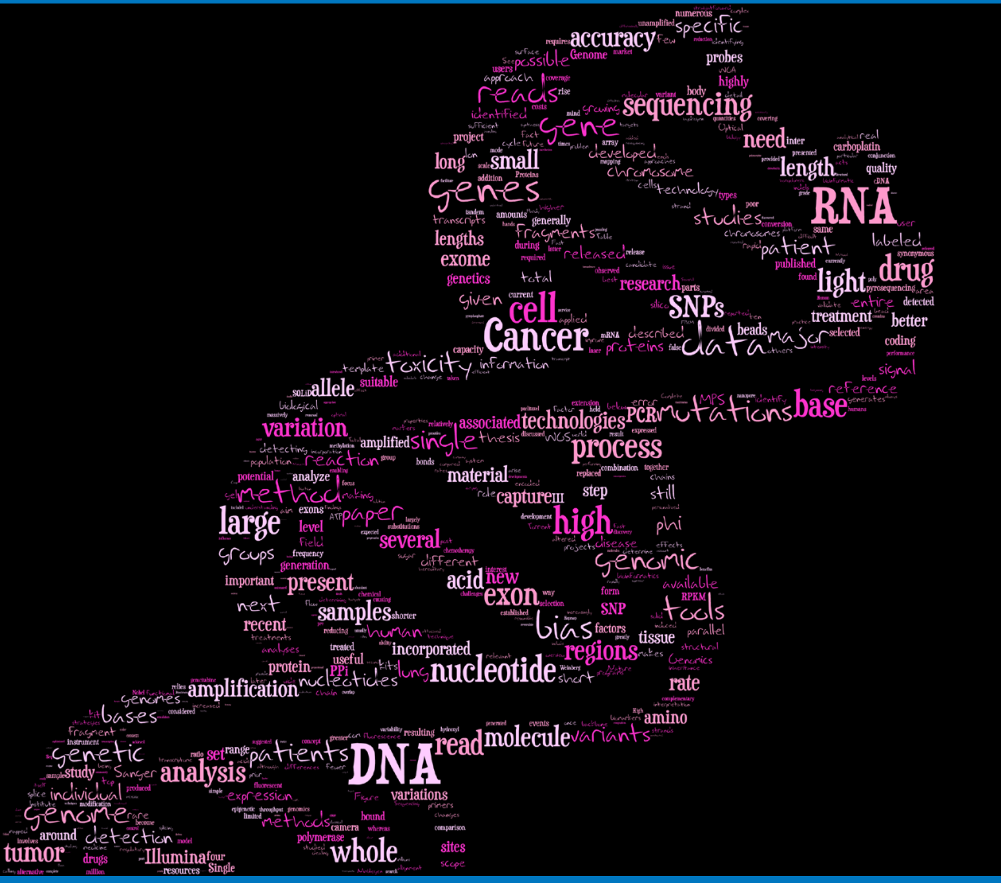
Analysis of genetic variations in cancer

JOHANNA HASMATS

KTH
VETENSKAP
OCH KONST

KTH Biotechnology



Doctoral Thesis in Biotechnology
Stockholm, Sweden 2012

Analysis of genetic variations in cancer

Johanna Hasmats

Royal Institute of Technology
School of Biotechnology
Department of Gene Technology

Stockholm 2012

Royal Institute of Technology
School of Biotechnology
Department of Gene Technology

Science for Life Laboratory
Tomtebodavägen 23B
171 65 Solna
Sweden

Printed by Universitetsservice US-AB
Drottning Kristinas väg 53B
100 44 Stockholm
Sweden

ISBN: 978-91-7501-450-0
TRITA-BIO Report: 2012:18
ISSN: 1654-2312

Cover illustration: DNA helix shaped tag cloud based on the entire content of this thesis (http://tagul.com)

Abstract

The aim of this thesis is to apply recently developed technologies for genomic variation analyses, and to ensure quality of the generated information for use in preclinical cancer research.

Faster access to a patients' full genomic sequence for a lower cost makes it possible for end users such as clinicians and physicians to gain a more complete understanding of the disease status of a patient and adjust treatment accordingly. Correct biological interpretation is important in this context, and can only be provided through fast and simple access to relevant high quality data.

Therefore, we here propose and validate new bioinformatic strategies for biomarker selection for prediction of response to cancer therapy. We initially explored the use of bioinformatic tools to select interesting targets for toxicity in carboplatin and paclitaxel on a smaller scale. From our findings we then further extended the analysis to the entire exome to look for biomarkers as targets for adverse effects from carboplatin and gemcitabine. To investigate any bias introduced by the methods used for targeting the exome, we analyzed the mutation profiles in cancer patients by comparing whole genome amplified DNA to unamplified DNA. In addition, we applied RNA-seq to the same patients to further validate the variations obtained by sequencing of DNA. The understanding of the human cancer genome is growing rapidly, thanks to methodological development of analysis tools. The next step is to implement these tools as a part of a chain from diagnosis of patients to genomic research to personalized treatment.

Keywords: *Cancer, Mutations, Variations, Single Nucleotide Polymorphism, DNA, RNA, Genome, PINK1, Massively Parallel Sequencing, Exome Sequencing, Toxicity*

Great spirits have always found violent opposition from mediocrities. The latter cannot understand it when a man does not thoughtlessly submit to hereditary prejudices but honestly and courageously uses his intelligence.

Albert Einstein, 1940

To the Strong and brave fighting cancer
To the Ones that have loved, and lost
To my Father, To my Brother

You challenge me, You inspire me, You encourage me, You strengthen me.

List of publications

This thesis is based on the following publications, which are referred to in the text by the corresponding Roman numerals (I-IV) and included in the appendix.

I. ***Hasmats J**., Green H., Werne Solnestam B., Zajac P., Huss M., Orear C., Validire P., Bjursell M., Lundeberg J. Validation of whole genome amplification for analysis of the p53 tumor suppressor gene in limited amounts of tumor samples. Biochemical and biophysical research communications, 2012. 425(2): p. 379-83.*

II. ***Hasmats J**., Gréen H., Orear C., Validire P., Huss M., Käller M., Lundeberg J. Assessment of whole genome amplification for sequence capture and massive parallel sequencing. Submitted, 2012*

III. ***Hasmats J**., Kupershmidt I., Rodríguez-Antona C., Su QJ., Khan MS., Jara C., Mielgo X., Lundeberg J., Gréen H. Identification of candidate SNPs for drug induced toxicity from differentially expressed genes in associated tissues. Gene, 2012. 506(1): 62-68.*

IV. ***Hasmats J**., Kupersmidt I., Edsgärd D., de Petris L., Lewensohn R., Alexeyenko A., Blackhall F., Besse B., Lindgren A., Sörenson S., Brandén E., Koyi H., Peterson C., Lundeberg J., Gréen H. Using whole exome sequencing to identify genetic candidates for carboplatin and gemcitabine induced toxicities. Manuscript, 2012*

All papers are reproduced with the permission from the respective copyright holders

Related publications

*Friboulet L., Barrios-Gonzales D., Commo F., Olaussen KA., Vagner S., Adam J., Goubar A., Dorvault N., Lazar V., Job B., Besse B., Validire P., Girard P., Lacroix L., **Hasmats J**., Dufour F., André F., Soria JC. Molecular Characteristics of ERCC1-Negative versus ERCC1-Positive Tumors in Resected NSCLC. Clinical Cancer Research, 2011. 17(17):5562-72.*

Contents

Introduction

The secret to solving the riddle of cancer lies in understanding the disease and using this understanding to find appropriate treatments. To achieve this, it will be necessary to interpret, analyze and understand the patient's genomic background due to its profound influence of the outcome of the treatment. However, the information required to do this is hidden inside the 1.8 m bases of DNA contained inside each cell of the human body.

This thesis focuses on some fundamental technologies used in genomic variation research, and presents studies that validate their use for drawing biological conclusions. I have utilized some of the latest technologies and bioinformatics tools to analyze sequence data, with the aim of identifying new biomarkers for predicting individual patients' toxicity responses to specific cancer treatments.

1

Genomes, genes, genomics, genetics

The interplay between hereditary traits and genes seems to be relatively straightforward at first glance, resembling material covered in introductory molecular biology courses. However, closer inspection quickly reveals a more complex picture. To properly understand what is known about this subject, it is necessary to define some key terms. The ***genome*** is the entirety of an organism's hereditary information (DNA), which is present in almost every cell of the body. The parts of the genome that encode proteins – the main workhorses of the body – are known as ***genes***, and it is estimated that we have 21224 of them (Ensembl release 68). Several approaches and techniques for studying genes and genomes have been developed, and are collectively termed **genomics** - another phenomenon of the "omics" era. [1] Genomics itself can be regarded as a sub-discipline of **genetics**; the science of genes, heredity and variation in living organisms. The studies reported in this thesis used several approaches to study the properties of cancer patients' genomes in order to better understand the genetic components of cancer.

1.1 The Central Dogma

Before the advent of classical genetics, it was generally believed that proteins carried the hereditary information. Classical genetics was founded on Mendel's law of inheritance, which was first outlined 1865 [2], and the chromosome theory of inheritance put forward by Boveri–Sutton a few years later. [3] In 1958, Francis Crick proposed the central dogma of molecular biology [4], which describes a one-way of hereditary information from DNA to RNA to protein within the cell. The theorem was presented in more detail in *Nature* in 1970 [5], further illustrating the complexity of the system. In brief, the central dogma holds that deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA), which carries information that is translated into amino acid chains with specific sequences (Proteins). This process relies on the reliable copying of DNA (replication), cutting (splicing) of RNA and finally the assembly of amino acid chains that subsequently fold into functional three-dimensional proteins. This relatively simple model is largely sufficient for prokaryotes, but in eukaryotes much of the regulatory information is hidden in other parts of the genome, such as non-coding DNA and feedback loops, as well as epigenetic factors and processes such as non-coding RNAs (ncRNAs) and post-translational modification of proteins.

DNA

Deoxyribonucleic acid (DNA) is arguably the most important molecule of life. Its double helical structure was discovered in 1953 by Watson and Crick. [6] DNA consists of a backbone made up of two sugar strands (deoxyribose phosphates) that are wrapped around one-another in a double helix and held together by phosphodiester bonds between the phosphate groups. The sugars are all oriented in the same way, with the 3' hydroxyl group of one being bound to the 5' hydroxyl of the next, making DNA a directional molecule. Each individual sugar moiety is attached to a single nitrogenous base known as a nucleotide, and the two intertwined strands of DNA are bound together by hydrogen bonds between the nucleotides. There are four nucleotides in DNA, each of which forms hydrogen bonds to a specific partner: adenine (A) binds to thymine (T), and cytosine (C) binds to guanine (G). It is these bases that encode the genetic information. The genes comprise the protein-coding sections of the genome, and account for approximately 1.2% of its total length in humans. [7] The DNA is efficiently folded and packed into chromosomes. Humans have 22 pairs of autosomal chromosomes and two sex-determining chromosomes. One set of 23 chromosomes is maternally inherited and the other is paternally inherited to form a diploid genome with 46 chromosomes in total.

The complete sequence (actually, the first draft) of the human genome was first published in 2001. [8, 9] This set the stage for the resequencing of its ~3 billion bases, enabling very large scale genomic analyses.

RNA

Discussions concerning the role of ribonucleic acid (RNA) in protein synthesis began in the late 1930's. Around twenty years later, Severo Ochoa was awarded the Nobel prize in medicine for his discovery of RNA synthesis.

RNA is similar to its precursor, DNA, although it is single stranded. In contrast to DNA, its sugar backbone is made from ribose and the nucleotide thymine is replaced with unmethylated uracil (U). Ribose has a hydroxyl group at the 2' position that is not present in the 2-deoxyribose that makes up the backbone of DNA. This makes RNA more prone to hydrolysis and thus less stable than DNA. In the cell, RNA is produced with a nucleotide sequence that is encoded by the section of DNA from which it is transcribed. There are several different kinds of RNA molecules, including coding mRNA (that holds the information about which proteins to synthesize), and non-coding RNAs such as tRNA (that transfers amino acids to be added to a polypeptide chain), regulatory RNA (RNAi, siRNA, microRNA), rRNA (that deciphers the mRNA in the ribosome) and several others. Transcription of the genetic blueprints encoded in the DNA generates messenger RNA (mRNA).

The entire set of transcripts produced by an organism is called its transcriptome. [10] In September 2012, the ENCODE consortium simultaneously released 30 papers in Nature, stating that approximately 60% of the human genome is transcribed into transcripts more than 200 nucleotides long. [11]

Protein

In 1838, the Swedish chemist Jöns Jacob Berzelius gave the name "proteins" to the macromolecules whose chemical composition was first described by his associate, Gerardus Johannes Mulder. [12]

Proteins consist of amino acid chains with variable lengths known as polypeptides. They are held together by peptide bonds and typically fold into complex tertiary structures. Most proteins are made up of 20 different amino acids, with the precise sequence of amino acids in each protein being determined by a series of three-base sequences known as codons in the DNA.

