are in common use today and thus warrant a closer look.

FASTA/FASTQ

The FASTA/FASTQ file format is a text file format developed at The Sanger Institute for representing genomic sequencing reads[21]. Each read in the file consists of four lines:

- The first is a unique identifier (that encodes some information about the sequencing process). For example - HWUSI-EAS100R:6:73:941:1973#0/1
- The second is the read sequence itself: $S = \{s_i : s \in \{A, C, G, T, N\}\}$. For example - ACGTCCCGTCCCTNTCCA
- The third is a + sign acting as a separator.
- The fourth is a set of per-base quality scores, represented on the Phred scale, that represent an estimate of the probability that the base has been called correctly. If the error probability is defined as ϵ then $Q_{phred} = -10 \times \log_{10}(\epsilon)$ and conversely $\epsilon = 10^{\frac{-Q_{phred}}{10}}$. In the actual file Phred scores are represented with ASCII characters from !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFG HIJKLMNOPQRSTUVWXYZ[\]_ˆàbcdefghijklmnopqrstuvwxyz{|}~ordered by increasing quality, where ! is the lowest possible quality and ~is the highest.

```
@809R9ABXX101220:5:48:9875:59863/1
TTTCTTTTCTTTTTCTTCCTTTTTTTTTTTTTTTTTTTTTTTTTGAGAAGAGATTTCTCACCTTTCTCCAACGCTCC
+
2?>DCBC=CD>@?BECAD:DBAB?BA>@AA>C?CCBCCCB###################################
@809R9ABXX101220:5:61:10970:200284/1
TTTCTTTTCTTTTTCTTCCTTTTTTTTTTTTTTTTTTTTTTTTTGGGATGGGGTTTTCCCCCTGTTTCCCGGGCGG
+
BABEECCCEEDCDCFEDFFEDDDDDDDDDDDDDDDDDDDDDCC#################################
@809R9ABXX101220:5:66:9966:82226/2
TCTTTTCTTTTCTTTTTCTTCCTTTTTTTTTTTTTTTTTTTTTTTTGAATGGGGTTCTCCTCCTTTTCCCCCGGGT
+
>AB@AADDBBCEDCCCCEECEEECCCCCCCCCCCCCCCCCCCCA>>###############################
@809R9ABXX101220:5:1:3553:170934/1
CTTTTCTTTTTCTTCCTTTTTTTTTTTTTTTTTTTTTTTTTGTAGTGGTGTTTCGTCCTGGTTCCCTCGCCTGCGCTC
+
2B;5>><;=;A@:6:>DAABB=6?6AB:6CA6>C7>C#######################################
@809R9ABXX101220:5:24:11915:38264/1
CTTTTCTTTTTCTTCCTTTTTTTTTTTTTTTTTTTTTTTTTGGGATGGGGTTTCGCTCCTTTTTCCCAGGCTGGGGTGC
+
BBBCCEECCCDFECFFEDDDDDDDDDDDDDDDDDDDDCC#####################################
@809R9ABXX101220:5:65:20977:60085/2
TTTTCTTCCTTTTTTTTTTTTTTTTTTTTTTGAATTGACGTTCCGCACCTGTCGCACGGCTTGCTCGCACCGTCGCC
+
=;;<=/;94??AABAA8BCABB7BA###################################################
@809R9ABXX101220:5:27:2026:120477/1
TTTTTTTTTTTTTTTTTTTTGAGAGGGGGTTTCGCCCCTGTTGCCCAGGCTGGACTGCCCCGTCCCCGGCCAAAAA
+
>>>>?@@@@@@@@@@@@AA#########################################################
```

Figure 2.3: Excerpt from a FASTQ file. /1 and /2 at the end of read ID indicate whether read is first or second in pair.

Since FASTQ is a large text file with highly repetitive content it is typically stored in a compressed manner.

SAM/BAM

The Sequence Alignment Map (SAM) text file format and its accompanying binary compressed version BAM was created for sequencing data storage and analysis in the context of the 1000 genomes project[105]. These files provide additional fields on top of the ones available in FASTA/FASTQ in order to relay information related to sequence alignment to a reference genome, and are the most widely used format for storing DNA sequencing data today. A SAM file consists of a header and a body. These are described below. All tables, descriptions, and definitions in this section are reproduced or adapted from the SAM file specification at https://samtools.github.io/hts-specs/SAMv1.pdf.

SAM Header Each line of the header begins with the character '`@`' followed by a header record (see Table 2.1). Each line is TAB-delimited and, apart from `@CO` lines, each data field follows the format '`TAG:VALUE`' where `TAG` is a two-character string that defines the format and content of `VALUE`.

Table 2.1: SAM file header record and column definition. Tags listed with '*' are required. (adapted from https://samtools.github.io/hts-specs/SAMv1.pdf)

Tag	Description
`@HD`	The header line. The first line if present.
`VN`*	Format version. *Accepted format*: `/^[0-9]+\.[0-9]+$/`.
`SO`	Sorting order of alignments. *Valid values*: `unknown` (default), `unsorted`, `queryname` and `coordinate`.
`GO`	Grouping of alignments, indicating that similar alignment records are grouped together but the file is not necessarily sorted overall. *Valid values*: `none` (default), `query` (alignments are grouped by QNAME), and `reference` (alignments are grouped by RNAME/POS).
`SS`	Sub-sorting order of alignments. Valid values are of the form *sort-order*`:`*sub-sort*, where *sort-order* is the same value stored in the `SO` tag and *sub-sort* is an implementation-dependent colon-separated string further describing the sort order. *Regular expression*: `(coordinate|queryname|unsorted)(:[A-Za-z0-9_-]+)+`
`@SQ`	Reference sequence dictionary. The order of `@SQ` lines defines the alignment sorting order.
`SN`*	Reference sequence name. The `SN` tags and all individual `AN` names in all `@SQ` lines must be distinct. The value of this field is used in the alignment records in RNAME and RNEXT fields. Regular expression: `[!-)+-<>-~][!-~]*`
`LN`*	Reference sequence length. *Range*: $[1, 2^{31} - 1]$
`AH`	Indicates that this sequence is an alternate locus.
`AN`	Alternative reference sequence names.
`AS`	Genome assembly identifier.
`DS`	Description. UTF-8 encoding may be used.
`M5`	MD5 checksum of the sequence.
`SP`	Species.
`UR`	URI of the sequence.
`@RG`	Read group. Unordered multiple `@RG` lines are allowed.
`ID`*	Read group identifier. Each `@RG` line must have a unique `ID`.
`BC`	Barcode sequence identifying the sample or library.
`CN`	Name of sequencing center producing the read.
`DS`	Description. UTF-8 encoding may be used.
`DT`	Date the run was produced (ISO8601 date or date/time).
`FO`	Flow order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read.
`KS`	The array of nucleotide bases that correspond to the key sequence of each read.
`LB`	Library.
`PG`	Programs used for processing the read group.
`PI`	Predicted median insert size.
`PL`	Platform/technology used to produce the reads. *Valid values*: `CAPILLARY`, `LS454`, `ILLUMINA`, `SOLID`, `HELICOS`, `IONTORRENT`, `ONT`, and `PACBIO`.
`PM`	Platform model. Free-form text providing further details of the platform/technology used.
`PU`	Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). Unique identifier.
`SM`	Sample. Use pool name where a pool is being sequenced.
`@PG`	Program.
`ID`*	Program record identifier. Each `@PG` line must have a unique `ID`. The value of `ID` is used in the alignment `PG` tag and `PP` tags of other `@PG` lines. `PG` IDs may be modified when merging SAM files in order to handle collisions.
`PN`	Program name
`CL`	Command line. UTF-8 encoding may be used.
`PP`	Previous `@PG-ID`. Must match another `@PG` header's `ID` tag. `@PG` records may be

Table 2.3: SAM file alignment record mandatory column definition. (adapted from https://samtools.github.io/hts-specs/SAMv1.pdf)

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SAM Body The body of a SAM file contains alignment records (see Section 2.2.2 on details of alignment algorithms). Each record has 11 mandatory fields. These fields always appear in the same order and must be present, but their values may be '0' or '*' (depending on the field) if the corresponding information is unavailable. Table 2.3 lists the mandatory fields in the SAM format: Alignment records represent the information contained in a sequencing read subsequent to it being aligned to a reference genome (see Figure 2.4).

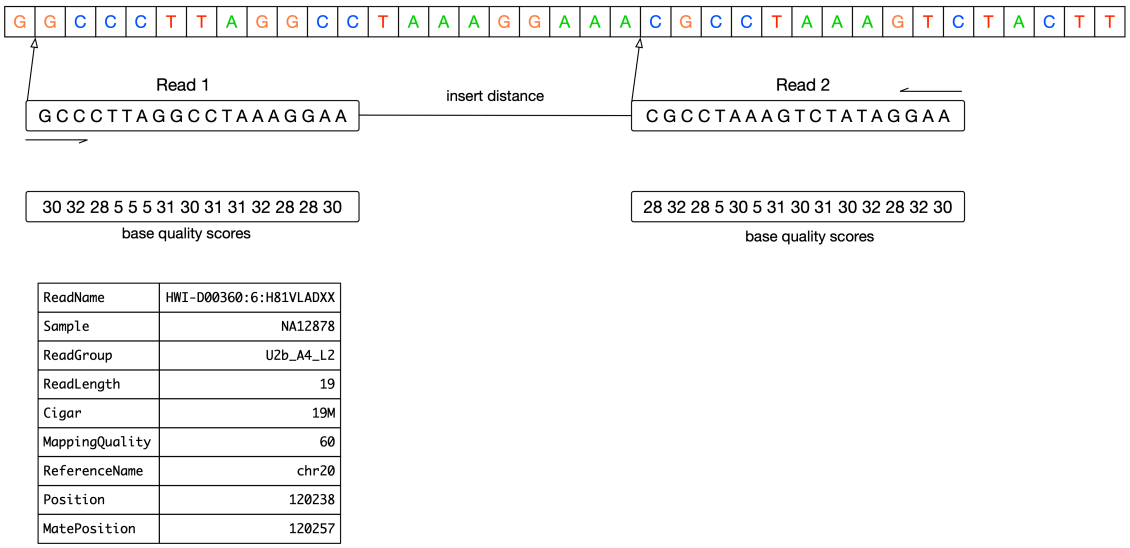


Figure 2.4: A read-pair that is aligned to the reference.

1. QNAME: Query template NAME. Reads/segments having identical QNAME are regarded to come from the same template. A QNAME '*' indicates the information is unavailable. In a SAM file, a read may occupy multiple alignment lines, when its alignment is chimeric or when multiple mappings are given.

2. FLAG: Combination of bitwise FLAGs. Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

- For each read/contig in a SAM file, it is required that one and only one line associated with the read satisfies 'FLAG `& 0x900 == 0`'. This line is called the *primary line* of the read.
- Bit 0x100 marks the alignment not to be used in certain analyses when the tools in use are aware of this bit. It is typically used to flag alternative mappings when multiple mappings are presented in a SAM.
- Bit 0x800 indicates that the corresponding alignment line is part of a chimeric alignment. A line flagged with 0x800 is called as a *supplementary line*.
- Bit 0x4 is the only reliable place to tell whether the read is unmapped. If 0x4 is set, no assumptions can be made about RNAME, POS, CIGAR, MAPQ, and bits 0x2, 0x100, and 0x800.
- Bit 0x10 indicates whether SEQ has been reverse complemented and QUAL reversed. When bit 0x4 is unset, this corresponds to the strand to which the segment has been mapped. When 0x4 is set, this indicates whether the unmapped read is stored in its original orientation as it came off the sequencing machine.
- Bits 0x40 and 0x80 reflect the read ordering within each template inherent in the sequencing technology used.[1] If 0x40 and 0x80 are both set, the read is part of a linear template, but it is neither the first nor the last read. If both 0x40 and 0x80 are unset, the index of the read in the template is unknown. This may happen for a non-linear template or when this information is lost during data processing.
- If 0x1 is unset, no assumptions can be made about 0x2, 0x8, 0x20, 0x40 and 0x80.
- Bits that are not listed in the table are reserved for future use. They should not be set when writing and should be ignored on reading by current software.

[1]For example, in Illumina paired-end sequencing, first (0x40) corresponds to the R1 'forward' read and last (0x80) to the R2 'reverse' read. (Despite the terminology, this is unrelated to the segments' orientations when they are mapped: either, neither, or both may have their reverse flag bits (0x10) set after mapping.)

3. RNAME: Reference sequence NAME of the alignment. If `@SQ` header lines are present, RNAME (if not '*') must be present in one of the `SQ-SN` tag. An unmapped segment without coordinate has a '*' at this field. However, an unmapped segment may also have an ordinary coordinate such that it can be placed at a desired position after sorting. If RNAME is '*', no assumptions can be made about POS and CIGAR.

4. POS: 1-based leftmost mapping POSition of the first CIGAR operation that "consumes" a reference base (see table below). The first base in a reference sequence has coordinate 1. POS is set as 0 for an unmapped read without coordinate. If POS is 0, no assumptions can be made about RNAME and CIGAR.

5. MAPQ: MAPping Quality. It equals $-10\log_{10}\Pr\{\text{mapping position is wrong}\}$, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.

6. CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description	Consumes query	Consu refere
`M`	0	alignment match (can be a sequence match or mismatch)	yes	yes
`I`	1	insertion to the reference	yes	no
`D`	2	deletion from the reference	no	yes
`N`	3	skipped region from the reference	no	yes
`S`	4	soft clipping (clipped sequences present in SEQ)	yes	no
`H`	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
`P`	6	padding (silent deletion from padded reference)	no	no
`=`	7	sequence match	yes	yes
`X`	8	sequence mismatch	yes	yes

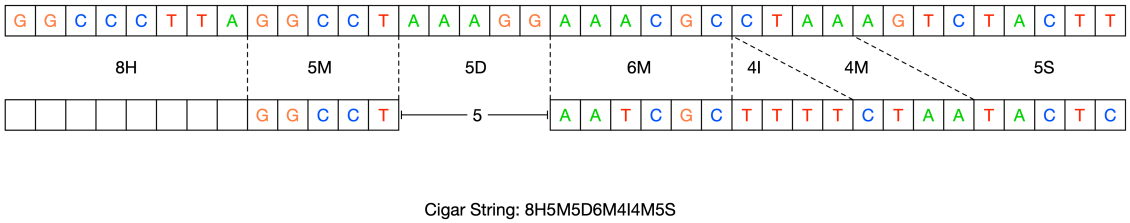


Figure 2.5: Example of an alignment CIGAR string.

- "Consumes query" and "consumes reference" indicate whether the CIGAR operation causes the alignment to step along the query sequence and the reference sequence respectively.
- `H` can only be present as the first and/or last operation.
- `S` may only have `H` operations between them and the ends of the CIGAR string.

- For mRNA-to-genome alignment, an `N` operation represents an intron. For other types of alignments, the interpretation of `N` is not defined.
- Sum of lengths of the `M/I/S/=/X` operations shall equal the length of SEQ.

7. RNEXT: Reference sequence name of the primary alignment of the NEXT read in the template. For the last read, the next read is the first read in the template. If `@SQ` header lines are present, RNEXT (if not '*' or '=') must be present in one of the `SQ-SN` tag. This field is set as '*' when the information is unavailable, and set as '=' if RNEXT is identical RNAME. If not '=' and the next read in the template has one primary mapping (see also bit 0x100 in FLAG), this field is identical to RNAME at the primary line of the next read. If RNEXT is '*', no assumptions can be made on PNEXT and bit 0x20.

8. PNEXT: 1-based Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. This field equals POS at the primary line of the next read. If PNEXT is 0, no assumptions can be made on RNEXT and bit 0x20.

9. TLEN: signed observed Template LENgth. If all segments are mapped to the same reference, the unsigned observed template length equals the number of bases from the leftmost mapped base to the rightmost mapped base. The leftmost segment has a plus sign and the rightmost has a minus sign. The sign of segments in the middle is undefined. It is set as 0 for single-segment template or when the information is unavailable.

10. SEQ: segment SEQuence. This field can be a '*' when the sequence is not stored. If not a '*', the length of the sequence must equal the sum of lengths of `M/I/S/=/X` operations in CIGAR. An '=' denotes the base is identical to the reference base. No assumptions can be made on the letter cases.

11. QUAL: ASCII of base QUALity plus 33 (same as the quality string in the Sanger FASTQ format). A base quality is the phred-scaled base error probability which equals $-10 \log_{10} \Pr\{\text{base is wrong}\}$. This field can be a '*' when quality is not stored. If not a '*', SEQ must not be a '*' and the length of the quality string ought to equal the length of SEQ.

BAM Files BAM files are SAM files that have been compressed with block gzip compression[98]. If the reads inside the BAM file are sorted in increasing order of reference coordinate, the BAM file supports random access via a supplementary BAM index (**.bam.bai*) file. Each block inside a BAM file is a separate gzip archive up to 64Kb in size with extra metadata to encode the read positions contained inside the block. The BAM index file contains offsets into the BAM file that correspond to particular genomic coordinate ranges. The BAM file this provides a binary compressed structure that supports indexed search on genomic coordinates. No other indexing, including by ID, is currently supported. The majority of DNA sequencing data in the world is currently stored as BAM files.

VCF Files

The Variant Call Format (VCF)[32] is a type of text file that was created in the context of the 1000 Genomes project to allow the representation of various types of genetic variation that are discovered via NGS experiments. VCF files describe variants in a refererence-relative manner - i.e. all genetic variation is shown with respect to a given reference genome build. VCF files have a tabular form and allow one to describe genetic variants as they occur in (or are absent from) a entire cohort of samples using a single file. This file format, despite several significant limitations, has become the widely adopted standard for representing genetic variation. All descriptions in this section have been reproduced or adapted from the 2011 Danecek et al. paper and the VCF specification ()

A VCF file consists of the following sections - Meta-information section describes the format and content of the data contained in the file, header section specifies the columns present in the file, data section contains the actual variant calls that are being described.

Meta-information The meta-information section contains a number of lines that describe the data section. The first line of this section is a *fileformat* line that specifies what version of the VCF spec the file adheres to. There can be any number of lines describing INFO fields. These fields are populated into the INFO column of the data section. A single INFO field may describe a single value or a tuple of values. These values can be Integer, Float, Flag, Character, or String. All of the INFO fields thus described are to be placed into the INFO column as a string of semi-colon-separated key-value pairs. The FILTER field describes any filters that have been applied to the variants. The FORMAT field describes the format of the genotype columns that are specified in the data section. The ALT field describes symbolic alternate alleles, that result from variants that have been called but not accurately genotyped. The *contig* field lists the contigs that the variants specified in the file have been called relative to (typically these are chromosome names from the reference sequence). The SAMPLE field specifies the samples that the variants map to. The PEDIGREE field specifies pedigree relationships between samples in the file.

Header The header section is a single tab-delimited line that lists what columns are present in the body. The mandatory columns are:

- #CHROM
- POS
- ID
- REF
- ALT

- QUAL
- FILTER
- INFO

If the VCF file contains genotypes then the INFO column is followed by a FORMAT column, followed by one column for each present sample where each column name is the respective sample ID and all sample IDs are unique within the file.

Data The data section of a VCF file contains the actual list of variants and their genotypes in all samples in tabular form where the columns align with the header section. The values are tab-separated. Any missing values are indicated with a ”.” (dot). The line contents are as follows:

CHROM - An identifier of the chromosome where the variant resides. The ID of the chromosome should match one of the *contig* entries in the meta-information section. All of the variants that belong to a single CHROM should exist in a single contiguous block of rows in a VCF file.

POS - 1-based position of the variant with respect to the reference chromosome specified in CHROM. Variant positions should be sorted numerically in increasing order.

ID - A semi-colon separated list of unique identifiers, where they exist. If no identifier exists a missing value should be indicated.

REF - the reference bases corresponding to the variant location where each base $b \in \{A, C, G, T, N\}$.

ALT - a comma separated list of alternative alleles. The alleles do not have to called in any of the samples. Value can be either a String of $b \in \{A, C, G, T, N, *\}$ or a missing value (when there is no variant).

QUAL - the PHRED scale variant quality. When ALT is present QUAL is $-10 \log_{10}(novariant)$ and when ALT is missing QUAL is $-10 \log_{10}(variant)$. If QUAL is unknown then missing value must be specified.

FILTER - encodes the filter status. If the variant passes all QC filters the value should be PASS. Otherwise there should be a semi-colon separated list of codes for filters that have not passed. If FILTER information is not available there should be a missing value.

INFO - A list of key-value pairs for the additional fields encoded for each variant as specified in the INFO lines of the metadata section. Keys without corresponding values may be used to indicate group membership. There is a number of frequently used reserved keys (see Table 2.4).

Key	Number	Type	Description
AA	1	String	Ancestral allele
AC	A	Integer	Allele count in genotypes, for each ALT allele, in the same order as listed
AD	R	Integer	Total read depth for each allele
ADF	R	Integer	Read depth for each allele on the forward strand
ADR	R	Integer	Read depth for each allele on the reverse strand
AF	A	Float	Allele frequency for each ALT allele in the same order as listed (estimated from primary data, not called genotypes)
AN	1	Integer	Total number of alleles in called genotypes
BQ	1	Float	RMS base quality
CIGAR	A	String	Cigar string describing how to align an alternate allele to the reference allele
DB	0	Flag	dbSNP membership
DP	1	Integer	Combined depth across samples
END	1	Integer	End position (for use with symbolic alleles)
H2	0	Flag	HapMap2 membership
H3	0	Flag	HapMap3 membership
MQ	1	Float	RMS mapping quality
MQ0	1	Integer	Number of MAPQ == 0 reads
NS	1	Integer	Number of samples with data
SB	4	Integer	Strand bias
SOMATIC	0	Flag	Somatic mutation (for cancer genomics)
VALIDATED	0	Flag	Validated by follow-up experiment
1000G	0	Flag	1000 Genomes membership

Table 2.4: Reserved INFO keys (from https://samtools.github.io/hts-specs/VCFv4.3.pdf).

If genotype information is present, the exact same definition is used for all samples. The FORMAT field specifies a column seprated list of the data types and their order that are present in the genotype columns. This is followed by one data block for each sample that contains the genotype data as described in the FORMAT column. All keys are optional but missing values should be indicated. There is a number of frequently used and reserved keys (see Table 2.5).

Field	Number	Type	Description
AD	R	Integer	Read depth for each allele
ADF	R	Integer	Read depth for each allele on the forward strand
ADR	R	Integer	Read depth for each allele on the reverse strand
DP	1	Integer	Read depth
EC	A	Integer	Expected alternate allele counts
FT	1	String	Filter indicating if this genotype was "called"
GL	G	Float	Genotype likelihoods
GP	G	Float	Genotype posterior probabilities
GQ	1	Integer	Conditional genotype quality
GT	1	String	Genotype
HQ	2	Integer	Haplotype quality
MQ	1	Integer	RMS mapping quality
PL	G	Integer	Phred-scaled genotype likelihoods rounded to the closest integer
PQ	1	Integer	Phasing quality
PS	1	Integer	Phase set

Table 2.5: Reserved genotype keys (from https://samtools.github.io/hts-specs/VCFv4.3.pdf).

The following keys are most important and frequently used:

GT - The genotype, encoded as allele values separated by / for unphased genotypes and | for phased genotypes. The values are 0 for reference allele, 1 for first allele listed in ALT, 2 for the second, and so on (for instance 0/1 for heterozygous variant in a diploid sample). When a call cannot be made the missing value is used (for instance ./.).

GL - Genotype likelihoods. A comma separated list of log_{10} likelihoods for all possible genotypes given the set of REF and ALT alleles at the locus.

GP - Genotype posterior probabilities, in the same order as GL field.

GQ - Genotype Quality in PHRED scale. Probability the genotype call is wrong given that the site is variant.

AD - Allele Depth. Per-sample read depth for each allele.

DP - Read Depth. $\sum_i AD_i$.

See Listing 20 for an example of a VCF file.

Remarks

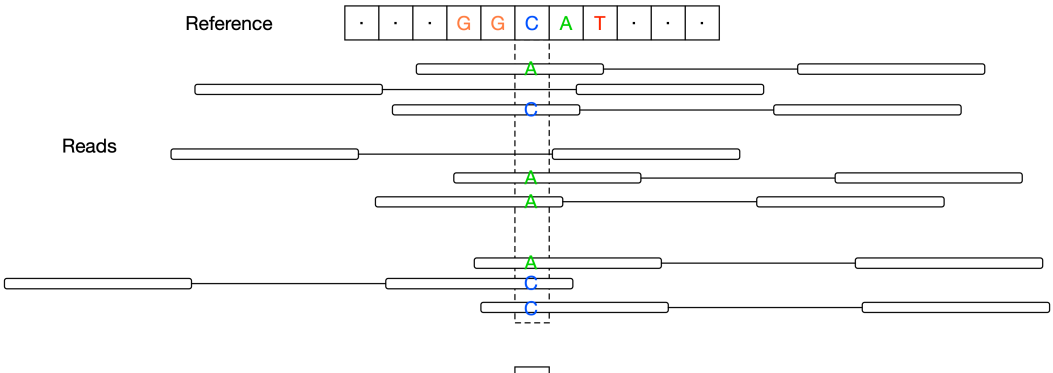
The SAM/BAM and VCF file formats have bencome the de-facto standards for representing sequencing read data and genetic variation respectively. Because of this status most computational tools that exist in this space either consume or produce one of these data-types and the data analysis process is influenced by the file access patterns inherent in these standards. This introduces a number of challenges and limitations that have held back scalability of the existing tools to larger data sets. Specifically:

- The focus on files for information storage and retrieval implies that sophisticated file management schemes must be deployed for successful management of large cohorts of samples. This includes concerns of data security, and data migration that must be implemented at the file-system level.
- Storage of sequence data at sample level of granularity in SAM/BAM files creates very large files that are typically greater than 150 GB in size per sample and are thus quite cumbersome to work with.
- Lack of usage of databases implies only basic indexing and querying schemes are possible for sequencing and variant data. Large amounts of data thus need to be loaded into memory in order to perform basic queries.
- Indexing only by genomic coordinate implies a coordinate-based data traversal mechanisms that process a genome linearly from beginning to the end.
- The multi-sample VCF format makes it easy to interpret all carriers for a single variant (by reading a single row), but is not well suited for interrogating all variants for a single individual (scanning all rows for a single column) in a text file.
- Storage of sequencing data in multiple per-sample files does not take advantage of extreme sequence similarity between samples at a given genomic locus, thus reducing the opportunities for data compression and driving up project costs for large scale sequencing efforts where data set size exceeds 1PB.

The stream-based approach adopted by the Rheos framework desribed in this work attempts to address many of the above issues.

2.2.2 Alignment

Genome alignment (also called mapping) is the process that, given a DNA sequencing read, finds the location in a reference genome that the read best matches to. This is an important process in a next generation sequencing pipeline because it provides a way of ordering the otherwise unordered collection of raw reads (by refrence coordinate) that can be used to find locations where the sample genome is different from the reference (see Figure 2.6).



Locus

A	
C	

Chromosome	2
Position	120238
Genotype	0/1
Genotype Quality	30
Genotype Likelihoods	-323.03,-99.29,-802.53
Ref	C
Alt	A
Read Depth	7
Ref Count	3
Alt Count	4
Sum of Ref Qual	92
Sum of Alt Qual	120

Figure 2.6: Collection of reads aligned to a reference genome and evaluated at a locus that contains a heterozygous SNP.

This process is equivalent to substring search where the string being searched is 3×10^9 characters long and there are 10^9 patterns of length 150 to be found. This process is complicated by the fact that reads from a sample have both sequencing errors and genuine genetic variation that make them differ from the reference, and the fact that certain regions of the genome can be highly repetititve, with the same sequence pattern occurring hundreds of times, making unique mapping challenging.

There are several key applications of alignment that involve sequences of varying length, require different properties, and may not all be best accomplished by the same algorithms. These are:

- Alignment of individual single-end and paired-end reads to a human reference sequence. Where most of the read is expected to match successfully. This is the most abundant use case and the one towards which most alignment algorithms and software implementations are geared.
- Alignment of split-reads to a reference sequence. Split-reads are those where a part of the read maps to one position in the reference sequence and another part of the read maps to a different part of the reference indicating that the read spans a genomic rearrangement. Reads that admit a split alignment may

come out unmapped from a normal alignment stage. This type of alignment is important for both germline structural variant calling and somatic structural variant calling where genomic aberrations are more common.

- Alignment of reads to a group of reference genomes for common pathogens (viruses, bacteria), as well as other common sample contaminating species.

- Alignment of assembled contigs to the reference sequence after local assembly. Some variant calling methods will perform local assembly of reads into a group of potential haplotypes and will perform alignment of these haplotypes to the reference sequence to identify where the actual variants are. The haplotypes may be anywhere from hundreds to millions of bases long.

- Alignment of reads to a group of alternative haplotypes. Variant callers may generate a list of candidate alternative haplotypes and reads need to be aligned to all of these in order to determine which haplotype is best supported by the reads data.

Initially, algorithms for mapping sequencing reads (such as Maq[103] and SOAP[106]) have used a hash table approach[140] (building up tables of subsequences) for finding promising approximate locations for a read within the reference, and then using the Smith-Waterman[160] dynamic programming algorithm for selecting the best matching location. These approaches, however, have proven to be quite slow with typical runtimes exceeding 72 hours on a single high coverage whole genome sample. For the past decade the best available genome alignment approaches in terms of balanced accuracy and speed of processing have been based on the Burrows-Wheeler Transform (BWT)[16] and the FM index[50]. Two of the most popular and widely used are Bowtie[87] and BWA[101]. With recent improvements in computer power, hash table based approaches, such as minimap2[97] and SNAP[190] have made a comeback and are again becoming competitive in the alignment space. A commercial tool called Novoalign (http://www.novocraft.com/products/novoalign/) has consistently been a top performer in terms of speed and accuracy, but the method is not publically available.

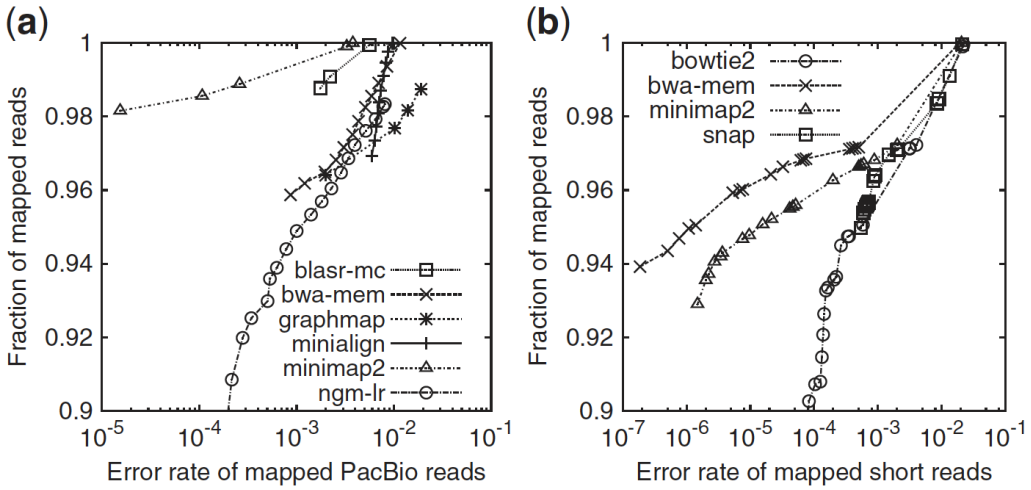


Figure 2.7: Comparison of several aligners on simulated a)long reads >1000 bp, and b) short reads 150bp.[97]

String search is a well studied field, and many genome aligners use the equivalence relationship between suffix trees, suffix arrays, and the BWT (see Figure 2.8) to construct searchable and compressed data structures suitable for genome aligment applications.

