EXOME ANALYSIS USING NEXT-GENERATION SEQUENCING DATA

MARIA FISCHER

DOCTORAL THESIS

Graz University of Technology
Institute for Genomics and Bioinformatics
Petersgasse 14, 8010 Graz, Austria

Graz, November 2010

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …….. ……..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…….. ……..
date (signature)

Abstract

The introduction of next-generation sequencing (NGS) technologies enables scientists to analyze millions of DNA sequences in a single run. The hereby produced gigabytes of raw data need to be further analyzed in order to gain biological meaningful results. Although NGS has lowered the cost for whole genome sequencing dramatically, its application for high-throughput screening studies still remains expensive. Exome sequencing provides a more cost effective approach where only the protein coding regions of a genome is utilized to find mutations which cause and maintain human diseases.

Spurred by NGS technologies, new efficient and well designed bioinformatics tools emerged which are addressing different tasks in the downstream analysis of NGS data. Since combining these tools into an analysis pipeline greatly facilitates the interpretation of NGS results, an exome sequencing pipeline was developed in this thesis which connects all necessary analysis steps into a unified application. The pipeline supports input data generated by the NGS platforms Illumina and ABI SOLiD™, handles correct execution of all integrated tools, and automatically distributes computational expensive tasks on a high-performance computing (HPC) cluster. It performs quality statistics on raw and processed reads, allows users to trim and filter sequence reads, and aligns the processed reads to a reference genome. Post alignment analysis includes the calculation of alignment statistics, region filtering, and the detection of variants resulting in a list of potential disease driving candidates. The developed pipeline was applied in a joint project with clinical research partners to detect potential causes for Mendelian disorders.

The integration of well established tools and newly developed promising algorithms into a unified solution eases the analysis of NGS data and may provide a valuable method for detecting and investigating therapeutical targets of diseases such as cancer and hereditary disorders.

Keywords: next-generation sequencing, exome analysis, pipeline development, high-performance computing, distributed analyses

Kurzfassung

Durch die Entwicklung von ‘Next Generation Sequencing’ (NGS) wurde die Analyse von Millionen von DNA Sequenzen in einem einzigen Sequenzierdurchlauf ermöglicht. Die auf diesem Wege gewonnen Rohdaten erfordern weitere Analysen um biologisch aussagekräftige Resultate zu liefern. Trotz drastisch gesunkener Kosten für die Sequenzierung vollständiger Genome, bleiben die absoluten Kosten für vergleichende Parameterstudien hoch. Exom-Sequenzierung bietet eine kosteneffizientere Methode, welche nur Eiweiß kodierende Regionen des Genoms zur Detektion von krankheitsauslösenden und -relevanten Mutationen im Menschen heranzieht.

Verschiedenste bioinformatische Werkzeuge wurden entwickelt, um die unterschiedlichsten Aufgaben der Analyse von NGS Daten zu bewerkstelligen. Die gegenwärtige Dissertation beschäftigt sich mit der Kombination einiger dieser Werkzeuge zu einer Analysekette, welche alle notwendigen Analysen in eine einheitliche Applikation vereint. Die hierbei erstellte Software unterstützt Exom-Sequenzdaten der NGS Plattformen Illumina und ABI SOLiD^{TM}, stellt die korrekte Ausführung aller Werkzeuge sicher und verteilt rechnerisch aufwendige Aufgaben auf Hochleistungsrechner. Sie berechnet Qualitätsmerkmale der Sequenzdaten, ermöglicht Trimmen und Filtern von Sequenzen und detektiert die Position der aufgearbeiteten Daten im Referenzgenom. Folgeanalysen beinhalten die Berechnung von Alignment Statistiken, das Filtern anhand der Position im Genom und die Detektierung von Mutationen, welche in eine Liste von potentiellen Krankheitsauslösern resultieren. Die entwickelte Software wurde bereits in einer Kooperation mit einem klinischen Forschungspartner zur Identifikation von potentiellen Ursachen von Erbkrankheiten angewandt.

Die Integration von etablierten sowie innerhalb der gegenwärtigen Arbeit neu entwickelten Algorithmen in eine einheitliche Software erleichtert die Analyse von NGS Daten und kann eine wertvolle Methode zur Detektierung und Erforschung von therapeutischen Targets für Erbkrankheiten und Krebs darstellen.