Proteins can undergo several post-translational modifications, including the addition of sugars to the amino acid backbone (glycosylation) as well as the addition of lipids, carbohydrates, acetyl groups (acetylation), phosphate groups (phosphorylation), and various other moieties. These changes generally alter the protein's chemical properties and can thereby affect its function. [13] Proteins have a wide variety of functions: some (known as enzymes) are catalysts, others are involved in cell signaling, and others still have structural roles, among other things. Together with alternative splicing of mRNAs, post-translational modification means that more than 100,000 different proteins can be produced from the human genome. This scope for the production of variant proteins plays a large role in determining the differences between individuals.

1.2 Variations

Before Mendel's theories became accepted, population genetics were generally discussed in terms of blended inheritance. However, Mendelian genetics and the theory of evolution made it clear that genomic variation existed and had to be accounted for, although for quite a while it was believed that the proteins were the factors of inheritance. [14]

The sequences of any two human genomes will only differ by ~0.1-1%. This raises a question: what is responsible for the phenotypic variation observed in human populations? Some traits, such as the natural color of one's hair or eyes, are clearly determined by genetic factors, but others (such as weight) are also influenced by environmental factors such as nutritional intake and physical activity levels. An individual's genetic makeup can be studied to determine how susceptible they are to certain diseases or how likely it is that they will respond well to specific treatments for conditions such as cancer. This could for example be done by considering simple measures such as deviations from the Hardy-Weinberg principle, which states that allele and genotype frequencies in a population will remain in equilibrium from generation to generation in the absence of disturbing influences. [15] Notably, alleles that satisfy Hardy-Weinberg could still mediate a risk to disease.

It is important to keep in mind that the modern definition of a genetic variation is a difference of one nucleotide or more relative to the sequence of a reference genome, which has become fixated in a population. Mutations, on the other hand, are rare and might be specific to a given phenotype (for example, they may be tumor specific), being acquired due to disease progression or other some other factor (e.g. UV radiation). Initiatives such as the Hapmap project, with the aim to catalogue the genetic variations present in different ethnic groups, map correlations between nearby variants and allele frequencies [16], have greatly increased the scope for eliminating noise due to population differences. Other examples such as the 1000 genome project [17], the sequencing of 200 exomes in Denmark [18], provide useful data. While the preferred database will depend on the aim of the study to be conducted, a recent comparison of the Hapmap and 1000 Genomes stated that the two resources have a 99 % overlap when a minor allele frequency filtering of 5% is applied. [19]

Singe nucleotide variation

A single nucleotide variation (SNV) is defined as a single nucleotide exchange in the genome; these mutations can be categorized into private, familial, rare and common of a population. The human mutation rate has been estimated to about one per one hundred million, however results published in 2011 indicate that on average, humans inherit 60 new mutations per generation from their maternal and paternal genomes. [20] (See Figure 1A)

Single nucleotide polymorphism

A single nucleotide polymorphism (SNP) is a change in a single base of the genome that is present in more than 1 % of a given population. There are currently 187,852,828 human SNPs listed (NCBI dbSNP Build 137 for Human). If the altered allele is present on both chromosomes of a pair, the SNP is said to be homozygous, whereas if one of the alleles is altered within a predefined range (usually 20-80% in observed sequence data), it is said to be heterozygous. A non-synonymous SNP gives rise to an amino acid change whereas a synonymous SNP is silent. Just as important, are SNPs that are situated in regulatory regions and splice sites. They influence gene expression by transcript instability and length. [21] Additionally, SNPs could affect chromatin modifications, methylation or transcription factor binding.

Structural variation

When the variations expand in size to cover larger regions, they are referred to as structural variations. Approximately 12% of all short variants (SNPs, indels, somatic mutations) are reported to be structural variants (Ensemble release 68).

Structural variations can be divided into a few categories. One is copy number variation (CNV), which encompasses deletions, insertions and duplications. Approximately 13% of the human genome originates from copy number variation. [22] Another class is inversions, where a chromosomal section is inverted so that its start and end points switch places. Finally, in translocations, a part of a chromosome is integrated into another chromosome.

A study of the impact of tandem duplications was presented in Nature in 2012, where mutations in these regions are associated with poor prognosis and thus a potential therapeutic target for future medicine. [23]

Rearrangements are often seen in cancers, and modern sequencing technology allows us to study them in more detail. [24] (See Figure 1B-1F)

Repeats

Sequential repeats of two or more nucleotides represent another form of genetic variation and are known as tandem repeats. If the length of the repeated element is between 10 and 60 basepairs, it is referred to as a minisatellite. If the repeated element is shorter than this, it is called a microsatellite or short tandem repeat. [25] Microsatellite instability is common in cancers [26], and is often detected in tumor cells with mismatch repair deficiency (MMR). [27, 28] (See Figure 1G)

Epigenetics

Epigenetic factors contribute substantially to the observed differences between individuals. By definition, (*epi-* (Greek: *επί*-over, above, outer) -*genetics*), epigenetic changes do not affect the sequence of the genome, but they can nevertheless be passed down over several generations of cell divisions. [29] The concept of the "epigenetic landscape" was first discussed in 1942 by Waddington, who used a marble as a metaphor for a cell taking a particular path through valleys of possible differentiation. [30] Epigenetic changes include cytosine methylation and histone modifications. Methylation of promoters is associated with repression, but methylation of genes is a sign of activation [31], whereas histone modification plays a role in transcriptional control. [32, 33] Allelic silencing (e.g expression of genes from either maternal or the paternal allele) was discovered in cancer cells [34], and in 2011, Hansen et al. reported that the variability of differentially methylated regions in the genomes of cancerous cells exceeded that in matched normal samples. [35]

Figure 1

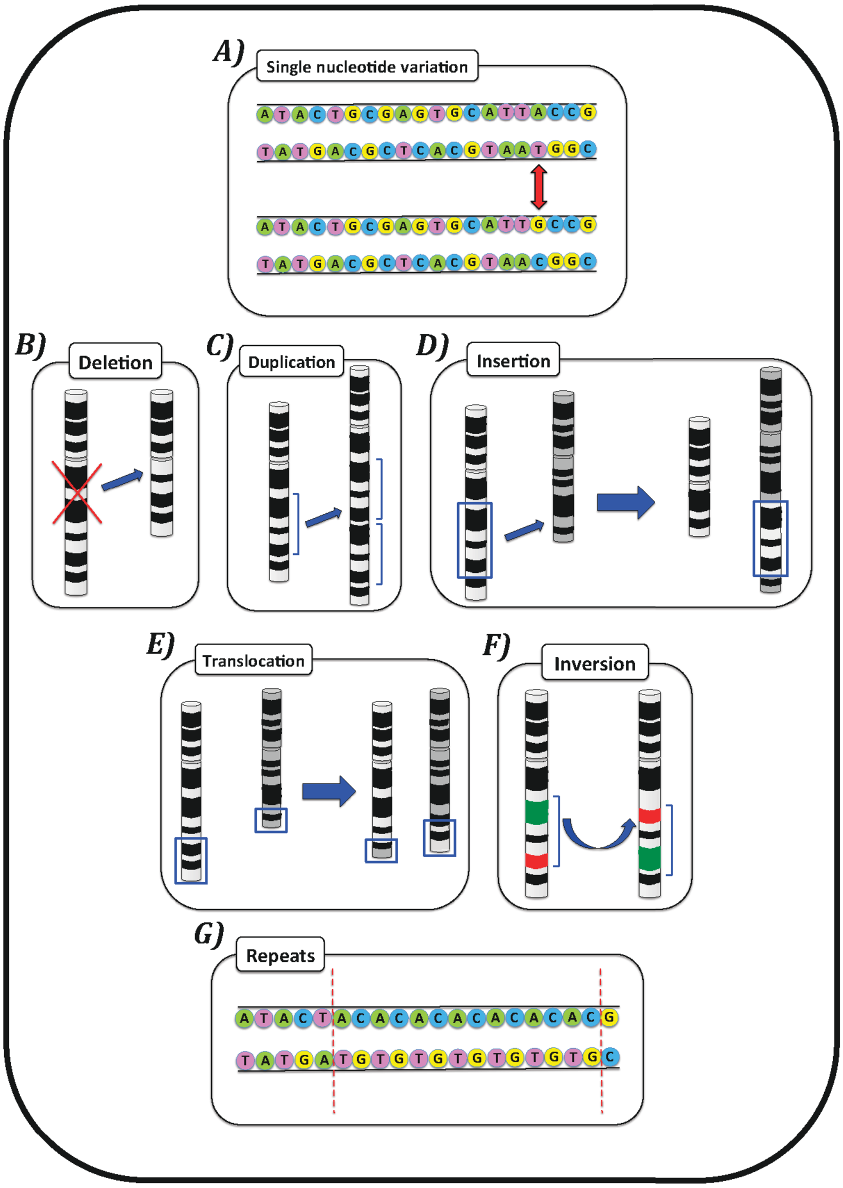

A)
Single nucleotide variation
ATACTGCGAGTGCATTACCG
TATGACGCTCACGTAATGGC
ATACTGCGAGTGCATTGCCG
TATGACGCTCACGTAACGGC
B)
Deletion
C)
Duplication
D)
Insertion
E)
Translocation
F)
Inversion
G)
Repeats
ATACTACACACACACACACG
TATGATGTGTGTGTGTGTGC

Figure 1.
A) Single nucleotide variation. A single nucleotide A (Adenine) has been replaced by a G, (Guanine), base pairing with T (Thymine) and C (Cytosine), respectively.
B) Deletion. A schematic overview of a larger deletion, the size range can vary between one base to several kb.
C) Insertion. A portion of a chromosome is integrated into another chromosome, reducing the size of the donor chromosome, and increasing the size of the receiving chromosome with corresponding size.
D) Duplication. A region of DNA is duplicated within the same chromosome, which could arise from erroneous homologous recombination for example.
E) Translocation. The unequal exchange of chromosomal parts between nonhomologuous chromosomes could for example result in gene fusions.
F) Inversion. When a part of a chromosome rearranges itself so that it is reversed end to end.
G) Repeats. An example of a tandem repeat consisting of two nucleotides A (Adenine) and C (Cytosine), that are directly adjacent.

2

Sequencing methods for detecting variations

Sequencing is the process of determining the order of nucleotides or bases in a length of DNA or RNA, as well as its methylation status in some cases. The first genome to be sequenced was that of bacteriophage φX174 in 1977, and subsequent developments in this field have revolutionized our understanding of genomes.

Different ‘generations’ of sequencing technologies have been developed: the early first generation methods, the more sophisticated and powerful second generation, and the current next generation methods. [36] For the remainder of the text in this thesis, I will use the term massively parallel sequencing (MPS) when referring to the latter.

MPS has enabled us to obtain deeper insights into the structure of the genome and a more detailed view of its properties, which can be studied in greater detail by resequencing.

2.1 Sanger sequencing

The first technique for “decoding” DNA was the chain terminating method described by Frederick Sanger in 1977, for which he was awarded the Nobel prize in 1980. An alternative method was proposed by Allan Maxam and Walter Gilbert, whose method of sequencing by chemical modification was also published in 1977. This method was appealing because it eliminated the need to clone fragments prior to the sequencing reaction, but due to its use of radiolabeled nucleotides and the difficulties of scaling up the process, it was less favored than Sanger's method. [37]

Chain termination sequencing is based on the extension of a DNA template using DNA polymerase, primers, and labeled chain terminating nucleotides known as dideoxynucleotidetriphosphates (ddNTPs). The ddNTPs, which terminate the elongation process, are present in the sequencing reaction mixture in small quantities, causing DNA fragments of varying lengths to be produced from a single template. This collection of fragments is then separated based on their size using polyacrylamide gel electrophoresis. Each column on the gel represents one of the four nucleotides, making it possible to read the DNA sequence by inspecting the size distribution of the fragments. The efficiency of the technique has been greatly increased since it was first developed. Notably, a capillary made of glass filled with a polymer to facilitate automated sequencing in a machine has replaced the gel. This is used in conjunction with visualization software to produce chromatograms showing the DNA sequence in a completely automated and hands-off process.

The Sanger method is currently regarded as the "gold standard" of sequencing methods, with reported error rates varying between 0.001-1% [38] It is therefore used to validate other sequencing methods in laboratories around the world. Its main drawbacks are the time taken for the analysis and its relatively high cost per sample, as well as the poor quality of the sequence information obtained for the first 40 or so bases of the targeted fragment.

