Rheos

A Service-Oriented Architecture for comprehensive stream-based genomic variant calling

Sergei Yakneen EMBL

Meyerhofstrasse 1,

69117 Heidelberg, Germany 19/07/2016

# Contents

1. [Introduction](#_bookmark0) 4
   1. [Context and Motivation](#_bookmark1) 4
   2. [Challenges and Problem Statement](#_bookmark2) 5
   3. [Proposed Solution](#_bookmark3) 7
   4. [Thesis Outline](#_bookmark4) 11
2. [Background and Related Work](#_bookmark5) 12
   1. [Genomics](#_bookmark6) 12
      1. [History of Genomics](#_bookmark7) 12
      2. [Next Generation Sequencing](#_bookmark8) 14
      3. [Genomics Studies](#_bookmark10) 15
      4. [Cancer Genomics](#_bookmark11) 15
      5. [Clinical Genomics](#_bookmark13) 17
   2. [Computational Pipelines for Next Generation Sequencing](#_bookmark14) 17
      1. [Base-calling](#_bookmark15) 17
      2. [Alignment](#_bookmark16) 17
      3. [Raw Data QA](#_bookmark17) 17
      4. [Germline SNP Calling](#_bookmark18) 17
      5. [Germline Indel Calling](#_bookmark19) 17
      6. [Germline Structural Variant Calling](#_bookmark20) 17
      7. [Variant Filtering](#_bookmark21) 17
      8. [Somatic SNP Calling](#_bookmark22) 17
      9. [Somatic Indel Calling](#_bookmark23) 17
      10. [Somatic Structural Variant Calling](#_bookmark24) 17
      11. [Germline Variant Annotation](#_bookmark25) 17
      12. [Somatic Variant Annotation](#_bookmark26) 17
      13. [de-novo Assembly](#_bookmark27) 17
   3. [Service Oriented Architectures](#_bookmark28) 17
   4. [Stream-based Systems](#_bookmark29) 17
3. [The Rheos Framework](#_bookmark30) 18
4. [Implementation](#_bookmark31) 19
5. [Experimental Validation](#_bookmark32) 20
6. [Discussion and Conclusion](#_bookmark33) 21

Sergei Yakneen 3

# Chapter 1 Introduction

## Context and Motivation

In the 17 years since the publication of the first draft human genome[[37](#_bookmark70)] 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 exper- imental 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[[68](#_bookmark101)], and the development of the Internet and cloud com- puting[[9](#_bookmark42)]. 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[[31](#_bookmark64)], while the Internet and cloud computing are democratising access to large-scale computational resources such that computa- tion 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[[35](#_bookmark68)], to identify causes and mechanisms of action of rare genetic diseases that affect a very small proportion of the population[[7](#_bookmark40)], and to elucidate and potentially target the genetic component of more common diseases such as cancer[[80](#_bookmark113)], heart disease[[8](#_bookmark41)], or dementia[[69](#_bookmark102)] 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, DNA sequencing as a whole, remains a largely unsolved problem. The pro- tein 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 co- horts at the scale of tens of thousands of individuals[[41](#_bookmark74)]. However, the other 99% of the human genome, it’s non-coding regions, contain crucial information such as gene regulatory elements[[13](#_bookmark46)] that are essential to our full understanding of the mecha- nisms 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 [[70](#_bookmark103)]. WGS studies at 100,000 participants scale that are planned for the coming years[[20](#_bookmark53)] 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[[40](#_bookmark73)] and certain types of can- cers[[65](#_bookmark98)] where a small set of genomic loci is interrogated via a gene panel[[4](#_bookmark37)] 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 poten- tial 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[[62](#_bookmark95)], but are generally not amenable to interrogation via gene panels.

The magnitude of the opportunity for improvement in the space of DNA sequenc- ing 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.

## Challenges and Problem Statement

Let’s examine the key challenges that need to be addressed in order to enable effi- cient 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[[71](#_bookmark104)].

**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[[54](#_bookmark87)].

**Data Formats** - The data formats used for storing genomic data are primarily large size character and binary files (FASTA, SAM, BAM, VCF)[[45](#_bookmark78), [15](#_bookmark48)] 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[[53](#_bookmark86)].

**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[[17](#_bookmark50)]. Each stage typ- ically 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 re- quiring 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 associ- ation with phenotypic variables of interest[[81](#_bookmark114)]. 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[[34](#_bookmark67)], 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[[78](#_bookmark111)]. Both parties wish to do so at minimal possible total cost – *C* = *cg* + *cs* + *ca* + *cr*, 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 computa- tional algorithms[[43](#_bookmark76), [86](#_bookmark119), [3](#_bookmark36), [51](#_bookmark84)] 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 con- tradictory 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 re- search 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.