Stichwörter: Next Generation Sequencing, Exomanalyse, Pipeline Entwicklung, Hochleistungsrechnen, verteilte Analysen

Publications

This thesis was based on following publications, as well as upon unpublished work:

Papers

Stocker G, **Fischer M**, Rieder D, Bindea GL, Kainz S, Oberstolz M, McNally J, Trajanoski Z.: iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis. BMC Bioinformatics 2009, 10:390 PMID: 19941647

Geigl JB, Obenauf AC, Waldispuehl-Geigl J, Hoffmann EM, Auer M, Hörmann M, **Fischer M**, Trajanoski Z, Schenk MA, Baumbusch LO, Speicher MR.: Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. Nucleic Acids Res. 2009, 37 PMID: 19541849

Aleksic K, Lackner C, Geigl JB, Auer M, Ulz P, **Fischer M**, Trajanoski Z, Otte M, Speicher MR.: Evolution of genomic instability in the DEN mouse hepatocellular carcinoma model. submitted (2010)

Fischer M, Stocker G, Ulz P, Speicher MR, Trajanoski Z.: Applied exome sequencing analysis for next-generation sequencing. in preparation (2010)

Conference proceedings and poster presentations

Stocker G, **Fischer M**, Rieder D, Bindea GL, McNally J, Trajanoski, Z.: iLAP: a novel work flow oriented approach for microscopy data management and protocol development. Focus on Microscopy 2008, Osaka-Awaji. 2008 Apr 13.

Fischer M, Haemmerle Hoefler G, Strauss JG, Trauner M, Zatloukal K, Trajanoski Z.: Identification of direct target genes of PPAR alpha in the liver. Doctoral Program Plus (DK-plus) W 1226-B18 Metabolic and Cardiovascular Disease - Review Hearing, Vienna. 2009 Oct 29.

Rieder D, Stocker G, **Fischer M**, Trajanoski Z, Scheideler M, Müller W, McNally J.: Spatial association of multiple coordinately expressed genes during cell differentiation. RECOMB Regulatory Genomics, Systems Biology, DREAM 2009, Boston. 2009 Dec 02.

Fischer M, Stocker, G, Trajanoski Z.: Analytical pipeline for next-generation exome sequencing data. 1st PhD Workshop of the GEN-AU Bioinformatics Integration Network III, Vienna. 2010 May 05.

Contents

Chapter 1

Introduction

1.1 Next-generation sequencing

The discovery of the use of dideoxy nucleotides for chain termination by Sanger et al. [1977] marked a milestone in the history of DNA sequencing. This concept provided a basis for the development of automated Sanger sequencing (Smith et al. [1986], Ansorge et al. [1987]) which has been the method of choice for DNA sequencing for almost 20 years. During this time, the technology has been enhanced to account for longer DNA fragments and for a higher level of parallelism. In its current stage, the technology supports simultaneous sequencing of 1000 base pairs (bp) per DNA fragment in 96 capillaries. Although this method achieved a limited level of parallelization, Sanger-based approaches have not been able to analyze DNA in a high-throughput manner.

Automated Sanger sequencing was the core technology of the Human Genome Project, which was funded in 1990 with the goal of determining all three billion base pairs making up the human genome. The project took ten years to produce first draft results (Lander et al. [2001], Venter et al. [2001]) and an additional three years to complete (Jasny and Roberts [2003]). During the project's final phase and early years thereafter numerous spin-off projects have been launched including the International HapMap Project and the prominent 1000 Genomes Project. The former project aimed at developing a haplotype map of the human genome which describes the common patterns of human DNA sequence variation (International HapMap Project [2006]). The latter concentrated on sequencing the genomes of at least one thousand anonymous participants from a number of different ethnic groups to provide a comprehensive resource on human genetic variation (1000 Genomes [2008]). Both projects were accompanied by the necessity for extensive sequencing. They led, together with a program aiming at the economic sequencing of complete,

high-quality, mammal-sized genomes (Service [2006], Mardis [2006]), to the development of new sequencing technologies. These so-called next-generation sequencing (NGS) technologies allowed sequencing at unprecedented speed in combination with low costs per base (illustrated in figure 1.1). As a consequence, the number of sequencing related data stored in public available databases has increased significantly over the last years and is expected to grow even faster (shown in figure 1.2). Taking advantage of the newly developed machines, the 1000 Genomes Project has accomplished yet to sequence the complete genome of 185 individuals from four populations, and to analyze targeted exons of 697 individuals from seven populations within only two years (Consortium et al. [2010]).

