Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
Nora Rieber, Master of Science
born in Karlsruhe, Germany

Oral examination: November 6th, 2013

Performance comparison of four human whole-genome sequencing technologies

Referees: Prof. Dr. Roland Eils
Prof. Dr. Holger Sültmann

"This new approach to DNA sequencing was I think the best idea I have ever had"

- Frederick Sanger

(In: Sequences, Sequences, and Sequences.
Annu Rev Biochem. 1988; 57:1-28.)

Abstract

After almost 30 years of inertia in the field of sequencing, the emergence of a whole range of so-called "next-generation" sequencing technologies has revolutionized the way we approach genomic and genetic research. Sequencing all 3 gigabases of a human genome, once a costly task of 13 years of international efforts, can now be done within a matter of days with a coverage of 30x and more, and comes with a price tag that is affordable for a middle-sized lab. Among the different next-generation sequencing machines developed over the course of the last 6 to 8 years, four instruments from three different companies have established themselves on the market for human whole-genome sequencing: Illumina's HiSeq2000, Life Technologies' SOLiD 4 and 5500xl SOLiD, and Complete Genomics' technology.

However, these next-generation sequencing platforms are still relatively new, and a comprehensive comparative assessment of their performance is lacking. For this purpose, the DNA of two tumor-normal pairs from medulloblastoma patients was sequenced individually to 30x coverage on each of the four instruments. The resulting data was analyzed with respect to its coverage distribution and biases over the genome, in particular GC bias, and regions without coverage as well as specific genomic regions were assessed. SNP calls on the different sequencing machines were compared, and the benefits of combining read information from different instruments were evaluated. Additionally, somatic mutations were analyzed.

The most striking result is the poor coverage of GC-rich regions by SOLiD 4 and 5500xl SOLiD, discouraging their use in particular for methylation experiments and exome sequencing. In contrast, Complete Genomics seems the least affected by GC content and shows the most comprehensive coverage of many genomic regions, except for short repeats. HiSeq2000 exhibits the most even genome-wide coverage distribution and the least sample-to-sample variation, while consistently achieving the highest sensitivity in SNP calling. A combination of read data from different technologies is shown to entail limited improvement in most cases, and is advisable only for very specific applications. Finally, the comparison of somatic variation confirms that calling somatic alterations is still a big challenge, which is due in particular to low allele frequency. In summary, this comparative study illustrates the assets and drawbacks of each individual machine and can be used as a guide to find the most suitable platform for a specific experimental goal.

Zusammenfassung

In den letzten Jahren hat das Aufkommen von sogenannten "next-generation-sequencing" Hochdurchsatz-Technologien die Analysemöglichkeiten im Bereich der Genomforschung vervielfacht. Die Sequenzierung der drei Gigabasen eines menschlichen Genoms, ein bisher kostspieliger Vorgang, der unter internationalen Bemühungen 13 Jahre in Anspruch nahm, ist nun innerhalb weniger Tage mit mehr als 30-facher Abdeckung möglich, und ist auch für Forschungseinrichtungen mittlerer Größe erschwinglich geworden. Unter den bisher entwickelten next-generation-sequencing Technologien haben sich vier Geräte von drei verschiedenen Firmen für den Einsatz an menschlichen Genomen etabliert: Illuminas HiSeq2000, Life Technologies' Solid 4 und 5500xl SOLiD Geräte, und Complete Genomics' Technologie.

Allerdings sind diese Plattformen noch immer relativ neu, und eine umfassende vergleichende Beurteilung ihrer Leistung fehlt. Zu diesem Zweck wurde die DNA zweier Tumor-Normal Paare von Medulloblastom-Patienten auf jedem der vier Geräte einzeln sequenziert. Die erhaltenen Daten wurden in Bezug auf die Verteilung der genomischen Abdeckung und auf etwaige Verzerrungen, insbesondere in GC-reichen und GC-armen Bereichen, untersucht. Nicht abgedeckte Bereiche und spezifische genomische Regionen wurden ebenfalls begutachtet. Die auf den unterschiedlichen Plattformen bestimmten SNPs wurden mithilfe eines Goldstandards verglichen und die etwaigen Vorteile einer Kombination alignierter "reads" verschiedener Technologien untersucht. Zusätzlich wurden somatische Mutationen analysiert.

Am hervorstechendsten ist die mangelhafte Abdeckung GC-reicher Genombereiche durch SOLiD 4 and 5500xl SOLiD, die stark gegen die Verwendung dieser Plattformen, insbesondere für Methylierungsexperimente und Exom-Sequenzierung spricht. Im Gegensatz dazu scheint Complete Genomics am wenigsten vom GC-Gehalt der Sequenz beeinflusst zu sein, und erreicht die umfassendste Abdeckung in vielen spezifischen genomischen Regionen, mit Ausnahme von kurzen Repeats. HiSeq2000 weist die gleichmäßigste genomweite Abdeckung und die geringste Variation zwischen den Proben auf. Weiterhin erreicht HiSeq2000 stets die höchste Sensitivität bei der SNP-Bestimmung. Eine Kombination der Daten verschiedener Technologien führt in den meisten Fällen nicht zu einer wesentlichen Verbesserung und ist nur für spezifische Anwendungen empfehlenswert. Schließlich bestätigt der Vergleich der somatischen Mutationen, dass das Bestimmen somatischer Variation immer noch eine große Herausforderung darstellt, vor allem aufgrund der niedrigen Allelfrequenzen. Zusammenfassend zeigt diese Studie die Vor- und Nachteile der einzelnen Sequenziergeräte auf und kann als Orientierungshilfe dienen, um die für eine bestimmte experimentelle Fragestellung am besten geeignete Plattform zu bestimmen.

Contents

List of Figures

List of abbreviations

AF	allele frequency
AUC	Area Under Curve
BAM	Binary Alignment/Map
bp	base pairs
BWT	Burrows Wheeler Transform
CE	capillary electrophoresis
CGI	CpG island
chr	chromosome
ddNTP	dideoxynucleotide triphosphate
del	deletion
DNBs	DNA nanoballs
FN	False Negative
FP	False Positive
Gb	Gigabase
HGP	Human Genome Project
ICGC	International Cancer Genome Consortium
IGV	Integrative Genomics Viewer
indel	insertion or deletion
ins	insertion
kb	kilobase
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Mb	Megabase
NGS	next-generation sequencing
PCR	Polymerase Chain Reaction
RAM	Random-Access Memory
RC	Rolling Circle
ROC	Receiver Operating Characteristic
rRNA	ribosomal RNA
scRNA	small cytoplasmic RNA
SINE	Short Interspersed Nuclear Element
snRNA	small nuclear RNA
srpRNA	signal recognition particle RNA
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
TF	Transcription Factor
TN	True Negative
TP	True Positive
tRNA	transfer RNA
VCF	Variant Call Format

1 Introduction

1.1 DNA sequencing

1.1.1 A short history of DNA sequencing

Sixty years after the structure of DNA was determined by James Watson and Francis Crick [1], a discovery which was honored with the Nobel Prize in 1962[1], the methods for the determination of individual DNA sequences have been subject to monumental transformation. The first DNA sequencing methods were published in the late seventies, the "Sanger sequencing" method by Sanger and Coulson [2, 3] quickly replacing the earlier technique by Maxam-Gilbert [4] as the latter used hazardous compounds and had a generally more complex technical setup [5]. The Nobel prize for chemistry was awarded for both methods in 1980[2].

In the 30 years following the publication of the Sanger method, this sequencing technique was the only broadly used protocol. The main characteristic of the method is the use of chain-terminating dideoxynucleotides (ddNTPs). The most recent protocol variant [6, 7] involves fluorescently labeled ddNTPs, and DNA sequence is read-out by capillary electrophoresis (CE). Initially, the DNA fragment to sequence is cloned into a plasmid vector and amplified through its transfection into bacterial cells. The resulting DNA is isolated and then replicated *in vitro* in four different reactions (each assessing the positions of one of the four DNA bases) using a DNA primer and a DNA polymerase. Each reaction takes place in presence of normal deoxynucleoside triphosphates as well as one of the four fluorescently labeled, chain-terminating dideoxynucleotides (either ddATP, ddCTP, ddGTP, or ddTTP). Because these do not contain a 3'-hydroxyl group, the replication is stopped upon their incorporation into the DNA chain. Each of the four replication reactions generates DNA fragment copies of different sizes, each of them terminated by a fluorescent ddNTP. These are then size-sorted by capillary electrophoresis, and the fluorescent patterns are read out via laser detectors, allowing the reconstruction of the underlying DNA sequence.

[1] `http://www.nobelprize.org/nobel_prizes/medicine/laureates/`
[2] `http://www.nobelprize.org/nobel_prizes/chemistry/laureates/`

Capillary electrophoresis was used for the initial sequencing of the full human genome. Launched in 1990, the completion of the Human Genome Project (HGP) took more than a decade, using up a budget of approximately 3 billion US dollars - essentially a dollar per base pair. In early 2001, the first draft human genome sequences were reported by the International Human Genome Sequencing Consortium [8] and by the company Celera Genomics [9]. In 2003, the HGP declared the human genome sequence "essentially complete", covering 99% of the euchromatic genome [10]. Regular updates to the assembly are given by the Genome Reference Consortium[3].

Establishing a human reference sequence had a massive impact on the research community, highlighting that all previous knowledge was extremely limited and that the human genome and its functions were far more complex than assumed [11]. As an example, protein-coding sequence was found to make up only 1-1.5% of the genome, which is in remarkable contrast to previous estimates, while the importance of gene regulation was strongly underestimated, as was the role of non-coding components like small RNAs. The sequence of the human genome enabled far more comprehensive and systematic approaches in the field of genetics and genomics, and its importance to the research community is largely reflected in the number of queries received by the main genome data servers, the European Bioinformatics Institute (EBI) alone recording around 9 million requests per day in 2012 [12].

1.1.2 Next-generation sequencing

The establishment of the human genome sequence was only the start of a major shift in the field. The year 2005 marked the advent of massively parallel, or "next-generation" sequencing. Both the company 454 Life Sciences [14] and the lab of George Church at Harvard Medical School [15] published new protocols which used a decreased reaction volume while allowing an impressive increase in the total number of DNA fragments assessed. In 2006 and 2007, respectively, the companies Illumina and Applied Biosystems (now Life Technologies) introduced novel next-generation sequencing instruments based on sequencing by synthesis and sequencing by ligation, respectively. Their capacity was orders of magnitude greater than that of previous methods, and provided the basis for an extremely sharp drop in sequencing cost (Figure 1.1) and the appointment of next-generation sequencing as "method of the year" by Nature Methods [16], emphasizing its relevance for life science research. Soon afterwards, a number of human whole genomes were sequenced on different platforms [17–20]. A project that had previously needed over ten years to complete with Sanger sequencing could now be completed in a fraction of the time, and was therefore welcomed by the research

[3] `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/`

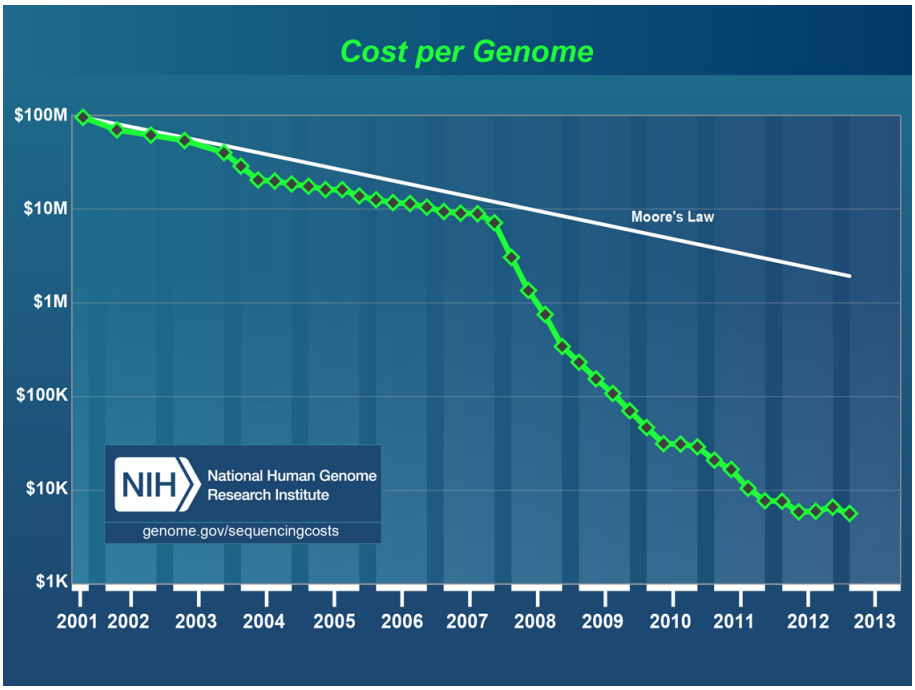

Figure 1.1: The cost of sequencing a human-sized genome from 2001 to 2013, as taken from the National Human Genome Research Institute of the USA (`http://www.genome.gov/sequencingcosts/`) on May 23rd, 2013. Costs for downstream analysis are not included. Moore's law (hypothetical data shown on the graph) reflects the observation that the processing power of computer hardware doubles roughly every two years. It is generally used as a way to assess technology improvements [13]. The sudden drop in sequencing costs and outperformance of Moore's law from January 2008 onwards corresponds to sequencing centers shifting from the Sanger sequencing method to next-generation sequencing platforms.

community as a novel, faster way to shed light onto our understanding of disease and treatment, and the genetic and epigenetic processes that make up the human being.

This cost-effective, faster way to sequence was mainly attained through the automation of a time-limiting and error-prone step of Sanger sequencing: the preparation of cloning libraries. Instead of amplifying the DNA to be sequenced in bacteria *in vivo*, as described above in section 1.1.1, it is randomly fragmented, adaptor-ligated, and then selectively amplified using PCR, with the positive side effect of avoiding cloning bias. Additionally, this enables the sequencing and detection of low-abundance reads and mutations, as the sequencing templates are derived from a single molecule.

The automation and reduction of the reaction volume result in a dramatically higher throughput. As an example, up-to-date whole-genome next-generation sequencing machines like the Illumina HiSeq2000 attain a throughput of roughly 55 Gb per day[4], while "traditional" CE-based Sanger sequencing produces around 1.35 Mb per day [6], over four orders of magnitude less. Currently, four different platforms from three companies are established for human whole-genome sequencing: the HiSeq2000 by Illumina (and the HiSeq2500, which was recently released) [17], Complete Genomics' platform [20], and the SOLiD 4 and 5500xl SOLiD by Life Technologies [18].

Illumina sequencers

Illumina is currently considered as "the most widely used sequencing technology" [21] and has released a number of upgraded platforms since its first one, the Illumina (Solexa) Genome Analyzer, was introduced to the market in 2006. Illumina sequencers use a polymerase-based sequencing-by-synthesis method [17].

The sequencing takes place in a glass flowcell (two for HiSeq2000 and upwards), each divided into eight lanes. Each separate lane can contain different sequencing material if needed. An overview of the steps followed is given in Figure 1.2 and 1.3. The surface of the flowcell is covered with covalently attached oligonucleotides. These are used to fix single-stranded DNA fragments onto the surface via end-ligated adapters. In order to later reach a sufficient signal during the sequencing step, these fragments are then replicated via bridge amplification PCR [17] and form so-called clusters of one and the same DNA fragment. The clusters are then exposed to fluorescently labeled nucleotides which bear a chemically inactivated 3'-hydroxyl group to prevent the incorporation of more than one base at a time. In the next step, the color emitted by each cluster is recorded and assigned to the corresponding nucleotide, whose 3' end is then

[4] `http://www.illumina.com/systems/hiseq_comparison.ilmn`

"unblocked". After this, the full cycle is repeated, and the sequence of each cluster of identical fragments is optically read out by a CCD camera, base by base until the maximum read length is reached. Currently, a read length of 2 x 100 bp can be reached with the Illumina HiSeq2000[5] when in paired-end mode, i.e., when both ends of the fragments are sequenced, and roughly 11 days are needed per run.

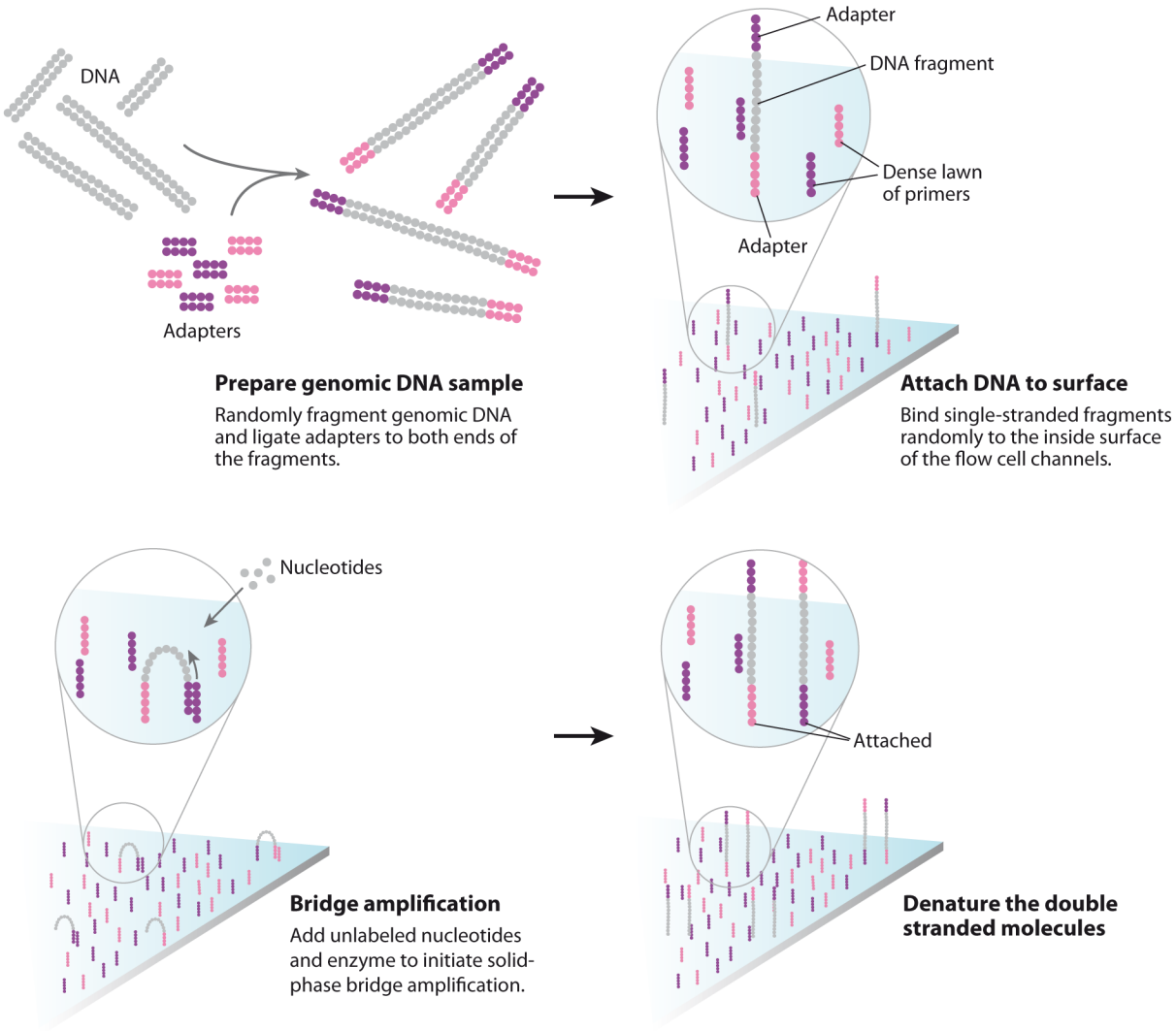

Figure 1.2: Preparation of the DNA clusters to be sequenced on an Illumina flow cell. Individual DNA fragments with ligated adapters are attached onto the surface and replicated into clusters of the same fragment via bridge amplification. Figure taken from Mardis *et al.* [22].

[5]See footnote 4

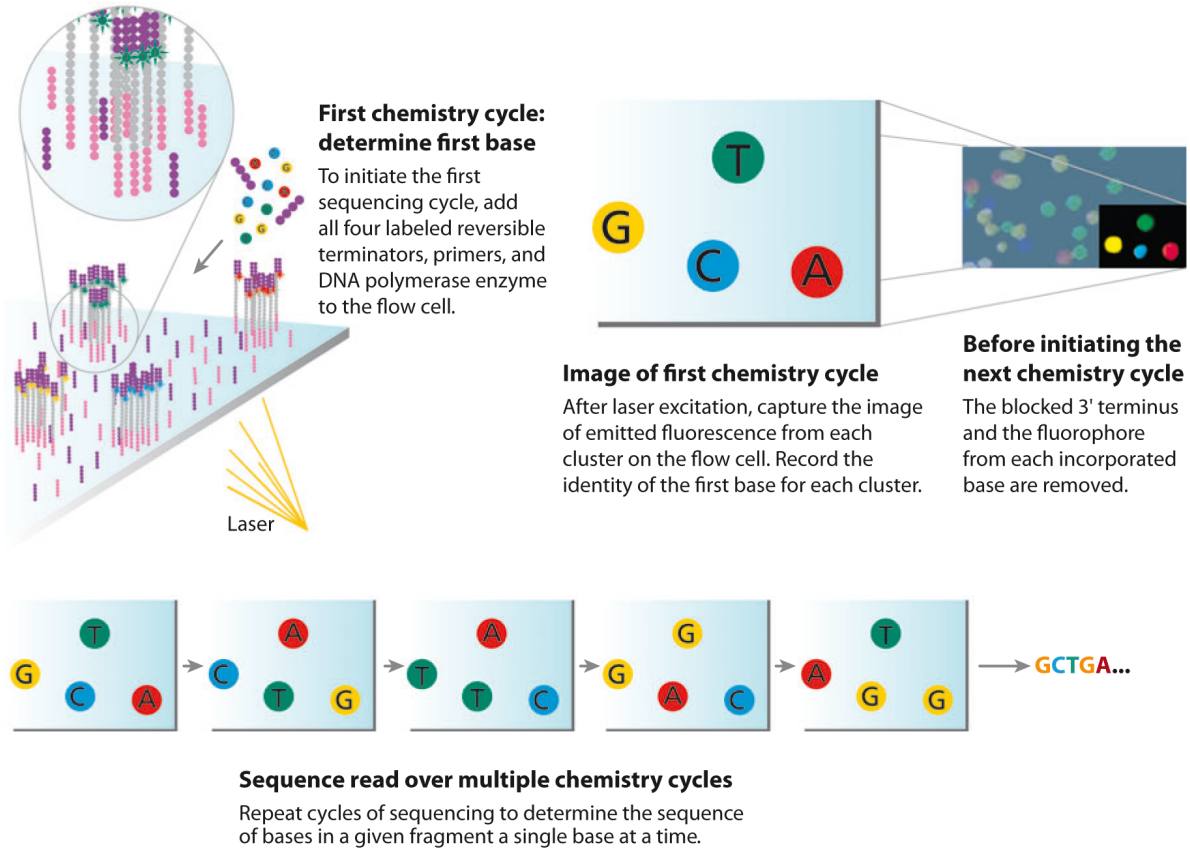

Figure 1.3: Illumina sequencing step. Fluorescently labeled nucleotides with a chemically blocked 3'-OH group are incorporated into the DNA fragment clusters. After read-out of the emitted fluorescence and assignment to the corresponding nucleotide, the 3' end of the incorporated nucleotide is unblocked and the cycle is repeated anew. Figure taken from Mardis *et al.* [22].

Complete Genomics

Complete Genomics Inc. was established in 2006 and published its first human whole-genome sequences in late 2009 [20]. Unlike other companies which offer sequencing instruments for sale, Complete Genomics is a sequencing facility selling sequencing and analysis as a proprietary service. The company focuses on human whole-genome sequencing and has developed both a novel sequencing method and software for the downstream analysis. They offer a Standard Sequencing Service as well as a Cancer Sequencing Service[6] and provide customers with reads and mapping as well as variants like SNVs, indels, and copy number variants.

Complete Genomics' sequencing relies on hybridization and ligation using a protocol allowing for extremely densely packed DNA fragments, and thus needing only very small reagent volumes. An overview of the process is given in Figure 1.4 (A). The fragments first go through a series of adapter insertions using restriction enzymes and are then circularized (Figure 1.4 (B)). The circularized sequencing fragments are around 400 bases long and contain 2 x 35 (10 + 10 + 10 + 5) bases of fragmented mate-pair reads from the original genomic DNA fragment. The reason for these short read fractions is that the sequencing by ligation approach as it is carried out by Complete Genomics does not allow for longer readouts[7].

The circularized templates are subsequently amplified with Φ29 polymerase, which generates hundreds of tandem copies of the fragments, called DNA nanoballs or DNBs (Figure 1.4 (C)), which are less than 200 nm in diameter [23]. These are then placed onto a photolithographically patterned silicon chip with aminosilane active sites placed 1 μm apart. Up to 3 billion DNBs can be placed on one 25 x 75 mm silicon substrate [23].

Each DNB site is then sequenced via combinatorial Probe-Anchor Ligation (cPAL). Using a set of standard and degenerate anchors, and pools of probes with four different fluorescent dye labels (one for each base), each position is read out independently after hybridization and ligation of the corresponding probe (Figure 1.4 (D)). The anchor and probe are washed away after each readout step. The fact that bases are read independently and in no fixed order avoids an accumulation of errors that can occur on other sequencing platforms.

[6] Complete Genomics Technology White Paper, `http://www.completegenomics.com/knowledge-center/whitepapers/`

[7] See footnote 6

Complete Genomics' standard genome coverage is generally higher than that of other platforms (50-80x vs. 30x), but the turnaround time for their service is about 90 to 120 days[8].

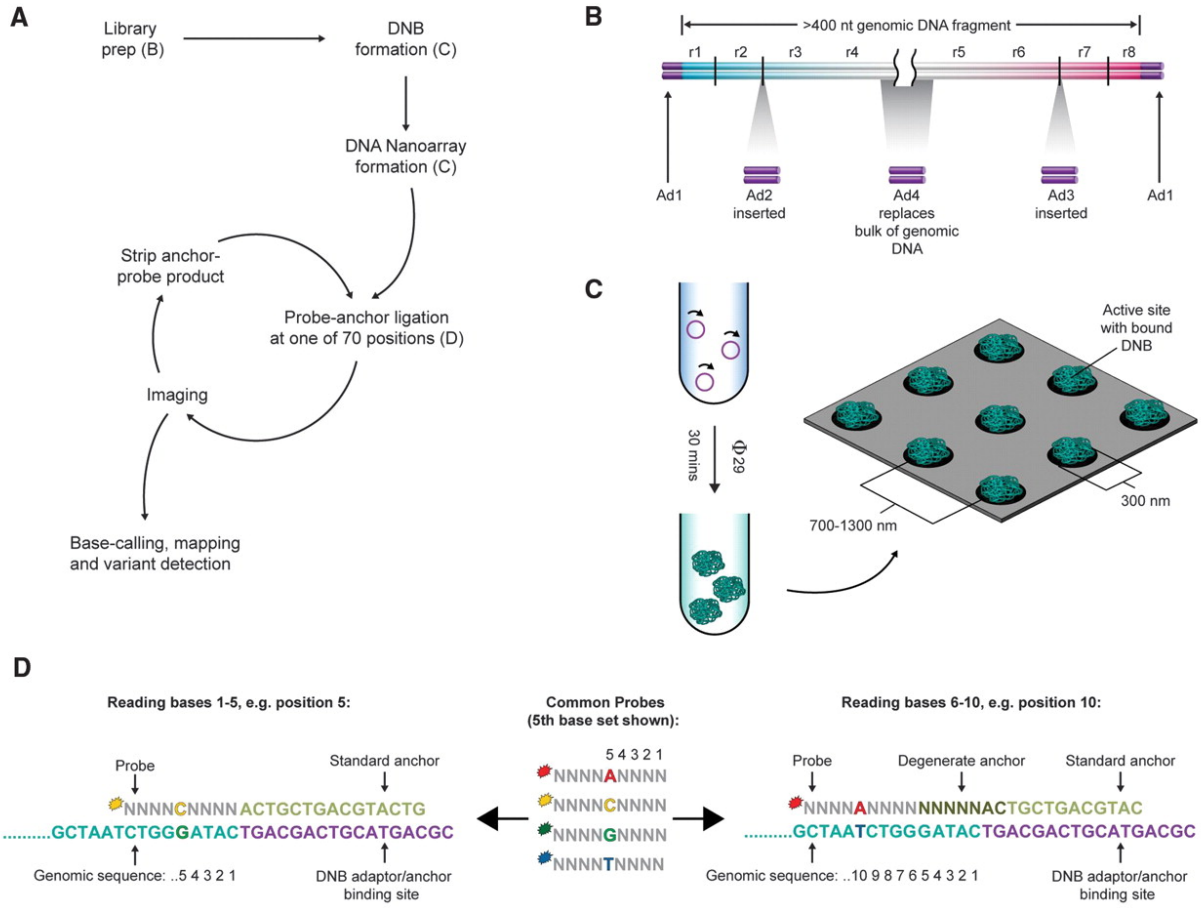

Figure 1.4: Complete Genomics sequencing setup. **(A)** Overview of the sequencing steps. **(B)** Adapter insertions into the genomic fragment to be sequenced. **(C)** Generation of DNA nanoballs via PCR and view of the silicon chips used for sequencing. **(D)** Example of combinatorial Probe-Anchor Ligation sequencing. Figure taken from Drmanac *et al.* [20].

[8] `http://www.completegenomics.com/FAQs/Complete-Genomics-Service-General-Information/`

SOLiD sequencers

First introduced in 2007, the SOLiD acronym stands for Sequencing by Oligo Ligation and Detection [18]. The DNA fragments to be sequenced are coupled to small magnetic beads covered with oligo adapters, and are subsequently amplified via emulsion PCR [24]. The beads are then attached to a glass slide inside a flow cell. For the 5500xl SOLiD platform, beads are replaced with direct amplification on the flow chips[9]. Each SOLiD instrument possesses 2 of these flow chips that can each take up to 8 (SOLiD 4)[10] or 12 (5500xl SOLiD)[11] separate samples per run.

