FREIE UNIVERSITAET BERLIN
FACHBEREICH MATHEMATIK UND INFORMATIK

Master thesis
in Bioinformatics

Detection of copy number variants in sequencing data

Kerstin Neubert
09/10/2010

Examiners:	Prof. Dr. Knut Reinert[1]
	Dr. Ralf Herwig[2]
Academic advisor:	Anne-Katrin Emde[1,2]

---

[1]Institut fuer Informatik, Freie Universität Berlin
[2]Max Planck Institute for molecular Genetics

Contents

1 Introduction **2**

2 Biological Foundations **4**

2.1 Definitions . . . 4
2.2 Copy number variation in the human genome . . . 5
2.3 Functional impact and disease association . . . 5

3 Background **7**

3.1 Detection of CNVs in microarray data . . . 7
3.1.1 Microarray technology . . . 7
3.1.2 Methods based on array data . . . 8
3.2 Detection of CNVs in sequencing data . . . 9
3.2.1 Second-generation sequencing (2GS) technology . . . 9
3.2.2 2GS data . . . 10
3.2.3 Mapping of short reads . . . 12
3.2.4 Methods based on depth of coverage (DOC) . . . 13
3.2.5 Methods based on paired reads (PEM) . . . 15

4 Methods **17**

4.1 Implementation of a CNV detection tool . . . 17
4.1.1 Programming flowchart . . . 18
4.1.2 Acquisition of DOC signals . . . 19
4.1.3 GC-Normalization . . . 20
4.1.4 Event calling . . . 22
4.1.5 Event Merging . . . 22
4.1.6 Selection of dimorphic events . . . 23
4.1.7 Filtering . . . 23
4.2 Implementation of a CNV simulation platform . . . 24
4.3 Interface setup for additional tools . . . 25
4.3.1 Input data conversions . . . 25
4.3.2 Parameter settings . . . 26
4.3.3 Sensitivity and Specificity calculations . . . 26
4.4 Data sets . . . 28
4.4.1 Simulated data . . . 28
4.4.2 European parent-child trio . . . 28

4.4.3 Tumor cell lines 28

5 Results **29**
5.1 Evaluation of the performance in synthetic data 29
5.2 Parameter sensitivity 30
5.3 Comparison with other tools 32
5.4 Application 33
5.4.1 1000 genomes project data 33
5.4.2 Tumor cell lines 34

6 Discussion **37**

7 Conclusions and outlook **39**

List of Figures

3.1 Example for pileups (marked with asteriks) in a SOLiD dataset visualized with the SOLiD alignment browser (http://solidsoftwaretools.com/gf/project/sab/) 11
3.2 Example for a multiread (a) and a uniquely mapped read (b) mapped to a reference genome in an alignment. 12
3.3 Illustration of depth of coverage in CNV regions 13
3.4 PEM: a. deletion and b. insertion of sequence in the sample genome with discordant paired ends, c. anchored split mapping 16

4.1 program overview 18
4.2 Histogram of the number of aligned sequence reads in 1 kb windows of chromosome 1 in NA12891; the expected Poisson distributions is drawn in red 20
4.3 GC dependency on counts of aligned sequence reads in chromosome 1 for NA12891 and sw620 21
4.4 CNV simulation platform 24

5.1 True positive rate (TPR) and false positive rate (FPR) w.r.t. coverage of copyDOC 30
5.2 Correlation plot between predicted and true variant size in 4809 deletions and 4656 duplications. 31
5.3 True positive rate (TPR) and false positive rate (FPR) given various fixed window sizes and coverage. 31
5.4 Comparison of performance of copyDOC and four published tools 33

List of Tables

2.1 Examples of disorders associated with CNVs 6

3.1 Overview of the three common 2GS sequencing platforms 9
3.2 Overview of available tools based on depth of coverage, RD = read depth, tsv = tabulator separated values, hg18 = NCBI human reference, build 36 15

4.1 Parameter for the CNV detection tool copyDOC 23
4.2 Parameter for the CNV simulation platform copySim with examples. 25
4.3 Parameter for the comparison of copyDOC with available tools in section 5.3 27

5.1 Mean and standard deviations for TPR and FPR of the implemented method (copyDOC) on data sets with different coverage 29
5.2 Mean and standard deviations of TPR for the implemented method (copyDOC) with window size 100 bp and 400 bp and other available tools. 32
5.3 Mean and standard deviations of FPR for the implemented method (copyDOC) with window size 100 bp and 400 bp and other available tools. 32
5.4 1000G trio dataset chromosome 1. 34
5.5 Detected events for the three runs with uniquely mapped reads. 34
5.6 Detected events for the three runs with all reads. 34
5.7 CNV concordance with DGV for program run with uniquely mapped reads. 35
5.8 CNV concordance with DGV for program run with all reads. 35
5.9 coverage of tumor cell line data. 36
5.10 Detected events for the tumor cell lines. 36
5.11 CNV concordance with array data. 36

1 Introduction

Genomic variation in humans occurs in diverse forms and sizes ranging from single nucleotide changes to large-scale structural variants of several megabases or even whole chromosomes (e.g. in Trisomie 21). Single nucleotide and copy number variants explain differences in individual phenotypes in humans (Stranger et al. 2007). Larger variants > 1 kb including balanced rearrangements (inversions, some translocations) and copy number variants (CNVs) result in rearrangement, amplification or deletion of subsequent genomic regions and consequently alter the structure of the genome. Interestingly, structural variants might alter gene function through diverse mechanism reviewed in Hurles et al. (2008): An increased copy number of a large (>100 kb) DNA segment that overlaps a gene might increase the expression level of that gene. Smaller variants can affect gene expression by reorganization of single functional units like regulatory elements (promoters, enhancers) or exons, e.g. a deletion of a promoter would abolish the expression of the consecutive gene.
Currently, whole-genome studies focus on the detection of copy number variants and other structural variants in the human genome to elucidate their impact on phenotypes and disease,their population differentiation and evolution. By now, a catalog of structural variants has been collected using microarrays and made publicly available in the database of genomic variants (http://projects.tcag.ca/variation/). Due to the limited resolution of microarrays by the number of probes on the array, the picture of the CNV map is still relatively diffuse. On a population genetic scale it is still unclear, especially for smaller CNVs < 5 kb, how many of them are expected to occur in one individual and to which extent they contribute to the individual phenotype. In the most recent study CNVs are expected to include 24 Mb sequence in a human genome, approximately 0.8% in healthy individuals (Conrad et al. 2010). Genomic rearrangements causing aberrant transcriptional events or dosage effects play also an important role in cancer development (Hanahan and Weinberg 2000). For example a gene that functions as a tumor supressor can be lost by a deletion of a huge DNA segment.
The first human genome assemblies, published by the Human Genome Project (HGP) and Celera Genomics, used the Sanger or dideoxy sequencing technique for the determination of the nucleotide sequence in DNA from different individuals (Nat 2004; Venter et al. 2001). In 2005 new sequencing technologies emerged with a drastic increase of cost-effective sequence throughput by a massive parallelization of sequencing. Three different platforms, with 454 sequencing being the first one, followed by Illuminas sequencing-by-synthesis and ABI/SOLiD's sequencing-by-ligation are currently well-established. These second-generation sequencing (2GS) or next-generation sequencing (NGS) technologies transformed genomic research through their diverse applications like whole-genome sequencing, targeted resequencing, RNA expression profiling (RNA-seq) and analysis of epigenetic modifications e.g. DNA methylation (Morozova and Marra 2008; Metzker 2010). Moreover, the 2GS technologies enable a throughout research on sequence and structural variation at base-pair resolution. The 1000 genomes project, launched by several companies worldwide, meets the challenge of sequencing several hundred genomes from different populations to infer a detailed map of human genetic variation.

The processing of second-generation sequencing data is labor-intensive because of the required preprocessing step, the mapping of the plentiful short reads to an assembled human genome with one of the available fast short-read alignment tools. In several publications a sequencing bias (GC-bias) in the data is reported (Dohm et al. 2008), which might require a normalization method. Finally, the analysis of copy number variants using mapped sequencing reads requires the development of sophisticated bioinformatics algorithms. Most published methods are based on paired reads, i.e. two sequenced fragments flanking the opposite ends of an insert in a genomic library. An unexpected distance of reads in a discordant pair indicates a deletion or duplication of a genomic region. Few methods use the depth of coverage signal of the mapped reads, which is also applicable to unpaired reads (single-end reads). Several methods originally developed for detection of CNVs in array data, have been adapted for application to 2GS data.

In this work I implement a method, which detects CNVs in 2GS data based on the depth of coverage (DOC) signal of mapped reads. I determine the DOC signal by counting reads in constant or dynamic windows along the genome. I use a statistical testing procedure, the event-wise testing (EWT) algorithm (Yoon et al. 2009) for the detection of significant copy number events in successive windows. I integrated the tool in Seqan, a C++ sequence analysis library, making use of its sequence-based data structures. Following up its inherent generic design principle relieves the further enhancement of the implemented workflow with related algorithms.
The necessary performance evaluation of the implemented program required test data. Therefore I implemented a simulation platform that generates artificial copy number variants on a given sequence and subsequently simulates sequencing reads on that sequence with substition errors. Furthermore I implemented interfaces for the comparative application of additional available tools, CNV-seq (Xie and Tammi 2009), DNAcopy (Olshen et al. 2004), SegSeq (Chiang et al. 2009) and SOLiD-CNV-Tool (McKernan et al. 2009), on the simulated data sets. Finally I apply the program on real sequencing data generated with two commonly used sequencing platforms (Illumina, SOLiD).

2 Biological Foundations

2.1 Definitions

The following biological definitions refer to Feuk et al. (2006) and Sharp et al. (2005).

Structural variant (SV). A structural variant is a polymorphic region of the DNA ($\geq$ 1 kb) that affects the structure of the genome in contrast to sequence variants (SNPs) that change the sequence of the nucleotides. Structural variants can be classified according to their influence on genomic copy count, in copy-number variants (CNVs) and copy-number invariant changes or balanced variants such as inversions and translocations.

Single nucleotide polymorphism (SNP). A single base substitution of one nucleotide with another observed in the general population at a frequency greater than 1%.

Copy-number variant (CNV). A DNA segment of at least 1 kb in size, that differs in copy number in two or more genomes within a species due to duplication or deletion events. Usually a sufficient large genomic region is expected to occur twice in the diploid human genome, i.e. the copy number is two.

Copy-number polymorphism (CNP). A CNV which appears in more than 1% in a population is outlined as a copy-number polymorphism (CNP).

InDel. A relative gain (insertion) or loss (deletion) of a DNA segment of one or more nucleotides in a genomic sequence.

Inversion. A DNA segment with reversed orientation with respect to the major sequence of a chromosome.

Segmental duplication. Segmental duplications or low-copy-repeats (LCRs) are DNA blocks (1 to 400 kb) that occur in more than one site within a haploid genome and are very similar (>90% sequence identity). They account for approximately 5% of the genome and frequently coincide with variable copies in different genomes and thus can be CNVs at the same time (Sharp et al. 2005) .

Translocation. A DNA segment with a modified position in the genome that has no gain or loss in DNA content is termed a translocation. The translocation can either occur within a chromosome (intra-chromosomal) or between different chromosomes (inter-chromosomal).

2.2 Copy number variation in the human genome

Several studies aimed to construct a map of copy number variants in the human genome to study their phenotype association. The first CNV map collected 1447 copy number variable regions (CNVRs), covering 360 Mb (12%) of the human genome using single-nucleotide polymorphism (SNP) genotyping arrays and clone-based comparative genomic hybridization (Redon et al. 2006). Due to an overestimation of CNV sizes in this study recent studies report that CNVs (>50 kb) affect less (approximately 0.5% in McCarroll et al. (2008)) of the genome than initially expected. In contrast, the influence of smaller CNVs (<5kb) is underestimated due to limited resolution of array-based methods. The extent to which two unrelated human genomes vary in copy number was currently estimated to be between 24 Mb and 60 Mb, spanning $0.8 - 2\%$ of the genome sequence (Conrad et al. 2010; Cooper et al. 2007). These variable regions overlap with approximately 400 protein coding genes (Conrad et al. 2010).
It is known, that CNVs are not distributed uniformly throughout the genome, but are enriched in regions close to telomeres, centromeres and simple tandem repeat sequences and segmental duplications (Nguyen et al. 2006; Cooper et al. 2007). Deletion, duplication and inversions of genomic segments can be formed by the mechanism of nonallelic homologous recombination between duplicated sequence blocks, which results in the observation, that copy number polymorphisms are enriched within hotspots of segmental duplications (Bailey et al. 2002; Sharp et al. 2005).

2.3 Functional impact and disease association

The nonrandom distribution of CNVs in the genome indicates that they might affect gene function. A genome-wide gene expression study in 270 lymphoblastoid cell lines showed an association of gene expression and large-scale (> 100 kb) copy number variation (Stranger et al. 2007). Genes involved in sensory perception (e.g. olfactory receptors), immune response and signaling are enrichend in published CNVs (Cooper et al. 2007). Large duplications or deletions are associated with specific inherited or sporadic (de novo rearrangement) genetic disorders and multifactorial diseases e.g. in Tab. 2.1 (Hurles et al. 2008; Freeman et al. 2006). For example the Charcot-Marie Tooth type 1 A (CMT1A) disease is caused by a 1.5-Mb tandem duplication on chromosome 17 resulting in three copies of the PMP22 gene (Lupski et al. 1991). In several complex diseases like autism spectrum discorder (ASD) (Pinto et al. 2010), schizophrenia (Xu et al. 2008), Parkinson (Singleton et al. 2003), HIV/AIDS susceptibility (Gonzalez et al. 2005) and cancer (Campbell et al. 2008) copy number variants were observed (Tab. 2.1).