2.2 Pyrosequencing

The development of pyrosequencing represented the advent of the second generation sequencing technologies. Pyrosequencing is based on detecting luminescence emitted by luciferase when a nucleotide is incorporated into the growing complementary strand during sequencing by synthesis. This phenomenon was discovered by Nyrén et al [39, 40] in 1986 and is driven by the pyrophosphate (PPi) released when the nucleotide is incorporated. The signal strength depends on the number of identical nucleotides incorporated into the growing strand sequentially. Individual nucleotides are added to the reaction mixture one by one, enabling the detection of distinct light signals arising from the release of PPi. The light-producing reaction involves the conversion of PPi into adenosine triphosphate (ATP) by ATP sulfyrase. The resulting ATP acts as a source of energy to drive the conversion of luciferin to oxyluciferin, which in turn produces visible light that can be detected with a camera. This reaction and the intensity of the emitted light are directly proportional to the amount of PPi released and thus to the number of nucleotides incorporated. The light signal is visualized on a pyrogram, which can be automatically produced by the instrument used to perform the sequencing. However, the relationship between the intensity of the light signal and the number of nucleotides incorporated becomes progressively less linear as the length of the homopolymeric sequence increases.

The read lengths attainable with pyrosequencing are shorter than those for Sanger sequencing. However, its cost, accuracy and time for small-scale genotyping and validation make it very competitive for such purposes. [41] Unlike its successors described below, it is not suitable for sequencing highly repetitive regions and is therefore not optimal for whole genome sequencing.

2.3 Massively parallel sequencing

One of the biggest recent breakthroughs in genomics came from the development of methods for using established sequencing technologies to analyze millions of reads in parallel. [42] This has provided important new insights and greatly increased the scope for exploring and understanding genomes. One of the first major achievements resulting from sequencing was the completion of the entire human genome in 2001 (more accurately, the first draft of the genome [8, 9]), after 13 years of effort in a collaborative project based on the efforts of numerous research groups around the world [43]. Since then, the cost per base of genome sequencing has decreased by a factor of more than 100 000.

The current bottleneck in massively parallel sequencing relates to the lack of suitable methods for handling the large amounts of data it generates, as discussed in chapter 3.

454

The 454 technology is an extension of pyrosequencing and was one of the first "Next Generation" massively parallel sequencing methods to be established, entering the market in 2005. [44, 45] At the time, it was distinguished by its vastly greater capacity compared to pyrosequencing. It is capable of about one million reads per run, with read lengths approaching those achieved with Sanger sequencing.

The process is conceptually similar to pyrosequencing. The DNA to be sequenced is broken up into blunt-ended fragments (DNA libraries) that are then bound to beads. Each bead contains only one DNA molecule, which is amplified through emulsion PCR. [46] After amplification, each DNA library (bead) is transferred to an individual well in a fiber optic microtiter plate. The well also contains DNA polymerase coupled to primed-template on beads, ensuring that the light reaction can proceed. The remaining reagents are provided via the instrument itself, and a CCD camera detects each light signal as a nucleotide is incorporated into the growing DNA fragment. While the intensity of the signal is proportional to the number of bases incorporated, the linearity of the relationship declines when the number of sequential identical bases is greater than 8. As a result, it can be difficult to interpret sequence data for homopolymers using this method.

As the first of the MPS technologies to reach the market, 454 sequencing was rather revolutionary when it was first introduced. However, its pre-eminence is coming to an end, largely because its capacity is on the lower end of current MPS technologies. [47] Its greatest strength is the very long reads it generates, which make it suitable for projects such as long amplicon sequencing. This can be very useful for detecting rare mutations in highly heterogenic cancer samples. It is also useful in another research area, metagenomics, where its accuracy is important for identifying sequences from individual species in highly diverse samples.

Illumina

In 2006, Illumina [48] released their first sequencing instrument [49] based on technology that had been developed since 2001 by the company Solexa. [50] It is based on adaptor ligation onto fragmented DNA, allowing the fragments to bind to primers that are attached to a solid surface. [51] Bridge amplification [52] is

then performed for each template so that clusters of clones originating from a unique DNA molecule can be formed. The addition of four discrete reversible terminator bases by passing a solution containing polymerase over the solid surface is followed by the removal of non-incorporated nucleotides. The power of this method stems from the fact that it can only perform single nucleotide extensions making it suitable for detecting mutations. It also has a very high throughput capacity, reducing its per base cost. Before each new cycle, a camera identifies the newly incorporated fluorescently labeled nucleotide and the 3' end terminator is removed, allowing the next nucleotide to be incorporated.
Sequencing based on reversible dye-termination has an enormous capacity for high throughput data generation, which is very beneficial for the rapid retrieval of genetic information. Its major drawbacks are its short read lengths and comparatively high error rates in homopolymeric regions. However, these must be weighed against the large number of users that can be accommodated, allowing the technology to provide fast solutions to common analytical challenges. The recent development of the bench top MiSeq instrument facilitates fast and efficient screening of selected targets, and ultra deep sequencing of amplicons, allowing for lower prevalence mutation detection. [53]

Solid

The SOLiD system was commercialized in in 2007 and is based on the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) method.[42] [54] The process involves attaching individual DNA fragments to microbeads and then performing emulsion PCR to ensure that each microbead is coated with a single cloned fragment. The beads are then covalently bound to a glass surface. A collection of all possible combinations of labeled octamers are then annealed and ligated, enabling detection by fluorescent imaging. The sequence is determined by the known identities of the first two bases of the octamer. Following detection, chemical removal of the three bases at the 5' end initiates the next cycle. After ten cycles, the generated fragment is removed from the bead, leaving room for the next fragment beginning at the next base.

In practice, this means that each base is called twice, with different color codings, reducing the difficulties of sequencing homopolymers. encountered with Illumina sequencing. However, base substitutions still constitute a problem for this technology. [55] Its high accuracy [56] has made it suitable for applications utilizing short read lengths such as Chip-Seq [57], and the 5500 SOLiD™ System has been used successfully in numerous studies on transcriptomics and epigenomics. [35, 58, 59] With their latest upgrade (Wildfire) the emulsion PCR step is eliminated and replaced with isothermal amplification, making it suitable for exome and RNA sequencing.

2.4 Contemporary technologies

The goal of single molecule sequencing of a full genome was first achieved in 2008, when the first M13 virus was sequenced, although previous attempts had been made already in 2003 by using fluorescence [60] [61]. The number of competing MPS technologies has grown dramatically since the concept was introduced; the main players in the field today are described below.

Semiconductor sequencing

Ion Torrent (currently Life technologies) [62] based their approach on detecting a molecule released during the enzymatic reaction that incorporates the new nucleotide into the growing complementary DNA sequence, as is done in pyrosequencing. However, rather than pyrophosphate, the ion torrent approach detects protons using a semiconductor. The signal is proportional to the number of protons released, which reflects the number of nucleotides incorporated. [63] As in pyrosequencing, homopolymers present a challenge for this method because indels interfere with the signal. [64]

Nanopore sequencing

Nanopore sequencing refers to the process of passing a DNA molecule through a small pore with a diameter in the size range of nanometers in order to determine its nucleotide sequence. The nanopore can be made from a number of different materials, including the protein α-hemolysin. When the DNA molecule passes through the pore, its physical characteristics change. This generates an electrical current if a potential is applied across the system.

The total cost of nanopore sequencing is expected to fall below $900/genome with their miniaturized bench top device MinION[65], making it commercially competitive with the other technologies. One of the major issues with this approach at present is its limited ability to achieve single nucleotide resolution for large-scale workflows. [66]

Single molecule fluorescent sequencing

In 2008, Helicos released their sequencing technology, which is based on single molecule fluorescence and is the first sequencing method that does not require amplification. [61]

In this method, DNA sequences longer than 1000 nucleotides are preferably fragmented, and purified using Solid-phase reversible immobilization (SPRI) beads. [67] They are then hybridized to primers attached on a flow cell and their 3' ends are then blocked by a terminal transferase to prevent extension prior to the initiation of the desired reaction. A labeled nucleotide is incorporated, and its fluorescent dye is detected by laser excitation and then cleaved off, permitting the incorporation of the next complementary nucleotide. The fluorescent tag on the new nucleotide is then excited with a laser and its fluorescence is detected using a camera built into the instrument.

The lack of an amplification step and size selection in this method constitutes a major advantage since it eliminates the scope for the introduction of amplification mutations and bias against difficult regions as well as reducing the hands-on time required for sequencing.

Because the method requires the presence of a 3' hydroxyl group, it is somewhat biased in favor of shorter fragments, and so it is advisable to fragment the DNA prior to sequencing.

Single-molecule real-time sequencing

Single-molecule real-time developed by Pacific Biosciences was first described in a *Science* paper in 2009 that presented a proof-of-concept for real time sequencing of a single molecule using zero-mode waveguide nanostructure arrays. [68]

The underlying technology is known as single molecule real-time (SMRT) sequencing and relies on the detection of a signal generated when a fluorescently labeled nucleotide is incorporated into a single DNA molecule using a laser situated at the bottom of a zero-mode waveguide well, in conjunction with the phi29 polymerase. During the process of nucleotide incorporation, the molecules are held in a narrow detection period for a comparatively long period of time, making it possible to record the sequence.

The development of techniques such as strobe sequencing and circularization of the DNA template has made this technique suitable for the analysis of structural variation. [69]

In 2010, a paper was published describing the use of this apparatus to analyze DNA methylation without prior modification. [70]

2.5 Service technologies

The rapid development of sophisticated sequencing technologies has created a market for service providers in this field. The use of such providers can increase the time and resources available to the end user. The outsourcing of routine experimental lab work and data analysis will probably become increasingly common in future, leaving the experts to focus exclusively on biological interpretation, saving both time and money.

Complete genomics

Complete Genomics offers several services related to the field of genetics, and launched a DNA nanoball sequencing service in 2010. The first step in this process involves fragmenting the DNA to select for sizes in the range of 400-500bp, and ligating adapters to circulate the DNA. The next step is to amplify the DNA template by rolling circle replication with phi29 DNA polymerase. The resulting circles of DNA are folded into small nanoballs (diameter ~300 nm), which then are repositioned onto a flow cell by adsorption. An oligonucleotide binding site next to the adapter sequence acts as an anchor for pools of fluorescent probes that are then added together with T4 DNA ligase. The probes consist of degenerate nucleotides except at certain positions, where the nucleotide is labeled. This makes it possible to determine which probe bound after the unbound material has been washed out. This cycle is repeated 10 times since the anchor is 10 nucleotides long and is thus completely replaced once the tenth cycle has been completed.

Increasing the density of sequencing arrays allows for an increased number of reads per flow cell, but PCR bias and short read lengths are still a problem when using this approach. [71]

In December 2011, the performance of the Complete Genomics service was compared to that of the Illumina platforms. It was concluded that the former seems to be more accurate based on transition-transversion rate (described in chapter 3.5) but also less sensitive in variant calling. Complete Genomics covers fewer bases, probably due to its shorter read lengths. [72]
Another comparison between Ion Torrent, Pacific Biosciences and Illumina MiSeq was carried out in 2012, where the ability to call variants from each platform was studied and they found that Ion Torrent was slightly better, but with higher false positive rate. [73]
A summary of the sequencing technologies is depicted in Table 1.

Table 1. Overview of sequencing technologies

Platform	Method	Read length	Capacity	Main bias	Advantages	Disadvantages	Application examples
454	Optical detection of fluorescence triggered by pyrophosphate release during base incorporation.	700-800bp	600-700 Mb	Indels	Fast, long read lengths, capacity	Errors in homopolymeric regions	De novo WGS, transcriptome sequencing, targeted resequencing (amplicon sequencing)
Illumina	Optical detection of fluorescence-labeled nucleotides.	2x100bp	600 Gb	Base substitut ions	Large user community, high throughput, low cost	Relatively low multiplex capacity, short reads	WGS, WES, transcriptome sequencing, SNP detection, epigenetic studies
SOLiD	Optical detection of fluorescence-labeled nucleotide octamers	75bp	300Gb	Base substitut ions, A-T bias	Improved accuracy with two base encoding	Short sequences slows down data process, complex analysis	WGS, WES , Chip-Seq in transcriptomics and epigenomics
Ion Torrent	Semiconductor detection of hydrogen ions released during nucleotide incorporation.	200bp	1 Gb	Indels	Fast, future technical development will improve, simple machine	Error rates, long sample preparation	Targeted sequencing
Nanopore	Measurement of conductivity changes when passing molecule through nanopore	>1000bp	-	Single base detection	Fast, low cost	Detection not at single nucleotide level	WGS
Helicos	Single molecule sequencing	2x50bp	35 Gb	Shorter fragment s	No amplification or size selection, short hands on time	High error rate, slow, high cost for increased accuracy	Ancient DNA, FFPE samples, DNA with high GCcontent, transcriptome sequencing
Pacific Biosciences	Zero mode wave guide	>2000bp	1-2 Gb	CG deletions	Real time single molecule sequencing, long reads	High error rate, high cost, need developments	Methylation detection, de novo WGS
Complete Genomics	Sequencing by ligation on balls of rolling circle amplification products	2x25bp	-	PCR bias	High density of DNA molecules on array	Short read lengths, multiple amplification cycles	Resequencing

3

Using genomics for analysis of cancer

Cancer is the second most severe cause of death worldwide, with 7.6 million cases per year. This is expected to rise to 13.1 million deaths per year in 2030 if current population trends persist. The progression of cancer is strongly influenced by the genetics of the patient, with structural variations and mutations being its major causes. Hanahan and Weinberg have suggested six contributing hallmarks of cancer that can be studied; sustained proliferative signaling, evasion of growth suppressors, resistance to cell death, the enablement of replicative immortality, induction of angiogenesis, and the activation of invasion and metastasis. Further research prompted them to add two further hallmarks: the reprogramming of energy metabolism and the evasion of immune destruction. [74]

Personalized medicine

The challenges of targeted treatment are tumor heterogeneity [75] and the difficulty of distinguishing between driver and passenger mutations [76, 77], resistance to drugs [78, 79], and the high costs and regulatory concerns from the Food and Drug Administration (FDA). The latter makes the process delicate and time-consuming, especially when ethical considerations are taken into account. Initiatives in large cancer projects create synergistic effects that allow the integration of expertise. Some such collaborations are listed in Table 2. One issue associated with such large projects is that they generally produce lists of mutated genes. A challenge for the future will be to convert this large body of data into something more useful for clinicians and thus ultimately for patients.

