THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Statistical assessment of somatic mutations and genomic variability using DNA sequence data

Anna Rehammar

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2016

Statistical assessment of somatic mutations and genomic variability using DNA sequence data
Anna Rehammar

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Author e-mail: anna.rehammar@chalmers.se

Typeset with LaTeX.
Printed in Göteborg, Sweden 2016

Statistical assessment of somatic mutations and genomic variability using DNA sequence data

Anna Rehammar

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg

Abstract

The development of new DNA sequencing techniques have made it possible to generate high-resolution genomic data at an unprecedented pace. However, the high dimensionality in combination with the substantial levels of technical errors and biological variability make the analysis challenging. Tailored statistical methods need therefore to be developed and applied in order to facilitate correct biological interpretation. The first two papers in this thesis are focused on finding tumor-specific (somatic) mutations in cancer, while in the third paper a new method to assess genomic variability in microbial communities is developed.

In *paper I*, the aim was to characterize somatic mutations in pheochromocytoma/paraganglioma, and to identify mutations that contribute to malignancy. Statistical analysis of exome sequencing data from nine replicated paired normal–tumor samples revealed 225 unique somatic mutations. A significantly higher rate of mutations was found in malignant compared to benign tumors. In addition, three genes with recurrent somatic mutations, exclusively located in malignant tumors, were identified.

In *paper II*, exome sequencing data was used to detect somatic mutations in 17 patients with acute myeloid leukemia. The identified mutations were evaluated as markers in a more sensitive analysis of remaining cancer cell levels after treatment. All but one of the studied patients were found to have potential markers in their somatic mutation profiles.

In *paper III*, a hierarchical Bayesian model for detecting genetic differences on nucleotide level between groups of microbial communities is proposed. The model is based on a Dirichlet-multinomial distribution and takes both within- and between-sample variability into account. The evaluation of the performance show that the model has a high sensitivity and maintains a low false positive rate even when the between-sample variability is high.

The thesis demonstrates the importance of dedicated statistical analysis and understanding of the error structure in DNA sequence data, in order to assure accurate identification of mutations and differences in genomic variability.

Keywords: DNA sequence data, exome sequencing, calling of somatic mutations, metagenomics, hierarchical Bayesian model, genomic variability

Acknowledgements

My first and foremost thanks goes to my supervisor Erik Kristiansson. Without your continuous support, trust and always so positive encouragement this thesis would not have happened. I am impressed by your knowledge, your constructive view, and your ability to always see possibilities and to prioritize in a good way. I would also like to thank my co-supervisor Frida Abel for sharing your biological knowledge and giving me the trust to be an important part in our project. Thank you Olle Nerman for taking your time to discuss and explain, for me, involved statistics and for welcoming me back as a PhD student after being away on other adventures.

I am grateful to all my co-authors for nice collaborations. Thanks to Erik Malmberg, Linda Fogelstrand, Tore Samuelsson, Andreas Muth, Bo Wängberg, Ola Nilsson and Yvonne Arvidsson for interesting and informative discussions and for giving me the opportunity to dig into your data. I also want to thank Marcela Dávila López for discussing bioinformatics with a smile. A special thanks to my PhD student partner from day one, Annica Wilzén, not only for all we have learnt together, but also for being a beloved soulmate. I cross my fingers and hope that we will work together again.

Many thanks to my colleagues at the mathematics department, and especially to the ones I get to visit castles with. To Viktor for being the best room mate, Anders for lunch walks and for sharing both happy and though moments in life, and to all the rest of the group for creating such a supporting working climate – Anna, Fanny, Fredrik, Johannes, Maria, Mariana, Marina and Tobias.

I am lucky to have my friends and relatives, thanks for being there when life takes unexpected turns. Thanks also to my parents for caring and for being such great grandparents so I can work some more without bad conscience.

Slutligen, tack till min älskade familj – min livskamrat Robert, våra fantastiska barn Alvin och Vilhelm, vår för alltid saknade Tage och vårt efterlängtade barn i magen (stanna gärna kvar där till efter seminariet!). Jag är så glad för att vi är ett team i glädje och sorg, för att ni alltid stöttar mig och sätter saker i perspektiv.

Till Vilhelm, som tio minuter efter att jag förklarat att jag behöver tid själv för att skriva om det jag jobbat med de senaste åren sticker in huvudet och frågar: Är du färdig nu mamma, så vi kan leka? Nu är jag färdig. För nu, och för en tid framöver. Till er andra; jag är snart tillbaka och även det ser jag fram emot!

Anna Rehammar

List of Papers

The licentiate thesis includes the following papers.

I. Wilzén, A.*, **Rehammar, A.***, Muth, A., Nilsson, O., Tesan Tomic, T., Wängberg, B. , Kristiansson, E., Abel, F. (2015). Malignant pheochromocytomas/paragangliomas harbor mutations in transport and cell adhesion genes. *Accepted for publication in International Journal of Cancer.*

II. Malmberg, E., Ståhlman, S., **Rehammar, A.**, Samuelsson, T., Alm, S.J., Kristiansson, E., Abrahamsson, J., Garelius, H., Palmqvist, L., Fogelstrand, L. (2015). Patient-tailored analysis of minimal residual disease in acute myeloid leukemia using next generation sequencing. *Submitted.*

III. **Rehammar, A.**, Nerman, O., Kristiansson, E. (2016). A hierarchical Bayesian model for assessing differential nucleotide composition between metagenomes. *Manuscript.*

* Authors contributed equally

Author contributions

I. Performed the bioinformatics work and the statistical analyses. Participated in study design, interpretation of the results and in drafting and editing the manuscript.

II. Performed the bioinformatics work and the statistical analyses for identification of somatic mutations and MRD candidates. Participated in editing the manuscript.

III. Designed the study, developed the model, implemented the model and the simulations, performed the analyses, drafted and edited the manuscript.