The SOLiD sequencing process consists of multiple sequencing rounds, as exemplified in Figure 1.5. After the annealing of a universal primer to the adapter, fluorescently labeled, semi-degenerate octamers are sequentially ligated to the DNA template. The distinctive feature of SOLiD platforms is their primary output in so-called "color space" (as opposed to base space)[12], an encoded form of the nucleotide sequence where four colors are used to represent 16 combinations of two bases, as shown in Figure 1.6. After the ligation of an octamer, the color emitted is recorded and the fluorophore and the end of the octamer are chemically cleaved to allow for the next ligation cycle. After a defined number of ligation cycles, the complementary strand is removed and a new sequencing round is started using a primer annealed one base further upstream. This means that the positions assessed by the octamers change in each round, as can be seen in Figure 1.5, and the sequencing stops once every base has been probed twice, i.e. is represented by two associated fluorophore colors.

The color-space data can then be decoded given prior knowledge of the leading base, usually the last base of the adapter. This strategy, called two-base encoding, allows to recognize certain sequencing errors, as depicted in Figure 1.5(b). Most importantly, a single point mutation in the DNA fragment sequenced will result in two adjacent color changes, while an error in color space will change the entire decoding of the remainder of the DNA fragment when assessed individually without a reference sequence. This property makes it easier to distinguish actual sequence changes in base space from sequencing errors in color space.

[9] http://tools.invitrogen.com/content/sfs/brochures/CO111759_5500_W_prelim_spec_sheet_FLR.pdf

[10] http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_078637.pdf

[11] http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html

[12] http://marketing.appliedbiosystems.com/images/Product_Microsites/Solid_Knowledge_MS/pdf/CSHL_Fu.pdf

Currently, paired-end read lengths of 50 bp + 35 bp, and 75 bp + 35 bp can be reached with the SOLiD 4 and 5500xl SOLiD platforms, respectively, and run times range from 10-20 days.

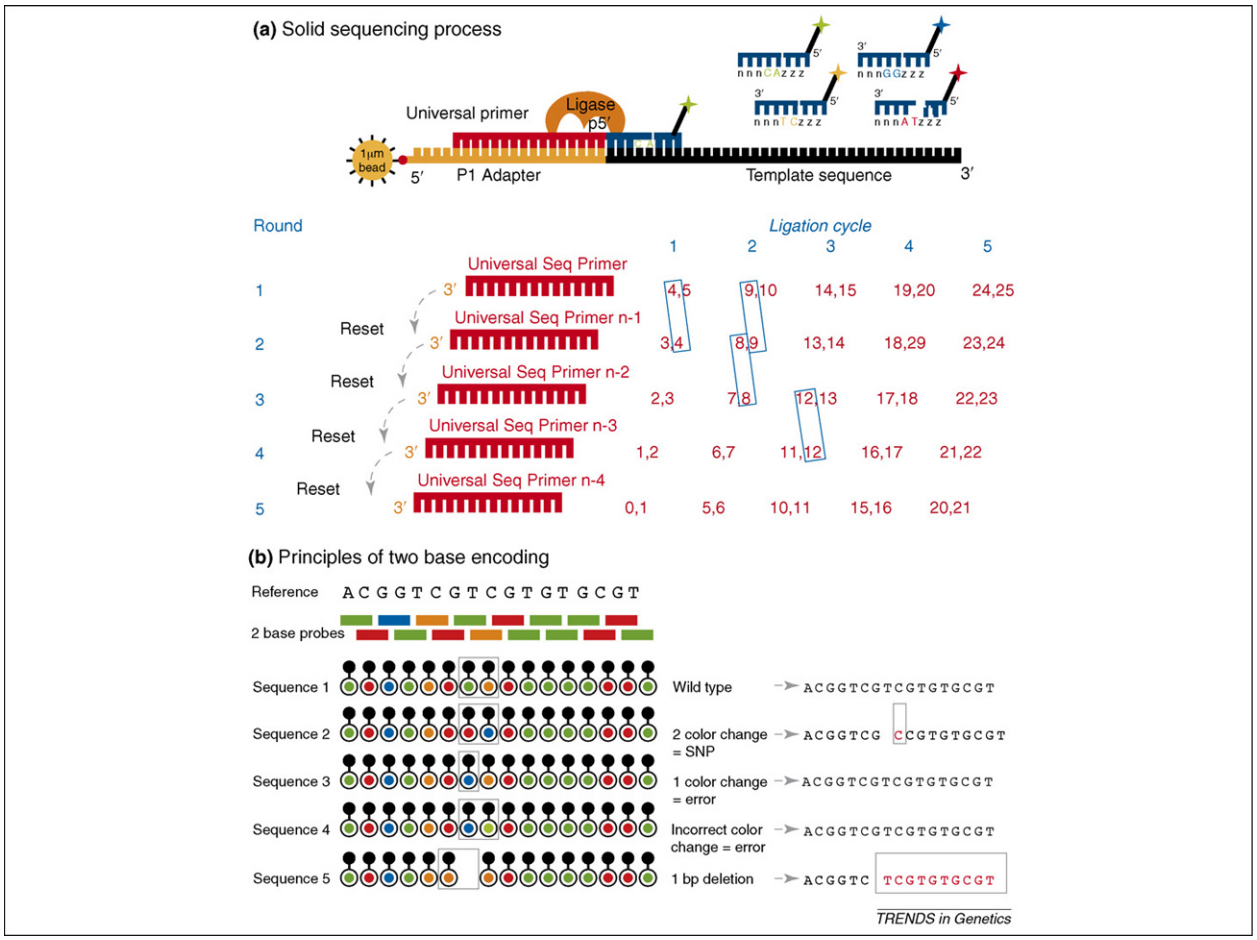

Figure 1.5: The SOLiD sequencing process. **(a)** Each sequencing round starts with the annealing of a universal primer and then the subsequent ligation of the fluorescently labeled octamers matching the underlying DNA sequence. The color of the fluorescent tag stands for a combination of the two non-degenerate bases of the octamer, as illustrated in Figure 1.6. After a defined number of ligation cycles, a new sequencing round is started with a primer placed one base further to the left. This is repeated until every base has been probed twice, as emphasized by the blue rectangles. **(b)** An example of possible color changes within the sequence and their associated meaning. Figure taken from Mardis *et al.* [25].

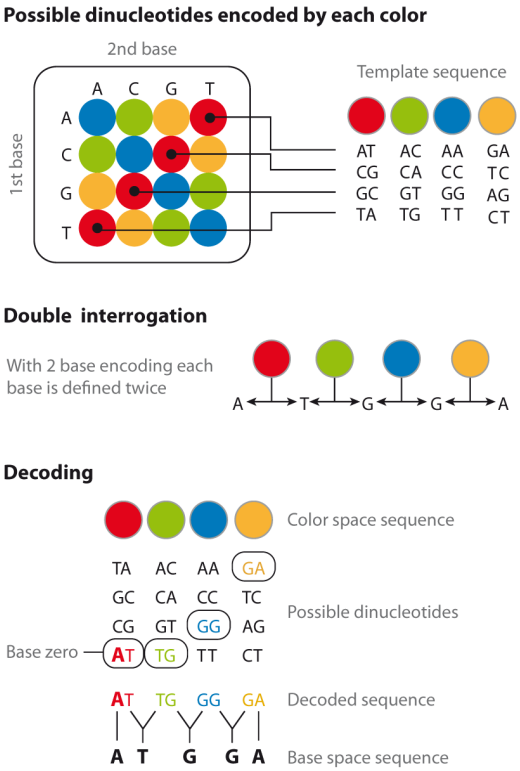

Figure 1.6: Principles of two-base encoding and color space used in SOLiD sequencing platforms. Each color stands for a succession of two bases. As a consequence, each base is interrogated twice. The alignment to a reference sequence allows the conversion back to base space and the identification of sequence variation and sequencing errors. Figure taken from Mardis *et al.* [22].

Known issues

One of the most prominent characteristics of next-generation sequencing is the short read length. While Sanger-based capillary sequencers produce read lengths in the range of 650 to 800 bp, the first Illumina and SOLiD instruments reached a read length of only 35-36 bp [25]. Although the accessible read length is rapidly growing among next-generation platforms, and even further among third-generation sequencing instruments (see section 1.1.2 below), a short read length is an obstacle to downstream analysis. The assembly of a genome from its sequenced reads, as it was performed with Sanger sequencing and the human genome (see section 1.1.1) is a challenging task with short reads, as this increases potential similarity between reads even if they do not belong to overlapping regions of the genome of interest [26]. For organisms with an already available genome sequence, this has led to the general practice of mapping reads onto the reference sequence instead of assembling them, as described in section 1.3.2. Both mapping and assembly processes can be improved by using paired-end reads, i.e. by sequencing both ends of the DNA fragment instead of one end only[13]. The information about the distance between these two reads is stored to decrease ambiguities during mapping or assembly.

Short read length is a problem in particular for DNA repeat regions, which represent a large fraction of many genomes, and especially of the human genome with a repeat content of at least 50% [27]. Repetitive elements take many forms, from just a few to millions of copies, and from just a few bases in length to millions of bases. Long regarded as "junk DNA", it has been found that repeats have a function in human evolution, can influence gene expression [28, 29], and play a role in a number of diseases [30–32]. In addition, short tandem repeats are used in genealogy and forensics as DNA fingerprints [33].

Another issue is known as "phasing noise" or "loss of synchrony". It describes the substitution errors that arise when clonally amplified DNA fragments which are fluorescently probed at the same time - as is the case in the Illumina or SOLiD protocols - get "out of phase". It is assumed that certain sequence patterns trigger loss of synchrony [34]. An additional factor which leads to low base quality at the read end is the fading of fluorescent signal intensity in higher cycle numbers, an issue which is supposedly caused by the limited yield of the elongation reaction [35].

Finally, next-generation sequencing is affected by polymerase-induced biases, leading to a non-uniform representation of the genome. This is especially problematic for

[13] `http://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf`

quantitative sequencing experiments like RNA-seq or copy number estimation, but can also lead to missed or erroneous variant calls.

During library preparation, i.e. the generation of fragments of DNA to be sequenced, a short PCR step is used to select for fragments successfully ligated to adapters [36]. Libraries with too little starting material or with a great variance in starting fragment size will result in a major fraction of duplicated fragments [37]. As long as this fraction is small, it is sufficient to remove reads or read pairs mapping at the exact same position.

The most prominent PCR-induced bias, which also arises during library preparation, is known as GC bias, i.e. the underrepresentation of GC-rich, but also GC-poor fragments, leading to decreased coverage in the corresponding genomic regions [38, 39]. Even though regions with a highly unbalanced base composition represent only a small part of the human genome, GC-rich regions have been shown to correlate with high gene content [40–42] and are frequently part of gene promoters [43].

The base pair error rate of next-generation sequencers is assumed to be higher than the error rate of Sanger sequencing. The Sanger error rates found in the literature range from 0.001% to 1%, depending on the read post-processing software [44]. Illumina error rates are usually found to be between 0.05% and 1%, while the error rates from Life Technologies' instruments are claimed by the company to be around 0.075% [45]. Complete Genomics have never released a base pair error rate[14], but state a variant call error rate of 1 variant per 100 kb [20].

Third-generation sequencing

A new generation of sequencing machines is currently emerging, denoted as "next-next-generation", "third-generation" or "benchtop sequencing". While some authors restrict the term to single-molecule sequencing protocols (as opposed to protocols using PCR-replicated libraries), it is widely used for the platforms introduced after the ultra-parallel whole-genome sequencing machines reviewed in this thesis. Beside the Illumina MiSeq, platforms like the Ion Torrent and Ion Proton [46] or the Pacific Biosciences sequencer [47] introduce innovative protocols that eventually promise longer reads - from 150 bp to over 1kb - and faster turnaround times (between a few hours and a day) at a lower overall cost [48]. These methods are still in their infancy and not yet capable of sequencing a full human genome in a single run, but Life Technologies recently reported "not seeing the end of capacity" for Ion Torrent and

[14] based on personal communication with Complete Genomics

Proton platforms [21].

The Ion sequencers are sequencing-by-synthesis machines based on semiconductor technology and function much like a pH-meter. They detect the proton that is released upon incorporation of a nucleotide into the growing strand, and thus do not require light, scanning and cameras for the detection process, saving a significant amount of time. The throughput has increased from 20-50 Mb/run on the first chips to 1 Gb/run (with a run time of 2-3 hours) on the current 318 chip. However, the platform still uses a "wash-and-scan" technology analogous to the second generation platforms.

The MiSeq, Illumina's new benchtop sequencer, is based on the sequencing-by-synthesis chemistry already in use in the previous Illumina machines. However, it is designed to have shorter run times and less throughput (1 Gb/run with a run time of roughly a day), attained through a smaller flow cell and faster microfluidics [49], and is geared towards clinical diagnostics and smaller laboratories.

Pacific Biosciences takes an entirely different approach, introducing single-molecule real-time (SMRT) sequencing. This is the first approach to observe not a high amount of replicated fragments, but a single DNA molecule as it is synthesized by DNA polymerase. This allows to circumvent many of the biases described earlier on, and even allows sequencing long stretches of short repeats [50]. This setup drastically reduces the sample preparation time and the amount of reagents needed, does not require time-intensive steps like scanning or washing, and allows significantly longer read-outs (over 1 kb). Pacific Biosciences sequencing uses metal chips with nanoscale wells, termed zero-mode waveguides, which contain a DNA polymerase fixed at the bottom. The small volume allows the detection of the incorporation of a single fluorescently labeled nucleotide into the growing DNA strand. Because the fluorescent dye is attached to the phosphate group, it is cleaved upon incorporation, and diffuses out of the well and into the dark quickly. The kinetic data of the polymerase can also be used to detect DNA modifications like methylation [51, 52]. Although the system allows real-time sequencing in a matter of hours, the throughput is currently low at only of 100 Mb/run, and the error rates are extremely high (12-13%), mostly consisting of insertions and deletions [48].

Another company which is developing a sequencing platform based on a single-molecule approach is Oxford Nanopore [53]. However, their device is still in the development phase and is not on the market yet.

1.2 Applications of next-generation sequencing

1.2.1 Experimental applications

The classical use of next-generation methods is the sequencing or re-sequencing of whole genomes, but the high throughput attained has allowed the development of a wide range of other experimental applications, gradually replacing genomic chips and arrays [54]. Beyond the discovery of point mutations and indels [55], whole-genome sequencing has been used to detect copy number variation [56, 57], structural variation [58], or DNA methylation using bisulfite conversion [59] (Figure 1.7). Next-generation sequencing is also opening the field of metagenomics [60] and enables the fast assembly of small genomes like the enterohemorrhagic *E. coli* strain which led to a severe outbreak of foodborne illness in Germany in 2011 [61].

RNA sequencing is used as a means to evaluate gene expression [62–64] (Figure 1.8 (a)) or for the analysis of non-coding small RNAs [65, 66]. A broad range of application also uses next-generation sequencing techniques to target specific parts of the genome, the most widely used being exome sequencing [67, 68] and DNA enrichment procedures to target DNA-protein interactions, chromatin structure or epigenetic marks, like ChiP-seq [69, 70] or MeDIP-seq [71] (Figure 1.8 (b)). The latter are both adaptions of DNA chip protocols to next-generation sequencing. A more complete overview of experimental applications can be found in Shendure *et al.* 2012 [72].

Next-generation sequencing in the clinic

Being able to sequence a human genome in a matter of days, at a non-prohibitive cost, makes next-generation sequencing applications interesting not only for fundamental research, but also in a clinical diagnostics setting. The idea is to use the information gained through whole-genome sequencing of patients for personalized medicine, i.e. for an assessment of expected patient response to different therapy options, and for an assessment of disease risk [74]. However, the requirements to fulfill are radically different for a "clinical-grade" whole genome. Up to now, there is no infrastructure and no standards or guidelines established for this [75], and sequencing data does not always lead to clear conclusions that can be immediately put into action [76], although disease gene panels exist for a small number of genes and for diseases with known genetic causes. In addition, the psychological aspect should not be left out, as whole-genome sequencing may incidentally unveil high risk factors for late-onset diseases with no available treatment, like Alzheimer's or Huntington's disease, or dementia [75, 77]. Also, there is a certain chance of false positives which should not be disregarded as this may have a severe impact on patients lives [78]. Pilot studies for integrating whole-

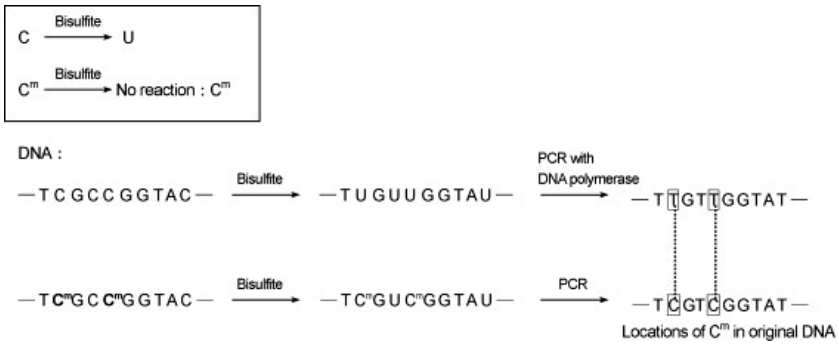

Figure 1.7: Bisulfite sequencing is used to analyze DNA methylation. DNA is subjected to a bisulfite treatment that converts unmethylated cytosines to uracil by deamination, but leaves methylated cytosines intact. In a subsequent PCR step, uracils are then converted to thymines, as they are thymine analogs. Figure taken from [73].

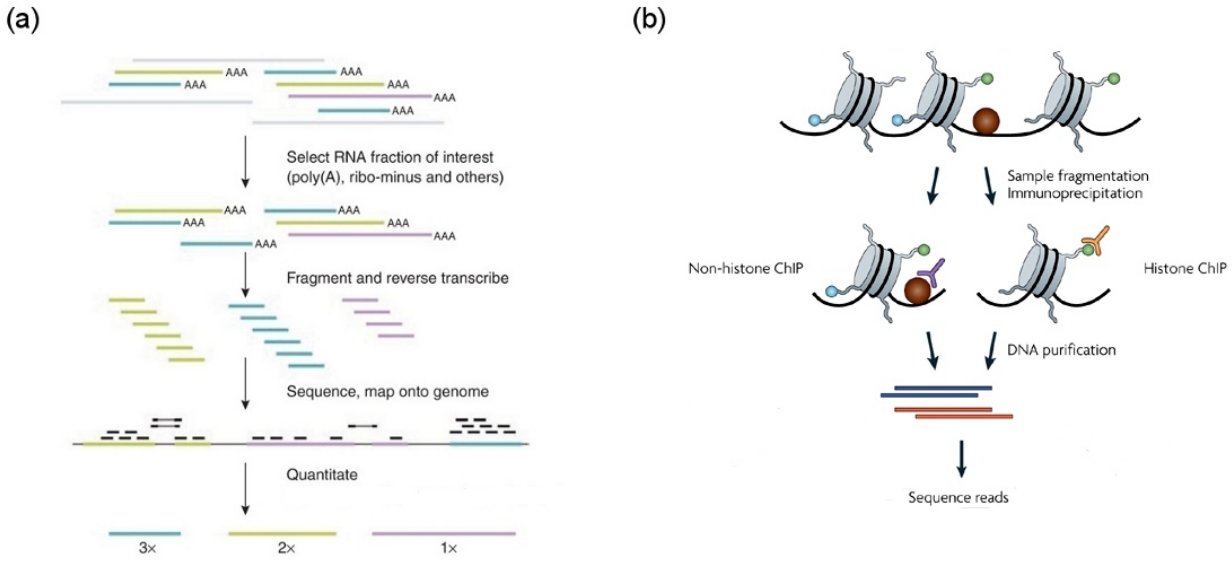

Figure 1.8: **(a)** RNA-seq for the analysis of gene expression. Coding RNA is isolated, usually by targeting its poly-A tail, and is then reverse transcribed to cDNA. The latter is then sequenced on a next-generation platform. After alignment, the sequenced reads can be used to quantitate the RNA content of the sample. Figure taken from [69]. **(b)** Chromatin immunoprecipitation (ChIP) is a method to analyze DNA-protein interactions. It is commonly used to locate the DNA binding sites of transcription factors, histones, or modified histones. The DNA fragments bound to the protein of interest are isolated via immunoprecipitation, and are then sequenced with a next-generation platform. A similar technique can be employed to identify sites of DNA methylation. Figure adapted from [70].

genome sequencing into clinical practice are underway [79, 80], and initiatives like the DKFZ HIPO project[15] promise to build a first scaffold for personalized medicine in oncology.

1.2.2 Large-scale projects

Sharply sinking costs and increased throughput have given rise to a number of prestigious large-scale international genomic projects, e.g., the 1000 Genomes project [55, 81] and the International Cancer Genome Consortium (ICGC) project [82], both launched in 2008. The recently completed Encyclopedia of DNA Elements (ENCODE) project [83, 84], although already planned in 2003, equally profited from the development of next-generation sequencing. All projects are aiming at sequencing a high number of genomes, either to investigate human genetic variation in healthy individuals (1000 genomes project) or cancer patients (ICGC), or to catalog functional elements of the human genome (ENCODE).

The 1000 genomes project aims to sequence entire genomes of a high number of individuals from different populations in order to assess their genetic diversity. While the initial project plan involved 1,000 individuals, this goal was extended to sequence around 2,500 samples, the sequencing results of which will be made available publicly. In order to lower costs, sequencing takes place at a comparatively low coverage of 4x. This is not enough to obtain a complete genotype of each individual, but allows to detect most genetic variants with a population frequency above 1%. These can then in turn be used for studies linking genetic variation and disease.

The International Cancer Genome Consortium is composed of 47 project teams from Asia, Australia, Europe and the Americas, and leads large-scale cancer genome studies of 50 cancer types. For each cancer type, the whole genome, the transcriptome and the epigenome of 500 patients are currently being sequenced to unravel oncogenic variation, which will allow to search for cancer origin and classify cancer subtypes in order to predict clinical outcome and develop better, individually tailored therapies.

The ENCODE project integrated a number of different technologies with the goal to attribute biochemical functions to genomic elements, especially outside the long established protein-coding regions, challenging the widespread assumption that the human genome predominantly consists of so-called "junk DNA". The insights from

[15] `http://www.bio-pro.de/magazin/index.html?lang=en&artikelid=/artikel/09012/index.html`

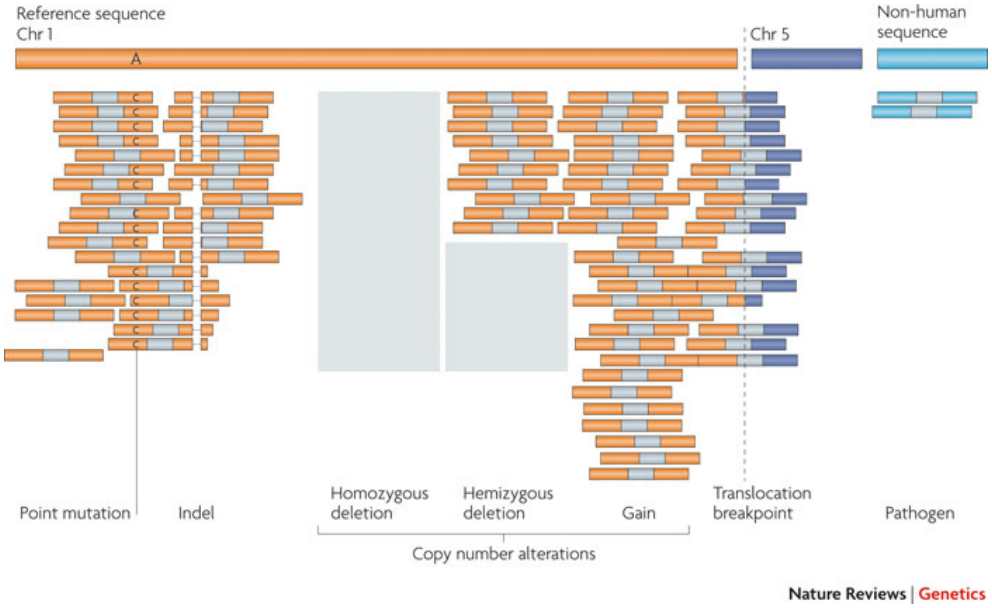

Figure 1.9: Genomic alterations playing a role in cancer, as seen in next-generation sequencing data. Figure taken from Meyerson *et al.* [86].

the ENCODE project were recently published simultaneously in 32 papers[16], the tenor being that over three quarters of the human genome can actually be transcribed. Using, among others, next-generation sequencing techniques like ChIP-seq and RNA-seq (see section 1.2.1), the ENCODE project studied different cell types and investigated the function, expression levels and localization of transcribed RNA, as well as factors influencing transcription, like transcription factor binding, histone modifications, DNA methylation, or general chromatin accessibility.

1.2.3 DNA sequencing in cancer

The availability of next-generation sequencing methods has lead to profound changes across many research fields. The search for the determinants of cancer has particularly benefited from the possibility of sequencing samples from hundreds of cancer patients, as is the case within the ICGC project (see previous section 1.2.2), as cancer is a consequence of often massive genomic alterations [85].

Figure 1.9 illustrates many of the changes that can occur in cancer genomes and the way they translate to whole-genome next-generation sequencing data: from simple point mutations to insertions and deletions, copy number changes, chromosomal translocations, or insertion of non-human, e.g. viral [87], sequence. In addition, epigenetic changes are also known to play an important role in cancer development [88]. These alterations are usually somatic, i.e. acquired in non-germ cells over decades, and

[16] `http://www.encodeproject.org`

are subject to natural selection. A tumor arises when cells carry unrepaired changes conferring a selective advantage, allowing them to proliferate [89]. These changes are termed "driver mutations", as opposed to "passenger mutations" that do not have an impact on cancer development.

Detecting mutations in cancer genomes is a challenge of its own. The search for somatic mutations normally involves the sequencing of a normal control tissue, so that two full genomes need to be sequenced per patient. Also, as cancer cells usually originate from several phases of clonal expansion, their genomes can be highly heterogeneous and may carry different driver mutations in parallel [90]. In addition, cancer tissue usually contains a certain fraction of normal, non-malignant cells, and copy number variation is a rather common event [86]. As a consequence, the allele frequency of somatic mutations is often drastically lower than variant allele frequencies in normal tissue, which makes them harder to detect and requires a high sequencing depth.

The ultimate goal of mutation detection in cancer is to advance diagnostics, prognostics, and treatment. As the treatments themselves are extremely aggressive, knowing driver mutations in advance allows for targeted therapy, meaning that patients likely to benefit from it can be identified upfront, whereas patients likely to be unresponsive will be spared from ineffective therapy. Tailored treatments often target mutated protein products, so that driver mutation detection can be used for drug discovery [91]. The most prominent example is imatinib, an inhibitor of a constitutively active tyrosine kinase in chronic myeloid leukemia, which has greatly improved the treatment of the disease [92].

1.3 Data analysis

1.3.1 "Big data" challenges

For a long time, the limiting factors for whole-genome sequencing have been cost and throughput, and sequencing was therefore only performed by large research centers like the Wellcome Trust Sanger Institute in the UK. At the National Human Genome Research Institute in the USA, a "Billion Basepair Celebration" was held in 1999 to celebrate the billionth base pair sequenced in the Human Genome Project[17], a number which can now be reached in a matter of hours. Meanwhile, the unprecedented throughput and sharp drop in sequencing costs allows even smaller labs to afford next-generation sequencing machines, leading to an unparalleled flood of genomic data.

[17] `http://www.genome.gov/10002105`

Often termed "big data", this phenomenon gives rise to a number of challenges related to data storage, transfer, processing and analysis, and also has implications for patient data privacy.

The output generated by next-generation sequencing machines has for some while outpaced Moore's law (Figure 1.1). While the transistor density on an integrated circuit is growing by a factor of 2 roughly every one to two years [93], for next-generation sequencing data, this has reached a factor of 10 every year since 2002 [94]. The European Bioinformatics Institute (EBI) alone, which runs one of the largest public repositories, stores two petabytes of next-generation sequencing data, with this amount doubling every twelve months [95]. As a consequence, new ways of storing, transferring, and analyzing data are needed. Some answers may lie in the design of smarter algorithms [96] and in the circumvention of data transfer needs, for example through cloud-based solutions containing both databases and analysis software. While the latter are beneficial especially to smaller research groups and institutions by eliminating the need for a computational infrastructure of their own, there may be access and privacy issues [12].

Privacy is a relatively new concern in this respect [97]. While genomic research has been heavily relying on raw, usually anonymized data shared via internet databases [98], some of which are open-access, a number of recent studies highlight that this may not be safe enough to protect the identity of study participants [99]. Particular attention was raised by Gymrek *et al.* for their recent study in which short tandem repeats on the Y chromosome were used in combination with publicly available information from genetic genealogy databases in order to identify sequenced individuals [100].

1.3.2 Analysis steps

The advent of next-generation sequencing required a whole new infrastructure for the downstream analysis of the machine's output, as previous solutions were not suitable any more, being too slow for the available throughput, and not adapted to the shorter read lengths or to new experiment designs (section 1.2.1).