Targeted therapy has fewer side effects for the patient, and has evolved from identifying a few targets with one aberration per gene in large patient cohorts towards massive genomic analyses with many markers in a few samples. We are relying more on *in silico* strategies to limit the experimental "space" in any given case by moving from discovery based on biomarker extraction to biological interpretation to computational analysis. This ultimately facilitates clinical investigations and trials that can rapidly validate a potential treatment. [80]

Table 2. Examples of cancer projects

Name	Collaborators	Number of samples available	Cancer types	Website(s)
The Cancer Genome Atlas (TCGA)	National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI)	6,866	> 20 types of cancer	http://cancergenome.nih.gov/
Cancer Genome Project[81]	Wellcome Trust Sanger Institute	542, 000 tumor samples [82]	Unlimited	http://www.sanger.ac.uk/genetics/CGP/
Total Cancer Care[83]	Moffitt cancer center, 19 sites across the United States	76,434	Unlimited	http://www.insidemoffitt.com/total-cancer-care.htm
International Cancer Genome Consortium (ICGC)	46 projects in 14 countries	3,561 cancer genomes	50 different cancer types	http://icgc.org/icgcm http://dcc.icgc.org/web/
1000 Chicago Cancer Genome Project	Institute for Genomics and Systems Biology (IGSB), University of Chicago	1,000	Breast, bladder, lung, prostate, head, neck, leukemias and lymphomas	http://www.igsb.org/research/cancer/

3.1 Methods for detecting and analyzing mutations in cancer.

In recent years, the process of detecting mutations has been greatly accelerated [84], reflecting underlying technological advances. [85, 86] The need for rational analytical strategies is clear given the increasing availability of patient material and the more rapid acquisition of sequencing data. [87] When searching for genetic biomarkers at the beginning of this era of personalized treatment, it is important to separate extremes by stratifying patients for biomarker discovery [80], to distinguish between genuine markers and false positives. It is also important to bear in mind potential bias in the reference genome that sequences are aligned against, since inter-individual variation can falsely be mistaken for acquired mutations. This can be is avoided if both normal and tumor tissue samples are available.[88, 89]

3.2 Whole genome sequencing of cancers

With the recent advances in sequencing technology and the associated reduction in costs, whole genome sequencing has become an increasingly viable option for a growing number of research labs. [90] Its major advantages for patients are that it can reduce the time required to identify a treatment and that the treatment selected will be chosen based on the patient's genetic makeup [87]. When dealing with disease causing events (mutations, structural variations) in regulatory regions and pseudogenes [91], whole genome sequencing provides a more complete view of the patient's disease than is obtained using methods such as genome wide association studies (GWAS). [92, 93] GWAS has been the method of choice for detecting disease causing variants across the genome. However, it can only be used to study variations in specific genes and thus may overlook potential rare variants, which is a prerequisite for personalized medicine. [94] Exome sequencing has a potential bias in that when searching for biomarkers, one can only obtain results in the genes included in the capture kits. However, low-frequency mutations are more easily detected than in WGS due to the method's more extensive coverage.

Various issues can present challenges for these methods, including low sample volume, sample handling (fresh frozen versus formalin fixed paraffin embedded tissues) and purity, tumor heterogeneity, variation in cancer progression and in the bioinformatic analyses applied. These all have to be considered carefully when selecting a sequencing platform.[41]

3.3 Chemotherapy and prediction of toxicity

Our constitutional genetic composition partially determines our response to chemotherapy. [95] It is becoming more and more evident that to be successful, personalized drug therapies will have to account for both the patient's constitutional genotype and the specific mutations involved in their tumor. The mutations in the tumor are widely accepted to determine which drug will be most effective for the patient. For example, in the case of lung cancer, Erlotinib is the preferred drug for patients with mutations in the EGFR gene [96], while Crizotinib is most beneficial for those with a mutation in the ALK gene. [97] However, most patients are still treated with classical chemotherapeutic drugs. These drugs are associated with adverse drug reactions and high variability in response even for patients with the same histology and grade of disease. Normal inter-individual genetic variability is considered to be largely responsible for these divergent responses. If a patient is likely to respond adversely to a particular drug, it may be necessary to identify an alternative treatment, adjust dosage, or employ some form of pre-medication. Conversely, a patient who is predicted to tolerate the drug can be given a higher dosage and may thus be treated more effectively. The dosages used in classical chemotherapy are based on body surface are and weight [98] and most drugs have a small therapeutic window, and so there is great interest in developing individualized treatments based on genetic markers for toxicity and pharmacokinetics. [99, 100] For example, there have been studies on the toxicity of anticancer drugs such as combinations of carboplatin/paclitaxel [101] and gemcitabine/carboplatin [80, 102-104]. These drugs are either transported into the cells by genes known to be highly polymorphic, or metabolized by such genes, hence interesting targets for studies on individual toxicity.

3.4 Target enrichment

In the absence of sufficient resources to study the entire genome, it is more efficient to focus on the functional parts of the genome to put findings in a more clinically relevant context. This can be achieved either by studying individual genes of interest, or looking at all genes at once (the exome), and thereby reducing the complexity of the analysis. [105] The main benefits of this approach are reduced costs and shorter analysis times, due to the sequencing of fewer bases and the associated reduction in the volume of data generated.

PCR

A major landmark within gene technology occurred in 1983, when Kary Mullis developed the polymerase chain reaction technique, or PCR, which can be used to amplify a particular DNA sequence from minute amounts of starting material. [106] This technique revolutionized molecular biology research at the time, and earned Mullis the Nobel Prize ten years later.

The components needed for the reaction are a DNA template covering the region of interest, primers complementary to the ends of each opposite strand of the template, DNA polymerase, nucleotides, magnesium, and an optimal buffer solution. Fragments of up to 40 kb have been successfully amplified. [107] The design of primers that will capture regions of interest is not always straightforward, but there are numerous online *in silico* PCR tools for testing primer performance. [108, 109] In addition, there are commercial PCR kits that offer good specificity and accuracy in mutation detection.[110] [111, 112]

Whole genome amplification

When dealing with a large number of target regions, PCR becomes much too time consuming and also requires lots of DNA. Under such conditions, whole genome amplification (WGA) is a superior alternative. Various technologies have been developed for this purpose, including primer extension preamplification (PEP) and degenerate oligonucleotide primed PCR (DOP-PCR). [113, 114] However, multiple displacement amplification (MDA) is generally a better choice for mutation analysis due to its higher accuracy, low bias and superior yields. [115] This method relies on the phi29 polymerase enzyme, which has better proofreading activity than the Taq polymerase used in PCR. Other advantages of this enzyme include the fact that it needs very little starting material and generates fragments with lengths of up to 10 kB. The reaction times can be fast (90 min) [116] and no thermal cycling is needed. The use of random hexamers requiring no primer design in combination with strand displacement makes this method appropriate for global analysis over the entire genome. As with all amplification methods, there is the potential for the introduction of mutations, but the introduction of bias can be avoided through careful sample selection. [117, 118]

Sequence capture

The advantage of exome capture is that it produces less data to analyze and the resulting analyses are more rapid. Its major drawbacks are its reliance on expensive hardware and the need for a relatively large quantity of input material required, which is not always possible for cancer samples. [119]

The first commercial capture kit entered the market in 2009, covering the 180000 exons distributed on the 30 million bases that make up the 1% of the genome that encodes proteins. [120] This technology is based on hybridization of DNA probes immobilized on a solid phased array, where unbound fragments are washed away, and the remaining fragment pool is enriched and eluted. Soon after this array based kit was introduced, another company released an in-solution kit that allowed for automated processing. Based on the NCBI Consensus CDS database (CCDS), this kit covered 50 Mb, with additional probes (or “baits”) around the exons and additional human non-coding RNAs and miRNAs.

The Haloplex method uses a different approach, namely the fragmentation of DNA by restriction enzymes in conjunction with biotinylated probes designed to hybridize to both ends of the fragments to form circular DNA strands that are ligated together. [121] The recently released TruSeq Exome Enrichment Kit from Illumina requires only 500 ng of input material. [122]

A recent comparison of the performance of exome capture kits provided by several vendors was published in Genome Biology (2011). It was reported that Nimblegen had a slightly better capture, and that all kits have difficulties in regions with high GC contents. [123] Interestingly, in another comparison between Nimblegen, Agilent and Illumina, Agilent and Illumina were able to detect a greater total number of variants due to Nimblegen’s coverage of fewer genomic regions by Nimblegen. [124]

The company 23andme has been providing raw exome data for less than 1000$ since September 2011, allowing anyone to submit their DNA. [125]

3.5 Data analysis

A natural consequence of the recent developments of sequencing platforms in combination with the dramatic reduction in costs is the production of very large quantities of data – on the order of gigabytes to terabytes. There has been an explosion in the number of analytical tools available to the research community for handling these large amounts of sequencing data, ranging from alignment programs to variant calling programs and programs for functional analysis to facilitate biological interpretation. It is not feasible to summarize them all in this thesis, but a brief overview of some of the tools that are relevant to the studies included in the thesis is provided below.

Mapping and variation calling

It is never straightforward to select an alignment method, and one must generally compromise between accuracy and speed. The alignment of reads against a reference genome is the essential first step in most sequencing experiments, in order to place them in a genomic context for subsequent analyses such as the identification of deviations and abnormalities. [126] The major concern is bias in the mapping and difficult genomic regions, which generates caveats in the final consensus sequence. Several groups have tried to improve the accuracy of this process, and numerous sophisticated tools have been developed. [127] One issue with the large amounts of sequence data produced by genomic sequencing is the requirement of enormous computer power, to the point that traditional Smith-Waterman based approaches such as BLAT and BLAST no longer have sufficient resources to handle the submitted queries. [128-130] Programs such as MOSAIK, Bowtie and BWA solve this problem for example by indexing the reference genome, making the process faster. [131-133]

Tools such as MOSAIK and MAQ look for a probable read match using hash tables, and then MOSAIK extend these hits by using Smith-Waterman algorithms. These tools are generally considered to have higher accuracy [132, 134] than faster tools based on Burrows-Wheeler algorithms [135] including BWA and Bowtie.

There is also a wide range of available SNV callers. Depristo et al [136] at the BROAD Institute have suggested a best practice protocol tailored for use with 1000genomes [137], that relies on the Genome Analysis Toolkit (GATK), currently one of the best tools for this purpose. [138]

For RNA sequencing, Trapnell et al from the Broad institute published a best practice document in Nature Protocols in 2012 based on the program Tophat [139] for alignment of transcripts in combination with Cufflinks [140] for expression level determination. [141] There are many additional programs available for these purposes as well.

The growing community of bioinformatics users benefits from open source sites such as Github (https://github.com), where they can obtain scripts prepared by others, and forums such as seqanswers.com where it is possible to rapidly consult other experts in the field.

Tools with convenient graphical user-friendly interfaces such as the CLC Genomic Workbench [142] and Avadis [143] make the process of sequence mapping and variation calling relatively straightforward for new users.

Quality analysis

There are several ways of assessing quality control for mutation analysis on each specific call at single nucleotide level, such as coverage, phred score and alignment quality. A simple measure of the accuracy of variations throughout a sequenced genome is the transition/transversion rate [117], where transitions are A<->G and C<->T substitutions, and transversions are A<->C, A<->T, C<->G and G<->T. Since there are twice as many possible transversions, the expected ratio is 0.5, however transitions occur more frequently and so the actual ratio in the exome is around three. (The ratio is around two in the genome). [144, 145] The ratio of homozygotes to heterozygotes also provides a quick measure of the data's quality, and is usually around two. [145, 146]

The number of heterozygous calls on the X and Y chromosomes in males provides a rough estimate of the number of false positives, which can arise due to duplications, recombinations in pseudoautosomal regions and sequencing errors. [147]

Functional analysis

High throughput sequencing experiments generate data for several thousand variations. Fortunately, tools have been developed for extracting relevant data from such files. [148] [149]

Numerous online bioinformatic tools are available for downstream genomic analyses, ranging from genome browsers (UCSC [150], Ensemble[151]), to integrated variation databases (Hapmap [152], dbSNP [153], SNPper[154]) to gene ontology and enrichment analysis tools (DAVID [155, 156], Ingenuity [157]). These are extremely useful for end users who need to simplify complex data to facilitate biological interpretation.