Contents

Chapter 1

Introduction

In all cells of all organisms the genetic material, the genome, contain information on how the cells should develop and function. Differences in the genome determine, together with the encountered environment, our traits, development and responses. Evolution of new functions and organisms is possible due to changes in the genome. The genetic material consists of DNA molecules, which are build up by long chains of four different building blocks. These are collectively called nucleotides and are denoted A, C, G and T. To be able to analyse how the information encoded in the DNA molecules govern biological processes, the information need to be read. That is, the type of nucleotide present at each position in the genome need to be mapped out. The term "sequencing" refer to the procedure of elucidating the order of the nucleotides in DNA molecules.

Until recently, sequencing was a time consuming and costly project. For example, when the first human genome was sequenced, it was a large collaborative project that had taken over 10 years to complete (Lander et al., 2001). Therefore, an early strategy for investigating the association of a property to variations in the human genome was to only read a very limited set of positions instead of the whole sequence. However, new innovative techniques for DNA sequencing, commonly referred to as next generation sequencing (NGS), have dramatically lowered the cost and effort for sequencing and revolutionized the ability to analyse genomes. Therefore, it is now possible to compare the information in the whole genome, or part of it, from many samples. The connection between genetic alterations and different properties, such as for example a disease, can thereby be investigated down to unprecedented resolution.

An impressive example of what now is possible is the whole-genome sequencing of 2,636 Icelanders reported by Gudbjartsson et al. (2015). The data is paired with other unique resources for the Icelandic population, such as a genealogy for the nation documented several hundred years back, access to na-

tionwide healthcare information and additional sequence data for over 100,000 Icelanders previously analysed at lower resolution. The article describes the landscape of genetic variants in the human genome in relation to, for example, functional annotation and gene position. In addition, three examples of connections between genetic variants and diseases found using the data are given, and additional such findings are reported in subsequent articles (Swaminathan et al., 2015; Oddsson et al., 2015).

The alterations in the nucleotide sequences giving rise to the genetic variability discussed above are commonly called mutations. The different genetic variants that mutations create are called alleles. Mutations can occur in several different ways. An exchange (also called a substitution) of one nucleotide for another, will here be denoted a single nucleotide variant (SNV). One or a few nucleotides can also be inserted or deleted from the DNA chain, such mutations are called insertions and deletions, respectively, or with a common name indels. These small-scale mutations are in focus in the work described in this thesis. Examples of such mutations are SNVs and indels in the *BRCA1* gene, changing the properties of the encoded protein and giving rise to increased risk for breast cancer (King et al., 2003). However, there are also mutations on a larger scale, with amplification or loss of larger parts up to whole chromosomes (a whole DNA molecule) or structural rearrangements within or between chromosomes. An example is the gain of an extra copy of chromosome number 21 or parts of it, giving Downs syndrome to the carrier. Although mutations can have damaging effects, it is important to note that they are a prerequisite for evolution and gain of new beneficial properties. One example is a mutation in the *FUT2* gene that give rise to immunity against winter vomiting disease (Thorven et al., 2005).

A common class of diseases that has genetic origin are different types of cancer, where mutations alter the normal functions of a cell and turns it into a cancer cell (Hanahan and Weinberg, 2011). Cancer is a heterogeneous group of diseases, that can have large differences in their genetic causes. Even within a specific type of cancer, such as breast or lung cancer, there are many combinations of mutations that can give rise to a tumor (Vogelstein et al., 2013). Which mutations a specific tumor harbor influence, for example, the aggressiveness of the disease and the ability for the tumor to metastasize (Armaghany et al., 2012; Brodeur et al., 1984). Furthermore, the response to treatment can be dependent on which mutations that are present in the tumor, and additional mutations can give rise to drug resistance during treatment (Garnett et al., 2012; Zahreddine and Borden, 2013; Nilsson et al., 2009). It is therefore an important field of research to characterize which mutations that cause different types of cancer and how they influence the progression and properties of the disease. This is required both to gain a more thorough understanding of tumor biology and to be able to develop better diagnostics and treatment. However,

the analysis is impeded by the fact that tumor cells have a higher mutation rate than normal cells. Many of those mutations acquired during tumor growth are so called passenger mutations and do not influence the progression of the disease (Martincorena and Campbell, 2015). Also, a mix of inherited mutations that exist in all cells of an individual and acquired mutations are often together causing the tumor development (Knudson, 1971).

Another area where genetic variability is studied is microbiology. Microorganisms are vital parts of ecosystems, but historically it has been hard to analyse the full complexity of microbial communities. The methods have been dependent on the ability to culture the studied organisms in a laboratory. A microbial community can consist of thousands of species and only a limited amount of those have been able to culture. However, with the advent of NGS techniques the field of metagenomics also burst forth. In metagenomics all the genetic material from a sample taken directly from the environment is sequenced, without any prior cultivation. Thereby the genetic variability, and hence the compositions of species and biological functions, and its connection to different conditions and properties can be investigated.

To better understand the processes in microbial communities is of great importance in many different fields, such as agriculture, waste water treatment and medicine. For example, bacteria exist practically everywhere, both in the environment and within humans. Often they contribute with important functions, such as in the digestion process in the gut. However, bacteria can also cause infections and we are dependent on having antibiotics to treat those infections. An emerging problem is bacteria that have become resistant to one or several types of antibiotics. This phenomena may turn infections that are today easily treatable into life-threatening ones. Also in the long run it can hamper our way of practice medical treatment. For example, effective antibiotics are important when doing surgery, to hinder infections in the wounds. Resistance towards antibiotics typically depends on changes in the genetic material of the bacteria. Mutations in protein coding genes is one way of acquiring resistance. For example, only three SNVs in the genome of the bacterium *Escherichia coli* is enough to make it highly resistant to certain types of antibiotics (Bagel et al., 1999). To be able to advance our understanding of the mechanisms behind antibiotic resistance, one important part is thus to examine which mutations that exists in bacteria and that are promoted under selection pressure from antibiotics.