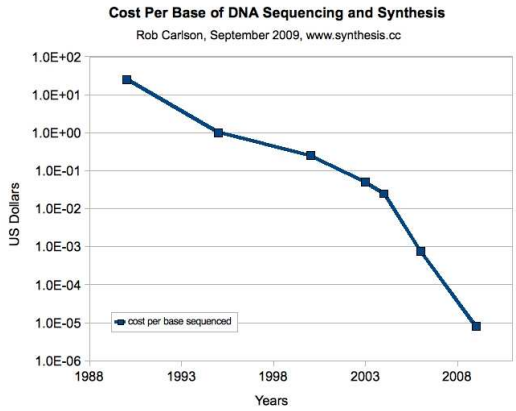

Figure 1.1: Per base cost development of DNA sequencing on a log scale. Prices have fallen from approximately $10 in the beginning of the 1990s to a fraction of a cent in 2009. Figure adapted from Carlson [2009].

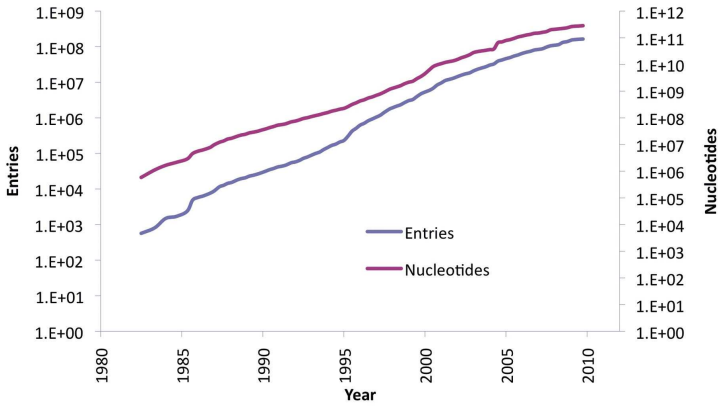

Figure 1.2: Growth rates of database entries and their corresponding number of nucleotides in the European Nucleotide Archive (ENA) shown in log scale. Figure taken from ENA [2010].

As NGS machines generate millions of short sequence reads per run, the bottleneck in sequencing shifted from sequence generation to data management and analysis. Data volume now

represents a major challenge for storage, backup, and analysis. New algorithmic approaches are required to overcome the drawbacks of short read lengths. The development of streamlined, highly automated pipelines for data analysis is critical for the transition from technology adoption to accelerated research and consequent publication (Koboldt et al. [2010]).

1.2 Exome sequencing

Since its early days, medical research has striven for identifying the causes of disorders with the ultimate goal of establishing therapeutic treatments and finding cures. Nowadays, whole genome sequencing (WGS) approaches are designed to discover genetic variations contributing to rare or common diseases. Despite the decrease in sequencing costs, the expenses for routinely obtaining and analyzing full genomes of a large number of individuals remain prohibitive (Hedges et al. [2009]). Alternative methods, focusing on only a fraction of the human genome, represent affordable approaches to identify potential disease-associated genetic variants. Sequencing all protein coding regions of the genome, also referred to as exome sequencing, is the promising candidate as

- it is believed that coding exons harbor most functional variations (Botstein and Risch [2003], Ng et al. [2008]),
- the exome constitutes only about 1% of the human genome requiring to sequence just approximately 30 mega bases (Mb) (Ng et al. [2009]),
- the whole exome sequencing effort is only 1/20 compared to WGS (Ng et al. [2010]), and
- single nucleotide polymorphisms (SNPs) occurring in coding regions, which are a resource for mapping complex genetic traits, are the most common causes for Mendelian disorders (Horner et al. [2010]).

1.3 Bioinformatics tools for exome sequencing data

Due to the high demand of analysis tools for exome sequencing data, numerous programs were developed to support aspects of a typical exome analysis workflow consisting of raw sequence quality control, sequence alignment, alignment postprocessing, and variant detection. TileQC (Dolan and Denver [2008]) and FastQC (Barbraham Bioinformatics [2009]) provide, among others, basic quality statistics for raw sequence data. Alignment programs such as SOAP2 (Li et al. [2009c]), BWA (Li and Durbin [2009]), and Bowtie (Langmead et al. [2009] were especially designed for mapping high amounts of short reads to a reference genome. The most popular tools

for alignment enhancement and variant detection are provided by the Genome Analysis Toolkit (GATK, McKenna et al. [2010]) which is used in the 1000 Genomes Project and The Cancer Genome Atlas (Broad Institute [2010a]).