## 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 tech- nology used, the underlying chemistry, and the cost of the reagents[[52](#_bookmark85)]. Improving these characteristics falls outside the scope of our discussion, and we assume the cost of the data generation component *cg* 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 evalu- ated 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[[44](#_bookmark77)]), by increasing the level of parallelisation within the computational pipeline i.e. paral- lelising steps that are currently sequential[[38](#_bookmark71)], or by utilising additional computa- tional resources, thereby increasing costs. The various components of cost, in turn, can be optimized by improved data storage and retrieval structures[[58](#_bookmark91)] (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[[50](#_bookmark83), [75](#_bookmark108)] 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 com- putational framework, called Rheos, that is based on the concepts of data streaming, cloud computing, and service orientation to provide a comprehensive toolset for ge- nomic 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 chal- lenges described in Section **??**. These are:

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

Service orientation[[21](#_bookmark54)] allows us to decompose the overall problem of comprehen- sively 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 inter- face 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 characteris- tics that makes explicit commitments about the service’s reliability, speed, etc[[59](#_bookmark92)]. 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 con- straints 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 (*cs* and *cr*). 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[[14](#_bookmark47)]. 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[[28](#_bookmark61)]. 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[[45](#_bookmark78), [26](#_bookmark59)], and VCF[[15](#_bookmark48)], 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[[19](#_bookmark52)].

Rheos adopts a data and event stream approach to accomplish scalable fine- grained communication between services[[55](#_bookmark88)]. This approach allows each service to listen to and produce data at the level of granularity that it needs to make deci- sions, 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[[32](#_bookmark65)], unlike file-based systems that need to wait for the entire sample to transmit before beginning. This approach can potentially enable real-time anal- ysis given sufficient allocation of computational resources[[1](#_bookmark34)]. 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[[22](#_bookmark55)]. 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[[27](#_bookmark60)]. Because genomic algorithms frequently make use of various summary statistics accumulated over the data-set[[60](#_bookmark93), [42](#_bookmark75)], 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[[16](#_bookmark49), [5](#_bookmark38)] 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[[45](#_bookmark78), [29](#_bookmark62), [11](#_bookmark44), [63](#_bookmark96)]. The algorithms then proceed by traversing the genome in coordinate-wise fashion from the beginning of chromosome 1 to the end of chromosome Y interrogat- ing each locus in turn by examining the set of reads that overlap that locus (a read pileup)[[45](#_bookmark78)]. 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[[82](#_bookmark115)], have been aligned[[44](#_bookmark77)], have been investigated for PCR duplicates[[75](#_bookmark108)], and have been sorted[[75](#_bookmark108)]. 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 align- ment 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[[84](#_bookmark117), [6](#_bookmark39)].

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[[45](#_bookmark78), [29](#_bookmark62)]. 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 im- mediately 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[[2](#_bookmark35)]. 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 applica- tions whereby sequencing is abandoned early when issues such as sample-swap[[23](#_bookmark56)], or contamination[[10](#_bookmark43)] are detected. In addition, when sufficient accuracy is reached based on observed data at a particular locus, the framework may choose to stop look- ing 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 en- able the large scale genomics studies of the future[[20](#_bookmark53), [36](#_bookmark69), [49](#_bookmark82)] in both research and clinical contexts by providing a toolset that allows for interpretation and compre- hensive 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 opti- mization 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 un- derlying 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.

## 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](#_bookmark5) 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 un- derlying statistical models and operational characteristics. The chapter concludes with a look at workflow frameworks that tie individual algorithms together into computational pipelines. Chapter [3](#_bookmark30) 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 investi- gate the theoretical properties of the algorithms that underlie Rheos-based genomic analysis. Chapter [4](#_bookmark31) describes the actual implementation of the Rheos framework’s components and investigates their operational characteristics. Chapter [5](#_bookmark32) is dedi- cated to the experimental evaluation of Rheos in comparison to other extant frame- works and algorithms using real genomic data. We conclude this work in Chapter [6](#_bookmark33) with a discussion of the results and an examination of the future direction of Rheos development.

# Chapter 2

**Background and Related Work**

## 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[**originofspecies**] and Gregor Mendel[[30](#_bookmark63)]. While genetics largely focuses on the study of single (or rel- atively small numbers of) genes - the *genotype*, and how genetic variation and mu-

tation 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.