Figure 1.10 shows typical analysis workflows for next-generation sequencing data. The instruments initially produce so-called "reads", i.e. short sequence fragments, and associated base quality values derived from base calling procedures, for example from

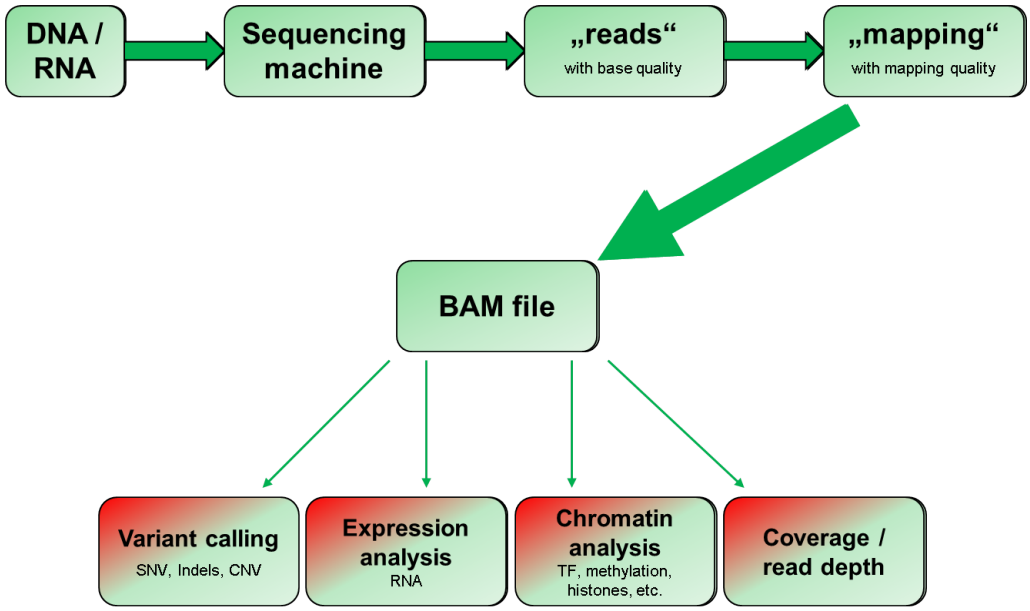

Figure 1.10: Typical analysis workflows for NGS data. DNA or RNA is fragmented, pre-processed and sequenced. The sequencing machine generates so-called reads, i.e. short stretches of read-out sequence. Each base is assigned a base quality. The reads are then aligned to a reference genome (if available) and the resulting information is stored in a standardized BAM file. Coverage is then evaluated to assess the success of the experiment, and the mapped data is analyzed according to the chosen experiment type: the most common being variant calling (SNVs, indels, or copy number variation for example), measurement of gene expression, or analyses of genomic regions previously enriched according to a certain criteria, e.g., CpG methylation, transcriptor factor binding, or histone binding, among others.

fluorescent readouts [101]. The standard format for base quality is Phred scores [102] which are computed as follows:

$$Q_{Phred} = -10 \cdot log_{10} P(error)$$

A base quality score of 20, for example, amounts to an error probability of 10^{-2}. However, a read in itself is almost useless before it is put into context, either by assembly, i.e. using overlaps between reads to reconstruct the genome of interest [26], or by mapping, i.e. aligning reads to a reference, which is the solution of choice when a reference genome is available [103]. Two major classes of mapping algorithms are used, either hash-table based (for example Novoalign[18] or Stampy [104]), or Burrows Wheeler transform (BWT)-based (for example BWA [105] or Bowtie [106]). Alignment methods need to choose a trade-off point between speed and accuracy: hash-table based methods are usually more accurate, while BWT-based methods are faster [101]. One of the currently most popular mapping algorithms for base-space data is BWA, as it is fast, freely available, easy to use, and produces reasonably accurate results [45].

After mapping, reads can be sorted into three different categories: unmapped (an acceptable match was not found in the reference genome), ambiguously or non-uniquely mapped (several equivalent or nearly equivalent matches were found), or uniquely mapped. Mapped reads are assigned a mapping quality based on the concordance between read and reference, and on the underlying base qualities, and this translates to the probability of misplacing the read. Like the base quality, the mapping quality is Phred-scaled. Ambiguously mapped reads are assigned a mapping quality of 0 and are usually not retained for downstream analysis [107]. Mapped reads are usually stored in standardized BAM (Binary Alignment/Map) files, along with alignment, mismatch and quality information [108].

Next-generation sequencing experiments are designed to reach a certain mean coverage or read depth, i.e. a certain redundancy in covering the genome, in order to compensate for sequencing errors and low allele frequency variants. The more evenly the coverage is spread across the genome, the more accurate are the analysis results. A mean genomic coverage of 30x is currently regarded as the standard for variant calling [17, 109]. After alignment, mapped reads are routinely checked for duplicates, i.e. reads with the exact same 5' mapping position. These reads arise from a single read during PCR, and not from different randomly fragmented genomes of multiple cells, as described in the section "Known issues" (section 1.1.2). A high number of duplicates is usually due to a low-complexity library and is to be avoided, as this may introduce unwanted bias and

[18] `http://www.novocraft.com/`

errors. As a consequence, in qualitative sequencing experiments, duplicate reads are removed after mapping [81, 110].

Programming languages and toolboxes used for analysis

The main toolboxes I used for data processing are the command-line based SAMtools [108] and Picard[19] which allow for fast and convenient manipulation of BAM files. All other computations were performed using the statistical programming language R[20], the high-level programming language Perl[21], as well as the Unix scripting language awk and bash shell scripts [111]. The main scripts are listed in the Technical Annex (Section 5) at the end of this thesis.

Statistical measures used

Sensitivity and specificity are common statistical terms used to evaluate the performance of binary classification [112]. Within this thesis, they were used to evaluate platform differences in SNP calls. Sensitivity describes the fraction of true positives – i.e. correctly identified events – among all positive events, i.e. correctly identified and incorrectly rejected events:

$$Sensitivity = \frac{TP}{TP + FN}$$

TP stands for True Positive, FN for False Negative. Specificity describes the fraction of true negatives – i.e. correctly rejected events – among all negative events, i.e. correctly rejected and incorrectly identified events:

$$Specificity = \frac{TN}{TN + FP}$$

TN stands for True Negative, FP for False Positive. Transferred to SNP calling, using a SNP array as a gold standard, sensitivity is the probability of correctly identifying an existing array SNP using the next-generation instrument data. Specificity is the probability of correctly assigning a non-mutated SNP array position using the next-generation instrument data. Specificity is directly related to the false positive rate: the higher the specificity, the smaller the false positive rate, i.e. the less incorrectly identified SNPs we have in the next-generation sequencing data:

$$False\ positive\ rate = 1 - Specificity$$

[19] `http://picard.sourceforge.net/`
[20] `http://www.r-project.org/`
[21] `http://www.perl.org/`

Similarly, the higher the sensitivity, the smaller the false negative rate, i.e. the less SNPs we are missing. In addition, ROC (receiver operating characteristic) curves can be plotted to depict the trade-off between sensitivity and false positive rate under a varying criterion. The higher the area under the curve (AUC), the better the performance.

1.4 Motivation and thesis outline

In the years 2008 and 2009, due to sharply dropping prices, novel next-generation sequencing platforms got adopted by a great number of research laboratories worldwide. Seemingly endless possibilities [6, 113] led to a considerable number of articles being published in scientific journals. However, as with the launch of every new technology[22], next-generation sequencing came with its own share of teething issues. Researchers became aware of new biases involved [36, 38, 114] and their likely impact on next-generation sequencing experiments and analysis results.

Initially, I had been working on data analysis within a project involving the transfer of a methylation assay to the next-generation sequencing methodology. We were confronted with a number of unforeseen issues involving an extremely high fraction of PCR artifacts (sometimes more than 90% duplicate reads), a very uneven distribution of read numbers between libraries, and some contamination problems. While some of the problems arising were obviously assay-based, optimization of the methylation enrichment protocol yielded little overall progress.

This experience sparked the project of benchmarking the different next-generation sequencing platforms available, evaluating the strengths and weaknesses of each of them. To this end, we used two tumor-normal pairs from medulloblastoma samples, sequenced on the four platforms currently commercially available, Illumina's HiSeq2000, Complete Genomics', and Life Technologies' SOLiD 4 and 5500xl SOLiD instruments. I analyzed the data from the four patient samples with regard to the coverage distribution and biases, in particular GC bias, and I investigated and compared regions without coverage as well as the coverage of specific genomic regions. In addition, the potential benefit of a combination of data from different platforms was examined and SNP calls compared to a gold standard across platforms. Finally, somatic SNVs and indels were called and compared.

While there are previous publications comparing next-generation sequencing platforms [115–118], none of them is comprehensive regarding platform choices, and most of them

[22] `http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp`

examine only one sequenced sample. Furthermore, the comparison is usually restricted to a few aspects only, e.g., SNV calling, and does not include a more in-depth analysis of genome-wide coverage distribution. The most recent study by Lam *et al.* [115] analyzes data from a single healthy subject sequenced with HiSeq2000 and Complete Genomics. They show that HiSeq2000 is more sensitive in SNV calling than Complete Genomics, and presume this may be due to HiSeq2000's longer reads which should lead to a better coverage of difficult regions. Additionally, they suggest using both technologies as a way to more comprehensive variant detection, where platform-specific variants are discarded or additionally validated using another technology.

1.4.1 Publication

I recently published a major part of my results in PLoS One together with my colleague Marc Zapatka [119].

2 Methods

Sections 2.1.1 and 2.2.2- 2.2.8 are partly taken from my first-author publication [119].

2.1 Data generation

2.1.1 Whole-genome sequencing

Two tumor/normal pairs obtained from the primary untreated tumor and whole blood of two pediatric medulloblastoma patients (MB14/BL14, female and MB24/BL24, male) were sequenced with Complete Genomics, HiSeq2000, SOLiD 4, and 5500xl SOLiD instruments. Whole-genome sequencing was carried out by the DKFZ Genomics and Proteomics Core Facility, except for Complete Genomics sequencing, which was done by the company itself, using their proprietary solution. Run information for each platform is given in Table 2.1.

All patient material was collected after obtaining written informed consent from participants and an ethical vote approving the study (Institutional Review Board: Ethics Committee of the Medical Faculty of Heidelberg University, Germany / Ethikkommission der Medizinischen Fakultät Heidelberg) according to ICGC guidelines (`www.icgc.org`).

HiSeq2000

High molecular weight genomic DNA was fragmented in a Covaris instrument (Woburn, MA, USA) to an average size of 400 nucleotides.

HiSeq2000 library preparation was performed using standard Illumina protocols and Illumina paired-end adapters. A PhiX kit v2 library (Illumina) was spiked into the libraries at a proportion of about 1% each. The total loading concentration was 7 pM. Amplification was performed in the cBOT (Illumina) using an Illumina TruSeq paired-end v2-cluster generation chemistry. For sequencing, 200 cycle TruSeq-v2-SBS chemistry was used and 2 x 101 cycles of sequencing were performed. Base calling was performed with Illumina RTA v1.10.36 software.

	Complete Genomics	HiSeq2000	SOLiD 4	5500xl SOLiD
DNA input amount	8-16 μg	1 μg	1 μg	1 μg
Read length (bp)	2 x 35 (5+10+10+10)	2 x 100	50+35	75+35
Fragment length (bp)	$\sim$ 400	400	230	230
Throughput	30-90 GB/day	55 GB/day	5-7 GB/day	20-30 GB/day

Table 2.1: Run information for each platform. Throughput information was obtained from the manufacturer's homepage.

SOLiD 4 and 5500xl SOLiD

High molecular weight genomic DNA was fragmented in a Covaris instrument (Woburn, MA, USA) to an average size of 230 nucleotides.

Genomic libraries were prepared following the manufacturer's standard instructions. Emulsion PCRs were performed using SOLiD^{TM}EZ Bead^{TM}Systems. SOLiD 4 sequencing was performed using Life Technologies standard protocols with 50/35 PE chemistry and model caller version MCC 4.04. 5500xl SOLiD sequencing was carried out using 75/35 PE chemistry following the manufacturers standard protocols and MCC 5500 1.0 software.

2.2 Data analysis

2.2.1 Hardware

Computations were performed on a high-performance computer cluster with 49 AMD opteron nodes running under Suse 11.4, each with up to 48 cores. The RAM configuration per node ranges from 16 to 256 GB, and swap space is up to 16 GB, depending on the node.

2.2.2 Read mapping

Sequences were aligned to the human reference genome (NCBI build 37/HG19, released in March 2009), available at `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/`. Due to the heterogeneous nature of the sequencing data, e.g., color space for SOLiD 4 and 5500xl SOLiD platforms, or split read structure for Complete Genomics' platform (see section 1.1.2), for each sequencer the best adapted alignment algorithm was used, following the broad experience gained within DKFZ ICGC projects [110,120,121], as well as personal communication with developers and application specialists from Life Technologies and Complete Genomics. Alignment filters were kept as similar as possible. Only uniquely mapping reads were considered.

For HiSeq2000, the reads were mapped by Natalie Jäger (Computational Oncology Group, Division of Theoretical Bioinformatics, DKFZ) using the Burrows Wheeler Aligner [105] v0.5.9-r16. For Complete Genomics, due to the specific nature of their sequencing data and due to their proprietary analysis algorithms, I relied on the company's alignment, as further methods were not available. For SOLiD 4 and 5500xl SOLiD, reads were aligned using Life Technologies' proprietary Lifescope 2.1 software. SOLiD 4 reads were aligned by the DKFZ Genomics and Proteomics Core Facility, 5500xl SOLiD reads partly by the Core Facility, partly by myself.

Duplicate reads pairs, i.e. read pairs with identical 5' coordinates and orientation, were removed using the Picard software tools v1.61 (`http://picard.sourceforge.net/`).

2.2.3 Coverage and downsampling

I computed the mean base coverage for each sample and platform after duplicate removal for all informative bases of the reference genome (excluding Ns) using a custom python script by Natalie Jäger. For comparison purposes, the BAM files were downsampled by randomly removing read pairs or singletons to reach 30x or 15x mean coverage, using a custom python script by Marc Zapatka (Division of Molecular Genetics, DKFZ).

Complete Genomics mapping files include reads mapped ("initial mapping files") and reads mapped by assembly at candidate regions deviating from the reference ("evidence files"). To allow for a fair comparison of coverage, only the initial mapping files were used when downsampling to 30x. For SNP and SNV comparisons, no downsampling was used, as explained in section 2.2.8.

Unless otherwise mentioned, all results correspond to 30x mean coverage, or for Complete Genomics to full coverage generated (for details see Table 2.2). An evaluation of Complete Genomics at full coverage is included in the analysis because, as explained in section 1.1.2, in place of selling sequencing instruments as is the case with Illumina and Life Technologies, Complete Genomics provides a proprietary sequencing solution with a usually higher coverage (40-80x) including their analysis of the results, e.g., variant calls. It is therefore not possible to purchase a lower coverage.

	Complete Genomics	**HiSeq2000**	**SOLiD 4**	**5500xl SOLiD**
MB14	45.46x	29.87x	30.0x	-
BL14	51.64x	34.06x	30.0x	-
MB24	51.76x	34.48x	30.0x	32.51x
BL24	50.0x	33.29x	30.0x	31.0x

Table 2.2: Average coverage information for each sample and platform assessed

2.2.4 Conversion of Complete Genomics data

Complete Genomics uses proprietary formats for their mapping and results files. I converted the initial mapping files and evidence files to the generic BAM format [108] using shell scripts and the Complete Genomics Analysis Tools (`http://www.completegenomics.com/analysis-tools/cgatools/`) v1.5.0.31, then merged and position-sorted the resulting BAM files with samtools [108]. Duplicates were removed using the Picard tool v1.61, as described in section 2.2.2.

2.2.5 Combination of sequencing data from different technologies

For the combination of data from different technologies, both Marc Zapatka and I merged aligned reads into single BAM files after base quality recalibration with the Genome Analysis ToolKit (GATK) [122] v1.3 (as described in section 2.2.8).

Marc Zapatka then called variants per chromosome using samtools mpileup and bcftools with the following command:

```
samtools mpileup -R -I -A -B -q 1 -Q $Q -r $chrom -ugf $REF $BAM |
bcftools view -vcgNI - | vcfutils.pl varFilter > result.vcf
```

Option `-A` was used to avoid skipping anomalous read pairs during variant calling, option `-B` disables Illumina-specific probabilistic realignment. Option `-I` skips indel calling. `-q 1` skips reads with mapping quality 0, `-Q $Q` sets the minimum base quality for a base to be considered, as described in section 2.2.8. `$REF` stands for the reference genome. bcftools and vcfutils.pl are used with the standard paramenters.

2.2.6 Coverage distribution and regions without coverage

I computed the per-base coverage and the regions without coverage from BAM files using samtools mpileup, a custom perl script, and BEDTools [123] v2.14.3. Only uniquely mapping reads were considered. Reference genome regions composed of undefined bases (Ns) as well as chr Y were not considered in the analysis.

Unless otherwise mentioned, a base was considered not covered if it was supported by less than three reads. The rationale behind this cutoff is that we argue 3 reads are the absolute minimum required to call a heterozygous variant - two reads with a non-reference base (to exclude sequence artifacts affecting only one read) and one with the reference base.

Base coverage in 1 kb windows was computed as the sum of the coverage per base using a custom perl script. GC content in 1 kb windows was computed as the percentage of GC dinucleotides per bin using custom perl scripts.

2.2.7 Functional regions

BED files with the genomic coordinates for CpG islands, CpG island shores, exons, segmental duplications, self chains (downloaded on 09/21/2011), promoters, repeats and mammalian conservation (downloaded on 12/19/2011) were taken from the UCSC Genome Bioinformatics Site (`http://genome.ucsc.edu/`).

CpG island shores were defined as 2 kb upstream and downstream of CpG islands [124]. Promoters were defined as 2 kb upstream and 500 bp downstream from the transcription start site. Intron coordinates were generated from exon coordinates using custom perl and shell scripts and BEDTools `complementBed` and `intersectBed`. BED files for different subcategories of repeats were generated by splitting the UCSC repeats file according to repeat type (DNA repeats, LINE, low complexity repeats, LTR, RC, RNA

repeats, rRNA, satellites, scRNA, simple repeats, SINE, snRNA, srpRNA, tRNA) using a custom R script. The coordinates for the Cancer Gene Census (downloaded on 05/31/2011) and genes from the Cosmic database [125] (downloaded on 11/09/2011) are taken from the Wellcome Trust Sanger Institute (`http://www.sanger.ac.uk/`).

Overlaps of regions without coverage with the genomic regions mentioned above, as well as the size of these overlaps, were computed with custom perl and shell scripts and with BEDTools `intersectBed`, `sortBed`, and `mergeBed`.

2.2.8 SNV calling

To increase base quality accuracy, Marc Zapatka and I performed base recalibration on HiSeq2000, SOLiD 4, and 5500xl SOLiD data, using the `CountCovariates` and `TableRecalibration` functions within the Genome Analysis ToolKit (GATK) [122] v1.3.

SNV calling was then done by Marc Zapatka using samtools v0.1.18 for HiSeq2000, SOLiD 4, and 5500xl SOLiD data. For SOLiD 4 and 5500xl SOLiD data, we were advised by Life Technologies to use samtools instead of their own analysis software LifeScope. Lifescope 2.1 was still used in addition. However, samtools yielded a better concordance with the Affymetrix SNP6 array used as a gold standard, improving the sensitivity for detecting SNPs identified by the array by one percent while gaining 0.02% in false positive rate.

Complete Genomics performed sequencing and data analysis using their proprietary pipeline (Software v2.0.1.5). Because the Complete Genomics Analysis Pipeline is not publicly available, it was not possible to downsample the entire data for direct SNV comparison.

For the validation of calls with an independent technology, the Affymetrix GenomeWide Human SNP Array 6.0 was used, which includes more than 906,600 SNPs. The arrays were hybridized and analyzed by The Centre for Applied Genomics (TCAG) in Toronto, Canada, according to the standard manufacturer protocols, as described in [126]. Briefly, the restriction enzymes NspI and StyI (New England Biolabs, Boston, MA) were used to digest 500ng of genomic DNA, which was then ligated to universal adapters and amplified using PCR. The resulting digested and amplified DNA was purified using polystyrene beads, then fragmented with DNaseI, labeled with biotin, and hybridized to the array. Arrays were then washed on Affymetrix fluidics stations

and scanned with the Gene Chip Scanner 3000 7G. Quality control was done within the Affymetrix Genotyping Console (GTC) using the recommended Affymetrix QC guidelines.

The array genotyping results were used by Marc Zapatka as the gold standard for the generation of receiver operating characteristic (ROC) curves for each of the four NGS platforms, using coverage at the SNP position as the independent variable. Samtools mpileup was used with the following settings, generating vcf files [127] split by chromosome (`$chrom`): `-AE` was used for HiSeq2000 data (using Illumina-specific probabilistic realignment), `-AB` for SOLiD and Complete Genomics data. Several quality cutoffs were tested (`$Q`: 1 and 13) and the cutoff selected that provided the largest AUC for the comparison with the SNP6 array.

For HiSeq2000, additional arguments were:

```
samtools mpileup -R -I -A -E -q 1 -Q $Q -r $chrom -ugf $REF $BAM |
bcftools view -vcgNI - | vcfutils.pl varFilter > result.vcf
```

and for the SOLiD platforms and Complete Genomics the following command was used:

```
samtools mpileup -R -I -A -B -q 1 -Q $Q -r $chrom -ugf $REF $BAM |
bcftools view -vcgNI - | vcfutils.pl varFilter > result.vcf
```

2.2.9 Detection of somatic SNVs

Complete Genomics

Complete Genomics already provides a comprehensive list of variation with every cancer genome they sequence, which includes somatic variations, called with a proprietary algorithm adapted to the structure of their data. For reasons stated above (see sections 2.2.2, 2.2.3), and after having consulted company experts, I relied on their somatic calls.

Somatic SNV calls were extracted from Complete Genomics' `somaticVcfBeta` file, which contains the full list of genomic variation, using quality criteria fixed according to Complete Genomics' Cancer Pipeline user manual (`http://media.completegenomics.com/documents/DataFileFormats+Cancer+Pipeline+2.0.pdf`) and with the assistance of the company's application specialists. I wrote a custom perl script retaining only the variants matching the following criteria:

- SNVs with somatic status `SS=Somatic`
- no `VQLOW` flag (stands for low call quality)
- a somatic score `CGA_SOMS` > -10 (stands for a high-confidence somatic variant)

The somatic SNVs chosen accordingly were converted to the generic vcf format with a custom perl script, and then annotated with the genomic categories described in section 2.2.7 using BEDTools `intersectBed` and a custom perl script.

HiSeq2000

For somatic SNV calling on HiSeq2000 data, I adapted an in-house calling pipeline written by Natalie Jäger, Matthias Schlesner and Barbara Hutter (Computational Oncology group, Division of Theoretical Bioinformatics, DKFZ) to the data. The pipeline uses publicly available software tools in combination with custom scripts and filtering steps.

SNVs were first called in the tumor samples using samtools mpileup and bcftools [108], taking into account only high quality reads and bases (minimum mapping quality: 30; minimum base quality: 13). Tumor samples often contain variants with a very small allele frequency due to contamination with normal tissue, copy number variation, and tumor heterogeneity [110]. To avoid missing these variants, bcftools parameters in the pipeline were adjusted so that one high quality non-reference base suffices for reporting a variant (parameter `-p 2`).

The ensuing high number of false positive SNV calls was corrected using different filters. The pipeline excluded positions covered by less than three reads in both tumor and control, with a somatic allele frequency below 5%, or with only one read containing the variant. Additionally, positions with strand bias, i.e. with reads supporting the variant found on one strand only, were screened, as this usually indicates a sequencing error. Bases in the immediate vicinity (+/- 10 bases) of these calls were checked for Illumina-specific error profiles [34] and the SNVs excluded if a match was found.

A pileup was then generated in the corresponding normal samples at the positions of the remaining SNV calls. These were divided into germline and somatic events according to the pileup information. An SNV was categorized as somatic if at most one read supporting the mutation per 30 reads was found in the corresponding normal sample.

Somatic SNVs were then annotated with the genomic categories described in section 2.2.7 using BEDTools `intersectBed` and a custom perl script.

SOLiD 4 and 5500xl SOLiD

SNVs were called in the tumor samples using both LifeScope 2.1., Life Technologies' proprietary analysis software, and samtools with the parameters described above for HiSeq2000. However, as the calling parameters were extremely lenient, even after filtering, samtools yielded a number of somatic SNVs so high (8 to 18 times more than for the corresponding HiSeq2000 samples) that I chose to proceed with LifeScope calls only and slightly more stringent quality parameters in order to take advantage of dibase encoding and to avoid an increased number of false positives.

LifeScope SNV calling parameters were fixed to a minimum mapping quality of 30 and a minimum base quality of 26. Since SOLiD 4 and 5500xl SOLiD have many regions with low coverage, a minimum coverage of 10 was required for SNV calls. Additionally, LifeScope computes a p-value for each SNV called. SNVs with a p-value higher than 0.05 were excluded. Finally, SNVs with an allele frequency below 5%, with only one read containing the variant, or with reads supporting the variant found on one strand only, were excluded. The resulting files were converted to the generic vcf format with a custom perl script.

As described above for HiSeq2000, a pileup of the normal sample was then used to identify somatic SNVs, and the somatic SNVs were then annotated with the genomic categories described in section 2.2.7 using BEDTools `intersectBed` and a custom perl script.

Validation

For each set of somatic SNVs called, I additionally removed all those called as germline in any of the technologies. Since resources for validation were limited, I chose the 13 somatic SNVs called by all four technologies, as well as, for each platform, 10 to 15 high-quality somatic SNVs that were called for this platform only. The chosen somatic SNVs were externally validated by Sanger sequencing.

2.2.10 Detection of somatic indels

Indel calling for HiSeq2000 was performed by Qi Wang (Computational Oncology Group, Division of Theoretical Bioinformatics, DKFZ). Indel calling for SOLiD 4 and

5500xl SOLiD, and indel extraction for Complete Genomics was done by myself. The subsequent somatic indel calling was performed by Qi Wang using her own custom shell and python scripts.

HiSeq2000

Indels were called in the tumor sample using the two-sample model from Pindel [128], an algorithm to detect insertions and deletions from paired-end short reads. Pindel was preferred to samtools because it allows to call larger indels (insertions: 1-20 bp; deletions: 1bp-10kb), which makes its results more comparable to those obtained for SOLiD 4 and 5500xl SOLiD using LifeScope, and those provided by Complete Genomics (both calling indels up to 50 bp, according to personal communication with Life Technologies and to the Complete Genomics FAQ, `http://www.completegenomics.com/FAQs/Variant-Calls-SNPs-and-Small-Indels/`).

The Pindel output was converted to vcf and compared to the pileup of the control sample using a custom script. A somatic indel was called when at least two reads supporting the indel were identified in the tumor sample, and no evidence of the indel was found in the control sample. We required a minimum coverage level of at least 3 reads in control for the position to be considered. The region around the indel (+/- 10 bp) was then scanned for its deviations from the reference. Regions with more than one indel or with a mismatch density - computed over all reads - higher than the average error rate of 0.01 (as assessed by the PhiX control kit, see also the HiSeq2000 User Guide, `http://support.illumina.com/documents/documentation/System_Documentation/HiSeq2000/HiSeq2000_User_Guide_15011190_P.pdf`) were considered too error-prone and excluded. Insertions corresponding to homopolymers of length > 5 were also excluded as they are known to have a high error rate [129].

Complete Genomics

Somatic indel calls were extracted from Complete Genomics' `somaticVcfBeta` file using the same procedure as described for somatic SNVs in section 2.2.9, "Detection of somatic SNVs".

SOLiD 4 and 5500xl SOLiD

Indels were called with LifeScope 2.3 using the small indel pipeline. Indels with reads supporting the variant found on one strand only were filtered out, the filtered results were then converted to the generic vcf format using a custom perl script. Somatic indels were then called as described above for HiSeq2000.

Validation

For each set of somatic indels called, all those called as germline in any of the technologies were removed. For validation, the only somatic indel called for all four technologies was chosen, as well as two somatic indels called both on HiSeq2000 and 5500xl SOLiD, two somatic indels called both on HiSeq2000 and Complete Genomics, and, for HiSeq2000, Complete Genomics and 5500xl SOLiD, 6 to 10 somatic indels called for this platform only. The chosen somatic indels were externally validated by Sanger sequencing.

2.2.11 Statistical tests

For the pairwise platform comparisons of GC bias, I used Kolmogorov-Smirnov tests for GC percentages below 25% and above 60%. Coverage input values were sampled from the loess curves. For the comparison of the coverage distribution between platforms, and for the comparison between platforms of the fraction without coverage for specific genomic regions, I used two-sample Students t-tests. For the comparison of the ROC curves, Marc Zapatka focused on the sensitivity, comparing the sensitivity between different technologies and samples with paired two-sample Student's t-tests.

Differences yielding p-values below or equal to 0.05 were considered significant. No p-values were computed for 5500xl SOLiD because of the small sample size (two samples).

2.3 Data access

All short-read sequencing data have been deposited at the European Genome-phenome Archive (EGA, `http://www.ebi.ac.uk/ega/`), which is hosted by the EBI, under accession number **EGAS00001000274**. The Affymetrix SNP6 array data has been deposited at Array Express (`http://www.ebi.ac.uk/arrayexpress/`) under accession number **E-MTAB-1159**. The main scripts used for this study are listed in the Technical Annex (Section 5) at the end of this thesis.

3 Results

Sections 3.1- 3.4 are partly taken from my first-author publication [119]. The analyses were performed using the sequenced whole-genome data from two medulloblastoma tumor-normal pairs, MB14/BL14 and MB24/BL24. The samples were sequenced with the four technologies assessed in this study, HiSeq2000, SOLiD 4, 5500xl SOLiD, and Complete Genomics. Unless otherwise mentioned, all results correspond to 30x mean coverage, or for Complete Genomics to full coverage generated. Complete Genomics data is included both at full coverage and at downsampled 30x coverage, because they provide a proprietary sequencing solution with a usually higher coverage (40-80x) including their analysis of the results. It is not possible to purchase a lower coverage.

In the first part of this comparison, I assess how evenly the reads from every platform are spread across the genome. A sample's coverage level and distribution are crucial for its analysis, as this can introduce considerable bias into the results, even more so in tumor samples, where contamination with normal tissue, copy number variation, and tumor heterogeneity can heavily influence variant allele frequency. For each of the four technologies and for each of the patient samples, I analyzed the read distribution with respect to the GC content of the underlying sequence (section 3.1), the general genome-wide distribution of coverage levels (section 3.2), as well as the coverage in specific genomic regions, like exons or CpG islands (section 3.3), and the size and fraction of regions without coverage (section 3.4). In addition, I examined whether a combination of reads from two different platforms can lead to a better coverage of certain genomic or functional regions (section 3.3).

We further investigated the differences between platforms in sensitivity and specificity of SNV calling, first by comparing SNV calls from platforms and from combinations of platforms to the results of a SNP array (sections 3.5 and 3.6). Then I called somatic variations in the cancer samples and analyzed the overlaps between platforms. A validation with Sanger sequencing experiments was performed (section 3.7). The SNP calling analysis in sections 3.5 and 3.6 was mainly conducted by Marc Zapatka (DKFZ), the somatic indel calling analysis in section 3.7.2 was mainly conducted by Qi Wang (DKFZ). These results are included here for completeness.

3.1 GC bias

As explained in section 1.1.2, the GC bias describes the dependence between coverage and GC contents, where both GC-rich and GC-poor regions are less well covered than regions with balanced base composition. Ideally, with no GC bias present, we would see a uniform distribution of coverage, independent of GC content. Figures 3.1 - 3.4 show the genomic coverage per 1 kb interval for each platform, sorted by GC content of that interval. Each figure corresponds to one of the patient samples.

A one-sample Kolmogorov-Smirnov test confirms that the data from each of the platforms significantly deviates from the uniform distribution, i.e. a GC bias is found for every platform. Resulting p-values for patient sample MB24 range between 4.4e-07 and 6.5e-12. To test for significant differences between platforms, I used two-sample Kolmogorov-Smirnov tests for low ($\leq$ 25% GC content per 1 kb bin) and high ($\geq$ 60%) GC contents. An overview of the resulting p-values is given in Table 3.1.