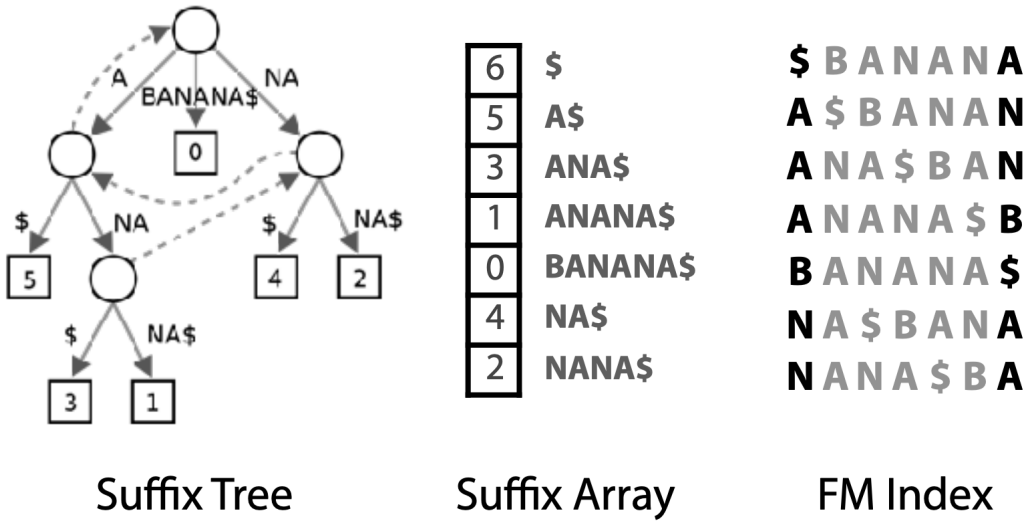


Figure 2.8: Equivalence between a Suffix Tree, Suffix Array, and the BWT matrix of the string BANANA.[1]

Suffix Tree A suffix *trie* T is a tree data structure that given a string $s = c_1, c_2, ..., c_n$ over an alphabet Σ, such that $c_i \in \Sigma$ encodes all of the suffixes of s from the root to the leaves. s is terminated with a sentinel character \$. For any character $c \in \Sigma$ and node $n_i \in T$, n_i has at most one child labeled c. Every leaf is the sentinel character \$. See Figure 2.9 for example.

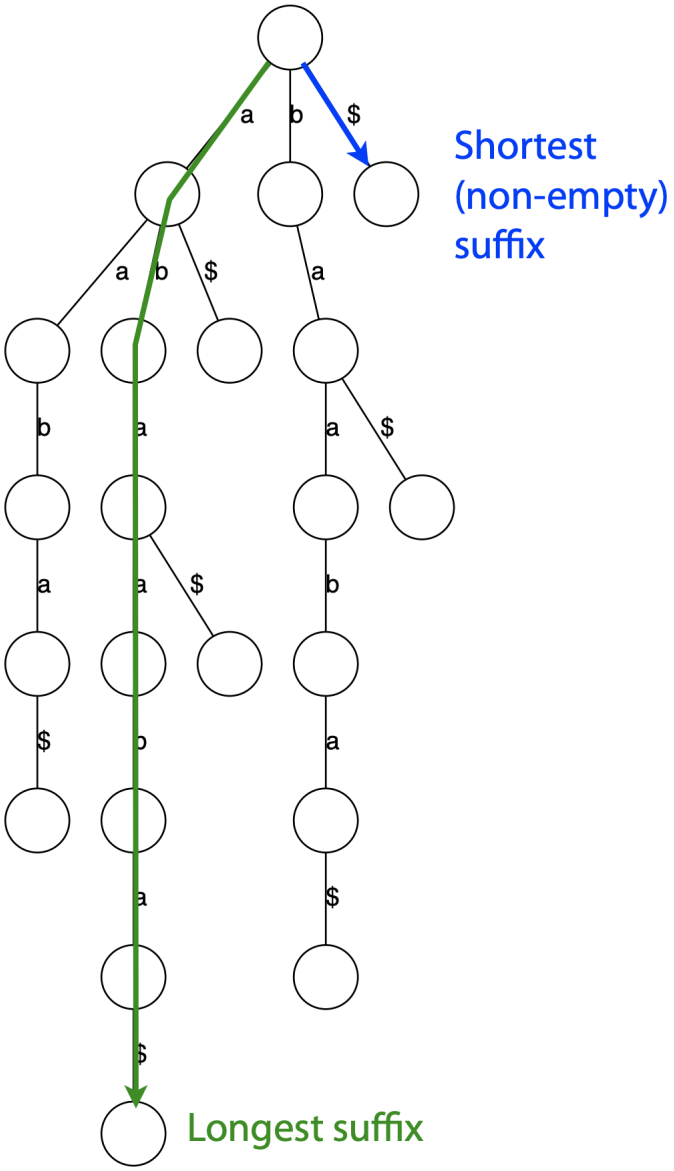


Figure 2.9: Suffix trie of string *abaaba*\$.[1]

Because every substring of s is a prefix of some suffix of s, given a search query q one can check whether q is a substring of s in $\mathcal{O}(|q|)$ time by progressively matching each character of q from the root of T until either all of q is matched (hence q is a substring of T) or at some point a character of q does not have a matching node in T, indicating that q is not a substring of T. The size (number of nodes) of the trie can grow with $\mathcal{O}(|s|^2)$

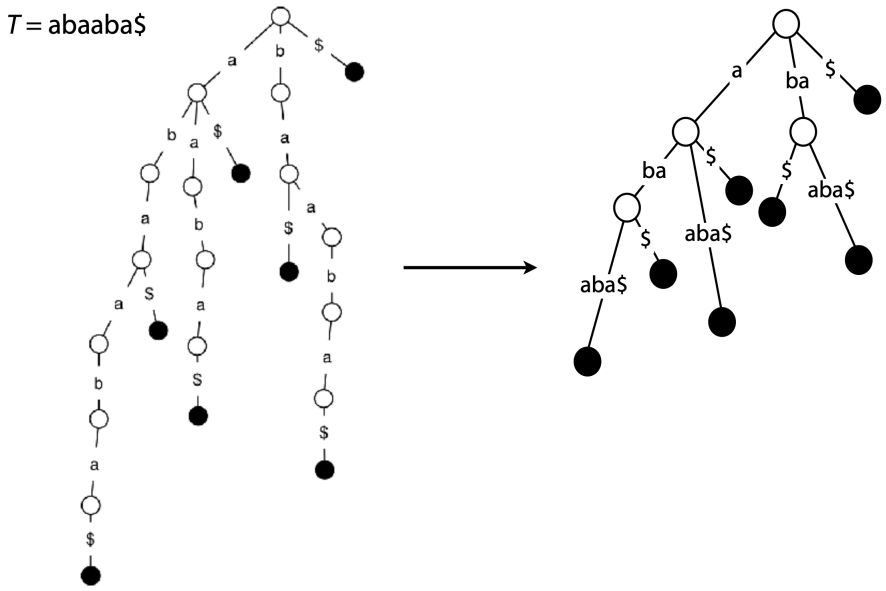


Figure 2.10: Create suffix tree by coalescing non-branching paths into single nodes.[1]

The size of the tree can be reduced by coalescing all non-branching subpaths of T into single nodes (see Figure 2.10).

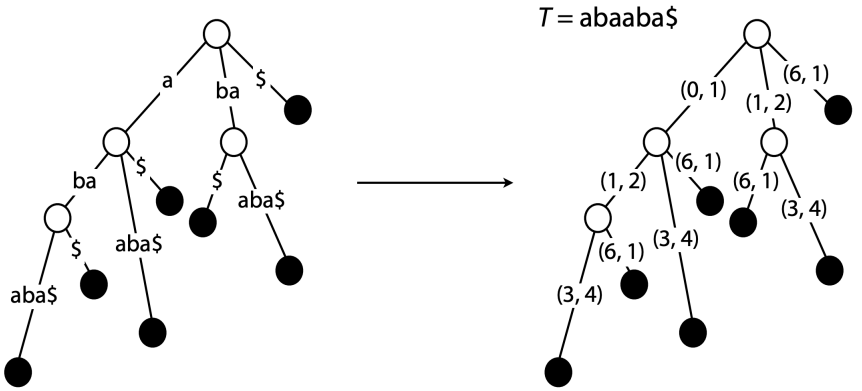


Figure 2.11: Reduce size by replacing substrings by offsets into the original string.[1]

The size can be further reduced to $\mathcal{O}(|s|)$ by replacing string nodes with pairs of offsets into the original string s (see Figure 2.11).

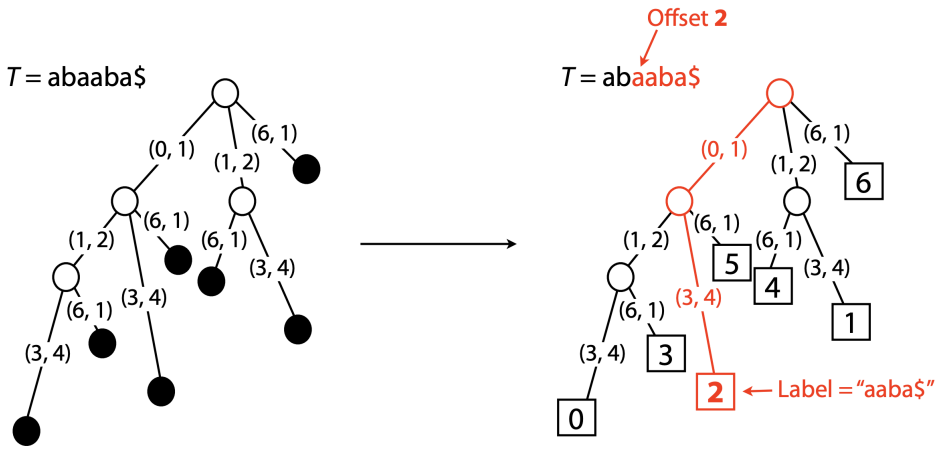


Figure 2.12: Aid search by storing index of substring in leaves.[1]

Searching can be aided by storing offsets of suffixes in the leaf node of the path spelling out that suffix. See Figure 2.12. The online construction algorithm by Ukkonen[169] allows creation of suffix trees in $\mathcal{O}(|s|)$ time and space. Although search against the suffix tree is very fast - $\mathcal{O}(|q|)$, the tree size is still an issue when considering applications to human genome sequencing. A suffix tree for a human would take occupy more than 45 GB of memory[1]. Furthermore, alignment of real reads is made more complex because reads have errors and encode genetic variation, thus inexact searching mechanisms are required.

Suffix Array Suffix arrays were developed by Manber and Myers[111] as a data structure that is equivalent to a suffix tree but occupies less space. Given a string $s = c_1, c_2, ..., c_n$, and letting $s[i, j]$ be a substring of s between indexes i and j, the suffix array SA of s contains integers providing the starting positions of suffixes of s sorted in lexicographical order, s.t. $\forall i \in [1, n] : s[SA[i-1], n] < s[SA[i], n]$. See Figure 2.8 for an exaple suffix array representing the string *banana*. Because the array is sorted, a simple binary search is possible that finds or rules out a match of query q in s in $\mathcal{O}(|q| \log |s|)$, although a more sophisticated search algorithm in $\mathcal{O}(|q| + \log |s|)$ is possible[111]. A suffix array can be constructed by a depth-first traversal of a suffix tree in $\mathcal{O}(|s|)$ time, with the best current algorithm due to Kärkkäinen[82]. Using the array can bring the space requirements for a human-sized genome search index down to about 16GB[1].

The BWT (see Figure 2.13) is constructed from an original string s by appending a special termination character \$ that is lexicographically smaller than all other characters and does not occur in s. Then a matrix of all cyclic rotations of $s\$$ is constructed and sorted in ascending lexicographical order (Figure 2.13 a). The last column of this matrix is the BWT, which can be used to reconstruct the original string and perform substring searches on it. This mechanism relies on the fact that all common prefixes of substrings occupy a contiguous block of rows in the BWT matrix (because of lexicographical ordering, Figure 2.13 c)), and the LF (last-first) property of the BWT. Namely, the character in the last column of the BWT directly

precedes the character in the first column of the BWT in the original string. And, the index of the occurrence of a character in the last column of the BWT is the same as the index of the occurrence of the same character in the first column of the BWT (see Figure 2.13 b)) i.e. the first occurrence of g in the last column corresponds to the first occurence of g in the first column. The third occurrence of a in the last column corresponds to the third occurrence of a in the first column, etc.

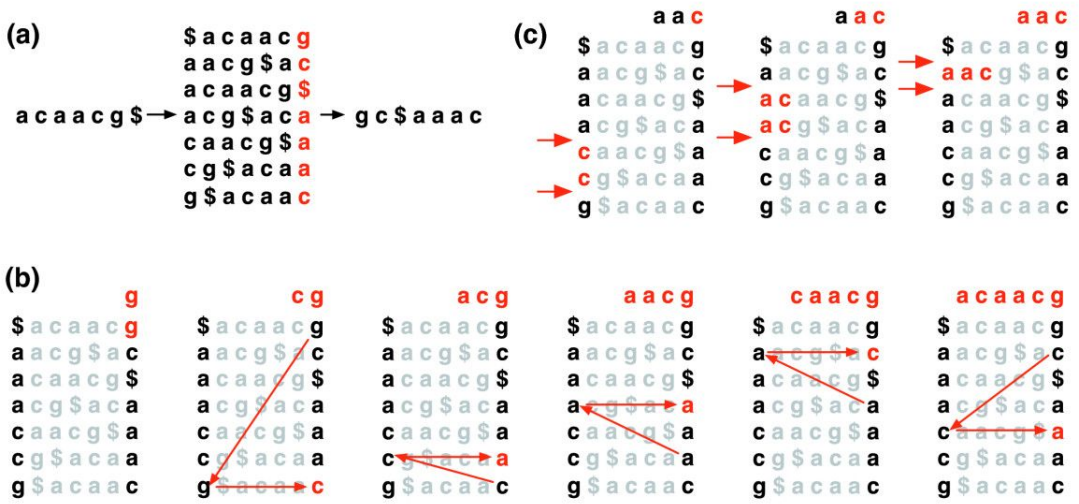


Figure 2.13: a) Creating the BWT of a string. b) All substrings with the same prefix occupy adjacent rows. c) The original string can be recovered from the BWT by using the LF rule.[87]

Alternatively the BWT can be constructed from the suffix array via:

$$BWT[i] = \begin{cases} s[SA[i]-1], & \text{if } SA[i] > 1 \\ \$, & \text{otherwise} \end{cases} \tag{2.1}$$

, per [158]

The entire original string can be recovered from the BWT in the following manner. The last column is the BWT. The first column is the lexicographically sorted BWT. Find the row that starts with \$. The character in the last column in that row is the last character of original string s (g in the example above). Find the first occurrence of that character in the first column. Look up the character in the last column of that row (c in the example above). This is the second-last character of s. Note the index of its occurrence in the last column (it's a second c in the example). Find the row with the same character and the same index in the first column (second c in the first column). The last character in that row is the second-last character of s. Repeat this process until the last character in a row is \$. This recovers the original string.

FM Index

Bowtie Bowtie is a popular aligner developed by Ben Langmead in 2009, with a focus on fast alignment of short reads[87], later followed up and ehnanced with Bowtie 2 for longer reads[85], and adapted massively parallel deployment on the AWS cloud with Crossbow[86]. Since the <50bp long short-reads that the original Bowtie was created for are no longer seen we focus our exposition on Bowtie 2.

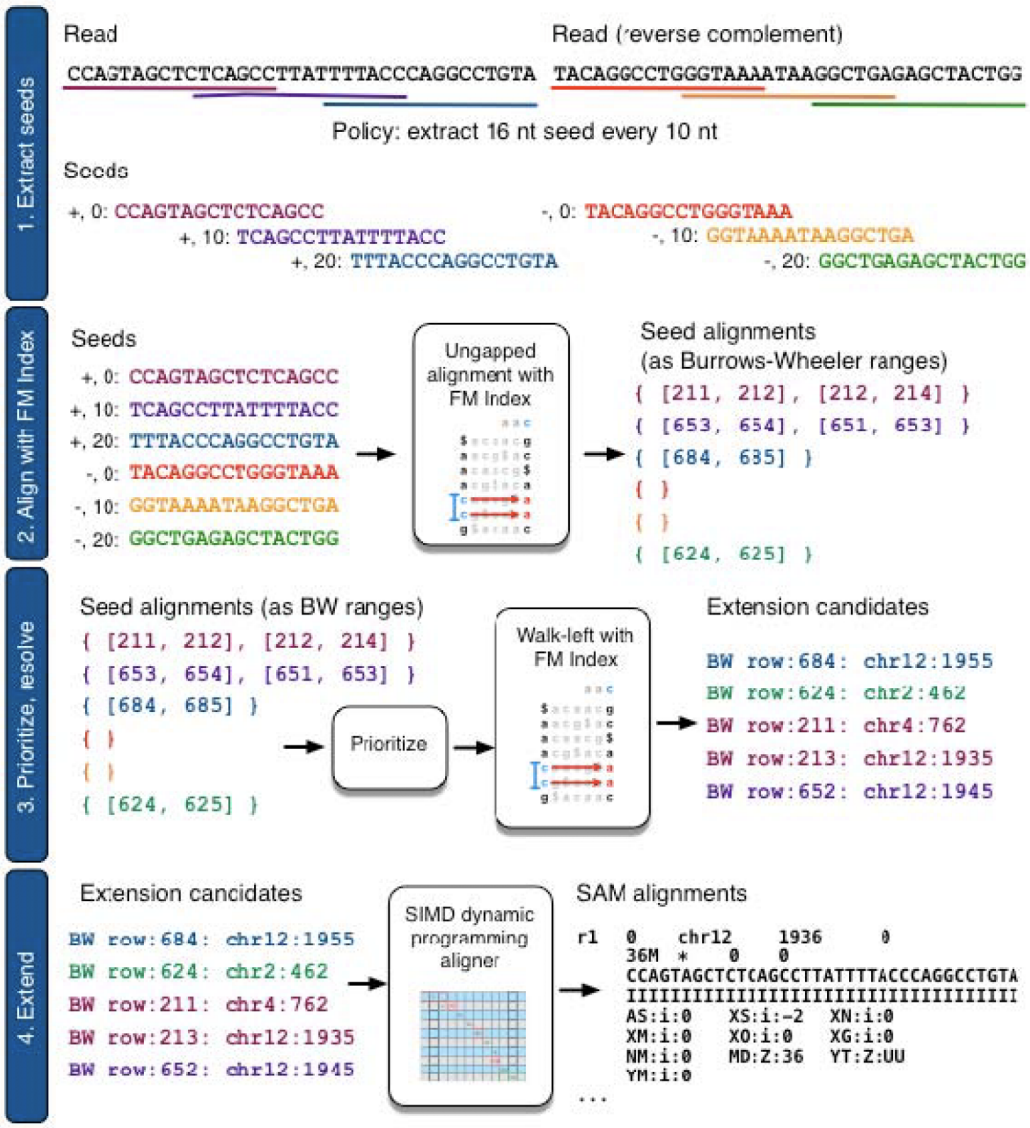


Figure 2.14: S
teps used by Bowtie 2 alignment.[85]

Figure 2.14 shows the sequence of steps used by the Bowtie 2 alignment process. One of the big challenges for aligners is the ability to deal with gapped alignment i.e. sequences in the reads that are inserted or deleted with respect to the reference. These may be the result of sequencing errors or genuine genetic variation present in the sample. Since the read sequence cannot be matched to the reference exactly, gapped alignment drastically increases the search space of possible matches that need to be considered, corresponding to different possible gap sizes. In order to mitigate the degrading effect of gapped alignment on performance, Bowtie adopts a seed-

and-extend mechanism where it performs ungapped exact alignments of substrings of the read that become seeds, followed by a gap-aware dynamic programming-based seed extension to obtain the final alignment. The index is built using a variant of [82] and occupies 2.2 GB of disk space for the human reference.

For alignment every read and its reverse complement are divided into overlapping subsequences that become the potential seeds. Each seed is aligned to the reference using the FM index in an ungapped fashion, but allowing a configurable number of mismatches. For a given read, if the FM index search fails to match the seed to the refrence at some location i it backtracks and attempts a base substitution at that position that can be successfully matched and in a way that maximizes the overall alignment quality. It performs up to k (configurable, default 1) such substitutions per read. Since the substitutions are made in a greedy manner the alignment result may not be globally optimal. To avoid potentially excessive backtracking Bowtie builds both a forward and a mirror index of the reference genome and attempts to match both ends of the read one after the other, with the forward and mirror index respectively attempting to obtain a high scoring alignment of both ends. The number of backtracking steps is further cut off at a hard threshold for performance reasons.

The FM index search produces sets of ranges between which each seed matches. Seeds are assigned priorities based on the number of potential matches, where seeds that have fewer potential matches (i.e. map more uiquely) are given higher priority. Seeds are then selected in priority order for exact resolution. Using the BWT LF property and the FM occurrence array the selected range is resolved to an actual maching location for a given seed. The seeds are then extended in both directions using an adaptation of SIMD-accellerated Smith Waterman alignment via dynamic programming. The striped Smith-Waterman algorithm[48] that Bowtie 2's alignment is based on allows end-to-end alignment in the neighbourhood of the seed with arbitrary numbers of mismatches and gaps and is base quality aware. When reads align to repetitive regions there may be many thousands of match locations for the seeds. To more accurately position these reads Bowtie employs a reseeding strategy where the seeds are selected from a read multiple times using a sliding window in an attempt to find a more accurate seed location.

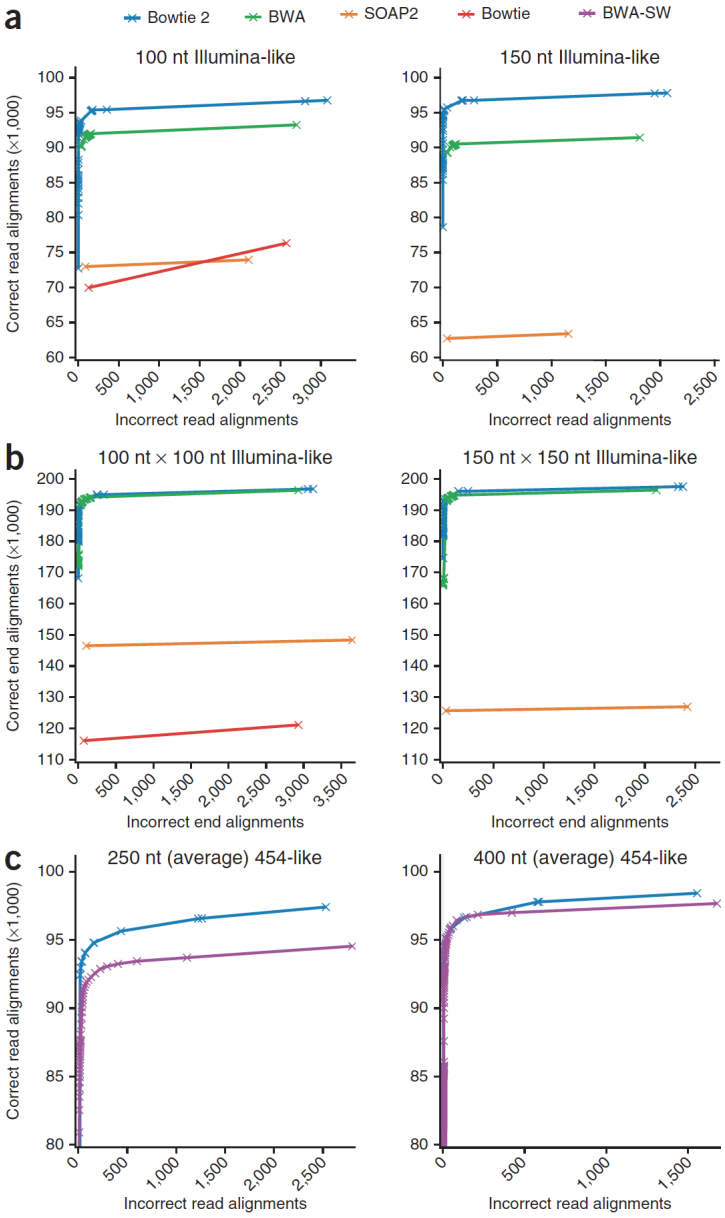


Figure 2.15: Performance comparison of Bowtie 2 to other callers on simulated data a) 100 and 150 bp unpaired, b) 100x100 and 150x150 paired, and c) 250 and 400 long reads.[85]

Bowtie uses information about paired-end reads to attempt to perform a more accurate alignment of the pairs. When the user supplies an expected DNA fragment length and read orientation and one read in a pair is fully aligned, Bowtie calculates a window of reference coordinates where the other read in the pair would be expected to reside and attempts to align the other read in that window using dynamic programing. If it's not able to produce a high quality alignment in this fashion it will attempt to align the other read using the full normal procedure.

A performance comparison by the tool's author Ben Langemead from [85] is shown in Figure 2.15. Bowtie 2 can be seen to be the same or slightly better than BWA and significantly better than Bowtie and SOAP2 on the simulated dataset.

BWA BWA is a tool by Heng Li that implements a series of algorithms for genome alignment based on the BWT and the FM index and described in a series of papers - [101], [100], [94]. Although the original BWA algorithm was focused towards reads that are fewer than 50 basepairs long and is no longer relevant, the subsequent BWA-SW, and BWA-MEM algorithms remain very popular. Since BWA-MEM is seen as the ultimate successor to the previous BWA algorithms we focus our discussion on this implementation of alignement. The following common notation is adopted (see Table **??**)

Table 2.6: BWA Common notation[95],[94]

Symbol	Description
$\Sigma = \{\$, A, C, G, T, N\}$	alphabet of DNA strings with lexicographical order $\$ < A < C < G < T < N$
$\$$	Is a sentinel character
N	Is an ambiguous DNA base
T	String: $T = a_0 a_1 \dots a_{n-1}$ with $a_{n-1} = \$$
$\|T\|$	Length of T including sentinels: $\|T\| = n$
$T[i]$	The i-th symbol in string T: $T[i] = a_i$
$T[i, j]$	Substring: $T[i, j] = a_i \dots a_j$
T_i	Suffix: $T_i = T[i, n-1]$
S	Suffix array; $S(i)$ is the position of the i-th smallest suffix
B	BWT: $B[i] = T[S(i) - 1]$ if $S(i) > 0$ or $B[i] = \$$ otherwise
$C(a)$	Accumul. count array: $C(a) = \|\{0 \leq i \leq n-1 : T[i] < a\}\|$
$O(a, i)$	Occurrence array: $O(a, i) = \|\{0 \leq j \leq i : B[j] = a\}\|$
$P \circ W$	String concatenation of string P and W
Pa	String concatenation of string P and symbol a: $Pa = P \circ a$
$\overline{P}$	Watson-Crick reverse complement of DNA string P

The *suffix array interval* $l^l(P), l^u(P)$ of a string P is defined as:

$$l^l(P) = \min \{k : P \text{ is the prefix of } T_{S(k)}\}$$
$$l^u(P) = \max \{k : P \text{ is the prefix of } T_{S(k)}\}$$
$$l^s(P) = l^u(P) - l^l(P) + 1 \text{ - the interval size}$$

Based on the definition of the FM index:

$$I^l(aP) = C(a) + O(a, I^l(P) - 1) \quad (2.2)$$
$$I^u(aP) = C(a) + O(a, I^u(P)) - 1 \quad (2.3)$$

and $I^l(aP) \leq I^u(aP)$ if and only if aP is a substring of T.

A string that terminates with a \$ is called a *text*. A text may have multiple sentinels and every sentinel has a different lexicographical rank i.e. given a text T if there exist $T[i] = \$$ and $T[j] = \$$, then $T[i] < T[j]$ iff $i < j$. Given an ordered set of texts their ordered string concatenation is called a *collection*. Given a set of DNA texts $R_0, \ldots, R_n$, let $T = R_0\overline{R_0}R_1\overline{R_1}, \ldots, R_{n-1}\overline{R_{n-1}}$ be a bidirectional collection of R. The FM index of T is called the FMD index, and the *bi-interval* of P is $[l^l(P), l^l(\overline{P}), l^s(P)]$. Using the fact that we can compute the suffix array interval of aP via Equation 2.2, and that $I^s(\overline{cP}) = I^s(cP)$ the full bi-interval of aP can be derived. This can be used, as in Algorithms 2 and 1 to bi-directionally extend a substring match P and its complement $\overline{P}$ in either direction.

Algorithm 1: Backward extension[95]

Input: Bi-interval $[k, l, s]$ of string W and a symbol a
Output: Bi-interval of string aW

Function BackwardExt($[k, l, s], a$) **begin**
  for $b \leftarrow 0$ **to** 5 **do**
    $k_b \leftarrow C(b) + O(b, k-1)$ $s_b \leftarrow O(b, k+s-1) - O(b, k-1)$
  $l_0 \leftarrow l$
  $l_4 \leftarrow l_0 + s_0$
  for $b \leftarrow 3$ **to** 1 **do**
    $l_b \leftarrow l_{b+1} + s_{b+1}$
  $l_5 \leftarrow l_1 + s_1$
  return $[k_a, l_a, s_a]$

Algorithm 2: Forward extension[95]

Input: Bi-interval $[k, l, s]$ of string W and a symbol a
Output: Bi-interval of string Wa

Function ForwardExt($[k, l, s], a$) **begin**
  $[l', k', s'] \leftarrow$BackwardExt($[l, k, s], \overline{a}$)
  return $[k', l', s']$

A MEM (*maximal exact match* is an exact pattern match to the index that cannot be extended in either direction, and a SMEM (*supermaximal exact match*) is a MEM that is not contained in other MEMs of the pattern. Building on match extension, Algorithm 3[95] provides a mechanism for finding SMEMs of a read in the reference sequence. BWA-MEM generates SMEMs that cover each position in a read, it then performs greedy SMEM filtering and chaining, linking together SMEMs that are located close to each other and filtering out chains that are contained in other longer chains. The list of seeds generated through this process is ranked by the length of its chain and the length of the seed itself.