As described above, the new sequencing techniques have made it possible to generate massive amounts of data and opened up a wealth of new opportunities for analysis of genomes. There are, however, also a number of new challenges related to data handling and analysis. After extraction of the genetic material from the studied sample, the DNA molecules are heavily fragmented and many such fragments are then sequenced rapidly and in parallel (Metzker, 2010). The

term "massively parallel sequencing" is hence often used for these techniques. Each region of interest in the genome is covered several times, by sequencing millions of DNA fragments coming from multiple cells. The reads (i.e sequenced fragments) are then mapped to a reference genome, generating piles of reads (Figure 1.1).

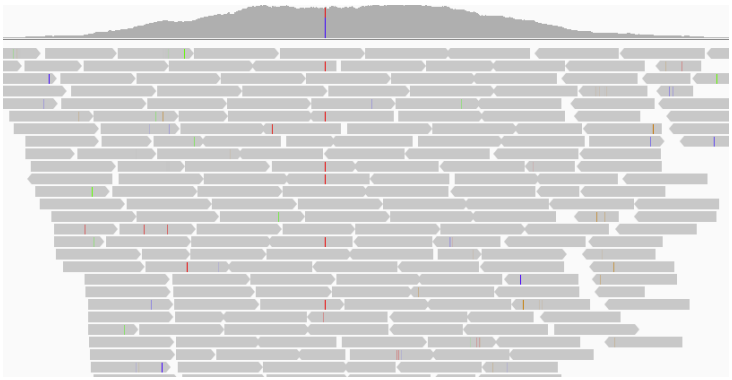

Figure 1.1: Sequenced DNA fragments are mapped to the human reference genome and viewed by the visualization tool Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013). The colored vertical lines represent positions where there are discrepancies (variant alleles) compared to the reference. On top, a histogram shows the number of times each position is read.

When identifying mutations using DNA sequence data, the task is to decide for which positions there actually are mutations and for which positions discrepancies from the reference genome only represent errors in the data. The methods need to be sensitive, to detect mutations also in regions with few sequenced fragments, and at the same time have a low false positive rate, due to the high dimensionality of the data. As an example, only sequencing the set of protein-coding genes in humans gives around 50 million positions and the whole human genome has over 3 billion nucleotides. To have both high sensitivity and specificity is however a non-trivial task, since the data contains considerable levels of noise coming from errors introduced during sample preparations and sequencing of the DNA, and from limitations in the bioinformatic data processing (Olson et al., 2015). As an example, in the sample preparation the DNA fragments need to be amplified. This can lead to insertion of wrong nucleotides and bias in what parts of the genome that are amplified (Aird et al., 2011). Furthermore, a typical read is around 100 nucleotides long and often contain multiple sites with sequencing errors. The relatively short fragment length and errors within the reads can make it hard to decide

where a read should be correctly placed along the reference. The problem is more severe in genome regions with repetitive patterns and for genes that are evolutionary closely related and hence may have similar nucleotide sequences (Treangen and Salzberg, 2012). Accumulation of incorrectly placed reads can lead to discrepancies from the reference genome that are artificial but look like true mutations. Moreover, the errors introduced during sequencing are both random and systematic. For example, errors occur more often at the end of reads and within certain patterns in the nucleotide sequence (Minoche et al., 2011). If not accounted for, these errors can lead to biased results. A quality score, related to the probability of an error, is estimated for each nucleotide and read placement. These scores can be utilized in bioinformatical algorithms and statistical models used to preprocess the reads and infer mutations. In conclusion, to be able to employ the NGS data and transform it into accurate information that can be used for new biological insights, statistical methods and competence are crucial. There is however a need for development of new and tailored computational and statistical methods to take full advantage of the potential in the data.

To summarize, the research activity utilizing next generation sequencing techniques have grown dramatically over the past years. Still we are only in the beginning of what can be investigated and to which accuracy. Improved bioinformatical and statistical methods are important to develop, as well as infrastructure to handle the vast amount of data The experimental methods continue to evolve, producing data with higher quality and new types of information. One example is techniques for sequencing the genetic material from single cells, instead of pools of cells that are done today, giving even higher resolution in the information (Shapiro et al., 2013). To implement sequencing techniques, and the possibilities they give, into daily clinical use in our hospitals is also an ongoing area. Altogether, we are in a position with unprecedented and evolving opportunities for studying gene function, biological diversity on different levels and evolution.

Chapter 2

Finding somatic mutations in exome sequencing data

In *paper I* and *paper II* the focus is on finding tumor-specific mutations in protein-coding regions (the exome) for the cancer types pheochromocytoma/ paraganglioma and acute myeloid leukemia, respectively. To complement the relatively brief method sections in *paper I* and *paper II*, the bioinformatical and statistical approaches used to preprocess the raw sequence data, identify candidate somatic mutations and filter for technical valid and biological important somatic mutations are summarized in the sections below.

In a tumor cell there is a mix of inherited (germline) mutations, and mutations that are specific to the tumor cells, denoted somatic mutations. The search for somatic mutations is, in important aspects, different from identifying germline mutations. One aspect concerns in what ratio sequenced fragments are expected to harbor discrepancies from the reference genome for a somatic mutation. In human cells, all but the sex chromosomes are inherited pairwise, giving two copies of each gene. In general, all the cells in the body contains the same genetic material. A germline mutation is therefore expected to on average show up in 50% (heterozygous) or 100% (homozygous) of the sequenced fragments. That is, the variant allele frequency (VAF) is expected to be 50% or 100%. However, in samples from tumor cells this is no longer the case. During development and growth of a tumor, new somatic mutations are acquired. Tumors are thus often heterogeneous and have subclones, meaning that the genetic material differs between groups of cells. While some somatic mutations can be common to all cells, due to an early mutation event or a large selective advantage, others exists only in subclones. Both types are important to find, in order to fully understand the genetic origin and which combinations that cause different properties. Furthermore, samples from tumors often contain normal

cells to a certain extent. Tumor cells can also mutate to have more or less than two copies of each gene. Altogether, this means that the ratio of chromosomes in the sample harboring a somatic mutation can take on values from, in principle, just above zero to 100%. Thus, the assumption on which VAF to expect in the data for a position with a somatic mutation must be relaxed.