Significant differences in GC bias were found between all platforms, except for SOLiD 4 vs. HiSeq2000 for a GC percentage above 60%. The most pronounced GC bias is found for Life Technologies' SOLiD 4 and 5500xl SOLiD, especially in regions with more than 60% GC content. HiSeq2000 shows a slightly reduced GC bias here (significant in two out of four samples: MB14 and BL14). Note that for HiSeq2000 sequencing, v2 chemistry was used for of all four samples. However, the latest release of v3 chemistry does not reveal a dramatic reduction in GC bias compared to the earlier v2 chemistry, as can be seen in Figure 3.5. The least GC bias for GC-rich regions by far is revealed by Complete Genomics, even when the higher mean coverage of around 50x (hereafter, "Complete Genomics") is computationally reduced to 30x mean coverage (hereafter, "Complete Genomics 30x") for comparison reasons.

At regions with GC content lower than 25%, 5500xl SOLiD and HiSeq2000 perform similarly with a generally lower bias than SOLiD 4 and Complete Genomics. In contrast to its behavior in GC-rich regions, Complete Genomics performs worst in GC-poor regions at downsampled 30x coverage. The GC bias at GC-rich and GC-poor regions, respectively, was consistently found across all four sequenced samples, except for patient sample BL14 where HiSeq2000 and Complete Genomics 30x perform similarly: the p-values of the Kolmogorov-Smirnov test are 0.9307 for %GC $\leq$ 25% and 0.4755 for %GC $\geq$ 60%, as listed in Table 3.1.

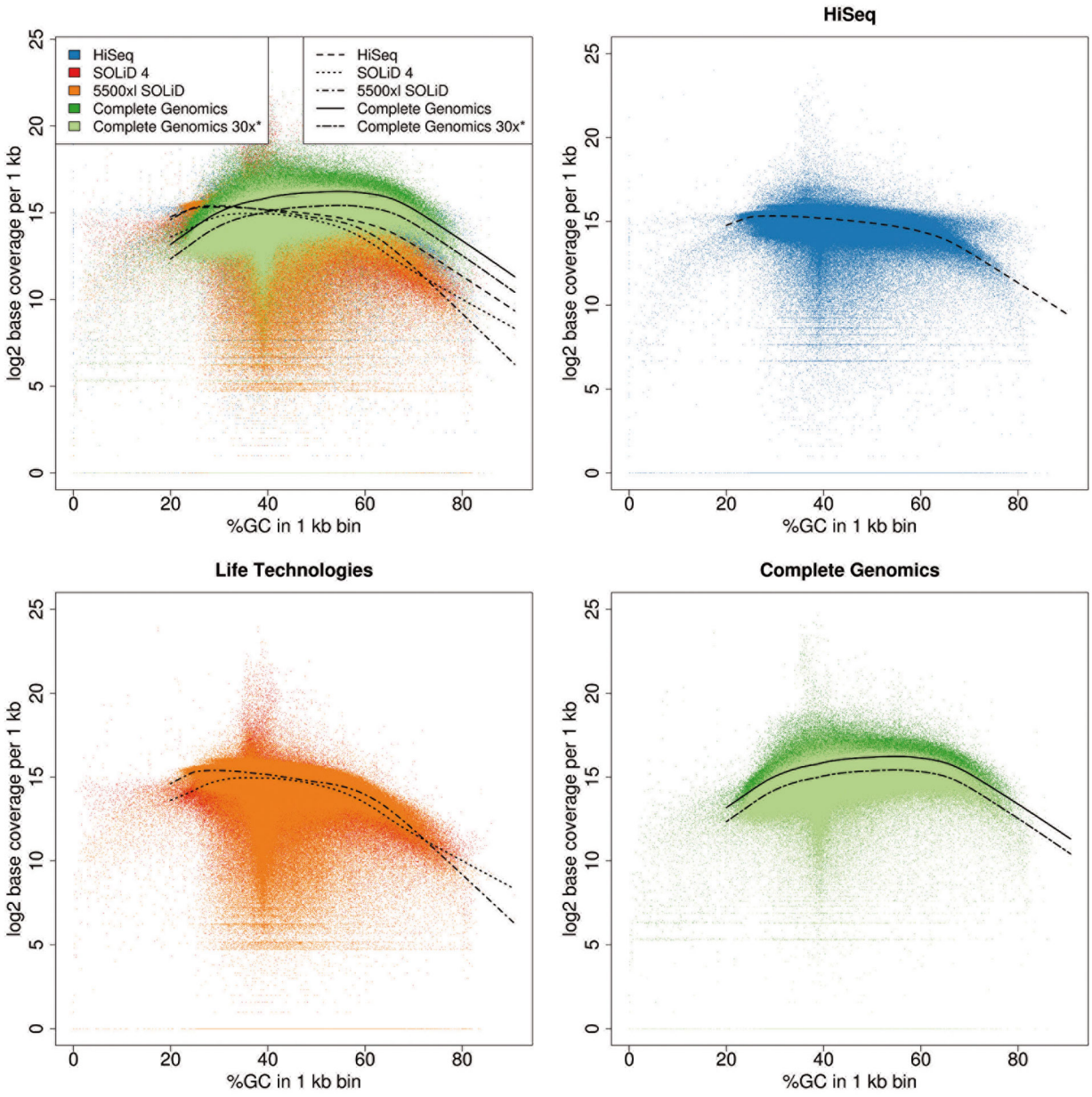

Figure 3.1: GC bias for each platform for sample MB24. Log2 base coverage in 1kb windows versus GC content in percent is shown for HiSeq2000, SOLiD 4, 5500xl SOLiD, and Complete Genomics data. The first panel shows an overlay of all four technologies. The upper right panel shows HiSeq2000 only (blue), the lower left SOLiD 4 and 5500xl SOLiD (red and orange, respectively), and the lower right Complete Genomics at full and downsampled 30x coverage (green and light green). Smoothed loess curves are fitted to each dataset to represent the local coverage trend.

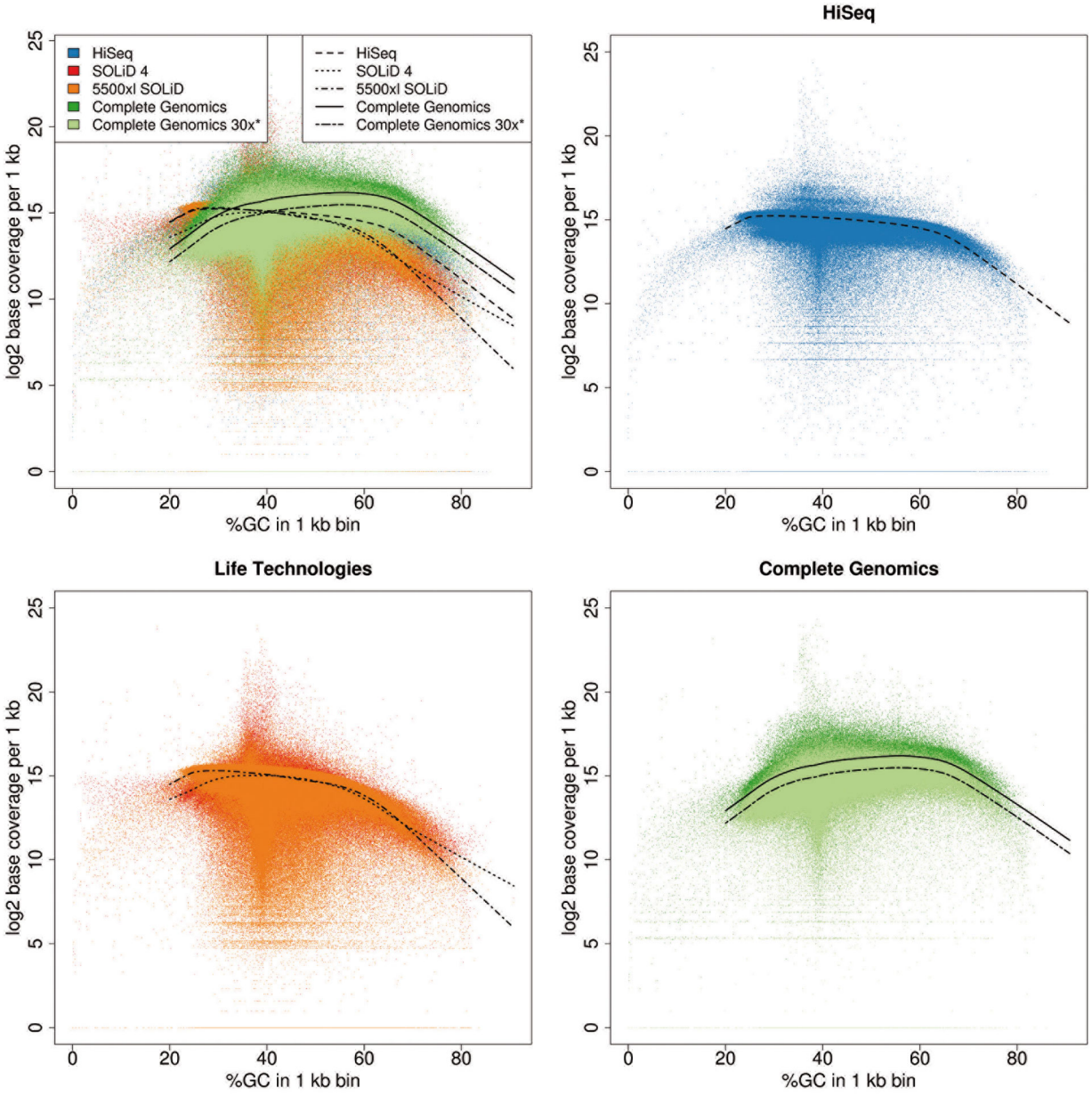

Figure 3.2: GC bias for each platform for sample BL24, plotted analogously to Figure 3.1.

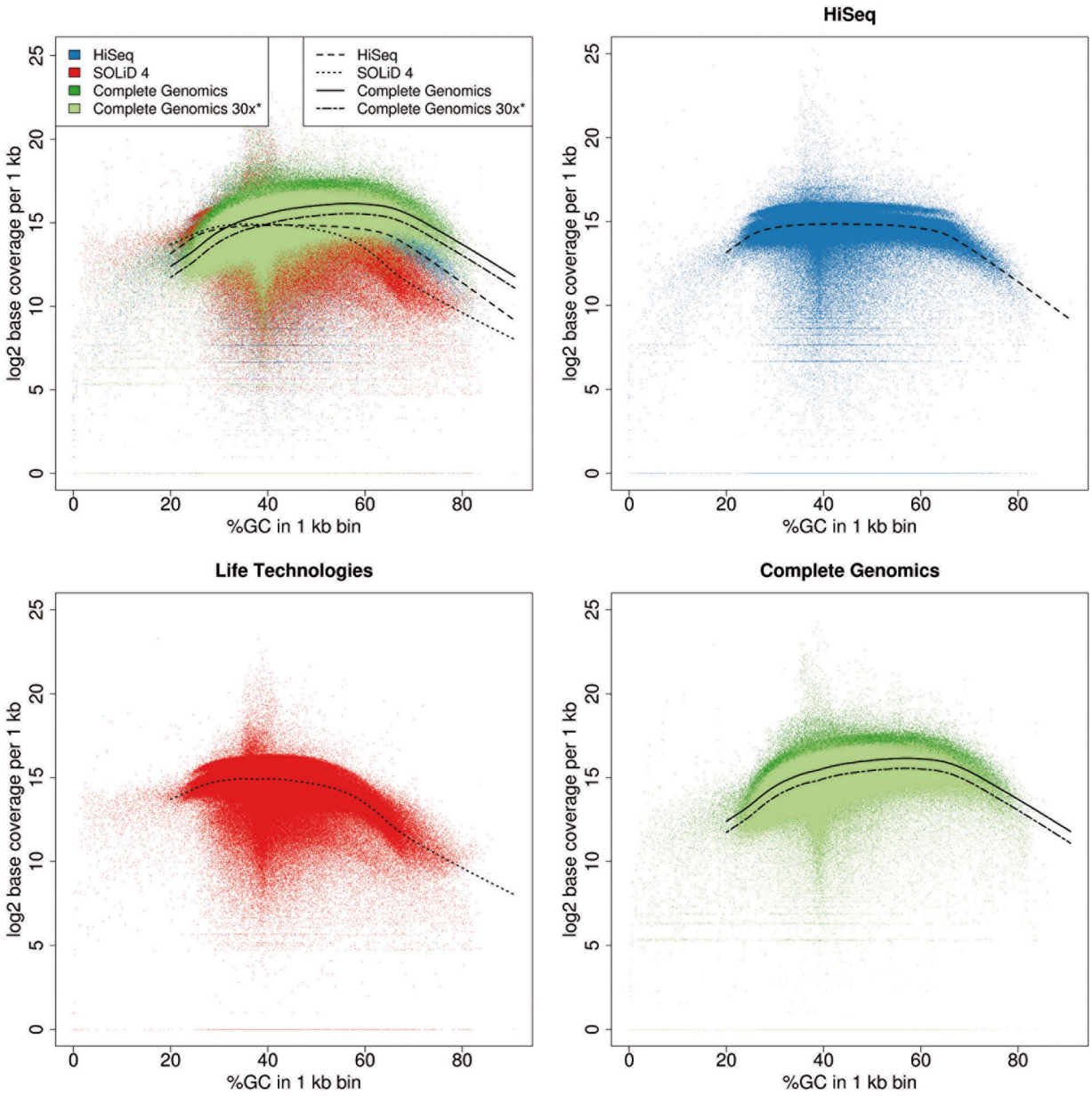

Figure 3.3: GC bias for each platform for sample MB14, plotted analogously to Figure 3.1. 5500xl SOLiD data is available only for samples MB24/BL24 (see Table 2.2).

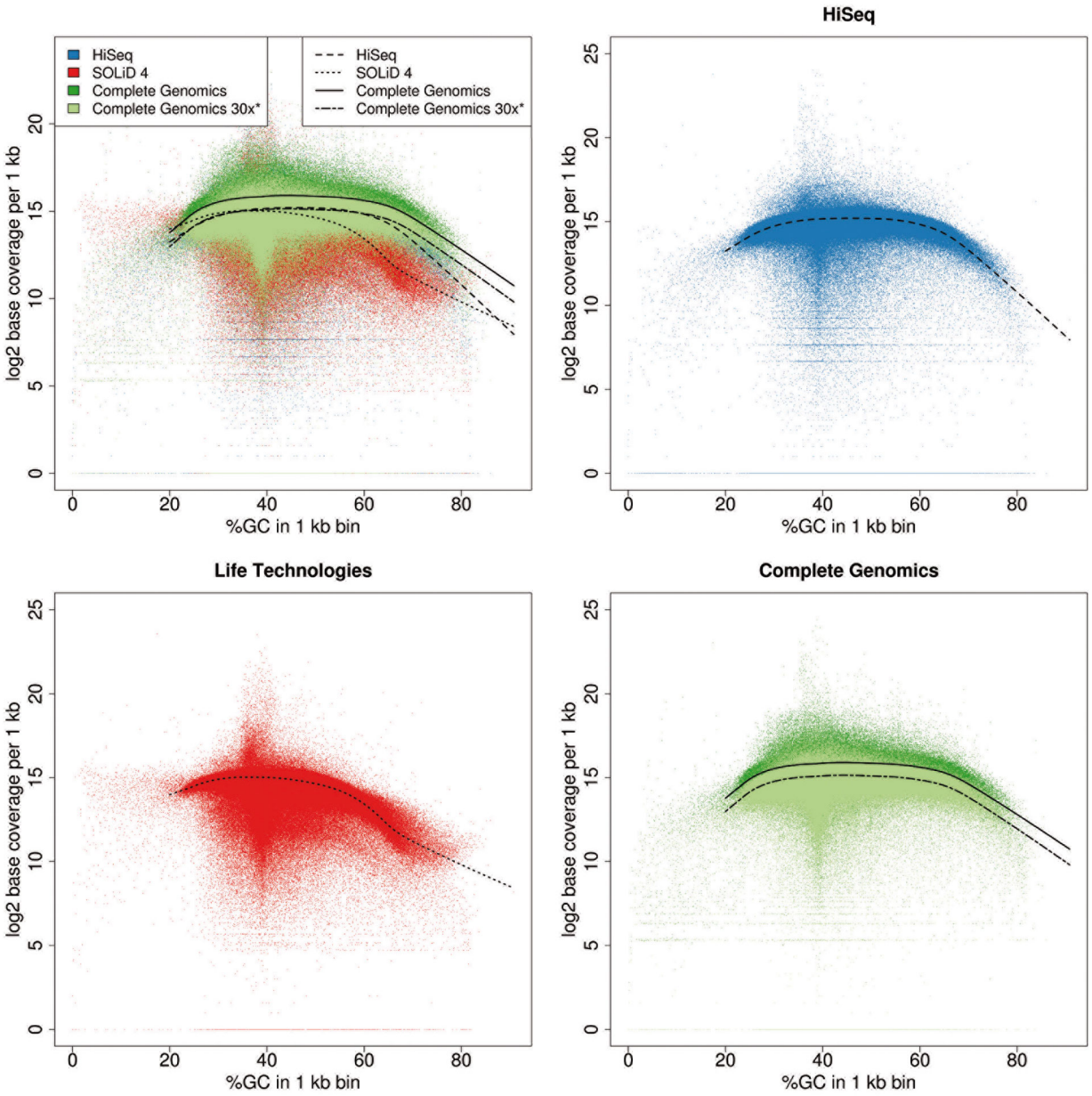

Figure 3.4: GC bias for each platform for sample BL14, plotted analogously to Figure 3.1. 5500xl SOLiD data is available only for samples MB24/BL24 (see Table 2.2).

		% GC in 1 kb bin $\leq$ 25%	% GC in 1 kb bin $\geq$ 60%
HiSeq2000 - CG 30x	MB14	***0.00217***	***0.01484***
	BL14	0.9307	0.4755
	MB24	***0.00217***	***0.03561***
	BL24	***0.00217***	***0.03561***
HiSeq2000 - SOLiD 4	MB14	0.474	***0.00561***
	BL14	***0.02597***	***0.03561***
	MB24	***0.00217***	0.0779
	BL24	***0.00217***	0.1558
CG 30x - SOLiD 4	MB14	***0.00217***	***3.352e-08***
	BL14	***0.02597***	***0.00016***
	MB24	***0.00217***	***3.964e-05***
	BL24	***0.00217***	***0.00016***

Table 3.1: P-values from Kolmogorov-Smirnov tests for pairwise platform comparisons of GC bias. CG 30x stands for Complete Genomics downsampled to 30x mean coverage. P-values below 0.05 are highlighted in bold and italic.

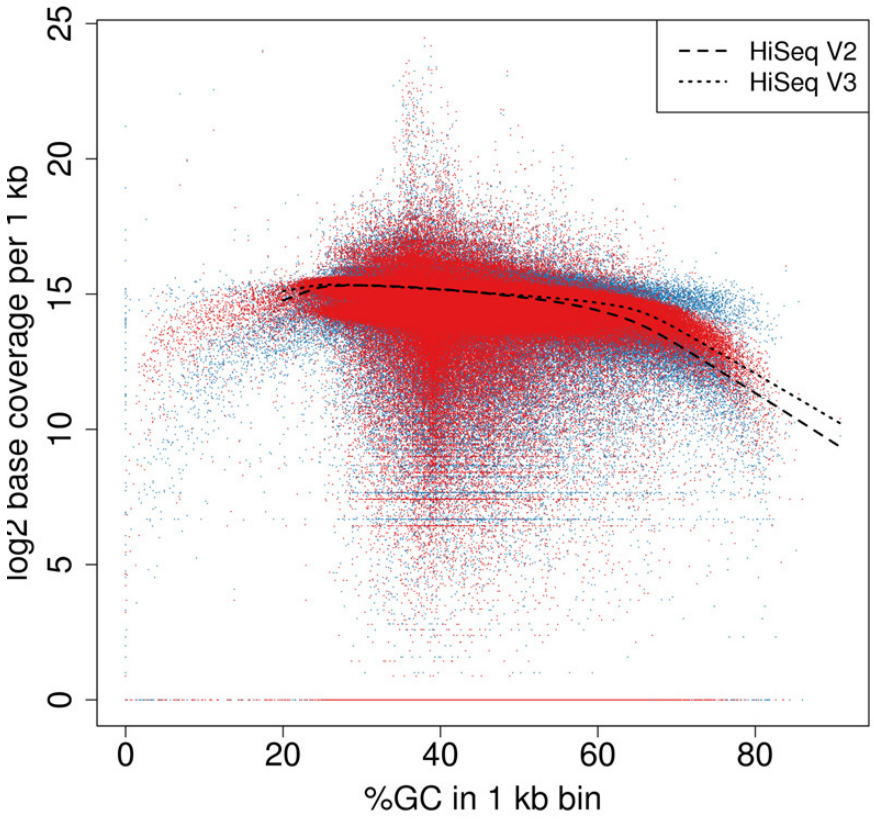

Figure 3.5: GC bias for HiSeq2000 with v2 chemistry versus HiSeq2000 with v3 chemistry. Log2 base coverage in 1kb windows is plotted versus GC content in percent. Smoothed loess curves are fitted to each dataset to represent the local coverage trend, analogously to Figure 3.1. Exemplary data from patient sample MB24 (v2, blue) is compared to another medulloblastoma patient sample (v3, red).

3.2 Coverage distribution

Looking at the genome-wide distribution of coverage levels, a number of striking differences between platforms can be seen (Figure 3.6). The most obvious divergence appears between Life Technologies' platforms, SOLiD 4 and 5500xl SOLiD, on one side, and HiSeq2000 and Complete Genomics on the other. At the same mean coverage, SOLiD 4 and 5500xl SOLiD show about 6 times more bases supported by less than 5 reads compared to HiSeq2000 and Complete Genomics. While the latter two show similar numbers in this respect, downsampling Complete Genomics to 30x for fairness of comparison shows a mean increase of almost factor 2.5 in bases supported by less than 5 reads. An average of these numbers across all samples is given in Table 3.2.

Coverage distribution is similar for Life Technologies' platforms SOLiD 4 and 5500xl SOLiD, with 5500xl SOLiD showing a slightly higher number of bases with higher coverage (20-60x). HiSeq2000 shows by far the narrowest coverage distribution compared to all other sequencing platforms, meaning its coverage is the most evenly distributed across the genome. Complete Genomics has the broadest coverage distribution, i.e. the highest deviations from the mean. Even for Complete Genomics downsampled to 30x mean coverage, the coverage distribution is still wider than the one resulting from HiSeq2000.

The cumulative coverage distribution, i.e. the percentage of the genome covered with at least `n` x, is shown in Figures 3.7 to 3.9, with Figure 3.7 showing the distribution for all samples and all platforms, Figure 3.8 showing the distribution of the sample means for each platform, and Figure 3.9 showing a magnified view of the sample means. These reveal that 5500xl SOLiD covers the smallest percentage of the genome, while HiSeq2000 and SOLiD 4 cover a similar and slightly higher fraction. However, the fraction of the genome covered for all three platforms is exceeded by Complete Genomics at both 30x and full coverage.

Further, higher variations in coverage distribution can be observed between the different samples sequenced by Complete Genomics compared to the other platforms, with the fraction of the genome covered with at least 30x differing up to about 15%, and even up to about 18% for the fraction of the genome covered with at least 50x (Figure 3.7). This is probably largely due to the differences in average coverage between Complete Genomics samples (see Table 2.2), but the between-sample variation can still be observed to a slightly lesser extent for low cumulative read depth at downsampled 30x coverage.

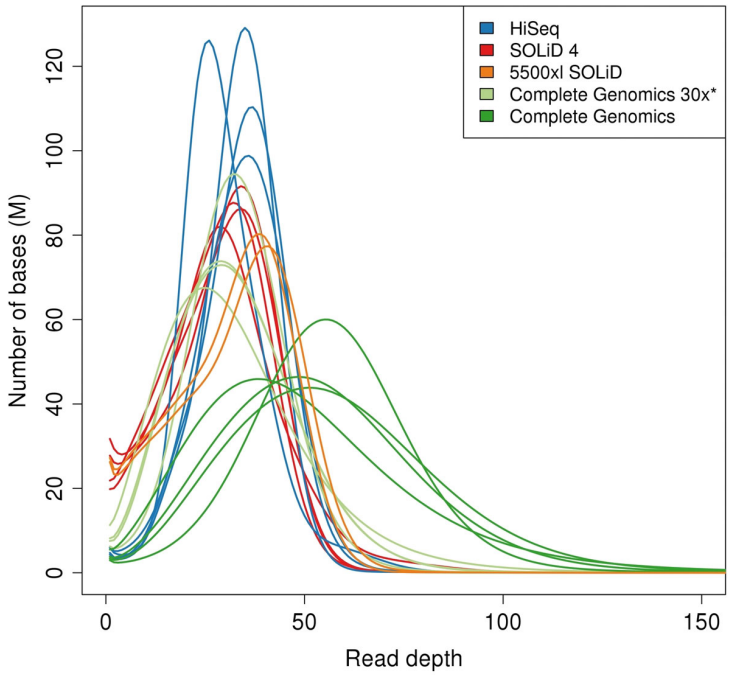

Figure 3.6: Distribution of genome-wide base coverage for each of the four platforms. Complete Genomics is shown at full coverage and at downsampled 30x coverage. Each curve corresponds to one sample.

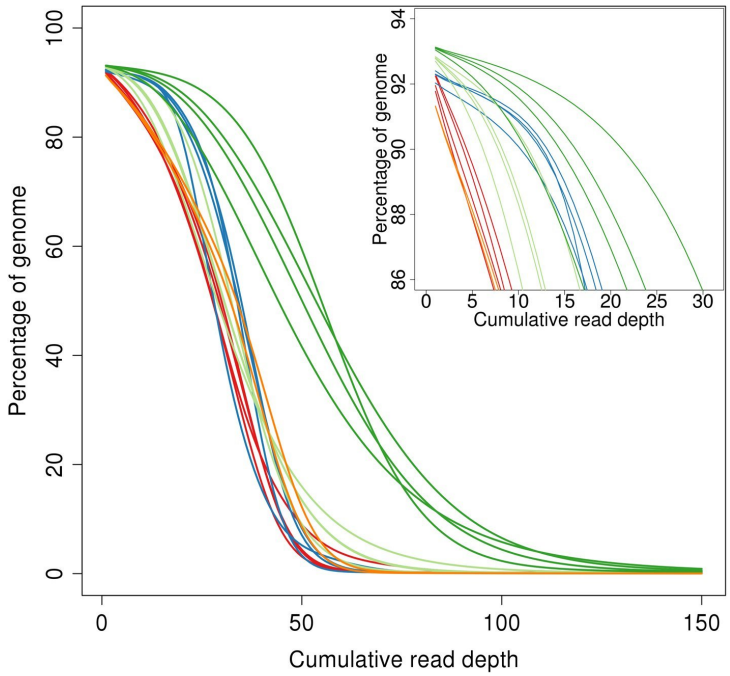

Figure 3.7: Percentage of genome covered by a given read depth. Each curve corresponds to one sample. The colors used correspond to the legend in Figure 3.6. The inset on the upper right shows a magnified view of the curves.

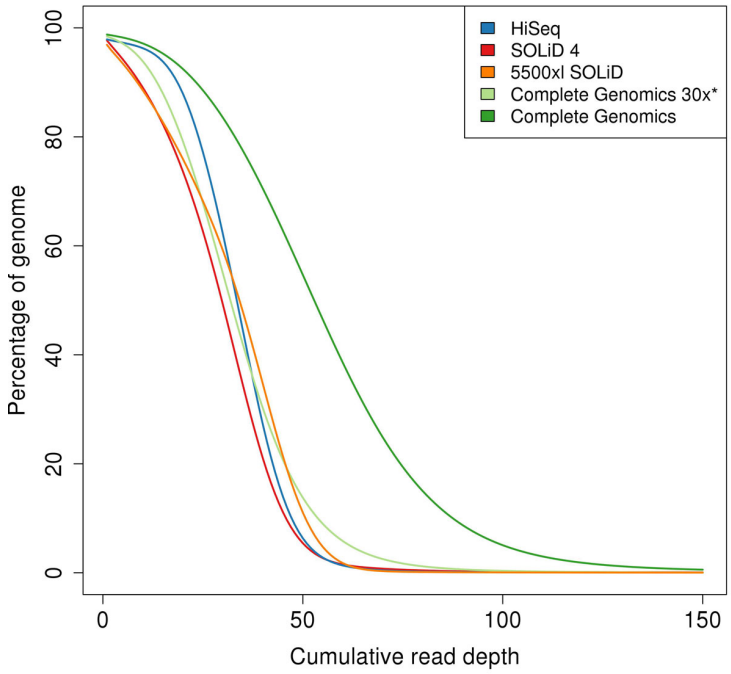

Figure 3.8: Mean percentage of genome covered by a given read depth. Each curve corresponds to the mean of all samples sequenced by one technology.

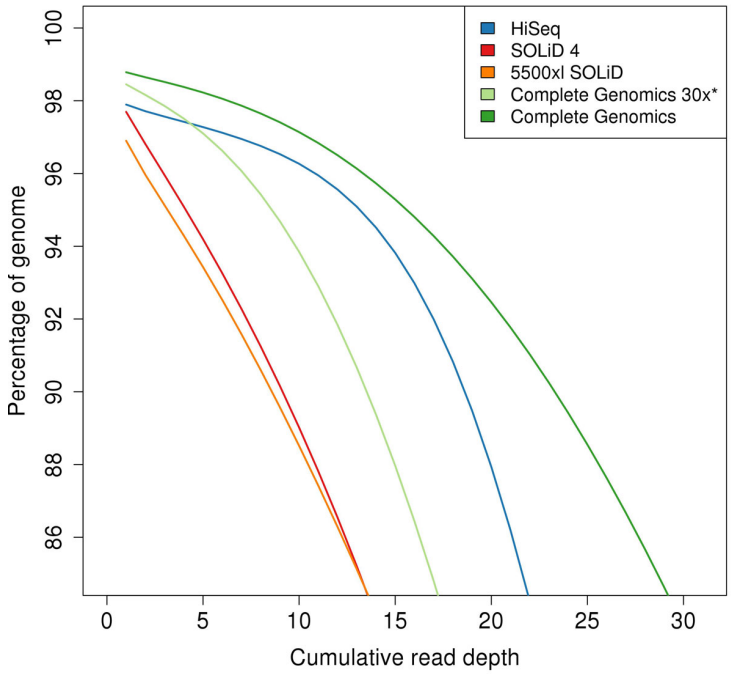

Figure 3.9: Magnified view of the mean percentage of genome covered by a given read depth as depicted in Figure 3.8. Each curve corresponds to the mean of all samples sequenced by one technology.