Currently, research groups have gone from investigating a limited set of candidate genes to generating lists of several thousand altered genes. The need for rational gene or variant selection strategies is evident [101] and it is increasingly important to consider the biological significance of any given change. A commonly used strategy is to annotate identified variants and select interesting candidates according to their severity or possible damage to phenotype. This is facilitated by resources such as Annovar. [158]

Large collaborative efforts such as the Danish 200 exome project [18] and the 1000 genomes project [17] have contributed tremendously in this area, generating annotations for many rare mutations. Planned projects such as the sequencing of Faroe Islands entire population of 50 000 (Fargen) will postulate a model for integrating whole genome sequencing into healthcare systems in other countries.

3.6 RNA sequencing

The sequencing technologies described above have made it possible to further expand sequence analyses by examining the transcriptome, and there is a growing number of RNA sequencing projects [159, 160]. This shift of focus into the world of transcriptomics, has shed light on RNA editing and gene fusions, and has established pipelines for determining gene expression levels, identifying differentially expressed genes, and detecting splice variants and non coding RNA.

There are two main principles in transcriptome shotgun sequencing; poly-A selection and ribosomal RNA depletion. The workflow in the former involves removal of the abundant ribosomal RNA by capturing mRNA through hybridization of the polyadenylated 3'-end to poly d(T) probes, followed by reverse transcription of the processed transcript to cDNA. The latter principle involves hybridizing the unwanted rRNAs to specific biotinylated probes [161], removing them using streptavidin-coated magnetic beads, and then converting the target RNA to cDNA. A comparison published by Cui et al. in 2010 suggested that the rRNA depletion approach works better for the analysis of additional non coding RNAs. [162] Realizing the need for a more complete analysis, and enable reads of low abundant transcripts, it could be useful to remove the most highly expressed transcripts.

A true RNA sequencing approach has been developed by Helicos, which has developed a system for direct RNA-sequencing without conversion to cDNA, (DRSTM), to avoid amplification and ligation bias. [163, 164]

Splice variants

Alternative splicing refers to the process of combining exons of a gene in more than one way to enable transcription of alternative transcripts, splice variants, that give rise to different isoforms of a protein encoded by a single gene. This creates an enormous diversity of proteins from a limited set of genes. Alternative splicing is a common event in humans, and it has been suggested that approximately 90% of genes are processed in this manner. [165] Together with transcriptional regulation, this mechanism determines tissue specific expression. [166]

The functional properties of the resulting protein are altered by how the RNA is spliced, and studies have shown that splice variants influence cancer. [167, 168] Additionally, mutations in splice sites in tumor suppressor genes that make their encoded proteins non-functional can be an important cause of cancer. [169]

Allele specific expression

Allelic imbalances in gene expression can arise from several factors, including genomic imprinting and common variation. [170] Fluctuations in gene expression due to variations within a gene (*cis*), in regulatory genes (*trans*), or some combination of these two events (*cis* by *trans*) are evenly distributed across the genome. [171] [172]

An associated problem with RNA sequencing is read mapping bias in regions with polymorphic sites, where the reference allele gets

more support and thus introduces bias. This is due to the fact that the reference genome is single stranded. Studies aiming to address this issue are ongoing, using strategies based on polymorphic loci in the reference [173], or improving the statistical framework by calibrating a model based on DNA sequencing. [174]
MMSEQ and AlleleSeq are examples of tools that use variation data to create more accurate references in order to improve mapping and identify allele-specific events.[175, 176]

Digital expression

Gene expression profiling has shifted from DNA microarrays to RNA sequencing [177], with improved accuracy and sensitivity, although challenges such as bias towards highly abundant transcripts has to be considered when performing such studies. [178]
Digital expression can be measured by quantitative counting of transcripts, normalized for exon length, which is commonly discussed in units of Reads Per Kilobase of exon model per Million mapped reads (RPKM). The RPKM value for a given experiment is given by:

$$RPKM = \frac{Total\ exon\ reads}{Mapped\ reads\ (millions) \times Exon\ length\ (KB)}$$

where *Total exon reads* is the number of reads covering an exon, an exon-exon junction, or exon-intron junction of a gene; *exon length* (KB) is the total length of all the exons of a gene, divided by 1000; and *mapped reads* (millions) is the number of reads that have been mapped to a gene, divided by 1 000 000 [179] [180].
As an alternative to RPKM, Fragments Per Kilobase of transcript per Million mapped reads (FPKM) is also widely used. If one of the two reads of a fragment in paired end RNA sequencing is of poor quality, the level of expression measured in RPKM might be skewed since that fragment will be counted once, whereas when using FPKM, both reads must be mapped. FPKM is calculated by counting fragments.

4

Present investigation

The main objective of the research presented in this thesis is to examine the application of recent technologies in genomic sequencing and their use to analyze variation in different cancer types. As recent technologies are readily available, the ability to rapidly screen the genomes of individual cancer patients enables clinicians and physicians to quickly identify appropriate and efficient treatments. The vast amount of data generated during such investigations requires systematic tools that can provide a reliable and high quality overview of the patient's genomic data even when dealing with small and highly heterogenic tumor samples, and the large individual differences between patients.

4.1 The papers

In paper I, we investigated the benefits of using a common amplification method, phi29 amplification, in a large patient cohort, focusing on the frequently mutated p53 tumor suppressor gene. In paper II, we extended this to cover the entire exome, by performing sequence capture using samples from 16 patients. In recognition of the need for useful bioinformatics tools for analyzing large genomic datasets, Paper III presents a successful search strategy for identifying relevant biomarkers for toxicity arising from chemotherapy applied to ovarian cancer patients. Building on this, in paper IV we performed sequence capture on DNA from blood samples originating from 32 lung cancer patients with varying degrees of toxicity, in order to analyze the relationships between genetic markers identified using the bioinformatics approaches described in paper III.

Validation of whole genome amplification for analysis of the p53 tumor suppressor gene in limited amounts of tumor samples

Molecular characterization of tumor biopsies has become increasingly important in the field of personalized cancer medicine, but is complicated by the fact that only small quantities of tumor tissue are usually available.

In paper I, we validated whole genome amplification on scarce genetic material required for downstream genetic analyses. We identified a total of 40 mutations in exons 5-8 in p53 by Sanger sequencing of whole genome amplified tumor tissue in 123 lung cancer patients. Using this result as a reference, we investigated the overlap between unamplified and whole genome amplified pools of the 123 tumor samples, focusing specifically on exon 7 due to its high mutation frequency. Exon 7 was amplified with PCR and cloned into *E. coli*. 80% of all mutations were recovered in the amplified DNA, compared to 65% in the unamplified material, suggesting that mutations are more easily detected when whole genome amplification of DNA was used. To further support our findings, we then simulated (*in silico*) the theoretical coverage over all exons for both unamplified and amplified DNA, and found that whole genome amplified DNA requires less coverage compared to unamplified samples in order to analyze mutations previously identified by Sanger sequencing of the individual samples. In conclusion, whole genome amplification can be used to increase the size of initially small samples, without altering their genetic composition.

Assessment of whole genome amplification for sequence capture and massive parallel sequencing

In paper II, we wanted to investigate the robustness of the findings presented in paper I, to see if they are valid for all genes. We extended the search area to the entire exome by using Sequence Capture on whole genome amplified DNA from 16 lung cancer patients. Tumor and healthy tissue samples were taken from all patients and sequenced using the HiSeq2000 from Illumina. In addition, we had access to RNA from 11 of the 16 patients, and performed RNA-seq, which was used to validate the mutations observed. We found an average overlap of 74% of mutations between unamplified and whole genome amplified DNA from tumor tissue in genes. Several quality checkpoints led us to suggest a strategy for selecting samples to include for genomic analysis. We observed that 89 % of all SNVs identified by sequencing tumor samples following WGA could be confirmed by sequencing unamplified material. WGA appears to contribute to a somewhat higher mutation frequency by introducing artefacts, but with RNA-sequencing of the right sequencing depth these could be corrected for.

Identification of candidate SNPs for drug induced toxicity from differentially expressed genes in associated tissues

With established reliability of the investigated technologies in papers I and II, we applied them in studies with closer biological context. In paper III, we use a meta-analysis search tool for mining large collections of high-throughput genomic datasets to identify candidate genes for predicting paclitaxel/carboplatin-induced myelosuppression and neuropathy.

We searched for expressed genes that were affected by drug exposure and were present in tissues associated with this toxicity. From the resulting list of candidate genes, we selected the ten top-ranked genes and identified 42 non-synonymous single nucleotide polymorphisms (SNPs) *in silico*. As a proof of concept, the selected SNPs were genotyped in 94 ovarian and lung cancer patients treated with carboplatin and paclitaxel. We observed variation in 11 SNPs, of which seven were present at a sufficient frequency for statistical evaluation. Of these seven SNPs, three were present in the genes ABCA1 and ATM, and found to be significantly or borderline associated with either thrombocytopenia or neuropathy.

Using whole exome sequencing to identify genetic candidates for carboplatin and gemcitabine induced toxicities.

Normal inter-individual is considered to be the main factor responsible for differences in the therapeutic effects of cancer chemotherapy and adverse drug reactions between individuals.

In paper IV, we addressed inter-individual variability by studying genetic variants to improve cancer drug therapies. We collected clinical data and DNA from 243 non-small cell lung cancer patients that had all been treated with carboplatin in combination with gemcitabine. The patients were divided into four groups according to their toxicity levels (Common Toxicity Criteria, CTC-grade).

We selected and compared 16 level 0-1 patients to 16 level 3-4 patients, and sequenced the exomes of these 32 'extremes' using an Illumina HiSeq2000.

For each patient, we identified 5000-6800 non-synonymous SNPs and around 100 indels with non-synonmous effects. We then compared the genotypes of these groups to identify potential genetic variants associated with carboplatin- and gemcitabine-induced myelosuppression.

We have selected six bioinformatic strategies for identifying optimal candidates:

1) Fisher's test on wild type vs. variant allele frequency in the two toxicity groups
2) Identification of genetic variants that are not in Hardy-Weinberg equilibrium and for which there is a genetic difference between the groups
3) Distribution among 60 *a priori* candidate genes from literature
4) Meta analysis of gene expression data using Nextbio
5) Pathway analysis using Ingenuity
6) Network analysis using Funcoup

The top 60 genetic variants will be validated using the Sequenom platform. At present, more than 350 lung cancer patients treated with gemcitabine and carboplatin are included in this study.

5

Future perspectives

The catalogue of technologies for investigating the molecular events underpinning cancer extends far beyond the scope of this thesis, but massively parallel sequencing will almost certainly continue to play a key role in cutting edge research in this field for the foreseeable future. Ethical issues could arise from the use of recent technologies, which enable the patient to consider their own data while even experts are unsure as to how it should be interpreted.