### History of Genomics

Each living cell is a bio-chemical machine that carries out a number of complex be- haviours 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 sub- divided 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. Un- derstanding 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 *se- quencing*. We are able to sequence DNA thanks to the pioneering work of researchers Rosalind Franklin[[25](#_bookmark58)], James Watson, and Francis Crick[[79](#_bookmark112)] who first elucidated the physical structure of DNA, then followed by the work of Fred Sanger[[67](#_bookmark100), [66](#_bookmark99)] who devised the first effective DNA sequencing method. The sequencing method allows

12

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 prob- lem.

While the entire length of the DNA of an organism ranges from several hun- dred 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[[61](#_bookmark94)] 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[[67](#_bookmark100)]. Reconstituting the original complete DNA sequence from partial overlaps between reads is thus a costly, time consuming, and computation- ally intensive problem known as *de-novo assembly*[[85](#_bookmark118)]. 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*[[43](#_bookmark76)] 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)[[56](#_bookmark89)] researchers are able to significantly reduce the amount of infor- mation 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 under- goes 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[[47](#_bookmark80)].

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[[37](#_bookmark70)] - an international effort to sequence and *de-novo* assemble the first complete human genome consisting of chromosomes 1-22, X, and Y and totalling approximately 3,000,000,000 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 ver- sion of the human genome was simultaneously published by a commercial company led by JC Venter[[77](#_bookmark110)], using a new sequencing method called shotgun sequencing[[76](#_bookmark109)],

a method that formed the basis for a new revolution in sequencing technology, now termed Next Generation Sequencing[[68](#_bookmark101)].

### Next Generation Sequencing

The Next Generation Sequencing methodology[[48](#_bookmark81)] 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[[44](#_bookmark77)]. Because NGS sequencing methods are prone to certain errors and biases[[18](#_bookmark51)], 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*[[57](#_bookmark90)], 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.