	Complete Genomics	Complete Genomics 30x	HiSeq2000	SOLiD 4	5500xl SOLiD
Total number of bases covered	2,826,524,353	2,817,003,995	2,801,114,390	2,795,379,490	2,772,621,192
Number of bases covered with less than 5 reads	15,938,617	38,555,229	17,727,532	100,145,774	99,297,132

Table 3.2: Number of bases covered on average across all samples, and average number of bases covered with less than five reads, for each platform assessed.

3.3 Coverage of genomic regions

To further evaluate the coverage differences between the different platforms, I analyzed the coverage in specific genomic and functional regions. Here, bases covered by fewer than three reads were considered as "not covered" or "without coverage", as explained in section 2.2.6.

Each of the four technologies has its strengths and weaknesses in covering different sections of the genome (Figure 3.10). Complete Genomics shows a similar coverage fraction for almost all regions, with a generally very low percentage ($< 2\%$) of bases not covered, both at 30x coverage and at full coverage. A comparably smaller covered fraction is observed only for regions containing a large number of short repeats, like simple repeats (24% uncovered at 30x coverage), low complexity repeats (11.9%), CpG islands (9.2%), and satellite repeats (3.7%). Overall, Complete Genomics performs better than all other technologies in this respect, except for simple repeat regions where it is surpassed by all three other platforms.

Comparative coverage of an exemplary simple repeat region is shown in Figure 3.11 (a). Almost no reads are mapped to this region by Complete Genomics. Read pairs with reads mapping to different chromosomes can be observed for HiSeq2000, SOLiD 4 and 5500xl SOLiD sequences, which reflects the difficulty of mapping reads to repeated sequences also for the latter three technologies. Interestingly, SOLiD 4 shows the highest coverage in this example, but also the largest number of differences from the reference genome.

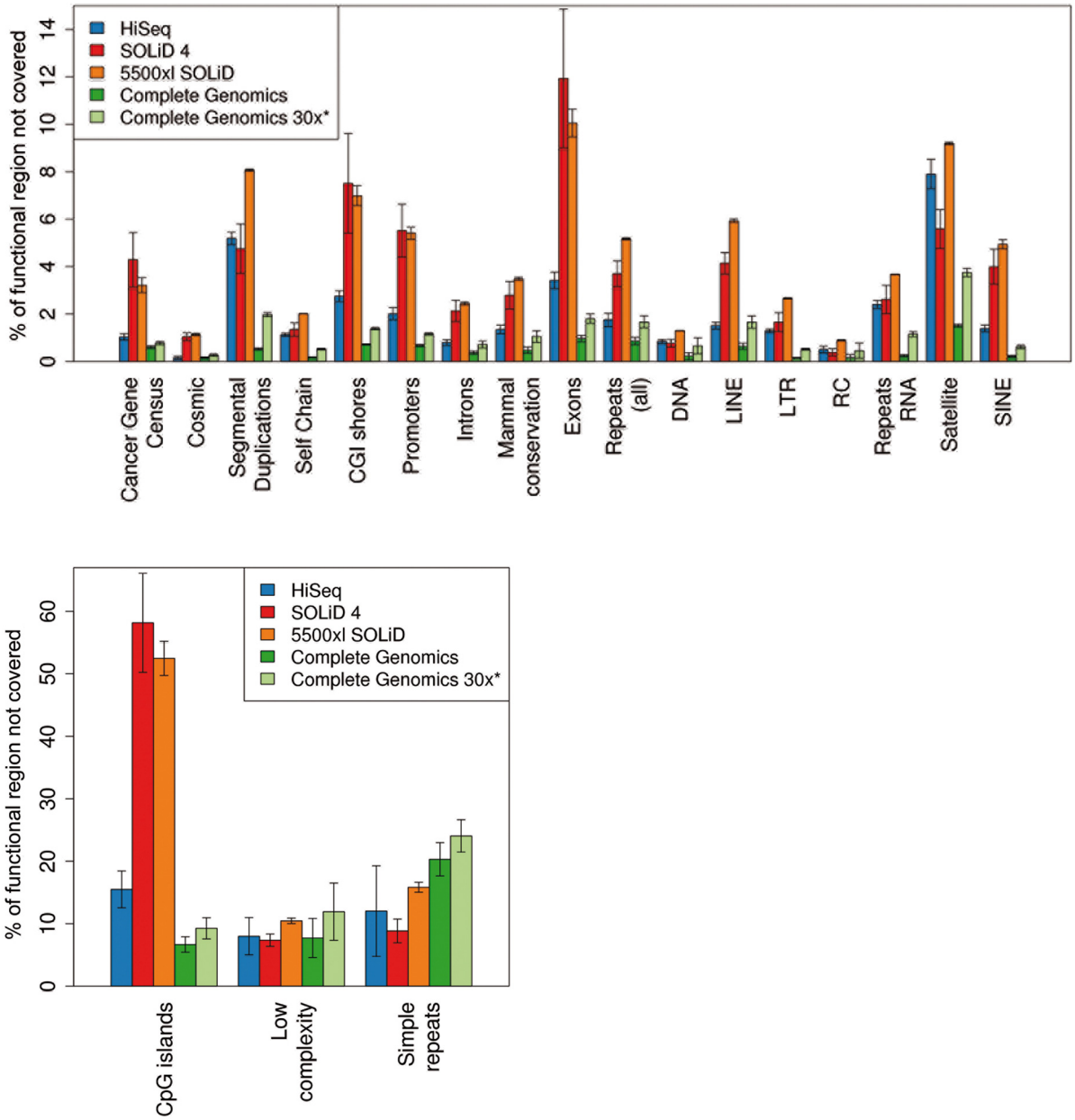

Figure 3.10: Mean percentage of uncovered bases across different genomic elements for each of the platforms. Bases covered with less than three reads were considered not covered. Note that reducing this threshold to 1 does not dramatically change the overall distribution of reads (see Figure 3.12). Error bars represent one standard deviation as obtained from analyzing all samples sequenced on one platform. DNA, LINE, Low complexity, LTR, RC, RNA, Satellite, Simple repeats and SINE are subcategories of Repeats (all). For better visibility, CpG islands, low complexity and simple repeats are plotted separately.

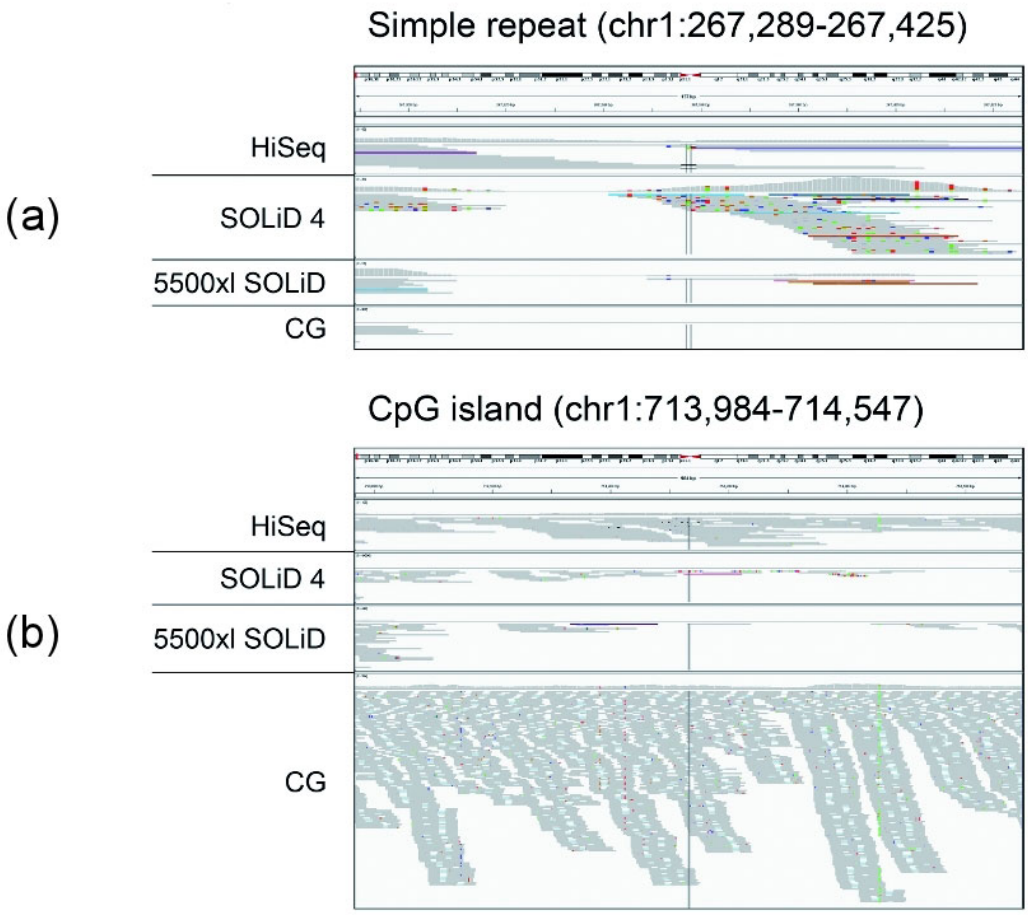

Figure 3.11: Visualization of read coverage for two exemplary genomic regions from patient sample MB24 by IGV [130] for HiSeq2000, SOLiD 4, 5500xl SOLiD and Complete Genomics. **(a)** A simple repeat region on chr1:267,289-267,425. **(b)** A CpG island on chr1:713,984-714,547.

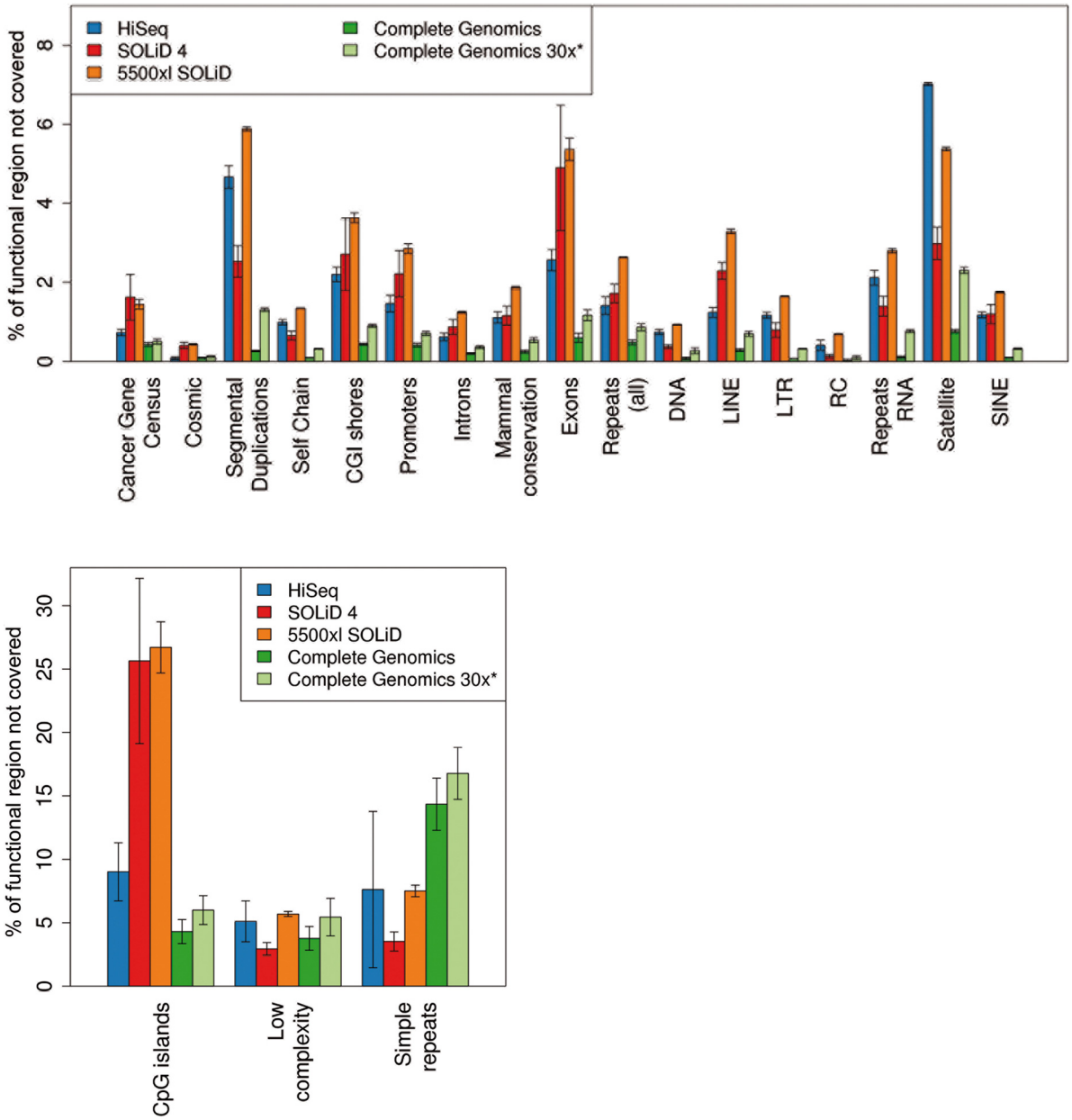

Figure 3.12: Mean percentage of uncovered bases across different genomic elements for each of the platforms. In this case, a base is considered not covered when it is covered by zero reads. The error bars represent one standard deviation as obtained from analyzing all samples sequenced on one platform. DNA, LINE, Low complexity, LTR, RC, RNA, Satellite, Simple repeats and SINE are subcategories of Repeats (all). For better visibility, CpG islands, low complexity and simple repeats are plotted separately.

SOLiD 4 and 5500xl SOLiD sequencing are most affected by GC content and consequently have by far the largest percentage of bases not covered in CpG islands (58.2% and 52.5%, respectively) and CpG island shores (7.5% and 7%, respectively). A t-test yields a p-value of 0.00069 (HiSeq2000 vs. SOLiD 4) and 0.0008 (Complete Genomics 30x vs. SOLiD 4) for CpG islands and a p-value of 0.019 (HiSeq2000 vs. SOLiD 4) and 0.01 (Complete Genomics 30x vs. SOLiD 4) for CpG island shores. For all platforms except for Complete Genomics, the fraction of CpG islands without coverage roughly doubles through our definition of an uncovered base (compare to Figure 3.12), showing that a large proportion of these regions is covered by less than 3 reads.

Coverage of an exemplary CpG island is shown in Figure 3.11 (b). Complete Genomics shows an impressive coverage of this region followed by HiSeq2000. The lowest coverage is present in SOLiD 4 and 5500xl SOLiD data.

Concordant with the differences in coverage of CpG regions, the exome coverage also shows dramatic differences between platforms with a mean difference in the fraction of bases not covered of factor 6.6 between Complete Genomics at 30x and SOLiD 4 (p-value 0.006). Overall, HiSeq2000 performs better than SOLiD 4 and 5500xl SOLiD in nearly all categories except for satellite regions (p-value 0.005 HiSeq2000 vs. SOLiD 4), and even outperforms Complete Genomics (both at full and at 30x coverage) in simple repeat regions (p-value 0.0386). SOLiD 4 performs slightly better than 5500xl SOLiD in repeat regions, while 5500xl SOLiD shows better coverage than SOLiD 4 in most other regions.

Interestingly, at the same mean 30x coverage, a combination of HiSeq2000 with 5500xl SOLiD data considerably decreases the uncovered fraction of certain repeat regions for both technologies, especially in satellites and simple repeats (Figure 3.13 (a) and (b)). Similarly, a combination of Complete Genomics data at full coverage with as little as 15x HiSeq2000 data (typically obtained with only one sequencing lane) shows a major increase of covered bases in simple repeats (Figure 3.13 (b)).

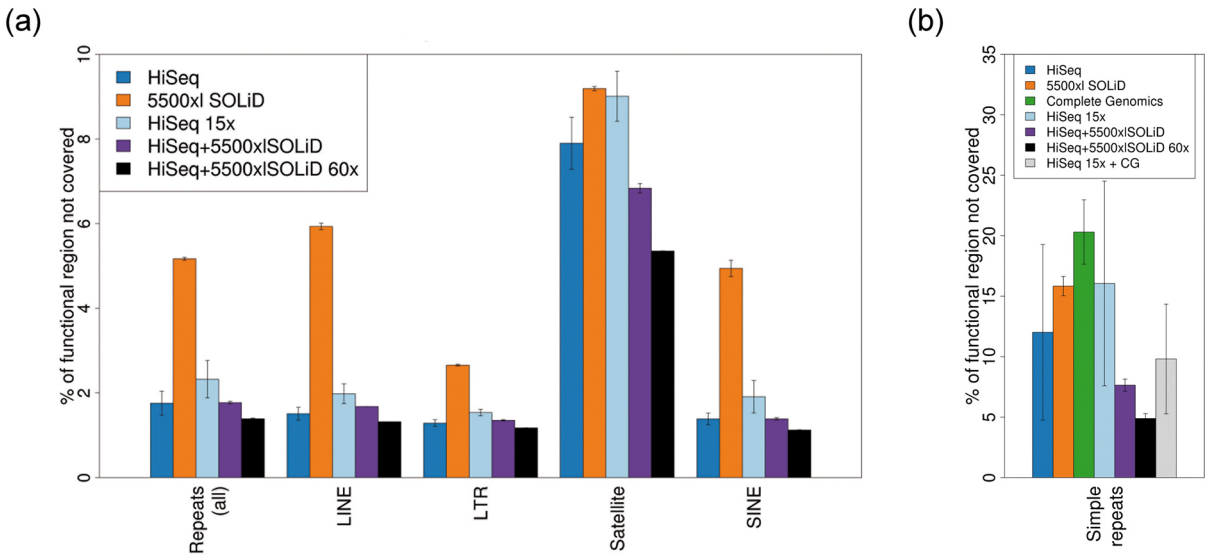

Figure 3.13: Mean fraction of uncovered bases across genomic elements for different combinations of technologies. Error bars represent one standard deviation as obtained from analyzing all samples in the group. **(a)** Mean fraction of uncovered bases for chosen repeat regions. Performance is compared to sequence data from single technology platforms. Only regions with observable differences are displayed. **(b)** Mean fraction of uncovered bases across simple repeat regions for different combinations of technologies. CG stands for Complete Genomics.

3.4 Regions without coverage

While the number of uncovered regions is similar for all platforms for larger-sized regions of 150 bp and above (see Figure 3.14), Life Technologies' platforms SOLiD 4 and 5500xl SOLiD show very high numbers of small regions without coverage compared to HiSeq2000 and Complete Genomics. The smaller the regions, the more pronounced are the differences between platforms, with HiSeq2000 performing better than Complete Genomics. 5500xl SOLiD shows slight improvement over SOLiD 4, except for extremely small regions of 1-2 bp, where the slight difference increases to a factor of 1.5 in the number of uncovered regions: on average, 384,304 uncovered regions of 1-2 bp are found for SOLiD 4, versus 260,252 regions for 5500xl SOLiD.

The size of the largest region without coverage is approximately 110,000 bp for all four platforms, except for HiSeq2000 whose largest region is 766,173 bp in length. However, this is due to the pseudoautosomal region on chromosome X and Y [131] and is a consequence of mapping differences: reads belonging to this region cannot be uniquely mapped to a reference genome containing the region on both chromosome X and Y and thus were automatically discarded before analysis (see section 2.2.2).

The fraction of the genome left without coverage (based on the reference genome excluding N's) at 30x coverage for HiSeq2000 and downsampled Complete Genomics is very similar (1.45% versus 1.61% on average across samples), both performing approximately 2.5 better in this respect than SOLiD 4 and 5500xl SOLiD. At 15x coverage, the difference between HiSeq2000 and the Life Technologies platforms is even more marked with a factor of approximately 3.5, suggesting that the latter can catch up at higher coverage. Notably, Complete Genomics at full coverage leaves only an average of 0.79% of the genome not covered.

3.5 SNP calling

After closely assessing coverage differences between the different technologies, we followed up with the identification of SNVs and the performance of the platforms in this respect. SNV calling is one of the major aims of sequencing projects, especially in cancer research.

To this end, Affymetrix SNP6 arrays were used as a gold standard for all samples. Affymetrix arrays are a sequencing-independent, well-established SNP calling

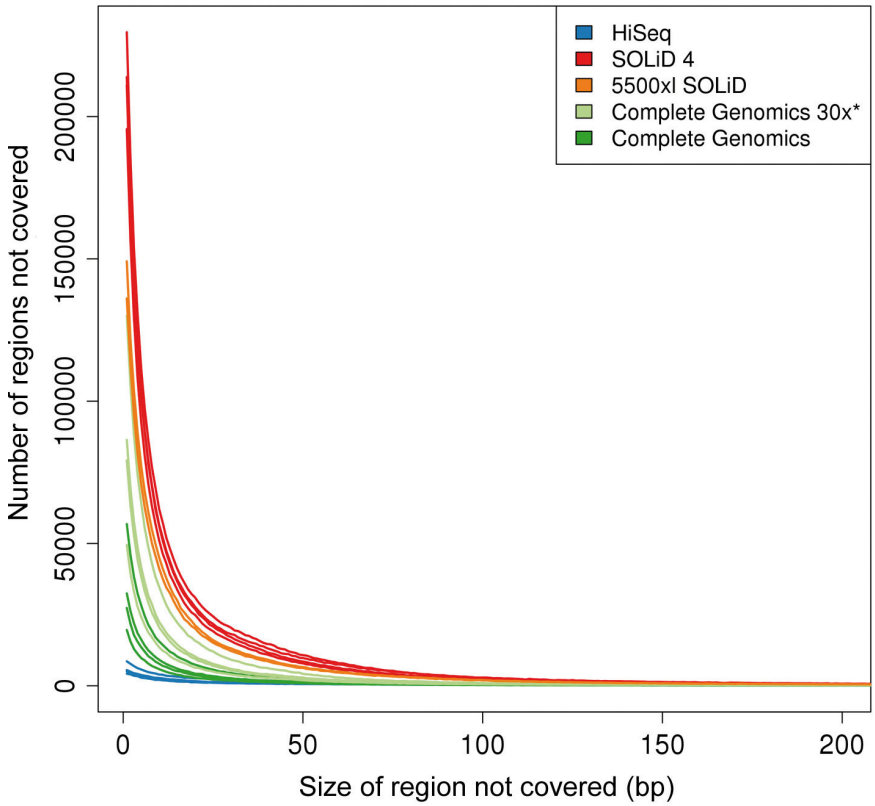

Figure 3.14: Size distribution of regions without coverage for all platforms and all samples, based on the reference genome excluding N's. Each curve corresponds to one sample. A base is considered uncovered when it is covered by less than three reads, as described in section 2.2.6. The x-axis is truncated at 200 bp.

technology[1]. To ensure that only minimal bias is introduced by the genomic positions assayed by the arrays, the distribution of array SNPs in different areas of the genome was examined. Figures 3.15 and 3.16 show an overview of the results.

Introns and regions of mammalian conservation show a similar representation on the SNP array and on the genome, and a high fraction of the regions contain at least one array SNP. Promoters, exons and CpG island (CGI) shores show a percentage of array SNPs similar to their fraction of the genome size. However, a large fraction of those regions does not contain SNPs measured on the array. Within CpG islands (covering 0.76% of the genome) the percentage of array SNPs (0.037%) is even a lot lower than expected from the size of CpG islands. Still, this corresponds to 1283 array SNPs falling into this genomic region, allowing a reasonable evaluation of the sequencing performance.

Repeat regions (Figure 3.16) show a similar representation on the SNP array and on the genome for most repeat types. Only SINEs are underrepresented (genome 27.1%, array SNPs 14.8% equivalent to 44659 SNPs), as are simple repeats, low complexity regions and satellite regions. However, this still results in a high number of array SNPs falling into these regions (respectively 1100, 374 and 758 array SNPs for simple repeats, low complexity regions and satellite regions).

Receiver operating characteristic (ROC) curves were computed for each platform and sample, as explained in section 2.2.8, to show the sensitivity and false positive rate (equivalent to `1 - specificity`) for SNV calling, using the Affymetrix SNP6 array as a reference (Figure 3.17 - 3.24). This SNP calling sensitivity should be considered an upper bound for somatic mutation calling, as cancer samples usually involve variant allele frequencies far below 50% (see section 1.2.3).

229 out of 907,551 SNPs, i.e. 0.025%, were not found on any of the four platforms, which may indicate that these SNPs are false positives on the SNP array. Looking at the ROC curves, the best platform in terms of sensitivity is HiSeq2000 (e.g., 99.15% for patient sample MB24, Figure 3.17 and 3.18), followed by Complete Genomics (e.g., 98.38% sensitivity for patient sample MB24). A t-test on sensitivity over all samples yields a p-value of 0.008651 for HiSeq2000 versus Complete Genomics. Notably, HiSeq2000 needs far less overall coverage than Complete Genomics to reach a comparable sensitivity: even at downsampled 15x coverage, HiSeq2000 reaches a

[1] `http://media.affymetrix.com/support/technical/other/snp6_array_publications.pdf`

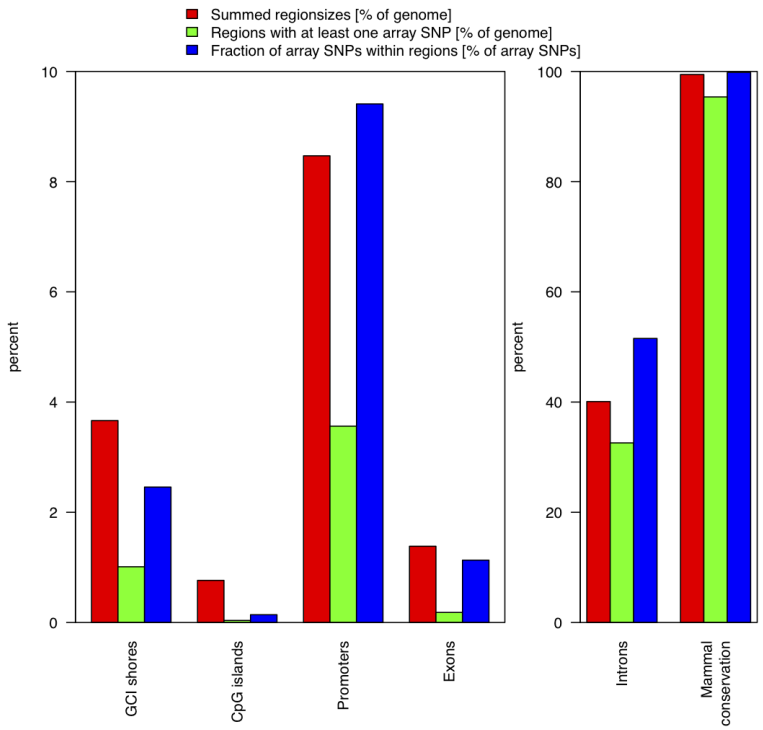

Figure 3.15: Percentage of genome covered by different genomic elements, in comparison to the distribution of Affymetrix array SNPs on these genomic elements, for patient sample MB24.

sensitivity of, e.g., 98.12% for patient sample MB24.

At the positions assayed by the SNP array, Complete Genomics shows a mean coverage considerably lower than its genome-wide mean coverage, while HiSeq2000 performs close to its overall mean coverage in this respect. E.g., for Complete Genomics on patient sample MB24, the SNP array positions are covered with 40x on average, versus 51.7x overall, while HiSeq2000's SNP array positions are covered with 32.1x. However, this still does not account for the fact that HiSeq2000 at 15x and Complete Genomics show a similar SNP calling sensitivity.

5500xl SOLiD and especially SOLiD 4 show a strongly reduced sensitivity in most samples. E.g., 5500xl SOLiD has a sensitivity of 96.80% on patient sample MB24, while SOLiD 4 reaches only 92.57% on the same sample. A t-test on all samples between SOLiD 4 and HiSeq2000 yields a p-value of 0.008324, and of 0.008189 between SOLiD 4 and Complete Genomics. However, this reduced sensitivity comes with a slightly lower false positive rate compared to HiSeq2000 (approximately 0.105-0.124% on patient sample MB24 for SOLiD 4 and 5500xl SOLiD, respectively).

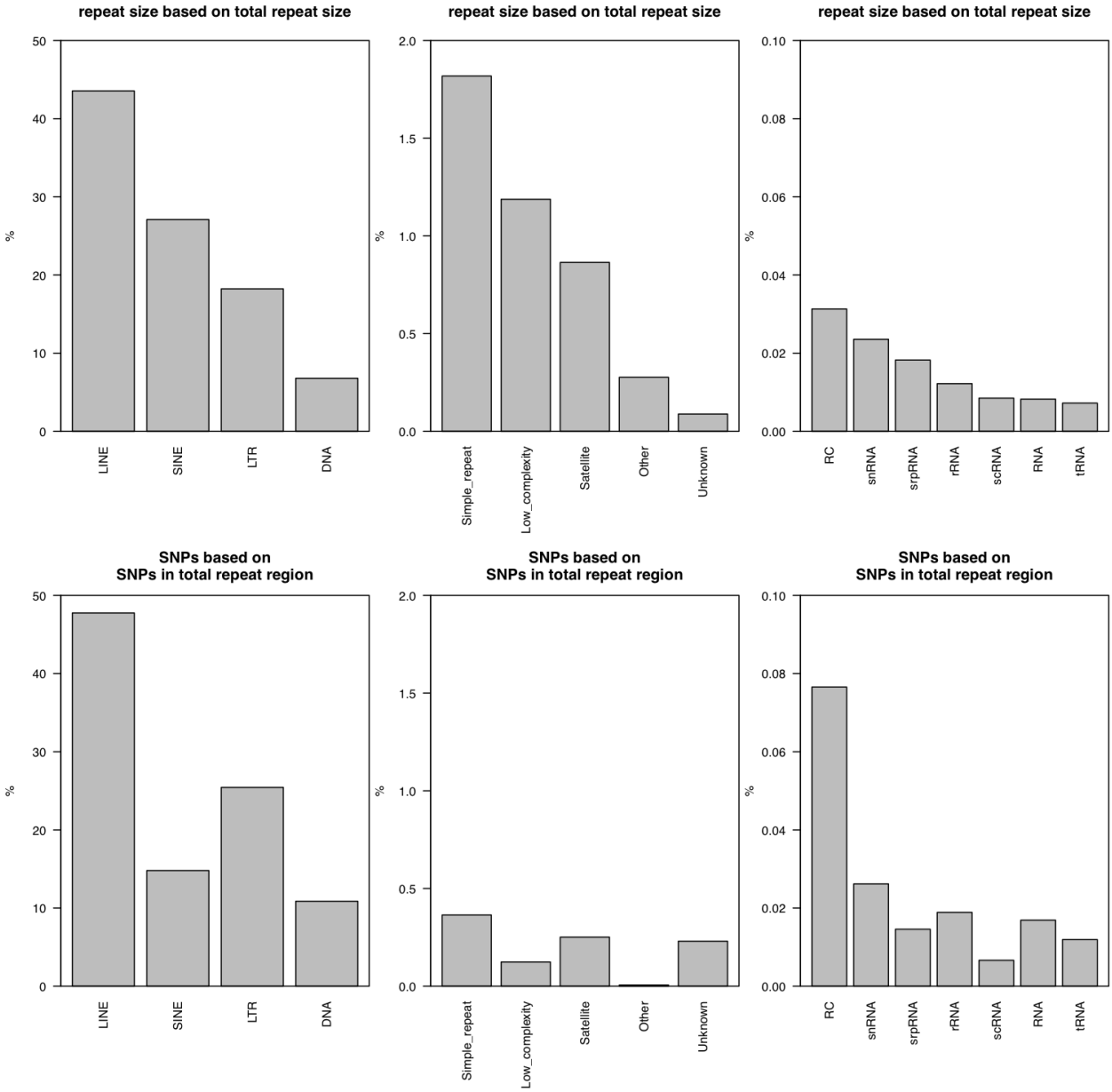

Figure 3.16: Distribution of Affymetrix array SNPs in repeat types analyzed for patient sample MB24. The percentage given in the upper panels for the size of each repeat region was computed in relation to the size of all genomic repeats. The percentage of array SNPs shown in the lower panels for specific repeat regions was computed in relation to the number of array SNPs in all genomic repeats.