Disease	location	Structural variant	References
Charcot-Marie-Tooth type 1A	17p12	duplication of PMP22	Lupskic et al. (1991)
Williams-Beuren syndrome	7q11.23	deletion of ELN and others	Ewart et al. (1993)
Autism	16p11.2	deletion of 16p11.2	Pinto et al. (2010) Marshall et al. (2008) Kumar et al. (2008)
Schizophrenia	16p11.2	microduplication of 28 genes	Xu et al. (2008); McCarthy et al. (2009)
Parkinson	4p15	triplication of SNCA	Singleton et al. (2003)
HIV susceptibility	17q	multi-allelic CNV of CCL3L1	Gonzalez et al. (2005)
Psoriasis	8p23.1	multiallelic CNV of Beta-defensins	Hollox et al. (2008)
Alzheimer	21q21	duplication of APP	Rovelet-Lecrux et al. (2006)

Table 2.1: Examples of disorders associated with CNVs

3 Background

3.1 Detection of CNVs in microarray data

3.1.1 Microarray technology

The development of DNA microarrays enabled the scanning of the human genome for variation in copy number at much higher resolution than in cytogenetic analysis using microscopy, which is limited to 5 Mb sized events. Array comparative genome hybridization (array-CGH) and single-nucleotide polymorphism (SNP) arrays have been used to detect copy-number variable regions with a size of 700 bp to several megabases in the genome (Carter 2007).

array-CGH

Clone-based comparative genomic hybridization (array-CGH) is based on the hybridization of differentially labelled test and reference DNAs in spotted clones on a glass slide. Resulting fluorescent signals are measured for each clone representing the relative amount of DNA at their location in the genome. An amplification or a deletion of a genomic region in the test genome relative to the reference genome is inferred through an increased or decreased hybridization intensity of the corresponding probes on the array. The resolution of this method depends on the number and size of the probes on the array. For example with array-CGH using BACs, which are usually 80-200 kb in length, large-scale copy number differences of $\geq$50 kb in two samples can be detected (Iafrate et al. 2004). Using oligonucleotide probes instead of clones substantially pushes the resolution of array-CGH up to 5 kb in high-density oligonucleotide arrays (e.g. HD2 array from NimbleGen). Representational oligonucleotide microarray analysis (ROMA-CGH) was developed to improve the poor signal-to-noise ratio of oligonucleotide arrays by reducing the complexity of the hybridized genomic DNA (Sebat et al. 2004). The resolution of array-CGH can be futher improved to almost nucleotide-level by the custom design of overlapping oligonucleotides in selected chromosomal regions resulting in ultra-high-resolution arrays (Gribble et al. 2007).

SNP genotyping arrays

Several studies adapt SNP genotyping arrays for the interrogation of genomic copy number variants. The Affymetrix SNP chips contain several matched and mismatched probe pairs (25 bp long) that cover a known SNP. Genomic DNA of one sample is hybridized to the array and the resulting signal intensities of the matched and mismatched probes are compared to those of another individual (or a group of

individuals) to detect copy number changes. Affymetrix extended its SNP genotyping arrays by non-polymorphic probes for interrogation of CNVs in genome areas not covered by SNPs and regions that include segmental duplications. The current Affymetrix SNP 6.0 array including 1.8 million probes was used to generate a human CNV map at 2 kb breakpoint resolution (McCarroll et al. 2008).

3.1.2 Methods based on array data

Different methods have been developed for the detection of CNVs in array intensity data. In this section I shortly explain two approaches that model the distribution of array intensity data (Clustering, hidden markov models) and two nonparametric methods (circular binary segmentation and mean-shift-based approach). In Dellinger et al. (2010) the performance of different approaches for SNP array data was evaluated.

Clustering approach. Korn et al. (2008) use multiple probes for specific interrogatation of a predefined CNP locus from a published CNV map (McCarroll et al. 2008) in SNP array data. Intensity signals are summarized for each CNP segment and clustered using a one-dimensional Gaussian mixture model (GMM) and prior information from previous experiments. The copy number of each locus in a sample is assigned by its membership to a cluster. This method requires knowledge of CNV loci in the human genome and a set of accurately genotyped CNVs for validation of results. Consequently, rare and *de novo* CNVs cannot be detected with this approach.

Hidden markov model (HMM). Several studies use approaches based on a stochastic model, the hidden markov model, for the detection of copy number variable regions in array data e.g. in Korn et al. (2008). It consists of hidden and observed states that represent the unknown copy number of probes in a sample and their normalized intensity measurements in the array. The parameters of the model, the emission and transition probablities are empirically estimated. The transition probabilities between the copy number states are chosen according to the expectation: For the transition from the state with normal copy number (expected to be 2) to another state a low probability is selected, otherwise a relative high probability is choosen. The distance of adjacent probes can be included in the model such that the transition probabilities of nearby probes is higher. This approch is applicable to data from both array platforms. The detection of multi-copy variants is limited with this approach, because with increased number of hidden states the computational complexity is multiplied.

Circular binary segmentation (CBS). Circular binary segmentation is a modification of a binary segmentation (Olshen et al. 2004). It uses the fact, that copy number variants are discrete gains or losses of DNA in contiguous segments of the genome that cover multiple array probes. The underlaying idea is to split each chromosome into regions of equal copy number and thereby overcome the noise in array data. Binary segmentation is a change-point method. Let $X_1, X_2, \ldots$ be a sequence of random variables. A change-point is defined as an index v that marks a shift in the distribution function F_0 of $X_1, \ldots, X_v$ to the distribution function F_1 in $X_{v+1}, X_{v+2}, \ldots$ until the next change-point. The array probes are searched for change-points in the log ratios of intensities in a test versus a control sample that belong to changes in copy number in the two data sets. These change-points separate segments with unequal copy number. The control sample is assumed to have a constant copy number, which is two in case

of autosomes. The binary segmentation procedure recursively applies a test statistic on the log ratio intensities of probes, that are indexed by their physical location on the chromosome, to identify change-points. It results in a partition of each chromosome into segments where the copy numbers are constant.

Mean-shift-based approach (MSB). The Mean-shift-based approach considers the array-CGH intenstiy data as sampled from a probability-density function (Wang et al. 2009). It uses a kernel-based approach, a method in pattern recognition, to estimate local gradients for this function. The kernel density estimation (or Parzen window method) is a method for estimating the probability density function (p.d.f., Ker (1995)). Essentially, MSB performs a discontinuity-preserving smoothing by iteratively shifting data points to the density maxima in the distribution of intensities. MSB segments each chromosome into regions of duplication and deletion.

3.2 Detection of CNVs in sequencing data

3.2.1 Second-generation sequencing (2GS) technology

Second-generation sequencing (2GS) or next-generation sequencing (NGS) are newer developments that followed the automated Sanger method, which was the first-generation technology. They substantially increase the sequence throughput through massively parallel sequencing of several million short reads (35-400 bp) resulting in lower costs per sequenced base pair. I illustrate here three established platforms - 454 (Roche Applied Science), Genome Analyzer (Illumina), SOLiD instrument (Applied Biosystems). For a detailed technical review of recent advanced 2GS technologies see Metzker (2010). The 2GS platforms use different combinations of methods for DNA template preparation, sequencing and imaging and produce different amounts of raw sequencing data (Tab. 3.1).

Platform	**template preparation**	**sequence reaction**	**read length (bp)**	**Run time (days)**	**Gb per run**
Roche/454's GS FLX Titanum	emPCR	pyrosequencing	400	0.35	0.45
Illunina/Solexa's GAII	solid-phase	reversible terminator	75 or 100	4^{F}, 9^{MP}	18^{F}, 35^{MP}
Life/APG's SOLiD 3	emPCR	sequencing by ligation	50	7^{F}, 14^{MP}	30^{F}, 50^{MP}

Table 3.1: Overview of the three common 2GS sequencing platforms, F = fragment, MP = mate-pair/paired-end, emPCR = emulsion PCR, GS = Genome Sequencer, GA = Genome Analyzer, SOLiD = Support Oligonucleotide Ligation Detection, APG = Agencourt Personal Genomics

Template preparation

The first step in template preparation includes the random shearing of genomic DNA into smaller pieces (e.g. 200-250 bp). These small sized DNA is transformed either in fragment templates or mate-pair templates. Subsequently these templates are attached to a solid surface or immobilized to a support. The imaging systems need multiplied fluorescent events, which is achieved by the amplification of

templates. Two common methods are *emulsion PCR (emPCR)* and *solid-phase* amplification. In emulsion PCR the DNA is captured onto beads with one DNA molecule amplified per bead (454, SOLiD). In solid-phase amplification the clonally amplified clusters are generated on a glass slide (Illumina/Solexa).

Sequencing and imaging

In the sequencing step the clonally amplified templates are sequenced simultaneously. The observed fluorescent signal is measured for each cycle. The *Cyclic reversible termination (CRT)* is a sequencing and imaging method that uses reversible terminators. Each cycle of CRT includes three steps, the incorporation of a single fluorescently labelled nucleotide by a DNA polymerase, imaging of the fluorescent signal to determine the identity of the nucleotide and cleavage of the terminating group and fluorescent dye. Illunia/Solexa uses a four-colour CRT for simultanous incorporation and imaging of the four different nucleotides, each labelled with a different dye.
Sequencing by ligation (SBL) distinguishes from CRT by using a DNA ligase instead of a DNA polymerase. A fluorescently labelled pobe is hybidized to its complementary sequence and joined by a DNA ligase to the adjacent primer. The identity of the ligated probe is determined by fluorescence imaging. Finally the fluorescent dye is cleaved to repeat the cycle. Applied Biosystems uses SBL with two-base-encoded probes in their support oligonucleotide ligation detection (SOLiD) platform.
In contrast to the other sequencing approaches, *pyrosequencing* does not modify nucleotides. Instead, a DNA polymerase is manipulated by single addition of a dNTP. The release of an inorganic pyrophospate is measured through conversion into light by enzymatic reactions with luciferase.

3.2.2 2GS data

Mate pairs/paired reads

The three widely-used 2GS-technologies (454, Genome Analyzer, SOLiD instrument) are able to generate paired reads that map at an approximately known distance in the human genome termed mate pairs or paired-end reads. Mate pairs are constructed when the ends of a long DNA fragment (>1kb) are sequenced. The ends of the 1 kb-fragment are tagged by an adaptor and circularized. Then the circularized DNA is randomly fragmented and the tagged junction fragments are purified and sequenced. In contrast, paired-end reads are generated by fragmentation of genomic DNA into short pieces (200-300 bp) and sequencing of both ends. The orientation of the paired reads is equal or opposite depending on the used library preparation protocol. The information about the expected approximate distance of paired reads in the genome can be used by algorithms for the detection of structrual variants. The mate pair and paired-end sequencing approach provide short-range and long-range pairing information.

Sequencing errors and coverage biases

Sequencing errors and coverage biases have been reported for different sequencing platforms. Sequencing error rates should be considered in the mapping of the reads to a genome. The coverage biases (GC-bias, pileups) may result in unequal sampling of certain genomic regions and thus require normalization methods.

Error rates. The different 2GS sequencing platforms are reported with different error rates, 3% for pyrosequencing (Roche/454) (Quinlan and Marth 2007), $1-2\%$ for Illumina/Solexa (Hillier et al. 2008) and 0.1% for SOLiD (McKernan et al. 2009). Substitutions are created during the amplification step (PCR) and might be mistakingly interpreted as sequence variants. The accuracy of the SOLiD platform is substantial higher compared with the other two platforms due to its inherent error correction by a double interrogation of each base (2-base encoding or 2BE). The sequencing error rates continue to improve for all platforms.

GC-bias. GC-poor and GC-rich genomic regions have been shown to be underrepresented in Illumina/Solexa and SOLiD sequencing data compared to regions with average GC content (Dohm et al. 2008; Hillier et al. 2008; Harismendy et al. 2009), which is probably caused by an amplification bias during template preparation (Metzker 2010). Shorter reads are more susceptible to such a chemistry bias due to local sequence composition e.g. GC content.

Pileups. If someone looks closer at the aligned short reads in a sequencing experiment some reads accumulate to noticeable huge perfectly identical pileups, e.g. in figure 3.1. These reads might be considered as PCR artefacts and should eventually be counted only once in an analysis that is based on a quantification of reads.