Algorithm 3: Finding super-maximal exact matches (SMEMs)[95]

Input: String P and start position i_0; $P[-1] = 0$
Output: Set of bi-intervals of SMEMs overlapping i_0

Function SUPERMEM1(P, x) **begin**
  Initialize Curr, Prev *and* Match *as empty arrays*
  $[k, l, s] \leftarrow [C(P[i_0]), C(\overline{P[i_0]}), C(P[i_0]+1) - C(P[i_0])]$
  for $i \leftarrow i_0 + 1$ **to** $|P|$ **do**
    if $i = |P|$ **then**
      Append $[k, l, s]$ *to* Curr
    else
      $[k', l', s'] \leftarrow$FORWARDEXT($[k, l, s], P[i]$)
      if $s' \neq s$ **then**
        Append $[k, l, s]$ *to* Curr
      if $s' = 0$ **then**
        break
      $[k, l, s] \leftarrow [k', l', s']$
  Swap array Curr *and* Prev
  $i' \leftarrow |P|$
  for $i \leftarrow i_0 - 1$ **to** -1 **do**
    Reset Curr *to empty*
    $s'' \leftarrow -1$
    for $[k, l, s]$ **in** Prev **do**
      $[k', l', s'] \leftarrow$BACKWARDEXT($[k, l, s], P[i]$)
      if $s' = 0$ **or** $i = -1$ **then**
        if Curr *is empty* **and** $i + 1 < i' + 1$ **then**
          $i' \leftarrow i$
          Append $[k, l, s]$ *to* Match
      if $s' \neq 0$ **and** $s' \neq s''$ **then**
        $s'' \leftarrow s'$
        Append $[k, l, s]$ *to* Curr
    if Curr *is empty* **then**
      break
    Swap Curr *and* Prev
  return Match

The seeds are extended using a banded affine-gap-penalty dynamic programming alignment implementation. A number of heuristics is deployed to limit the search space. For paired-end mapping, BWA-MEM performs Smith-Waterman alignment in a window of $[\mu - 4\sigma, \mu + 4\sigma]$ from a mapped read, when its mate in unmapped. When selecting how to match paired alignments BWA-MEM uses a score of $S_{i,j} = S_i + S_j - \min\{-a \log_4 P(d_{i,j}, U\}$, where S_i and S_j are individual read alignment scores, $d_{i,j}$ is the insert distance between two reads, $P(d_{i,j})$ is the probability of observing $d_{i,j}$ under a normal model of insert sizes, and U is a sensitivity threshold. Figure 2.16 shows a comparison performed by BWA-MEM author between BWA-MEM and other popular aligners. BWA-MEM is seen to outperform all aligners except the commercial tool Novoalign on accuracy and be in the top two tools for speed.

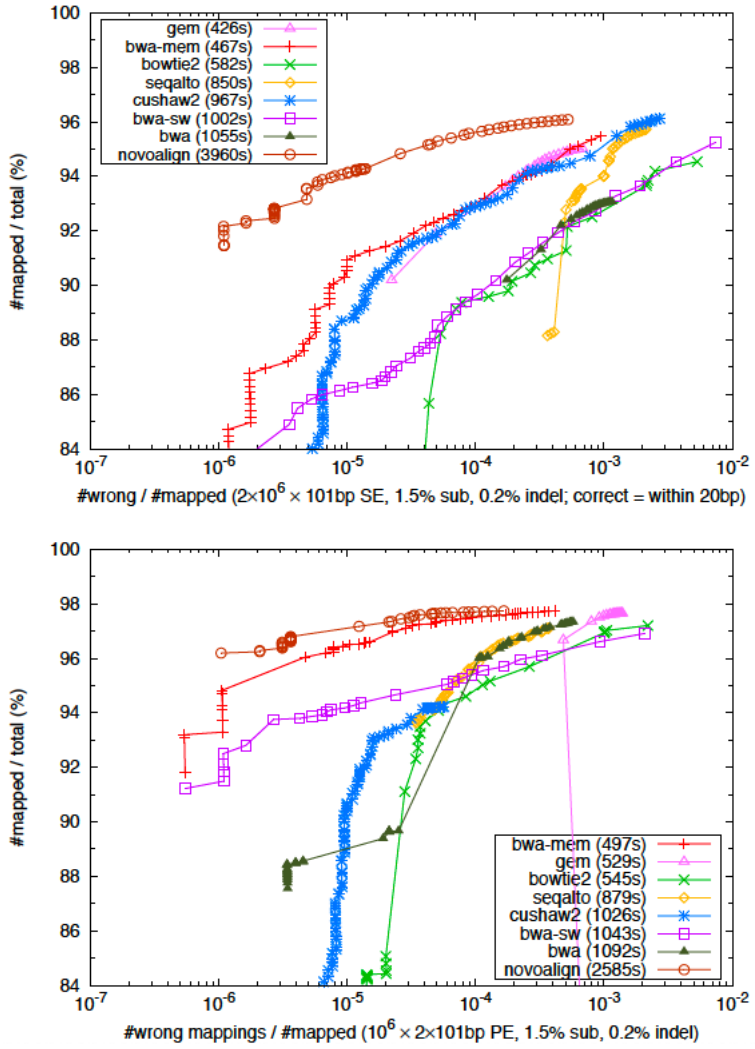


Figure 2.16: P
erformance of multiple aligners on simulated single-end and paired-end 101 bp reads.[94]

Minimap2 Bowtie 2 and BWA-MEM that have been presented so far represent state-of-the-art algorithms for the alignment of short (<500 bp) reads to a reference genome using an FM-index. There are, however, several legitimate cases for the alignment of much longer reads to one or several references, and for finding overlaps between groups of long reads. New DNA sequencing technologies, including Single Molecule Real-Time (SMRT)[141] from Pacific Biosciences Inc., and Oxford Nanopore Technologes' (ONT)[107], have been producing reads that range from 1000 bp to $> 10^6$ bp albeit with a much higher error rate than those from Illumina short reads. These long reads can be used to resolve structural variation encountered in repeat-enriched areas of the genome that are not possible to uniquely resolve with shorter Illumina-based reads[145]. Additionally, tools such as the GATK HaplotypeCaller (see Section 2.2.4) create locally assembled haplotypes (that can be thousands of basepairs long) and align these to the reference sequence to detect variants. Yet, tools that have been developed primarily for short read alignment (like BWA-MEM) either crash or perform exceedingly slow when aligning long read

sequences. Minimap 2[97] has been created to solve the efficient alignment problem for long reads using hashmap-based approach. All of the formulas in this section have been reproduced from the 2018 paper by Heng Li[97].

Minimap2 uses a hashmap approach based on building a database of representative k-mers, called minimizers, for a reference sequnce, and searching against this database. Minmizers have been introduced by Roberts et al.[143] as a way of reducing the storage required for comparing biological sequences. The intuitive notion is that if one is comparing two strings with a significant overlap, extracting and comparing a set of representative k-mers from both strings will produce a match between the k-mers that can then be used as the basis for a seed-and-extend algorithm for more accurate matching. In this case, rather than storing all k-mers, one may store only the minimizers and use them when searching for seeds.

Given a string $T_i = a_1a_2...a_n$ where $a_i \in \Sigma = \{A, C, G, T, N\}$, a k-mer triple is a tuple (s, i, p) that stores the k-mer string s, index i that identifies T_i and p the start index of s in T_i. Assume there exists an order on elements of Σ. If we consider w consecutive k-mers, covering $w+k-1$ letters of T_i, then the smallest k-mer is called a (w, k) minimizer of T_i (See Figure 2.17). If two strings have a substring of length $w + k - 1$ in common, then they have a (w, k) minimizer in common.

Position	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10	11	12
Sequence	2	3	1	0	3	4	3	4	2	6	4	7	2	8	1	4	7	5	1
k-mers	2	3	1					4	2	6	4	7	2	8					
with		3	1	0					**2**	**6**	**4**	**7**	**2**	**8**	**1**				
minimizer			1	0	3					6	4	7	2	8	1	4			
in				**0**	**3**	**4**					4	7	2	8	1	4	7		
bold					3	4	3					7	2	8	1	4	7	5	
	(a)							(b)					2	8	1	4	7	5	1

Figure 2.17: E
xamples of minimizers - a) (5,3)-minimizer, b) (6,7)-minimizer.[143]

Minimap2 builds a list of of minimizers of the reference sequence and stores them in a hash table where the key is the hash of the minimizer string and the value is the list of offsets into the reference where that minimizer is found. For each query sequence that needs to be aligned, minimap2 builds minimizers of the query and searches the reference hash table for them. The matches that are found become alignment anchors and sets of anchors that are in the same order and orientation in the query and the referece are joined into chains. The gaps between anchors in a chain are then filled using dynamic programming with a 2-piece affine gap penalty to produce a final base-level alignment.

On simulated long reads (Figure 2.7 a)), minimap2 is more accurate than other aligners and is up to 30 times faster. On real SMRT reads minimap2 was 70 times faster than short-read aligners. On short reads (Figure 2.7 b)), minimap2 was less accurate than BWA-MEM but was up to 3 times faster. Using data from a synthetic diploid cell-line[104] minimap2 shows a higher false negative rate compared to BWA-MEM (FNR 2.6% vs. 2.3%), but a lower false positives per million bases (FPPM,

7.0 vs. 8.8) for SNPs, and similar performance for indels.

2.2.3 Raw Data QC

The data that is generated as part of NGS experiments can have widely varying quality, and may suffer from systematic biases introduced by the experimental protocol employed, the technology used, as well as individual events such as sample contamination[88, 5, 38, 19]. The effect of low quality data such as sequencing errors on downstream analysis can be quite significant, not only because it may introduce false positive variants in the final output, but also because it can produce a large number of potentially variant sites that need to be evaluated and will consume a large amount additional computational resources to eventually rule out. A wide variety of computational methods exist for examining the data after it has been generated in order to identify and fix or filter out low quality data. The most widely used tools include Picard[137], FastQC[9], QC Toolkit[134], QC-Chain[193], and FASTX-Toolkit[61]. These tools evaluate data at two granularities - read-level, and sample-level. Some of these can act on sequencing data pre-alignment, while others either work post-alignment or actively make use of a built-in aligner. The QC process is generally organized into a QC pipeline (see Figure 2.18). Furthermore, some tools, like FastQC and Picard tend to only collect various QC metrics and generate reports, leaving the filtering based on these metrics up to the user, while other tools, such as QC Toolkit and QC-Chain perform the actual filtering themselves.

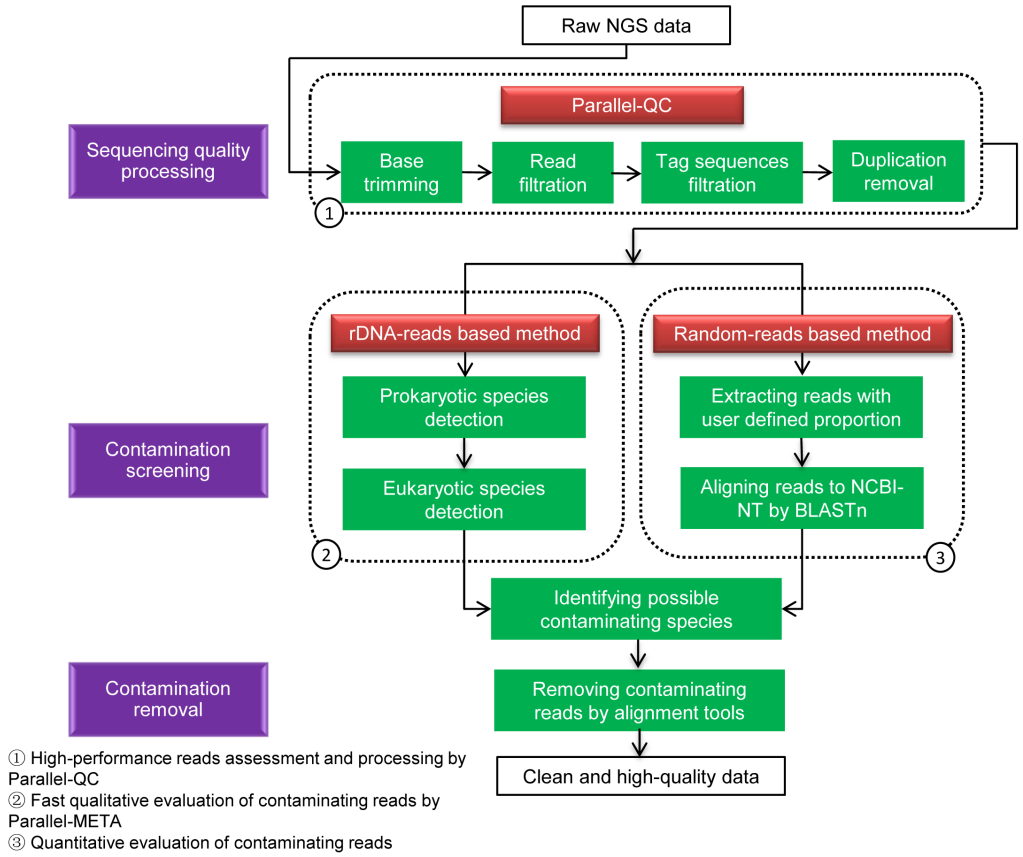


Figure 2.18: The QC-Chain pipeline involves trimming low quality bases, removal of adapter sequences, removal of duplicate reads, and contamination detection.[193].

At individual read level the following QC measures are of interest:

Base Quality Distribution - Reads in FASTQ format have a base-quality score associated with each nucleotide which serve as estimates of the probability that the base has been called correctly by the base-caller software. Bases towards the end of Illumina reads tend to suffer from deteriorating quality([38] and may not be usable for downstream analysis.

Adapter Sequence Presence - The NGS library preparation protocol involves ligating an adapter sequence to the end of DNA fragments in order to bind the fragment to the sequencing flowcell. Sometimes the sequencing process does not stop at the end of the actual DNA molecule and also sequences the adapter. It is necessary to detect and trim out these adapter sequences as they do not belong to the genome[13].

Duplicate Detection - in NGS libraries that use PCR amplification a particular DNA fragment may be sequenced multiple times resulting in increased apparent, and depleted actual coverage of that region of the genome, possibly influencing downstream variant calling efforts. Both Picard[137] and Samtools[93] have utilities for detecting PCR duplicates, although the additional benefit of this processing step may be marginal[40].

Sample Contamination/Sample Swap - Sample contamination during library preparation may result in the presence of foreign DNA in the generated sequence. The foreign sequence may originate from the same or from foreign species. Furthermore, entire samples may be swapped or mislabelled (for instance, cancer samples with tumour from one patient, but normal sample from another patient). Contamination may be detected by aligning reads to a panel of potentially contaminating species' reference genomes[193], or when a surprising number of variants is found in a given genome.

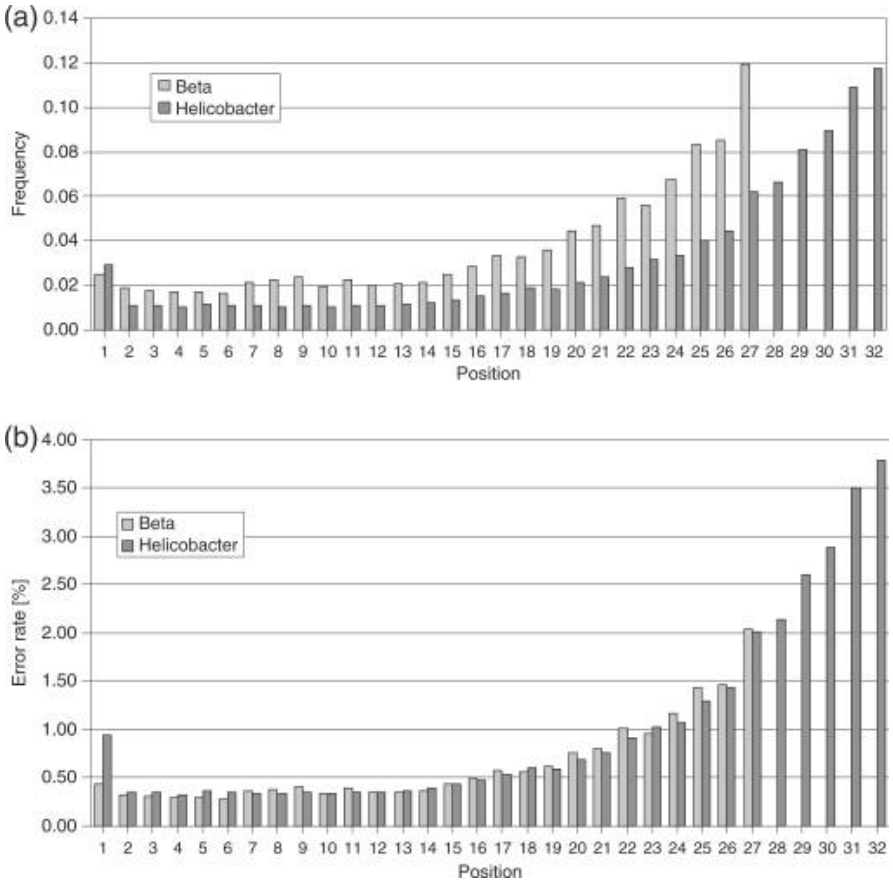


Figure 2.19: Frequency of wrong base calls in Illumina reads based on position in the read from 5' to 3'.[38].

At the sample level the following QC measures are important:

Insert Distribution - summary statistics (mean, median, standard deviation) related to the distribution of the distance between paired-end reads.

Per-base Quality Distribution - distribution of base qualities for each position in a read, aggregated over all reads.

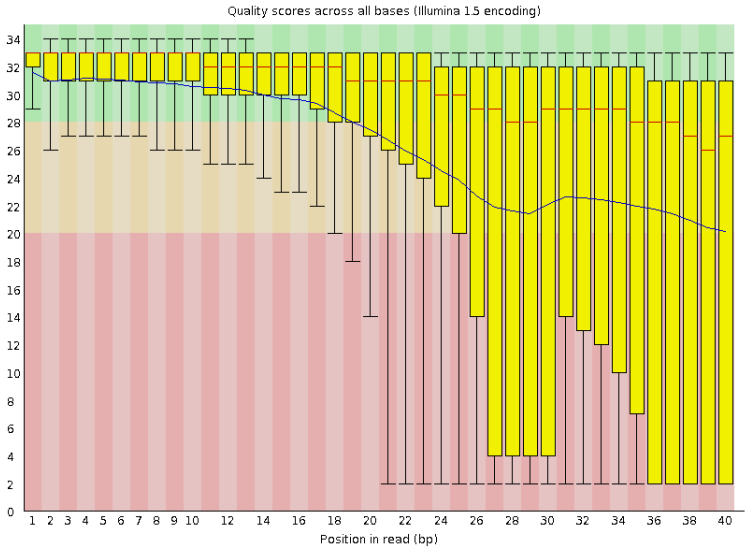


Figure 2.20: from https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Per-read Quality Distribution - distribution of average read qualities.

Per-base Sequence Content Distribution - distribution of nucleotide frequencies for each position in a read, aggregated over all reads.

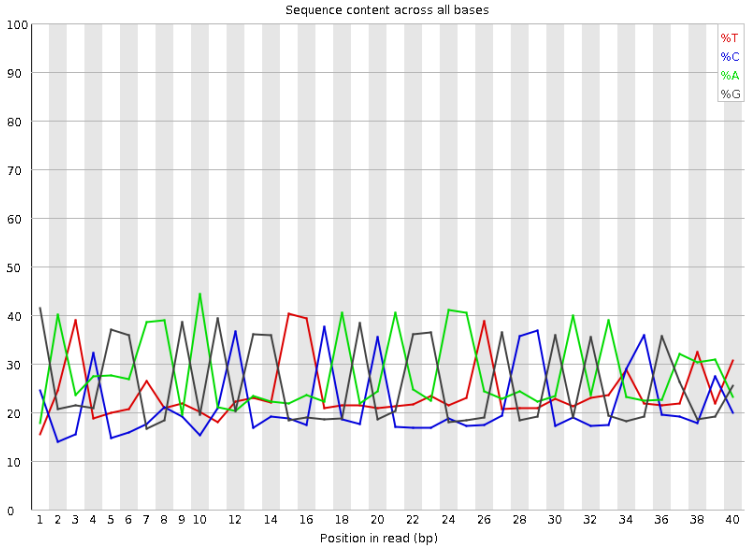


Figure 2.21: from https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Per-read GC Content Distribution - distribution of average GC content per read.

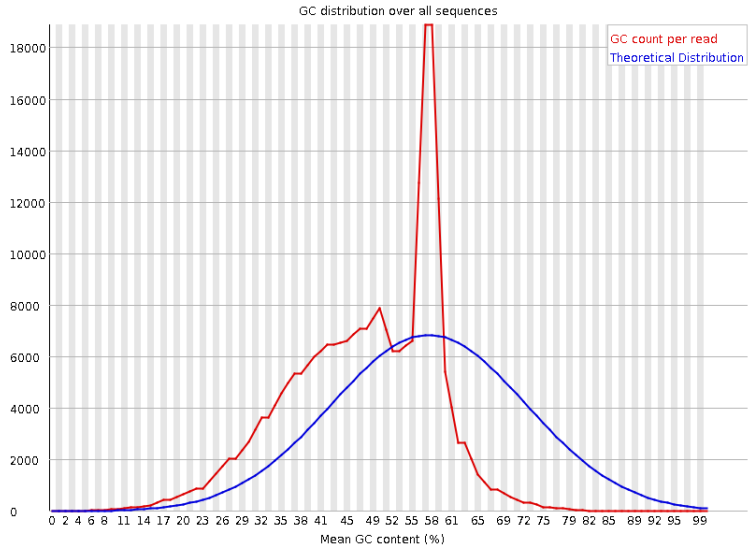


Figure 2.22: from https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Read-length Distribution - summary statistics of read lengths.

Per-base N Content - distribution of uncalled bases for each position in a read, aggregated over all reads.

Sequence Duplication Distribution - distribution of counts for duplicated sequences.

Overrepresented Sequence Distribution - frequency of sequence fragments that occur more frequently that expected.

Per-base Adapter Content Distribution - frequency of adapter sequence presence for each position in a read, aggregated over all reads.

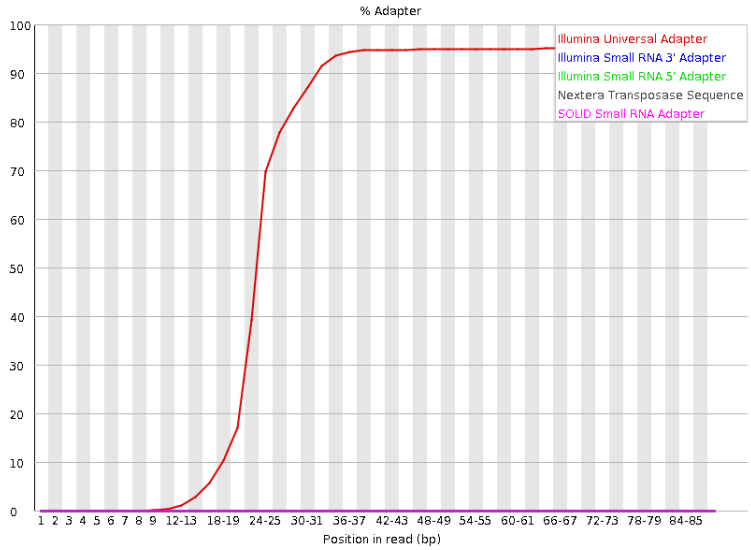


Figure 2.23: from https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Many other QC metrics can be of interest and are listed on the individual tools' pages (for instance https://broadinstitute.github.io/picard/picard-metric-definitions.html). Once collected, these metrics can be used for visualization, manual curation, threshold-based filtering, or machine learning approaches to QC.

2.2.4 Germline SNP Calling

Single Nucleotide Polymorphisms or SNPs are locations in an individual's genome where that individual differs from the reference sequence at a single position. The reference sequence is haploid i.e. it provides a single base (for instance T) at every genomic location, whereas the human genome is diploid (there are two copies of each chromosome, and thus two bases at each location) for chromosomes 1-22, and chromosome X for females, while being haploid for chromosomes X and Y for males. Thus, at each genomic location, the human genome may be:

Homozygous Reference - When both alleles carried by the individual at that location match the reference.

Heterozygous - When one allele matches the reference and one is different from the reference.

Homozygous Alternate - When both alleles are the same and different from the reference.

Multiallelic - When both alleles are different from the reference and are different from each other[72].

SNPs are the most common type of genomic variant, with every individual carrying over 3 million SNPs on average[156]. Furthermore, the presence of certain SNPs is strongly associated with disease[27], where some SNPs are known to be causative[77], while others, are merely associated with a disease phenotype[150]. A large number of scientific studies[71] and clinical practice[186] is thus enabled by efficient and comprehensive characterization of the gamut of human SNPs to assess their contribution to disease risk, see Figure 2.24.

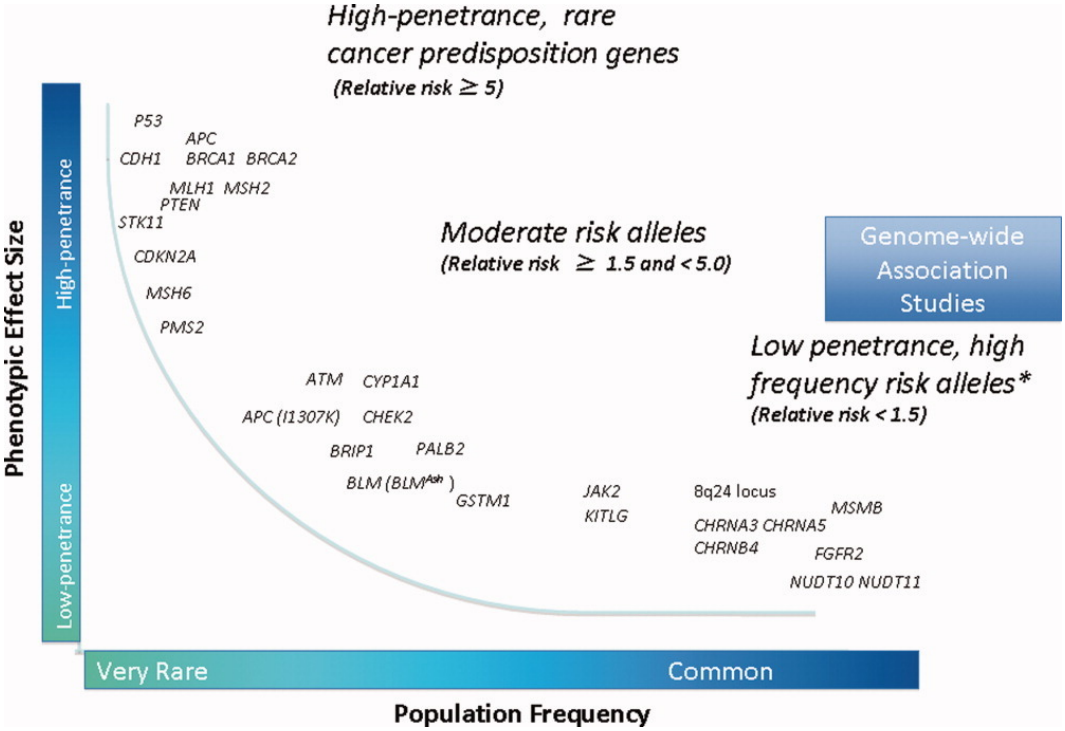


Figure 2.24: Distribution of mutations by population frequency against phenotypic effect size.[180].

There are a number of methods that have been used for assessing SNPs with the aid of microarry technology[69], but here we focus on methods that make use of Next Generation Sequencing (NGS). Since the primary data type generated by NGS is a sequencing read, most presently used methods for SNP detection rely on investigating the collection of sequencing reads that overlap each genomic locus and comparing the observed data to the reference sequence. It is important to distinguish two typically separate activities that take place as part of SNP calling - variant calling, and genotyping. Variant calling attempts to locate positions in the sample genome where that sample is different from the reference, whereas genotyping attempts to assign an actual genotype (e.g. homozygous-alternate), along with a measure of confidence, to each putative variant. We present several of the key computational methods currently used in SNP calling with additional detail. These are:

- samtools
- freebayes
- GATK
- platypus

These tools have been selected because they have been developed independently, at different institutions, and have been repeatedly demonstrated to produce consistent and high-quality results. See Figure 2.25 for a recent comparison.

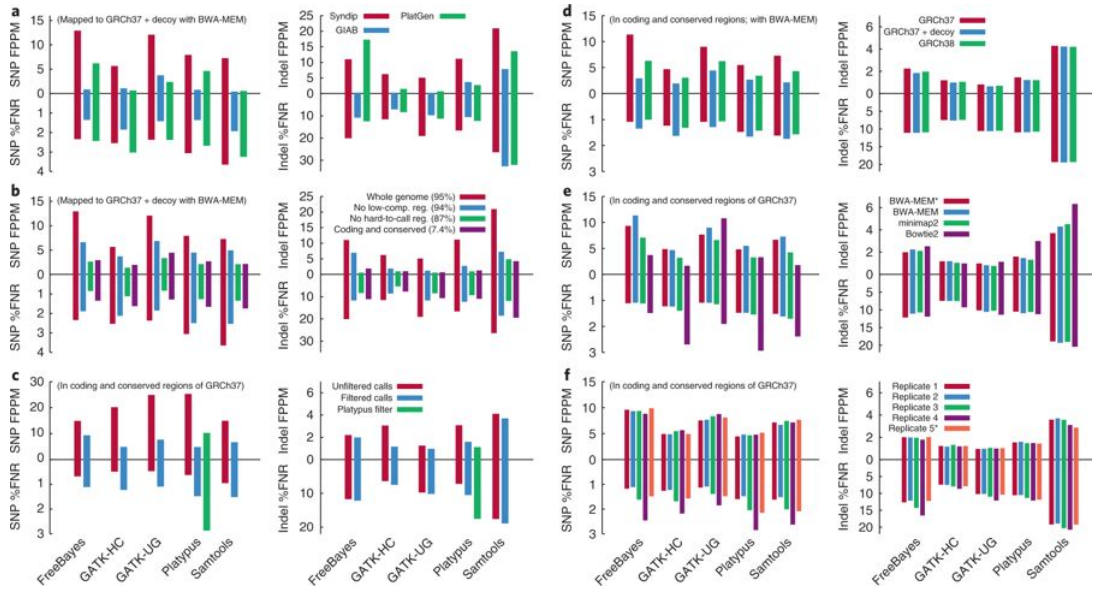


Figure 2.25: Comparison of samtools,freebayes, GATK, and platypus on three benchmark data sets - Syndip, GIAB, and PlatGen. Here FPPM - number of false positives per megabase of sequence, and FNR - false negative rate = 100xFN/(TP+FN).[104].

samtools

Samtools[105] is a software package for genomic data processing developed by Heng Li et al. in the context of the 1000 Genomes Project[24] and implemented as a C program with a CLI interface. This tool has enjoyed continued and widespread use in the bioinformatics community for the purposes of small variant calling (including SNPs). All of the mathematical results in this section are reproduced from the 2011 paper by Li[93] that describes the method, as well as a set of mathematical notes made available separately by Li[96] in 2010.

Although the samtools framework could be extended to support calling multi-allelic sites, the framework, as-published, has been developed for calling only bi-allelic variants. Table 2.7 contains commonly used definitions.

It is additionally assumed that there are n individuals being sequenced with the i-th individual having ploidy m_i (typically 2 in practice). At a particular genomic locus, the sequence read data for the i-th individual is d_i and the genotype is g_i, an integer in $[0, m_i]$, counting the number of reference alleles in the individual at that locus. Furthermore, it is assumed for simplicity that data at individual genomic loci are independent (which isn't necessarily true), as are sequencing and mapping errors between loci and individuals.