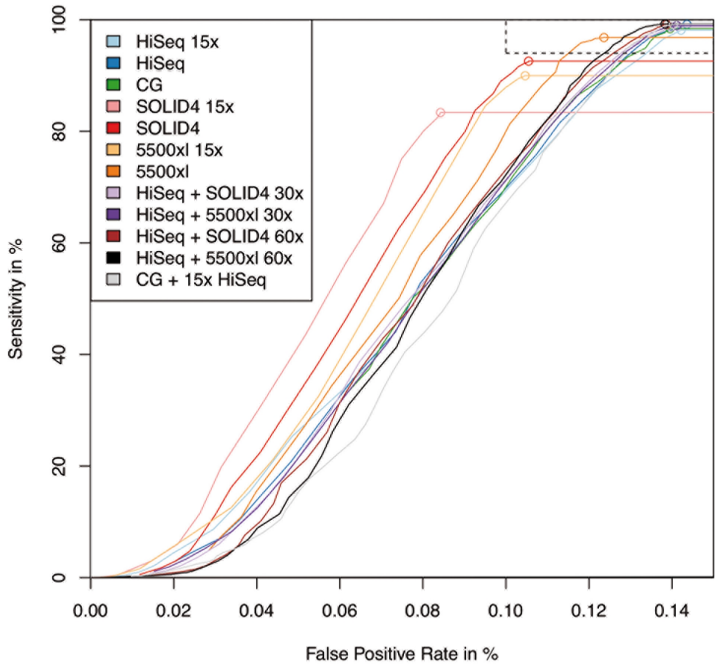

Figure 3.17: ROC curves comparing sensitivity and specificity of SNV calling for all platforms on sample MB24. The false positive rate (1 - specificity) is plotted from 0-0.15. All curves have reached their plateau at that point and will continue as straight lines.

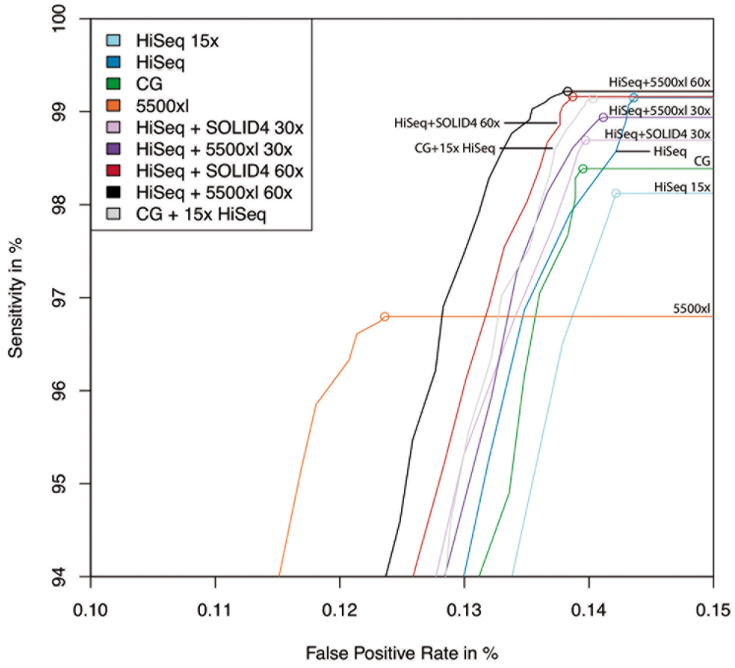

Figure 3.18: Magnified view of the ROC curves for sample MB24 as indicated by the dashed frame in Figure 3.17. Curves that do not appear in this plot reached their plateau below the sensitivity cutoff chosen for this window.

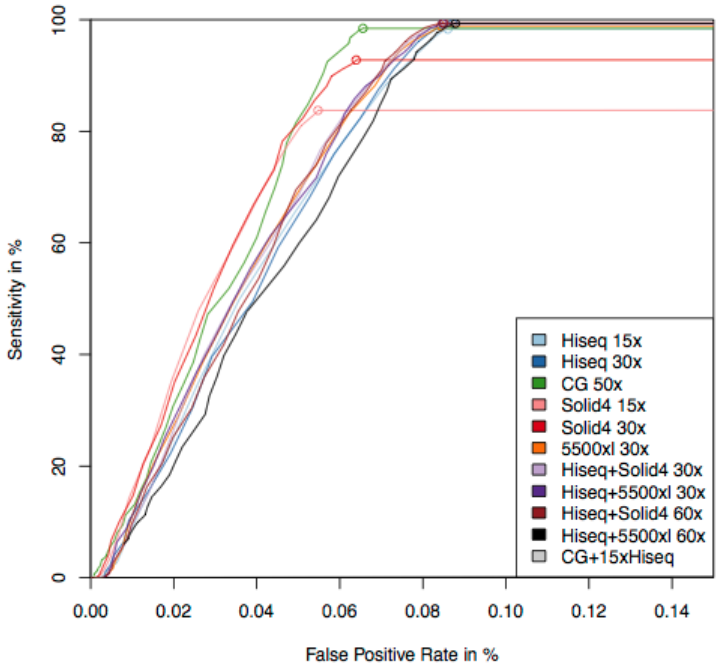

Figure 3.19: ROC curves for SNV calling on patient sample BL24, plotted analogously to Figure 3.17.

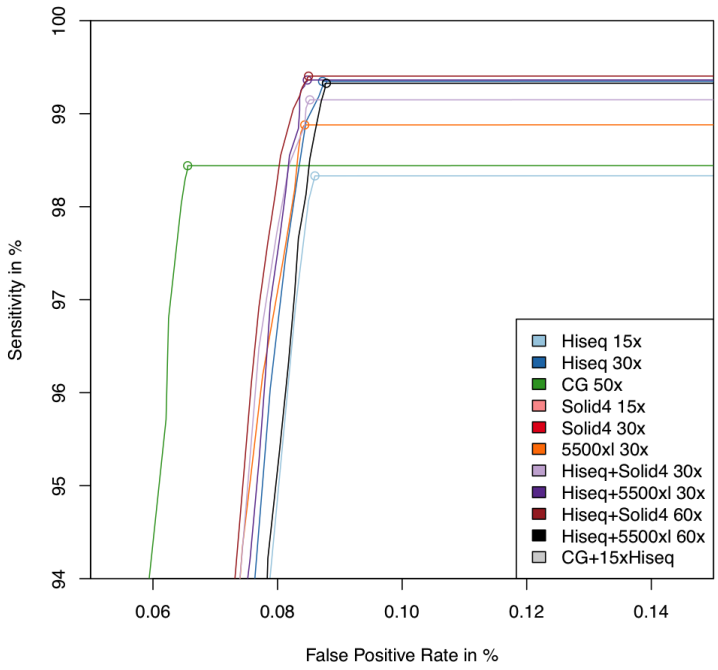

Figure 3.20: Magnified view of ROC curves for SNV calling on patient sample BL24, plotted analogously to Figure 3.18.

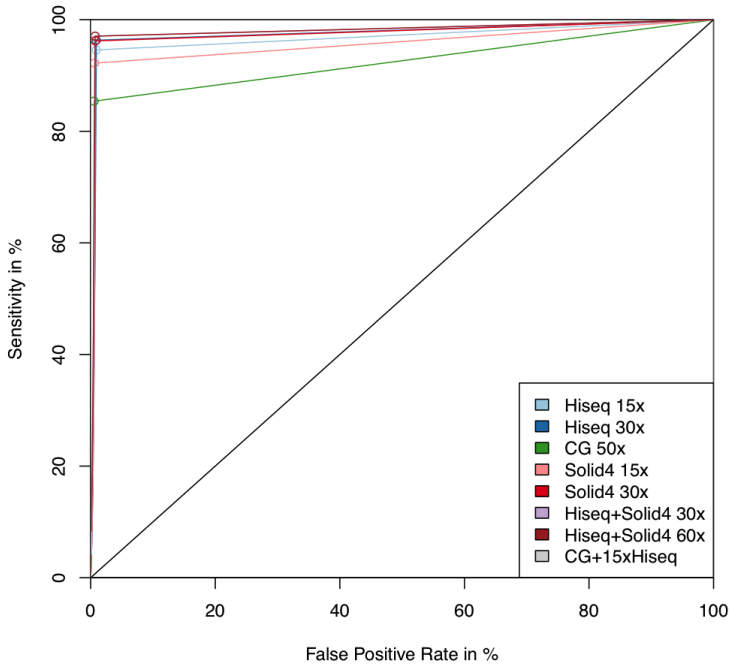

Figure 3.21: ROC curves for SNV calling on patient sample MB14, plotted analogously to Figure 3.17.

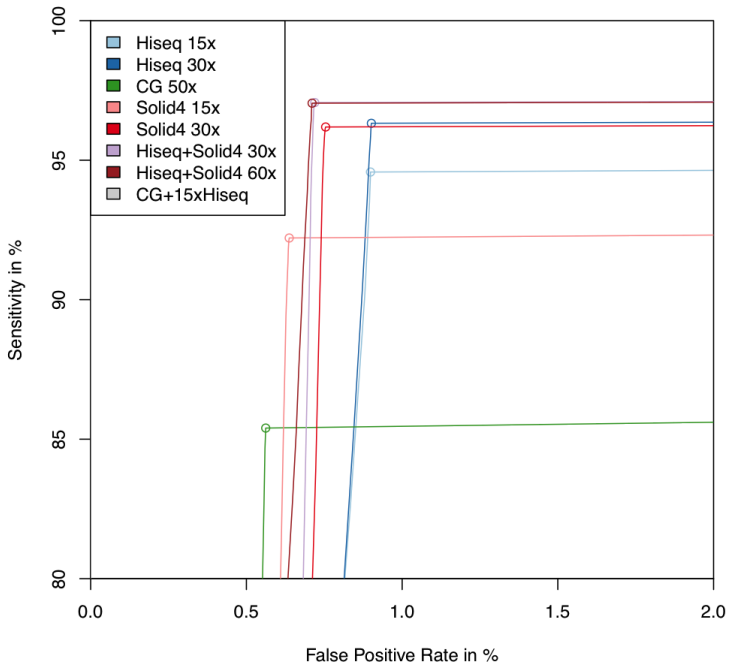

Figure 3.22: Magnified view of ROC curves for SNV calling on patient sample MB14, plotted analogously to Figure 3.18.

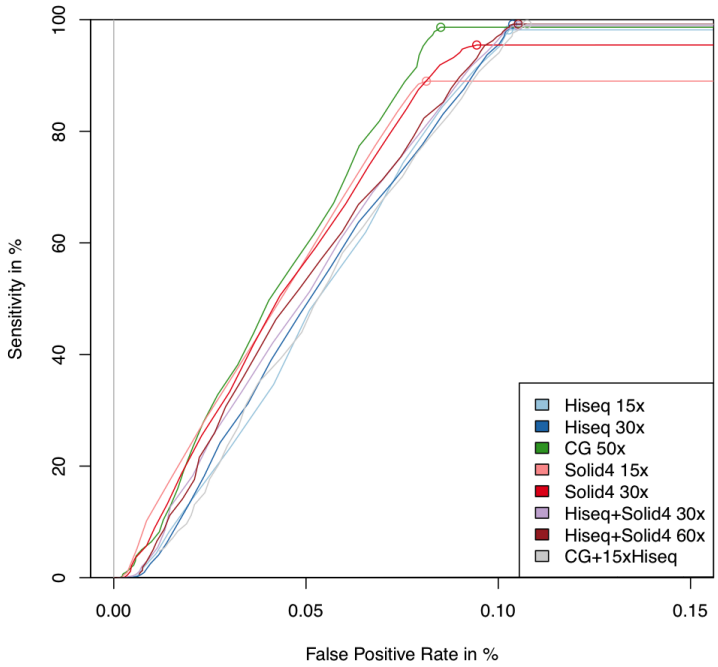

Figure 3.23: ROC curves for SNV calling on patient sample BL14, plotted analogously to Figure 3.17.

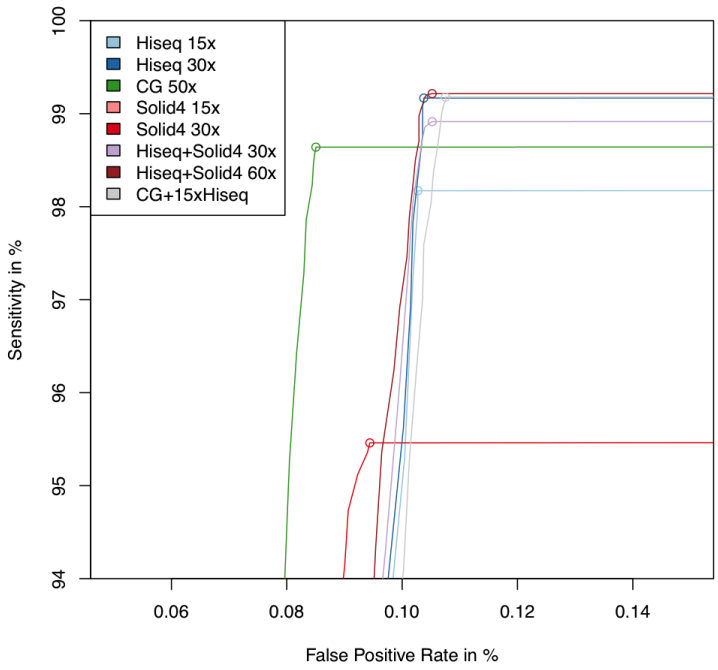

Figure 3.24: Magnified view of ROC curves for SNV calling on patient sample BL14, plotted analogously to Figure 3.18.

3.6 Combination of sequencing technologies

Additionally, we assessed whether combining sequencing data from different technologies would improve SNV calling results, uniting the strengths and compensating for the weaknesses of each platform. As expected, combining the data sets as they are, and thus reaching a coverage twice as high, e.g., with HiSeq2000 at 30x plus 5500xl SOLiD at 30x, yields a sensitivity and specificity slightly higher than any other technology alone (Figure 3.18).

However, when comparing SNV results at a similar coverage, i.e. with a combination of platforms at a total coverage of 30x, it is hardly possible to reach a sensitivity higher than the one achieved with HiSeq2000 sequencing alone. This result can be confirmed on all samples (see Figures 3.20, 3.22, and 3.24). Combining HiSeq2000 with 5500xl SOLiD data, at 15x each, yields good results. The sensitivity decreases only slightly compared to HiSeq2000 at full coverage, and specificity slightly increases above the level reached by 5500xl SOLiD. However, the decrease in sensitivity (0.17%) is far higher than the increase in specificity (0.0025%).

An interesting result is that for Complete Genomics sequencing, adding HiSeq2000 data at only 15x, a coverage that currently corresponds to only one lane of HiSeq2000 sequencing, can improve the SNV calling performance of Complete Genomics both in sensitivity and specificity. The increase in sensitivity reached here is 0.73% compared to Complete Genomics alone, which corresponds to a p-value of 0.008692, at a slightly increased specificity.

3.7 Somatic variation calling

Going beyond the calling of SNVs within one sample, we attempted to estimate the performance of the different platforms in cancer genome studies by calling somatic SNVs and somatic indels on the two tumor/normal pairs, MB14/BL14 and MB24/BL24.

3.7.1 Somatic SNVs

After calling somatic SNVs as described in section 2.2.9, I first examined the overlap between platforms. An overview of the number of somatic SNVs called for each technology is given in Table 3.3. Generally, patient sample MB14 seems to contain more somatic SNVs than MB24. However, it is striking to see that, for one and the same sample, these numbers are extremely different between platforms, ranging from 71 somatic SNVs on SOLiD 4, to 1599 on Complete Genomics for sample

pair MB24/BL24. In addition, the between-sample ratio also varies a lot between technologies, with 18.8 times more somatic SNVs found in patient sample MB14 (1333) than in sample MB24 (71) for SOLiD 4, while the ratio is 6.8 for HiSeq2000 and only 1.7 for Complete Genomics. As we know from other studies of the same samples that MB24 is tetraploid [110], this indicates that the platforms differ in their ability to pick up variants with low allele frequency.

	HiSeq2000	**SOLiD 4**	**5500xl SOLiD**	**Complete Genomics**
MB14	3950	1333	-	2962
MB24	584	71	470	1599

Table 3.3: Number of somatic SNVs called for each sample, for each of the four technologies.

The pairwise overlap of somatic SNVs between platforms is shown in Figure 3.25 and 3.26. In addition, the number of somatic SNVs called by all four platforms concordantly is only 456 for MB14 and only 13 for MB24. These numbers are drastically below any expectation. If anything, the concordance is a bit better for sample MB14, even though the number of somatic SNVs called is higher than in MB24.

Comparison to validation experiments

To further investigate the reasons for the drastic lack of overlap between somatic SNVs on different platforms, I had the opportunity to have a small number of SNVs externally validated with Sanger sequencing. For validation, variants from patient sample MB24 were chosen in order to include 5500xl SOLiD in the comparison, and because this sample seems to contain fewer somatic SNVs than sample MB14.

The 13 somatic SNVs that were concordantly called by all four technologies were selected for Sanger validation, as well as, for each platform, 10 to 15 somatic SNVs that were not called on any other platform, preferentially those called with high confidence. An overview of validation results is given in Table 3.4. The individual variants and their validation status are listed in Tables 3.5 (concordant SNVs selected for validation), 3.6 (HiSeq2000), 3.7 (SOLiD 4), 3.8 (5500xl SOLiD), and 3.9 (Complete Genomics). The reason why the number of SNVs validated is different for each platform is that some of the PCR reactions for validation failed.

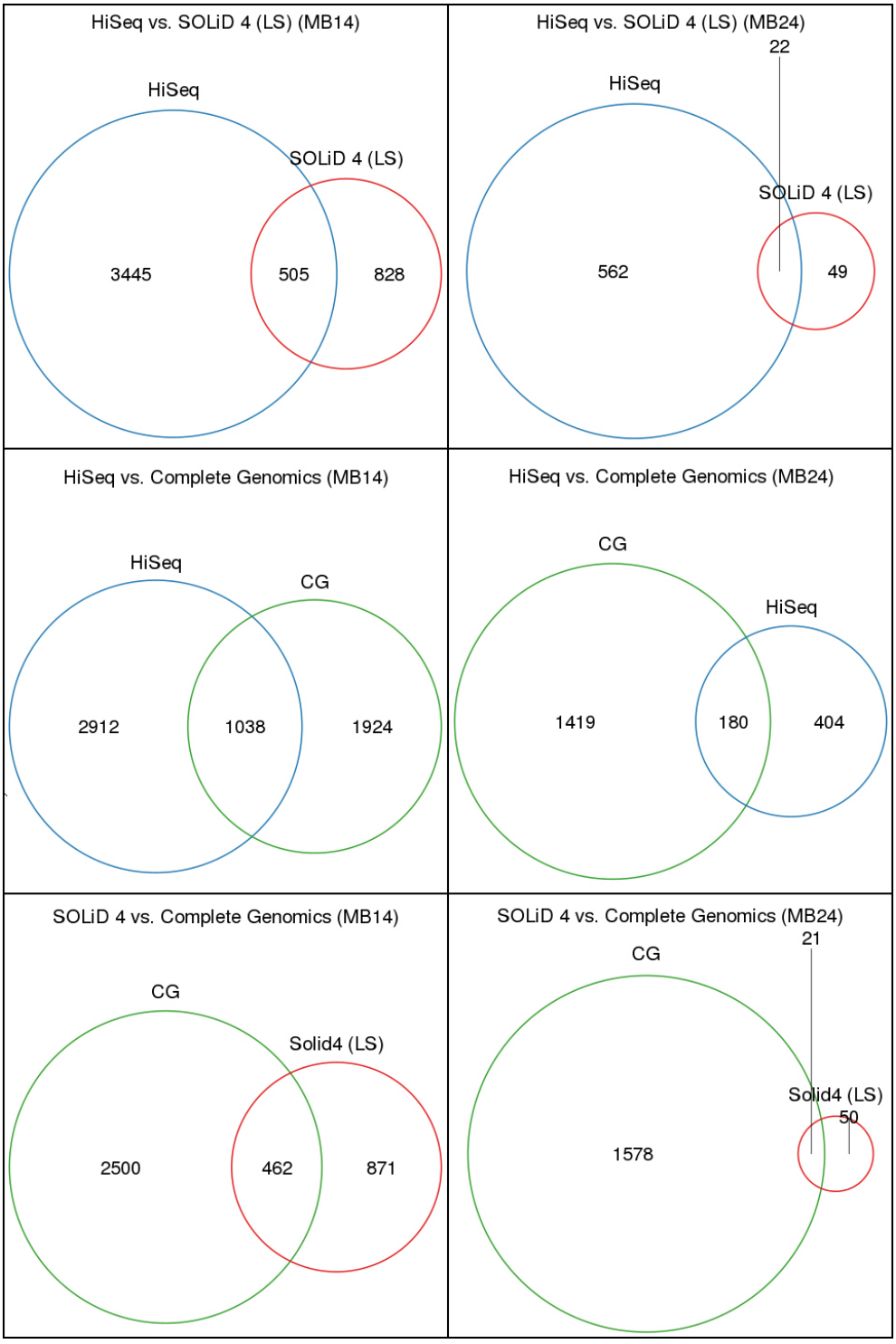

Figure 3.25: Overlap of somatic SNVs called by different technologies. The left-hand side panels stand for patient sample MB14, the right-hand side panels for patient sample MB24.

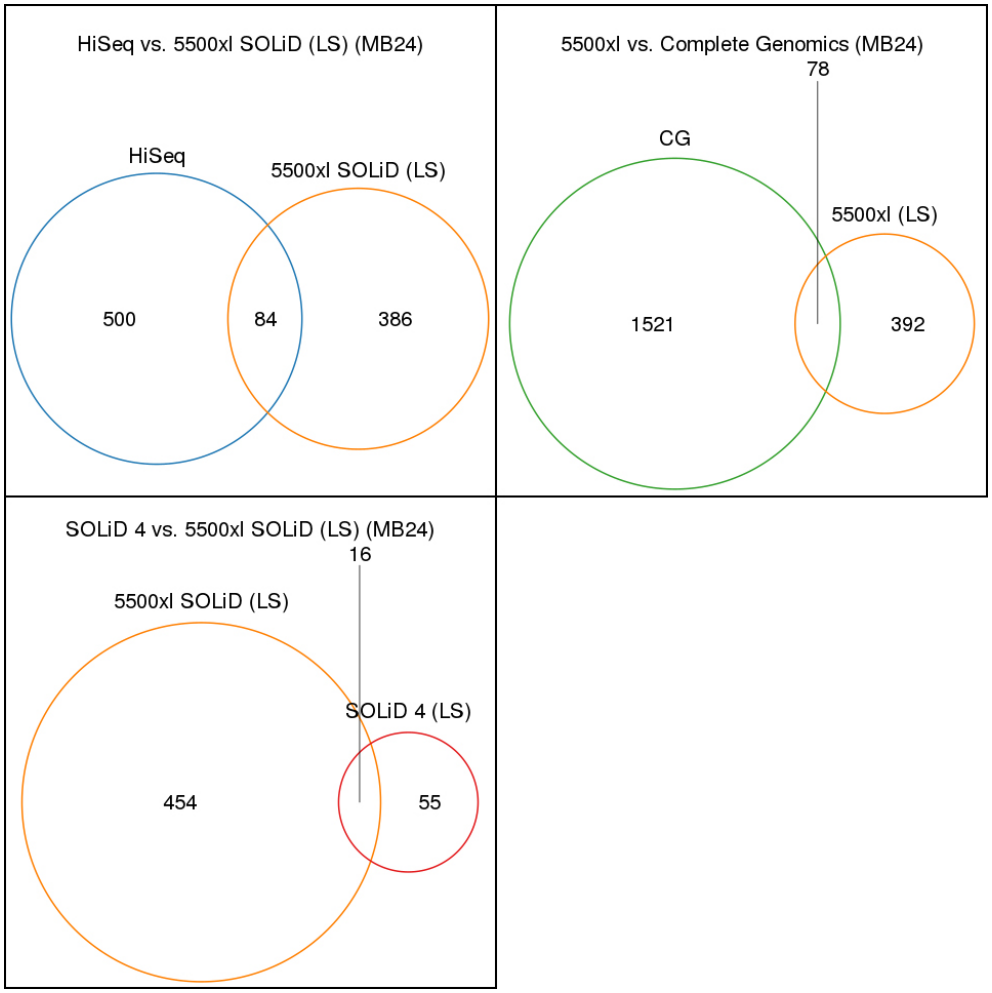

Figure 3.26: Overlap of somatic SNVs called by different technologies for patient sample MB24.

Concordant	9/9 true somatic changes
HiSeq2000 only	2/7 true somatic changes
SOLiD 4 only	0/14 true somatic changes
5500xl SOLiD only	0/11 true somatic changes
Complete Genomics only	1/9 true somatic changes

Table 3.4: Somatic SNV results: number of true somatic changes out of the total number tested, as validated by Sanger sequencing.

The validation results show that we can be confident that a variant called by all platforms is real: all variants that could be tested are true somatic changes. However, variants called by one platform only have disastrous validation rates: only 1 out of 9 variants could be validated for Complete Genomics and only 2 out of 7 for HiSeq2000, while not even a single somatic SNV called by either SOLiD 4 or 5500xl SOLiD was true. If we extrapolate, this would mean that most platform-specific somatic SNV calls are false.

Subsequently, I examined the true and false somatic SNVs in order to understand why these variants were only picked up by one technology.

SOLiD "flanking SNV" artifact

The most obvious cause for false positive somatic SNVs can be seen in Life Technologies' platforms and is a consequence of color space SNV calling (see section 1.1.2). Figure 3.27 illustrates what we call the "flanking SNV" artifact: the false positive somatic SNV is flanked by two SNVs positioned on either side. In Life Technologies' color space, SNVs are discriminated from sequencing errors through dibase encoding: each base is linked to each of its neighbors by a common color code, which results in a certain number of valid dicolor changes relative to the reference, corresponding to SNVs. Any deviation from these changes is considered a sequencing error [22, 132].

However, the event of an SNV followed by a reference base followed by a second SNV, as depicted in Figure 3.27, results not in two, but in four subsequent colors differing from the reference. This means that even though the base in the middle is a reference base, it is represented by two colors differing from the reference (the second and third color changes) and is therefore considered as an SNV. This assumption, in turn, renders the first and the last of the four color changes, which correspond to the flanking real SNVs, invalid: the false call of an SNV between the real SNVs has already reverted

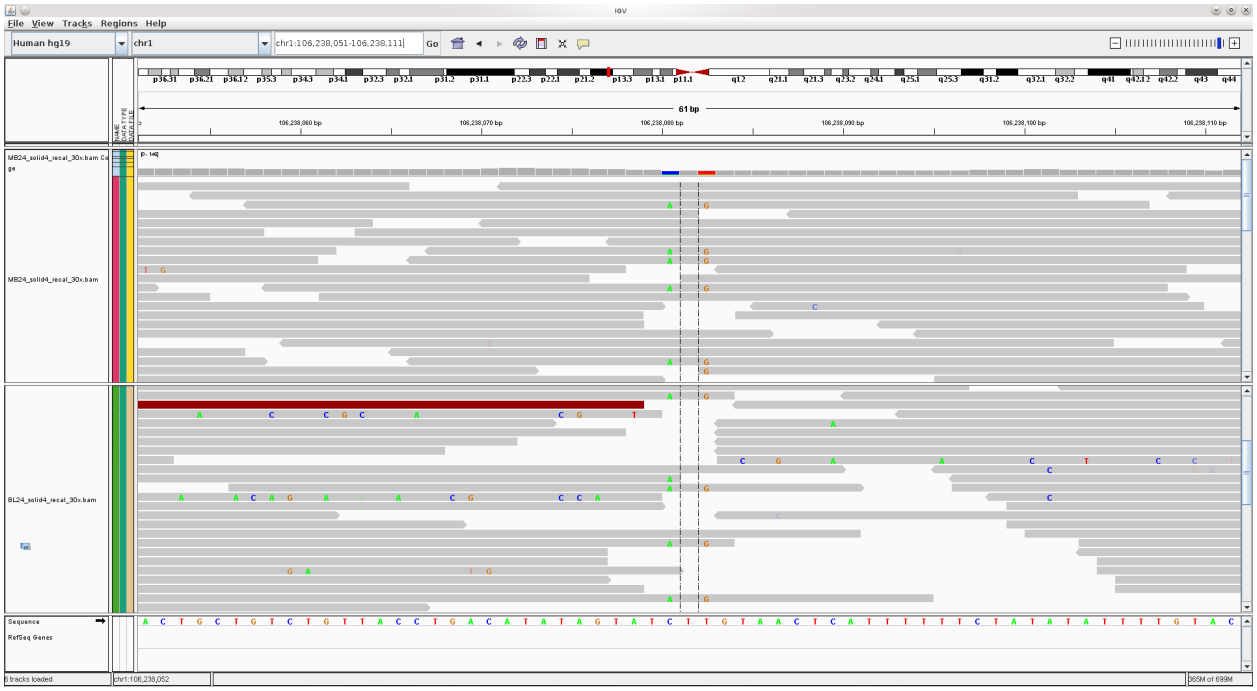

Figure 3.27: An example of the "flanking SNV" artifact seen in Life Technologies' platforms. This is an IGV screen shot for SOLiD 4 and patient sample MB24.

the real SNVs to reference bases.