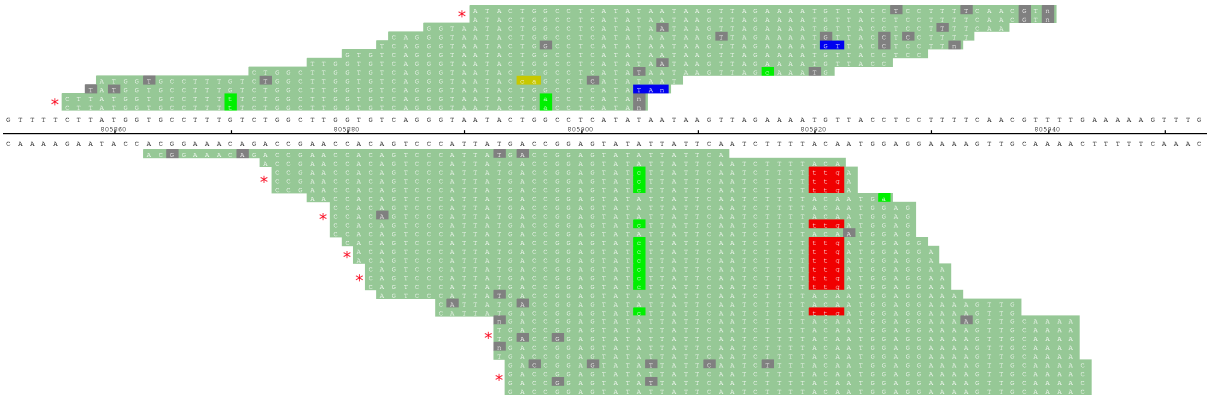

Figure 3.1: Example for pileups (marked with asteriks) in a SOLiD dataset visualized with the SOLiD alignment browser (http://solidsoftwaretools.com/gf/project/sab/)

Current sequencing-based methods rely on the indirect comparison through one assembled reference genome for detection of copy number variants in different individuals. The de-novo assembly of their genomes for a direct comparison of the sequences using short reads is a challenging task that requires very high coverage (at least 30x, see Li et al. 2010) and sophisticated algorithms. Li et al. (2010) successfully assembled the human genome sequences of an Asian and African individual using single-end and paired-end short reads (35-75 bp) at an average coverage of 71x and 40x of the NCBI human reference genome using a de Bruijn graph and a supercomputer with 512 Gb memory to handle the huge number of short reads. As long as de-novo assembly of genomes is still unaffordable for structural variant detection in multiple genomes, methods use different signatures that result in the mapping of short reads to a reference genome (Medvedev et al. 2009).

3.2.3 Mapping of short reads

Initially, each sequencing read is aligned to a reference sequence allowing mismatches and eventually small InDels. The mass data in second-generation sequencing requires efficient mapping algorithms that associate each read with at least one position in the genome. There are already several tools available (listed in http://lh3lh3.users.sourceforge.net/NGSalign.shtml).
Due to the highly repetetive structure of the genome it is likely that reads are sampled from a redundant sequence by the sequencing process. Even if large repetetive portions (telomeres, centromeres) are excluded from the mapping, there remain smaller redundant structures e.g. *Alu* elements, LINE repeats and segmental duplications (SDs). A mapping algorithm will detect multiple good alignments that map these reads below the allowed sequencing error threshold. This results in the multiread assignment problem, that is to choose one position from multiple good alignments of an ambiguously mapped read. Some mapping tools try to adress this problem additionally. For example the Maq algorithm simply assigns the multiread randomly to one of its shared positions. But the existing solutions are dissatisfying, because they cannot guarantee that the selected position is indeed the originating one for the read.
The multiread assignment problem can be avoided by discarding sequence reads that map to multiple genomic loci and use exclusively reads that fit in some definition as *uniquely mapped*:

Def. *uniquely mapped reads*
Given an alignment of a sequence read to a reference genome that allows at most e mismatches. The aligned read is considered as *uniquely mapped* if its second best alignment has more than e mismatches when compared to its best alignment.

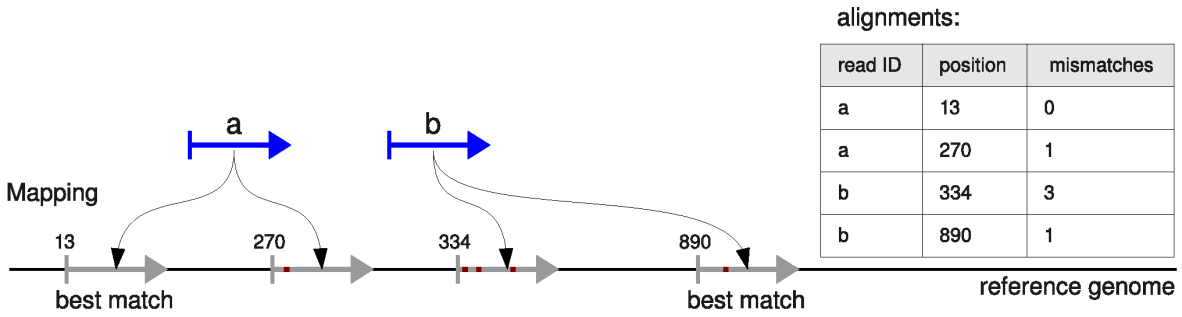

Figure 3.2: Example for a multiread (a) and a uniquely mapped read (b) mapped to a reference genome in an alignment.

An illustration of uniquely mapped reads is shown in Fig. 3.2. In this example reads are mapped with at most 2 mismatches. Read *a* has two good alignments with zero and one mismatch that are below the given error threshold. Thus read *a* is not a uniquely mapped read. Read *b* aligns to two positions with one or three errors, but its second best position would be excluded by the used mapping algorithm, because the number of mismatches exceeds the allowed error threshold. As a consequence, read *b* is a uniquely mapped read and is assigned the genomic position of its best match. The limitation to uniquely mapped reads has some disadvantages. Considerable information in repetive regions is lost and the coverage is notably reduced in a non-uniform fashion. This results in an underrepresentation

of mapped reads in repetetive regions e.g. segmental duplications.
The multiread issue might influence the results for the analysis of copy number variation that makes use of the depth of coverage signal. All published tools focus on uniquely mapped reads or choose the read with the best match with fewest mismatches from all reported alignments. Futher methods based on paired-end mapping (PEM) use additional information about the pairing of the two reads to reliably detect their location in the genome. Discrepancies in the distance of the reads inform about potential deletions or duplications.

3.2.4 Methods based on depth of coverage (DOC)

Methods based on depth of coverage along the genome are similar to the array-based methods. Instead of determining the copy number by quantifying a probe signal relative to a reference, they estimate the copy number by quantifying the amount of reads in a sample in windows. Assuming the sequencing method samples sequence reads uniformly from the genome, the number of reads aligning to a region approaches a Poisson distribution with mean μ proportional to the size of the region and its copy count in the genome. According to this the probability of exactly k observed reads in a fixed genomic region is

$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!} \tag{3.1}$$

A deleted region is expected to have less or no reads mapped to it and a duplicated more reads (Fig. 3.3 a,b). For sufficiently high sequence coverage the Poisson distribution can be approximated by a normal distribution. The majority of algorithms based on depth of coverage build statistical models on the read distribution.

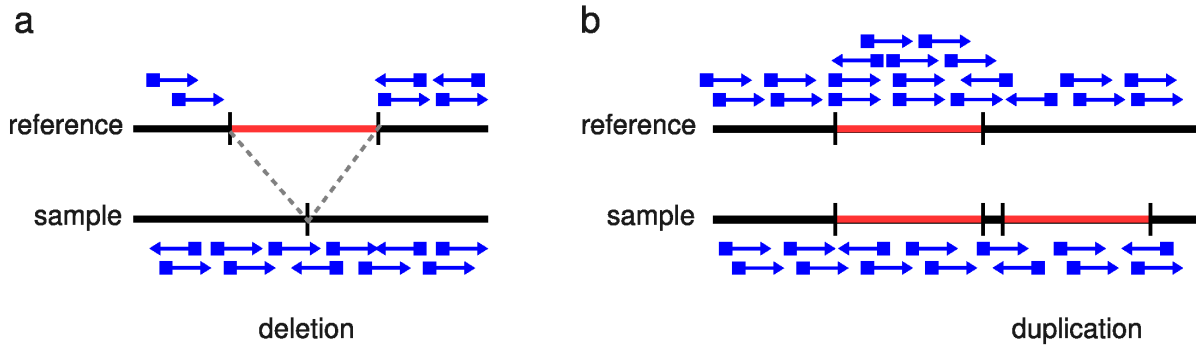

Figure 3.3: Illustration of depth of coverage in CNV regions (red). A deletion in a sample sequence results in zero coverage relative to the reference (a) and a duplication in the sample results in twice the expected coverage (b)

CNV-seq. Xie and Tammi (2009) use a simple robust statistical model for the sequencing process based on log-ratios of read counts in constant overlapping windows in two samples. The window size is determined based on the coverage of the input data and the threshold parameters for log-ratios and p-values. The program is available as a R package (http://tiger.dbs.nus.edu.sg/cnv-seq/).

DNAcopy. Circular binary segmentation (see sec. 3.1.2) was originally developed for the segmentation of chromosomes in regions with equal copy number using array data (Olshen et al. 2004). For its application to 2GS data, the sequencing data must be processed to achieve log ratios of read counts. Campbell et al. (2008) divided the genome into nonoverlapping, adaptive windows based on a fixed number of in silico mapped reads with high uniqueness (with approx. 15 kb mappable sequence) and counted the number of mapped reads in each window. Then they applied DNAcopy to the log ratio of the read counts in two datasets. The tool is available from the Bioconductor project [1].

EWT. Yoon et al. (2009) developed a method called event-wise testing (EWT) that uses significance testing for duplication and deletion events on intervals of non-overlapping windows on the genome. They iteratively enlarge the event size starting with two adjacent windows. For larger events less stringent significance thresholds are used. The iterations stop when the significance threshold reaches a cutoff. The authors did not publish the implemented Java program.

SegSeq. Chiang et al. (2009) implemented a method called local change-point analysis. They established a test statistic based on local differences of log ratios in read counts of a test and control data set. Peaks in this local difference statistic that exceed a signifcance threshold are identified as candidate breakpoints for copy number changes. The resulting segments are iteratively joined by eliminating candidate breakpoints. The authors applied the method for the detection of copy number alterations in Illumina sequencing data from a tumor and normal sample. The Matlab source code of SegSeq is accessible from the Broad Institute (http://www.broadinstitute.org/mpr/publications/projects/Computational_Biology/SegSeq_1.0.1.tar.gz).

SOLiD-CNV-Tool. The SOLiD-CNV-Tool employs a hidden markov model that represents the copy number (0-7) as discrete hidden states and the coverage signal as observations on variable-length windows. The window sizes depend on the uniqueness or mappability of the sequence (McKernan et al. 2009). Thus in regions that include repetetive or redundant sequence larger windows are used. The tool is written in C, available in http://solidsoftwaretools.com/gf/project/cnv/.

The four available tools that are based on depth of coverage (CNV-seq, DNAcopy, SegSeq, SOLiD-CNV-Tool) require different input data formats (Tab. 3.2). Most tools output the position, copy number and p-value for each predicted CNV, except DNAcopy that outputs location of DNA segments with the same copy number and their mean read depth.

[1] http://www.bioconductor.org /packages/2.3/bioc/html/DNAcopy.html

Tool	Algorithm	Input (format)	Output	Ref.
CNV-seq	statistical model on log ratios	location of best matches (tsv)	position, log ratio, *p*-values of CNVs	Xie and Tammi (2009)
DNAcopy	Circular binary segmentaion	RD log ratios in windows (tsv)	segments, mean RD	Olshen et al. (2004)
SegSeq	Local change point analysis	location of mapped reads (tsv)	position, copy number, *p*-value of CNVs	Chiang et al. (2009)
SOLiD-CNV-Tool	Hidden markov model	mapped SOLiD reads @hg18 (gff2/gff3)	position, copy number, *p*-value of CNVs	McKernan et al. (2009)

Table 3.2: Overview of available tools based on depth of coverage, RD = read depth, tsv = tabulator separated values, hg18 = NCBI human reference, build 36

3.2.5 Methods based on paired reads (PEM)

Sequence-based methods have extensively made use of mate-pair or paired-end reads for the analysis of structural variation in sequencing data (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008; Bentley et al. 2008). The paired-end mapping (PEM) approach uses the fact, that the two paired reads must map to the reference sequence with a fixed distance according to the size of the insert in the used library.

Analysis of discordant read pairs

Discordant read pairs with an enlarged distance of the mapped paired ends reveal a potential deletion in the sample genome. An unexpectedly short distance of the ends points to an insertion of sequence at the spanned locus in the sample genome relative to the reference genome (Fig. 3.4 a,b). Insertions which exceed the size of the sequenced fragment are difficult to detect with this method. If the inserted sequence occurs at another location in the genome, a linking of the paired reads between the location of the insertion and the inserted sequence can be detected. If the inserted sequence is not contained in the reference genome, a read spanning a breakpoint will not map resulting in a hanging insertion. A novel inserted sequence can be determined by a local assembly of the breakpoints within the variant region (Chen et al. 2009). Clustering strategies as decribed in Tuzun et al. (2005) classify discordant read pairs with a mapped distance more than 2 *sd* apart the mean insert size with similar size and location. This method enables the estimation of the breakpoint location of a copy number variant.

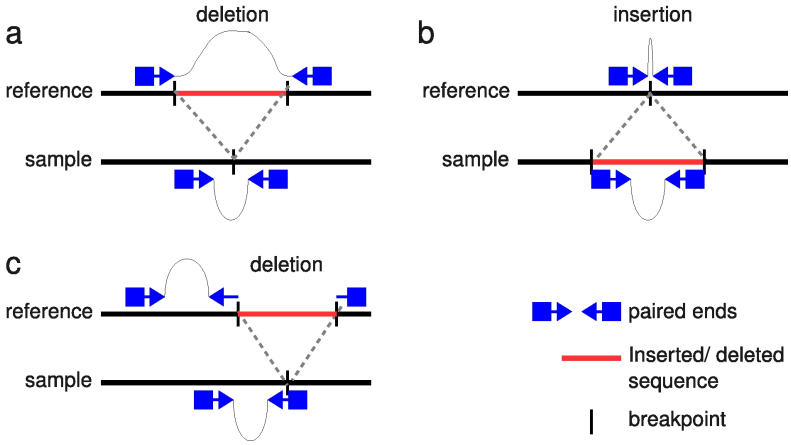

Figure 3.4: PEM: a. deletion and b. insertion of sequence in the sample genome with discordant paired ends, c. anchored split mapping

Split-read alignment

If a read spans a deletion breakpoint in the sample genome, it is splitted in its prefix and suffix mapping at different locations in the reference genome (Fig. 3.4c). Ye et al. (2009) uses the mapped pair of a split read as an anchor to reduce the search space for the mapping of the splitted read. With the anchored split mapping the exact breakpoints of large deletions can be determined with base-pair resolution.