Whole genome sequencing

The cost of whole genome sequencing has fallen dramatically, and is thus accessible to most researchers. If no mutations or biomarkers are detected in the exome, the natural next step would be to extend the analysis to whole genome sequencing (before epigenetic studies), and to search for major structural changes such as rearrangements. [181]

As yet, there are no tools for efficiently evaluating all of the information generated from such studies and it is hard for most researchers to keep pace with the developments in sequencing technology. This is expected to become less severe as the outsourcing [182] of bioinformatic analyses becomes increasingly common, reducing the time invested by non-specialists as well as the costs incurred and the volume of tissue required. In addition, community based solutions to common issues can be obtained from online forums, which are rapidly becoming increasingly powerful and useful resources. [183]

Genetic research on cancer

One of the dominant trends in cancer research is the use of recently developed DNA sequencing technologies to explore and analyze cancers. Such studies have produced long lists of 'cancer genes' [82] but have not identified any common hot spot genes [184]. It is clear that cancer is a very genetically complex disease that is affected by both extensive genetic variation and epigenetic factors, as well as the patient's environment. This makes it a difficult disease to study. [185] Thirty years ago, it was expected that the problem of cancer could be solved relatively quickly [186], but we are still far away from a deeper understanding. The identification of mutated cancer causing genes has dominated research since that time. [187]

More recently, the availability of increasingly large bodies of data from numerous studies has shifted the emphasis of research to targeted treatments and personalized drugs. Both the identification of optimal treatments and diagnoses have been facilitated by mutation profiling. [188] The massive output of cancer-related data in recent years has prompted the development of new methods for its integration and analysis on a large scale. [189, 190]

Single cell/molecule analysis

In tandem with the development of more sophisticated techniques for unbiased amplification of genetic material and sequencing technologies that require ever smaller quantities of input material, single cell analyses are becoming increasingly important components of disease progression analyses. For example, the detection of circulating tumor cells from complex samples has been simplified and commercial kits are becoming available to researchers. [191-193]

Global analysis towards personalized medicine

It has become more evident that genetic network interactions play an important role in the interplay between vertical inhibition (along pathways), horizontal inhibition (across pathways) and differential inhibition (synergy effects) in cells. Pathway analyses are therefore becoming part of the routine analysis pipeline in both genetics and proteomics. With collaborations between research groups working with different molecular platforms on the same patient biopsy [194], global analysis of DNA, RNA and proteins will ultimately increase the signal to noise ratio of all platforms. [195] It is more beneficial for the patient if this kind of communication can make relevant data rapidly available to clinicians, ensuring that they efficiently receive the optimal treatment. Recent advances in technology have made it possible to explore a patient's entire genetic map in detail, allowing progression from monotherapies to sophisticated combinatorial therapy. Eleven new cancer drugs have been approved in 2012 alone [196], and in the near future there will be 500 compounds hitting hundreds of targets. (Comment by Michael Pellini, Foundation Medicine, WIN conference, Paris 2012)

Abbreviations

A	Adenine
ATP	Adenosine triphosphate
C	Cytosine
CCD	Charge-coupled device
CCDS	Consensus coding sequence
cDNA	Complementary DNA
ChIP-Seq	Chromatin immunoprecipitation sequencing
CNV	Copy number variation
CTC	Common Toxicity Criteria
ddNTP	Dideoxynucleotidetriphosphate
DNA	Deoxyribonucleic acid
DOP-PCR	Degenerate oligonucleotide primed PCR
DRSTM	Direct RNA-sequencing technologies
FDA	Food and Drug Administration
FFPE	Formalin fixed paraffin embedded
FPKM	Fragments per kilobase of transcript per million mapped reads
G	Guanine
GWAS	Genome wide association studies
KB	Kilobase
MDA	Multiple displacement amplification
MPS	Massively parallel sequencing
mRNA	Messenger RNA
NCLC	Non-small cell lung cancer
ncRNAs	Non coding RNA
NGS	Next generation sequencing
PCR	Polymerase chain reaction
PEP	Primer extension preamplification
PPi	Pyrophosphate
RNA	Ribonucleic acid
RNAi	RNA interference
rRNA	Ribosomal RNA
RPKM	Reads per kilobase of exon model per million mapped reads
SCLC	Small cell lung cancer
siRNA	Small interfering RNA
SMRT	Single molecule realtime
SNP	Single nucleotide polymorphism

SNV	Single nucleotide variation
SOLiD	Sequencing by oligonucleotide ligation and detection
SPRI	Solid phase reversible immobilization
T	Thymine
tRNA	Transfer RNA
U	Uracil
WES	Whole exome sequencing
WGA	Whole genome amplification
WGS	Whole genome sequencing

Acknowledgements

Det här är ju min bok, så jag får skriva vadsomhelst väl? Alla bläddrar ju hit först så lika bra att bli långrandig. Jag tycker att tackandet är viktigast, eftersom jag ALDRIG hade klarat det här annars! Det är för övrigt en klassiker att man gör detta mitt i natten, klockan är nu två…

Joakim, min handledare, vilken resa det har varit! (Ja förutom de faktiska till Paris, Åre, Champoluc, Milano, Madrid, Manchester, Baltimore, Davos och Paris igen då). Tack för att du gav mig en plats i det härliga GeneTechgänget där jag kunde få lov att utvecklas genom utmaningar, uppmaningar och till sist några framgångar. Du inspirerar till utveckling! Min bihandledare **Magnus** som har varit stöttande på alla plan, vad hade jag gjort utan dig? **Henrik G**, vilket tålamod du har och tack för att du visat hur pekskrivande går till, ingen har lärt mig så mycket om tabeller som du. **Afshin** tack för granskning och optimism. **Lars,** ett eller två möten fick vi i alla fall till va? En sann connoisseur inom bioinformatiken, tack för inspiration och ventilation i svåra stunder. Resterande **PI**s inom Skolan för Bioteknologi, tack för att ni får denna arbetsplats att frodas inom vetenskap och glädje! Alla kollegor på **ScilifeLab** och **Albanova**, fortsätt vara glada och öppensinniga, det är ni som gör att man orkar! **Bubbarna Marcus** och **Johan**, vilket intro ni gav mig med massa skratt samt att ni varnade för att doktorera… R frågor och hets, allmän bubbighet och innerumsbandy. Tack Johan också för cancerkoll. **Genetech** bestående av nya och gamla, vilka härliga resor och workshops vi har gjort! Tänk att jag skulle lära mig åka skidor trots allt! Det är skönt att veta, att GT alltid är på plats på jobbet, trots sena kvällar och mer sällan tidiga morgnar så är jag aldrig själv! Tack! **Beata,** hur ska jag nu klara mig när jag slutar? Allt skvaller med en så snäll person som du, jag hade aldrig kunnat så mycket om orkidéer och möbelbeklädning om det inte vore för dig, och du har varit ett STORT stöd! Jag hoppas verkligen att vi trots våra olika personligheter behåller vår vänskap ☺ Får se om jag lyckas vika den där scarfen snart? **Pawel**, tack för våra eviga skvallerstunder, resan till Milano med efterskratt när vi lurade Kimi, pepparkaksmutationsletande en hel jul och nyår, samt att du lärde mig noggrannhet inom protokollskrivande. **Erik Buffert**, tack för alltifrån Pythontips till Tyresögemenskap, anekdoter om katter, Aurora borealis illustration, men framförallt ditt goda humör! Ska

vi cykla nu? **Pat** tack för många skratt med finsk dialekt, cancerkunskap och för att du visade att det är ok att jobba sent (nu är hon 03:03), **Mårten** du har lärt mig labbande, och framförallt nördhumor á la 1337, men även goda råd när jag stod inför ett vägskäl, **Maya** we had opposite diets and I still keep track of you, **Anders** jag kommer alltid jämföra, men nu är jag nog brunare än dig, **Fredrik**, varje gång jag ser en ko tänker jag på dig, när kör vi snowboard nästa gång? Ingen kan bära upp ett par klackskor som du på labbet. **Anna L** tack för trevligt rumssällskap i Davos, och fortsätt med dina bilder på mobilen, de är fantastiska. **Kicki** tack för att du lärde mig sterilteknik och kritiskt tänkande, likaså en fantastisk resa till Davos. Du har en förmåga att se människor, fortsätt med det. **Sverre, Yajing, Sanja, Benjamin, Julle, Linn, et al** keep contributing to science! **GrabbRummet** med bland andra **Jorge, Malin, Erik V** och **BonusDaniel**, vi hade bästa stämningen! **Mikael** tack för hjälp med både statistik, bioinformatik och RNA, snabbare och mer informativ mailrespons får man leta efter, du är en riktig pärla. **Paul**, on your last day you said to me "You shouldn't do science, you should go out there in the world and have fun!" You Romanian grey, popped zit, from the a-hole of the world, our lunches were BY FAR the best I've ever had! It was a privilege making people leave the table with you, it gave me hope of (un) decency in the world ;) **Luisa**, I'm so glad I got to know you, not only for some way too detailed new year's partying, but also help with this thesis and calming of nerves, and yes, GREAT booty conversations. **Bioinformatikerna** för alla frågor inklusive **Uppmax**, **Daniel E** för varierat kollande och massa skriptande, **Bahram** tack för allt sekvenserande och för att du tillhandahåller en avslappnad tillvaro, **Martina** för knasfrågorna på slutet, **Max** för sekvensande och tabellkoll, **Emma** tack för härligt killsnack och 11 timmars promenerande i Paris, introduktion av kaffe med choklad, jag ser fram emot mer av allt ☺, **Viktoria** och **Montana**, mina två grymma specialarbetstjejer! Ingen kan spränga en Pyro som vi en söndagskväll :D **Maria** och **Janne,** tack för grymma resor och konferenser, shoppingturer och lyckosamma projekt. **Mats** för databaskonstruktion och de bästa grundläggande frågorna av alla. Konstigt nog kommer du alltid med positiva resultat? Thanks to the **Chemores consortium** with great characters like the entrepreneur **Vladimir**, **Manfred** with your unlimited knowledge, **Johan** for making it possible, **Ilya** for great teamwork, **Rudi** for breaking chairs and the rest of the crew.

Bänkpressarna: Henrik S, nu är jag äntligen här jag med! Du är en av de mest intelligenta jag känner, för du förvaltar både kunskap och värme i ett, har alltid nära till skratt, respektfull och hjälpsam! Du är en fin vän och det var grymt kul att få labbassa med dig! Vem kommer alltid hungrig med släpande steg kl 11:50 dagligen haha! **Charlotte** vilket äventyr detta har varit! Tack för en fin vänskap med massor av skratt och luncher, doktorerandet hade inte varit detsamma utan dig! **Sebastian,** thank you for a great friendship with Lucia concerts and letting me your toastmaster! Those ghost icons will haunt you forever ;) **Johan** jag kom nästan ikapp din sida av våra inverterade matlåder där ett tag! Bättre renoveringshistorier kommer jag aldrig höra, kan du dra den igen?

Hojgalningar (aka min andra familj) såsom **Krille,** hur många mil har vi promenerat nu? Alltifrån pepp, matlagning till Hackathons, du är en oändlig energigivare! Nu drar vi till Vietnam, Kambodja och Laos!! Du är verkligen en underbar person och jag är SÅ glad att jag lärt känna dig. Du om NÅGON har stöttat mig under skrivandet av denna bok, och när det var som tyngst i deffen, hade ALDRIG klarat det utan dig ska du veta! **Bobo**, min nalle, mitt stöd, alltid lika usel, tack för att du alltid får mig på gott humör med allehanda medel. **Pucken,** du är nog lite av en mentor med dina fantastiska råd som fick mig frisk i hela 7 månader!! USLA ordvitsar som ingen annan förstår, vilken underbar Frankrikeresa det blev! **Speedy** våra allehanda intellektuella diskussioner på varierande nivå är alltid lika uppfriskande, jag tror du kör snabbast i stan, och "studieresa med inspektion av ugnar" kommer jag garva mig harmynt av. **PMC,** TACK för alla underbara avbrott i form av långa hojturer, för att ni tog mig under era vingar, munhuggningar, slaskbrickor, garv som fick mig att glömma bort allt vad stress och avhandlingsskrivande heter, nattliga race, kramar när jag som mest behövde det, klotningar, meck, optimism, bad, bandagar, utveckling: Man ska titta dit man ska! Höj blicken!! Ni är helt underbara, och kan nu briljera med att troligen vara det enda (? - referens saknas) mcgäng som finns omnämnt i en akademisk avhandling, jag sa ju det ;) Ska vi festa nu? <3 Fitnesspinglorna **Dimman, Johan, Nelli, Alexandra, Nina, Alina, Lina Lynx och Stefan J,** mina peppande optimister under TUNGA energilösa timmar på gymmet. Jag gjorde det! Jag tävlade i SM i Bodyfitness och skrev avhandling samtidigt! Jag skulle ALDRIG klarat detta utan er! **Emma** (Stolpen) var ska jag börja? Du kom in i mitt liv när det var som mest turbulent, och har stått ut med mig sen dess? Att jag fann en så fin och underbar person som stöttar mig i ALLT från jobb till fitness till hojpåhitt är för mig ofattbart, snart sitter vi där på verandan vi diskuterade i Turkiet och reflekterar över hur bra vi har det :D Du betyder så mycket för mig <3 **Hannah** min kära vän sen 18 år, du har inspirerat mig under tunga timmar, för jag vet att du har åstadkommit mer, och framförallt gav du mig min underbara gudson **Noah,** vad vi har gått igenom mycket i livet tillsammans! <3 **Åsa** (Ratten) min underbara bästis, som stöttar mig i allt, ord kan inte beskriva vad du betyder för mig. Det skulle kräva en avhandling i sig. Gotland, Göteborg, fitnesstävlingar, hojåkande, skratt gråt, pyssel, halv sex, skk, Uffe, vår vänskap är det bästa jag nånsin haft! Jag är så oerhört stolt över dig och älskar dig så mycket!! <3 <3 <3 **Hilkka**, KIITOS kaikesta, ja että tulit auttamhaan mua kuin mulla oli vaikeutuksia. Nyt sä oot mun äiti, kun mä olin pienenä mä aina luulin että sä olit mun oikea äiti, ja nyt se on voimassa! **Simo**, nyt mä on keksinyt sitä ilopilleria! Eikooki mä oon niin "Söde och Snäll"? **Leena ja Raija**, mun systerit, joka kesä me ottaamme sitä poskikuva, mä toivon että se jatkaa koko meirän elämä! **Martti, Tuula, Pekka ja muut**, ihana että te ootte kuka te oot, älä koskaan muuttuu. mä on AINA tuntenut lämpöösesti tervetuloa teirän kans! Te ootte mun extraperhe! Faster **Lena,** en till Doktor Hasmats nu då? **Farmor** som idag fyller 100 år!