Another aspect is directly connected to the definition of a somatic mutation; that it should not be present in normal cells. A paired experimental design, including samples from both normal and tumor cells from each patient, is therefore needed. For each position, the tumor and normal sequence data is compared and if variant alleles are present in the tumor sample but not in the normal sample, a candidate position for a somatic mutation is found (Figure 2.1).

Figure 2.1: Sequenced DNA fragments aligned to the human reference genome and viewed by Integrative Genomics Viewer, where variant alleles are shown by colored letters. Variant alleles are detected in the tumor sample but not in the normal, i.e. a candidate somatic mutation.

2.1 Preprocessing of the data

The purpose of the preprocessing of the data, before the actual identification of the mutations, is to correct or at least compensate for errors introduced during sample preparation, sequencing and mapping to the reference genome.

We start from the point where we have access to the reads, i.e. sequenced DNA fragments, and the quality value for each sequenced nucleotide (denoted Q). The scores are related to the probability of an sequencing error, P, according to

$$Q = -10 \log_{10} P.$$

The first step is to filter the data based on the quality values, to ensure that reads with overall low quality is discarded. Also, the quality often drops towards the end of the reads, and such stretches can be trimmed of during the filtering

step. In *paper I* and *paper II* we used the tool PRINSEQ to perform quality filtering (Schmieder and Edwards, 2011).

Then, the reads are mapped (aligned) to the reference genome. That is, the original position in the genome for each sequenced DNA fragment is to be found. A number of different algorithms for aligning reads have been developed (Li and Homer, 2010). In *paper I* and *paper II* we used BWA (Li and Durbin, 2010) in paired-end mode, as recommended in the Best Practices developed at the Broad Institute (Van der Auwera et al., 2013). Paired-end refers to the type of sequencing performed, where the DNA fragment size is aimed at being at least twice the length of a read and then the fragment is sequenced from both edges. In this way the mapping is enhanced, since information from both reads in a pair can be utilized. A mapping quality score is assigned to each read, indicating how well the read matched the reference sequence. Reads who matched several intervals in the reference equally well are flagged, by giving a mapping quality score of zero.

The DNA amplification used in the sample preparation can lead to the same fragment being sequenced twice or more, especially when low amount of DNA is used as input material. To not account for the same information several times, such duplicated reads needs to be removed. We used a tool called Picard (http://broadinstitute.github.io/picard) to mark read-pairs with the same genomic starting positions for both reads as duplicates. In duplicate marking, all reads from the same sample preparation must considered simultaneously. It is worth noting that when using formalin-fixed paraffin-embedded (FFPE) tumor material as in *paper I*, the levels of duplicates were in general much higher than for fresh-frozen (SF) material. This is likely due to the additional rounds of amplification that were needed in the sample preparation of the FFPE material. Also, the Picard algorithm left to a larger extent duplicated reads unmarked, due to inconsistent mapping of one of the reads in the pair, in the data from FFPE samples. This produced a significant amount of false positive somatic mutations in the FFPE material. We removed those by adding a down-stream filter requiring the mutations to be found in several different positions in the supporting reads. This problem was also noticed and solved similarly by another study utilizing FFPE material (Yost et al., 2012).

In regions with insertion or deletions in the sequenced sample, the mapping algorithm often have a hard time deciding whether to include indels or mismatched nucleotides in the alignment, especially at the end of reads. Each read-pair is mapped independently of the others, which can produce inconsistent decisions for different reads at the same position. The process of correcting for such inconsistency is called indel realignment. Intervals that need to be corrected are searched for and all reads in such an interval are realigned together (DePristo et al., 2011). For a paired design with samples from normal and tumor cells, it is important to perform indel realignment with all reads from one

patient included at the same time. Otherwise, different consensus decisions may be taken in the tumor and the normal samples, creating false positives when inferring somatic mutations.

The quality scores for the nucleotides are used extensively in the algorithms for identifying mutations, and in the subsequent filters. However, the quality scores contain systematic errors associated to for example sequencing machine cycle and sequence context. A process called "base recalibration" is shown to effectively reduce the bias in the quality scores (DePristo et al., 2011). In *paper I* and *paper II* we applied base recalibration taking sequence context, sequencing cycle, original base quality score and read group ID into account (Van der Auwera et al., 2013). The read group ID gathers reads from the same sample preparation and machine lane together.

2.2 Identification of candidate somatic mutations

To identify genomic positions harboring a mutation in the sample, positions with mutations need to be differentiated from those where discrepancies in the matching of the reads to the reference genome only represents noise in the data. A statistical method for identifying somatic mutations need to be sensitive, both to be able to detect low-frequency mutations and mutations in regions with low sequence coverage. At the same time specificity is important, because of the many positions to be considered and the considerable levels of noise in the data. Also, a mutation found in the tumor need to be classified as somatic or germline, for which the normal sample is utilized. Typically a statistical model is used to identify candidate somatic mutations, followed by filtering the candidate list to further remove false positives. The first step is described in this section, while the filtering part is the topic of the following section.

One method that has been used to identify candidate somatic mutations, especially in early studies, is a simple comparison between mutation lists from tumor and normal samples (Pleasance et al., 2010). One starts with using a method for identifying germline mutations, such as for example Unified Genotyper (DePristo et al., 2011), on the normal and tumor sample separately. Then the list of mutations in the normal sample is subtracted from the list of mutations in the tumor sample. One major disadvantage with this method is that low-frequency mutations in the tumor are in risk of being missed, since the statistical model, incorrectly, assume heterozygous or homozygous (VAF 0.5 or 1.0) mutations. Further, all germline mutations that are missed in the normal sample but detected in the tumor will show up as false positive somatic mutations.