Because of the above independence assumptions the joint likelihood function of the data observed for all individuals factors as a product of individual likelihood

Table 2.7: Samtools common definitions

Symbol	Description
n	Number of samples
m_i	Ploidy of the i-th sample ($1 \leq i \leq n$)
M	Total number of chromosomes in samples: $M = \sum_i m_i$
d_i	Sequencing data (bases and qualities) for the i-th sample
g_i	Genotype (the number of reference alleles) of the i-th sample ($0 \leq g_i \leq m_i$)[1]
ϕ_k	Probability of observing k reference alleles ($\sum_{k=0}^{M} \phi_k = 1$)
$P(A)$	Probability of an event A
$\mathcal{L}_i(\theta)$	Likelihood function for the i-th sample: $\mathcal{L}_i(\theta) = P(d_i\|\theta)$

functions:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathcal{L}_i(\theta) \tag{2.4}$$

Suppose that a single sample i represents an individual of ploidy m_i and a given locus is covered by k reads. The sequencing data d_i is composed of an array of bases where each element has value 1 representing the reference allele and is 0 otherwise.

$$d_i = (b_1, \ldots, b_k) = (\underbrace{1, \ldots, 1}_{l}, \underbrace{0, \ldots, 0}_{k-l})$$

The error probability of the j-th base is ϵ_j, which is taken to be the larger between sequencing and mapping errors for that read. Under the independence assumptions above:

$$\mathcal{L}_i(0) = P(d_i|0) = \prod_{j=1}^{l} \epsilon_j \prod_{j=l+1}^{k} (1 - \epsilon_j) = \left(1 - \sum_{j=l+1}^{k} \epsilon_j + o(\epsilon^2)\right) \prod_{j=1}^{l} \epsilon_j \tag{2.5}$$

$$\mathcal{L}_i(m_i) = P(d_i|m_i) = \left(1 - \sum_{j=1}^{l} \epsilon_j + o(\epsilon^2)\right) \prod_{j=l+1}^{k} \epsilon_j \tag{2.6}$$

For $0 < g_i < m_i$:

$$\begin{aligned}
\mathcal{L}_i(g_i) = P(d_i|g_i) &= \sum_{a_1=0}^{1} \cdots \sum_{a_k=0}^{1} \Pr\{d_i|B_1 = a_1, \ldots, B_k = a_k\} \Pr\{B_1 = a_1, \ldots, B_k = a_k|g\} \\
&= \sum_{\vec{a}} \left(\frac{g}{m}\right)^{\sum_j a_j} \left(1 - \frac{g}{m}\right)^{k - \sum_j a_j} \cdot \prod_j p_j(a_j) \\
&= \left(1 - \frac{g}{m}\right)^k \prod_j \sum_{a=0}^{1} p_j(a) \left(\frac{g}{m-g}\right)^a \\
&= \left(1 - \frac{g}{m}\right)^k \prod_{j=1}^{l} \left(\epsilon_j + \frac{g}{m-g}(1-\epsilon_j)\right) \prod_{j=l+1}^{k} \left(1 - \epsilon_j + \frac{\epsilon_j g}{m-g}\right) \\
&= \left(1 - \frac{g}{m}\right)^k \left\{ \left(\frac{g}{m-g}\right)^l + \left(1 - \frac{g}{m-g}\right) \left(\sum_{j=1}^{l} \epsilon_j - \sum_{j=l+1}^{k} \epsilon_j \right) + o(\epsilon^2) \right\}
\end{aligned}$$

where $a = \sum_j a_j$ and

$$p_j(a) = \begin{cases} \epsilon_j & a = 1 \\ 1 - \epsilon_j & a = 0 \end{cases}$$

In particular, for a diploid sample ($m = 2$), the likelihoods for $g = 0, 1, 2$ are

$$\mathcal{L}(0) = \prod_{j=1}^{l} \epsilon_j \prod_{j=l+1}^{k} (1 - \epsilon_j) \tag{2.7}$$

$$\mathcal{L}(1) = \frac{1}{2^k} \tag{2.8}$$

$$\mathcal{L}(2) = \prod_{j=1}^{l} (1 - \epsilon_j) \prod_{j=l+1}^{k} \epsilon_j \tag{2.9}$$

For instance, taking $g_i = 2$ (i.e. the true genotype is homozygous-reference) as an example, and under above independence assumptions, the likelihood of observing the data d_i is the likelihood of sampling l reads without error (the reads match the reference) and $k - l$ reads with error (the reads do not match the reference).

Let $\Phi = \{\phi_k\}$ for $k \in [0, m_i]$ be a prior distribution of genotype probabilities (a model from population genetics, such as Wright-Fisher, can be used, or an empirical distribution from antoher study), the actual genotype for individual i at the given locus is estimated via Bayes' Rule as:

$$\hat{g}_i = \operatorname*{argmax}_{g_i} \Pr\{G_i = g_i|d_i, \Phi\} = \operatorname*{argmax}_{g_i} \frac{P(d_i|g_i)\phi_k}{\sum_{h_i} P(d_i|h_i)\phi_h} = \operatorname*{argmax}_{g_i} \frac{\mathcal{L}(g_i)\phi_k}{\sum_{h_i} \mathcal{L}(h_i)\phi_h} \tag{2.10}$$

The variant quality is defined as:

$$Q_{\text{var}} = -10 \log_{10} \Pr\{G = m_i|d_i, \Phi\} \tag{2.11}$$

i.e. the Phred-scaled probability that the locus is homozygous reference given the observed data. The locus is called *variant* if the variant quality exceeds a certain pre-defined threshold.

Equations 2.10 and 2.11 thus represent the key computations corresponding to the activities of SNP genotyping and variant calling that are performed by samtools for each genomic locus in a sample.

freebayes

freebayes[58] is a C software package implemented by E. Garrison for discovery and genotyping of SNPs, indels, and other small variants that builds on a previous method by G. Marth[116]. Where small means that the variant length is smaller than the size of a sequencing read. Unlike samtools, which looks at all genomic loci independently, freebayes uses local sequence context to guide detection and genotyping of variants, additionally freebayes builds variant haplotypes[31] i.e. groups of variants that are inherited together on the same DNA molecule. All of the figures and methematical formuls in this section are reproduced from the 2012 preprint by Garrison et al[58].

Table 2.8: Freebayes common definitions

Symbol	Description
n	Number of samples
m_i	Copy number of the i-th sample ($1 \leq i \leq n$)
M	Total number of copies of each locus in all samples: $M = \sum_i m_i$
$\{b_j : j \in [1, K]\}$	Set of K distinct alleles present among M copies
$\{C_j : j \in [1, K]\}$	Corresponding set of allele counts for each b_K
$\{f_j : j \in [1, K]\}$	Corresponding set of allele frequencies for each b_K
G_i	Unphased genotype of individual i
$\{b_{i_j} : i \in [1, n], j \in [1, k_i]\}$	Set of k_i distinct alleles of G_i
$\{c_{i_j} : i \in [1, n], j \in [1, k_i]\}$	Set of k_i allele counts for each b_{i_j}
$\{f_{i_j} : i \in [1, n], j \in [1, k_i], f_i = c_i/k_i\}$	Set of allele frequencies of G_i
$B_i : \lvert B_i \rvert = m_i$	Multiset of alleles equivalent to G_i
$R_i = \{r_j : j \in [1, s_i]\}$	Set of s_i sequencing observations for each sample i s.t. there are $\sum_{i=1}^{n} \lvert R_i \rvert$ reads at a given locus
$\{q_j : j \in [1, s_i]\}$	Mapping quality, i.e. probability that read r_j is mis-mapped against the reference

For a set of n individuals the conditional probability of a combination of genotypes

given the observed data is assessed simultaneously as:

$$P(G_1, \ldots, G_n | R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n) P(R_1, \ldots, R_n | G_1, \ldots, G_n)}{P(R_1, \ldots, R_n)} \tag{2.12}$$

$$P(G_1, \ldots, G_n | R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n) \prod_{i=1}^{n} P(R_i | G_i)}{\sum_{\forall G_1, \ldots, G_n} P(G_1, \ldots, G_n) \prod_{i=1}^{n} P(R_i | G_i)} \tag{2.13}$$

For a single sample at a particular locus there are R_i reads, and k_i observed alleles - $B'_i = b'_1, \ldots, b'_{k_i}$, which correspond to $b_1, \ldots, b_i$ underlying alleles represented at the given locus. For each observed allele b'_i there is a corresponding count of observations o_f s.t. : $\sum_{j=1}^{k_i} o_j = s_i$ and each b'_i corresponds to a true allele b_i. The probability of a single observation b'_i given a genotype in a single sample is:

$$P(b'_i | G) = \sum_{\forall (b_i \in G)} f_i P(b'_i | b_i) \tag{2.14}$$

where f_i is the allele frequency of b_i in G. The authors introduce the following approximation:

$$P(b' | b) = \begin{cases} 1 & \text{if } b' = b \\ P(error) & \text{if } b' \neq b \end{cases} \tag{2.15}$$

where $P(error)$ is derived from the base quality score from the sequencing read. Using the above approximation the probability of observing the data R_i given the genotype is estimated as:

$$P(R_i | G) \approx \binom{s_i}{o_1, \ldots, o_{k_i}} \prod_{j=1}^{k_i} f_{i_j}^{o_j} \prod_{l=1}^{s_i} P(b'_l | b_l) \tag{2.16}$$

In order to evaluate the posterior probability of a particular combination of genotypes given the data, the authors derive:

$$P(G_1, \ldots, G_n | f_1, \ldots, f_k) = \binom{M}{c_1, \ldots, c_k}^{-1} \prod_{i=1}^{n} \binom{m_i}{c_{i_1}, \ldots, c_{i_{k_i}}} \tag{2.17}$$

counting the number of ways of selecting a set of unphased genotypes G_i given the allele frequency spectrum f_k, and

$$P(f_1, \ldots, f_k) = P(a_1, \ldots, a_M) = \frac{M!}{\theta \prod_{z=1}^{M-1} (\theta + z)} \prod_{j=1}^{M} \frac{\theta^{a_j}}{j^{a_j} a_j!} \tag{2.18}$$

using Ewens' sampling formula[47], where θ is the population mutation rate and allele frequencies f_i are transformed to frequency counts $a_1, \ldots, a_M : \sum_{c=1}^{M} ca_c = M$ where each a_f is the number of alleles in $b_1, \ldots, b_k$ whose allele count in the sample set is c.

Once variants are initially detected using Equation 2.13, the method continues to build local haplotypes grouping variants that are inherited together in dynamically-sized windows based on a distance threshold between successive variants. Each group of variants is combined into a haplotype observation H_i, with an assigned quality score that is the minimum of the supporting reads' mapping quality and the minimum base quality of the variant allele observations (bases that span the variants). The size of the window is determined via an iterative process where an initial variant is used as the seed, and all of the reads that overlap it are added. If these reads overlap any other variants, then these are added, along with any reads that overlap them, and so on, until no new variants can be added.

Once the window is determined, the method uses a gradient descent algorithm to find a MAP estimate of the genotype for each sample. It starts with the maximum likelihood estimate for each sample's genotype given the observed data, and then attempts to find a genotype assignment that has a higher posteriror probability across all samples.

GATK

The GATK[36] is not actually a tool that's built solely for SNP calling, but is instead a comprehensive framework for genomic data analysis that includes tools for data pre-processing and QA, SNP and indel calling, CNV calling, and SV calling, as well as post-processing and filtering. It is implemented as a Java program and the latest version makes use of the in-memory computing engine Apache Spark for efficient computation over large data sets. Here we focus on the SNP calling aspects of the framework. There are two components in the GATK that deal with SNP calling, the UnifiedGenotyper, and the HaplotypeCaller. The UnifiedGenotyper is an older component that has been superceded by the HaplotypeCaller for all practical purposes and we focus our attention on it. All of the mathematical formulas and figures, as well as some of the descriptions, in this section have been reproduced from the 2010 manuscript by McKenna et al.[119], the 2011 manuscript by DePristo et al.[36], and the GATK website[2]

The GATK Best Practices pipeline (see Figure 2.26) is a set of best practices published by The Broad Institute that describe how to best use their software. In the context of this section we are primarily interested in the processes that occur in the middle panel of this figure that deal with processing aligned reads to call and genotype variants that will then be used for post-processing and filtering.

The HaplotypeCaller which is the tool that is primarily used for SNP calling takes a continued refinement approach, where the data is processed in the following sequential steps:

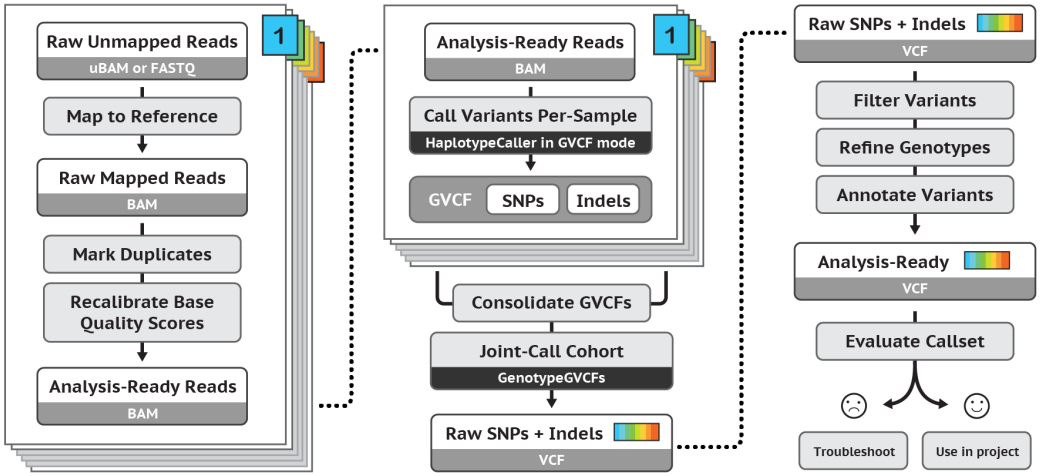


Figure 2.26: GATK Best Practices pipeline[2]

Define active regions - The program determines which regions of the genome it needs to operate on (active regions), based on the presence of evidence for variation.

Determine haplotypes by assembly of the active region - For each active region, the program builds a De Bruijn-like graph to reassemble the active region and identifies what are the possible haplotypes present in the data. The program then realigns each haplotype against the reference haplotype using the Smith-Waterman algorithm in order to identify potentially variant sites.

Determine likelihoods of the haplotypes given the read data - For each active region, the program performs a pairwise alignment of each read against each haplotype using the PairHMM algorithm. This produces a matrix of likelihoods of haplotypes given the read data. These likelihoods are then marginalized to obtain the likelihoods of alleles for each potentially variant site given the read data.

Assign sample genotypes - For each potentially variant site, the program applies Bayes' rule, using the likelihoods of alleles given the read data to calculate the likelihoods of each genotype per sample given the read data observed for that sample. The most likely genotype is then assigned to the sample.

Active regions are defined by targeting loci with high quality reads that are different from the refernce and surrounded by soft-clipped bases (bases in the read that couldn't be aligned by the alignment algorithm). The region is then smoothed by a Gaussian kernel and passed onto variant calling if its profile score exceeds a pre-defined threshold.

For each active region the GATK performs local assembly by first building a de-Bruijn graph[23] using just the reference sequence for the active region. The graph is enhanced by comparing each read from the active region and creating nodes in the graph where the read differs from the graph. Edge weights between pairs of nodes are increased when a read passes through that edge. The resulting graph must

meet complexity requirements based on the ratio of non-unique to unique kmers and presence of cycles (that are a sign of repetitive sequence). When the graph does not meet complexity requirements it is automatically rebuilt with successively increased kmer sizes. If a sufficiently complex graph cannot be built, the region is discarded because the method is not able to produce a high quality variant call in this case. Once the graph is built it is refined by pruning out paths that are not supported by a sufficient number of reads. By traversing the paths in the graph and selecting those with a high aggregate weight the algorithm builds a list of possible local haplotypes within the active region. Each candidate haplotype is then re-aligned to the reference sequence using Smith-Waterman[160] alignment to produce a list of potential variant sites that includes SNPs and other potential variants.

Reads from the active region are aligned to each haplotype using a Hidden Markov Model via the PairHMM algorithm[39] to produce a likelihood of each read given a haplotype. The set of haplotype likelihoods is transformed to a set of likelihoods per allele where for each allele at a given site the the highest scoring likelihood for a given read and a given haplotype that supports that allele is selected. Under a read independence assumption the total likelihood of the allele is computed as the product of read likelihoods.

For each variant locus the probability of each genotype is calculated using:

$$P(G|D) = \frac{P(G)P(D|G)}{\sum_i P(G_i)P(D|G_i)} \tag{2.19}$$

where D is the set of reads, G_i is the set of possible genotypes, and under the assumptions of a diploid sample, independent read observations, and independent errors. The prior probability of genotypes $P(G)$ is assumed to be uniform.

Based on the diploid and independence assumptions:

$$P(D|G) = \prod_j \left(\frac{P(D_j|H_1)}{2} + \frac{P(Dj|H_2)}{2} \right) \tag{2.20}$$

where $P(D_j|H_n)$ is the likelihood of read given the haplotype perviously obtained.

The genotype with the highest $P(G|D)$ is emitted as the genotype for that variant.

Platypus

Platypus[142] is a Python and C package by Rimmer et al., that provides SNP, indel, and other small and medium (upt to 1kb) variant calling capabilities utilizing a Bayesian statistical framework built on top of reference-free local assembly of haplotypes. All of the figures and methematical formuls in this section are reproduced from the 2014 Nature Genetics manuscript by Rimmer et al. See Figure 2.27 for the general sequence of data processing steps in platypus.

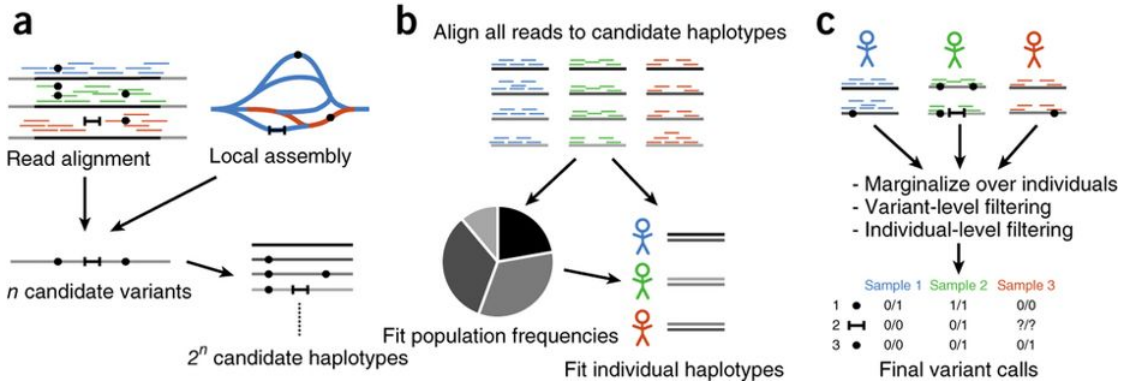


Figure 2.27: Three stages of data processing in Platypus[142]

Platypus works by loading batches of reads into memory from a BAM file. 100 kb (kilobases) are loaded at a time. Low quality reads are filtered out. Sites that are flagged as potential variants by the aligned become candidate variants. The local assembly step operates on successive windows of 1.5 kb in size, including all reads that map to the window as well as their mates, irrespective of where or if they map. As a first step, the algorithm builds a colored de Bruijn graph[78] from the reads and the reference sequence. Candidate variant haplotypes are produced by finding paths that diverge from the reference and come back to the reference by performing a Depth First Search from the graph node that diverges and until the reference is reached again. All such paths are recorded and contain putative SNPs, MNVs, indels, and larger rearrangements that are up to the window-size in length. A prior of 0.33×10^{-3} is used for SNPs under the assumption that SNPs occur in humans at a density of 10^{-3} per base and every non-reference allele is equally likely. Smaller windows are created by combining groups of candidate variants that are within 15 basepairs of each other, as long as window size is smaller than read-length and there are fewer than 8 variants in a group.

Candidate haplotypes are generated by taking all possible combinations of candidate variants (assuming diploid sample), resulting in 2^n haplotypes. For a window with 8 variants there are 256 possible haplotypes. The haplotype likelihood $P(r|h)$ is calculated by aligning reads to each haplotype using a Hidden Markov Model. Platypus uses the Viterbi algorithm[52] to compute the most likely path through the HMM as an approximation of the actual likelihood as a matter of optimization. After haplotype likelihoods are computed for all combinations of reads and haplotypes Platypus uses Expectation Maximization to estimate the frequency of each haplotype under the following model:

$$\mathcal{L}(R|\{h_i, f_i\}_{i=1...a}) = \prod_{\substack{samples,\\ s}} \sum_{\substack{haplotypes,\\ i,j}} f_i f_j \prod_{\substack{reads,\\ r \in R_s}} \left(\frac{1}{2} p(r|h_i) + \frac{1}{2} p(r|h_j) \right) \tag{2.21}$$

where f_i is frequency of haplotype h_i, a is the number of considered alleles, R is all reads, R_s reads in sample s.

The posterior probability of a variant v given the data is computed as:

$$P(v|R) = \frac{P(v)\mathcal{L}(R|\{h_i, f_i\}_{i=1...a})}{P(v)\mathcal{L}(R|\{h_i, f_i\}_{i=1...a}) + (1 - P(v))\mathcal{L}(R|\{h_i, \frac{f_i}{1-F_v}\}_{i \in I_v})} \tag{2.22}$$

here the likelihood of data R given all haplotypes is compared to the likelihood of R given those haplotypes that do not include v. I_v is the set of haplotype indices such that h_i does not contain v, and $F_v = \sum_{i \in I_v} f_i$. A variant is called when its posterior exceeds a pre-defined threshold. Genotype likelihoods for a variant are calculated by marginalizing over the genotypes at other variants within the window, and the best likelihood is selected as the genotype.

Remarks

Although samtools, freebayes, GATK, and Platypus are not the only tools that have been successfully used for germline SNP calling in the past 10 years, they have enjoyed sustained popularity and continued use in large scale genomics projects such as 1000 Genomes Project, ExAC[92], PCAWG[161] and others. They thus can serve as primary examples of how the problem of calling germline SNPs is currently approached within the field, as well as providing a benchmark for new methods that attempt to solve similar problems. There are several key distinctions in the approaches that these tools take to the overall problem, yet a number of key similarities exist that we outline here and adopt as a general framework for tackling the SNP calling problem in our work.

Site Selection The problem of selecting sites in the genome for detecting variants has variety of approaches from evaluating every single site independently as one that potentially harbors a variant (as in samtools) to the varied windowing approaches used by the other variant callers. Looking at sites independently has the benefit of simplicity, while the windowing approach, at the expense of additional computation, allows a more comprehensive evaluation of a genomic region that is not limited to SNPs but can also support the detection of other types of variants. Thus, even though an independent site approach may be acceptable in an initial implementation, some form of windowing is desired in order to benefit from knowledge about the surrounding sequence context. In general, the interestingness of a site for the purposes of detecting a potential variant is universally linked to the presence of high quality reads spanning a particular locus that disagree with the reference sequence at that locus.

Haplotype Construction Samtools has the simplest model here as it does not attempt to construct haplotypes at all. This limits the ability of samtools to accurately represent genomic architecture, and prevents it from being able to supply

phasing information for variants, which is of interest. Freebayes has the next simplest model, where putative haplotypes are constructed directly from the observed read sequences with the observation window. GATK and Platypus actually perform local assembly in the window to come up with an arrangement of reads that is free of artifacts[99] associated with alignment to the reference. Although local assembly appears to improve the ability of these callers to accruately represent sequence variants (especially non-SNPs) this processing step introduces a significant impact on the overall processing cost, mostly incurred from the resource-intensive HMM-aided alignment of reads to the putative haplotypes.

Allele Frequency Spectrum Estimation A distribution of allele frequencies in a population is of interest as it can be used as a prior in the calculation of genotype likelihoods for the samples under analysis and local and population-specific allele frequencies can vary significantly from values implied by generic population-genetics models. Most callers provide capabilities for estimating the Allele Frequency Spectrum by Expectation Maximization or similar approaches (Equation (5) in [93], not covered here, equation 2.18 for freebayes above). Alternatively, a non-informative prior (as in GATK) can be used.

Genotyping Pretty universally across the methods, genotyping is set up as a Bayesian inference selecting a Maximum Likelihood Estimate or a Maximum A Posteriori estimate of the genotype (for a single site or a haplotype) given the reads data, and taking into account the probability of errors dervied from read base quality and mapping quality scores supplied by the aligner. See Eq. 2.10 for samtools', Eq. 2.13 frebayes', Eq. 2.19 GATK's, Eq. 2.21 Platypus' model setups. This model thus remains highly relevant for any new development in the space.

2.2.5 Germline Indel Calling

2.2.6 Germline Structural Variant Calling

Structural Variants (SVs) are medium- to large-size alterations (typically >50bp in length) of the genomic sequence that fall into a several broad categories (see Figure 2.28) including insertions, deletions, tandem and interspersed duplications, inversions, translocations, mobile element insertions, as well as complex rearrangements that constitute a combination of the above classes or are otherwise difficult to classify.

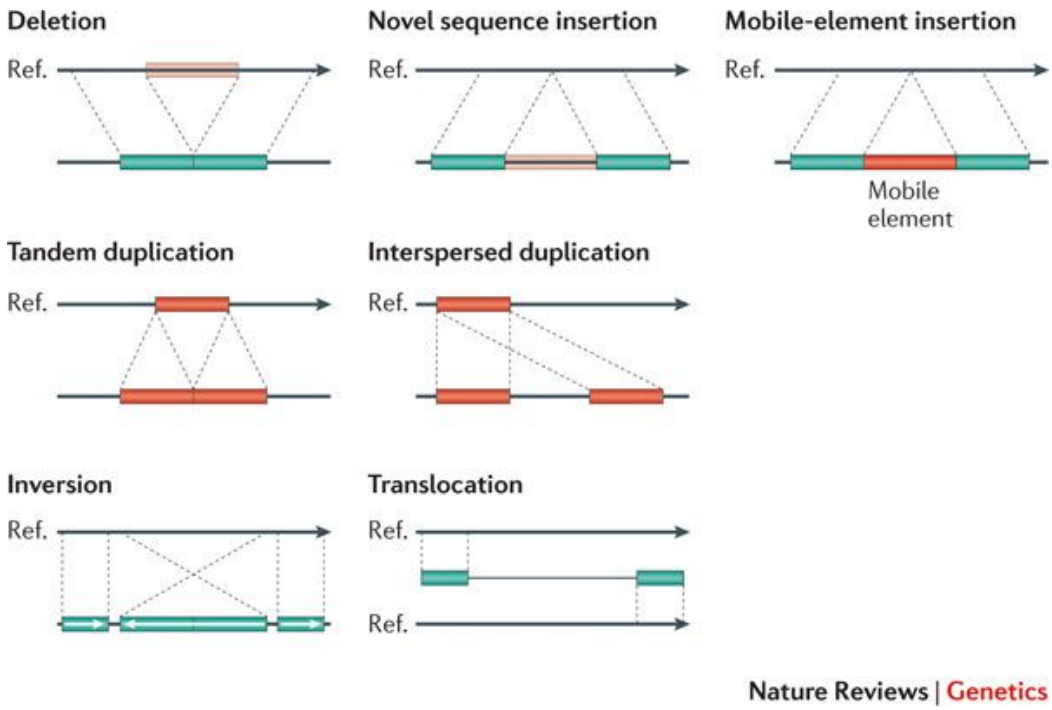


Figure 2.28: Classes of structural variation.[7].

Accurate detection of SVs remains an open challenge since both sensitivity and specificity performance of SV calling is an order of magnitude worse than for SNP calling. For instance, in the latest data release of the 1000 Genomes Project an estimated sensitivity of 88% was achieved for deletions, 65% for duplications, 32% for inversions[165] and False Discovery Rate of up to 89%autocitemills2011mapping. A key challenge that impacts calling accuracy is small read size relative to the size of the variants. Since SVs can be hudreds of kilobases in length and typical Illumina reads are only 150-500 baspeairs long, accurate reconstruction relies on an agglomeration of reads in order to produce the variants. Structural Variant detection approaches typically make use of several sources of evidence (see Figure 2.29) for the detection of "breakpoints" - regions of the genome where a DNA double-strand break is detected and sequence is either inserted or excised. The breakpoints are then interpreted to produce the most likely variants that they represent.

Discordantly-mapped read pairs are pairs of sequencing reads that the alignment algorithm has mapped at a distance that is statistically significantly larger or smaller than the average read-pair distance for that sample, or that have an unusual read orientation, since read-pairs are supposed to be sequencing from two ends of a molecule towards each other. Reads that map closer than they are supposed to are indicative of an insertion, reads that map farther are indicative of a deletion, and reads that map in an unusual orientation are indicative of an inversion.

Split-reads are read-pairs where one of the reads maps properly but the aligner is not able to map the other read because different pieces of the read map to different locations in the genome indicating that this read spans a breakpoint. Distance

between the split read pieces and their mapping orientation are informative of the type of breakpoint that the read spans.

Read-depth Regions have a higher than normal or lower than normal read depth (count of reads spanning a region) are indicative of increased or reduced copy number respectively. Care must be taken to distinguish actual SVs from areas of repetitive sequence where the same simple sequence pattern is repeated many times. Typically aligners are not able to accurately resolve such areas and map many reads to the same set of coordinates making it appear like a duplication.

Assembly-based approaches perform local assembly in a reference-free manner to reconstruct the variants encoded in the sample sequence that the aligner may struggle with resolving because of a large number or increased complexity of rearrangements.