Given the vast amount of provided tools, choosing an appropriate set of analysis software to obtain high quality results from NGS data is a very challenging and complex task. To overcome this problem, exome analysis pipelines were developed by several groups including Eck et al. [2010], CLCbio GenomicsWorkbench (CLCbio [2010]), and NextGENe (Softgenetics [2010]). However, these pipelines are either incomplete, designed to process only Illumina sequencing data, or not freely available to the scientific community. Therefore, a freely available exome analysis pipeline providing a streamlined exome analysis for SE and PE data generated by Illumina or ABI SOLiD^{TM}platforms would be a major contribution to the field and tremendous help for human geneticists.

1.4 Objectives

The aim of this thesis was to develop and evaluate a pipeline for the detection of SNPs and deletion/insertion polymorphisms (DIPs) within DNA sequences obtained by targeted re-sequencing of a genome's entire set of protein coding regions. The pipeline should be capable of handling NGS data produced by Illumina and ABI SOLiD™ platforms. The application should then be tested with real biological data obtained by clinical research partners.

The specific goals were to develop a pipeline which

- allows processing different kinds of input data to analyze reads encoded either in nucleotide or color space and supports single-end (SE) as well as paired-end (PE) data
- generates measurements for input quality evaluation as well as for sequence alignment and sequence capturing efficiency characterization
- includes automatic detection of SNPs and DIPs and supports SNP ranking
- determines and splits homo- and heterozygous variant calls

Chapter 2

Results

The results of this thesis are presented in two sections. The first consists of a universal pipeline for exome sequencing and its associated pipeline tools, the second describes the obtained pipeline analysis results of two biological samples.

2.1 Exome sequencing analysis pipeline

The main objective of this thesis was the development of a pipeline for investigating exome sequencing data generated by Illumina and ABI SOLiD^{TM}NGS devices (see section 4.1.2 for a detailed description). Therefore, a highly configurable exome sequencing analysis pipeline was developed which meets the challenging requirements of NGS data-handling while still being an easy to use program. The pipeline is based on the Java Platform, Enterprise Edition (JEE) and uses high performance computing (HPC) approaches in conjunction with an inhouse developed cluster application programming interface (API). Since the application distributes and executes all computationally expensive analysis tasks on an HPC infrastructure and provides an easy to use command line interface, the pipeline can be used universally from every PC connected to the HPC server. Furthermore, the pipeline supports the analysis of single-end (SE) as well as paired-end (PE) data and provides the following analysis components (see figure 2.1):

- **read preprocessing** - for checking and enhancing read quality
- **sequence alignment** - for creating and refining read alignment
- **alignment statistics** - for verifying read alignment quality and statistics
- **variant detection** - for identifying and filtering variants

The pipeline requires as input sequence reads, their corresponding base calling quality values, and a list of the re-sequenced exon positions, specifying the exome.