To improve the performance, a number of dedicated statistical methods for identifying somatic mutations have recently been developed (Roth et al., 2012;

Cibulskis et al., 2013; Larson et al., 2012; Saunders et al., 2012; Koboldt et al., 2012). Several articles comparing these methods have been published, for example Wang et al. (2013) and Xu et al. (2014). For the subset of these methods where the statistical model is set up under the assumption of only heterozygous and homozygous mutations, the sensitivity to detect low-frequency mutations is shown to be reduced.

To identify candidate SNVs in *paper I* and *paper II*, we used a method called MuTect (Cibulskis et al., 2013). It allows for a continuous range of possible frequencies for the sought somatic mutations in its statistical model. To detect genomic positions with a mutation in the tumor sample, MuTect applies a Bayesian classifier. Two alternative models are considered for each position harboring variant alleles in the data, one denoted $L(M_f^m)$ assuming that a variant allele m with frequency f is present in the sample and one denoted $L(M_0)$ assuming that no variant alleles truly exist in the sample. The likelihood for each model is calculated based on the sequence data, taking the read nucleotides and their quality scores into account. For details on calculation of the likelihoods, see Online Methods in Cibulskis et al. (2013). The ratio of the likelihoods times the prior probability for each model is calculated and compared to a decision threshold $\log_{10} \delta_T$:

$$\log_{10} \frac{L(M_f^m)P(m,f)}{L(M_0)(1-P(m,f))} \geq \log_{10} \delta_T.$$

The choice of δ_T tells how many times more confident one wants to be in the model with a mutation, to declare it as a candidate mutation. By assuming a constant prior probability $P(m,f)$, the equation can be rearranged to

$$\log_{10} \frac{L(M_f^m)}{L(M_0)} \geq \theta_T,$$

where θ_T is a constant depending, on δ_T and $P(m,f)$, that can be tuned to achieve different sensitivity. When the performance of MuTect was evaluated in Cibulskis et al. (2013), a δ_T of 2 and a prior probability for a somatic mutation of 3×10^{-6} were chosen, yielding a threshold of $\theta_T = 6.3$. In *paper I* and *paper II* we instead choose to set $\theta_T = 8$, representing both a lower prior probability for somatic mutations in the studied tumor types and a higher ratio of the likelihoods needed to call a mutation.

For each position with a candidate mutation in the tumor, a similar method is used for the normal data in order to classify the mutation as somatic or germline. The mutation frequency in the model with a germline mutation is assumed to be 0.5 (assuming heterozygosity). To assure that there is convincing evidence for *not* having a germline mutation at the position, ten times higher likelihood for the model without a mutation is required to classify a candidate mutation as somatic. In addition, a filter for the maximum number,

or proportion, of variant alleles that are allowed to be observed in the normal sample is added. In *paper I* and *paper II* we choose to reject a candidate somatic mutation if three or more variant alleles, or a proportion above 8%, were observed in the normal sample.

To identify candidate indels in *paper I* and *paper II*, we used a combination of two methods, Varscan2 and Strelka (Koboldt et al., 2012; Saunders et al., 2012). In Varscan2, all positions in the normal and tumor samples are first inspected separately to see if there are a larger proportion of variant alleles in the data than a user-defined threshold. In *paper I* and *paper II* we set the threshold to 0.05. For positions where the threshold is exceeded in the tumor but not in the normal, the Fisher's exact test is used to test if there is evidence for a significant difference in allele frequency between tumor and normal. In Strelka, a Bayesian approach is instead used. Briefly, the VAF in the normal sample is modeled as a mixture of heterozygous/homozygous genotypes and noise. The VAF in the tumor sample is modeled as a mixture of the normal sample and additional somatic variation. Thereby, a continuous range of possible frequencies for the sought somatic mutations are allowed for, and base qualities are taken into account. For full details about the statistical model used in Strelka, see Saunders et al. (2012).

Finally, it is worth noting that before applying the statistical models described in this section, all methods have their own prior filtering regarding which positions that have enough covering data to be evaluated and which reads that are of sufficient quality to be used. For each method utilized in *paper I* and *paper II*, we used the default settings (Cibulskis et al. (2013), http://varscan.sourceforge.net/, Saunders et al. (2012)).

2.3 Filtering of candidate somatic mutations

The methods used to identify candidate somatic mutations operates at data from one position at the time, assuming that the sequencing errors are random and independent, and further that all reads are aligned correctly. These assumptions are in general not met. For example, reads can be aligned at the wrong place or with wrong decision where to incorporate mismatches/indels, and sequence errors tend to accumulate for certain preceding sequence patterns. Thus, the whole error structure in the data is complex and not fully captured by the models and the list of candidate somatic mutations often contains a high rate of false positives. The statistical models described above are therefore in general complemented with different approaches for filtering the list of candidate somatic mutations. For example, MuTect has multiple numbers of filters implemented that we applied to the lists of candidate somatic mutations in *paper I* and *paper II* (Cibulskis et al., 2013). Important steps of the filtering includes removing mutations at positions with proximal gaps, i.e. where the

aligned reads spanning the position harbor surrounding indels, and mutations where the variant alleles mainly sit at the start or end of reads. Furthermore, positions where the mapping quality scores are low for the reads supporting the mutation or indicate that many of the reads could equally well have been placed at another region, are also excluded. Another example is the removal of mutations with strand bias, i.e where mismatches are seen mainly in one read direction and thus can be assumed to be dependent on sequence context.