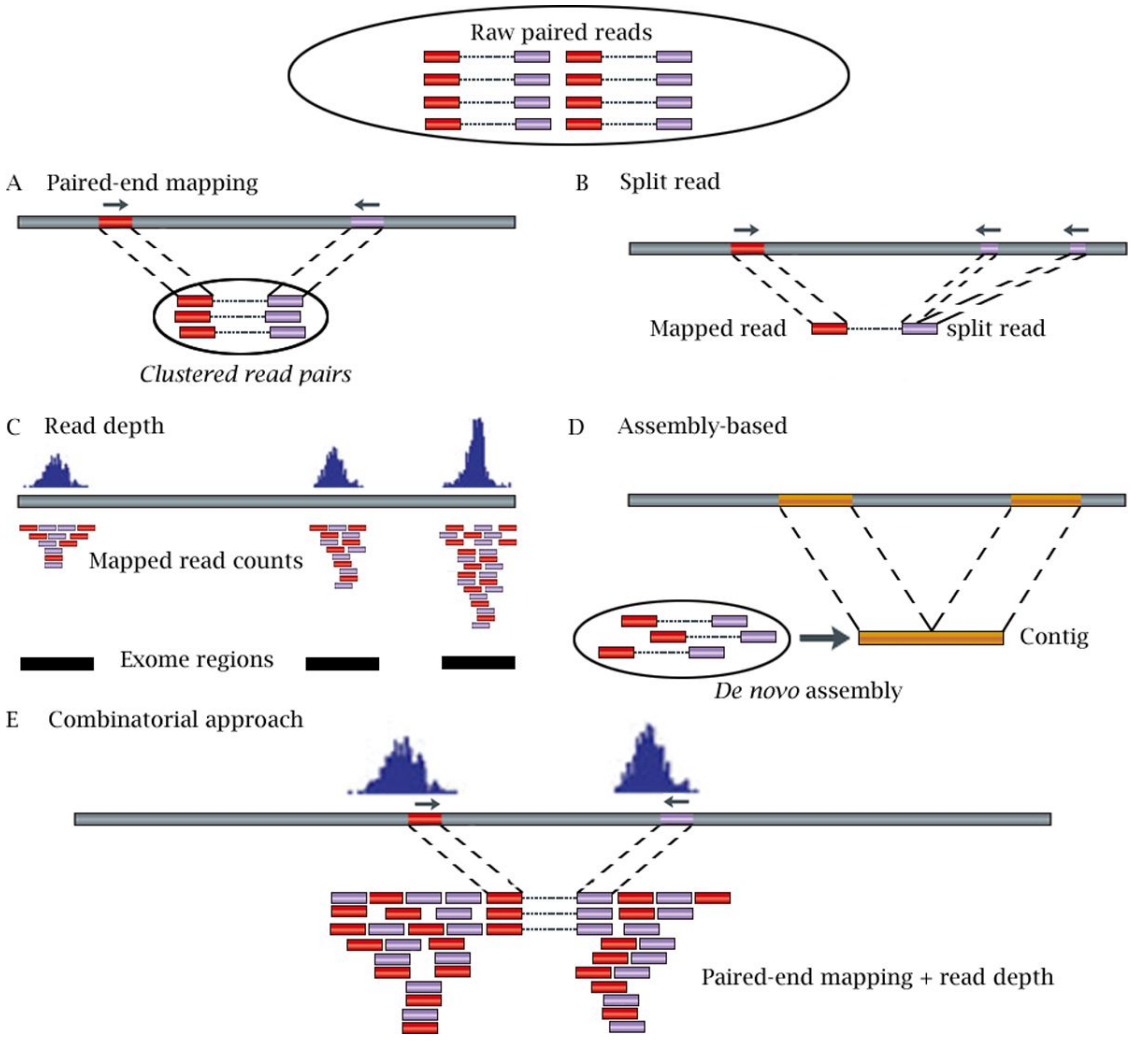


Figure 2.29: Sources of evidence for the presence of Structural Variants.[192].

We look at several SV callers in closer detail.

Delly

Delly[139] is a structural variant calling tool built in 2011 by T. Rausch et al. in the context of the 1000 Genomes Project. Delly uses discordantly-mapped reads and split-reads as sources of evidence for the presence of SVs, is able to characterize a broad set of variants and has been implemented in C++. All of the mathematical formulas and figures in this section are reproduced from the 2012 mauscript by Raush et al.

The discordantly-mapped reads component of Delly begins by computing the median and standard deviation of the read insert size (distance between the ends of the two reads in a pair), as well as the default read orientation by sampling a pre-defined number of reads from the BAM file. Read-pairs that have an insert size farther than 3 standard deviations from the median are considered as evidence for structural variation. This induces a minimum SV size detectable by Delly. See Figure 2.30 for the variant types implied by each insert size deviation and read-pair orientation.

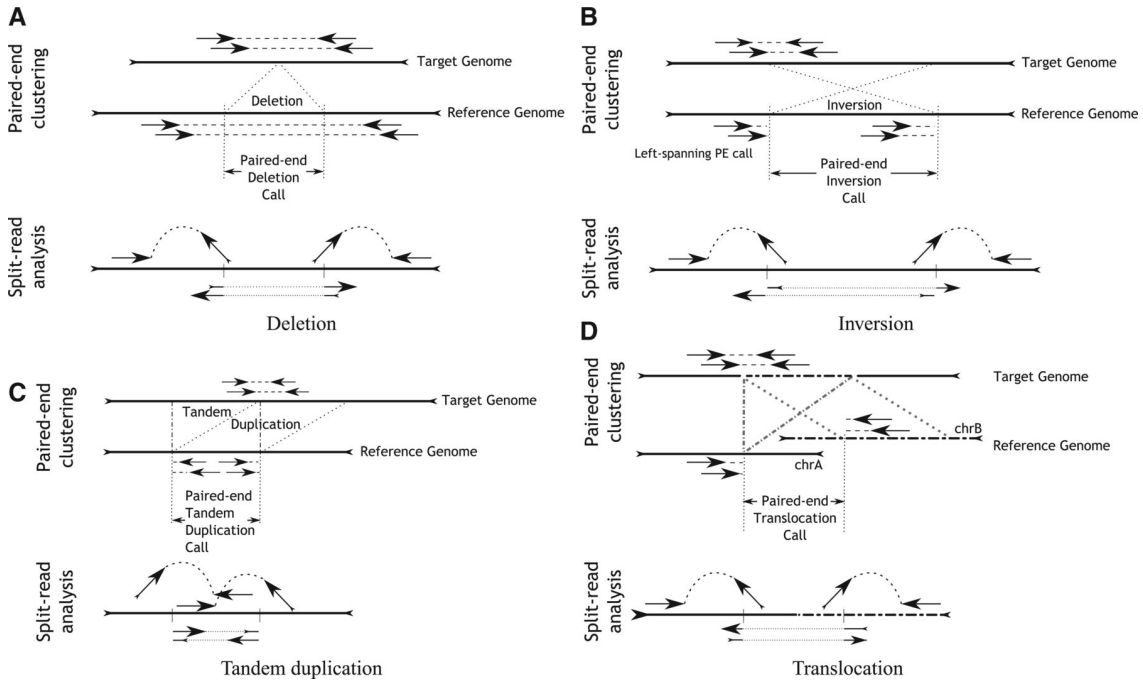


Figure 2.30: Variant classes detectable by Delly based on their read-pair and split-read signatures.[139].

Using the list of discordantly-mapped read-pairs Delly builds an undirected, weighted graph to indicate which read-pairs support the same variant. In the graph $G(V, E)$, a read-pair p_i corresponds to a node $v_i \in V$ and an edge $e_{v_i,v_j} \in E$ connects two nodes that support the same SV. The weight $w(e_{v_i,v_j})$ is the absolute value of the difference between the SV sizes predicted by v_i and v_j. v_i and v_j support the same variant when the corresponding read pairs have the same read orientation and the absolute difference between the left and right ends of the two reads is less than the expected insert size. Variants are identified by computing connected components C_i

of G. When components are not fully connected, Delly sorts the edges in each such component by weight and attempts to find a maximal clique M_i within C_i using edge with the smallest weight as the seed of the clique. The clique is extended by searching for the next smallest edge for which one of the vertices is already in the clique M_i, and requiring that $M_i \cup \{v_l, v_m\}$ is also a clique, until no further vertices can be added. The vertices in M_i are reported as the read pairs supporting that SV. Each rearrangement type is analyzed independently in this manner.

All of the rearrangements identified in the discordantly-mapped read analysis are refined using split-read analysis. The reads used for split-read analysis are those where one read in the pair maps to the reference and the other is unmapped. All such reads within a distance of 2 standard deviations of the median insert size from the breakpoint are considered up to a configurable maximum of 1000. Delly splits each unmapped read into kmers and aligns the kmers to the reference sequence spanning the SV.

Lumpy

Lumpy[90] is an SV caller developed in 2013 by R. Layer et al. and is a package implemented in C++. All of the mathematical formulas and figures presented in this section are reproduced from the 2014 publication by R. Layer et al.

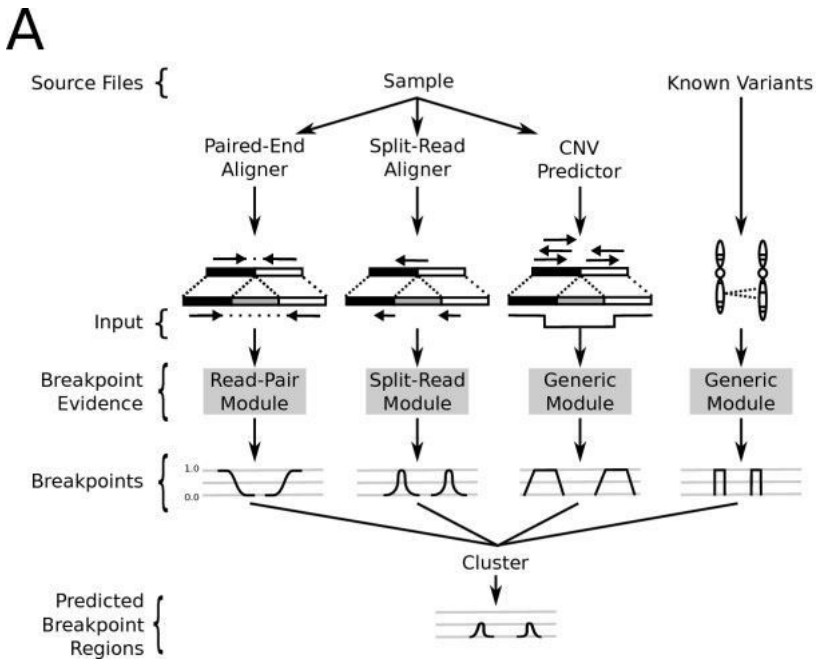


Figure 2.31: Lumpy calling model integrates several signals from a single sample.[90].

Lumpy defines an SV breakpoint as a pair of bases that are adjacent in the sample under study but not in the reference genome. Furthermore, each breakpoint is represented as a pair of probability distributions that span the predicted breakpoint regions and represent the uncertainty about the precise location of the breakpoint.

Lumpy integrates multiple signals (see Figure 2.31) to update the probability distributions that represent each breakpoint based on different kinds of evidence provided. A breakpoint is a tuple $b = \langle E, l, r, v \rangle$ where E is the evidence, $b.l$ and $b.r$ are the left and right intervals each having start and end coordinates, $b.l.s$ and $b.l.e$ for example. $b.l.p$ is a vector of length $b.l.e - b.l.s$ where each $p[i]$ is the probablity that $b.l.s + i$ is the true location of the breakpoint. $b.v$ is the breakpoint class (insertion, deletion, etc.). Overlapping breakpoints are merged together. The intervals that contain 95% of the probability density, as well as the Maximum Likelihood Estimates of the location of each variant are returned.

The paired-end analysis looks at read pairs $\langle x, y \rangle$ where each read is aligned to the reference genome as $R(x) = \langle c, o, s, e \rangle$ where c is the chromosome, $o \in +, -$ indicates alignment orientation, s and e represent the alignment start and end respectively. It is assumed that x and y align uniquely and that $R(x).s < R(x).e < R(y).s < R(y).e$. $S(x) = \langle o, s, e \rangle$ is defined to be the alignment of x with respect to the sample's (unknown) genome. Read pairs are assumed to be aligned with $R(x).o = +, R(y).o = -$ and having $R(y).e - R(x).s$ approximately equal to the library preparation fragment size. Read pairs that align outside the expected parameters for orientation and distance constitute evidence for structural variant breakpoints, in particular reads with the same or switched orientation, and pairs that align at a distance shorter or longer than the fragment size. Expected fragment length is estimated from the mean fragment length in the sample, along with its standard deviation. Breakpoint variety is determined as follows:

When $R(x).c = R(y).c$, the variety is inferred from read orientation. $R(x).o = R(y).o$ implies an inversion. $R(x).o = -, R(y).o = +$ implies a tandem duplication, $R(x).o = +, R(y).o = -$ implies a deletion. Insertions are not presently supported. When $R(x).c \neq R(y).c$ the breakpoint is labelled an interchromosomal rearrangement.

$\langle x, y \rangle$ are mapped to $b.l$ and $b.r$ as follows. By convention x is assumed to map to l and y to r. It is assumed that there exists a single breakpoint b between x and y. Orientation of x determines whether l is upstream or downstream from x. Thus, if $R(x).s = +$, then the breakpoint begins after $R(x).e$. The length of the breakpoint interval is proportional to expected fragment length and its standard deviation. The distance of one end of the breakpoint from x is assumed to be less than expected insert size L plus $v_f s$ - a configurable number of standard deviations. The probability that position i in the breakpoint interval l is part of the actual breakpoint is estimated as the probability that x and y span i, which is true when the fragment that had produced $\langle x, y \rangle$ is longer than the distance from x to i. Thus, the quantity of interest is $P(S(y).e - S(x).s > i - R(x).s)$ for $R(x).o = +$ and $P(S(y).e - S(x).s > R(x).e - i)$ for $R(x).o = -$. This quantity is estimated from the empirical distribution of fragment sizes collected from the sample.

Lumpy does not perform its own split-read alignment instead relying on the results of a split-read aligner like YAHA[49]. Every split-read is a DNA fragment X that consits of two or more sub-fragments $x_i, ..., x_j$ that do not map to adjacent locations of the reference. Each sub-fragment pair $\langle x_i, x_{i+1} \rangle$ is evaluated independently depending on how it maps with respect to the reference $R(x)$. When

$R(x_i).o = R(x_{i+1}).o$ the event is marked as either a deletion or a tandem duplication, where $R(x_i).s < R(x_{i+1}).s$ implies a deletion, and otherwise a tandem duplication. When orientations do not match, the event is marked an inversion. To account for potential inaccuracies in the location of the breakpoint with respect to the split-read pair, each split-read pair is mapped to two breakpoint intervals l and r that are centered at the split. The probability is modeled to be highest at the split with an exponential decrease towards the edges, with a configurable width parameter.

SvABA

SvABA[176] is a germline and somatic SV caller by J. Walla et al., created in 2016. Svaba is implemented in C++ and relies on genome-wide local assembly for variant detection. Internally SvABA makes use of SGA[159] and BWA-MEM[101] for assembly and read mapping respectively.

Figure 2.32: The SvABA variant callng method[176].

SvABA performs local de-novo assembly in 25-kbp windows tiled across the genome with a 2kb overlap. First reads that may be indicative of variation are extracted from a BAM file, these include reads with high-quality soft-clipped bases, discordant reads (with non-standard orietnation or with insert size greater than four standard deviations away from the sample mean insert size, determined using a sample of 5 million reads with top and bottom 5% of insert sizes truncated), unmapped reads, reads with unmapped mates, and reads with insertions or deletions indicated in the CIGAR string. Low quality reads that are marked as PCR duplicates, have failed QC, and reads with repeats of >20 basepairs are filtered out. Discordant reads are realigned to the reference with BWA-MEM and those with an available nondiscordant alignment of >70% of the maximum alignment score are discarded. Reads with many different (>20) high quality candidate alignments are also discarded. Remaining discordant reads are clustered based on orientation and insert-size and assembled using SGA.

The contigs produced by SGA are aligned to the reference genome using BWA-MEM and examined for potential variants. Contigs with an alignment that has fewer than 30 nonaligned bases and no alignment gaps are considered reference. Indels are called when the alignment has gaps, and SVs called when the resulting alignment is multi-part. In order to evaluate read support for the putative variants, reads from the windows are aligned to both the reference and the assembled contigs using BWA-MEM. Reads are considered matching to the contig when the alignment score for the read to the contig is greater than the alignment of the read to the reference and is >90% of the length of the match. Reads that have an alignment that is up to 8 bases to the left or right of a putative variant are consdered supporing the variant.

Remarks

We looked at Delly, Lumpy, and SvABA in the context of germline SV calling. Although there are some differences in the details of the approaches taken by the tools' authors there are consistent similarities as well, which additionally carry over to other SV callers such as Pindel[187], and Manta[18], that have not been presented here. These approaches rely heavily on paired-end reads and select those that are mapped uncommonly far apart or close together. These are clustered or optionally locally assembled to produce putative breakpoints. Breakpoint locations are further refined by split-read analysis, using those reads that are unmapped by regular read mapping software. These reads are broken down into kmers and each kmer is aligned separately to non-adjacent locations in the genome. Alternative haplotypes may be constructed using these reads and alignments to the reference and the alternative haplotypes scored for genotyping purposes. A new method for structural variant calling should thus focus on making the best use of these two data types (discordant and split reads), while possibly also making use of read depth information to be competitive with the current generation of best callers.

2.2.7 Variant Filtering

2.2.8 Somatic SNP Calling

Mutect - Muse - Pindel -

2.2.9 Somatic Indel Calling

2.2.10 Somatic Structural Variant Calling

2.2.11 Germline Variant Annotation

Annovar - Variant Effect Predictor -

2.2.12 Somatic Variant Annotation

2.2.13 de-novo Assembly

Velvet - Abyss -

2.3 High Performance , High Throughput, and Cloud Computing

The practice of performing large scale scientific computation on supercomputers or clusters of commodity hardware can be split into two notions - High Performance Computing (HPC) and High Throughput Computing (HTC).

The European Grid Infrastructure defines these as follows[59]:

> HPC - A computing paradigm that focuses on the efficient execution of compute intensive, tightly-coupled tasks. Given the high parallel communication requirements, the tasks are typically executed on low latency interconnects which makes it possible to share data very rapidly between a large numbers of processors working on the same problem. HPC systems are delivered through low latency clusters and supercomputers and are typically optimised to maximise the number of operations per seconds. The typical metrics are FLOPS, tasks/s, I/O rates.

> HTC - A computing paradigm that focuses on the efficient execution of a large number of loosely-coupled tasks. Given the minimal parallel communication requirements, the tasks can be executed on clusters or physically distributed resources using grid technologies. HTC systems are typically optimised to maximise the throughput over a long period of time and a typical metric is jobs per month or year.

Although early High Performance Computing efforts (1960's - 1980's) relied on supercomputers with a shared memory model[147], where all of the memory was shared between multiple processors, by the late 1980's machines with a distributed memory model[129], where each processor has its own memory, started gaining ground, forming the basis for the modern HPC cluster.

The software interface that the user has to a HPC/HTC cluster typically takes the shape of a queueing system such as PBS[70] or LSF[194] where the user writes a script that submits a series of jobs to the queueing system. The jobs can invoke software that is installed by the IT department that manages the cluster. The user is not able to install any software and has limited visibility into the runtime performance characteristics of the jobs they submit.

Cloud computing has emerged in the early 2000's enabled by improvements in hardware virtualization which was driven by the adoption of Virtual Private Networks, and the desire to commercialize access to compute capacity as a utility[17].

The National Institute of Standards and Technology provides a standard definition of cloud computing that encompasses several areas of this domain - Essential Characteristics, Service Models, and Deployment Models[121].

The Essential Characteristics of a cloud are as follows:

On-demand self-service - End-user can independently manage infrastructure without involving the service provider.

Broad network access - Cloud resources are available on the network via a set of standard protocols.

Resource pooling - Service providers dynamically assign virtual infrastructure in a multi-tenant environment based on consumer demand.

Rapid elasticity - Resources can be elastically provisioned and discarded according to customer requirements.

Measured service - Resource usage by end users is measured and transparently provided back to the user by the service provider.

It is the self-service and broad network access characteristics that set cloud computing apart from traditional HPC computing the most.

Service Models include:

Infrastructure as a Service (IaaS) - This service allows the the user to provision and control virtualized infrastructure such as VMs and networks.

Platform as a Service (PaaS) - This service allows the user to deploy their application onto virtualized hardware but not to control the management of the infrastructure.

Software as a Service (SaaS) - This service allows the user to make use of applications that are deployed on virtualized hardware but not to manage the applications or the infrastructure itself.

The Deployment Models covered by the NIST definition are as follows:

Private Cloud - Operated privately by a single organization and not accessible on a public network.

Community Cloud - Established for use by a particular community of users with a common interest.

Public Cloud - Established for general use by the public.

Hybrid Cloud - A collection of cloud entities that use one of the other deployment models but allow application portability.

The first publicly available commercial cloud computing platform has been developed by Amazon.com and launched in August, 2006 in the form of two services - Elastic Compute Cloud (EC2), and Simple Storage Service (S3). Cloud offerings by

Microsoft and Google followed in 2010, and 2012. This early lead has allowed Amazon to capture the majority of the public cloud computing market, earning $2.57 billion USD in Q1 2016 revenue.

One of the main drawbacks, however, of using Amazon's or another proprietary cloud solution is the issue of "vendor lock-in" i.e. inability to easily switch infrastructure providers should the customer wish to do so, because of the amount of software relying on the proprietary cloud provider protocols. Another key reason for avoiding public clouds is the necessity to store sensitive data. This issue applies both to the commercial enterprise (with industries such as banking, and payments) and scientific domains (especially genomics and medicine) where handling of sensitive patient data is restricted based on both technical security, as well as ethical considerations[83].

To help alleviate these concerns an open-source cloud platform called Openstack was launched in 2010 jointly by Rackspace Hosting and NASA[152]. Openstack provides most of the same features that are provided by Amazon Web Services and other commercial cloud providers as free open-source tools. These include:

- Infrastructure
- Networking
- Identity Management
- Block Storage
- Object Storage
- Managed Databases
- Queues
- Monitoring

Openstack deployments for the basis for most academic private and community clouds such as EBI Embassy Cloud[28], University of Chicago Open Science Data Cloud[65], Cancer Genome Collaboratory[188], and Helix Nebula[117]. Because these clouds implement the security measures necessary when handling patient data they are a system of choice for large scale bioinformatics analyses.

2.4 Workflow Systems

The focus on workflow stems the work of Frederick Taylor (1856-1915) and Henry Gantt (1861-1919) on the improvement and automation of industrial processes, also known as "scientific management"[166]. One of the key techniques that were devised at the time and served as the prototype for future workflows were "time and motion

studies"[11] where employees were observed as they performed repetitive cycles of work in order to determine standard execution times and sequences of steps. As this field evolved over the course of the 20th century it gave rise to several other related fields such as Operations Management, Business Process Management, and Lean Manufacturing.

In 1993 an international consortium was formed with the purpose of defining the standards related to workflows and workflow management systems. This consortium is called the Workflow Management Coalition (WfMC). One of the key specifications produced by the WMC in 1995 is The Workflow Reference Model[73]. This document provides two basic definitions that illuminate the scope and purpose of workflow systems:

> Workflow - The computerised facilitation or automation of a business process, in whole or part.

> Workflow Management System - A system that completely defines, manages and executes "workflows" through the execution of software whose order of execution is driven by a computer representation of the workflow logic.

A number of standards have been produced for workflow definition, many of them are XML-based[154]. Notable examples include:

XPDL - Was developed by the WfMC, currently at version 2.2, as of 2012. Uses an XML dialect to express process definitions

BPML - Developed by the Object Management Group (OMG) using XML. Deprecated as of 2008 in favour of BPEL.

BPEL/BPEL4WS - Developed by Organization for the Advancement of Structure Information Standard (OASIS). Uses XML format. Adopted by Microsoft and IBM for their workflow products -

Graphically, workflow definitions are typically expressed using a Petri-Net[136] or Business Process Model and Notation (BPMN), the latter borrowing its structure from UML activity diagrams. A set of workflow definition design patterns exists to guide workflow creation[37]. A workflow engine is responsible for ingesting workflow definitions, generating their graphical representation, and allowing the user to execute the workflow definitions on suitable hardware.

As initially the focus of workflow systems research and development has been on process improvement within commercial enterprises there exists a large pool of workflow engine implementations targeted at that sector. Some of these are:

jBPM - An open-source workflow engine that is based on the Java platform and is currently owned by Red Hat.

Activiti - An open-source workflow engine that has been developed by previous jBPM developers.

Oracle BPEL Process Manager - A commercial workflow engine acquired by Oracle from Collaxa in 2004, now integrated into the rest of the Oracle portfolio.

Websphere Process Server - Commercial workflow engine that is part of IBM's Business Process Manager suite.

Although these tools have gained wide adoption in the enterprise community they have had limited success within scientific circles. Instead, several open-source workflow management systems exist that have been purpose-built for the scientific domain, and especially bioinformatics. These include:

Kepler[108] - A Java-based WfMS built on top of the Ptolemy II[35] execution engine.

Taverna[130] - A Java-based WfMS originally built by myGrid, currently under incubation at Apache Software Foundation.

Galaxy[60] - A Python-based WfMS developed specifically for bioinformatics applications with a focus on GUI-driven development of workflows.

Curcin et al[30] provide a head-to-head comparison of six scientific workflow systems including Taverna and Kepler, whereby Taverna is described as primarily being aimed at researchers who wish to build scientific workflows from web services utilizing a proprietary XML dialect called SCUFL which implements a DAG model of workflows. The primary execution environment for a Taverna workflow is on a grid or an HPC cluster. Kepler implemented a different methodology, whereby workflow modelling, which is taken on by Actors, is separated from workflow execution, taken on by Directors. An Actor knows only about its inputs, the computation that it needs to perform, and the output that it needs to produce, while Directors provide different models of execution, such as Synchronous Data Flow, Process Network, Continuous Time, and Discrete Event.

The Galaxy workflow framework has a specific focus on bioinformatics analyses and comes with a large library of community-developed bioinformatics workflows. The user creates and executes workflows via a web-based GUI where pre-installed tools and scripts can be laid out into a pipeline. The primary deployment environment for Galaxy is on an institutional HPC cluster although a separate component allows the deployment of a Galaxy instance on Amazon Web Services[3].

2.5 Service Oriented Architectures

2.6 Stream-based Systems

Chapter 3

The Butler Framework - Requirements and Architecture

Chapter 4

The Butler Framework - Implementation and Experimental Validation

Chapter 5

The Rheos Framework

In this chapter we describe a software framework called Rheos, which demonstrates an approach for reasoning about large genomic datasets utilizing concepts of service-orientation and data streaming in contrast with traditional genomic data analysis frameworks[36] that take a procedural batch-based approach. Rheos' focus on service-orientation and streaming allow the users to make active tradeoff decisions between analysis time, cost, and quality as well as setting up precise operational Service Level Agreements, both between Rheos components, and between Rheos and external systems, as we describe in detail below.

5.1 General Framework Design

As already discussed in Chapters 1 and 2, the general problem consists of collecting DNA samples from a population of individuals under study, sequencing these samples using Next Generation Sequencing techniques, identifying the mutations that are present, annotating their functional impact and utilizing the obtained data in a downstream data analysis with research or clinical decision-making goals. While there is a great variety of possible downstream analyses that may be performed depending on the individual goals of the analyst, there is a fairly well established set of steps for processing of the raw NGS data into a set of annotated variants, and it is these steps that we target with this work. The typical approach that is in widespread use today is to collect a batch of samples and then process each sample individually with a sequence of individual tools, that may be described via a higher-level workflow construct (such as in Figure 2.26, or using a framework like Butler, as described in Chapters 3, 4). There are, however, a number of factors that leave room for improvement in this model. These improvements lie along a set of dimensions that we describe briefly in the Introduction via a utility function $U_i = C_i + T_i + A_i$ for sample $i \in [1, N_s]$ that needs to be optimized, and that we describe in more detail here.

We use the following definitions throughout the text:

Table 5.1: Rheos common definitions

Symbol	Description
N_p	Number of people
$P = \{p_i : i \in [1, N_p]\}$	Set of individuals under study
N_s	Number of samples
$S = \{s_i : i \in [1, N_s]\}$	Set of sequenced DNA samples. Each individual can have one or more samples.
A_i	Accuracy score of analysis for sample i (precise definition of Accuracy TBD)
$C_i = c_{g_i} + c_{s_i} + c_{a_i} + c_{r_i}$	Cost score of data generation, storage, analysis, and retrieval respectively
T_i	Time score to process sample i
$U_i = C_i + T_i + A_i$	A utility function for individual i that penalizes high cost, high processing time, and low accuracy
$U = \sum_{i=1}^{N_s} U_i$	Overall utility of processing N_s samples through Rheos.

5.2 Data Streaming Architecture

The overall technical architecture of the Rheos system is set up as a Service Oriented Architecture (SOA)[155] which is an information system architecture paradigm where the overall problem that the system is trying to solve is broken down into a collection of loosely-coupled components called services. Each service has a well defined interface of inputs that it accepts and outputs that it produces. Services can be combined and orchestrated together to produce the overall desired output for the system. A key distinguishing feature of this architectural approach is that each service can be individually optimized to fulfill its contract most efficiently helping break down some of the performance limitations brought about by the necessity to simultaneously tackle competing constraints in more monolithic information system designs. Additionally, within a services framework, the dependencies between separate services can be negotiated not only in terms of service interfaces, but also in terms Service Level Agreements which constitute Quality of Service promises made by one service to its dependents[76]. Because it is unlikely that a service designer will be able to accurately foresee all of the demands that will be placed on a service during its lifetime the SLAs provide a valuable feedback framework through which the service can be evaluated as it operates in production, as well as serving as a basis for negotiating evolving requirements between dependent services.

While general web services can support any data processing paradigm, in the Rheos framework we adopt a data streaming approach[10]. In this approach we assume that the input to any service is a randomly ordered sequence of messages $M = m_1, m_2, ...$ where each message represents a fact about the underlying domain that the service reasons over, as well as some metadata, including an identifier, and a variety of timestamps of interest. The content of each message may provide a

datum, such as the measurement of a quantity of interest, or signal that a particular event has taken place. It is in general assumed that the data stream is infinite in size, that messages may arrive out of order, and that any message that is placed in the stream is observed at most once, and may, in fact, never be observed. Messages are typically not sent directly from one service to another, instead the transfer of messages is mediated by a queuing system using a publish-subscribe[44] model. Under this model each queue acts as a *topic*. Message producers can publish data to the topic, and message consumers subscribe to receive messages from the topic. A message is consumed from the queue only after all of the subscribed consumers have seen it. End users retrieve information from the system via a set of User Interfaces that support both push (notifications) and pull (querying) models of data retrieval. A more detailed description of the architectural aspects of the system follows:

5.2.1 Service-Oriented Data Streaming Model

A data stream $M_{s,d} = m_1, m_2, ...$ is a sequence of datagrams transmitted over the network with the following properties:

- The stream has a source s and a destination d.
- A message m in the stream is a tuple of the form $(header, payload)$, where:
  - *header* is a tuple of the form $(id, \dots)$ that holds at minimum a unique identifier id for messages, and may hold additional metadata.
  - *payload* is an arbitrary data structure that holds the informational content of the message.
- $|M| = \infty$ by assumption.
- Messages may not arrive at destination d in the same order that they were sent from source s.
- If $t_{i,s}$ is the time message m_i leaves the source s and $t_{i,d}$ is the time of arrival at destination d, then $\sup_i\{t_{i,d} - t_{i,s}\} = \infty$, i.e. a given sent message may never arrive at its destination.

A service $S = \{o_i\}$ is a collection of operations o_i that act on one or more input data streams $\{M_i\}$ to produce one or more transformed output data streams $\{M_j\}$. Specifically:

An operation O is a tuple of the form:

$$O = (i, o, p, f) \tag{5.1}$$

where:

- $i = \{M_j, j \in [0, K]\}$ is a set of $K \geq 0$ input data streams.

- $o = \{M'_j, j \in [0, L]\}$ is a set of $L \geq 0$ output data streams.
- $f : M^K \mapsto M'^L$ is a transformation function that produces messages m' in the output streams based on messages m observed in the input streams.
- $p = \{p_i\}$ is a set of potentially optional query parameters.

There are several distinct categories of operations that a service can perform on a set of input streams. We describe these here:

Windowing Function - Service S observes a sliding window, which is a sample of size n of messages from stream M_i and computes a summary statistic (see Figure 5.1) over the sample which is meant to be an approximation of the corresponding population parameter.