Pappa (Ekan) – Jag skulle ju ALDRIG bli teknolog eller doktor! Du och farfar (som gick ut KTH 1936) introducerade mig till matteproblem vid 5 års ålder och på den vägen bar det. Du har aldrig tvingat in mig till KTH (trots att jag är 4e generationen efter farmors pappa som gick ut KTH i slutet på 1800talet), utan bara stil(l)samt läst Ny teknik och alltid varit överanalytisk. Våra sofistikerade ordvitsar kommer jag alltid bära med mig och jag skryter om dig än, du är både en av de mest intelligenta jag känner till, snällaste och drar de lägsta vitsarna.

Mika (Tårtan), Du har följt mig hela mitt liv både som extraförälder och bästa vän, tror ingen syskonrelation kan komma i närheten av det band vi delar! Finlandsresor, nattliga övningskörningar, moppeskjuts mitt i natten i skogen i Kauhava, besöket på Hawaii. Undrar om du kommer i tid till disputationsfesten?

Biceps & **Triceps**, mina underbara skitmaskiner som jamade sig igenom hela avhandlingsskrivandet så jag blev galen… Tänk vilken glädje två ligister kan skänka <3

References

1. Naidoo, N., et al., *Human genetics and genomics a decade after the release of the draft sequence of the human genome.* Human genomics, 2011. **5**(6): p. 577-622.
2. Mendel, G., *Experiments in Plant Hybridization.* http://www.mendelweb.org/Mendel.html, 1865.
3. Theodor Boveri, W.S., *Chromosome theory of inheritance.* http://www.genomenewsnetwork.org/resources/timeline/1902_Boveri_Sutton.php, 1902.
4. Crick, F., *On protein synthesis.* Symp. Soc. Exp. Biol, 1958. **12**: p. 138–63.
5. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
6. Watson, J.D. and F.H.C. Crick, *Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-738.
7. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
8. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
9. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
10. Pertea, M., *The Human Transcriptome: An Unfinished Story.* Genes, 2012. **3**(3): p. 344-360.
11. Bernstein, B.E., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
12. Mulder, G.J., *Over Proteine en hare Verbindingen en Ontledingsproducten.* Natuur- en scheikundig Archief 6: 87–162., 1838.
13. Nussinov, R., et al., *Allosteric post-translational modification codes.* Trends in biochemical sciences, 2012.
14. Lutz, F.E., *Combinations of Alternative and Blending Inheritance.* Science, 1908. **28**(714): p. 317-8.
15. Hardy, G.H., *Mendelian Proportions in a Mixed Population.* Science, 1908. **28**(706): p. 49-50.
16. Consortium, T.I.H., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

17. Project, G., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
18. Li, Y., et al., *Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants.* Nature genetics, 2010. **42**(11): p. 969-72.
19. Buchanan, C.C., et al., *A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data.* Journal of the American Medical Informatics Association : JAMIA, 2012. **19**(2): p. 289-94.
20. Conrad, D.F., et al., *Variation in genome-wide mutation rates within and between human families.* Nature genetics, 2011. **43**(7): p. 712-4.
21. Lalonde, E., et al., *RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression.* Genome Research, 2011. **21**(4): p. 545-54.
22. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease.* Annual review of medicine, 2010. **61**: p. 437-55.
23. Smith, C.C., et al., *Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia.* Nature, 2012. **485**(7397): p. 260-3.
24. Campbell, P.J., et al., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nature genetics, 2008. **40**(6): p. 722-9.
25. Gulcher, J., *Microsatellite markers for linkage and association studies.* Cold Spring Harbor protocols, 2012. **2012**(4): p. 425-32.
26. Wu, X., et al., *Causal link between microsatellite instability and hMRE11 dysfunction in human cancers.* Molecular cancer research : MCR, 2011. **9**(11): p. 1443-8.
27. Masuda, K., et al., *Relationship between DNA Mismatch Repair Deficiency and Endometrial Cancer.* Molecular biology international, 2011. **2011**: p. 256063.
28. Cicek, M.S., et al., *Quality assessment and correlation of microsatellite instability and immunohistochemical markers among population- and clinic-based colorectal tumors results from the Colon Cancer Family Registry.* The Journal of molecular diagnostics : JMD, 2011. **13**(3): p. 271-81.

29. Pujadas, E. and A.P. Feinberg, *Regulated noise in the epigenetic landscape of development and disease.* Cell, 2012. **148**(6): p. 1123-31.
30. Waddington, C.H., *The epigenotype. 1942.* International journal of epidemiology, 2012. **41**(1): p. 10-3.
31. Chou, A.P., et al., *Identification of Retinol Binding Protein 1 Promoter Hypermethylation in Isocitrate Dehydrogenase 1 and 2 Mutant Gliomas.* Journal of the National Cancer Institute, 2012.
32. Ke, X.S., et al., *Genome-wide profiling of histone h3 lysine 4 and lysine 27 trimethylation reveals an epigenetic signature in prostate carcinogenesis.* PloS one, 2009. **4**(3): p. e4687.
33. Enroth, S., et al., *Cancer associated epigenetic transitions identified by genome-wide histone methylation binding profiles in human colorectal cancer samples and paired normal mucosa.* BMC cancer, 2011. **11**: p. 450.
34. He, L., et al., *Hypervariable allelic expression patterns of the imprinted IGF2 gene in tumor cells.* Oncogene, 1998. **16**(1): p. 113-9.
35. Hansen, K.D., et al., *Increased methylation variation in epigenetic domains across cancer types.* Nature genetics, 2011. **43**(8): p. 768-75.
36. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends in genetics : TIG, 2008. **24**(3): p. 133-41.
37. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(2): p. 560-4.
38. Hoff, K.J., *The effect of sequencing errors on metagenomic gene prediction.* BMC genomics, 2009. **10**: p. 520.
39. Ronaghi, M., M. Uhlen, and P. Nyren, *A sequencing method based on real-time pyrophosphate.* Science, 1998. **281**(5375): p. 363, 365.
40. Nyren, P., *The history of pyrosequencing.* Methods in Molecular Biology, 2007. **373**: p. 1-14.
41. Daniels, M., et al., *Whole genome sequencing for lung cancer.* Journal of thoracic disease, 2012. **4**(2): p. 155-63.
42. Brenner, S., et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nature biotechnology, 2000. **18**(6): p. 630-4.

43. Lander, E.S., *Initial impact of the sequencing of the human genome.* Nature, 2011. **470**(7333): p. 187-97.
44. Mardis, E.R., *Next-generation DNA sequencing methods.* Annual review of genomics and human genetics, 2008. **9**: p. 387-402.
45. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-380.
46. Williams, R., et al., *Amplification of complex gene libraries by emulsion PCR.* Nature methods, 2006. **3**(7): p. 545-50.
47. Liu, L., et al., *Comparison of next-generation sequencing systems.* Journal of biomedicine & biotechnology, 2012. **2012**: p. 251364.
48. Illumina, http://www.illumina.com/.
49. Bennett, S.T., et al., *Toward the $1000 human genome.* Pharmacogenomics, 2005. **6**(4): p. 373-382.
50. Bennett, S., *Solexa Ltd.* Pharmacogenomics, 2004. **5**(4): p. 433-8.
51. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-59.
52. Adessi, C., et al., *Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms.* Nucleic Acids Research, 2000. **28**(20): p. E87.
53. Harismendy, O., et al., *Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing.* Genome Biology, 2011. **12**(12): p. R124.
54. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**(5741): p. 1728-1732.
55. Chan, E.Y., *Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery.* Methods in Molecular Biology, 2009. **578**: p. 95-111.
56. SOLiD, *White paper: Demonstration of increased accuracy with Exact Call Chemistry (ECC).* http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf, 2011.
57. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nature reviews. Genetics, 2009. **10**(10): p. 669-80.

58. Inaki, K., et al., *Transcriptional consequences of genomic structural aberrations in breast cancer.* Genome Research, 2011. **21**(5): p. 676-87.
59. Ruan, X. and Y. Ruan, *Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET).* Methods in Molecular Biology, 2012. **809**: p. 535-62.
60. Braslavsky, I., et al., *Sequence information can be obtained from single DNA molecules.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(7): p. 3960-4.
61. Harris, T.D., et al., *Single-molecule DNA sequencing of a viral genome.* Science, 2008. **320**(5872): p. 106-109.
62. Technologies, L., *Ion Torrent.* http://www.iontorrent.com/.
63. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-52.
64. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature. **475**(7356): p. 348-352.
65. Technologies, O.N., *MinION Press release February 2012.* http://www.nanoporetech.com/news/press-releases/view/39, 2012.
66. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Human Molecular Genetics, 2010. **19**(R2): p. R227-40.
67. DeAngelis, M.M., D.G. Wang, and T.L. Hawkins, *Solid-phase reversible immobilization for the isolation of PCR products.* Nucleic Acids Research, 1995. **23**(22): p. 4742-3.
68. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-138.
69. Ritz, A., A. Bashir, and B.J. Raphael, *Structural variation analysis with strobe reads.* Bioinformatics, 2010. **26**(10): p. 1291-8.
70. Flusberg, B.A., et al., *Direct detection of DNA methylation during single-molecule, real-time sequencing.* Nat Meth, 2010. **7**(6): p. 461-465.
71. Drmanac, R., et al., *Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.* Science, 2010. **327**(5961): p. 78-81.

72. Lam, H.Y., et al., *Performance comparison of whole-genome sequencing platforms.* Nature biotechnology, 2012. **30**(1): p. 78-82.
73. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC genomics, 2012. **13**: p. 341.
74. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
75. Lindberg, J., et al., *Exome Sequencing of Prostate Cancer Supports the Hypothesis of Independent Tumour Origins.* European urology, 2012.
76. Bignell, G.R., et al., *Signatures of mutation and selection in the cancer genome.* Nature, 2010. **463**(7283): p. 893-8.
77. Muller, F.L., et al., *Passenger deletions generate therapeutic vulnerabilities in cancer.* Nature, 2012. **488**(7411): p. 337-42.
78. Diaz, L.A., Jr., et al., *The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers.* Nature, 2012. **486**(7404): p. 537-40.
79. Ding, L., et al., *Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.* Nature, 2012. **481**(7382): p. 506-10.
80. Hasmats J., K.I., Edsgärd D., de Petris L., Lewensohn R., Alexeyenko A., Blackhall F., Besse B., Lindgren A., Sörenson S., Brandén E., Koyi H., Peterson C., Lundeberg J., Gréen H. , *Using whole exome sequencing to identify genetic candidates for carboplatin and gemcitabine induced toxicities.* Manuscript, 2012.
81. Quail, M.A., et al., *A large genome center's improvements to the Illumina sequencing system.* Nature methods, 2008. **5**(12): p. 1005-10.
82. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Research, 2011. **39**(Database issue): p. D945-50.
83. Fenstermacher, D.A., et al., *Implementing personalized medicine in a cancer center.* Cancer journal, 2011. **17**(6): p. 528-36.
84. Majewski, J. and D.S. Rosenblatt, *Exome and whole-genome sequencing for gene discovery: the future is now!* Human mutation, 2012. **33**(4): p. 591-2.

85. Link, D.C., et al., *Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML.* JAMA : the journal of the American Medical Association, 2011. **305**(15): p. 1568-76.
86. Navin, N., et al., *Inferring tumor progression from genomic heterogeneity.* Genome Research, 2010. **20**(1): p. 68-80.
87. Welch, J.S., et al., *Use of whole-genome sequencing to diagnose a cryptic fusion oncogene.* JAMA : the journal of the American Medical Association, 2011. **305**(15): p. 1577-84.
88. Fujimoto, A., et al., *Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators.* Nature genetics, 2012. **44**(7): p. 760-4.
89. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome Research, 2012. **22**(3): p. 568-76.
90. Tabor, H.K., et al., *Genomics really gets personal: how exome and whole genome sequencing challenge the ethical framework of human genetics research.* American journal of medical genetics. Part A, 2011. **155A**(12): p. 2916-24.
91. Kalyana-Sundaram, S., et al., *Expressed pseudogenes in the transcriptional landscape of human cancers.* Cell, 2012. **149**(7): p. 1622-34.
92. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.
93. Zhang, J., et al., *The impact of next-generation sequencing on genomics.* Journal of genetics and genomics = Yi chuan xue bao, 2011. **38**(3): p. 95-109.
94. Jeck, W.R., A.P. Siebold, and N.E. Sharpless, *Review: A Meta-Analysis of GWAS Studies and Age-Associated Diseases.* Aging cell, 2012.
95. Robert, J., et al., *Predicting drug response and toxicity based on gene polymorphisms.* Critical reviews in oncology/hematology, 2005. **54**(3): p. 171-96.
96. Galvani, E., et al., *Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors: current status and future perspective in the development of novel irreversible inhibitors for the treatment of mutant non-small cell lung cancer.* Current pharmaceutical design, 2012.