However, the filters added to each method do not cover all systematic errors that can occur. The paired design used when calling somatic mutations means that data from each tumor sample is compared to data from the normal sample in the same patient. The aim is primarily to exclude germline mutations, however technical artifacts present in both both tumor and normal data are also captured. To remove false positives due to rare but systematic position-specific errors, not only the paired normal sample but all of the normal samples can be utilized. In *paper I* and *paper II* we used an approach where we screened all the normal samples at all positions where candidate somatic SNVs were identified. If two or more samples failed the normal criteria (at most 2 reads or 8% of the reads harboring the variant allele) at a specific position, the corresponding candidate somatic mutation was excluded. During the analysis of the data in *paper I* and *paper II* we also noticed that mutations or variant alleles were identified recurrently at certain positions for samples sequenced under the same conditions. That is, position-specific sequencing errors correlated to sequencing machine (e.g HiScanSQ or NextSeq) and its settings, including chemistry version, were present. An important aspect of the study design is therefore to handle and sequence paired samples together. Furthermore, to fully utilize the screening of normal samples, it is of value to have access to as many other samples as possible sequenced under similar conditions.

We have now arrived at a list of somatic mutations that are evaluated and filtered from a technical perspective. One remaining question is which of the somatic mutations that influence the disease in a crucial way and which that are merely passenger mutations. Worth noting is that in *paper I* this is an important part of the aims, while the functional consequences is not in focus in *paper II* since the search there is rather for a genetic marker with the property of being present in all tumor cells.

A first step to elucidate the importance of the mutations is to annotate them with respect to gene name, location in functional elements, if any amino acid substitution occur and the (germline) population frequency of the mutation. In *paper I* and *paper II* we used the tool ANNOVAR to annotate the list of somatic mutations (Wang et al., 2010). Mutations with a population frequency below 1% were excluded from the lists. In the *paper II*, these common mutations are removed due to higher likelihood of being missed germline mutations, and thus unsuitable as tumor cell markers. This also applies to the case in *paper I*, and

in addition the probability is low that a mutation common in the population causes the rare cancer disease that is studied. For the mutations located in protein-coding parts, only mutations resulting in a change in the amino acid chain of the encoded protein (nonsynonymous mutations) were kept in *paper I*. A change in the amino acid chain is a prerequisite for altering the function of a protein, but different changes affects the protein structure and function to different extents. As a further guidance to the functional consequences of the somatic mutations found in *paper I*, we also annotated the mutations with the scores from five different functional prediction algorithms (Liu et al., 2013).

A strong criterion for influence on the disease is whether a gene is mutated recurrently, i.e. has somatic mutations in several patients. However, when analysing large collection of samples or tumors with high mutation rate, recurrent mutations in a gene can occur just by chance, especially for large proteins. There are statistical methods testing for the hypothesis that a gene exhibits more mutations than expected according to the background mutation rate (Raphael et al., 2014). In *paper I* such tests on gene level were not applicable, due to the low somatic mutation rate in combination with a heterogeneous disease and rather few samples. We instead high-lighted all the genes that harbored recurrent mutations, with the exception of genes that were previously suggested to often represent false positives in cancer studies due to large size or high mutation frequency (Lawrence et al., 2013).

Chapter 3

Summary of papers

3.1 Paper I – Malignant pheochromocytomas/ paragangliomas harbor mutations in transport and cell adhesion genes

Pheochromocytoma (PCC) and paraganglioma (PGL) are rare neuroendocrine tumors, located in the adrenal medulla or extra-adrenal paraganglia. Just over 10% of the patients with a primary PCC/PGL tumor develop malignant disease. The prognosis for patients with malignant disease is poor and the metastases may occur several years after removing the primary tumor. Thus long-term surveillance of PCC/PGL patients is required. This is emphasized by the fact that although some factors that may indicate a higher risk of future malignancy are known, there is currently no reliable way to predict if a primary tumor will metastasize or not. Inherited mutations predisposing for PCC/PGL have been characterized, but less is known about additional somatic events leading to tumor progression and malignancy.

In *paper I*, the aim was to investigate somatic mutations in benign and malignant PCCs/PGLs tumors and identify somatic mutations that contribute to the malignant transformation. Exome-sequencing of paired samples (normal–tumor) from four patients with benign and five patients with malignant tumors was performed. Two biological replicates were taken from each tumor, one from fresh-frozen (SF) and one from formalin-fixed paraffin embedded (FFPE) material. The raw sequencing data was quality-trimmed, aligned to the human reference genome, marked for duplicates, realigned patient-wise and base recalibrated. Somatic SNVs and indels were then identified, annotated and filtrated.

The resulting landscape of somatic mutations included 225 unique mutations, located in 215 genes, and with an median variant allele frequency of

0.27. The average mutation rate per sample was 0.54 mutations/megabase, placing the mutation rate of PCC/PGL tumors in the lower range compared to other cancer types. A significantly higher rate of mutations in malignant tumors compared to benign ones was seen. Four genes had somatic mutations in more than one patient; *HRAS*, *MYCN*, *MYO5B* and *VCL*. Mutations in *HRAS* were found in benign sporadic cases, similar to the findings in previous studies of PCC/PGL. Recurrent mutations in *MYCN*, *MYO5B* and *VCL* are however novel findings in PCC/PGL and were exclusively found in malignant PGL cases. Out of these three mutations, *MYCN* is a previously known oncogene. *MYO5B* and *VCL* have functions related to cell migration, an important mechanism for malignant potential in tumors. When screening publicly available PCC/PGL datasets, three additional *MYO5B* mutations were found, two in patients with malignant disease and one in a tumor displaying pathological risk factors for malignancy.

The overlap between SF and FFPE samples was in general high, with on average 58% of the mutations found in SF samples also present in corresponding FFPE samples. This exemplifies the usefulness of FFPE material in exome-sequencing studies for somatic mutations. Also, the unique mutations identified in each sample confirms the heterogeneity of tumors and shows that biological replicates contribute to a more complete picture of the landscape of somatic mutations.