In an evaluation of the different methods the specificity and sensitivity for the detection of copy-number changes, their ability to predict the copy number, its size and the exact location of the breakpoints might be considered. Microarray-based technologies are limited to an a-priori focus on specific probe content on the array and have thus a lower resolution and sensitivity compared to sequencing-based methods. In addition, the accurate copy number of multi-copy variants can not be inferred by array-based methods, because of oversaturation effects.
Second-generation sequencing (2GS) methods can achieve base-pair resolution with sufficient high coverage. In contrast to methods based on depth of coverage, methods based on paired reads enable the detection of breakpoints (the accurate boundaries of a copy number event). Using a combined library approach with different insert sizes increases the resolution of the breakpoint mapping.

4 Methods

I implemented a program for detection of CNVs in 2GS data based on depth of coverage in C++ (*copyDOC*), using the open source C++ Sequence Analysis Library SeqAn [1] as platform. To evaluate the performance of the implementation I established a simulation platform for generating synthetic CNVs in a template DNA sequence and subsequent simulation of 2GS sequencing data on that sequence (*copySim*). This platform was used comparing copyDOC with four published tools that use different approaches based on depth of coverage. Therefore it was necessary to setup interfaces for data conversion of the Sequence Alignment/Map (SAM) format to the respective input format and for evaluation of results.
In the following two parts I explain in detail the copyDOC tool and the copySim environment. Futhermore I illustrate the interface setup and parameter settings. To assess the usability of copyDOC in practice I applied it to real 2GS data sets from Illumina/Solexa and SOLiD platform. Results will be shown in chapter 5.

4.1 Implementation of a CNV detection tool

In most biological applications, the experimentator is interested in copy number variable regions in two samples e.g. from different individuals or tissues. Sometimes, however, a second dataset is not available as a control. I designed the copyDOC program such that it can be run with or without a control data set. Prerequisite to the CNV detection is a preprocessing of the reads, the mapping to a reference genome using one of the available alignment tools [2]. I use SAM as input format for aligned sequence reads, because it is supported by most alignment tools (e.g. Bowtie, BWA, mrFAST, RazerS, SHRiMP). In the following sections I first give an overview of the programming flowchart and then explain single steps.

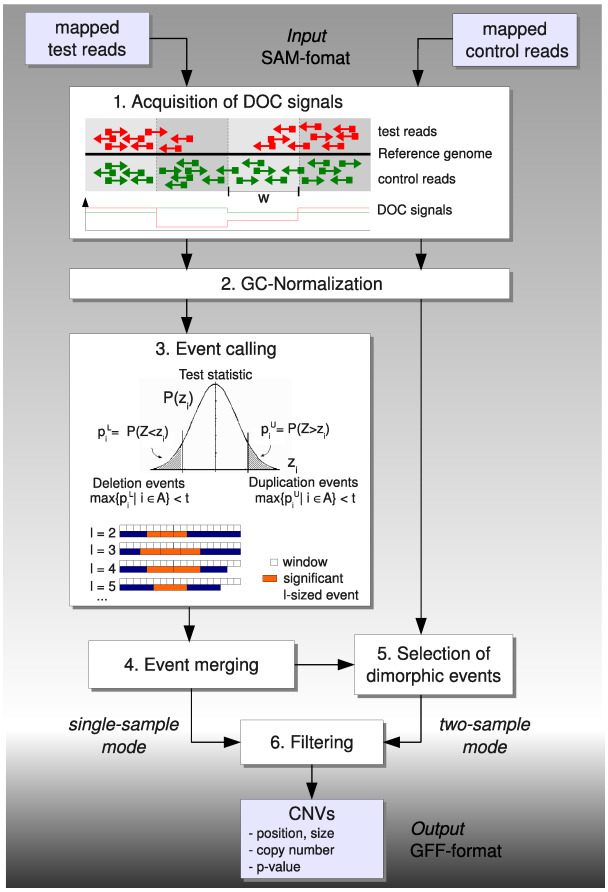

Figure 4.1: program overview

4.1.1 Programming flowchart

1. **Acquisition of DOC signals**

   Depth of coverage signals are determined by counting start positions of mapped reads in fixed or dynamic windows along the genome.

2. **GC-Normalization**

   A simple median-based GC-Normalization is applied to DOC signals (optional).

3. **Event calling**

   For event calling the Event-wise testing (EWT) algorithm (Yoon et al. 2009) is used that builds a test statistic on the empirical distibution of normalized DOC signals. It evaluates lower and upper tail probabilities (p_i^L and p_i^U) based on the test statistic for each window i in a sequence. Then it searches each chromosome for consecutive windows in an interval A with maximum tail probabilities below a given adaptive threshold t that indicate deletion events ($max\left\{p_i^L | \in A\right\} < t$)

[1] http://www.seqan.de/
[2] http://lh3lh3.users.sourceforge.net/NGSalign.shtml)

or duplication events ($max\left\{p_i^U | \in A\right\} < t$).

4. **Event merging**

   I merge events that are at most d_{max} bases apart. Small events (at most s_{max} base pairs large) are gouped by nearest neighbor clustering and merged.

5. **Selection of dimorphic events**

   If a control dataset is given a t-test is applied to DOC signals in the test and the control dataset in order to select dimorphic events.

6. **Filtering**

   Finally, CNVs are filtered by user-defined parameters (e.g. size, significance threshold) and results are exported in GFF-format.

4.1.2 Acquisition of DOC signals

The first essential step in the pipeline is the determination of depth of coverage signals in windows. The number of aligned sequence reads in windows is expected to follow a Poisson distribution with mean proportional to the size of the region and the copy number. To confirm this I plotted the read count distributions for NA12891 (1000 Genomes project data) in the alignable portion of the human chromosome 1 (excluding gaps) in Fig. 4.2. In this dataset multireads are included and their position was taken randomly from all possible alignments by the Maq alignment tool. I fitted a Poisson distrubution based on the average number of sequence reads in that dataset. In contrast to the expected Poisson distibution the density of reads in real data is inhomogeneous. This is presumably explained by biological variation or coverage biases of the sequencing process.

The DOC signal of the sequencing data is determined by counting start positions of reads in constant or dynamic-sized windows. Whether only uniquely mapped reads or multireads are used depends on the input SAM file (all reads are processed) or the given mapping quality threshold (reads below the threshold are not counted).

Constant windows

For constant-sized windows their size must be choosen according to the coverage of the dataset. Xie and Tammi (2009) adapts the window size to the average genome coverage and user defined threshold parameters. In the copyDOC program the user chooses an adequate window size.

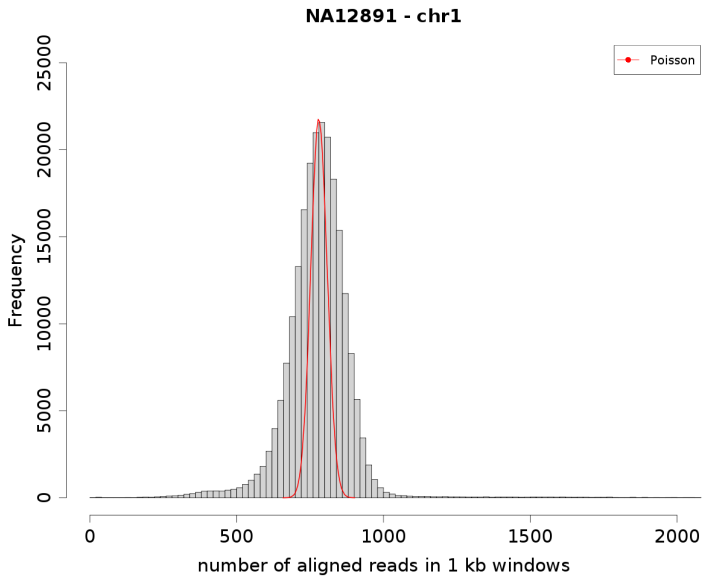

Figure 4.2: Histogram of the number of aligned sequence reads in 1 kb windows of chromosome 1 in NA12891; the expected Poisson distributions is drawn in red

Dynamic windows

If uniquely mapped reads are used, less reads map in repetetive regions in the genome resulting in low read counts in the corresponding windows. Therefore it might be useful to choose dynamic window sizes that depend on the repetetiveness or uniqueness of the sequence. I added this as an optional feature. As additional input, the dynamic windows feature requires a mappability file in wiggle-format that informs about whether a genomic position can be mapped uniquely for a given read length and number of allowed errors. This file cannot be precomputed, because it depends on the currently used genome assembly, the read length and allowed error rate in the mapping process. It can be created by fragmentation of a genome with given read length and remapping the reads to it with a defined error threshold. The information about unambigously mapped positions must be written in the wiggle-file. Dynamic windows are choosen such that the number of uniquely mappable positions in a window is constant and equal to a user-defined parameter.

4.1.3 GC-Normalization

A GC-bias, that is a nonlinear dependency between the G+C percentage and the number of aligned reads, was observed in Illumina/Solexa and SOLiD data (Dohm et al. 2008; Hillier et al. 2008; Harismendy et al. 2009). I examined this in two different data sets from Illumina and SOLiD. Therefore I

computed the G+C percentage of all possible reads in each 1 kb alignable window of chromosome 1. I plotted counts of aligned reads in each window in relation to the estimated G+C percentage of the reads in that window (Fig. 4.3). In NA12891 data set I can not infer a nonlinear relationship. In the SOLiD dataset sw620 there seems to be a nonlinear dependency between read counts and G+C content in 1 kb windows.

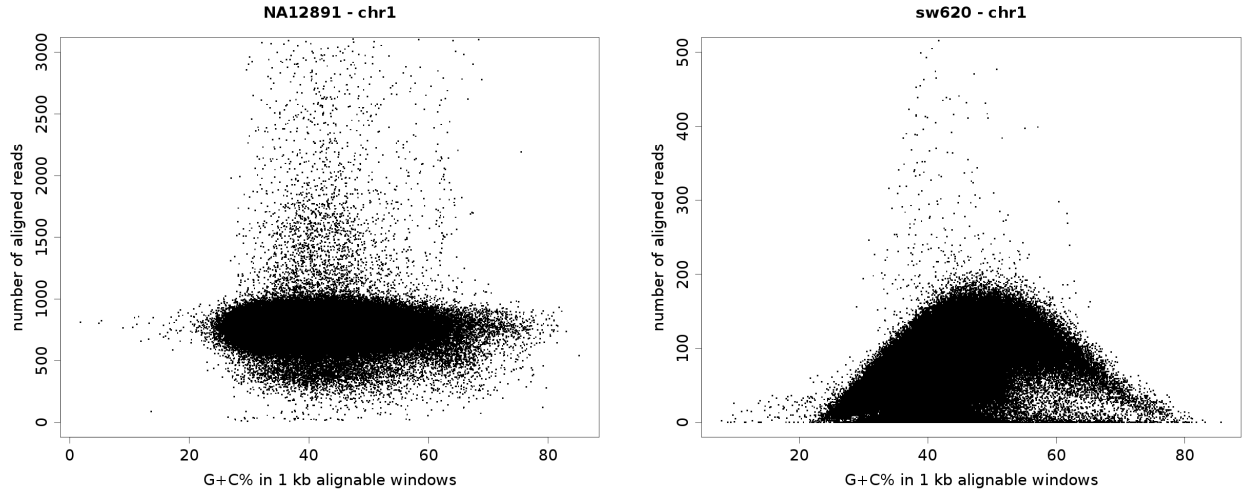

Figure 4.3: GC dependency on counts of aligned sequence reads in chromosome 1 for NA12891 and sw620

The optional normalization for GC-bias is done by a simple median based normalization. For each window the G+C content is ascertained using the reference sequence. For different G+C intervals from 1% upto 100% the median read count of all windows with the same approximated G+C content is computed. Yoon et al. (2009) adjust the read count of window i, r_i, by the median read count of windows with the same G+C-content (m^i_{GC}) as window i relative to the overall median read count m, resulting in the normalized read count $\tilde{r}_i$:

$$\tilde{r}_i = r_i \frac{m}{m^i_{GC}} \tag{4.1}$$

To be more robust with respect to windows with extreme G+C content I implemented a smoothed version of the median-based normalization, that uses a sliding averge of median read counts of the adjacent windows $i-1$ and $i+1$ of window i:

$$\tilde{r}_i = r_i \frac{m}{\frac{1}{4}m^{i-1}_{GC} + \frac{1}{2}m^i_{GC} + \frac{1}{4}m^{i+1}_{GC}} \tag{4.2}$$

4.1.4 Event calling

If the coverage in windows is sufficiently high, the poisson distriubution of reads can be approximated by a normal distribution. Therefore it is crucial that the choosen window size is not too small. The Event-wise testing (Yoon et al. 2009) uses a test statistic based on the normal distribution for the empirical DOC signals. The first step of EWT is the conversion of the DOC signal in each window to z-scores by substracting the mean read count of all windows and dividing by the standard deviation. Copy number variable regions result in lower or higher coverage signals than expected. Regions containing amplified sequence are expected to show high z-scores, which are unlikely to occur, with small upper-tail probability $P(Z > z_i)$. Regions that contain deletions have negative z-scores and are also unlikely with small lower-tail probability. For all windows the upper- and lower-tail probability is computed. Then statistical tests are performed separately for deletion and duplication events. If the maximum lower tail probability of consecutive windows is smaller than a significance threshold t, a duplication event is detected. This is done analogously for duplication events. The search is started with 2-sized events and iteratevely adds windows to the interval to evaluate the maximal tail probability. With increasing event sizes I use already computed probabilities dynamically for the detection of the maximum probability. The threshold increases with larger event size l, i.e. becoming less stringent for larger events, with L being the number of windows in a chromosome and FPR the fixed false positive rate:

$$t = \left(\frac{FPR}{\frac{L}{l}} \right)^{\frac{1}{l}} \tag{4.3}$$

The iteration over l is stopped when t reaches 0.5, an arbitrary treshold reported by Yoon et al. (2009).