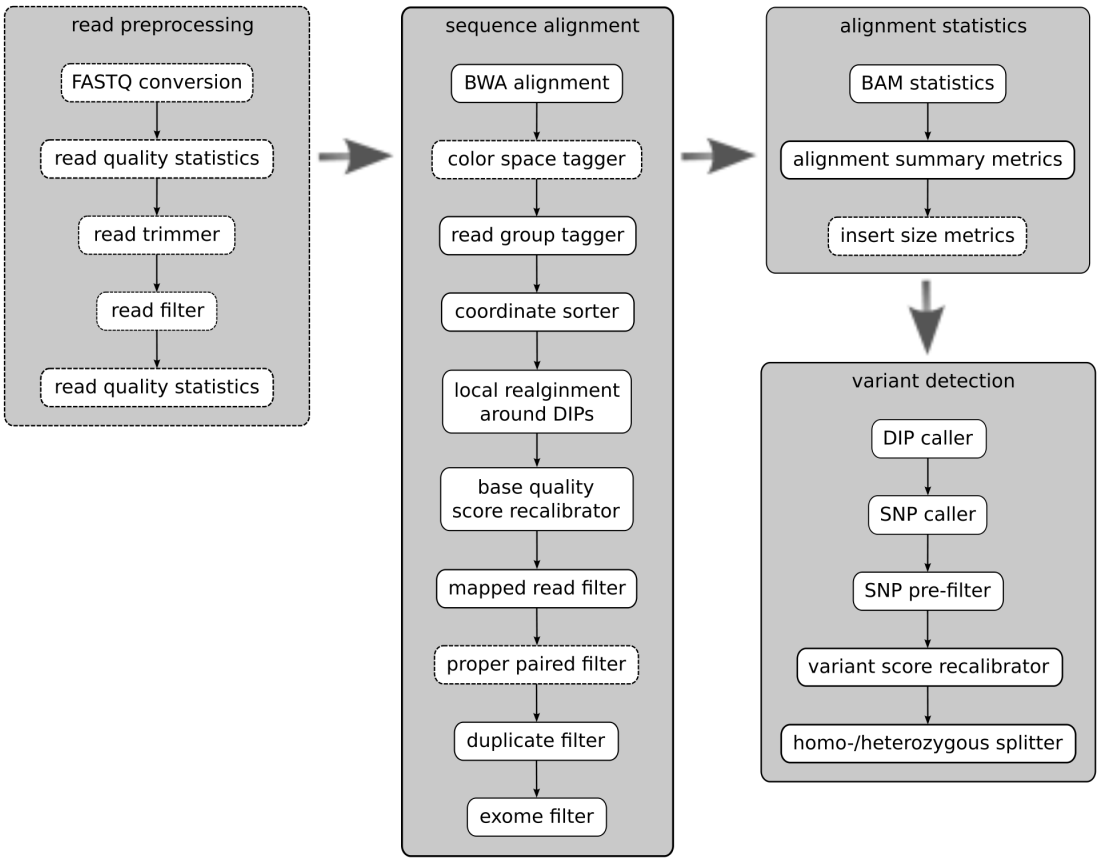

Figure 2.1: Exome sequencing analysis pipeline workflow. Dashed boxes indicate optional analysis steps.

2.1.1 Read preprocessing

This analysis component was developed to offer a first overview of the sequence reads, to allow the user to convert data to standardized file formats, and to enhance the overall read quality. All analysis steps conducted within this component operate on reads stored in the FASTQ file format (definition given in section 4.6.1) and are highly configurable to meet the needs of the different NGS devices and library preparation methods.

FASTQ conversion

Exome sequencing studies conducted on various NGS devices can currently produce three different FASTQ file formats (see section 4.3 for a detailed discussion). The traditional Sanger FASTQ

format is seen as a de facto standard in sequencing and is the file format of choice for submitting sequence data into NCBI's Sequence Read Archive (SRA, Wheeler et al. [2008]). Therefore, the pipeline converts data from Illumina 1.3- and Illumina 1.3+ to Sanger FASTQ.

Read quality statistics

To support quality evaluation of a sequencing run, the following read and read quality characteristics are calculated and reported by the pipeline:

- *number of sequenced reads*
- *read length information* - this characteristic includes minimum, first quartile, median, mean, third quartile, maximum, boxplot and histogram of the read lengths (see figure 2.2).

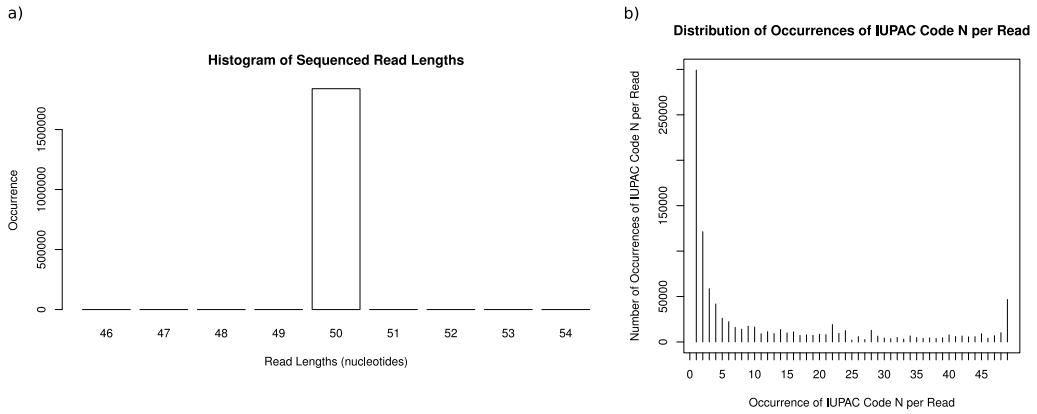

Figure 2.2: Read quality statistics. a) Histogram representing the read length distribution of a sequencing run before read trimming. b) Plot illustrating frequencies of unidentified base calls per sequence read before trimming.