These false calls in the tumor are not filtered out during somatic SNV calling, as the comparison to the normal sample is based on a pileup in base space and not on color space. Since there is no mutation at the corresponding position in the normal sample, the SNV in the tumor is considered somatic, although it is simply falsely called.

A closer look at the somatic SNVs called in Life Technologies' platforms, beyond those chosen for validation, shows that this is their main source of error. For SOLiD 4, 10 out of 14 somatic SNVs that were Sanger-validated (71,4%) and 21 out of the 31 remaining (67,7%) result from this artifact. For 5500xl, 8 out of 11 somatic SNVs that were Sanger-validated (72,7%) carry the artifact, and out of 27 further somatic SNVs checked manually, 16 carry it (59,3%). This means that roughly 2/3 of the somatic SNVs called in Life Technologies' platforms are false positives generated by a calling issue.

Other technologies

The next question addressed was why the three platform-specific somatic SNVs that were confirmed by Sanger sequencing (1 on Complete Genomics, 2 on HiSeq2000, as listed in Table 3.4) were not picked up by the other technologies.

The two confirmed HiSeq2000-specific variants have a rather low allele frequency (0.2 and 0.22, respectively). They are actually initially identified as somatic by Complete Genomics, but with an extremely low probability score (flagged as SQLOW), which is why they are filtered out. On the SOLiD platforms, the first variant is actually visible in the tumor BAM file on IGV, but is not called, probably because of the low allele frequency. The second variant is not visible in the SOLiD BAMs due to low coverage of the region.

The confirmed Complete Genomics-specific variant is also initially identified in HiSeq2000, but is later filtered out because it lies within a sequence region containing the Illumina-specific error pattern GGC [34]. On the SOLiD platforms, the variant is visible in the tumor BAM on IGV, but is not called, which may be due to the low allele frequency (0.24).

In summary, the three Sanger-confirmed platform-specific somatic SNV calls are generally hard to call because of low allele frequency, or because the region they fall in which may not be covered well enough in SOLiD platforms, or contains a known Illumina error pattern.

A closer look at the false somatic SNVs assessed by Sanger sequencing also reveals miscellaneous and sometimes unclear reasons for their call. Very often, the SNVs are located in regions that are difficult to map to, like poly-A stretches or segmental duplications. Another reason for falsely called somatic SNVs are mutations that are also present in control, but are not called because the coverage or the allele frequency are too low. Finally, a few cases seem to be sequencing errors, where a mutation can clearly be seen in the tumor BAM file on IGV, even though the Sanger validation could not identify any.

chr	pos	dbSNP	ref	alt	genomic regions	hom/ het	validation result
chr1	172157950	-	A	G	Introns Mammal conservation Promoters LINE repeats	het	confirmed
chr1	237345250	-	C	T	Introns Mammal conservation	het	confirmed
chr3	7589499	-	C	T	Introns Mammal conservation	het	confirmed
chr3	14882459	rs116407913	G	A	Introns Mammal conservation	het	sequencing failed
chr3	106758969	-	A	G	Mammal conservation LINE repeats	het	sequencing failed
chr4	594999	-	G	C	Mammal conservation LINE repeats Self chain	het	confirmed
chr6	108328601	-	C	T	Mammal conservation Promoters LTR Repeats	het	sequencing failed
chr7	109072602	-	T	A	Mammal conservation LINE Repeats	het	sequencing failed
chr8	27427015	-	G	A	Mammal conservation	het	confirmed
chr9	9216163	-	C	T	Introns Mammal conservation SINE Repeats	het	confirmed
chr12	28477282	-	A	G	Introns Mammal conservation Promoters	het	confirmed
chr15	39798057	-	C	A	Mammal conservation	het	confirmed
chr16	9975836	-	G	A	Introns Mammal conservation	het	confirmed

Table 3.5: Overview of the somatic SNVs concordant between all four platforms and their Sanger sequencing validation results. For each somatic SNV the chromosome, position, dbSNP identification (if present), reference and alternative allele, genomic regions, zygosity and Sanger validation result is given.

chr	pos	dbSNP	ref	alt	genomic regions	hom/ het	validation result	alt. allele frequency
chr1	43754699	-	G	T	Mammal conservation Promoters LINE repeats	het	TRUE	0.22
chr2	114188440	-	C	T	Introns Mammal conservation Promoters LTR repeats Segmental duplication Self chain	hom	FN	0.83
chr4	23622862	-	A	C	Simple repeats	hom	FP	1
chr6	33747921	-	G	A	Exons Mammal conservation Promoters	het	TRUE	0.2
chr9	68380340	rs77836632	G	T	CpG islands Mammal conservation Segmental duplications Self chain	het	FP	0.24
chr19	4512945	rs75031432	C	T	Exons Mammal conservation	het	FN	0.27
chr19	4512946	rs79662071	A	G	Exons Mammal conservation	het	FN	0.27

Table 3.6: Overview of the HiSeq2000 somatic SNVs chosen for Sanger validation, and their validation results. For each somatic SNV the chromosome, position, dbSNP identification (if present), reference and alternative allele, genomic regions, zygosity, Sanger validation result, and alternative allele frequency is given. "False positive" (FP) stands for an SNV that was not found in the tumor with Sanger sequencing, "false negative" (FN) for an SNV that is found both in tumor and control with Sanger sequencing.

chr	pos	dbSNP	ref	alt	genomic regions	hom/ het	validation result	alt. allele frequency
chr1	106238081	-	T	G	Mammal conservation LINE repeats	het	Flanking SNV artifact	0.46
chr3	19041671	rs112827577	T	A	Mammal conservation	het	Flanking SNV artifact	0.32
chr5	11008475	rs113518426	T	C	Introns Mammal conservation Promoters LTR repeats	het	Flanking SNV artifact	0.44
chr6	29819600	-	G	A	CGI shores Mammal conservation Self chain	het	Flanking SNV artifact	0.24
chr7	53857155	-	G	A	Mammal conservation	het	Flanking SNV artifact	0.47
chr7	152104381	rs113774675	A	G	Introns Mammal conservation Self chain	het	Flanking SNV artifact	0.4
chr9	6932312	rs112868579	G	A	Introns Mammal conservation	het	Flanking SNV artifact	0.33
chr12	31826443	-	G	A	Introns Mammal conservation Promoters LINE repeats	het	Flanking SNV artifact	0.36
chr13	102878420	-	G	A	Introns Mammal conservation	het	Flanking SNV artifact	0.26
chr15	21183913	rs115553418	T	C	Mammal conservation Promoters Segmental duplications Self chain	het	FN	0.36
chr15	85114547	rs12899953	T	C	CGI shores Mammal conservation Promoters Segmental duplications Self chain Simple repeats	het	FN	0.5
chr21	9907222	-	T	C	CGI shores Introns Mammal conservation Segmental duplications Self chain	het	FN	0.29
chr22	47384036	rs113412201	A	G	Introns Mammal conservation DNA repeats	het	Flanking SNV artifact	0.28
chr22	49544337	rs6009515	T	A	Mammal conservation Self chain	het	FP	0.43

Table 3.7: Overview of the SOLiD 4 somatic SNVs chosen for Sanger validation, and their validation results. For each somatic SNV the chromosome, position, dbSNP identification (if present), reference and alternative allele, genomic regions, zygosity, Sanger validation result, and alternative allele frequency is given. 'False positive' stands for an SNV that was not found in the tumor with Sanger sequencing, 'false negative' for an SNV that is found both in tumor and control with Sanger sequencing.

chr	pos	dbSNP	ref	alt	genomic regions	hom/ het	validation result	alt. allele frequency
chr1	16842448	rs3982351	G	A	CGI shores Mammal conservation Segmental duplications Self Chain	het	FP	0.25
chr3	10168966	rs113587531	T	C	Mammal conservation Promoters	het	Flanking SNV artifact	0.32
chr8	69145468	rs34931257	A	G	Mammal conservation Promoters	het	Flanking SNV artifact	0.53
chr8	125315212	-	A	T	CGI shores Mammal conservation Segmental duplications Self chain	het	Flanking SNV artifact	0.45
chr10	105356443	-	C	G	Introns Mammal conservation Promoters Simple repeats	het	Flanking SNV artifact	0.32
chr16	33942851	rs76303220	C	A	CpG islands Mammal conservation Segmental duplications Self chain	het	FP	0.38
chr16	78921414	-	A	G	Introns Mammal conservation Simple repeats	het	Flanking SNV artifact	0.54
chr17	31963897	rs111564603	C	A	Introns Mammal conservation Self chain Simple repeats	het	Flanking SNV artifact	0.42
chr17	76253591	-	G	A	Mammal conservation Promoters Simple repeats	het	FN	0.55
chr19	52099247	-	C	T	CGI shores Mammal conservation Promoters LINE repeats Self chain	het	Flanking SNV artifact	0.29
chr20	6065730	rs112086590	T	C	Exons Mammal conservation	het	Flanking SNV artifact	0.4

Table 3.8: Overview of the 5500xl SOLiD somatic SNVs chosen for Sanger validation, and their validation results. For each somatic SNV the chromosome, position, dbSNP identification (if present), reference and alternative allele, genomic regions, zygosity, Sanger validation result, and alternative allele frequency is given. 'False positive' stands for an SNV that was not found in the tumor with Sanger sequencing, 'false negative' for an SNV that is found both in tumor and control with Sanger sequencing.

chr	pos	dbSNP	ref	alt	genomic regions	hom/ het	validation result	alt. allele frequency
chr1	17322619	-	A	G	Exons Mammal conservation	het	FP	0.21
chr1	36907562	-	C	A	Introns Mammal conservation Promoters SINE repeats	het	FP	0.21
chr1	47076668	-	A	C	CGI shores Introns Mammal conservation	het	FP	0.21
chr3	100170628	-	A	G	Exons Mammal conservation	het	FN	0.2
chr4	166199389	-	T	A	Exons Mammal conservation Segmental duplications Self chain	het	FN	0.34
chr10	129905945	-	T	G	Exons Mammal conservation Self chain	het	FP	0.21
chr15	102226182	-	C	T	Exons Mammal conservation	het	TRUE	0.24
chr17	16876781	-	G	A	Mammal conservation Promoters LINE repeats Self chain	het	FP	0.27
chr20	61595238	-	T	G	Introns Mammal conservation Promoters	het	FP	0.21

Table 3.9: Overview of the Complete Genomics somatic SNVs chosen for Sanger validation, and their validation results. For each somatic SNV the chromosome, position, dbSNP identification (if present), reference and alternative allele, genomic regions, zygosity, Sanger validation result, and alternative allele frequency is given. 'False positive' stands for an SNV that was not found in the tumor with Sanger sequencing, 'false negative' for an SNV that is found both in tumor and control with Sanger sequencing.

3.7.2 Somatic indels

The overall number of somatic indels (Table 3.10) shows that, similarly to the somatic SNVs, the differences between platforms are pronounced. Most variations are called in HiSeq2000 (966), followed by 5500xl SOLiD (630), Complete Genomics (457) and finally SOLiD 4 (37). However, the ratio between insertions and deletions is rather similar for HiSeq2000, 5500xl SOLiD and SOLiD 4, ranging from 0.56 for HiSeq2000 to 0.72 for 5500xl SOLiD, which is consistent with the fact that insertions are generally harder to call than deletions [129]. For Complete Genomics however, this trend is reversed: for 333 insertions called in patient sample MB24, only 124 deletions are found.

The overlap between platforms, both for insertions and for deletions, is even lower than for somatic SNVs (Figure 3.28). Generally, SOLiD 4 seems to have absolutely no indels in common with the other platforms, while only one deletion and zero insertions are shared by the three other platforms. The number of somatic insertions or deletions shared by two platforms is slightly higher (e.g., 4 deletions and 18 insertions for Complete Genomics and HiSeq2000, and 3 deletions and 4 insertions for Hiseq2000 and 5500xl SOLiD), but still represents only an extremely tiny fraction of the total number of indels per platform.

		HiSeq2000	**SOLiD 4**	**5500xl SOLiD**	**CG**
MB24	**ins**	346	15	265	333
	del	620	22	365	124

Table 3.10: Number of somatic indels called for patient sample MB24, for each of the four technologies. "ins" stands for insertion, "del" for deletion. CG stands for Complete Genomics.

Comparison to validation experiments

We had the opportunity to have a small number of somatic indels from patient sample MB24 externally validated with Sanger sequencing. The 5 indels that were identified on more than one platform were selected for validation (see Table 3.15), as well as 7 to 10 indels called by one technology only (Tables 3.12 (HiSeq2000), 3.13 (5500xl SOLiD), and 3.14 (Complete Genomics)). An overview of the results of the Sanger validation can be found in Table 3.11.

As expected, indels called on at least two different technologies are real, whereas not a single one of those called by a single platform could be confirmed by Sanger sequencing.

This suggests that indel calling is not very accurate, regardless of technology, and that somatic indel calling is an even bigger challenge. Indeed, for HiSeq2000, SOLiD 4, and 5500xl SOLiD, the majority of platform-specific indels assessed by Sanger sequencing are false negatives, meaning the indel was found in control, making it a germline indel. This is different for Complete Genomics, which has a majority of false positive indels among those assessed by Sanger sequencing, i.e. of indels found neither in the tumor nor in the control. This indicates that the predominant difficulty lies in the calling of somatic indels for all platforms except Complete Genomics, which seems to call many false positive indels overall.

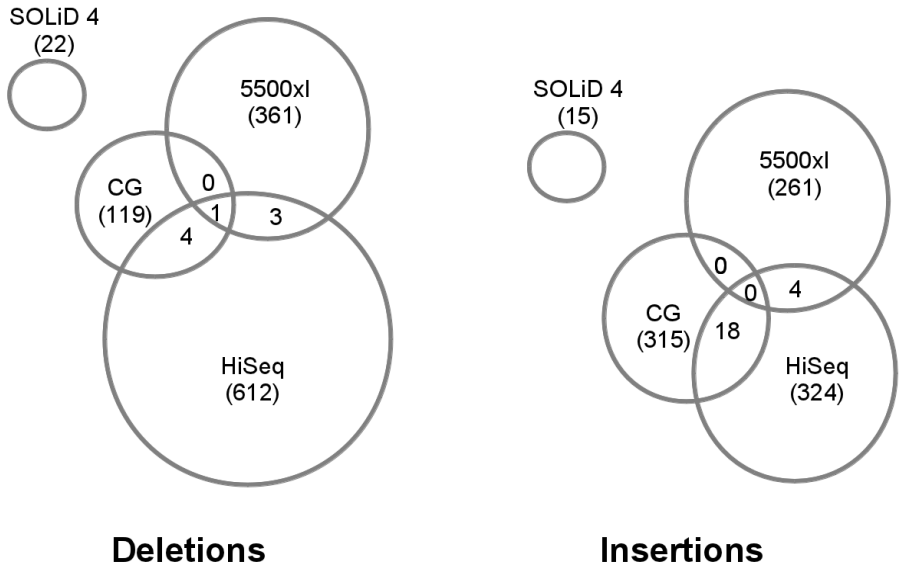

Figure 3.28: Overlap of somatic insertions and deletions found across platforms for patient sample MB24. The plot includes platform-specific somatic SNVs called as germline by other platforms.

HiSeq2000 + 5500xl SOLiD + CG	1/1 true somatic changes
HiSeq2000 + 5500xl SOLiD	2/2 true somatic changes
HiSeq2000 + CG	2/2 true somatic changes
HiSeq2000 only	0/9 true somatic changes
5500xl SOLiD only	0/7 true somatic changes
CG only	0/10 true somatic changes

Table 3.11: Somatic indels results: number of true somatic changes out of the total number tested, as validated by Sanger sequencing. CG stands for Complete Genomics.

chr	pos	ref	alt	allele frequency	validation result
chr1	26895673	AAG	A	0,2	False negative
chr1	54429616	G	GC	0,4	False negative
chr2	96657159	C	CG	0,364	False negative
chr2	201929338	ATTAT	A	0,111	False negative
chr4	171115658	AC	A	0,136	False negative
chr10	127595259	G	GA	0,162	False positive, or allele freq too low
chr11	20499360	CA	C	0,3	False negative
chr15	77246836	GGA	G	0,125	False negative
chr18	15322955	ATTTATAC	A	0,091	False positive

Table 3.12: Overview of the HiSeq2000 somatic indels chosen for Sanger validation, and their validation results. For each somatic indel the chromosome, position, reference and alternative allele, allele frequency and Sanger validation result is given. 'False positive' stands for an indel that was not found in the tumor with Sanger sequencing, 'false negative' for an indel that is found both in tumor and control with Sanger sequencing.

chr	pos	ref	alt	allele frequency	validation result
chr1	143475780	ACCAATTTTGT	ACCAATTTGT	0,545	False negative
chr2	132776038	AAGACTCTAG	AAGACTAG	0,333	False negative
chr5	115202417	CTAAGAGA	CTGA	0,057	False positive, or allele freq too low
chr7	8010576	TA	TAT	0,071	False negative
chr15	20083350	AC	ACTTAC	0,116	False negative
chr15	20942204	TTCAACCATACAT	TTCAACAT	0,389	False negative
chr21	31123516	TATG	TATGTG	0,077	False positive, or allele freq too low

Table 3.13: Overview of the 5500xl SOLiD somatic indels chosen for Sanger validation, and their validation results. For each somatic indel the chromosome, position, reference and alternative allele, allele frequency and Sanger validation result is given. 'False positive' stands for an indel that was not found in the tumor with Sanger sequencing, 'false negative' for an indel that is found both in tumor and control with Sanger sequencing.

chr	pos	ref	alt	allele frequency	validation result
chr1	247104579	CAAA	C	0,308	False negative
chr2	228593117	TGCG	T	0,291	False positive
chr5	3125712	T	TTGTTGTTGTTTGTTTCTTTTTGTTGTTGA	0,625	False positive
chr6	160956039	A	ATGATGGTGGTAG	0,292	False positive
chr10	134326339	A	ACTGAGCCCTACTCTCATCCCCAACGACCCAGG	0,731	False positive
chr14	24522271	GCA	G	0,333	False negative
chr16	6773733	A	AT	0,258	False negative
chr19	5852410	CT	C	0,187	False positive, or allele freq too low
chr22	16202448	A	AC	0,346	False positive
chr22	48930058	A	ACATGAATAAGTCAGGTGC GGTGGGTCAGGCAGGTGC	0,433	False positive

Table 3.14: Overview of the Complete Genomics somatic indels chosen for Sanger validation, and their validation results. For each somatic indel the chromosome, position, reference and alternative allele, allele frequency and Sanger validation result is given. 'False positive' stands for an indel that was not found in the tumor with Sanger sequencing, 'false negative' for an indel that is found both in tumor and control with Sanger sequencing.

chr	pos	ref	alt	allele frequency	validation result	platforms found in
chr5	39297910	GA	G	0,147	Confirmed	HiSeq2000, 5500xl SOLiD
chr11	123375285	TC	T	0,143	Confirmed	HiSeq2000, 5500xl SOLiD
chr3	967815	TC	T	0,171	Confirmed	HiSeq2000, Complete Genomics
chr4	82373501	TG	T	0,129	Confirmed	HiSeq2000, Complete Genomics
chr18	730060	GTTTAA	G	0,176	Confirmed	HiSeq2000, 5500xl SOLiD, Complete Genomics

Table 3.15: Overview of the somatic indels found by several platforms that were chosen for Sanger validation, and their validation results. For each somatic indel the chromosome, position, reference and alternative allele, allele frequency, Sanger validation result and platforms found on is given. 'False positive' stands for an indel that was not found in the tumor with Sanger sequencing, 'false negative' for an indel that is found both in tumor and control with Sanger sequencing.

4 Discussion

The recent advent of massively parallel sequencing technologies has paved the way to a better assessment and thus a better understanding of genomics and genetics. Once an extremely costly and time-intensive task, large-scale human genome sequencing is only a few steps away from becoming a routine laboratory task. However, the state-of-the-art platforms currently available are still relatively new and thus not thoroughly evaluated, and there are a number of caveats to be considered and uncovered.

The goal of this thesis is a comparison of the four whole-genome sequencing instruments currently established on the market: Illumina's HiSeq2000, Life Technologies' SOLiD 4 and 5500xl SOLiD, and Complete Genomics' technology. To this end, I used four whole-genome samples (two tumor-normal pairs) from two pediatric medulloblastoma patients, sequenced once on each of the four platforms. For each sequencing machine, I presented an extensive assessment of coverage distribution and bias, in particular GC bias, a comparison of SNV calls with regard to a SNP array gold standard, as well as an assessment of the potential benefits of combining mapped reads from different technologies. Additionally, somatic mutation calls (SNVs and indels) from different platforms were evaluated. This study highlights the advantages and drawbacks of the individual platforms while considerably extending previous comparative studies, as it includes all four platforms, uses a more comprehensive approach, and assesses multiple samples instead of one, allowing more compelling results.

4.1 Coverage assessment

Coverage, i.e. the redundancy with which each base in the sequenced genome is covered, should ideally be evenly distributed across the genome. Instead it fluctuates substantially due to different factors [133], one of the most common being the underlying base or sequence composition. An even and sufficient coverage of all genomic regions is crucial for reliable downstream analysis for a number of reasons. First, it minimizes the impact of sequencing errors: unless it is a systematic error due, for example, to a specific sequence pattern [34], the same sequencing error is highly unlikely to appear at the same genomic position in many different reads

(although the overall number of sequencing errors, assuming a constant error rate, increases with coverage). As a consequence, the higher the coverage, the higher is the confidence in the results of a downstream analysis. A mean coverage of 30x is currently agreed as the standard for whole-genome sequencing (see section 1.3.2).

In addition, sufficient coverage is essential for accurately calling variants with low allele frequency. These are extremely common in cancer samples, due to tumor heterogeneity, tissue heterogeneity, and copy number variations (see section 1.2.3). Localized lack of coverage or extremely low coverage, regardless of the reason, may lead to the downstream analysis missing or erroneously reporting important variants, while fluctuating coverage is an issue particularly in quantitative sequencing experiments like RNA-seq, ChIP-seq, or CNV-seq. Coverage biases are often found in specific genomic regions of interest, like in GC-rich or GC-poor regions, which are often located in or near genes and gene promoters, or in repeat regions which may have a number of different functions, as explained in section 1.1.2.

My results show that GC bias, i.e. a significant drop in coverage in GC-rich, but also GC-poor regions, is present in all samples and for all platforms analyzed, but is most pronounced for SOLiD 4 and 5500xl SOLiD instruments, in particular in regions with a GC content above 60%. A slightly better coverage of GC-rich regions can be observed with HiSeq2000, but the least (although present) GC bias is achieved with Complete Genomics. This is in contrast to GC-poor regions ($\leq$ 25% GC content), where Complete Genomics shows the highest deviation from a uniform distribution, together with SOLiD 4, while HiSeq2000 and 5500xl SOLiD exhibit a substantially better coverage. This confirms results from earlier studies [38, 117] which clearly show the presence of a GC bias in several next-generation sequencing platforms, notably in data from Illumina's GAII instrument, but stands in contrast to the results from Suzuki *et al.* [116] who claim finding "no striking GC bias" for SOLiD sequencers.

The overall coverage distribution proved to be surprisingly diverse, with HiSeq2000 showing the narrowest peak, i.e. the most even distribution. All three other platforms, when observed at comparable coverage levels, display a relatively high number of bases with very low (< 5) coverage. In addition, between-sample variation for Complete Genomics was shown to be higher than for other technologies, even at similar mean coverage, therefore hampering a comparison of variant calls between samples, as variant call sensitivity is related to coverage. Also, up to a cumulative coverage of around 40x, Complete Genomics covers a smaller genome fraction compared with HiSeq2000 at the same mean coverage. This is consistent with Lam *et al.*'s remark that the less uniform

the coverage, the higher mean coverage is needed to reach a certain read depth level for most of the genome [115].

There is also major variation between sequencers in the uncovered fraction of specific genomic and functional regions, the most striking being in CpG islands. Consistent with the GC bias mentioned above, both SOLiD 4 and 5500xl SOLiD have major difficulties covering CpG islands, leaving over half of them uncovered. This offers a straightforward explanation for the major difficulties encountered in the analysis of data from methylation experiments described in section 1.4, as these were conducted on the SOLiD 4 platform. Also, a large part of CpG islands is covered with less than 3 reads for both HiSeq2000 and the SOLiD sequencers, although HiSeq2000 performs better in this respect than its predecessor, Illumina's GAII [117]. The good performance of Complete Genomics might seem counter-intuitive, as CpG islands contain a high number of repeated sequences, and Complete Genomics' read length is by far the shortest of all four platforms. However, as exemplified by Benjamini and Speed [38], the GC bias is dependent on the GC content of the full fragment and not just the sequenced read. Complete Genomics' complex protocol uses short fragmented sequences interspersed with adapter sequences, therefore lowering the GC content of the fragment and making it less susceptible to GC bias.

Generally, Complete Genomics, even at downsampled 30x coverage, shows the smallest uncovered fractions for most genomic regions considered, although it is usually closely followed by HiSeq2000. Its weakness lies in the coverage of short repeats, in particular simple repeats, which is most likely due to the shortness and split structure of the reads, as the other technologies perform better in this respect. This also fits with the observation of Lam *et al.* [115] that, compared to concordant SNVs, a high fraction of platform-specific SNVs are located within simple repeats and low complexity repeats, suggesting that these false positive calls are due to mapping difficulties.

Overall, it is the SOLiD 4 and 5500xl platforms which display the most shortcomings, SOLiD 4 even more than 5500xl, often leaving considerable fractions of genomic elements not covered. Exon coverage, which is of paramount importance in sequencing, is particularly problematic in this respect, with around 10-12% of uncovered bases for the SOLiD sequencers.

A combination of mapped reads from two different sequencers, as proposed by Nothnagel *et al.* as an attempt to lower the false positive rate for SNV calling [118], showed limited improvement in the uncovered fractions of problematic regions. The attempt to combine

the strengths of two platforms in this respect proved to be useful for only few of the genomic regions considered, in particular specific repeat regions.

4.2 Variant detection comparison

SNVs called genome-wide were compared to the SNP calls obtained from Affymetrix arrays, which were used as a gold standard to compute sensitivity and false positive rate. HiSeq2000 consistently reaches the highest sensitivity on all samples, closely followed by Complete Genomics, even though the mean coverage for HiSeq2000 is considerably lower than Complete Genomics' coverage. In fact, even with half the coverage, i.e. 15x, HiSeq reaches a sensitivity very close to Complete Genomics at full coverage. It is interesting to note that Complete Genomics' coverage at the SNP positions assessed is substantially lower than the sample's mean coverage (e.g., 40x vs. 51.7x for patient sample MB24), while this is not the case for HiSeq2000. Still, this does not justify that HiSeq2000 at 15x reaches the sensitivity of Complete Genomics at full coverage. The sensitivity of 5500xl SOLiD, and SOLiD 4 in particular, is far behind the sensitivity reached by HiSeq2000 and Complete Genomics, although the specificity is slightly increased.

The comparison of reliability of concordant and discordant SNV calls conducted by Lam *et al.* on HiSeq2000 and Complete Genomics data [115] shows results consistent with our findings. An earlier study by Suzuki *et al.* [116] mentions similar detection performance for Illumina and SOLiD platforms, but these result relate to data from very early sequencing machines (Illumina's Genome Analyzer and SOLiD's first platform). Altogether, the results denote a strong preference for Complete Genomics and HiSeq2000 in any research setting focusing on sensitive yet specific results. This is especially valid for cancer research which relies on the detection of variants with low allele frequency.

A combination of mapped reads from two different technologies shows that the sensitivity reached by HiSeq2000 is hard to outperform. The sensitivity of 5500xl SOLiD data can be strongly increased by combining 15x coverage from this machine with 15x coverage from HiSeq2000, at the cost of a small loss of specificity, but will not reach the sensitivity of HiSeq2000 alone. Interestingly however, the sensitivity of Complete Genomics calls can be significantly increased by adding as little as one lane (corresponding to around 15x mean coverage) of HiSeq2000 sequencing. This supports the suggestion by Lam *et al.* [115] that combining sequence data from different platform can help boosting their strengths for SNV calling.

4.3 Evaluation of the detection of somatic mutations

As evidenced in section 1.2.3, calling somatic mutations in cancer samples is a challenging task for a number of reasons. The cells of tumor tissues can contain very heterogeneous genomes, the samples are usually admixed with a certain amount of normal tissue, and copy number variation and polyploidy occur frequently, the latter being a common event in pediatric medulloblastoma patients [110]. All this leads to a low allele frequency (i.e., below 50%) for many somatic mutations, meaning that there may be only a handful of reads carrying the mutation of interest, and that these need to be distinguished from sequencing errors. This emphasizes once again the importance of sufficient coverage (section 4.1).

In addition, somatic variation is identified by comparing the sequenced tumor sample to the sequence of a normal, unaffected control sample. This holds an additional source of error, since many false positive calls in the tumor sample will result in a somatic mutation call, as we are unlikely to find the exact same error in control. This means that, regardless of their origin, a large part of the false positive calls within the tumor will be kept when calling somatic SNVs. Because there are usually few somatic mutations in comparison to the total number of mutations called – i.e., germline and somatic –, the fraction of false positive calls increases when moving on to somatic calling. This is particularly evident for the "flanking SNV" artifact I identified within SOLiD data (see section 3.7.1) which affects a drastic 2/3 of SOLiD's somatic SNV calls, but it is to be expected that this factor will also affect the other platforms, although to a lesser extent.

Besides false positive somatic mutation calls, comparison to a control normal tissue also introduces false negative calls, i.e. germline mutations that are picked up in the tumor sample but missed in the control, for example due to a low allele frequency and/or low coverage, low base quality or mapping quality, or strand bias. Taken together, all three factors mentioned above – low allele frequency in cancer samples, false positive calls stemming from sequencing errors, and false negative calls – increase the error rate compared to one-sample SNV calling in a non-cancer context, i.e. without classification into somatic and germline.