97. La Madrid, A.M., et al., *Targeting ALK: a promising strategy for the treatment of non-small cell lung cancer, non-Hodgkin's lymphoma, and neuroblastoma.* Targeted oncology, 2012.
98. Du Bois, D. and E.F. Du Bois, *A formula to estimate the approximate surface area if height and weight be known. 1916.* Nutrition, 1989. **5**(5): p. 303-11; discussion 312-3.
99. Garcia-Donas, J., et al., *Single nucleotide polymorphism associations with response and toxic effects in patients with advanced renal-cell carcinoma treated with first-line sunitinib: a multicentre, observational, prospective study.* The lancet oncology, 2011. **12**(12): p. 1143-50.
100. Phan, V.H., et al., *An update on ethnic differences in drug metabolism and toxicity from anti-cancer drugs.* Expert opinion on drug metabolism & toxicology, 2011. **7**(11): p. 1395-410.
101. Hasmats, J., et al., *Identification of candidate SNPs for drug induced toxicity from differentially expressed genes in associated tissues.* Gene, 2012. **506**(1): p. 62-8.
102. Tanaka, M., et al., *Gemcitabine metabolic and transporter gene polymorphisms are associated with drug toxicity and efficacy in patients with locally advanced pancreatic cancer.* Cancer, 2010. **116**(22): p. 5325-35.
103. Chew, H.K., et al., *Phase II studies of gemcitabine and cisplatin in heavily and minimally pretreated metastatic breast cancer.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2009. **27**(13): p. 2163-9.
104. Kiyotani, K., et al., *A genome-wide association study identifies four genetic markers for hematological toxicities in cancer patients receiving gemcitabine therapy.* Pharmacogenetics and genomics, 2012. **22**(4): p. 229-35.
105. Myllykangas, S. and H.P. Ji, *Targeted deep resequencing of the human cancer genome using next-generation technologies.* Biotechnology & genetic engineering reviews, 2010. **27**: p. 135-58.
106. Mullis, K.B., F.A. Faloona, and W. Ray, *[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction*, in *Methods in Enzymology*1987, Academic Press. p. 335-350.
107. Cheng, S., et al., *Effective amplification of long targets from cloned inserts and human genomic DNA.* Proceedings of the

National Academy of Sciences of the United States of America, 1994. **91**(12): p. 5695-9.
108. UCSC, *In silico PCR.* http://genome.csdb.cn/cgi-bin/hgPcr.
109. NCBI, *In silico PCR.* http://www.ncbi.nlm.nih.gov/projects/e-pcr/.
110. BioRad, *QX100 Droplet Digital PCR system.* http://biorad-ads.com/green/LSG-WW-GXD-Droplet-Digital-PCR/?gclid=CJ-lzcPdm7ICFcwtmAodImAANA.
111. Qiagen, *Type-it Mutation Detect PCR Kit.* http://www.qiagen.com/products/type-itmutationdetectpcrkit.aspx - Tabs=t1.
112. Biosystems, A., *Competitive Allele-Specific TaqMan® PCR (castPCR).* http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/castpcr.printable.html.
113. Telenius, H., et al., *Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer.* Genomics, 1992. **13**(3): p. 718-25.
114. Zhang, L., et al., *Whole genome amplification from a single cell: implications for genetic analysis.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(13): p. 5847-51.
115. Pinard, R., et al., *Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.* BMC genomics, 2006. **7**: p. 216.
116. Healthcare, G., *illustra GenomiPhi V2 DNA Amplification Kits Datasheet.* https://http://www.gelifesciences.com/gehcls_images/GELS/Related Content/Files/1314774443672/litdoc28408753AB_20110831103140.pdf.
117. Hasmats J. , G.H., Orear C. , Validire P. , Huss M. , LundebergJ. , *Assessment of whole genome amplification for sequence capture and massive parallel sequencing. Manuscript.* 2012.
118. Hasmats, J., et al., *Validation of whole genome amplification for analysis of the p53 tumor suppressor gene in limited amounts of tumor samples.* Biochemical and biophysical research communications, 2012. **425**(2): p. 379-83.

119. Ku, C.S., et al., *Exome sequencing: dual role as a discovery and diagnostic tool.* Annals of neurology, 2012. **71**(1): p. 5-14.

120. Nimblegen, *SeqCap EZ Human Exome Library v3.0 Datasheet.* http://www.nimblegen.com/products/lit/06593518001.pdf. 2011.

121. Agilent, *Haloplex.* http://www.halogenomics.com/haloplex/how-it-works.

122. Illumina, *Truseq Exome Enrichment Kit Datasheet.* http://www.illumina.com/documents/products/datasheets/datasheet_truseq_exome_enrichment_kit.pdf. 2012.

123. Sulonen, A.M., et al., *Comparison of solution-based exome capture methods for next generation sequencing.* Genome biology, 2011. **12**(9): p. R94.

124. Clark, M.J., et al., *Performance comparison of exome DNA sequencing technologies.* Nature biotechnology, 2011. **29**(10): p. 908-14.

125. 23andme, *Personal exome.* https://http://www.23andme.com/exome/.

126. Horner, D.S., et al., *Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing.* Briefings in Bioinformatics, 2010. **11**(2): p. 181-197.

127. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Briefings in Bioinformatics, 2010. **11**(5): p. 473-83.

128. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

129. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

130. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* Journal of Molecular Biology, 1981. **147**(1): p. 195-197.

131. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.

132. Mosaik, http://bioinformatics.bc.edu/marthlab/Mosaik. (*retreived* 2011-10-06).

133. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
134. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Research, 2008. **18**(11): p. 1851-1858.
135. Burrows, M.W., D.J. , *A block-sorting lossless data compression algorithm. .* Technical report 124, Digital Equipment Corporation, 1994.
136. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-498.
137. Altshuler, D., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-1073.
138. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-303.
139. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
140. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature biotechnology, 2010. **28**(5): p. 511-5.
141. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature protocols, 2012. **7**(3): p. 562-78.
142. *CLC Genomics Workbench* http://www.clcbio.com, 2012.
143. Intelligence, S.S., *Avadis.* http://www.avadis-ngs.com/, 2012.
144. Collins, D.W. and T.H. Jukes, *Rates of transition and transversion in coding sequences since the human-rodent divergence.* Genomics, 1994. **20**(3): p. 386-96.
145. Guo, Y., et al., *Exome sequencing generates high quality data in non-target regions.* BMC genomics, 2012. **13**: p. 194.
146. Bainbridge, M.N., et al., *Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities.* Genome Biology, 2011. **12**(7): p. R68.

147. Pelak, K., et al., *The characterization of twenty sequenced human genomes.* PLoS genetics, 2010. **6**(9).
148. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
149. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
150. Kuhn, R.M., D. Haussler, and W.J. Kent, *The UCSC genome browser and associated tools.* Briefings in Bioinformatics, 2012.
151. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Research, 2012. **40**(Database issue): p. D84-90.
152. *Hapmap.* http://hapmap.ncbi.nlm.nih.gov/.
153. *SNP Database.* http://www.ncbi.nlm.nih.gov/SNP/.
154. Riva, A. and I.S. Kohane, *A SNP-centric database for the investigation of the human genome.* BMC Bioinformatics, 2004. **5**: p. 33.
155. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Research, 2009. **37**(1): p. 1-13.
156. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2009. **4**(1): p. 44-57.
157. *Ingenuity.* http://www.ingenuity.com.
158. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Research, 2010. **38**(16): p. e164.
159. Ha, K.C., et al., *Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines.* BMC medical genomics, 2011. **4**: p. 75.
160. Tariq, M.A., et al., *Whole-transcriptome RNAseq analysis from minute amount of total RNA.* Nucleic Acids Research, 2011. **39**(18): p. e120.
161. Braasch, D.A. and D.R. Corey, *Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA.* Chemistry & biology, 2001. **8**(1): p. 1-7.

162. Cui, P., et al., *A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing.* Genomics, 2010. **96**(5): p. 259-65.

163. Thompson, J.F. and K.E. Steinmann, *Single molecule sequencing with a HeliScope genetic analysis system.* Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.], 2010. **Chapter 7**: p. Unit7 10.

164. Kanamori-Katayama, M., et al., *Unamplified cap analysis of gene expression on a single-molecule sequencer.* Genome Research, 2011. **21**(7): p. 1150-9.

165. Pal, S., R. Gupta, and R.V. Davuluri, *Alternative transcription and alternative splicing in cancer.* Pharmacology & therapeutics, 2012.

166. Noh, S.J., et al., *TISA: tissue-specific alternative splicing in human and mouse genes.* DNA research : an international journal for rapid publication of reports on genes and genomes, 2006. **13**(5): p. 229-43.

167. Wang, L., et al., *A novel DNMT3B subfamily, DeltaDNMT3B, is the predominant form of DNMT3B in non-small cell lung cancer.* International journal of oncology, 2006. **29**(1): p. 201-7.

168. Cao, X., et al., *Upregulation of VEGF-A and CD24 gene expression by the tGLI1 transcription factor contributes to the aggressive behavior of breast cancer cells.* Oncogene, 2012. **31**(1): p. 104-15.

169. Zhang, L., et al., *BRCA1 R71K missense mutation contributes to cancer predisposition by increasing alternative transcript levels.* Breast cancer research and treatment, 2011. **130**(3): p. 1051-6.

170. Palacios, R., et al., *Allele-specific gene expression is widespread across the genome and biological processes.* PloS one, 2009. **4**(1): p. e4150.

171. Main, B.J., et al., *Allele-specific expression assays using Solexa.* BMC genomics, 2009. **10**: p. 422.

172. Lo, H.S., et al., *Allelic variation in gene expression is common in the human genome.* Genome Research, 2003. **13**(8): p. 1855-62.

173. Vijaya Satya, R., N. Zavaljevski, and J. Reifman, *A new strategy to reduce allelic bias in RNA-Seq readmapping.* Nucleic Acids Research, 2012.

174. Skelly, D.A., et al., *A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data.* Genome Research, 2011. **21**(10): p. 1728-37.
175. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Molecular systems biology, 2011. **7**: p. 522.
176. Turro, E., et al., *Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads.* Genome Biology, 2011. **12**(2): p. R13.
177. Fang, Z., J.A. Martin, and Z. Wang, *Statistical methods for identifying differentially expressed genes in RNA-Seq experiments.* Cell & bioscience, 2012. **2**(1): p. 26.
178. Kuznetsov, V.A., G.D. Knott, and R.F. Bonner, *General statistics of stochastic process of gene expression in eukaryotic cells.* Genetics, 2002. **161**(3): p. 1321-32.
179. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
180. *RPKM formula.* http://www.clcbio.com/manual/genomics/Definition_RPKM.html.
181. Ross, J.S. and M. Cronin, *Whole cancer genome sequencing by next-generation methods.* American journal of clinical pathology, 2011. **136**(4): p. 527-39.
182. World, B.-I., *The Value of Outsourcing Bioinformatics.* http://www.bio-itworld.com/BioIT_Article.aspx?id=105789&terms=2007, 2011.
183. Meyer, P., et al., *Industrial methodology for process verification in research (IMPROVER): toward systems biology verification.* Bioinformatics, 2012. **28**(9): p. 1193-201.
184. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.
185. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
186. Clark, R.L., *Cancer 1980: achievements, challenges, and prospects.* Cancer, 1982. **49**(9): p. 1739-45.
187. Futreal, P.A., et al., *A census of human cancer genes.* Nature reviews. Cancer, 2004. **4**(3): p. 177-83.

188. Kohlmann, A., V. Grossmann, and T. Haferlach, *Integration of next-generation sequencing into clinical practice: are we there yet?* Seminars in oncology, 2012. **39**(1): p. 26-36.
189. D'Antonio, M., et al., *Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes.* Nucleic Acids Research, 2012. **40**(Database issue): p. D978-83.
190. Kawaji, H., et al., *Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation.* Nucleic Acids Research, 2011. **39**(Database issue): p. D856-60.
191. Veridex, *Circulating Tumor Cell Seach.* http://www.veridex.com/CellSearch/CellSearchHCP.aspx.
192. Hou, Y., et al., *Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm.* Cell, 2012. **148**(5): p. 873-85.
193. Xu, X., et al., *Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor.* Cell, 2012. **148**(5): p. 886-95.
194. Nijhawan, D., et al., *Cancer vulnerabilities unveiled by genomic loss.* Cell, 2012. **150**(4): p. 842-54.
195. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* The New England journal of medicine, 2012. **366**(10): p. 883-92.
196. *FDA approved drugs.* http://www.centerwatch.com/drug-information/fda-approvals/, 2012.