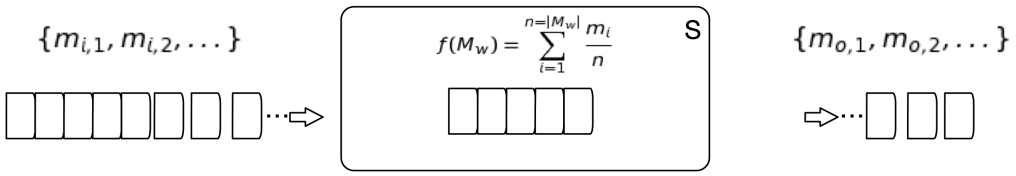


Figure 5.1: Service S computes a summary statistic over a window of messages from stream M

Decorator Function - Service S observes messages m_i and applies a function that augments (decorates) each message with additional attributes (see Figure 5.2) producing augmented messages m_o as output.

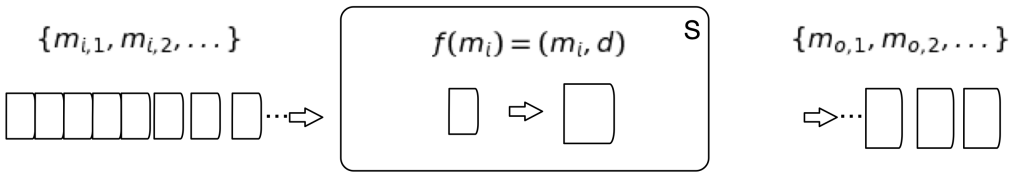


Figure 5.2: Service S augments messages from M with an additional set of attributes.

Filter Function - Service S observes messages m_i and applies a function $f : M \mapsto \{True, False\}$ that evaluates to a boolean value (see Figure 5.2). Only messages that map to $True$ are emitted as output.

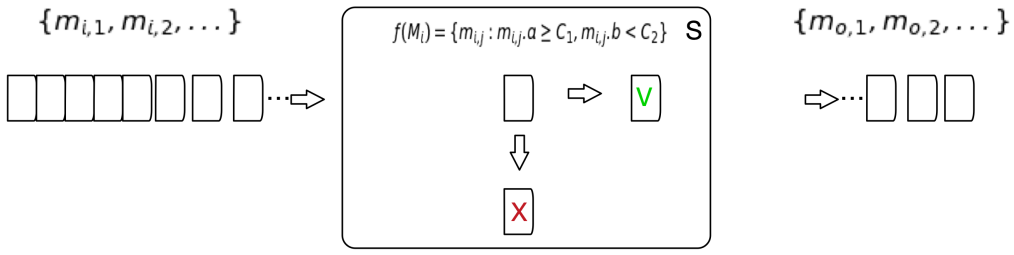


Figure 5.3: Service S filters messages from input stream M and only allows through those that pass the filtering condition.

Aggregator Function - Service S observes messages from N different streams $\{M_j : j \in [1, N]\}$ and applies a function $f : M_i^N \mapsto M_o$ that aggregates messages from these streams to produce its output (see Figure 5.4). Because aggregation happens over groups of messages that may not all arrive at the same time the service S requires a mechanism for keeping local state so that it can accumulate messages that have already arrived while waiting for those that are necessary to compute f yet have not been observed. The statefulness requirement of this type of service places an extra level of complexity (related to state-management and request routing) as well as inherent scalability limitations compared to stateless services[131].

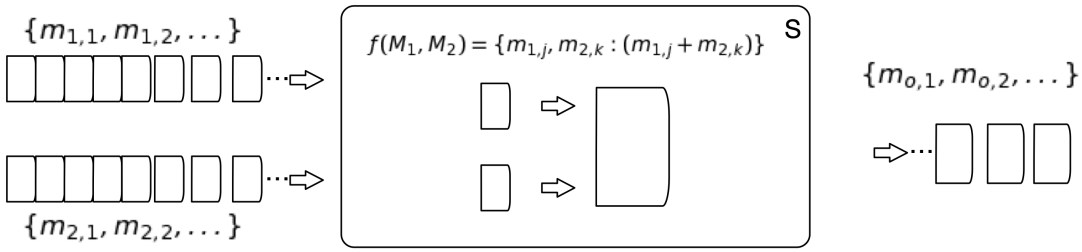


Figure 5.4: Service S integrates messages from multiple input streams M_i to produce an aggregated output stream M_o via f.

Local State Aggregator Function - Service S observes an input stream M_i which it integrates with a local (non-stream) queryable data store (see Figure 5.5). Messages m_i are integrated with query results q_i to produce an output stream M_i. This type of service also requires management of state and scalability concerns similar to the Aggregator service, especially when the local data store is itself distributed.

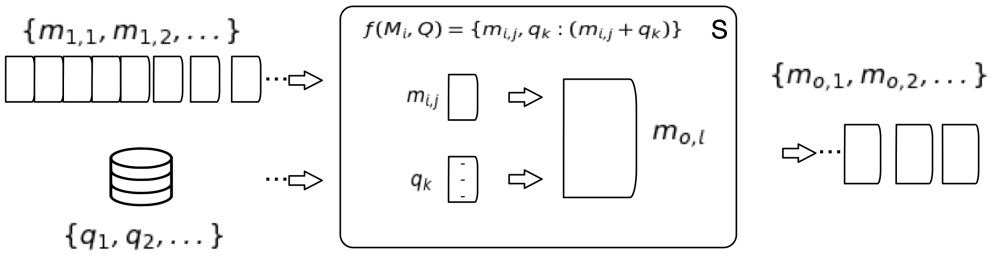


Figure 5.5: Service S aggregates m_i with query results q_i obtained from a local data store.

Persistence Function - Service S observes messages m_i and is responsible for persisting them to a data store where their contents can later be queried (see Figure 5.6). Although persistence of data to, and subsequent querying of data from, a store, such as a database, are comparatively more expensive operations than immediate reasoning over a live data stream, such mechanisms are necessary for situations where data may need to be accessed multiple times, or where data may need to be retained for audit purposes.

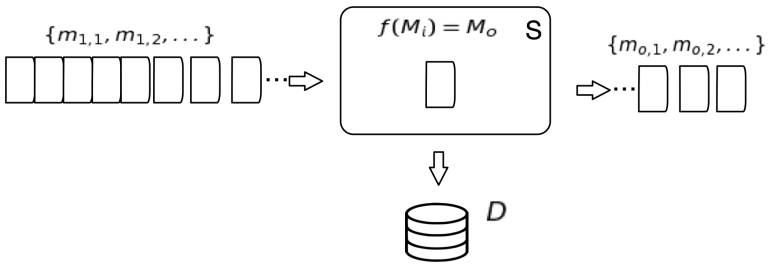


Figure 5.6: Service S processes messages m_i into persistent storage. The output stream M_o contains persistence confirmation and error events.

Query Function - Service S observes a stream of queries Q_i. The queries are fulfilled against a data store D and the results emitted via the output stream M_o.

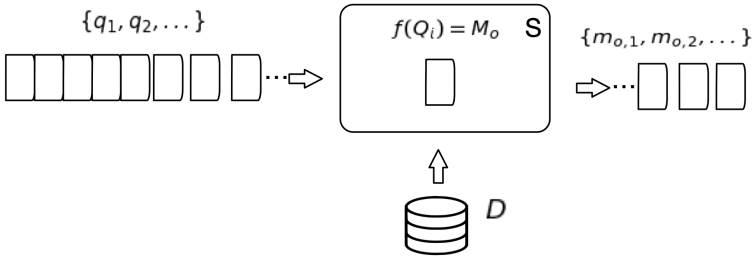


Figure 5.7: Service S filters messages from input stream M and only allows through those that pass the filtering condition.

The basic operations above can be combined to produce arbitrarily complex logic

on data streams.

One of the key advantages of a service-oriented approach is that, because services are typically constantly executing, it naturally lends itself to an examination of the system's runtime characteristics. This applies to both service-internal characteristics that are related to each operation a service performs, as well as to external characteristics that relate to the contracts a service establishes with its dependencies. We consider both of these.

For each given operation $o_i \in S$ it is instrumental to understand the resource requirements of the operation on typical inputs and limiting factors that affect the efficiency with which the operation can be performed by S. Of particular interest are the per-operation profiles of:

- CPU utilization
- RAM
- Secondary storage
- Network utilization

If o_i is a long-running operation that takes multiple seconds to complete on average, a detailed distribution in time of each metric above may be necessary. If the operation can be completed at a sub-second rate then summary statistics (min, max, mean, median, inter-quartile range, 90th, and 99th percentiles) may be sufficient. This level of understanding is necessary in order to make sure that the service can adequately deal with the incoming message stream while the messages are first loaded into memory, since subsequent retrieval from secondary storage is several orders of magnitude slower and may cause further delays in processing. If o_i is stateless, i.e. it does not require the storage and retrieval of any local state that depends on the content of each arriving message $m_j \in M_i$, then the service S can be scaled "horizontally"[171] with respect to o_i. Given that the performance-limiting condition of o_i is known (CPU, memory, etc.), the ability of S to efficiently deal with fluctuations in the rate of incoming messages M_i can be successfully achieved simply by adding and removing servers that execute S (see Figure 5.8), which can be done automatically[113]. If o_i is stateful and requires access to databases, or predictable request routing via sessions, then horizontal scalability may not be possible and thus, detailed understanding of the performance profile and performance-limiting conditions of o_i is even more important as vertical scaling of services is more expensive and challenging to accomplish, and may increase system complexity by necessitating data partitioning, for example[171].

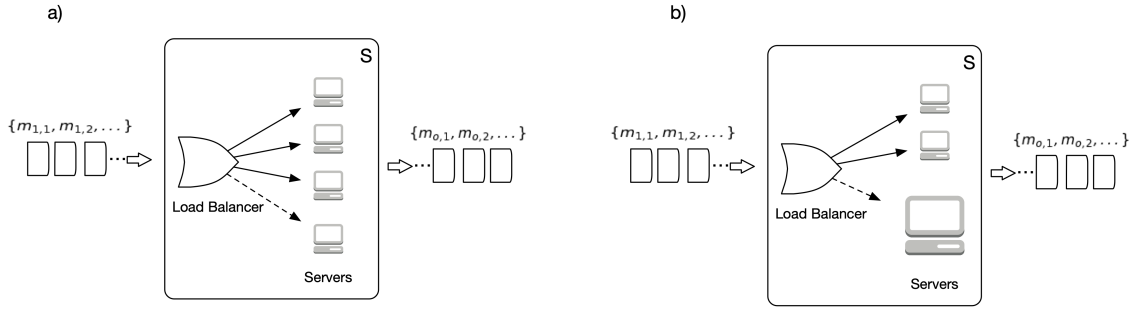


Figure 5.8: a) In horizontal scaling new servers are added and removed (dashed arrow) behind a load balancer as the rate of data stream M_i fluctuates. b) In vertical scaling more powerful servers need to be launched (dashed arrow) to replace smaller servers (with potential service outage) when the rate of M_i increases beyond capacity.

Assume service S implements operation o supported by n physical servers $V = \{v_j : j \in [1, n]\}$ by consuming a stream of incoming messages M_i. For a suitable time increment t, let $r_{M_i} = |M_i|/t$ be the incoming message arrival rate, and $r_{M_{o,j}} = |M_{o,j}|/t$ be the processing rate for server v_j. If $r_{M_i} > \sum_{j=1}^{n} r_{M_{o,j}}$, then S will not be able to adequately process all of the incoming messages from M_i and messages will either be lost or need to be backlogged while more servers are added to S to deal with the incoming message rate. Since commissioning new servers takes considerable time and the timing and magnitude of increases in r_{M_i} may be unpredictable, serious information loss may result if measures are not put in place to mitigate the message rate fluctuations.

A queue is the mechanism that we put in place to address this concern (see Figure 5.9). A queue Q is a message buffering system which consists of a set of n "topics" $P = \{p_i : i \in [1, n]\}$, where each topic is a tuple of the form $p_i = (D_p, B_p, C_p)$. Here $D_p = \{d_i : i \in [1, k]\}$ is a set of k data producers that put messages into Q, B_p is a message buffer of max capacity N_{max} dictated by underlying server hardware characteristics, containing a sequence of messages $\{m_t, m_{t-1}, m_{t-2}, ..., m_1\}$ that are accessible in a Fist In First Out (FIFO) manner, and $C_p = \{c_i : i \in [1, l]\}$ is a set of l consumers that are interested in observing messages from p_i.

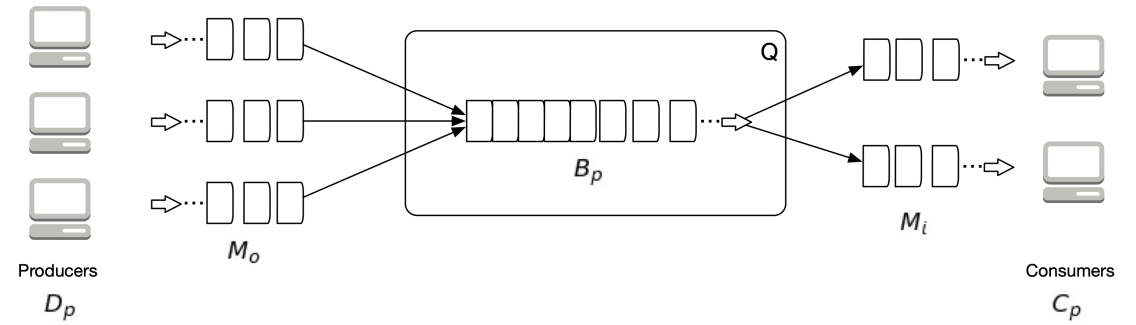


Figure 5.9: A
queue Q establishes a message buffer B_p between a set of message producers D_p and consumers C_p, for a given topic p.

Messages arrive into a particular topic of Q from all producers D_p, and are marked safe for deletion only when all of the subscribed consumers C_p have observed a particular message. Thus, for topic p_i, the incoming message rate is $R_i = \sum_{j=1}^{k} r_{M_{o,j}}$ i.e. the sum of the message processing rates for all of the producers for this topic. The queue message processing rate $R_o = \min_{i \in [1,l]} \{r_{M_{o,i}}\}$ is the slowest message processing rate among all consumers. Assuming $R_i > R_o$ and that there are N messages presently in B_p, there remains $t = \frac{N_{max} - N}{R_i - R_o}$ time before queue overflow occurs. The situation should then be remedied by allocating additional hardware to Q or those services S_i whose consumers are slowest, until the condition $R_o \geq R_i$ can be reliably maintained. If the queue does reach its maximum capacity overflow measures need to be put in place. Depending on the data stream in question data loss may or may not be acceptable. If data loss is acceptable then overflow messages can be simply discarded. If data loss is not acceptable then producers must block waiting for additional queue capacity to become available. This not only degrades performance locally, but can have a drastic effect on the entire system if the effects are allowed to percolate trough the complex distributed system. As message rates evolve through time with system load, the scheme above sets up a framework for flow control and hardware allocation within the architecture.

When designing a service-oriented system the interfaces of operations provided by the service are of utmost importance as they define the capabilities that the service offers to its clients. Of secondary, but also significant, importance is the set of Service-Level Agreements (SLAs)[184] that a service advertises. These SLAs are a set of commitments that a service makes to its clients that describe the operational characteristics of the service, such as:

Availability - Guarantees related to the service uptime, maintenance outages, disaster recovery, etc.

Throughput - The number of requests serviced per unit time.

Latency - The delay between a request being sent and a response being received.

Abandonment Rate - Proportion of requests that are never answered.

Error Rate - Proportion of well-formed requests that result in an error.

Based on the SLAs that are advertised by a given service, the services that depend on it can make assumptions about expected runtime behavior, and take action when expectations are not met. Furthermore, when requirements evolve and features are added to or removed from a service, the impact on the advertised SLAs helps communicate the full effect of the changes. Lastly, the costs of operating a service are more clearly understood through the SLA framework, where improvements to a particular SLA metric, such as Transactions-Per-Minute (TPM) can be transparently traced to a corresponding increase in operational costs.

The set of services $\{S\}$ that communicate over data streams $\{M_{s,d}\}$, mediated by a set of queues $\{Q\}$ with a set of established SLAs $\{L_s\}$ together form the overall framework of Rheos that is used to tackle the challenges of large-scale genomic data processing in a manner the enables active tradeoffs between the competing constraints of cost, time, and accuracy.

5.3 Domain-specific Problems

Having laid out the general data-streaming service-oriented architecture of Rheos in the previous section we now turn to a discussion of the set of actual domain-specific problems that need to be solved within the data-streaming paradigm in order to enable the comprehensive genomic characterization of large cohorts of samples within Rheos, as we have set out to do. We make use of the flow of data types from the most raw to the most refined (see Figure 5.10) to illustrate the challenges that need to be solved during transformation of the input data between each successive stage, first in summary form, and then in full detail, below.

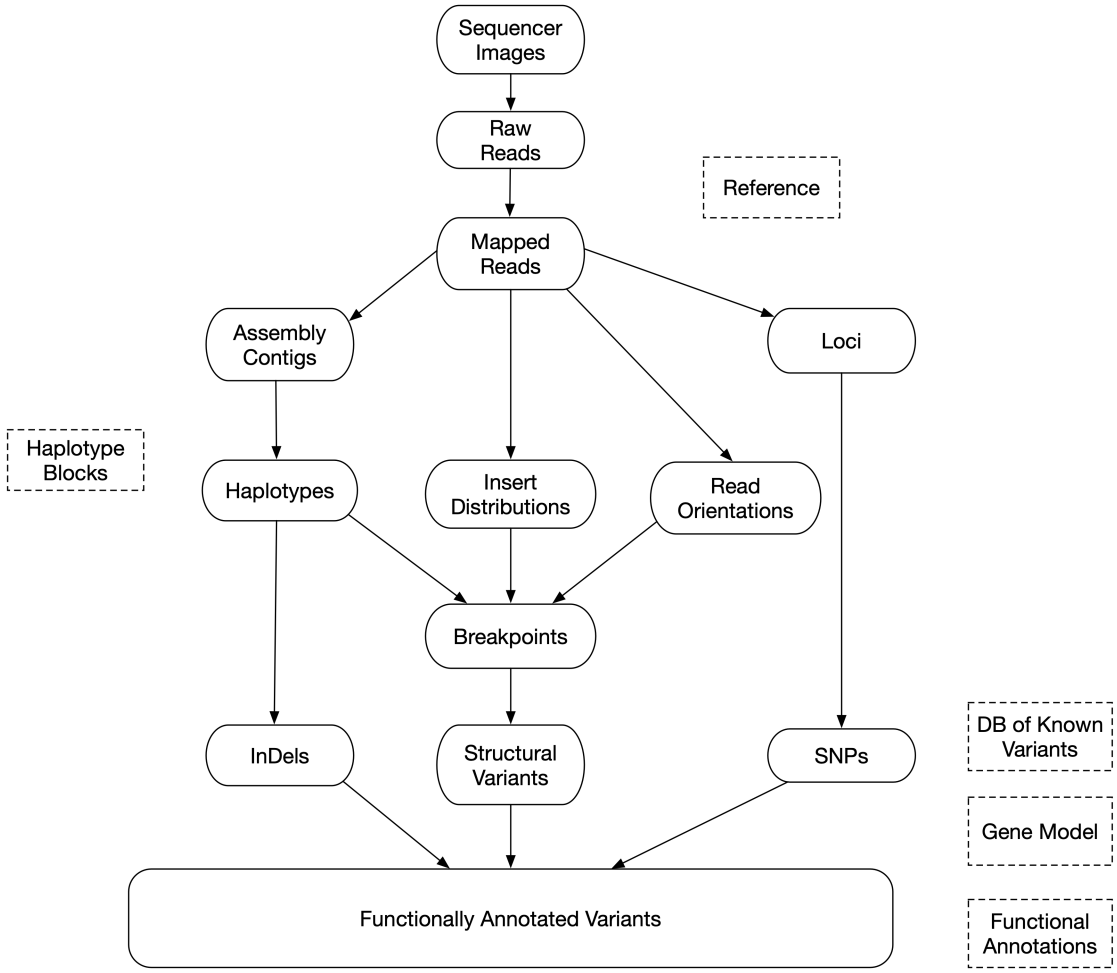


Figure 5.10: The conceptual flow of data types within Rheos from the most raw - Sequencer Images, to the most refined - a set of Functionally Annotated Variants.

The most raw data type that is produced from a sequencing experiment is the set of raw image files generated by the sequencer. Although, conceptually, processing of the raw images could also be accomplished within Rheos, it is presently outside of the scope of this work. Instead, we assume the most basic data type to be raw sequencing reads, as found in a FASTQ[21] file.

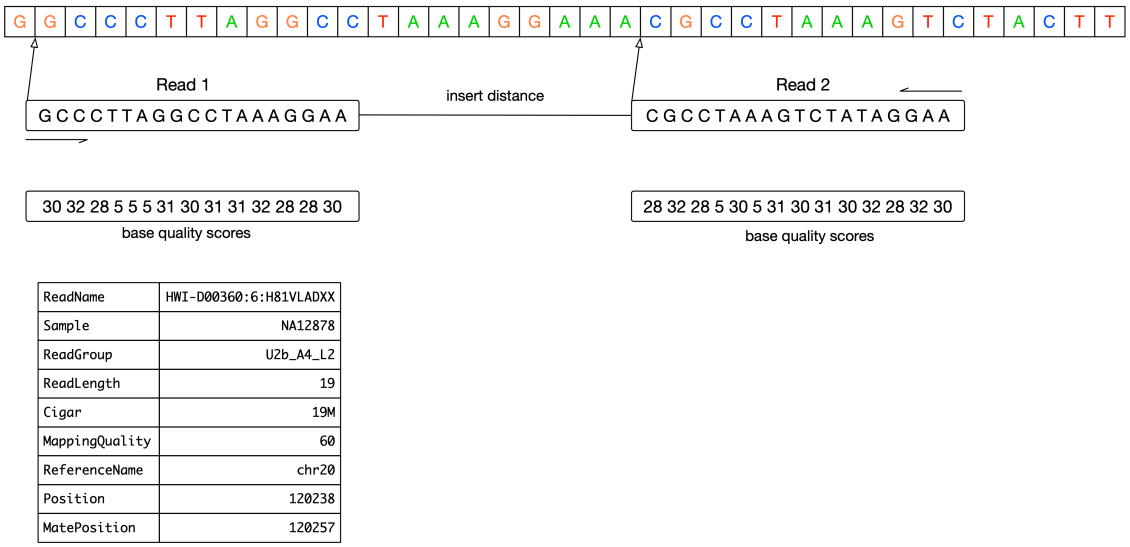


Figure 5.11: A read-pair that is aligned to the reference.

Each read is a tuple of the form:

$$r = (s\_id, r\_id, b, q, f_p) \tag{5.2}$$

where:

- s_id - is the sample ID, which is unique among all samples.
- r_id - is the read ID, which is unique among all reads for that sample.
- $b = \{b_1, b_2, ..., b_n\}$ - is the sequence of DNA bases, where $b_i \in \{A, C, G, T, N\}$.
- $q = \{q_1, q_2, ..., q_n\}$ - is the set of PHRED-scaled base quality scores corresponding to the probability that the base has been called incorrectly. See discussion on FASTQ format in Section 2.2.1 for details.
- $f_p \in \{True, False\}$ - is a boolean flag indicating whether this read is the first read in a pair.

5.3.1 Read QC Metrics

Section 2.2.3 of the Background chapter discusses various metrics of interest that are based on observations of read data and the tools that are used to collect them.

Here we describe how to collect the most typical metrics in the streaming paradigm of Rheos. As before, there are per-read metrics such as Base Quality Distribution, and Adapter Sequence Presence, as well as per-sample metrics such as Average GC Content, Insert Distribution, Read Length Distribution, and others. The utility of these metrics is to be able to set up filters for low quality data as well as input for downstream variant-calling models (see Section 2.2.6 for example).

Assume that we are observing a stream $M_{raw} = \{m_i : m_i = (header, payload)\}$ of read messages where the payload is a read r as defined above. Under the assumptions of Section 5.2 we know that the number of elements in the stream is unbounded. We are able to straightforwardly calculate incremental estimates for metrics such as mean, variance, max, and min, but require more sophisticated structures for computing estimates of rank statistics such as median and other quantiles to maintain operations in bounded space. We use the following update rules for min, max, mean, and variance[181] calculations:

$$min_k(M) = \begin{cases} m_k, & \text{if } m_k < min_{k-1}(M) \\ min_{k-1}(M), & \text{otherwise} \end{cases} \tag{5.3}$$

$$max_k(M) = \begin{cases} m_k, & \text{if } m_k > max_{k-1}(M) \\ max_{k-1}(M), & \text{otherwise} \end{cases} \tag{5.4}$$

$$\mu_k(M) = \mu_{k-1}(M) + \frac{m_k - \mu_{k-1}}{k} \tag{5.5}$$

$$\sigma_k^2(M) = \frac{\sigma_{k-1}^2(M) + (m_k - \mu_{k-1}(M))(m_k - \mu_k(M))}{k-1} \tag{5.6}$$

In order to set up the mechanisms to answer quantile queries the following definitions are used[57]:

- Given a set S of size n, and a quantile $\phi \in [0, 1]$, return $v \in S$ whose rank in sorted S is ϕn.
- An ϵ-approximate ϕ-quantile is a value v whose rank $r*(v) \in [n(\phi-\epsilon), n(\phi+\epsilon)]$.
- A quantile summary is $Q = q_1, q_2, ..., q_l : q_1 \leq q_2 \leq \cdots \leq q_l, q_i \in S, i \in [1, l]$ where each q_i has rank at least $rmin_Q(q_i)$ and at most $rmax_Q(q_i)$ in S, and $rmax_Q(q_1) \leq \epsilon|S|$, and $rmin_Q(q_l) \geq (1-\epsilon)|S|$.
- A quantile summary $Q(\epsilon)$ is ϵ-approximate if it can be used to answer any quantile query with ϵ-accuracy.

We use two approaches for computing quantile summaries, one due to Greenwald and Khanna[63] is able to compute the quantile summary using $O(log(\epsilon n)/\epsilon)$

space, and the other by Shrivastava et al.[157] computes the quantile summary in $O(log(M)/\epsilon)$ when the values are integers in range $[1, M]$. Both algorithms work for a scenario where one node sees all of the data in a stream, but can also be generalized to topologies where the stream is observed by multiple nodes in parallel.

We provide several examples of QC queries of interest that are specified on a data stream:

Average Base Quality - As an assessment of the individual quality of each read we are interested in the average base quality so that we can filter out reads that are of low quality as a whole. We use a Decorator Function construct from Section 5.2.1.

Input: $M_{raw} = \{m_i : m_i = (header, payload)\}$ where $m.payload = r = (s\_id, r\_id, b, q, f_p)$ as in 5.2. Operation: $q_{av} = f(r) = \frac{\sum_{i \in [1,|r|]} r.q_i}{|r|}$
Output: $M_{out} = \{m_i : m_i = (header, payload)\}$ where $m.payload = r = (s\_id, r\_id, b, q, f_p, q_{av})$

Base Quality Distribution - The distribution of base quality scores per base position of a read and per sample are of interest to investigate the presence of systemic biases in base quality scores as a function of the position within the read.

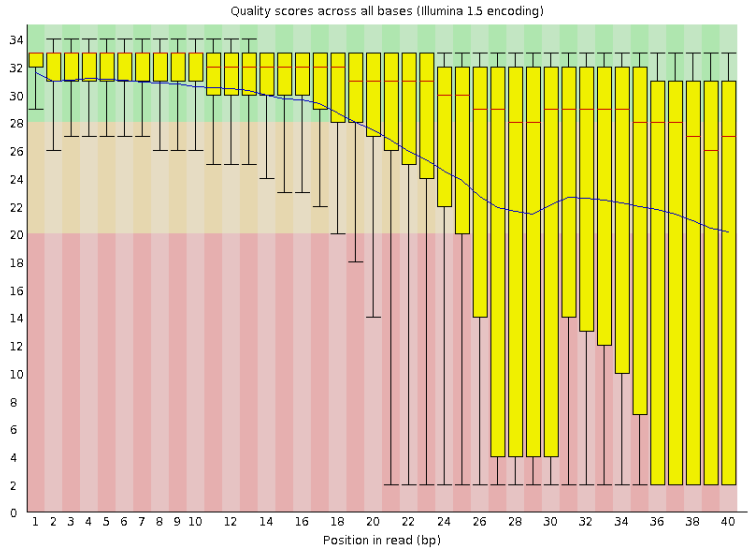


Figure 5.12: Distribution of base qualities per read position, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc. Large quality drop-off can be see towards the end of the read.

Because PHRED-scaled quality scores are integers that fall in a fixed range $q \in [0, 96]$ building quantile summaries using the q-gram[157] approach is the most space-efficient. Because base quality scores need to be aggregated over many reads and tracked for many samples, a service that implements this functionality needs to keep local state, and the operation to update the quantile summaries based on incoming reads follows the Local State Aggregator pattern from Section 5.2.1.

Since the local state required for storing the quantile summaries may not fit in memory and may need to be persisted to disk, updating the summaries may be too expensive to do for every single read that is observed in a read input stream. Instead, reads may be buffered into a set of reservoirs, triggering an update of the quantile summaries when the reservoir is full. The contents of the reservoir would then be purged and an updated set of quantile summaries $Q_s, bqg = \{q_i : i \in [1, max\_bases]\}$, where each q_i is a quantile summary corresponding to the Base Quality Distribution at a particular read position, issued to the output stream (see Algorithm 4).

Input: $M_{raw} = \{m_i : m_i = (header, payload)\}$ where $m.payload = r = (s\_id, r\_id, b, q, f_p)$ as in 5.2.
Operation: $f(r) = updateQuantileSummaries(r)$
Output: $M_{out} = \{m_i : m_i = (header, payload)\}$ where $m.payload = (s\_id, Q_{bqd})$.

Algorithm 4: Updating quantile summaries for Base Quality Distribution.

```
Function updateBQDQuantileSummaries(r) begin
    reservoir ←getReservoir(r.s_id)
    reservoir.addNewRead(r)
    if reservoir.isFull then
        summaries ←getQuantileSummaries(r.s_id)
        for read in reservoir do
            for index, read.q in read do
                updateQGram(summaries[index], read.q)          // per [157]
        purgeReservoir(reservoir)
        outputQuantileSummary(r.s_id, summaries)
```

Insert Size Distribution - The insert size distribution is an important metric because it is not only indicative of the overall quality of a sample's data, but it is also used by structural variant calling to find read-pairs that map abnormally far apart (indicating a deletion), or abnormally close together (indicating an insertion). Calculating this metric requires a stream of read-pairs, where both reads have been successfully mapped to the reference genome. Given a mapped read-pair (r_1, r_2) where each read has a beginning coordinate $r.pos$ and an end coordinate $r.end$, the insert size is $l = r_2.end - r_1.pos$. We are interested in the mean, variance and quantiles of the insert size distribution. Because the insert length can be any size the quantile summary method of Greenwald and Khanna[63] is most appropriate for the quantiles. Read pairs are observed on the input stream and buffered in per-sample reservoirs. When a reservoir is full the read pairs are used to update and output an appropriate insert size distribution mean, variance, and quantile summary.