4.1.5 Event Merging

The detected events are restricted in size by the stopping criterion of the event-wise testing procedure. Therefore a merging step is necessary to fuse nearby events. I established a merging procedure for predicted events of the same type (deletion or duplication) in two steps. First, overlapping and nearby events that are at most d_{max} apart are collapsed. Local variability of coverage might result in small events ($< 1kb$) that infact belong to the same variant but could not be merged, because they are more than d_{max} apart. I group small events (at most s_{max} base pairs large) into clusters by nearest neighbor clustering with at most r_{max} base pairs distance of the events in one cluster. The events of each cluster are joined to their spanned region. The merging process can be influenced by the user with the d_{max}, s_{max} and r_{max} parameters (200 bp, 500 bp and 500 bp in default setting).

4.1.6 Selection of dimorphic events

If two samples are given, it is useful to detect relative copy number changes in both, because this step removes falsely detected events that are in fact repeats or assembly errors in the reference genome. A t-test is used in order to determine whether the read counts of merged events are significantly different in test and control dataset. If this is the case, a (dimorphic) copy number variant is called. This step is omitted if no control sample is given (one-sample mode, Fig. 4.1).

4.1.7 Filtering

The filtering step is applied to reduce the number of false positives in the results. The results can be filtered by their p-values according to the test statistic on DOC signals or by the p-value of the t-test for dimorphic events. They can also be filtered by the ratio of the DOC signal in the event call relative to the overall mean DOC signal (copy number ratio) and the absolute difference of copy numbers in dimorphic events. Filter parameters are listed with their default values in tab. 4.1. In the pipeline this step is optional and the unfiltered event set is also exported by the program.

Parameter	Description	Default value
Main options		
fasta	sequences in FASTA format	hg18.fa
wig	wiggle file (uniquely mapped positions of sequence)	hg18.wig
controlId	control sample Id	0
tag-length	read length of sequence reads	50
Algorithm options		
window-size	size of fixed windows	100
dynamic	use of dynamic-sized windows	0
mappable-positions	minimal number of uniquely mappable positions (if dynamic is set)	100
mapping-qual	minimal mapping quality of reads	0
normalize-gc	apply GC-normalization procedure	0
merge-dist	maximal distance for merging adjacent events	200
small-event-size	maximal size of small events for clustering	500
small-event-dist	maximal distance of small events for clustering	500
Filter parameter		
max-pval	significance level for CNVs	$1e-6$
dimorph-p-val	t-test p-value for dimorphic events	0.001
difference-threshold	absolute difference threshold of copy number for dimorphic events	0.5
cn-ratio	minimal copy number ratio to base level	0.75

Table 4.1: Parameter for the CNV detection tool copyDOC

4.2 Implementation of a CNV simulation platform

To evaluate the performance of the implemented tool I established a simulation environment for CNVs on sequencing data (*copySim*). For simplicity, CNVs are assumed to be uniformly distributed in the genome with random size range and copy number. CopySim consists of three main steps, the simulation of CNVs, the manipulation of a random or user-defined template sequence and the simulation of single-end sequencing reads (Fig. 4.4).

First, deletions and duplications are sampled randomly on a given template sequence with a user-

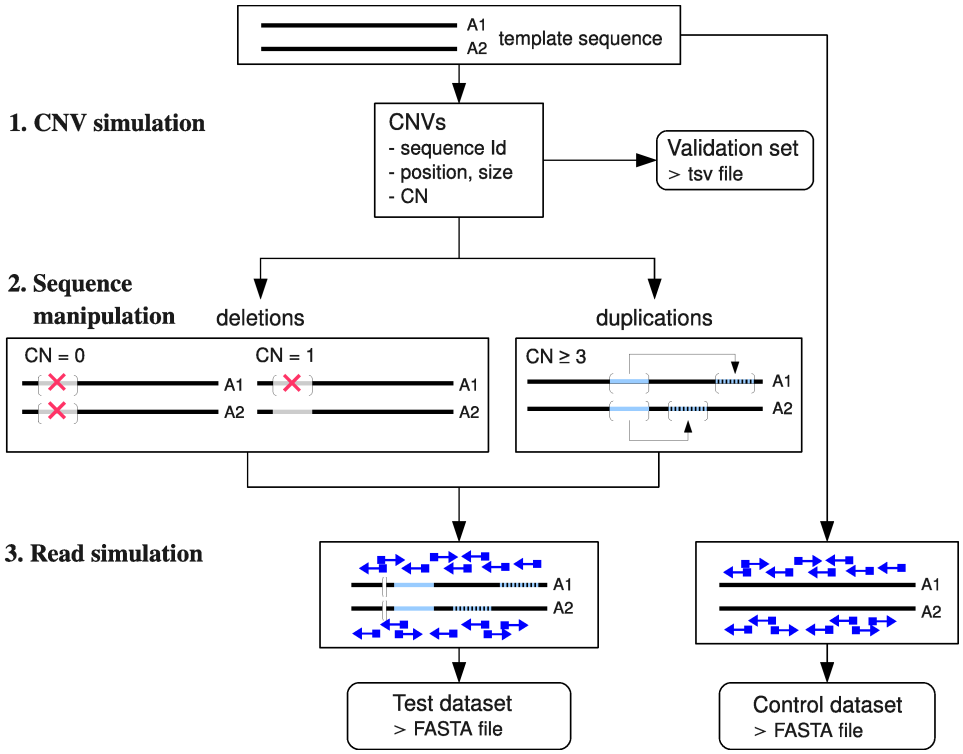

Figure 4.4: Workflow of the CNV simulation platform CopySim that includes three steps: 1. **CNV simulation.** deletions and duplications are sampled on a sequence with random size and copy number (CN) and saved in the validation dataset file. 2. **Sequence manipulation.** The simulated variants are introduced in one (CN = 1, CN = 3) or two copies (CN = 0, $CN \geq 3$) of the template sequence. **3. Read simulation.** Fragment reads are sampled on the manipulated and unaffected sequence and saved in FASTA-format.

defined size range and copy number range. These CNVs are introduced in two copies of the template sequences, that represent the two alleles of a genome. Deletions are simulated either in both alleles, given a copy number (CN) of zero or in one randomly chosen allele ($CN = 1$). Duplications with $CN \geq 3$ are added by insertion of sequence at $CN - 2$ random, nonoverlapping positions in one allele. In the last step single-end reads are simulated on the manipulated and original sequence with user-defined parameters for mismatch probabilities, number of reads, read length and maximal errors in one

read. The read is sampled with randomly chosen allele, position and orientation (forward or reverse). The output of the simulation platform consists of a test and a control read dataset in FASTA-format and a file that contains the validation set for the simulated CNVs (tsv-format). The simulation platform can be used with different settings that are defined by the parameters listed in Tab. 4.2 e.g. maximal variant size (at least 1 kb), number of variants.

Parameter	Description	Example
num-cnvs	number of simulated reads	100
min-size	minimal variant size in bp	1000
max-size	maximal variant size in bp	10000
min-copy-number	minimal copy number	0
max-copy-number	maximal copy number	4
tag-length	read length of sequence reads	50
num-reads	number of simulated reads	1000000
max-errors	maximal number of sequencing errors in a read	2
error-dist	text-file with error frequency at each position in read	uniform error 0.01 (1%)
source-length	length of random sequence if template sequence is not given	10000000 (10 Mb)

Table 4.2: Parameter for the CNV simulation platform copySim with examples.

4.3 Interface setup for additional tools

4.3.1 Input data conversions

I applied four tools to the detection of copy number variants in sythetic data (CNV-seq, DNAcopy, SegSeq and the SOLiD-CNV-Tool). They implement different approaches (section 3.2.4) and use various data input formats that require adaptation of the input. Therefore I produced a script for each tool that converts the SAM alignment format to its input format, which I describe in the following section.

CNV-seq. CNV-seq requires files that contain the best mapping locations for each sequence read with chromosome and genomic location separeated by tabulator (best-hit location file). SAM files are converted in best-hit location files using an awk command that writes the third and fourth column of the SAM alignment file in the best-hits file.

DNAcopy. DNAcopy was originally used with log ratios of intensity values in microarray data. To use it with sequence data I computed log ratios of read counts in fixed windows for the test (r_t) and

control dataset (r_c), normalized by the total coverage of each dataset (N_t and N_c):

$$nlogR = log2\left(\frac{r_t}{r_c}\right) * \frac{N_c}{N_t} \tag{4.4}$$

Then I used the normalized log ratios ($nlogR$) as input for the DNAcopy package in R and followed the instructions for the Genome Segmentation Program *segment* [3].

SegSeq. SegSeq requires chromosome, genomic position and strand of aligned reads for a test and control dataset and a textfile that contains the file locations. I wrote a perl scripts that converts a SAM file in the SeqSeq input format and creates the textfile with the file information. SegSeq is implemented in Matlab.

SOLiD-CNV-Tool. The SOLiD tool was implemented exclusively for SOLiD data and in the current version it is limited to NCBI build 36. It uses SOLiD GFF files as input format that contain uniquely mapped reads. It excludes CNVs within a defined distance of the centromeres and telomeres. Therefore the location of the p-arm and q-arm of chromosomes must be defined in an input file (cmap-file). I created a perl script that converts a SAM file into SOLiD GFF format. I omitted quality information from the aligned reads, because they are not used by the SOLiD tool.

4.3.2 Parameter settings

In table 4.3 I specify all parameters used in the comparison of copyDOC with other tools in section 5.3) on the synthetic dataset. In most cases I used default parameters, otherwise I choosed parameters that seemed to be most suitable.

4.3.3 Sensitivity and Specificity calculations

CNV-seq, SegSeq, SOLiD-CNV-Tool output the start and end positions of the predicted variants, but DNAcopy produces a list of DNA segments with mean and standard deviation of log ratios of read counts. From this list I filtered segments with size at most 1 Mb and mean larger than 0.1. Only SOLiD-CNV and copyDOC infer a copy number for the predicted variant. SegSeq outputs a copy ratio for the two datasets and CNV-seq and DNAcopy log ratios of coverage. The results were evaluated by determination of true positive rates (TPR) and false positive rates (FPR), which I define in the following

[3] http://www.bioconductor.org/packages/2.3/bioc/manuals/DNAcopy/man/DNAcopy.pdf

Tool	Parameter	Description	Value
CNV-seq	–minimum-windows-required	minimum number consecutive windows	1
	–log2-threshold	threshold for log2 values	1
	–p-value	*p*-value threshold	0.00001
	–genome-size	genome size required for approximation of window size	62435964
copyDOC	–window-size	size of fixed windows	100
	–dimorph-p-val	significance level for dimorphic events	0.001
	–difference-threshold	absolute difference threshold for dimorphic events	0.5
DNAcopy	–alpha	significance level for the test to accept change-points	0.008
	–nperm	number of permutations used for p-value computation	50000
	–p.method	method for p-value computation	hybrid
	–undo.splits	specifys how change-points ar undone, e.g. "sdundo" undoes splits that are less than SDs apart	"sdundo"
	–undo.SD	number of SDs between means to keep a split if undo.splits="sdundo"	1
SegSeq	–W	size of local windows (i.e. number of consecutive normal reads)	400
	–a	number of false positive candidate breakpoints for initialization	1000
	–b	number of false positive segments for termination	10
SOLiD-CNV	–coverage-format	format for the aligned reads	GFF
	–trim-distance	distance in kb to be trimmed from chromosome ends	0
	–window-size	window size	1000

Table 4.3: Parameter for the comparison of copyDOC with available tools in section 5.3

section.

Def. *True positive.* A predicted variant is considered as a true positive, if it overlaps with one CNV in the validation set.

Def. *True positive rate (TPR).* The percentage of true positives relative to the number of variants in the validation set.

Def. *False positive.* A predicted variant is considered as a false positive if none of the true variants are overlapping with it.

Def. *False positive rate (FPR).* The percentage of false positives relative to the total number of predicted variants.

The copy number is not verified for the determination of the TPR, because it is not given by all tools.

4.4 Data sets

4.4.1 Simulated data

To evaluate the performance of the implemented tool, more precisely its sensitivity and specifity for the detection of CNVs in sequencing data, I generated synthetic data using the simulation platform from section 4.2. It consists of 100 duplications and deletions ranging from 1 kb to 10 kb with different copy numbers (0,1,3,4) on chromosome 20 (US National Center for Biotechnlogy Information build 36 reference sequence). I simulated 50-bp reads from chromosome 20 with at most 2 sequencing errors and remapped them to chromosome 20 using Bowtie with parameter –y –best –strata –chunkmbs 256 –m 1 –k 1 –l 50 –n 2 –f –S. With this parameter setting only uniquely mapped reads are reported. With different amounts of simulated reads (1-10 million) I obtained data sets with 0.7-7x haploid coverage on chromosome 20. To increase statistical power, each simulation was repeated 100 times.