The somatic SNV calling results show pronounced differences. The number of somatic SNVs called varies strongly between platforms on the same sample, as does the between-sample ratio from platform to platform. Overall, each platform presents more somatic SNV calls for patient sample MB14 than for patient sample MB24. This is very probably due to the fact that MB24 is tetraploid, and suggests that the ability to

detect variants with low allele frequency differs between platforms.

Both the overlap between results from different platforms, and the Sanger validation rates for somatic SNVs specific to one platform are extremely low, MB14 performing slightly better in this respect than MB24. While somatic SNVs called by all platforms were all validated by Sanger sequencing (9 out of 9 tested), none of the SOLiD 4-specific and 5500xl SOLiD-specific somatic SNVs chosen for Sanger resequencing could be validated (0 out of 14 and 0 out of 11, respectively), and only very few of the somatic SNVs found exclusively with HiSeq2000 or Complete Genomics sequencing were actually verified (2 out of 7 and 1 out of 9, respectively). This suggests that using more than one platform for somatic mutation calling may be beneficial for weeding out false calls and increasing specificity, as a large majority of platform-specific calls seem to be false. This is consistent with Lam *et al.*'s finding on standard mutation calling in a healthy individual that "concordant SNVs have high accuracy and platform-specific SNVs have a high false positive rate" [115], an effect that is intensified for somatic events.

As mentioned above, a major portion of somatic SNV calls detected with SOLiD sequencing are false calls arising through the "flanking SNV" artifact. These make up not only roughly 2/3 of the total somatic calls, but also around 80-100% of the number of non-overlapping calls in pairwise platform comparisons, depending on the instrument used and sample studied. Assuming a low error rate for concordant calls between two platforms, this supports the conclusions from the SNP comparison showing a comparatively low false positive rate for the SOLiD platforms.

A closer look at the three Sanger-validated, true platform-specific variants in Complete Genomics and HiSeq2000 data shows that the main reasons why they were missed in the other platforms are indeed low allele frequency and/or low coverage, the latter being a problem especially for the SOLiD platforms with their more variable coverage distribution. False calls not validated by Sanger sequencing, on the other hand, mostly fall into the categories of "false positive" and "false negative" calls illustrated above, although they are often found in error-prone regions like homopolymer stretches or repeats. It is interesting to note that almost all of the Sanger-assessed false calls on the SOLiD platforms, not considering the "flanking SNV" artifacts, fall into GC-rich regions like promoters, CpG islands, and CpG islands shores, which is consistent with the poor coverage of SOLiD platforms in these regions. The influence of analysis tools on these results is further discussed in section 4.4.

4.3.1 Somatic indel calling

Somatic indel calling results show even less overlap between platforms and an even lower validation rate compared to somatic SNVs. The general issues stated above are valid for all somatic events, but however, indel calling in itself is already an extremely challenging task [129, 134]. Indel calling depends very strongly on the abilities of the mapper to allow gapped alignments, as reads containing indels are more difficult and less straightforward to map than reads containing base errors or SNVs, especially in the case of longer insertions. Very often, reads with indels will be mapped with mismatches instead of a gap, and even with a gapped alignment, the indel may not be placed at its exact location due to, for example, the presence of repetitive elements or insufficient quality. Finally, some instruments, including Illumina's platforms, have difficulties accurately detecting the length of homopolymers [129].

Indel detection is a topic of current research and is far less advanced than SNV calling due to its higher complexity. Solutions to the indel issue include, for example, analyses that do not consider reads independently [134].

4.4 Influence of mapping and detection software

An important point when comparing platforms is acknowledging the influence of the chosen analysis software on the results, as this is what essentially gives meaning to the raw data. At first sight, using the exact same software for every platform and every analysis step may seem like the most straightforward solution for a cross-platform comparison. However, there are a number of caveats to consider in our case. While Illumina platforms essentially follow the previously used consensus of a base-by-base read-out of DNA fragments, both Complete Genomics and the SOLiD platforms introduce new concepts that require adapted handling of the raw data. Both the use of color space instead of base space and the use of fragmented reads with interspersed adapters (see section 1.1.2) hold advantages, like added error correction for SOLiD through two-base encoding, or a better coverage of GC-rich locations for Complete Genomics. However, using the same software for all platforms compared will not only prevent taking advantage of these properties, but will also heavily penalize all platforms but one. As an example, my experience showed that mapping SOLiD data in base space results in about 50% to 65% less mapped reads in comparison to mapping in color space.

For this reason, both for mapping and for variant calling, we used the software tools we consider best adapted for each platform. While we are aware that using different software choices for different sequencing machines will introduce bias, our

understanding is that this bias will be considerably smaller than the bias introduced by using the same options for all sequencers, especially for mapping. Our criteria for choosing the analysis software include comparison of results (e.g., AUC for ROC curves), extensive shared experience of several DKFZ research groups on different types of sequencing data, particularly within ICGC projects [110, 120, 121], personal communication with experts from Life Technologies and Complete Genomics, published algorithm comparisons [104, 105, 117, 135], and commonness of use within the next-generation sequencing community.

For the mapping of reads from Illumina's HiSeq2000 platform, BWA [105] was used for a number of reasons. BWA is one of the most widely used mappers for Illumina data, and also the mapper of choice for ICGC Illumina data. It was shown to have one of the best overall performances for Illumina data [117], has a good trade-off between speed and accuracy [135], and is relatively easy and straightforward to use.

For SOLiD data, the mapping algorithm was required to map in color space, which greatly restricted the available choices. The main options were the LifeScope aligner by Life Technologies[1], NovoalignCS by Novocraft[2], BFAST [136], and SHRiMP [137]. BWA and Bowtie [106], which are sometimes stated as color space compatible mappers, actually dropped their color space support. SHRiMP was rejected as an option due to its extremely long run times [117], as was BFAST, which has issues with the pairing of SOLiD reads[3,4]. NovoalignCS is a commercial mapper and was therefore disregarded. Although Novocraft offers a version that is free for academic and non-commercial use, it does not support multithreading, which makes it rather slow in comparison. LifeScope (previously BioScope) comes with the sequencing machine, and the latest version is free, meaning most SOLiD data users have access to it. It is the most widely used SOLiD mapper and was retained as our mapping choice for the SOLiD machines.

There is currently only one aligner which takes the specific nature of Complete Genomics' data into consideration - the company's own proprietary software. None of the analysis methods routinely used by Complete Genomics are available for use outside the company, and Complete Genomics does not use any of the established standard data file formats. In addition, the processes of mapping and calling variants are very intertwined, as the company uses re-assembly at locations differing from the reference

[1] http://de-de.invitrogen.com/site/de/de/home/Products-and-Services/Applications/Sequencing/Next-Generation-Sequencing/Data-Analysis-Solutions-for-Next-Generation-Sequencing/LifeScope-Genomic-Analysis-Solutions/LifeScope-Genomic-Analysis-Software.html

[2] http://www.novocraft.com/wiki/tiki-index.php?page=NovoalignCS

[3] http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main_Page

[4] http://www.nilshomer.com/index.php?title=BFAST_with_BWA

genome. However, one of Complete Genomics' unique selling points is the inclusion of an extensive analysis of the generated data, a feature that is useful to customers lacking the considerable infrastructure and knowledge needed for downstream analysis, making it even less likely for users to apply their own downstream analysis. For this reason, Complete Genomics' analysis results were used, including mapping and variant calling results. Quality filters were applied to the somatic calls, as advised in personal communication with company experts.

There is a wide range of available Illumina-centered SNV calling algorithms based on different methodologies [101], the most commonly used methods being samtools/bcftools [108] and GATK [122]. As they differ only minimally in methodology, samtools was used for calling SNVs because it was already established in our group. It was used for the comparison to the SNP array for both HiSeq2000 and SOLiD data, as for the latter it improved sensitivity (as computed from the SNP array) by 1% at a hardly altered specificity, when compared to the results obtained with LifeScope.

For calling somatic SNVs, I had access to a very comprehensive and well-tuned pipeline that is routinely used in ICGC projects [110]. Except for Illumina-specific filtering steps, the procedure used after SNV calling for determining the somatic/germline status of each mutation was kept the same for HiSeq2000 and SOLiD data. Preliminary results from benchmarking studies[5] show that different calling procedures will yield very different total numbers of somatic SNVs, sometimes differing by an order of magnitude. This encouraged us to use a setup as similar as possible for HiSeq2000, SOLiD 4, and 5500xl SOLiD data.

Overall however, it is hardly possible to disentangle the influence of analysis software choices from the effects of using a particular sequencing platform, as these are often intricately linked. As an example, lack of sufficient coverage was identified as an important cause of missed calls, but the reasons for lack of coverage are manifold, and are often both an issue of the platform and of the downstream analysis: mapping difficulties can be due to low base or read quality, due to specific sequencing biases, due to the mappability of the region, due to a short read length, or due to too stringent alignment settings.

[5] currently unpublished; talk by Ivo Gut, ICGC meeting, Heidelberg, December 10th, 2012

4.5 Conclusions and outlook

This comparative study reveals the strong and weak points of the sequencing machines and provides an indication of the preferred platform to use, depending on the aims of the experiment. The most striking result is the poor performance of the two SOLiD sequencers in a GC-related setting, disapproving their use especially for methylation assessment and exon sequencing. This also explains the major problems that occurred during the methylation experiments performed with SOLiD 4 machines (see section 1.4). Combining raw data from different technologies proved to be only of limited use and is indicated only for very specific applications that require good coverage of specific genomic regions while retaining high SNP calling sensitivity. The latter proved to vary strongly between platforms and once again places the SOLiD platforms, in particular SOLiD 4, at the bottom end. The assessment of somatic SNVs showed that calling somatic mutations is still a big challenge for many different reasons. It requires high, and preferably evenly distributed coverage throughout the genome in order to ease the discovery of mutations with low allele frequency. Using the calls from several platforms was shown to increase the certitude of a true somatic call. Comprehensive benchmarking experiments are needed for a better understanding of the issues raised by somatic calling, a task that is beginning to be tackled by large-scale sequencing consortia [138].

Laboratory parameters like the required amount of starting DNA, sequencing costs, or turnaround time, were not considered in this comparison. Although they might be decisive for choosing a particular sequencing platform, they tend to vary strongly over time. However, it is worth pointing out that HiSeq2000 currently has by far the fastest turnaround time among the sequencing machines considered, and that Complete Genomics requires a comparatively high amount of starting material, prohibiting its use for experiments with a very limited DNA amount available. Furthermore, there is a lack of adapted downstream analysis algorithms for Complete Genomics' data, except for their own proprietary software, and while the company's analysis service is useful for smaller labs without bioinformatics support or strong computational infrastructure, it is not flexible enough in other cases and does not match with the requirements of a comprehensive analysis.

Different methods for downstream analysis were explicitly not reviewed here, as this has already been done, and instead the algorithms best adapted to each platform were chosen in order to give consideration to the heterogeneity of the data generated by the different instruments. While a larger sample size per platform would have given more

statistical power, this is currently difficult to achieve because of budgetary issues.

New, "third" generation platforms currently emerge, which confirms the trend towards platforms generating longer reads, a development that will help circumvent many of the issues encountered with current next-generation sequencing instruments. Once a sufficient throughput can be obtained, they will compete with current technology. After the launch of the Ion Proton sequencer, Life Technologies recently promised the "1000$ genome" in the foreseeable future [139]. Pacific Biosciences is developing the promising SMRT (single-molecule real-time) sequencing, a technology which offers much longer reads and less bias, as well as modified base detection (see section 1.1.2). The throughput is currently too low for human-sized genomes and the error rate is still high, but these issues are expected to be resolved in the future, opening the field for a more comprehensive assessment of human genetics and genomics.

5 Technical annex

The major scripts used for this thesis are included on a CD. A description can be found below.

Coverage analysis

- `coverageQc.py`
  determines genome-wide coverage
- `mpileup.sh`
  runs samtools mpileup per chromosome, giving coverage for each covered genomic position
- `mpileup_coverage_distrib.sh`
  concatenates mpileup single-chromosome files and computes coverage distribution using mpileup files
- `sortn_uniqc.py`
  used by `mpileup_coverage_distrib.sh` for coverage computation
- `genome_coverage_plots.R`
  plots coverage distributions
- `generate_bed_of_uncovered_regions_refgenome.pl`
  computes coordinates of Ns in reference genome
  (used for `generate_bed_of_uncovered_regions.pl` and `generate_bed_of_uncovered_regions_0-2.pl`)
- `generate_bed_of_uncovered_regions.pl`
  generates, for each technology, a BED file of uncovered regions
- `generate_bed_of_uncovered_regions_0-2.pl`
  generates, for each technology, a BED file of uncovered regions (meaning, in this case, base coverage < 3)
- `my.genomehg19`
  chromosome sizes for reference genome version HG19; used by the scripts above
- `run_compute_uncovered_regions_overlap_with_bed_files_30x.sh`
  starts `compute_uncovered_regions_overlap_with_bed_files.sh`

- `compute_uncovered_regions_overlap_with_bed_files.sh`
  computes the overlap of uncovered regions with genomic and functional regions
- `compute_BED_bp_size.pl`
  computes the size of the overlaps (used for `compute_uncovered_regions_overlap_with_bed_files.sh`)
- `compute_uncovered_regions_overlap_with_bed_files_cleanupFiles.sh`
  concatenates results from `compute_uncovered_regions_overlap_with_bed_files.sh`
- `xls_plot_with_R.R`
  plots, for each platform and for each genomic/functional region, the percentage of the region that is not covered
- `xls_plot_with_R_functions.R`
  functions used in `xls_plot_with_R.R`
- `uncovered_regions_stats_30x.R`
  plots the size distribution of the regions not covered for each technology
- `plot_vsGC_tumorcontrol.R`
  plots the GC bias

SNP analysis

- `ROC_functions.r`
  identifies concordance between array and the different platforms, computes and plots ROC values plots

Somatic analysis

- `CG_somaticSNVsToVcf.pl`
  conversion of Complete Genomics' somaticVcfBeta file format to generic vcf file format
- `CG_extract_somatic_SNVs.pl`
  extracts somatic SNV calls from Complete Genomics calls
- `LT_filter_SNPs_gff3.pl` and `LT_diBayesSNP_GFF3_2vcf.pl`
  filtering and conversion from gff3 to vcf file format for LifeScope SNV calls
- `LT_split_converted_vcf_per_chr.pl`
  splits up LifeScope SNV calls by chromosome
- `./natalie_snvcalling/` and `./matthias_snvcalling/`
  somatic SNV calling pipeline

- `CG_extract_germline_SNVs.pl`
  extracts germline SNVs for Complete Genomics
- `filter_somaticSNVs_by_germlinelist_alltechs.pl`
  filters somatic SNVs (removes all those that are called as germline in any of the technologies)
- `overlap_concordant_discordant_snvs_2ndversion.pl`
- `overlap_concordant_discordant_snvs_all4techs.pl`
  computes somatic SNV overlaps between platforms
- `vennDiag_concordant_discordant.R`
  plots Venn diagrams of the somatic SNV overlaps between platforms
- `discordant_somSNVs_only_1_tech.pl`
  computes somatic SNVs called by only one platform
- `overlap_concordant_discordant_snvs_withfctregions_2ndversion.pl`
  annotate the somatic SNVs with functional regions

Bibliography

[1] Watson J and Crick FH. Molecular structure of nucleid acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[2] Sanger F, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. *P Natl Acad Sci USA*, 74:5463–5467, 1977.

[3] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, and Smith M. Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature*, 265:687–695, 1977.

[4] Maxam AM and Gilbert W. A new method for sequencing DNA. *P Natl Acad Sci USA*, 74:560–564, 1977.

[5] Pesole G and Saccone C. *Handbook of comparative genomics: principles and methodology*. Wiley-Liss, New York, 2003.

[6] Mardis ER. Anticipating the 1,000 dollar genome. *Genome Biol*, 7:112, 2006.

[7] Slatko BE, Albright LM, Tabor S, and Ju J. DNA sequencing by the dideoxy method, Chapter 7: Unit 7.4a. In *Curr Protoc Mol Biol*. 2001.

[8] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, M Doyle, and FitzHugh W *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[9] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, and Holt RA *et al.* The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[10] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

[11] Lander ES. Initial impact of the sequencing of the human genome. *Nature*, 470:187–197, 2011.

[12] Marx V. Biology: The big challenges of big data. *Nature*, 498:255–260, 2013.

[13] Nagy B, Farmer JD, Bui QM, and Trancik JE. Statistical basis for predicting technological progress. *PLoS One*, 8:e52669, 2013.

[14] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, and Chen Z *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[15] Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, and Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–1732, 2005.

[16] Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods*, 5:16–18, 2008.

[17] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, and Bignell HR *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[18] McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, and Lee CC *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, 19:1527–1541, 2009.

[19] Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, and Zhang J *et al.* The diploid genome sequence of an Asian individual. *Nature*, 456:60–65, 2008.

[20] Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, and Yeung G *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327:78–81, 2010.

[21] Glusman G. Clinical applications of sequencing take center stage. *Genome Biol*, 14:303, 2013.

[22] Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[23] Reid C. Company profile: Complete Genomics Inc. *Future Oncol*, 7:219–221, 2011.

[24] Dressman D, Yan H, Traverso G, Kinzler KW, and Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *P Natl Acad Sci USA*, 100:8817–8822, 2003.

[25] Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 24:133–141, 2008.

[26] Nagarajan N and Pop M. Sequence assembly demystified. *Nat Rev Genet*, 14:157–167, 2013.

[27] de Koning AP, Gu W, Castoe TA, Batzer MA, and Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*, 7:e1002384, 2011.

[28] Jurka J, Kapitonov VV, Kohany O, and Jurka MV. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet*, 8:241–259, 2007.

[29] Britten RJ. Transposable element insertions have strongly affected human evolution. *P Natl Acad Sci USA*, 107:19945–19948, 2010.

[30] López Castel A, Cleary JD, and Pearson CE. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol*, 11:165–170, 2010.

[31] Orr HT. Unstable nucleotide repeat minireview series: a molecular biography of unstable repeat disorders. *J Biol Chem*, 284:7405, 2009.

[32] Albrecht A and Mundlos S. The other trinucleotide repeat: polyalanine expansion disorders. *Curr Opin Genet Dev*, 15:285–293, 2005.

[33] Gymrek M, Golan D, Rosset S, and Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*, 22:1154–1162, 2012.

[34] Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, and Takahashi H *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*, 39:e90, 2011.

[35] Erlich Y, Mitra PP, delaBastide M, McCombie WR, and Hannon GJ. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods*, 5:679–682, 2008.

[36] Kozarewa I and Turner DJ. Amplification-free library preparation for paired-end Illumina sequencing. *Methods Mol Biol*, 733:257–266, 2011.

[37] Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, and Turner DJ. A large genome center's improvements to the Illumina sequencing system. *Nat Methods*, 5:1005–1010, 2008.

[38] Benjamini Y and Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, 40:e72, 2012.

[39] Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, and Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, 12:R18, 2011.

[40] Vinogradov AE. DNA helix: the importance of being GC-rich. *Nucleic Acids Res*, 31:1838–1844, 2003.

[41] Zoubak S, Clay O, and Bernardi G. The gene distribution of the human genome. *Gene*, 174:95–102, 1996.

[42] Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, and Bernardi G. The distribution of genes in the human genome. *Gene*, 100:181–187, 1991.

[43] Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, and Andrau JC. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res*, 22:2399–2408, 2012.

[44] Hoff KJ. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, 10:520, 2009.

[45] Gundry M and Vijg J. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutat Res*, 729:1–15, 2012.

[46] Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, and Edwards M *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348–352, 2011.

[47] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, and Bettman B *et al.* Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, 2009.

[48] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, and Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13:341, 2012.

[49] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, and Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*, 30:434–439, 2012.

[50] Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, and Hagerman PJ. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res*, 23:121–128, 2013.

[51] Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, and Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*, 7:461–465, 2010.

[52] Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, and Jabado OJ *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol*, 30:1232–1239, 2012.

[53] Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, and Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*, 4:265–270, 2009.

[54] Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet*, 11:31–46, 2010.

[55] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[56] Xie C and Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10:80, 2009.

[57] Simpson JT, McIntyre RE, Adams DJ, and Durbin R. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics*, 26:565–567, 2010.

[58] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, and Du L *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318:420–426, 2007.

[59] Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, and Jaffe DB *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454:766–770, 2008.

[60] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, and Bäckhed F. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 8:241–259, 2013.

[61] Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, and Zhang W *et al.* Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One*, 6:e22751, 2011.

[62] Wang Z, Gerstein M, and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10:57–63, 2009.

[63] Ozsolak F and Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 12:87–98, 2011.

[64] Marioni JC, Mason CE, Mane SM, Stephens M, and Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18:1509–1517, 2008.

[65] Lu C and Shedge V. Construction of small RNA cDNA libraries for high-throughput sequencing. *Methods Mol Biol*, 729:141–152, 2011.

[66] Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, and Plasterk RH. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet*, 38:1375–1377, 2006.

[67] Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, and McCombie WR. Genome-wide in situ exon capture for selective resequencing. *Nat Genet*, 39:1522–1527, 2007.

[68] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, and Eichler EE *et al.* Targeted capture and massively

parallel sequencing of 12 human exomes. *Nature*, 461:272–276, 2009.

[69] Pepke S, Wold B, and Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*, 6:S22–S32, 2009.

[70] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10:669–680, 2009.

[71] Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, and Tomazou EM *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, 26:779–785, 2008.

[72] Shendure J and Lieberman AE. The expanding scope of DNA sequencing. *Nat Biotechnol*, 30:1084–1094, 2012.

[73] Hayatsu H. The bisulfite genomic sequencing used in the analysis of epigenetic states, a technique in the emerging environmental genotoxicology research. *Mutat Res*, 659:77–82, 2008.

[74] Conti R, Veenstra DL, Armstrong K, Lesko LJ, and Grosse SD. Personalized medicine and genomics: challenges and opportunities in assessing effectiveness, cost-effectiveness, and future research priorities. *Med Decis Making*, 30:328–340, 2010.

[75] Biesecker LG, Burke W, Kohane I, Plon SE, and Zimmern R. Next-generation sequencing in the clinic: are we ready? *Nat Rev Genet*, 13:818–824, 2012.

[76] Highnam G and Mittelman D. Personal genomes and precision medicine. *Genome Biol*, 13:324, 2012.

[77] Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, and Mandl KD. Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res*, 22:421–428, 2012.

[78] Kohane IS, Hsing M, and Kong SW. Taxonomizing, sizing, and overcoming the incidentalome. *Genet Med*, 14:399–404, 2012.

[79] Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, Kalyana-Sundaram S, Sam L, Balbin OA, and Quist MJ *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med*, 3:111ra121, 2011.

[80] Biesecker LG. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. *Genet Med*, 14:393–398, 2012.

[81] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[82] International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464:993–998, 2013.

[83] ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, 9:e1001046, 2011.

[84] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[85] Stratton MR, Campbell PJ, and Futreal PA. The cancer genome. *Nature*, 458:719–724, 2009.

[86] Meyerson M, Gabriel S, and Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11:685–696, 2010.

[87] Talbot SJ and Crawford DH. Viruses and tumours - an update. *Eur J Cancer*, 40:1998–2005, 2004.

[88] Dawson MA and Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell*, 150:12–27, 2012.

[89] Salk JJ, Fox EJ, and Loeb LA. Mutational heterogeneity in human cancers: origin and consequences. *Annu Rev Pathol*, 5:51–75, 2010.

[90] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, and Esposito D *et al.* Tumour evolution inferred by single-cell sequencing. *Nature*, 472:90–94, 2011.

[91] Stuart D and Sellers WR. Linking somatic genetic alterations in cancer to therapeutics. *Curr Opin Cell Biol*, 21:304–310, 2009.

[92] Druker BJ. Translation of the Philadelphia chromosome into therapy for CML. *Blood*, 112:4808–4817, 2008.

[93] Moore GE. Progress in digital integrated electronics. *IEEE International Electron Devices Meeting*, pages 11–13, 1975.

[94] Berger B, Peng J, and Singh M. Computational solutions for omics data. *Nat Rev Genet*, 14:333–346, 2013.

[95] EMBL-European Bioinformatics Institute. EMBL-EBI annual scientific report 2012, 2013.

[96] Adams DJ, Berger B, Harismendy O, Huttenhower C, Liu XS, Myers CL, Oshlack A, Rinn JL, and Walhout AJ. Genomics in 2011: challenges and opportunities. *Genome Biol*, 12:137, 2011.

[97] Schadt EE. The changing privacy landscape in the era of big data. *Mol Syst Biol*, 8:612, 2012.

[98] Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, and Austin CP *et al.* Prepublication data sharing. *Nature*, 461:168–170, 2009.

[99] Rodriguez LL, Brooks LD, Greenberg JH, and Green ED. The complexities of genomic identifiability. *Science*, 339:275–276, 2013.

[100] Gymrek M, McGuire AL, Golan D, Halperin E, and Erlich Y. Identifying personal genomes by surname inference. *Science*, 339:321–324, 2013.

[101] Nielsen R, Paul JS, Albrechtsen A, and Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, 12:443–451, 2011.

[102] Ewing B and Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8:186–194, 1998.

[103] Trapnell C and Salzberg SL. How to map billions of short reads onto genomes. *Nat Biotechnol*, 27:455–457, 2009.

[104] Lunter G and Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*, 21:936–939, 2011.

[105] Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

[106] Langmead B, Trapnell C, Pop M, and Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10:R25, 2009.

[107] Li H, Ruan J, and Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18:1851–1858, 2008.

[108] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079, 2009.

[109] Mwenifumbo JC and Marra MA. Cancer genome-sequencing study design. *Nat Rev Genet*, 14:321–332, 2013.

[110] Jones DT and Jäger N *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488:100–105, 2012.

[111] Kernighan BW and Pik R. *The Unix Programming Environment*. Prentice Hall, 1984.

[112] Vidakovic B. *Statistics for Bioengineering Sciences*. Springer, New York, 2011.

[113] Kahvejian A, Quackenbush J, and Thompson JF. What would you do if you could sequence everything? *Nat Biotechnol*, 26:1125–1133, 2008.

[114] Lassmann T, Hayashizaki Y, and Daub CO. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics*, 27:130–131, 2011.

[115] Lam HY, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, and Gerstein MB *et al.* Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol*, 30:78–82, 2011.

[116] Suzuki S, Ono N, Furusawa C, Ying BW, and Yomo T. Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS ONE*, 6:e19534, 2011.

[117] Wang W, Wei Z, Lam TW, and Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep*, 1:55, 2011.

[118] Nothnagel M, Herrmann A, Wolf A, Schreiber S, Platzer M, Siebert R, Krawczak M, and Hampe J. Technology-specific error signatures in the 1000 Genomes Project data. *Hum Genet*, 130:505–516, 2011.

[119] Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, and Lichter P *et al.* Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. *PLoS ONE*, 8:e66621, 2013.

[120] Rausch T, Jones DT, Zapatka M, and Stütz AM *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, 148:59–71, 2012.

[121] Weischenfeldt J, Simon R, Feuerbach L, Schlangen K, Weichenhan D, Minner S, and Wuttig D *et al.* Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell*, 23:159–170, 2013.

[122] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, and DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20:1297–1303, 2010.

[123] Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 2010.

[124] Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, and Webster M *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*, 41:178–186, 2009.

[125] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, and Menzies A *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*, 39:D945–D950, 2011.

[126] Northcott PA, Shih DJ, Peacock J, Garzia L, Morrissy AS, Zichner T, Stütz AM, Korshunov A, Reimand J, and Schumacher SE *et al.* Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, 488:49–56, 2012.

[127] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, and Sherry ST *et al.* The variant call format and VCFtools. *Bioinformatics*, 27:2156–2158, 2011.

[128] Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25:2865–2871, 2009.

[129] Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, and Durbin R. Dindel: accurate indel calls from short-read data. *Genome Res*, 21:961–973, 2011.

[130] Thorvaldsdóttir H, Robinson JT, and Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14:178–192, 2013.

[131] Helena Mangs A and Morris BJ. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Curr Genomics*, 8:129–136, 2007.

[132] McLaughlin SF, Peckham HE, Ranade SS, Lee CC, Clouser CR, Manning JM, Hendrickson CL, Zhang L, Dimalanta ET, and Sokolsky TD *et al.* Large-scale SNP detection via ligation-based dibase sequencing across multiple HapMap individuals: NA18507, NA19240, and NA12878. Poster at Annual Meeting of the American Association for Cancer Research (AACR), Denver, CO, USA, April 2009. http://tools.invitrogen.com/content/sfs/posters/cms_065548.pdf.

[133] Vega VB, Cheung E, Palanisamy N, and Sung WK. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PLoS One*, 4:e5241, 2009.

[134] Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, and Smith KS *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res*, 23:749–761, 2013.

[135] Ruffalo M, LaFramboise T, and Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27:2790–2796, 2011.

[136] Homer N, Merriman B, and Nelson SF. BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, 4:e7767, 2009.

[137] Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, and Brudno M. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5:e1000386, 2009.

[138] Kim SY and Speed TP. Comparing somatic mutation-callers: beyond Venn diagrams. *BMC Bioinformatics*, 14:189, 2013.

[139] DeFrancesco L. Life Technologies promises $1,000 genome. *Nat Biotechnol*, 30:126, 2012.

Acknowledgements

This is the opportunity to express my heartfelt gratitude to the people who have been involved, directly or indirectly, in this doctoral thesis. Without their invaluable support and contributions, I would not be where I stand today.

I would like to thank my first advisor, Prof. Dr. Roland Eils, for giving me the opportunity to work in his group and for his support and guidance throughout this project, as well as my group leader Dr. Benedikt Brors for his scientific assistance and continuous encouragement. I also want to thank Prof. Dr. Holger Sültmann for agreeing to be my second referee on short notice, as well as Prof. Dr. Stefan Wiemann and Dr. Dirk Grimm for being my examiners.

I would like to acknowledge my collaborators on the project, Marc Zapatka and Qi Wang, for countless inspiring discussions and insights, and David Jones and Bärbel Lasitschka for their helpful work. A big thank you goes to all members of the Computational Oncology group, in particular Natalie Jäger, Barbara Hutter, and Dilafruz Juraeva, who have been amazing in all respects. I am grateful for the invaluable support of Rolf Kabbe and Karlheinz Groß, who have been maintaining the complex computational infrastructure needed for this project and were always willing to help, and the support of Corinna Sprengart and Manuela Schäfer who have demonstrated great kindness, patience and fantastic organizational skills.

A very special thank you goes to my family and friends for their limitless encouragement through the good times and the bad.

To my parents and sister, for their unconditional love and support.
To my fellow fire juggling friends, for keeping me grounded and in touch with reality.
To Claire Vidalie, for opening my eyes.
To Niki, for making me a better person.
To Jeanette, for her enthusiasm and dreams.

To Alexei, for everything.

Erklärung

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 26. August 2013 .....

(Nora Rieber)