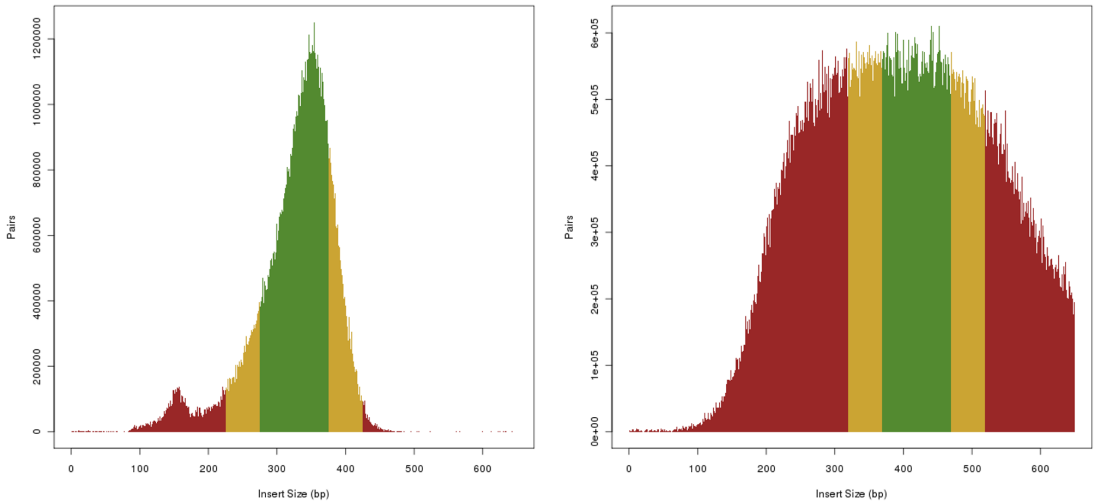


Figure 5.13: Distribution of insert sizes from two ICGC pancreatic cancer patients DO35138 and DO22154.[163]

Input: $M_{pair} = \{m_i : m_i = (header, payload)\}$ where $m.payload = (r_1, r_2)$, and $r = (s\_id, r\_id, b, q, f_p)$ as in 5.2.
Operation: $f(r) = updateInsertSizeDistribution(s\_id, r_1, r_2)$
Output: $M_{out} = \{m_i : m_i = (header, payload)\}$ where $m.payload = (s\_id, \mu_{isd}, \sigma^2_{isd}, Q_{isd})$.

Algorithm 5: Updating metrics for Insert Size Distribution.

```
Function updateInsertSizeDistribution(s_id, r1, r2) begin
    pairReservoir ←getReservoir(s_id)
    pairReservoir.addNewReadPair(r1, r2)
    if pairReservoir.isFull then
        summary ←getQuantileSummary(s_id)
        mu ←getMu(s_id)
        sigmaSq ←getSigmaSq(s_id)
        for (r1, r2) in pairReservoir do
            insertSize ←r2.end − r1.pos
            updateQuantileSummary(summary, insertSize)        /* per [63] */

            newMu ←updateMu(mu, insertSize)                   /* using Eq. 5.5 */

            newSigmaSq ←updateSigmaSQ(sigmaSq, mu, newMu, insertSize)
             /* using Eq. 5.6 */

        purgeReservoir(pairReservoir)
        outputInsertSizeDistribution(s_id, newMu, newSigmaSq, summary)
```

Other QC metrics, such as those measuring GC Content Distribution, Read Length Distribution, etc. can be collected analogously.

5.3.2 Alignment

5.3.3 Local Assembly

5.3.4 Simple SNP Calling

5.3.5 Assembly-based Variant Calling

5.3.6 Variant Filtering

5.3.7 Variant Annotation

5.3.8 Variant Output

Describe specific problems that need to be addressed by the system in a stream-based formulation - how to go from a collection of raw reads to a set of annotated variants: map reads, perform QC filtering, model loci, assemble alternative haplotypes, call variants, filter variants, annotate, produce output.

- Perform read QC
- Align reads to reference
- Collect read stream statistics
- Assemble local read contigs
- Model individual loci
- Call variants (maybe need to split by variant types)
- Genotype variants
- Filter variants
- Annotate variants
- Output variants

5.4 Services of Rheos

Provide a mapping from the domain-specific problems onto a particular implementation in the data streaming architecture. List services, their responsibilities, contracts, etc.

Metadata - Take in metadata related to patients, samples, files, etc. In: Metdata records, Out: ingestion confirmation events.

Read Streaming - Take data from outside the system (file, web service, etc) and turn it into a standard stream. In: files, external streams, Out: internal read stream.

Read Persistence - Store reads on disk, index. In: read stream, Out: persistence confirmation events.

Read Statistics - Look at read stream and calculate various approximate stats of interest - insert size, GC-bias, etc. In: read stream, Out: running stats of interest

QC - Compute QC score for reads. In: reads, Out: reads with QC score

Read Filtering - Filter out low quality reads based on configured parameters. In: reads with QC score, Out: filtered reads.

Read Mapping - Align reads to reference genome. In: stream of reads, Out: streams of mapped, unmapped, split reads.

Local Assembly - Local assembly of reads into candidate haplotypes. In: stream of aligned reads, Out: Updated haplotypes event.

Haplotype Persistence - Storage and lookup of candidate haplotypes. In: stream of reads, Out: persistence confirmation events, stream of haplotypes.

Variant Calling - Evaluate candidate haplotypes for presence of variation. In: haplotype update events, Out: variant update events.

Variant Persistence - Storage and lookup of variants. In: stream of reads, Out: persistence confirmation events, stream of variants.

Genotyping - Genotype variant sites. In: variant update stream, Out: genotype update events.

Variant Filtering - Filter out low quality variants. In: stream of variants, Out: filtered stream of variants.

Variant annotation - Annotate variants for functional impact. In: stream of variants, Out: stream of annotated variants.

Output variants - Format variants for external output. In: stream of variants, Out: files, external variant stream.

Notification - Notify the user when events of interest occur. In: any stream, Out: stream of notifications.

5.5 Proof of concept implementation

Describe actual implementation efforts. Focus on very basic use case (take already mapped reads from a file, turn them into a stream, use stream to call SNPs, maybe some Indels/SVs). Demonstrate some comparison metrics compared to other callers. Demonstrate some service-level metrics (throughput etc.)

5.6 Conclusions

Rheos is great!

Chapter 6

Discussion and Conclusion

6.1 Validation and Conclusion

We have deployed Butler in a production setting at the EMBL/EBI's Embassy Cloud in a configuration that utilizes 1500 CPUs, 6 TB RAM, 1 PB of Isilon storage accessed over NFS, and 40 TB of block-storage. Furthermore, we have built a series of workflows that facilitate the large-scale cancer genomics analyses carried out by the Germline Working Group of the Pan Cancer Analysis of Whole Genomes project, including:

- Germline SNV discovery
- Germline SNV joint-genotyping
- Germline SV genotyping
- Variant Filtration
- Sample submission

Using these workflows wSe have carried out a number of analyses on a 725TB data set of 2834 cancer patients' DNA samples consuming a total of 546,552 CPU hours. Each analysis took no longer than two weeks to complete and utilized only 1.5% - 2.2% of the overall compute capacity for management overhead. On several occasions we were able detect large scale cluster instability and program crashes utilizing the Operational Management system and take corrective action with a minimal impact on overall cluster productivity.

Subsequent to the success of these analyses several research groups from the European Bioinformatics Institute, Ontario Institute for Cancer Research, Francis Crick Institute, and the Centre for Genomic Regulation have expressed their interest in utilizing Butler for their own large scale analyses in the cloud.

Based on the adherence of the Butler design and implementation to the stated set of requirements, and sustained successful production operation in a large scale deployment on a multitude of scientific analyses of significant scope and size, we conclude that the Butler framework is an effective tool for large scale scientific workflow management in the cloud.

6.2 Future Direction

Butler has been created to facilitate scientific analyses at scale and we have demonstrated that it is able to successfully perform at the level required for today's big data initiatives in the genomics domain. There are projects on the horizon, however, that are one to two orders of magnitude larger than the current biggest projects, these include the UK's 100,000 Genomes Project[118], and the US Precision Medicine Initiative[22] (with up to 1,000,000 genomes). This means that in order to not have to proportionately increase the timeline for theses projects the computational infrastructure will have to be scaled up instead. It is thus imperative for Butler's continued relevance to be able to ascertain the framework's performance level at 1 or 2 orders of magnitude larger than the current 1500 core empirically obtained result. The most immediate opportunity to do so will come up in 2017 when the EMBL/EBI's Embassy Cloud will be upgraded to 5000 CPU cores and Butler has been invited to take part in the stress-testing of the upgraded cloud.

It is important to grow the library of workflows that are readily available for the Butler system to make the framework more appealing to new users. The Technical Working Group of the PCAWG project is in the process of migrating all of the main computational pipelines that have been used in the project into Docker[122] containers. Although the workflows that have been developed for the Germline Working Group have not yet been ported to Docker, Airflow, the workflow system underlying Butler has support for running Docker containers. Thus, a key next step for growing the library of Butler workflows lies in the adaptation of the core PCAWG workflows to be able to easily run them on a Butler instance. This would allow Butler to offer a comprehensive set of next generation sequencing workflows that are used for cancer genomics analysis.

Deploying Butler to a larger variety of environments will confirm the multi-cloud purpose of the framework and allow for the development of a richer set of configuration and provisioning profiles, as necessitated by the differences between deployment environments. On the basis of the already completed analyses for the PCAWG Germline Working Group, the Butler framework has also been selected to help deliver the science demonstrator work packet of the European Open Science Cloud Pilot[45] initiative that is launching in 2017. Additionally, de.NBI - The German Network for Bioinformatics Infrastructure[74] which is working to establish a German academic cloud computing environment for bioinformatics research will be using Butler to deliver a number of new bioinformatics pipelines on its cloud in 2017.

Thus, over the course of the next 12 months the focus of Butler development will be on supporting improved scalability, developing a richer set of computational pipelines and operating in a number of new cloud computing environments. These steps should result in a more robust, feature rich, and useful tool.

Appendix A

Code Listings

Listing 1: Terraform configuration of a worker VM

```
provider "openstack" {
        user_name = "${var.user_name}"
        password = "${var.password}"
        tenant_name = "${var.tenant_name}"
        auth_url = "${var.auth_url}"
}

resource "openstack_compute_instance_v2" "worker" {
          image_id = "${var.image_id}"
        flavor_name = "s1.massive"
        security_groups = ["internal"]
        name = "${concat("worker-", count.index)}"
        network = {
                uuid = "${var.main_network_id}"
        }
        connection {
                user = "${var.user}"
                 key_file = "${var.key_file}"
                 bastion_key_file = "${var.bastion_key_file}"
                 bastion_host = "${var.bastion_host}"
                 bastion_user = "${var.bastion_user}"
                 agent = "true"

        }
        count = "175"
        key_pair = "${var.key_pair}"
        provisioner "remote-exec" {
                inline = [
                        "sudo mv /home/centos/saltstack.repo
                        ↪  /etc/yum.repos.d/saltstack.repo",
                        "sudo yum install salt-minion -y",
```

```
                        "sudo service salt-minion stop",
                        "echo 'master: ${var.salt_master_ip}' | sudo tee  -a
                        ↪  /etc/salt/minion",
                        "echo 'id: ${concat("worker-", count.index)}' | sudo tee
                        ↪  -a /etc/salt/minion",
                        "echo 'roles: [worker, germline, consul-client]' | sudo
                        ↪  tee -a /etc/salt/grains",
                        "sudo hostname ${concat("worker-", count.index)}",
                        "sudo service salt-minion start"
                ]
        }
}
```

Listing 2: Terraform configuration of a security group

```
resource "openstack_compute_secgroup_v2" "internal" {
        name = "internal"
        description = "Allows communication between instances"
        #SSH
        rule {
                from_port = 22
                to_port = 22
                ip_protocol = "tcp"
                self = "true"
        }
        #Saltstack
        rule {
                from_port = 4505
                to_port = 4506
                ip_protocol = "tcp"
                self = "true"
        }
}
```

Listing 3: Salt Pillar for specifying test data location.

```
test_data_sample_path: /shared/data/samples

test_data_base_url: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/

test_samples:
  NA12874:
    -
      - NA12874.chrom11.ILLUMINA.bwa.CEU.low_coverage.20130415.bam
      - 88a7a346f0db1d3c14e0a300523d0243
```

```
    -
      - NA12874.chrom11.ILLUMINA.bwa.CEU.low_coverage.20130415.bam.bai
      - e61c0668bbaacdea2c66833f9e312bbb
```

Listing 4: Using Salt Mine to look up a server's IP Address.

```
consul-client:
  service.running:
    - enable: True
    - watch:
      - file: /etc/opt/consul.d/*
{%- set servers = salt['mine.get']('roles:(consul-server|consul-bootstrap)',
↪  'network.ip_addrs', 'grain_pcre').values() %}
{%- set node_ip = salt['grains.get']('ip4_interfaces')['eth0'] %}
# Create a list of servers that can be used to join the cluster
{%- set join_server = [] %}
{%- for server in servers if server[0] != node_ip %}
{% do join_server.append(server[0]) %}
{%- endfor %}
join-cluster:
  cmd.run:
    - name: consul join {{ join_server[0] }}
    - watch:
      - service: consul-client
```

Listing 5: Using Top File to map States to Roles.

```
base:
  '*':
    - consul
    - dnsmasq
    - collectd
  'G@roles:monitoring-server':
    - influxdb
    - grafana
  'G@roles:job-queue':
    - rabbitmq
```

Listing 6: Collectd configuration for metrics collection.

```
# Read metrics about cpu usage
[[inputs.cpu]]
  ## Whether to report per-cpu stats or not
  percpu = true
```

```
  ## Whether to report total system cpu stats or not
  totalcpu = true
  ## If true, collect raw CPU time metrics.
  collect_cpu_time = false
  ## If true, compute and report the sum of all non-idle CPU states.
  report_active = false


# Read metrics about disk usage by mount point
[[inputs.disk]]
  ## By default, telegraf gather stats for all mountpoints.
  ## Setting mountpoints will restrict the stats to the specified mountpoints.
  # mount_points = ["/"]

  ## Ignore some mountpoints by filesystem type. For example (dev)tmpfs (usually
  ## present on /run, /var/run, /dev/shm or /dev).
  ignore_fs = ["tmpfs", "devtmpfs", "devfs"]


# Read metrics about disk IO by device
[[inputs.diskio]]
  ## By default, telegraf will gather stats for all devices including
  ## disk partitions.
  ## Setting devices will restrict the stats to the specified devices.
  # devices = ["sda", "sdb"]
  ## Uncomment the following line if you need disk serial numbers.
  # skip_serial_number = false
  #
  ## On systems which support it, device metadata can be added in the form of
  ## tags.
  ## Currently only Linux is supported via udev properties. You can view
  ## available properties for a device by running:
  ## 'udevadm info -q property -n /dev/sda'
  # device_tags = ["ID_FS_TYPE", "ID_FS_USAGE"]
  #
  ## Using the same metadata source as device_tags, you can also customize the
  ## name of the device via templates.
  ## The 'name_templates' parameter is a list of templates to try and apply to
  ## the device. The template may contain variables in the form of '$PROPERTY' or
  ## '${PROPERTY}'. The first template which does not contain any variables not
  ## present for the device is used as the device name tag.
  ## The typical use case is for LVM volumes, to get the VG/LV name instead of
  ## the near-meaningless DM-0 name.
  # name_templates = ["$ID_FS_LABEL","$DM_VG_NAME/$DM_LV_NAME"]


# Get kernel statistics from /proc/stat
[[inputs.kernel]]
```

```
  # no configuration


# Read metrics about memory usage
[[inputs.mem]]
  # no configuration


# Get the number of processes and group them by status
[[inputs.processes]]
  # no configuration


# Read metrics about swap memory usage
[[inputs.swap]]
  # no configuration


# Read metrics about system load & uptime
[[inputs.system]]
  # no configuration
```

Listing 7: Filebeat Prospector configuration.

```
[collectd]
  enabled = true
  bind-address = ":8096"
  database = "metrics"
  retention-policy = ""
  batch-size = 5000
  batch-pending = 10
  batch-timeout = "10s"
  read-buffer = 0
  typesdb = "/usr/share/collectd/types.db"
```

Listing 8: TICKscript for alerting on CPU value.

```
// Parameters
var info = 70
var warn = 80
var crit = 90
var infoSig = 2.5
var warnSig = 3
var critSig = 3.5
var period = 10s
```

```
var every = 10s

// Dataframe
var data = stream
  |from()
    .database('metrics')
    .retentionPolicy('default')
    .measurement('cpu_value')
    .where(lambda: "type" == 'percent' AND "type_instance" == 'idle')
  |eval(lambda: 100 - "value")
    .as('used')
  |window()
    .period(period)
    .every(every)
  |mean('used')
    .as('stat')

// Thresholds
var alert = data
  |eval(lambda: sigma("stat"))
    .as('sigma')
    .keep()
  |alert()
    .id('{{ index .Tags "host"}}/cpu_value')
    .message('{{ .ID }}:{{ index .Fields "stat" }}')
    .info(lambda: "stat" > info OR "sigma" > infoSig)
    .warn(lambda: "stat" > warn OR "sigma" > warnSig)
    .crit(lambda: "stat" > crit OR "sigma" > critSig)

// Alert
alert
  .log('/tmp/cpu_alert_log_2.txt')
```

Listing 9: TICKscript for handling dead VMs.

```
{% raw %}
var db = 'telegraf'
var rp = 'autogen'
var measurement = 'system'
var groupBy = ['host']
var whereFilter = lambda: TRUE
var period = 30s
var name = 'Host Deadman'
var idVar = name + ':{{.Group}}'
var blah = '{{index .Tags "host"}}'
var message = 'The host {{index .Tags "host"}} is offline as of {{.Time}}.'
var messageN = 'The host {{index .Tags "host"}} is back online at {{.Time}}.'
```

```
13 var idTag = 'alertID'
14 var levelTag = 'level'
15 var messageField = 'message'
16 var durationField = 'duration'
17 var outputDB = 'chronograf'
18 var outputRP = 'autogen'
19 var outputMeasurement = 'alerts'
20 var triggerType = 'deadman'
21 var threshold = 0.0
22 var data = stream
23     |from()
24         .database(db)
25         .retentionPolicy(rp)
26         .measurement(measurement)
27         .groupBy(groupBy)
28         .where(whereFilter)
29
30 var trigger = data
31     |deadman(threshold, period)
32         .stateChangesOnly()
33         .message('{{ if eq .Level "CRITICAL" }}' + message + '{{else}}' +
           ↪  messageN + '{{end}}')
34         .id(idVar)
35         .idTag(idTag)
36         .levelTag(levelTag)
37         .messageField(messageField)
38         .durationField(durationField)
39         .slack()
40         .channel('#embassyalerts')
41 {% endraw %}
42         .exec('butler_healing_agent', 'relaunch-worker', '-t', '{{
           ↪  pillar['terraform_files'] }}', '-s', '{{ pillar['terraform_state']
           ↪  }}', '-v', '{{ pillar['terraform_vars'] }}', '-p', '{{
           ↪  pillar['terraform_provider'] }}')
43 {% raw %}
44 trigger
45     |eval(lambda: "emitted")
46         .as('value')
47         .keep('value', messageField, durationField)
48     |influxDBOut()
49         .create()
50         .database(outputDB)
51         .retentionPolicy(outputRP)
52         .measurement(outputMeasurement)
53         .tag('alertName', name)
54         .tag('triggerType', triggerType)
55
56 trigger
```

```
    |httpOut('output')
{% endraw %}
```

Listing 10: Butler healing agent code for restarting the Airflow Scheduler.

```
def call_command(command, cwd=None):
  try:
      logging.debug("About to invoke command: " + command)
      my_output = check_output(command, shell=True, cwd=cwd, stderr=STDOUT)
      logging.debug("Command output is: " + my_output)
      return my_output
  except CalledProcessError as e:
      logging.error("An error occurred! Command output is: " +
      ↪  e.output.decode("utf-8") )
      raise

def is_critical(level):
  return level == "CRITICAL"

def parse_alert_data():
  return json.loads(sys.stdin.read())

def get_host_name(alert_data):
  return alert_data["data"]["series"][0]["tags"]["host"]

def restart_service(host, service_name):
  call_command("pepper {} service.restart {}".format(host, service_name), None)

def parse_args():
  my_parser = argparse.ArgumentParser()

  sub_parsers = my_parser.add_subparsers()

  common_args_parser = argparse.ArgumentParser(
      add_help=False, conflict_handler='resolve')

  restart_airflow_scheduler_parser = sub_parsers.add_parser(
      "restart-airflow-scheduler", parents=[common_args_parser],
      ↪  conflict_handler='resolve')

  ↪  restart_airflow_scheduler_parser.set_defaults(func=restart_airflow_scheduler_command)

def restart_airflow_scheduler_command(args, alert_data):
  if is_critical(alert_data["level"]):
      restart_service("-G 'roles:tracker'", "airflow-scheduler")
```

Listing 11: Butler healing agent code for relaunching a failed VM.

```
def call_command(command, cwd=None):
  try:
      logging.debug("About to invoke command: " + command)
      my_output = check_output(command, shell=True, cwd=cwd, stderr=STDOUT)
      logging.debug("Command output is: " + my_output)
      return my_output
  except CalledProcessError as e:
      logging.error("An error occurred! Command output is: " +
      ↪  e.output.decode("utf-8") )
      raise

def is_critical(level):
  return level == "CRITICAL"

def parse_alert_data():
  return json.loads(sys.stdin.read())

def get_host_name(alert_data):
  return alert_data["data"]["series"][0]["tags"]["host"]

def restart_service(host, service_name):
  call_command("pepper {} service.restart {}".format(host, service_name), None)

def parse_args():
  my_parser = argparse.ArgumentParser()

  sub_parsers = my_parser.add_subparsers()

  common_args_parser = argparse.ArgumentParser(
    add_help=False, conflict_handler='resolve')

  relaunch_worker_parser = sub_parsers.add_parser(
      "relaunch-worker", parents=[common_args_parser],
      ↪  conflict_handler='resolve')
  relaunch_worker_parser.add_argument(
      "-t", "--terraform_location", help="Location of the terraform definition
      ↪  files.",
      dest="terraform_location", required=True)
  relaunch_worker_parser.add_argument(
      "-s", "--terraform_state_location", help="Location of the terraform state
      ↪  file.",
      dest="terraform_state_location", required=True)
  relaunch_worker_parser.add_argument(
      "-v", "--terraform_var_file_location", help="Location of the terraform vars
      ↪  file.",
      dest="terraform_var_file_location", required=True)
  relaunch_worker_parser.add_argument(
      "-p", "--terraform_provider", help="The terraform provider to use.",
```

```
44          choices = provider_list,
45          dest="terraform_provider", required=True)
46      relaunch_worker_parser.set_defaults(func=relaunch_worker_command)
47
48      def is_key_present(key_data, host_name):
49      parsed_key_data = json.loads(key_data)
50      return_data = parsed_key_data["return"][0]["data"]["return"]
51
52      if "minions" in return_data:
53          return_vals = return_data["minions"]
54          for val in return_vals:
55              if val == host_name:
56                  return True
57
58      return False
59
60  def locate_minon_key(host_name):
61      minion_connect_try = 1
62      while minion_connect_try <= MINION_CONNECT_MAX_RETRIES:
63          logging.info("Attempt #{} of {} to retrieve minion key for host {} from the
            ↪  master.".format(minion_connect_try, MINION_CONNECT_MAX_RETRIES,
            ↪  host_name))
64          key_data = call_command("pepper --client=wheel key.name_match
            ↪  match={}".format(host_name))
65          logging.debug("Retrieved key data: " + key_data)
66          if is_key_present(key_data, host_name):
67              return True
68          else:
69              logging.debug("Key data for host {} not found at time {}. Sleeping for
                ↪  {} seconds.".format(host_name, datetime.now(),
                ↪  MINION_CONNECT_SLEEP_PERIOD))
70              time.sleep(MINION_CONNECT_SLEEP_PERIOD)
71              minion_connect_try = minion_connect_try + 1
72
73      return False
74
75  def relaunch_worker_command(args, alert_data):
76      if is_critical(alert_data["level"]):
77          host_name = get_host_name(alert_data)
78
79
80          tf_location = args.terraform_location
81          tf_state_location = args.terraform_state_location
82          tf_var_file_location = args.terraform_var_file_location
83          tf_resource = provider_resource_lookup[args.terraform_provider]
84          worker_number = host_name.split("-")[1]
85
86          call_command("pepper --client=wheel key.delete match={}".format(host_name))
```

```
        call_command("terraform taint -lock=false -state={}
        ↪  {}.worker.{}".format(tf_state_location, tf_resource, worker_number),
        ↪  tf_location)
        call_command("terraform apply -lock=false -state={} --var-file {}
        ↪  -auto-approve".format(tf_state_location, tf_var_file_location),
        ↪  tf_location)

        locate_minon_key(host_name)

        call_command("pepper '*' mine.update")
        call_command("pepper {} state.apply dnsmasq".format(host_name))
        call_command("pepper {} state.apply consul".format(host_name))
        call_command("pepper {} state.highstate".format(host_name))
```

Listing 12: Consul service definition for PostgreSQL.

```
{
        "results_base_path": "/shared/data/results/discovery/",
        "results_local_path": "/tmp/discovery/",
        "freebayes": {
                "mode": "discovery",
                "flags": "--min-repeat-entropy 1
                ↪  --report-genotype-likelihood-max"
        }
}
```

Listing 13: Source code for the freebayes workflow.

```
from airflow import DAG
from airflow.operators import BashOperator, PythonOperator
from datetime import datetime, timedelta

import os
import logging
from subprocess import call

import tracker.model
from tracker.model.analysis_run import *
from tracker.util.workflow_common import *


def run_freebayes(**kwargs):

    config = get_config(kwargs)
    logger.debug("Config - {}".format(config))
```

```
18
19      sample = get_sample(kwargs)
20
21      contig_name = kwargs["contig_name"]
22      contig_whitelist = config.get("contig_whitelist")
23
24
25      if not contig_whitelist or contig_name in contig_whitelist:
26
27          sample_id = sample["sample_id"]
28          sample_location = sample["sample_location"]
29
30          result_path_prefix = config["results_local_path"] + "/" + sample_id
31
32          if (not os.path.isdir(result_path_prefix)):
33              logger.info(
34                  "Results directory {} not present,
                    ↪  creating.".format(result_path_prefix))
35              os.makedirs(result_path_prefix)
36
37          result_filename = "{}/{}_{}.vcf".format(
38              result_path_prefix, sample_id, contig_name)
39
40          freebayes_path = config["freebayes"]["path"]
41          freebayes_mode = config["freebayes"]["mode"]
42          freebayes_flags = config["freebayes"]["flags"]
43
44          reference_location = config["reference_location"]
45
46          if freebayes_flags == None:
47              freebayes_flags = ""
48
49          if freebayes_mode == "discovery":
50              freebayes_command = "{} -r {} -f {} {} {} > {}".\
51                  format(freebayes_path,
52                         contig_name,
53                         reference_location,
54                         freebayes_flags,
55                         sample_location,
56                         result_filename)
57          elif freebayes_mode == "regenotyping":
58              variants_location = config["variants_location"]
59
60              freebayes_command = "{} -r {} -f {} -@ {} {} {} > {}".\
61                  format(freebayes_path,
62                         contig_name,
63                         reference_location,
64                         variants_location[contig_name],
```

```
                        freebayes_flags,
                        sample_location,
                        result_filename)
        else:
             raise ValueError("Unknown or missing freebayes_mode -
             ↪  {}".format(freebayes_mode))

        call_command(freebayes_command, "freebayes")

        compressed_sample_filename = compress_sample(result_filename, config)
        generate_tabix(compressed_sample_filename, config)
        copy_result(compressed_sample_filename, sample_id, config)
    else:
        logger.info(
            "Contig {} is not in the contig whitelist,
            ↪  skipping.".format(contig_name))


default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': datetime.datetime(2020, 01, 01),
    'email': ['airflow@airflow.com'],
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 1,
    'retry_delay': timedelta(minutes=5),
}

dag = DAG("freebayes", default_args=default_args,
          schedule_interval=None, concurrency=10000, max_active_runs=2000)


start_analysis_run_task = PythonOperator(
    task_id="start_analysis_run",
    python_callable=start_analysis_run,
    provide_context=True,
    dag=dag)


validate_sample_task = PythonOperator(
    task_id="validate_sample",
    python_callable=validate_sample,
    provide_context=True,
    dag=dag)

validate_sample_task.set_upstream(start_analysis_run_task)

```

```
complete_analysis_run_task = PythonOperator(
    task_id="complete_analysis_run",
    python_callable=complete_analysis_run,
    provide_context=True,
    dag=dag)

for contig_name in tracker.util.workflow_common.CONTIG_NAMES:
    freebayes_task = PythonOperator(
        task_id="freebayes_" + contig_name,
        python_callable=run_freebayes,
        op_kwargs={"contig_name": contig_name},
        provide_context=True,
        dag=dag)

    freebayes_task.set_upstream(validate_sample_task)

    complete_analysis_run_task.set_upstream(freebayes_task)
```

Listing 14: Saltstack state for workflow deployment.

```
current_runs = run_session.query(Configuration.config[("sample","
↪  sample_id")].astext).\
        join(AnalysisRun, AnalysisRun.config_id == Configuration.config_id).\
        join(Analysis, Analysis.analysis_id == AnalysisRun.analysis_id).\
        filter(and_(Analysis.analysis_id == analysis_id, AnalysisRun.run_status
        ↪  != tracker.model.analysis_run.RUN_STATUS_ERROR)).all()

available_samples = sample_session.query(PCAWGSample.index.label("index"),
↪  sample_id.label("sample_id"), sample_location.label("sample_location")).\
        join(SampleLocation, PCAWGSample.index == SampleLocation.donor_index).\
        filter(and_(sample_location != None, sample_id.notin_(current_runs))).\
        limit(num_runs).all()
```

Listing 15: Butler Analysis configuration for SNP genotyping.

```
{
        "variants_location": {
            "1": "/freebayes.chr_1.sites.snv_indel.annot.final.vcf.gz",
            "2": "/freebayes.chr_2.sites.snv_indel.annot.final.vcf.gz",
            "3": "/freebayes.chr_3.sites.snv_indel.annot.final.vcf.gz",
            "4": "/freebayes.chr_4.sites.snv_indel.annot.final.vcf.gz",
            "5": "/freebayes.chr_5.sites.snv_indel.annot.final.vcf.gz",
            "6": "/freebayes.chr_6.sites.snv_indel.annot.final.vcf.gz",
            "7": "/freebayes.chr_7.sites.snv_indel.annot.final.vcf.gz",
            "8": "/freebayes.chr_8.sites.snv_indel.annot.final.vcf.gz",
```

```
            "9": "/freebayes.chr_8.sites.snv_indel.annot.final.vcf.gz",
            "10": "/freebayes.chr_10.sites.snv_indel.annot.final.vcf.gz",
            "11": "/freebayes.chr_11.sites.snv_indel.annot.final.vcf.gz",
            "12": "/freebayes.chr_12.sites.snv_indel.annot.final.vcf.gz",
            "13": "/freebayes.chr_13.sites.snv_indel.annot.final.vcf.gz",
            "14": "/freebayes.chr_14.sites.snv_indel.annot.final.vcf.gz",
            "15": "/freebayes.chr_15.sites.snv_indel.annot.final.vcf.gz",
            "16": "/freebayes.chr_16.sites.snv_indel.annot.final.vcf.gz",
            "17": "/freebayes.chr_17.sites.snv_indel.annot.final.vcf.gz",
            "18": "/freebayes.chr_18.sites.snv_indel.annot.final.vcf.gz",
            "19": "/freebayes.chr_19.sites.snv_indel.annot.final.vcf.gz",
            "20": "/freebayes.chr_20.sites.snv_indel.annot.final.vcf.gz",
            "21": "/freebayes.chr_21.sites.snv_indel.annot.final.vcf.gz",
            "22": "/freebayes.chr_22.sites.snv_indel.annot.final.vcf.gz",
            "X": "/freebayes.chr_X.sites.snv_indel.annot.final.vcf.gz",
            "Y": "/freebayes.chr_Y.sites.snv_indel.annot.final.vcf.gz"
        },
        "results_base_path":
        ↪  "/shared/data/results/regenotype_freebayes_discovery/",
        "results_local_path": "/tmp/regenotype_freebayes_discovery/",
        "freebayes": {
                "mode": "regenotyping",
                "flags": "-l"
        }
}
```