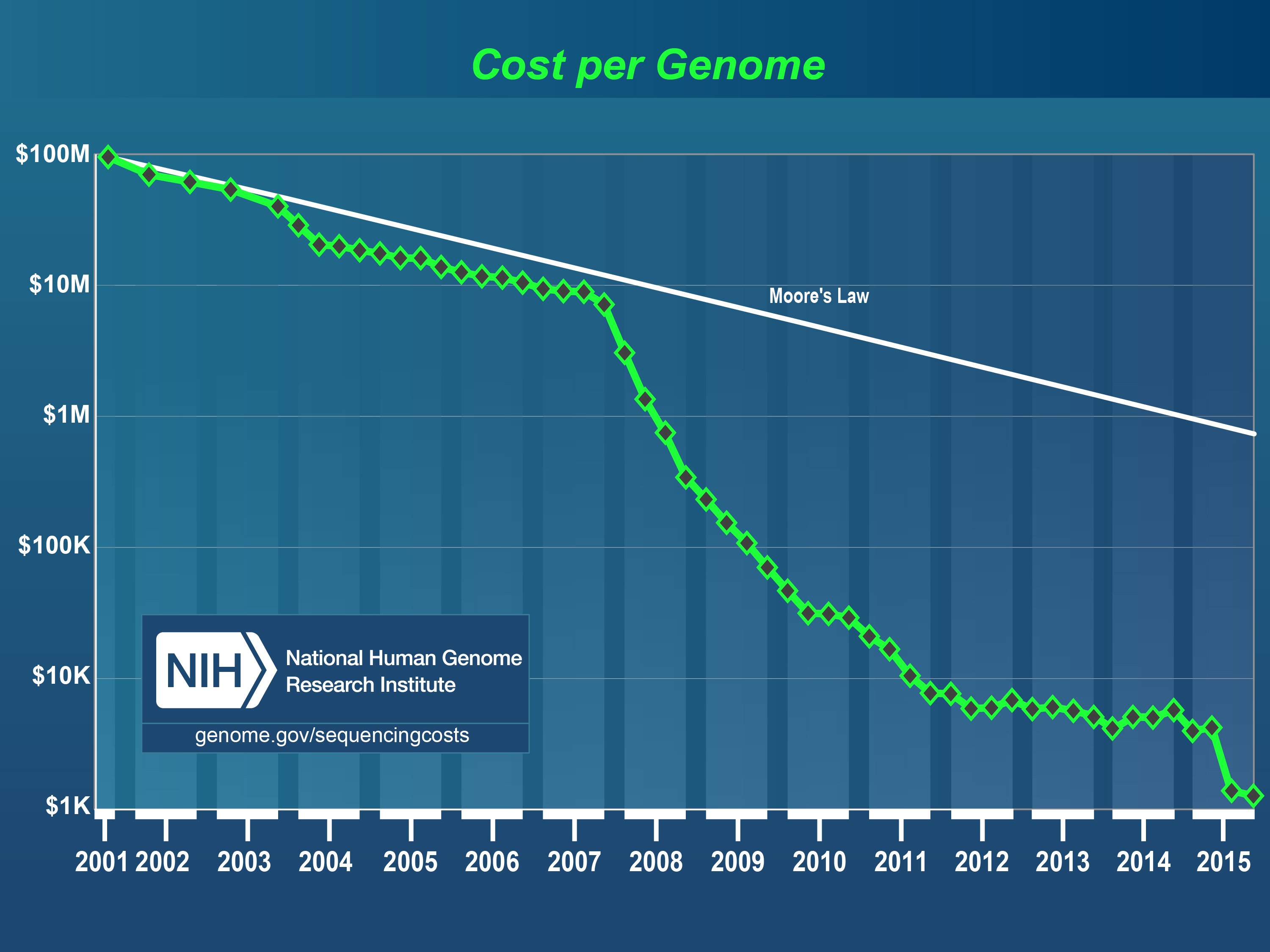


Figure 2.1: Cost of DNA sequencing[[73](#_bookmark106)]

Figure [2.1](#_bookmark9) 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[[12](#_bookmark45)], and the large scale sequencing of the Icelandic population[[31](#_bookmark64)].

### Genomics Studies

### Cancer Genomics

Cancer is a genetic disease that has an extremely high burden on the human pop- ulation. In 2012, the global incidence of new cases worldwide has been estimated as 14.1 million, and deaths at 8.2 million[[74](#_bookmark107)]. The economic cost of cancer to the European Union has been estimated at 126 billion euro in 2009[[46](#_bookmark79)], and in the US

$124.5 billion USD in 2010[[83](#_bookmark116)]. Because of the genetic nature of the disease studying genomes of cancer patients helps uncover the mechanisms behind the development and evolution of cancer[[72](#_bookmark105)].

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 mu- tation rate, increased proliferation, anchorage independent growth, and resisting cell death[[33](#_bookmark66)] . Only certain mutations, however, contribute to the development of cancer, while others are benign. Cancer genomics studies aim to identify and char- acterize those mutations that are cancer drivers and play a role in the formation or progression of tumours[[72](#_bookmark105)].

Studying cancer genomes is more complex and expensive than studying the genomes of healthy individuals because each patient requires that two DNA samples are col- lected - 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 popu- lation[[64](#_bookmark97)].

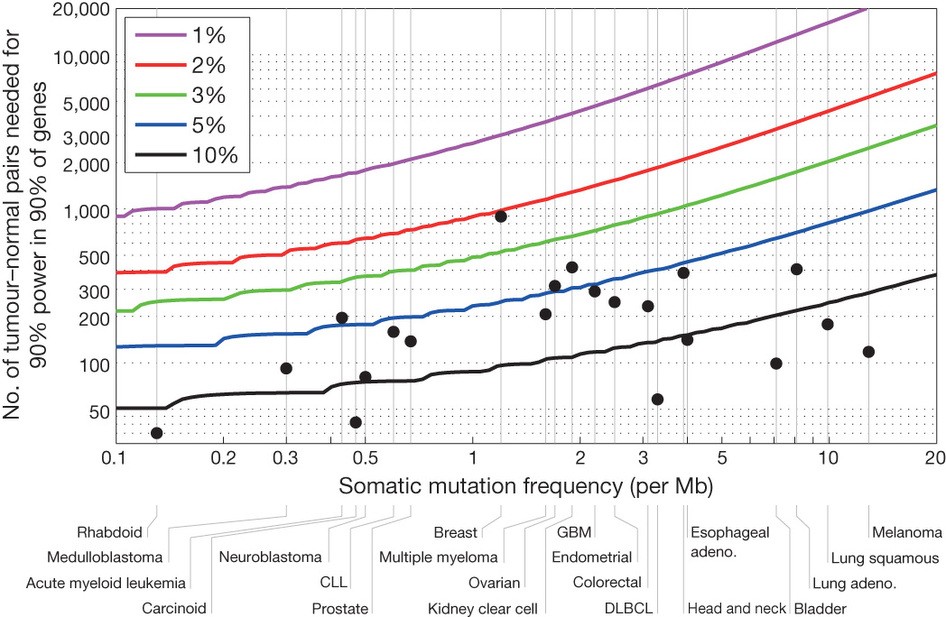


Figure 2.2: Sequencing sample size required by mutation rate[[39](#_bookmark72)].