4.4.2 European parent-child trio

For each of the three samples I downloaded approximately 150 million paird-end sequences of length 36 to 41 bp that were mapped to chromosome 1 from the US National Center for Biotechnlogy Information (NCBI) Short Read archive. The whole dataset constitutes about 30-fold sequence coverage of the human genome. This dataset was sequenced with the Illumina 1 G Analyzer and mapped to the NCBI build 37 reference using MAQ.

4.4.3 Tumor cell lines

I used two SOLiD sequencing datasets that were derived from tumor cell lines (unpublished data from Dr.Dr.M.R.Schweiger). The cells were taken from tumor tissue and metastases of a patient with colorectal cancer (American Tissue Culture Collection).The 50mer reads were aligned with the Applied Biosystems mapping tool iMAP v 0.2.5.3 in classic mode allowing 5 mismatches per read on NCBI hg18 reference genome. Both datasets contain exclusively uniquely mapped reads that have a 4-fold genomic coverage on hg18.

5 Results

5.1 Evaluation of the performance in synthetic data

For a first test of the implemented tool sythetic data (described in sec. 4.4.1) were used that make some simplifying assumptions about the composition of sequencing data e.g. uniform distribution of reads on the sequence. The evaluation of sensitivity and specifity for detection of CNVs was done by computing the average number of predicted copy number variants that overlap a simulated variant (true positives) and the average number of predicted variants that are not included in the simulated data (false positives) in 100 random duplications and deletions on chromosome 20. I repeated the analysis for different physical coverages on chromosome 20 in the range of 0.7 to 7-fold. At 7-fold coverage, the implemented method is able to predict 96% of the variants with 18% false positives relative to all predictions (FPR), see Tab. 5.1. I obtained these results with the unfiltered call set. With reduced coverage the sensitivity decreases to 35% at 0.7-fold coverage. I split the simulated CNVs in size bins to

coverage	mean TPR (%)	sd TPR (%)	mean FPR (%)	sd FPR (%)
0.7	34.5	7.5	60.8	7.9
1.4	55.1	5.0	39.6	7.7
2.8	80.8	3.7	23.3	4.4
4.3	90.4	2.9	21.9	3.6
5.7	93.4	2.7	18.6	4.1
7.1	95.6	2.1	18.2	3.9

Table 5.1: Mean and standard deviations for TPR and FPR of the implemented method (copyDOC) on data sets with different coverage

analyze the sensitivity for detection of deletions and duplications at different variant size (Fig. 5.1).The true positive rate drops substantially for variants that are smaller than 2 kb to around 70% at 7-fold coverage. Deletions (copy number 0-1) are slightly easier to predict than duplications (copy number 3-4). But there are also many more falsely predicted deletions than duplications. The most false positives are smaller than 2 kb.

Furthermore I examined the correlation of predicted and true variant size at 7-fold coverage (Fig. 5.2).

CopyDOC can approximate roughly the true variant size of deletions, but is unable to determine the size for duplications.

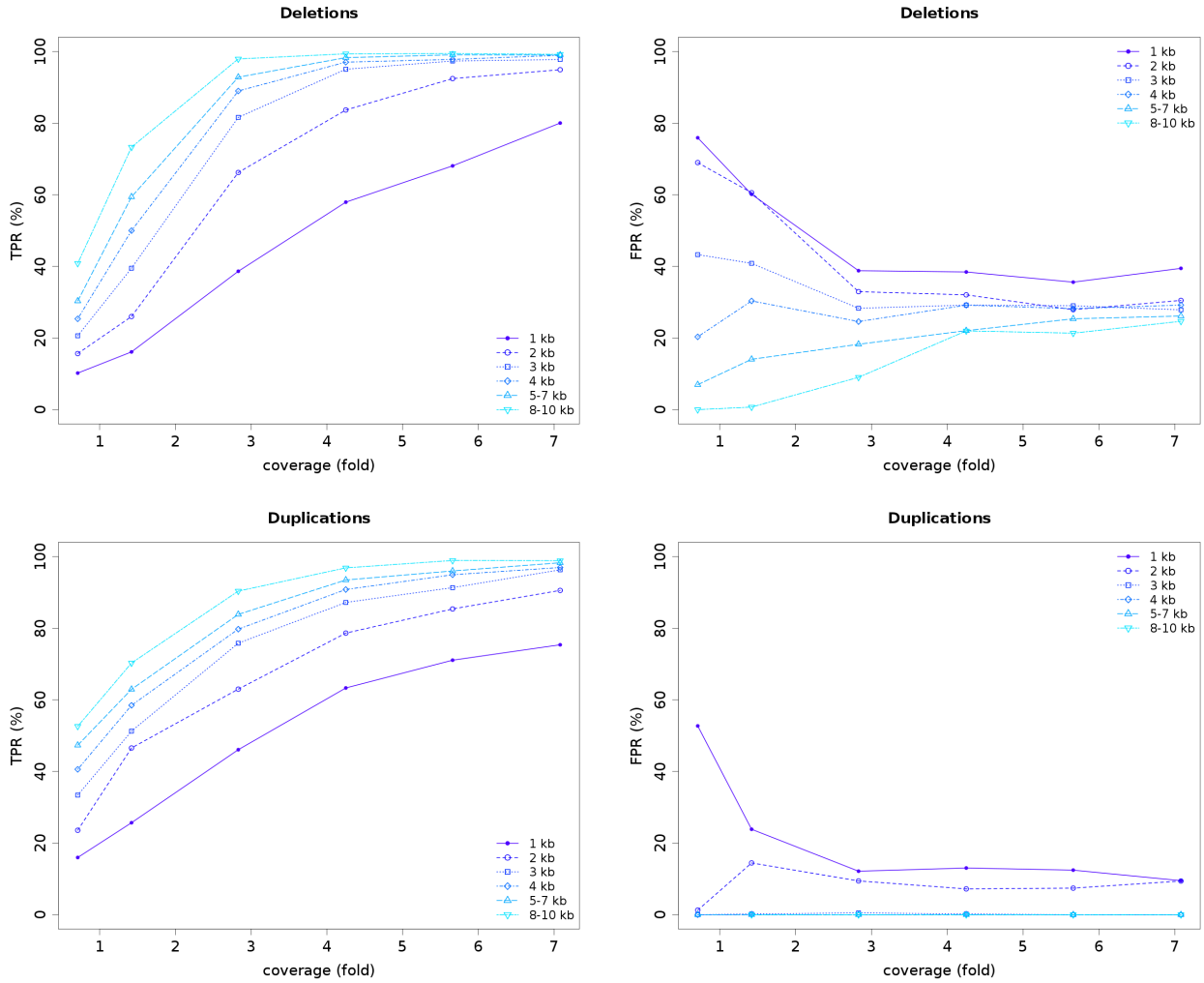

Figure 5.1: True positive rate (TPR) and false positive rate (FPR) w.r.t. coverage of copyDOC at a window size of 100 bp for deletions and duplications; CNVs are grouped in different size bins: 1 kb ($<$ 2 kb), 2 kb ($\geq$ 2 and $<$ 3 kb), 3 kb ($\geq$ 3 and $<$ 4 kb), 4 kb ($\geq$ 4 and $<$ 5 kb), 5-7 kb ($\geq$ 5 and $<$ 7 kb), 8-10 kb ($\geq$ 8 and $<$ 10 kb). TPR for each bin was computed relative to the number of true variants of that size range. FPR was determined by counting the false positives for a given size bin and divide it by the number of all predicted variants of this size.

5.2 Parameter sensitivity

The sensitivity of the implemented tool strongly depends on the window size parameter as can be seen in Fig. 5.2 for different coverages. The sensitivity is almost cut by half with a tenth of the window

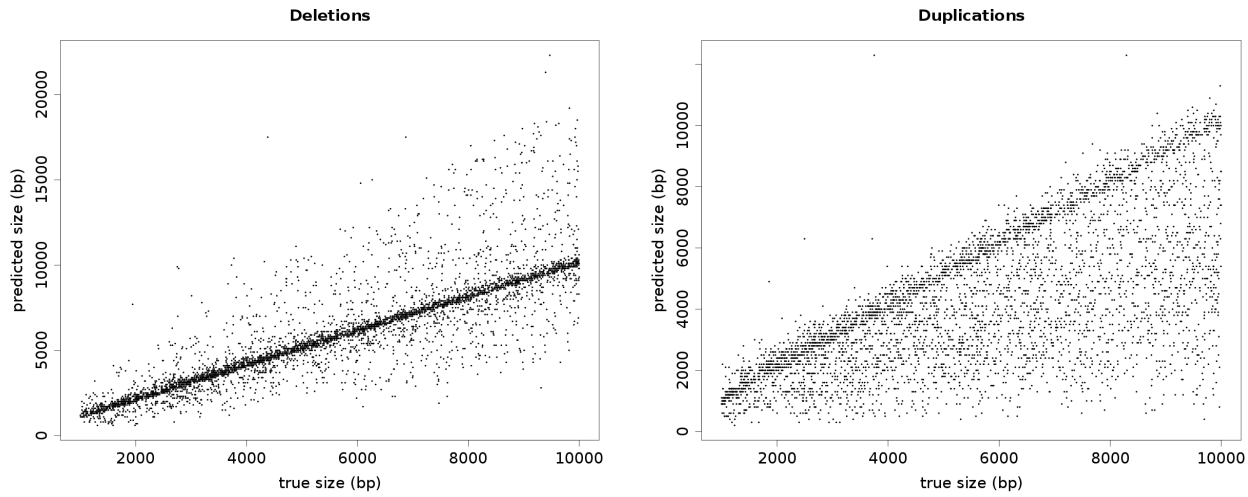

Figure 5.2: Correlation plot between predicted and true variant size in 4809 deletions and 4656 duplications.

size at 7-fold coverage. This observation might be due to small variants in the simulated size mixture. Theoretically, the program is not able to detect variants that are smaller than twice the window size, because searched events span at least two adjacent windows (see methods, sec. 4.1.4). The specificity of copyDOC strongly depends on the window parameter at low coverage ($0.7x$, $1.4x$), see Fig. 5.2. With higher coverage (>4) the window size has only marginal influence on the specificity.

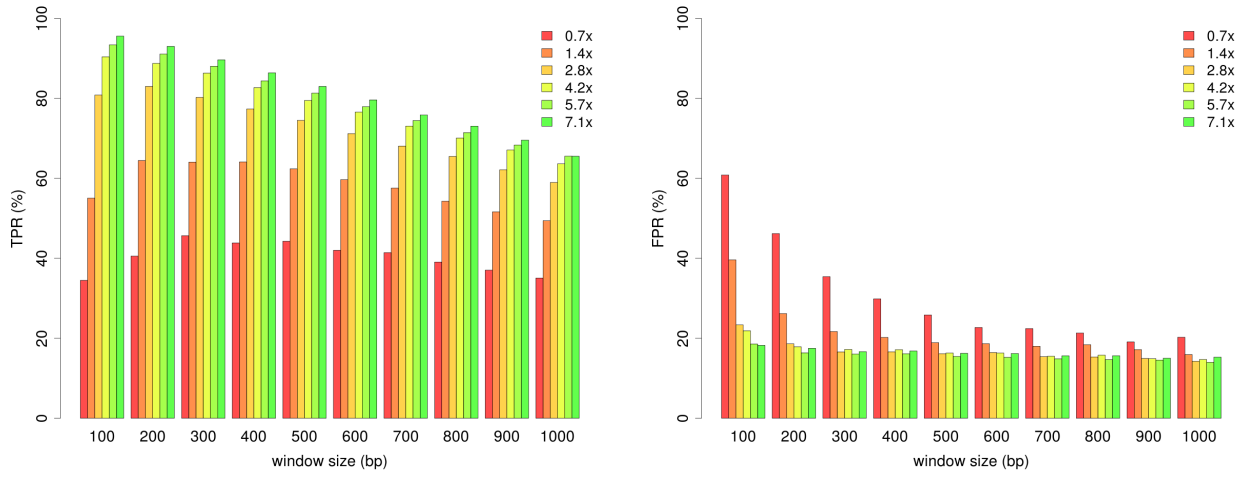

Figure 5.3: True positive rate (TPR) and false positive rate (FPR) given various fixed window sizes and coverage.