Listing 16: Butler Workflow configuration for Data Submission.

```
{
        "gnos": {
                "ebi": {
                        "url": "https://gtrepo-ebi.annailabs.com"
                },
                "osdc_icgc": {
                        "url": "https://gtrepo-osdc-icgc.annailabs.com"
                },
                "osdc_tcga": {
                        "url": "https://gtrepo-osdc-tcga.annailabs.com"
                }
        },
        "rsync": {
                "flags": "-a -v --remove-source-files"
        }
}
```

Listing 17: Butler Analysis configuration for Data Submission.

```
{
        "gnos": {
                "ebi": {
                        "key_location":
                        ↪  "/home/airflow/.ssh/sergei_pcawg_gnos_icgc.pem"
                },
                "osdc_icgc": {
                        "key_location":
                        ↪  "/home/airflow/.ssh/sergei_pcawg_gnos_icgc.pem"
                },
                "osdc_tcga": {
                        "key_location":
                        ↪  "/home/airflow/.ssh/sergei_bionimbus_gnos_may.pem"
                }
        },
        "metadata_template_location": "/opt/pcawg-germline/workflows/gtupload-wo
        ↪  rkflow/analysis_template.xml",
        "submission_base_path":
        ↪  "/shared/data/results/freebayes_discovery_gnos_submission/",
        "destination_repo_mapping": {
                "ICGC": "ebi",
                "TCGA": "osdc_tcga"
        }
}
```

Listing 18: Python code for the run_delly function which implements the functionality of the delly_genotype task inside the Butler Delly Workflow.

```
def run_delly(**kwargs):

    config = get_config(kwargs)
    sample = get_sample(kwargs)

    sample_id = sample["sample_id"]
    sample_location = sample["sample_location"]

    result_path_prefix = config["results_local_path"] + "/" + sample_id

    if (not os.path.isdir(result_path_prefix)):
        logger.info(
            "Results directory {} not present,
            ↪  creating.".format(result_path_prefix))
        os.makedirs(result_path_prefix)

    delly_path = config["delly"]["path"]
    reference_location = config["reference_location"]
```

```
    variants_location = config["variants_location"]
    variants_type = config["variants_type"]
    exclude_template_path = config["delly"]["exclude_template_path"]

    result_filename = "{}/{}_{}.bcf".format(
        result_path_prefix, sample_id, variants_type)

    log_filename = "{}/{}_{}.log".format(
        result_path_prefix, sample_id, variants_type)

    delly_command = "{} call -t {} -g {} -v {} -o {} -x {} {} > {}".\
        format(delly_path,
               variants_type,
               reference_location,
               variants_location,
               result_filename,
               exclude_template_path,
               sample_location,
               log_filename)

    call_command(delly_command, "delly")

    copy_result(result_filename, sample_id, config)
```

Listing 19: Butler Delly Workflow analysis configuration to genotype deletions.

```
{
        "variants_location": "/delly_deletion_sites/del.sites.bcf",
        "results_base_path":
        ↪  "/shared/data/results/delly_germline_deletions_14_07_2016/",
        "results_local_path": "/tmp/delly_germline_deletions/",
        "variants_type": "DEL"

}
```

Listing 20: Example of a VCF file (from https://samtools.github.io/hts-specs/VCFv4.3.pdf).

```
{
  ##fileformat=VCFv4.3
  ##fileDate=20090805
  ##source=myImputationProgramV3.1
  ##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
  ##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb
  ↪  2da,species="Homo
  ↪  sapiens",taxonomy=x>
```

```
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ
↪  0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50
↪  0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB
↪  GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60
↪  0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4
↪  0/2:17:2 1/1:40:3
```

Bibliography

[1] URL: http://www.langmead-lab.org/teaching-materials/.

[2] URL: https://software.broadinstitute.org/gatk/documentation.

[3] Enis Afgan et al. "Galaxy CloudMan: delivering cloud compute clusters". In: *BMC bioinformatics* 11.12 (2010), p. 1.

[4] Charu C Aggarwal. *Data streams: models and algorithms*. Vol. 31. Springer Science & Business Media, 2007.

[5] Daniel Aird et al. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries". In: *Genome biology* 12.2 (2011), R18.

[6] Tyler S Alioto et al. "A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing". In: *Nature communications* 6 (2015), p. 10001.

[7] Can Alkan, Bradley P Coe, and Evan E Eichler. "Genome structural variation discovery and genotyping". In: *Nature Reviews Genetics* 12.5 (2011), p. 363.

[8] Carmen J Allegra et al. "American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti–epidermal growth factor receptor monoclonal antibody therapy". In: *Journal of clinical oncology* 27.12 (2009), pp. 2091–2096.

[9] Simon Andrews et al. "FastQC: a quality control tool for high throughput sequence data". In: (2010).

[10] Brian Babcock et al. "Models and issues in data stream systems". In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM. 2002, pp. 1–16.

[11] Ralph M Barnes. "Motion and time study." In: (1949).

[12] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[13] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[14] Kym M Boycott et al. "Rare-disease genetics in the era of next-generation sequencing: discovery to translation". In: *Nature Reviews Genetics* 14.10 (2013), pp. 681–691.

[15] Benoit G Bruneau. “The developmental genetics of congenital heart disease”. In: *Nature* 451.7181 (2008), pp. 943–948.

[16] Michael Burrows and David J Wheeler. “A block-sorting lossless data compression algorithm”. In: (1994).

[17] Rajkumar Buyya et al. “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility”. In: *Future Generation computer systems* 25.6 (2009), pp. 599–616.

[18] Xiaoyu Chen et al. “Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications”. In: *Bioinformatics* 32.8 (2015), pp. 1220–1222.

[19] Kristian Cibulskis et al. “ContEst: estimating cross-contamination of human samples in next-generation sequencing data”. In: *Bioinformatics* 27.18 (2011), pp. 2601–2602.

[20] Kristian Cibulskis et al. “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples”. In: *Nature biotechnology* 31.3 (2013), pp. 213–219.

[21] Peter JA Cock et al. “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants”. In: *Nucleic acids research* 38.6 (2009), pp. 1767–1771.

[22] Francis S Collins and Harold Varmus. “A new initiative on precision medicine”. In: *New England Journal of Medicine* 372.9 (2015), pp. 793–795.

[23] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. “How to apply de Bruijn graphs to genome assembly”. In: *Nature biotechnology* 29.11 (2011), p. 987.

[24] 1000 Genomes Project Consortium et al. “A map of human genome variation from population-scale sequencing”. In: *Nature* 467.7319 (2010), pp. 1061–1073.

[25] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pp. 57–74.

[26] International HapMap Consortium et al. “The international HapMap project”. In: *Nature* 426.6968 (2003), p. 789.

[27] Wellcome Trust Case Control Consortium et al. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”. In: *Nature* 447.7145 (2007), p. 661.

[28] Charles E Cook et al. “The european bioinformatics institute in 2016: Data growth and integration”. In: *Nucleic acids research* 44.D1 (2016), pp. D20–D26.

[29] Georgiana Copil et al. “Multi-level elasticity control of cloud services”. In: *International Conference on Service-Oriented Computing*. Springer. 2013, pp. 429–436.

[30] Vasa Curcin and Moustafa Ghanem. “Scientific workflow systems-can one size fit all?” In: *2008 Cairo International Biomedical Engineering Conference*. IEEE. 2008, pp. 1–9.

[31] Mark J Daly et al. “High-resolution haplotype structure in the human genome”. In: *Nature genetics* 29.2 (2001), p. 229.

[32] Petr Danecek et al. “The variant call format and VCFtools”. In: *Bioinformatics* 27.15 (2011), pp. 2156–2158.

[33] Charles Darwin. *On the origin of species*. D. Appleton and Co., 1871. DOI: 10.5962/bhl.title.28875.

[34] Mayur Datar et al. “Maintaining stream statistics over sliding windows”. In: *SIAM journal on computing* 31.6 (2002), pp. 1794–1813.

[35] J Davis II et al. *Overview of the Ptolemy project*. Tech. rep. ERL Technical Report UCB/ERL, 1999.

[36] Mark A DePristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. In: *Nature genetics* 43.5 (2011), pp. 491–498.

[37] Wil MP van Der Aalst et al. “Workflow patterns”. In: *Distributed and parallel databases* 14.1 (2003), pp. 5–51.

[38] Juliane C Dohm et al. “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing”. In: *Nucleic acids research* 36.16 (2008), e105–e105.

[39] Richard Durbin et al. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[40] Mark TW Ebbert et al. “Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches”. In: *BMC bioinformatics* 17.7 (2016), p. 239.

[41] Genomics England. “The 100,000 genomes project”. In: *The* 100 (2016), pp. 1–2.

[42] Thomas Erl. *Service-oriented architecture (SOA): concepts, technology, and design*. 2005.

[43] Opher Etzion, Peter Niblett, and David C Luckham. *Event processing in action*. Manning Greenwich, 2011.

[44] Patrick Th Eugster et al. “The many faces of publish/subscribe”. In: *ACM computing surveys (CSUR)* 35.2 (2003), pp. 114–131.

[45] *European Open Science Cloud | Open Science - Research and Innovation - European Commission*. URL: http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud (visited on 10/31/2016).

[46] Kelly R Ewen et al. “Identification and analysis of error types in high-throughput genotyping”. In: *The American Journal of Human Genetics* 67.3 (2000), pp. 727–736.

[47] Warren J Ewens. “The sampling theory of selectively neutral alleles”. In: *Theoretical population biology* 3.1 (1972), pp. 87–112.

[48] Michael Farrar. “Striped Smith–Waterman speeds database searches six times over other SIMD implementations”. In: *Bioinformatics* 23.2 (2006), pp. 156–161.

[49] Gregory G Faust and Ira M Hall. "YAHA: fast and flexible long-read alignment with optimal breakpoint detection". In: *Bioinformatics* 28.19 (2012), pp. 2417–2424.

[50] Paolo Ferragina and Giovanni Manzini. "Opportunistic data structures with applications". In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE. 2000, pp. 390–398.

[51] Simon A Forbes et al. "COSMIC: exploring the world's knowledge of somatic mutations in human cancer". In: *Nucleic acids research* 43.D1 (2015), pp. D805–D811.

[52] G David Forney. "The viterbi algorithm". In: *Proceedings of the IEEE* 61.3 (1973), pp. 268–278.

[53] Rosalind E Franklin and Raymond G Gosling. "Molecular configuration in sodium thymonucleate". In: *Nature* 171 (1953), pp. 740–741.

[54] Markus Hsi-Yang Fritz et al. "Efficient storage of high throughput DNA sequencing data using reference-based compression". In: *Genome research* 21.5 (2011), pp. 734–740.

[55] Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. "Mining data streams: a review". In: *ACM Sigmod Record* 34.2 (2005), pp. 18–26.

[56] Hector Garcia-Molina, Frank Germano, and Walter H Kohler. "Debugging a distributed computing system". In: *IEEE Transactions on Software Engineering* 2 (1984), pp. 210–219.

[57] Minos Garofalakis, Johannes Gehrke, and Rajeev Rastogi. *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2016.

[58] Erik Garrison and Gabor Marth. "Haplotype-based variant detection from short-read sequencing". In: *arXiv preprint arXiv:1207.3907* (2012).

[59] *Glossary V1 - EGIWiki*. URL: `https://wiki.egi.eu/wiki/Glossary_V1` (visited on 10/28/2016).

[60] Jeremy Goecks, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". In: *Genome biology* 11.8 (2010), p. 1.

[61] A Gordon and GJ Hannon. "Fastx-toolkit". In: *FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit* 5 (2010).

[62] Christopher Greenman et al. "Patterns of somatic mutation in human cancer genomes". In: *Nature* 446.7132 (2007), pp. 153–158.

[63] Michael Greenwald and Sanjeev Khanna. "Space-efficient online computation of quantile summaries". In: *ACM SIGMOD Record*. Vol. 30. 2. ACM. 2001, pp. 58–66.

[64] Mendel Gregor. "Versuche uber Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines in Brunn". In: 4 (1865), pp. 3–47.

[65] Robert L Grossman et al. "An overview of the open science data cloud". In: *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. ACM. 2010, pp. 377–384.

[66] Daniel F Gudbjartsson et al. “Large-scale whole-genome sequencing of the Icelandic population”. In: *Nature genetics* 47.5 (2015), pp. 435–444.

[67] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[68] Douglas Hanahan and Robert A Weinberg. “Hallmarks of cancer: the next generation”. In: *cell* 144.5 (2011), pp. 646–674.

[69] Michael J Heller. “DNA microarray technology: devices, systems, and applications”. In: *Annual review of biomedical engineering* 4.1 (2002), pp. 129–153.

[70] Robert L Henderson. “Job scheduling under the portable batch system”. In: *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer. 1995, pp. 279–294.

[71] Joel N Hirschhorn and Mark J Daly. “Genome-wide association studies for common diseases and complex traits”. In: *Nature Reviews Genetics* 6.2 (2005), p. 95.

[72] Alan Hodgkinson and Adam Eyre-Walker. “Human triallelic sites: evidence for a new mutational mechanism?” In: *Genetics* 184.1 (2010), pp. 233–241.

[73] David Hollingsworth. “The workflow reference model”. In: (1995).

[74] *Home*. URL: `https://www.denbi.de/` (visited on 10/31/2016).

[75] Eun Pyo Hong and Ji Wan Park. “Sample size and statistical power calculation in genetic association studies”. In: *Genomics & informatics* 10.2 (2012), pp. 117–122.

[76] David B Ingham, Fabio Panzieri, and Santosh K Shrivastava. “Constructing dependable Web services”. In: *Advances in Distributed Systems*. Springer, 2000, pp. 277–294.

[77] Vernon M Ingram et al. “Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin”. In: *Nature* 180.4581 (1957), pp. 326–328.

[78] Zamin Iqbal et al. “De novo assembly and genotyping of variants using colored de Bruijn graphs”. In: *Nature genetics* 44.2 (2012), p. 226.

[79] Mark Jobling, Matthew Hurles, and Chris Tyler-Smith. *Human evolutionary genetics: origins, peoples & disease*. Garland Science, 2013.

[80] Peter A Jones and Stephen B Baylin. “The epigenomics of cancer”. In: *Cell* 128.4 (2007), pp. 683–692.

[81] Jocelyn Kaiser and Jennifer Couzin-Frankel. “Biden seeks clear course for his cancer moonshot”. In: *Science* 351.6271 (2016), pp. 325–326.

[82] Juha Kärkkäinen and Peter Sanders. “Simple linear work suffix array construction”. In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2003, pp. 943–955.

[83] Bartha Maria Knoppers and Ruth Chadwick. “Human genetic research: emerging trends in ethics”. In: *Nature Reviews Genetics* 6.1 (2005), pp. 75–79.

[84] Eric S Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (2001), pp. 860–921.

[85] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359.

[86] Ben Langmead et al. “Searching for SNPs with cloud computing”. In: *Genome biology* 10.11 (2009), R134.

[87] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome biology* 10.3 (2009), R25.

[88] Martin Lauss et al. “Monitoring of technical variation in quantitative high-throughput datasets”. In: *Cancer informatics* 12 (2013), CIN–S12862.

[89] Michael S Lawrence et al. “Discovery and saturation analysis of cancer genes across 21 tumour types”. In: *Nature* 505.7484 (2014), pp. 495–501.

[90] Ryan M Layer et al. “LUMPY: a probabilistic framework for structural variant discovery”. In: *Genome biology* 15.6 (2014), R84.

[91] Hane Lee et al. “Clinical exome sequencing for genetic identification of rare Mendelian disorders”. In: *Jama* 312.18 (2014), pp. 1880–1887.

[92] Monkol Lek et al. “Analysis of protein-coding genetic variation in 60,706 humans”. In: *Nature* 536.7616 (2016), pp. 285–291.

[93] Heng Li. “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data”. In: *Bioinformatics* 27.21 (2011), pp. 2987–2993.

[94] Heng Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: *arXiv preprint arXiv:1303.3997* (2013).

[95] Heng Li. “Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly”. In: *Bioinformatics* 28.14 (2012), pp. 1838–1844.

[96] Heng Li. “Mathematical Notes on SAMtools Algorithms”. In: (2010).

[97] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 1 (2018), p. 7.

[98] Heng Li. “Tabix: fast retrieval of sequence features from generic TAB-delimited files”. In: *Bioinformatics* 27.5 (2011), pp. 718–719.

[99] Heng Li. “Towards better understanding of artifacts in variant calling from high-coverage samples”. In: *Bioinformatics* (2014), btu356.

[100] Heng Li and Richard Durbin. “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 26.5 (2010), pp. 589–595.

[101] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[102] Heng Li and Nils Homer. “A survey of sequence alignment algorithms for next-generation sequencing”. In: *Briefings in bioinformatics* 11.5 (2010), pp. 473–483.

[103] Heng Li, Jue Ruan, and Richard Durbin. “Mapping short DNA sequencing reads and calling variants using mapping quality scores”. In: *Genome research* (2008), gr–078212.

[104] Heng Li et al. “A synthetic-diploid benchmark for accurate variant-calling evaluation”. In: *Nature methods* 15.8 (2018), p. 595.

[105] Heng Li et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[106] Ruiqiang Li et al. “SOAP: short oligonucleotide alignment program”. In: *Bioinformatics* 24.5 (2008), pp. 713–714.

[107] Hengyun Lu, Francesca Giordano, and Zemin Ning. “Oxford Nanopore MinION sequencing and genome assembly”. In: *Genomics, proteomics & bioinformatics* 14.5 (2016), pp. 265–279.

[108] Bertram Ludäscher et al. “Scientific workflow management and the Kepler system”. In: *Concurrency and Computation: Practice and Experience* 18.10 (2006), pp. 1039–1065.

[109] Ramon Luengo-Fernandez et al. “Economic burden of cancer across the European Union: a population-based cost analysis”. In: *The lancet oncology* 14.12 (2013), pp. 1165–1174.

[110] David Malkin et al. “Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms”. In: *Science* (1990), pp. 1233–1238.

[111] Udi Manber and Gene Myers. “Suffix arrays: a new method for on-line string searches”. In: *siam Journal on Computing* 22.5 (1993), pp. 935–948.

[112] Teri A Manolio. “Genomewide association studies and assessment of the risk of disease”. In: *New England Journal of Medicine* 363.2 (2010), pp. 166–176.

[113] Ming Mao and Marty Humphrey. “Auto-scaling to minimize cost and meet application deadlines in cloud workflows”. In: *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*. IEEE. 2011, pp. 1–12.

[114] Elaine R Mardis. “Next-generation DNA sequencing methods”. In: *Annu. Rev. Genomics Hum. Genet.* 9 (2008), pp. 387–402.

[115] Ronald Margolis et al. “The National Institutes of Health’s Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data”. In: *Journal of the American Medical Informatics Association* 21.6 (2014), pp. 957–958.

[116] Gabor T Marth et al. “A general approach to single-nucleotide polymorphism discovery”. In: *Nature genetics* 23.4 (1999), p. 452.

[117] Vivien Marx. “Biology: The big challenges of big data”. In: *Nature* 498.7453 (2013), pp. 255–260.

[118] Vivien Marx. “The DNA of a nation”. In: *Nature* 524.7566 (2015), pp. 503–505.

[119] Aaron McKenna et al. “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. In: *Genome research* 20.9 (2010), pp. 1297–1303.

[120] William McLaren et al. “The ensembl variant effect predictor”. In: *Genome biology* 17.1 (2016), p. 122.

[121] Peter Mell and Tim Grance. “The NIST definition of cloud computing”. In: (2011).

[122] Dirk Merkel. “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux Journal* 2014.239 (2014), p. 2.

[123] C Mohinudeen et al. “An Overview of Next-Generation Sequencing (NGS) Technologies to Study the Molecular Diversity of Genome”. In: *Microbial Applications Vol. 1*. Springer, 2017, pp. 295–317.

[124] Fruzsina Molnár-Gábor et al. “Computing patient data in the cloud: practical and legal considerations for genetics and genomics research in Europe and internationally”. In: *Genome Medicine* 9.1 (2017), p. 58.

[125] Paul Muir et al. “The real cost of sequencing: scaling computation to keep pace with data generation”. In: *Genome biology* 17.1 (2016), p. 53.

[126] Shanmugavelayutham Muthukrishnan et al. “Data streams: Algorithms and applications”. In: *Foundations and Trends® in Theoretical Computer Science* 1.2 (2005), pp. 117–236.

[127] Michael W Nachman. “Single nucleotide polymorphisms and recombination rate in humans”. In: *TRENDS in Genetics* 17.9 (2001), pp. 481–485.

[128] Rasmus Nielsen et al. “Genotype and SNP calling from next-generation sequencing data”. In: *Nature Reviews Genetics* 12.6 (2011), pp. 443–451.

[129] Bill Nitzberg and Virginia Lo. “Distributed shared memory: A survey of issues and algorithms”. In: *Distributed Shared Memory-Concepts and Systems* (1991), pp. 42–50.

[130] Tom Oinn et al. “Taverna: a tool for the composition and enactment of bioinformatics workflows”. In: *Bioinformatics* 20.17 (2004), pp. 3045–3054.

[131] David Oppenheimer and David A Patterson. “Architecture and dependability of large-scale internet services”. In: *IEEE Internet Computing* 6.5 (2002), pp. 41–49.

[132] Stavros Papadopoulos et al. “The TileDB array data storage manager”. In: *Proceedings of the VLDB Endowment* 10.4 (2016), pp. 349–360.

[133] Mike P Papazoglou. “Service-oriented computing: Concepts, characteristics and directions”. In: *Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*. IEEE. 2003, pp. 3–12.

[134] Ravi K Patel and Mukesh Jain. “NGS QC Toolkit: a toolkit for quality control of next generation sequencing data”. In: *PloS one* 7.2 (2012), e30619.

[135] Jaume Pellicer, Michael F Fay, and Ilia J Leitch. “The largest eukaryotic genome of them all?” In: *Botanical Journal of the Linnean Society* 164.1 (2010), pp. 10–15.

[136] James L Peterson. “Petri net theory and the modeling of systems”. In: (1981).

[137] *Picard toolkit*. `http://broadinstitute.github.io/picard/`. 2018.

[138] Erin D Pleasance et al. “A comprehensive catalogue of somatic mutations from a human cancer genome”. In: *Nature* 463.7278 (2010), p. 191.

[139] Tobias Rausch et al. “DELLY: structural variant discovery by integrated paired-end and split-read analysis”. In: *Bioinformatics* 28.18 (2012), pp. i333–i339.

[140] Knut Reinert et al. “Alignment of next-generation sequencing reads”. In: *Annual review of genomics and human genetics* 16 (2015), pp. 133–151.

[141] Anthony Rhoads and Kin Fai Au. “PacBio sequencing and its applications”. In: *Genomics, proteomics & bioinformatics* 13.5 (2015), pp. 278–289.

[142] Andy Rimmer et al. “Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications”. In: *Nature genetics* 46.8 (2014), pp. 912–918.

[143] Michael Roberts et al. “Reducing storage requirements for biological sequence comparison”. In: *Bioinformatics* 20.18 (2004), pp. 3363–3369.

[144] Nicola D Roberts et al. “A comparative analysis of algorithms for somatic SNV detection in cancer”. In: *Bioinformatics* (2013), btt375.

[145] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. “The advantages of SMRT sequencing”. In: *Genome biology* 14.6 (2013), p. 405.

[146] Dan Robinson et al. “Integrative clinical genomics of advanced prostate cancer”. In: *Cell* 161.5 (2015), pp. 1215–1228.

[147] Richard M Russell. “The CRAY-1 computer system”. In: *Communications of the ACM* 21.1 (1978), pp. 63–72.

[148] Fred Sanger and Alan R Coulson. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. In: *Journal of molecular biology* 94.3 (1975), 441IN19447–446IN20448.

[149] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[150] Wataru Satake et al. “Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson’s disease”. In: *Nature genetics* 41.12 (2009), p. 1303.

[151] Stephan C Schuster. “Next-generation sequencing transforms today’s biology”. In: *Nature* 200.8 (2007), pp. 16–18.

[152] Omar Sefraoui, Mohammed Aissaoui, and Mohsine Eleuldj. “OpenStack: toward an open-source solution for cloud computing”. In: *International Journal of Computer Applications* 55.3 (2012).

[153] Dennis J Selkoe. “Amyloid β-protein and the genetics of Alzheimer’s disease”. In: *Journal of Biological Chemistry* 271.31 (1996), pp. 18295–18298.

[154] Robert Shapiro. “A technical comparison of XPDL, BPML and BPEL4WS”. In: *Cape Visions* (2002).

[155] Mary Shaw and David Garlan. *Software architecture*. Vol. 101. Prentice Hall Englewood Cliffs, 1996.

[156] Hui Shen et al. “Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians”. In: *PLoS One* 8.4 (2013), e59494.

[157] Nisheeth Shrivastava et al. “Medians and beyond: new aggregation techniques for sensor networks”. In: *Proceedings of the 2nd international conference on Embedded networked sensor systems*. ACM. 2004, pp. 239–249.

[158] Jared T Simpson and Richard Durbin. “Efficient construction of an assembly string graph using the FM-index”. In: *Bioinformatics* 26.12 (2010), pp. i367–i373.

[159] Jared T Simpson and Richard Durbin. “Efficient de novo assembly of large genomes using compressed data structures”. In: *Genome research* 22.3 (2012), pp. 549–556.

[160] Temple F Smith and Michael S Waterman. “Comparison of biosequences”. In: *Advances in applied mathematics* 2.4 (1981), pp. 482–489.

[161] Lincoln D Stein et al. “Data analysis: create a cloud commons”. In: *Nature* 523 (2015), pp. 149–151.

[162] Zachary D Stephens et al. “Big data: astronomical or genomical”. In: *PLoS biology* 13.7 (2015), e1002195.

[163] Zachary D Stephens et al. “Simulating next-generation sequencing datasets from empirical mutation and sequencing models”. In: *PloS one* 11.11 (2016), e0167047.

[164] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. “The cancer genome”. In: *Nature* 458.7239 (2009), pp. 719–724.

[165] Peter H Sudmant et al. “An integrated map of structural variation in 2,504 human genomes”. In: *Nature* 526.7571 (2015), pp. 75–81.

[166] Frederick Winslow Taylor. *Scientific management*. Routledge, 2004.

[167] *The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI)*. URL: `https://www.genome.gov/sequencingcosts/` (visited on 11/04/2016).

[168] Lindsey A Torre et al. “Global cancer statistics, 2012”. In: *CA: a cancer journal for clinicians* 65.2 (2015), pp. 87–108.

[169] Esko Ukkonen. “On-line construction of suffix trees”. In: *Algorithmica* 14.3 (1995), pp. 249–260.

[170] Geraldine A Van der Auwera et al. “From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline”. In: *Current protocols in bioinformatics* (2013), pp. 11–10.

[171] Luis M Vaquero, Luis Rodero-Merino, and Rajkumar Buyya. “Dynamically scaling applications in the cloud”. In: *ACM SIGCOMM Computer Communication Review* 41.1 (2011), pp. 45–52.

[172] J Craig Venter et al. “The sequence of the human genome”. In: *science* 291.5507 (2001), pp. 1304–1351.

[173] J Craig Venter et al. “Shotgun sequencing of the human genome”. In: *Science* 280.5369 (1998), pp. 1540–1542.

[174] Kathleen A Vermeersch and Mark P Styczynski. “Applications of metabolomics in cancer research”. In: *Journal of carcinogenesis* 12 (2013).

[175] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. "Next-generation sequencing: from basic research to diagnostics". In: *Clinical chemistry* 55.4 (2009), pp. 641–658.

[176] Jeremiah A Wala et al. "SvABA: genome-wide detection of structural variants and indels by local assembly". In: *Genome research* (2018).

[177] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[178] James D Watson, Francis HC Crick, et al. "Molecular structure of nucleic acids". In: *Nature* 171.4356 (1953), pp. 737–738.

[179] John N Weinstein et al. "The cancer genome atlas pan-cancer analysis project". In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[180] Jeffrey N Weitzel et al. "Genetics, genomics, and cancer risk assessment: state of the art and future directions in the era of personalized medicine". In: *CA: a cancer journal for clinicians* 61.5 (2011), pp. 327–359.

[181] BP Welford. "Note on a method for calculating corrected sums of squares and products". In: *Technometrics* 4.3 (1962), pp. 419–420.

[182] Danielle Welter et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic acids research* 42.D1 (2013), pp. D1001–D1006.

[183] Justin P Whalley et al. "Framework For Quality Assessment Of Whole Genome, Cancer Sequences". In: *bioRxiv* (2017), p. 140921.

[184] Philipp Wieder et al. *Service level agreements for cloud computing*. Springer Science & Business Media, 2011.

[185] K Robin Yabroff et al. "Economic burden of cancer in the United States: estimates, projections, and future research". In: *Cancer Epidemiology Biomarkers & Prevention* 20.10 (2011), pp. 2006–2014.

[186] Yaping Yang et al. "Clinical whole-exome sequencing for the diagnosis of mendelian disorders". In: *New England Journal of Medicine* 369.16 (2013), pp. 1502–1511.

[187] Kai Ye et al. "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads". In: *Bioinformatics* 25.21 (2009), pp. 2865–2871.

[188] Christina K Yung et al. "ICGC in the cloud". In: *Cancer Research* 76.14 Supplement (2016), pp. 3605–3605.

[189] Shelemyahu Zacks. *The theory of statistical inference*. Vol. 34. Wiley New York, 1971.

[190] Matei Zaharia et al. "Faster and more accurate sequence alignment with SNAP". In: *arXiv preprint arXiv:1111.5572* (2011).

[191] Daniel R Zerbino and Ewan Birney. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". In: *Genome research* 18.5 (2008), pp. 821–829.

[192] Min Zhao et al. “Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives”. In: *BMC bioinformatics* 14.11 (2013), S1.

[193] Qian Zhou et al. “QC-Chain: fast and holistic quality control method for next-generation sequencing data”. In: *PloS one* 8.4 (2013), e60234.

[194] Songnian Zhou. “Lsf: Load sharing in large heterogeneous distributed systems”. In: *I Workshop on Cluster Computing*. Vol. 136. 1992.