3.2 Paper II – Patient-tailored analysis of minimal residual disease in acute myeloid leukemia using next generation sequencing

Acute myeloid leukemia (AML) is the most common form of acute leukemia in adults. The initial treatment is based on chemotherapy, and to guide the choice of treatment intensity, risk stratification tools are of great importance. One of the most important factors for risk stratification is early response to treatment. The response is monitored by analysing the levels of minimal residual disease (MRD), i.e. the amount of remaining leukemic cells. This analysis also has an important role in monitoring patients in remission after treatment with a high risk of relapse. Today, multiparameter flow cytometry (MFC) is the most commonly used method for MRD analysis. It utilizes the immunophenotype of the leukemic cells, i.e. the set of expressed proteins connected to the cell membrane. The technique has several disadvantages, including a potential shift in the immunophenotype, which can lead to false negative results. Instead genetic aberrations in the leukemic cells can be utilized for MRD analysis. However, the genetic heterogeneity of the disease means that there is no limited set of recurrent genetic variants that can be used in all cases.

3.3. Paper III – A hierarchical Bayesian model for assessing differential nucleotide composition between metagenomes.

In the work described in *paper II* the aim was to identify leukemia-specific mutations in patients with AML and evaluate their suitability for patient-tailored MRD analysis. For a mutation to be suitable for MRD analysis and avoid false negative results due to subclonality, it is important that the mutation is both present in all leukemic cells and does not exist in any other cells. To get the profiles of somatic mutations in individual patients, leukemic cells and normal lymphocytes were isolated from 17 patients with AML using fluorescence activated cell sorting. The two fractions were then exome sequenced. After data preprocessing, identification of candidate somatic mutations and filtering to remove potential false positives, in total 262 somatic SNVs and indels were found. The majority of the mutations had a variant allele frequency (VAF) around 0.5, corresponding to being present as heterozygous mutations in all leukemic cells. A comparison of the observed VAF distribution to a simulated distribution was done, taking sequencing depth into account and assuming heterozygosity for all mutations in the simulation. Although it showed an overall correspondence, some mutations had lower VAF than expected from the simulation. To remove mutations not likely to be present in all leukemic cells, a 95% confidence interval around the VAF of each mutation was calculated. Mutations where the interval was below 0.50 were excluded. In total 191 leukemia-specific mutations passed this filtering and were thus considered as candidates for MRD analysis. All patients but one had MRD candidates in their somatic mutation profile (median 11 per case, range 0-25).

To detect the low frequencies of mutations that are desirable in MRD analysis, targeted deep sequencing, where specific parts of the genome is selected and sequenced to a high depth, can be utilized. The technique was used on follow-up samples from a patient with AML. Four mutations from the set of previously identified MRD candidates for the patient were analysed. The results showed that this approach for MRD analysis was more sensitive than the ordinary MFC method. Furthermore, when the MFC method failed to correctly capture the relapse after 10 months, due to a change in immunophenotype for a majority of the leukemic cells, all four of the somatic mutations were detected with a high mutation load.

3.3 Paper III – A hierarchical Bayesian model for assessing differential nucleotide composition between metagenomes.

In *paper III* the focus is shifted from humans to metagenomes and from detecting mutations in one individual at the time to comparing groups of samples from different conditions. A metagenome consists of all the genetic material in an environmental sample, which can be a complex mixture of thousands

of species. The development of new DNA sequencing techniques have revolutionized the way we can study such microbial communities, with access to information about the genetic material from in principle all species down to nucleotide resolution. This has been utilized for studying both the human microbiome and its connection to diseases, as well as the structure and diversity of environmental microbial communities under different conditions. Still, the detailed information on which genetic variants that are selected for in microbial communities under different experimental, medical and environmental conditions remains to a large extent to be studied. However, the statistical analysis of genetic variants in metagenomes and comparison between conditions is challenging. The data exhibits considerable levels of noise and the biological variability between communities is often large. Both high sensitivity and specificity is needed, especially since the datasets typically contains few samples and are high-dimensional.

The work in *paper III* is part of an ongoing study with an overall aim to develop a method that is able to make accurate inference about changes in nucleotide composition, i.e. genetic variants, between groups of metagenomes from different conditions. A hierarchical Bayesian model for the observed nucleotide counts at each genomic position is proposed. The model is based on a Dirichlet-multinomial distribution, where the multinomial part accounts for the within-sample variability arising from picking DNA fragments at random for sequencing. The Dirichlet distribution models the sample-specific nucleotide composition and hence accounts for the between-sample variability. The performance of the proposed model is evaluated for simulated data and compared to using the ordinary χ^2-test. The results show that the model has a high sensitivity to detect positions with a difference in nucleotide composition. The false positive rate is kept at a low level even with high levels of between-sample variability, which is not the case for the χ^2-test. The ability to differentiate between positions with and without effect is thus considerable improved. The study demonstrates the importance of methods that models the biological and technical variability encountered in metagenomics data, in order to be able to make accurate inference about differential nucleotide composition between conditions.

Chapter 4

Future work

The work described in *paper III* is an ongoing study. A model for assessing differential nucleotide composition, based on the Dirichlet-multinomial distribution, is proposed and evaluated for simulated data.

As future work, we aim to improve the estimation of the position-specific overdispersion parameters, by using a shrinkage approach. This will be implemented through adding a common prior distribution for the overdispersion parameters to the model. We will evaluate different choices of prior distribution, based on the fit to the overdispersion distribution in real data and the estimated impact on model performance. In addition, we will consider alternative ranking scores, with the aim to find a score that is more robust towards the combination of high overdispersion and skewed nucleotide compositions

The evaluation of the current model was done for two different nucleotide compositions and effects. However, real data may contain many more possible configurations of nucleotide compositions and effects. Also, a mix of those and different levels of biological and technical variability will be encountered. We will therefore perform a more extensive evaluation of the model with a larger set of composition and effect combinations. The ranking ability when mixing positions with different overdispersion will also be investigated.