5.3 Comparison with other tools

I compared the implemented method with four published CNV detection tools, namely CNV-seq, DNAcopy, SegSeq and SOLiD-CNV. Since these tools are developed for different sequencing platforms and applications, a straightforward benchmark test is difficult and goes beyond this work. I used them for an additional performance test of my implementation. Each tool was run on the simulated data from section 4.4.1 using default parameters in most cases (exceptions in sec. 4.3.2) and the true positive rate (TPR) and false positive rate (FPR) was evaluated for all tools.

coverage	copyDOC w=100	copyDOC w=400	CNV-seq	DNAcopy	SegSeq	SOLiD-CNV
0.7	34.5 ± 7.5	43.8 ± 5.6	7.1 ± 3.4	0.8 ± 0.8	7.3 ± 4.0	19.6 ± 4.4
1.4	55.1 ± 5.0	64.1 ± 3.7	27.1 ± 4.8	5.7 ± 2.4	34.9 ± 5.3	43.0 ± 4.9
2.8	80.8 ± 3.7	77.4 ± 4.3	48.9 ± 4.4	46.5 ± 5.4	65.7 ± 4.3	59.2 ± 4.7
4.3	90.4 ± 2.9	82.7 ± 3.7	60.1 ± 4.8	66.6 ± 5.4	79.6 ± 3.9	67.7 ± 4.3
5.7	93.4 ± 2.7	84.4 ± 3.6	66.0 ± 4.8	76.5 ± 4.9	86.0 ± 3.7	72.7 ± 4.1
7.1	95.6 ± 2.1	86.4 ± 3.3	70.6 ± 4.9	84.8 ± 3.9	90.4 ± 2.8	76.2 ± 4.6

Table 5.2: Mean and standard deviations of TPR for the implemented method (copyDOC) with window size 100 bp and 400 bp and other available tools.

coverage	copyDOC w=100	copyDOC w=400	CNV-seq	DNAcopy	SegSeq	SOLiD-CNV
0.7	60.8 ± 7.9	29.9 ± 6.7	48.2 ± 15.7	2.3 ± 14.0	13.9 ± 31.6	20.5 ± 10.5
1.4	39.6 ± 7.7	20.2 ± 5.3	37.2 ± 7.8	7.8 ± 13.1	6.1 ± 5.8	19.7 ± 6.6
2.8	23.3 ± 4.4	16.6 ± 4.8	39.9 ± 6.4	17.7 ± 6.8	10.8 ± 6.0	17.7 ± 6.7
4.3	21.9 ± 3.6	17.1 ± 4.3	40.6 ± 5.8	20.7 ± 5.0	13.3 ± 5.0	17.8 ± 5.0
5.7	18.6 ± 4.1	16.1 ± 4.9	43.3 ± 3.3	22.6 ± 5.3	13.7 ± 5.0	15.9 ± 5.3
7.1	18.2 ± 3.9	16.8 ± 4.5	44.3 ± 3.2	25.8 ± 4.6	15.5 ± 4.3	16.6 ± 4.7

Table 5.3: Mean and standard deviations of FPR for the implemented method (copyDOC) with window size 100 bp and 400 bp and other available tools.

The implemented tool (copyDOC) detected the most variants compared with the other tools at a coverage $\geq$ 3-fold (see Tab. 5.2 and Fig. 5.4). At 7-fold coverage and using 100 bp windows copyDOC detected 96% of the simulated variants, using 400 bp windows 86%. The other tools predicted 70.6% (CNV-seq), 85% (DNAcopy), 90.4% (SegSeq) and 76.2% of the variants at the same coverage. The larger sensitivity of copyDOC is likely due to the smaller window size compared to those of the other approaches: SegSeq (400 bp consecutive reads in control), SOLiD-CNV (1 kb) and CNV-seq (adaptive windows e.g. 1 kb at 7-fold and 10 kb at 0.7-fold coverage). The high sensitivity of copyDOC comes with the expense of a poor specificity at low coverage (1-2x) e.g. 40% FPR at 1.4-fold coverage (tab. 5.3). The specificity is considerably improved with a window size of 400 bp (20% at 1.4x). SegSeq performs better with respect to false positives compared to copyDOC (16% and 19% at 7-fold coverage) and the other tools. If the sensitivity and specificity is considered at the same time the performance of copyDOC

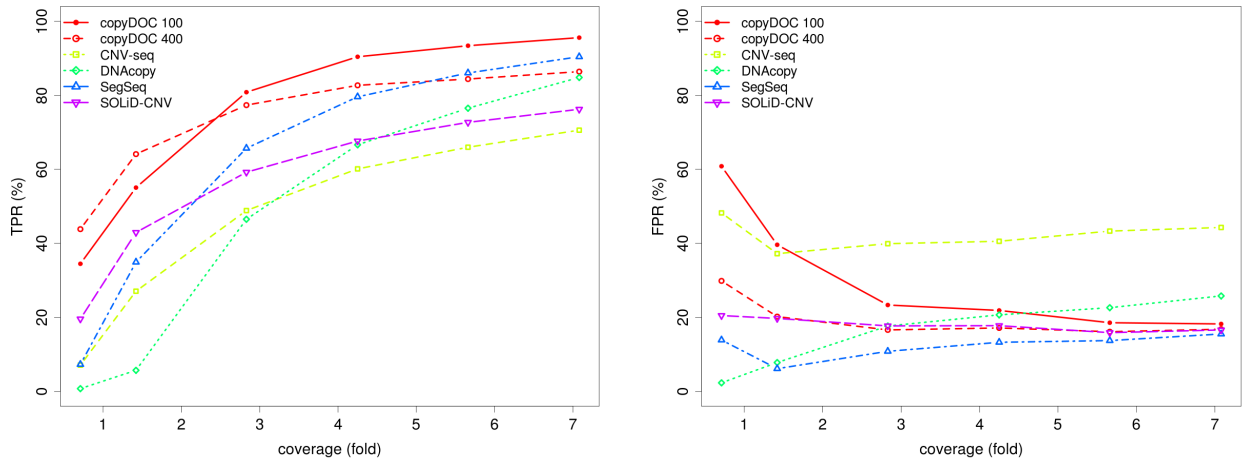

Figure 5.4: True positive rate (TPR) and false positive rate (FPR) w.r.t coverage for the implemented tool at window size 100 bp (copyDOC 100) and 400 bp (copyDOC 400) and four published tools (CNV-seq, DNAcopy, SeqSeg, SOLiD-CNV) with parameter settings from 4.3.2.

with a window size of 400 bp is comparable to that of SegSeq at a coverage of at least 3-fold. A low coverage is problematic for the specificity of copyDOC independent of the window size.

5.4 Application

5.4.1 1000 genomes project data

In order to analyse its performance on real data copyDOC was applied to data from chromosome 1 of the parent-child trio (1000 genomes project data). Each dataset covers chromosome 1 at approximately 30-fold with all mapped reads and 20-fold with uniquely mapped reads (tab. 5.4). I used the program in two-sample mode with copyDOC parameters –window-size 100, –mapping-qual 30 –normalize-gc (otherwise defaults from Tab. 4.1) to detect copy number variants that differ in the two samples. This parameter setting results in exclusion of reads with a mapping quality below 30. The filtered reads include only uniquely mapped reads. The program took 5h 20 min to process two datasets with 200 million reads each. The bottleneck in the pipeline of copyDOC is the first step, the determination of depth of coverage signals from the data, because of the import of the huge amount of data. The program called about 10000 events for each run, the majority of them are duplication events (tab. 5.5). The filtering step removes half of this set resulting in about 2000 deletions and duplications in chromosome 1.

dataset	reads (all)	coverage on chr1	reads mq≥30	coverage unique reads on chr1
NA12878	201240699	29.3x	158896695	23.1x
NA12891	186100776	27.1x	146307842	21.3x
NA12892	164033228	29.3x	124335270	18.1x

Table 5.4: 1000G trio dataset chromosome 1.

test	control	deletion events	duplication events	filtered deletions	filtered duplications
NA12878	NA12891	867	8927	732	1637
NA12878	NA12891	2737	8927	2487	2132
NA12891	NA12892	2711	8926	2461	2132

Table 5.5: Detected events for the three runs with uniquely mapped reads.

For a second run I used the same datasets with no filtering by mapping quality. The results are shown in table 5.6. CopyDOC predicts less CNVs (about overall 1000) than in the previous parameter setting. This is notably striking in the predicted duplications (about 600 versus 9000).

test	control	deletions	duplications	filtered deletions	filtered duplications
NA12878	NA12891	493	417	285	120
NA12878	NA12892	1703	593	949	262
NA12891	NA12892	1415	582	742	253

Table 5.6: Detected events for the three runs with all reads.

I counted the number of predicted CNVs that overlap an entry in the database of genomic variants (DGV) for both runs. Using uniquely mapped reads the overlap of predicted CNVs that are at least 10 kb large with DGV is at most 56% (Tab. 5.7). It is lower ($31-39\%$), if also smaller CNVs (≤ 5 kb) are considered. The overlap with DGV is significantly higher in the second run for all datasets, e.g. 83% overlap with DGV entries of CNVs ≥ 10 kb in NA12891/NA12892 (Tab. 5.8). Including multireads has a considerable influence on the result in this dataset.

5.4.2 Tumor cell lines

I tested copyDOC in single-sample mode with two whole-genome SOLiD datasets from tumor cell lines (sw480 and sw620, unpublished data from Dr.Dr.M.R.Schweiger). The mapped reads were exported in GFF format by the Applied Biosystems mapping tool iMAP and converted in SAM format using the GFF conversion tool (matogff) from Applied Biosystems. The datasets consist of uniquely mapped reads on NCBI hg18 (about 4-fold coverage, tab. 5.9).

I applied copyDOC on both datasets with 1 kb windows and GC-Normalization. It predicted 4704 and 2997 deletions in the filtered call sets of sw480 and sw620 (Tab. 5.10). In both datasets more duplications

test	control	min mapqual	min CNV size	CNVs (filtered)	copyDOC vs. DGV (%)
NA12878	NA12891	30	2 kb	2344	722/2344 (31%)
		30	5 kb	412	133/412 (32%)
		30	10 kb	79	39/79 (49%)
NA12878	NA12892	30	2 kb	4450	1480/4450 (33%)
		30	5 kb	901	349/901 (39%)
		30	10 kb	135	76/135 (56%)
NA12891	NA12892	30	2 kb	4434	1477 (33%
		30	5 kb	893	348 (39%)
		30	10 kb	134	75 (56%

Table 5.7: CNV concordance with DGV for program run with uniquely mapped reads.

test	control	min mapqual	min CNV size	CNVs (filtered)	copyDOC vs. DGV (%)
NA12878	NA12891	0	2 kb	404	294/404 (73%)
		0	5 kb	244	197/244 (81%)
		0	10 kb	129	110/129 (85%)
NA12878	NA12892	0	2 kb	1210	743/1210 (61%)
		0	5 kb	412	308/412 (75%)
		0	10 kb	179	149/179 (83%)
NA12891	NA12892	0	2 kb	994	658/994 (66%)
		0	5 kb	404	303/404 (75%)
		0	10 kb	177	147/177 (83%)

Table 5.8: CNV concordance with DGV for program run with all reads.

than deletions were detected (9405 and 6122). I validated the results with array data (Affymetrix SNP 6.0 array, unpublished data from Dr.Dr.M.R.Schweiger). The array data includes very large variants, on average 15 Mb large. I determined the percentage of predicted CNVs from array data that overlap the copyDOC result with equal variant type (deletion or duplication/amplification), see Tab. 5.11. If I select predicted CNVs that are at least 10 kb large they are consistent with 79% and 82% of the CNVs in the array dataset. The concordance is lower when variants $\geq$ 100 kb are considered (50% and 57%). This is due to the fact that the predicted CNVs are larger in the array dataset compared with the result on sequencing data.

dataset	uniquely mapped reads	coverage on hg18
sw480	218117242	3.5x
sw620	237796278	3.9x

Table 5.9: coverage of tumor cell line data.

Dataset	Deletions	Duplications	Filtered deletions	Filtered duplications
sw480	6525	10188	4704	9405
sw620	3462	6549	2997	6122

Table 5.10: Detected events for the tumor cell lines.

Dataset	Min CNV size	CNVs (filtered)	Array result	Array vs. copyDOC (%)
sw480	2 kb	14109	58	46/58 (79%)
	10 kb	14064	58	46/58 (79%)
	50 kb	14064	58	43/58 (74%)
	100 kb	583	54	27/54 (50%)
sw620	2 kb	9119	78	65/78 (83%)
	10 kb	9113	78	64/78 (82%)
	50 kb	1734	75	56/75 (75%)
	100 kb	826	69	39/69 (57%)

Table 5.11: CNV concordance with array data.

6 Discussion

The evaluation of sensitivity of the implemented program (copyDOC) demonstrated that it is able to detect copy number variants with high sensitivity in 2GS data with a TPR of $81-96\%$ at a coverage of $2.8-7.1x$ (see Tab.5.1). The number of false positives at this coverage range is relatively high ($18-23\%$). At low coverage ≤ 2 the sensitivity is decreased to 34% and there are much more false positives (61%). The program is not able to infer the variant size for duplications (Fig. 5.2), i.e. the true size is underestimated. Since the sensitivity for detection of duplications is similar to that of deletions (Fig. 5.1), the poor size prediction is due to insufficient merging of duplication events. Using the simulated data by the implemented copySim environment, that inserts duplications randomly in the template sequence, it is in general difficult to predict the exact breakpoints of the duplication in the sequence. The window parameter has a significant influence on the number of correctly predicted variants. It has a strong impact on the number of false positives in low coverage data ($\leq$ 2-fold), see Fig. 5.3.

In comparison with four published tools that are also based on analysis of depth of coverage, the copyDOC program performs relatively good with respect to sensitivity and specificity (Tab. 5.2, Tab. 5.3, Fig. 5.4). The parameters for the evaluated tools were choosen according to unsystematic performance tests, but they might not represent the optimal parameter setting for this dataset. For the CNV-seq program I was not able to find a parameter setting that decreases the high FPR at 40%. Except the window size I did not examine the influence of filter parameters in copyDOC, which might improve specificity. In practical applications the user would not test different parameters and is unable to determine the FPR. Thus parameters should not have a major influence on the result.

The number of false positives seems to be very high (around 20%). This can be explained by repeats that are contained in the used un-masked template sequence (chromosome 20), which are not removed by the copySim platform. The number of predictable variants might also be influenced by repeats, as far as the analysis is done on uniquely mapped reads.