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)[[24](#_bookmark57)], each mutation has a low chance of being present in any given tumour. Figure [2.2](#_bookmark12) 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.

### Clinical Genomics

## Computational Pipelines for Next Generation Sequencing

### Base-calling

### Alignment

### Raw Data QA

### Germline SNP Calling

### Germline Indel Calling

### Germline Structural Variant Calling

### Variant Filtering

### Somatic SNP Calling

### Somatic Indel Calling

### Somatic Structural Variant Calling

### Germline Variant Annotation

### Somatic Variant Annotation

### de-novo Assembly

## Service Oriented Architectures

## Stream-based Systems

# The Rheos Framework

**Implementation**

**Experimental Validation**

**Discussion and Conclusion**

# Bibliography

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

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

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

[4] 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 re- ceptor monoclonal antibody therapy”. In: *Journal of clinical oncology* 27.12 (2009), pp. 2091–2096.

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

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

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

[8] Benoit G Bruneau. “The developmental genetics of congenital heart disease”.

In: *Nature* 451.7181 (2008), pp. 943–948.

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

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

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

[12] 1000 Genomes Project Consortium et al. “A map of human genome variation from population-scale sequencing”. In: *Nature* 467.7319 (2010), pp. 1061–1073.

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

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

[15] Petr Danecek et al. “The variant call format and VCFtools”. In: *Bioinformatics*

27.15 (2011), pp. 2156–2158.

[16] Mayur Datar et al. “Maintaining stream statistics over sliding windows”. In:

*SIAM journal on computing* 31.6 (2002), pp. 1794–1813.

[17] 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.

[18] 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.

[19] Richard Durbin et al. *Biological sequence analysis: probabilistic models of pro- teins and nucleic acids*. Cambridge university press, 1998.

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

2.

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

[22] Opher Etzion, Peter Niblett, and David C Luckham. *Event processing in ac- tion*. Manning Greenwich, 2011.

[23] 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.

[24] Simon A Forbes et al. “COSMIC: exploring the world’s knowledge of so- matic mutations in human cancer”. In: *Nucleic acids research* 43.D1 (2015), pp. D805–D811.

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

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

[27] Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. “Min- ing data streams: a review”. In: *ACM Sigmod Record* 34.2 (2005), pp. 18–26.

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

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

[30] Mendel Gregor. “Versuche uber Pflanzen-Hybriden. Verhandlungen des natur- forschenden Vereines in Brunn”. In: 4 (1865), pp. 3–47.

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

[32] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and tech- niques*. Elsevier, 2011.

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

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

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

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

[37] Eric S Lander et al. “Initial sequencing and analysis of the human genome”.

In: *Nature* 409.6822 (2001), pp. 860–921.

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

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

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

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

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

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

[44] 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.

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

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

[47] 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.

[48] Elaine R Mardis. “Next-generation DNA sequencing methods”. In: *Annu. Rev.*

*Genomics Hum. Genet.* 9 (2008), pp. 387–402.

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

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

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

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

[53] 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.

[54] 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.

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

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

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

[58] Stavros Papadopoulos et al. “The TileDB array data storage manager”. In:

*Proceedings of the VLDB Endowment* 10.4 (2016), pp. 349–360.

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

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

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

[66] 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.

[67] 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.

[68] Stephan C Schuster. “Next-generation sequencing transforms today’s biology”.

In: *Nature* 200.8 (2007), pp. 16–18.

[69] Dennis J Selkoe. “Amyloid *β*-protein and the genetics of Alzheimer’s disease”.

In: *Journal of Biological Chemistry* 271.31 (1996), pp. 18295–18298.

[70] Lincoln D Stein et al. “Data analysis: create a cloud commons”. In: *Nature*

523 (2015), pp. 149–151.

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

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

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

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

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

[76] J Craig Venter et al. “Shotgun sequencing of the human genome”. In: *Science*

280.5369 (1998), pp. 1540–1542.

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

[78] 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.

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

[80] John N Weinstein et al. “The cancer genome atlas pan-cancer analysis project”.

In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[81] 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.

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

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

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

[85] 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.

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