When evaluating the performance of a model, using data simulated under the model will give to optimistic results. In reality, all the assumptions of the model and in the simulations are not fulfilled and the structure and complexity of real data is thus different. The advantage of simulated data is that the true answer to the inference problem is known, which is typically not the case in real data. We will therefore also evaluate the model performance using resampling of real metagenomic data (similar to Jonsson et al. (2016)). The idea is to have a larger collection of metagenomes and randomly select two subsets of those. An effect is then added to a part of the positions in one group, according to some model. In this way properties such as gene abundance, sequence depth

distribution, nucleotide composition and overdispersion will be according to real data. Still, the procedure can be repeated many times and the model evaluated for sensitivity and specificity. Furthermore, using resampled data will give us the opportunity to try out different choices in the preprocessing of the data and in the filtering of candidate positions, and adopt those to suit real metagenomic data.

Finally, we will apply the model on publicly available metagenomic data. In particular, we aim to investigate how the selective pressure of antibiotics affects the genotype distribution in gut metagenomes.

References

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18.

Armaghany, T., Wilson, J. D., Chu, Q., and Mills, G. (2012). Genetic Alterations in Colorectal Cancer. *Gastrointestinal Cancer Research : GCR*, 5(1):19–27.

Bagel, S., Hüllen, V., Wiedemann, B., and Heisig, P. (1999). Impact of gyrA and parC mutations on quinolone resistance, doubling time, and supercoiling degree of Escherichia coli. *Antimicrobial Agents and Chemotherapy*, 43(4):868–875.

Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E., and Bishop, J. M. (1984). Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science (New York, N.Y.)*, 224(4653):1121–1124.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam,

A. T., Davies, H., Stevenson, J. A., Barthorpe, S., Lutz, S. R., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O'Brien, P., Boisvert, J. L., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H. G., Chang, J. W., Baselga, J., Stamenkovic, I., Engelman, J. A., Sharma, S. V., Delattre, O., Saez-Rodriguez, J., Gray, N. S., Settleman, J., Futreal, P. A., Haber, D. A., Stratton, M. R., Ramaswamy, S., McDermott, U., and Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575.

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., Sverrisson, J. T., Gretarsdottir, S., Walters, G. B., Rafnar, T., Thjodleifsson, B., Bjornsson, E. S., Olafsson, S., Thorarinsdottir, H., Steingrimsdottir, T., Gudmundsdottir, T. S., Theodors, A., Jonasson, J. G., Sigurdsson, A., Bjornsdottir, G., Jonsson, J. J., Thorarensen, O., Ludvigsson, P., Gudbjartsson, H., Eyjolfsson, G. I., Sigurdardottir, O., Olafsson, I., Arnar, D. O., Magnusson, O. T., Kong, A., Masson, G., Thorsteinsdottir, U., Helgason, A., Sulem, P., and Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5):435–444.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *To appear in BMC Genomics*.

King, M.-C., Marks, J. H., Mandell, J. B., and New York Breast Cancer Study Group (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science (New York, N.Y.)*, 302(5645):643–646.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–823.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage,

D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld,

A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, 28(3):311–317.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595.

Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483.

Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation*, 34(9):E2393–2402.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science (New York, N.Y.)*, 349(6255):1483–1489.

Metzker, M. L. (2010). Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, 11(1):31–46.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, 12(11):R112.

Nilsson, B., Nilsson, O., and Ahlman, H. (2009). Treatment of gastrointestinal stromal tumours: imatinib, sunitinib – and then? *Expert Opinion on Investigational Drugs*, 18(4):457–468.

Oddsson, A., Sulem, P., Helgason, H., Edvardsson, V. O., Thorleifsson, G., Sveinbjörnsson, G., Haraldsdottir, E., Eyjolfsson, G. I., Sigurdardottir, O., Olafsson, I., Masson, G., Holm, H., Gudbjartsson, D. F., Thorsteinsdottir, U., Indridason, O. S., Palsson, R., and Stefansson, K. (2015). Common and rare variants associated with kidney stones and biochemical traits. *Nature Communications*, 6:7975.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., and Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics*, 6:235.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A., and Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196.

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine*, 6(1):5.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., Marra, M. A., Aparicio, S., and Shah, S. P. (2012). JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(7):907–913.

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)*, 28(14):1811–1817.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, 27(6):863–864.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews. Genetics*, 14(9):618–630.

Swaminathan, B., Thorleifsson, G., Jöud, M., Ali, M., Johnsson, E., Ajore, R., Sulem, P., Halvarsson, B.-M., Eyjolfsson, G., Haraldsdottir, V., Hultman, C., Ingelsson, E., Kristinsson, S. Y., Kähler, A. K., Lenhoff, S., Masson, G., Mellqvist, U.-H., Månsson, R., Nelander, S., Olafsson, I., Sigurðardottir, O., Steingrimsdóttir, H., Vangsted, A., Vogel, U., Waage, A., Nahi, H., Gudbjartsson, D. F., Rafnar, T., Turesson, I., Gullberg, U., Stefánsson, K., Hansson, M., Thorsteinsdóttir, U., and Nilsson, B. (2015). Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nature Communications*, 6:7213.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192.

Thorven, M., Grahn, A., Hedlund, K.-O., Johansson, H., Wahlfrid, C., Larson, G., and Svensson, L. (2005). A homozygous nonsense mutation (428g–>A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *Journal of Virology*, 79(24):15351–15355.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1):36–46.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 11(1110):11.10.1–11.10.33.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–1558.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164.

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K. B., Pao, W., and Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome medicine*, 5(10):91.

Xu, H., DiCarlo, J., Satya, R. V., Peng, Q., and Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC genomics*, 15:244.

Yost, S. E., Smith, E. N., Schwab, R. B., Bao, L., Jung, H., Wang, X., Voest, E., Pierce, J. P., Messer, K., Parker, B. A., Harismendy, O., and Frazer, K. A. (2012). Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Research*, 40(14):e107.

Zahreddine, H. and Borden, K. L. B. (2013). Mechanisms and insights into drug resistance in cancer. *Frontiers in Pharmacology*, 4:28.