The copyDOC program could be successfuly applied to real datasets from Illumina and SOLiD. For the trio dataset from 1000 genomes project the results were dependent on whether unique reads or all reads of the dataset was used resulting in $49-56\%$ and $83-85\%$ overlap of the predicted variants ($\geq$ 10 kb)

with entries in DGV (Tab. 5.7, Tab. 5.8). A second application in two tumor cell line datasets confirmed that the program can get on a single dataset. The evaluation of the results with a true positive set based on array data resulted in 79% and 82% concordance with respect to the array data set.

7 Conclusions and outlook

In this work a program for detection of CNVs in sequencing data based on depth of coverage was implemented in C++ (copyDOC). Single steps in the pipeline, the acquisition of DOC signals in windows, the event calling and merging are implemented using generic programming techniques that enable the future integration of other algorithms in the pipeline. Furthermore, a testing environment was implemented, the copySim platform, which is very useful for testing and evaluation of different algorithms. CopyDOC was successfully applied to synthetic and real data using constant sized windows. Dynamic windows, that adapt according to the local mappability of the sequence, are implemented in the pipeline, but could not be tested in this work. They might be advantageous in datasets that contain uniquely mapped reads. However, CNVs have been shown to be overrepresented in segmental duplications (Nguyen et al. 2006; Cooper et al. 2007) and by a general exclusion of multireads those CNVs might be difficult to ascertain. In the application of copyDOC to a 1000 genomes dataset the overlap of predicted variants was considerable higer using multireads compared to uniquely mapped reads. Thus there is a requirement for tools that can handle multireads.

Futher improvements of copyDOC might be done for the CNV calling algorithm and the merging step. For example the program workflow could be tested with a direct comparison of the DOC signals in two datasets via log ratios instead of appling a t-test on DOC signals in the two datasets. CopyDOC and copySim could be used as platform for the implementation and evaluation of futher CNV detection algorithms.

Acknowledgements

I thank Dr. Ralf Herwig for providing the opportunity to do this work in his group at the Max-Plack Institue for Molecular Genetics and support during my thesis. I thank Prof. Knut Reinert for supervision and support of this work and Anne-Katrin Emde for supervision, valuable advice and help in all stages of this work. Futhermore I would like to thank Marcus Albrecht and Anja Thorman for helpful comments and correction of the manuscript. I thank Dr. Dr. Michal-Ruth Schweiger for providing the tumor cell line datasets.

Bibliography

(1995). *Kernel smoothing*. Chapman & Hall. 9

(2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45. 2

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science*, 297(5583):1003–7. 5

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Cheetham, R. K., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Catenazzi, M. C. E., Chang, S., Cooley, R. N., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fajardo, K. V. F., Furey, W. S., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Jones, T. A. H., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ng, B. L., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Pinkard, D. C.,

Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Rodriguez, A. C., Roe, P. M., Rogers, J., Bacigalupo, M. C. R., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Sohna, J. E. S., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9. 15

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles, M. E., Edwards, P. A., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–29. 5, 14

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*, 39(7 Suppl):S16–21. 7

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009). Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6(9):677–81. 15

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, 6(1):99–103. 3, 14, 15

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12. 2, 5

Cooper, G. M., Nickerson, D. A., and Eichler, E. E. (2007). Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, 39(7 Suppl):S22–9. . 5, 39

Dellinger, A. E., Saw, S.-M., Goh, L. K., Seielstad, M., Young, T. L., and Li, Y.-J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res*, 38(9):e105. 8

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 36(16):e105. 3, 11, 20

Ewart, A. K., Morris, C. A., Atkinson, D., Jin, W., Sternes, K., Spallone, P., Stock, A. D., Leppert, M., and Keating, M. T. (1993). Hemizygosity at the elastin locus in a developmental disorder, williams syndrome. *Nat Genet*, 5(1):11–6. 6

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97. **The article gives an overview of all types of structural variants in the human genome, including microsopic and submicroscopic variants like copy-number variants (deletions, duplications), insertions, inversions and translocations and the methods which are used to discover them.** 4

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., and Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome Res*, 16(10):949–961. 5

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O'Connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., Sen, L., Dolan, M. J., and Ahuja, S. K. (2005). The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science*, 307(5714):1434–40. 5, 6

Gribble, S. M., Kalaitzopoulos, D., Burford, D. C., Prigmore, E., Selzer, R. R., Ng, B. L., Matthews, N. S. W., Porter, K. M., Curley, R., Lindsay, S. J., Baptista, J., Richmond, T. A., and Carter, N. P. (2007). Ultra-high resolution array painting facilitates breakpoint sequencing. *J Med Genet*, 44(1):51–8. 7

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70. 2

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32. 11, 20

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., Magrini, V. J., Richt, R. J., Sander, S. N., Stewart, D. A., Stromberg, M.,

andTodd Wylie, E. F. T., Schedl, T., Wilson, R. K., and Mardis, E. R. (2008). Whole-genome sequencing and variant discovery in c. elegans. *Nat Methods*, 5(1179):183–188. 11, 20

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P. C. M., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J. A. L., and Schalkwijk, J. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet*, 40(1):23–5. 6

Hurles, M. E., Dermitzakis, E. T., and Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends Genet*, 24(5):238–45. 2, 5

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee., C. (2004). Detection of large-scale variation in the human genome. *Nat Genet*, 36(10):949–951. 7

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64. 15

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M., and Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–6. 15

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J., and Altshuler, D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*, 40(10):1253–60. 8

Kumar, R. A., KaraMohamed, S., Sudi, J., Conrad, D. F., Brune, C., Badner, J. A., Gilliam, T. C., Nowak, N. J., Cook, E. H. J., Dobyns, W. B., and Christian, S. L. (2008). Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet*, 17(4):628–38. 6

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Songgang, Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272. 11

Lupskic, J. R., de Oca-Lunaa, R. M., Slaugenhaupte, S., Pentaoa, L., Guzzetta, V., Traskf, B. J., Saucedo-Cardenasa, O., Barkerg, D. F., Killiand, J. M., Garciah, C. A., Chakravartie, A., and Patel, P. I. (1991). Dna duplication associated with charcot-marie-tooth disease type 1a. *Cell*, 66(2):219–32. 5, 6

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B., Fiebig, A., Schreiber, S., Friedman, J., Ketelaars, C. E. J., Vos, Y. J., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., Sloman, L., Summers, A., Gibbons, C. A., Teebi, A., Chitayat, D., Weksberg, R., Thompson, A., Vardy, C., Crosbie, V., Luscombe, S., Baatjes, R., Zwaigenbaum, L., Roberts, W., Fernandez, B., Szatmari, P., and Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet*, 82(2):477–88. 6

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 40(10):1166–74. 5, 8

McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., Perkins, D. O., Dickel, D. E., Kusenda, M., Krastoshevsky, O., Krause, V., Kumar, R. A., Grozeva, D., Malhotra, D., Walsh, T., Zackai, E. H., Kaplan, P., Ganesh, J., Krantz, I. D., Spinner, N. B., Roccanova, P., Bhandari, A., Pavon, K., Lakshmi, B., Leotta, A., Kendall, J., Lee, Y.-H., Vacic, V., Gary, S., Iakoucheva, L. M., Crow, T. J., Christian, S. L., Lieberman, J. A., Stroup, T. S., Lehtimaki, T., Puura, K., Haldeman-Englert, C., Pearl, J., Goodell, M., Willour, V. L., Derosse, P., Steele, J., Kassem, L., Wolff, J., Chitkara, N., McMahon, F. J., Malhotra, A. K., Potash, J. B., Schulze, T. G., Nothen, M. M., Cichon, S., Rietschel, M., Leibenluft, E., Kustanovich, V., Lajonchere, C. M., Sutcliffe, J. S., Skuse, D., Gill, M., Gallagher, L., Mendell, N. R., Craddock, N., Owen, M. J., O'Donovan, M. C., Shaikh, T. H., Susser, E., Delisi, L. E., Sullivan, P. F., Deutsch, C. K., Rapoport, J., Levy, D. L., King, M.-C., and Sebat, J. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet*, 41(11):1223–7. 6

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek,

J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., Vega, F. M. D. L., and Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, 19(9):1527–41. 3, 11, 14, 15

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, 6(11 Suppl):S13–20. 11

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46. 2, 9, 11

Morozova, O. and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–64. **This paper gives an overview of the broad applicability of the three established 2GS methods (454, Illumina, ABI/SOLiD) in functional genomics and shortly introduces the underlying sequencing technologies focussing on their advantages and drawbacks in comparison to state-of-the-art Sanger sequencing**. 2

Nguyen, D.-Q., Webber, C., and Ponting, C. P. (2006). Bias of selection on human copy-number variants. *PLoS Genet*, 2(2):e20. 5, 39

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–72. 3, 8, 14, 15

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bader, G. D., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bolte, S., Bolton, P. F., Bourgeron, T., Brennan, S., Brian, J., Bryson, S. E., Carson, A. R., Casallo, G., Casey, J., Chung, B. H. Y., Cochrane, L., Corsello, C., Crawford, E. L., Crossett, A., Cytrynbaum, C., Dawson, G., de Jonge, M., Delorme, R., Drmic, I., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B. A., Folstein, S. E., Fombonne, E., Freitag, C. M., Gilbert, J., Gillberg, C., Glessner, J. T., Goldberg, J., Green, A., Green, J., Guter, S. J., Hakonarson, H., Heron, E. A., Hill, M., Holt, R., Howe, J. L., Hughes, G., Hus, V., Igliozzi, R., Kim, C., Klauck, S. M., Kolevzon, A., Korvatska, O., Kustanovich, V., Lajonchere, C. M., Lamb, J. A., Laskawiec, M., Leboyer, M., Couteur, A. L., Leventhal, B. L., Lionel, A. C., Liu, X.-Q., Lord, C., Lotspeich, L., Lund, S. C., Maestrini, E., Mahoney, W., Mantoulan, C., Marshall, C. R., McConachie, H., McDougle, C. J., McGrath, J., McMahon, W. M., Merikangas, A., Migita, O., Minshew, N. J., Mirza, G. K., Munson, J., Nelson, S. F., Noakes, C., Noor,

A., Nygren, G., Oliveira, G., Papanikolaou, K., Parr, J. R., Parrini, B., Paton, T., Pickles, A., Pilorge, M., Piven, J., Ponting, C. P., Posey, D. J., Poustka, A., Poustka, F., Prasad, A., Ragoussis, J., Renshaw, K., Rickaby, J., Roberts, W., Roeder, K., Roge, B., Rutter, M. L., Bierut, L. J., Rice, J. P., Salt, J., Sansom, K., Sato, D., Segurado, R., Sequeira, A. F., Senman, L., Shah, N., Sheffield, V. C., Soorya, L., Sousa, I., Stein, O., Sykes, N., Stoppioni, V., Strawbridge, C., Tancredi, R., Tansey, K., Thiruvahindrapduram, B., Thompson, A. P., Thomson, S., Tryfon, A., Tsiantis, J., Engeland, H. V., Vincent, J. B., Volkmar, F., Wallace, S., Wang, K., Wang, Z., Wassink, T. H., Webber, C., Weksberg, R., Wing, K., Wittemeyer, K., Wood, S., Wu, J., Yaspan, B. L., Zurawiecki, D., Zwaigenbaum, L., Buxbaum, J. D., Cantor, R. M., Cook, E. H., Coon, H., Cuccaro, M. L., Devlin, B., Ennis, S., Gallagher, L., Geschwind, D. H., Gill, M., Haines, J. L., Hallmayer, J., Miller, J., Monaco, A. P., Nurnberger, J. I. J., Paterson, A. D., Pericak-Vance, M. A., Schellenberg, G. D., Szatmari, P., Vicente, A. M., Vieland, V. J., Wijsman, E. M., Scherer, S. W., Sutcliffe, J. S., and Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–72. 5, 6

Quinlan, A. R. and Marth, G. T. (2007). Primer-site SNPs mask mutations. *Nat Methods*, 4(3):192. 11

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–54. 5

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Meur, N. L., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T., and Campion, D. (2006). App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*, 38(1):24–6. 6

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–8. 7

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., and Eichler, E. E. (2005).

Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet*, 77(1):78–88. 4, 5

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., Lincoln, S., Crawley, A., Hanson, M., Maraganore, D., Adler, C., Cookson, M. R., Muenter, M., Baptista, M., Miller, D., Blancato, J., Hardy, J., and Gwinn-Hardy, K. (2003). alpha-Synuclein locus triplication causes parkinson's disease. *Science*, 302(5646):841. 5, 6

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavare, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–53. 2, 5

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V., and Eichler, E. E. (2005). Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–32. 15

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., and8 Mel Simon and9 Carolyn Slayman and0 Michael Hunkapiller, R. J. R., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I.,

Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351. 2

Wang, L.-Y., Abyzov, A., Korbel, J. O., Snyder, M., and Gerstein, M. (2009). MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res*, 19(1):106–17. 9

Xie, C. and Tammi, M. T. (2009). Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10(1):80. 3, 13, 15, 19

Xu, B., Roos, J. L., Levy, S., van Rensburg, E. J., Gogos, J. A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet*, 40(7):880–5. 5, 6

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–71. 16

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, 19(9):1586–92. 3, 14, 18, 21, 22

Selbständigkeitserklärung

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig und nur unter Zuhilfenahme der erwähnten Hilfsmittel angefertigt habe.

Datum

Unterschrift