- *base call comparisons* depending on the position within the read - two separate graphs are used for illustrations (shown in figure 2.5). The first one prints the median base calling quality values separately for each nucleotide. The second chart shows the absolute occurrence of a certain nucleotide per position within the read. Both figures include all nucleotide calls represented in the IUPAC codes A, C, G, T, and N.
- *GC, AT, and N content in percent*
- *characteristics concerning unidentified base calls (N) within the reads*
  - number of reads containing no unidentified base calls
  - amount of reads consisting solely of unidentified base calls
  - histogram describing the distribution of number of unidentified base calls within one read (illustrated in figure 2.2)

- plot of the *base calling quality distribution based on the position within the read* for each nucleotide - the quality value distribution is presented in a normalized heatmap where the amount of quality values per position is color encoded (see figure 2.3).

All statistics are calculated on the HPC back-end of the pipeline using the R programming language (R Development Core Team [2010]) and printed to a PDF file.

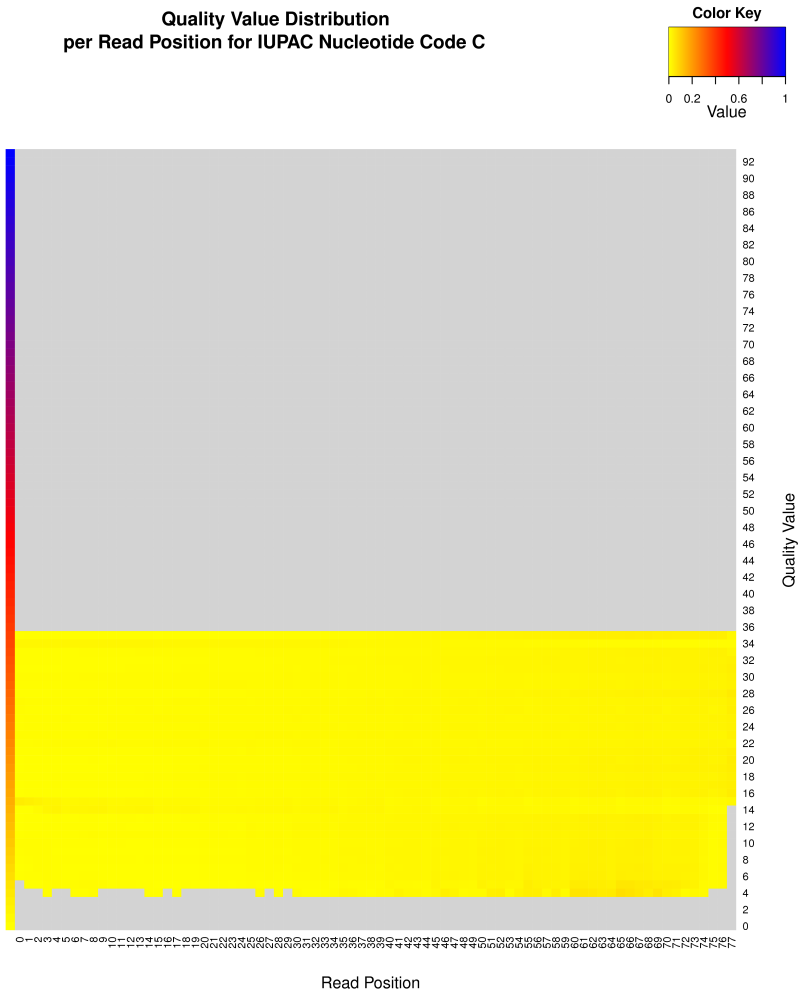

Figure 2.3: Example of base calling quality heatmap showing the quality distribution per read position for cytosines. The color keys on the upper right side and on the vertical left stripe encode the values $\in]0; 1]$. Light gray areas mark quality values that did not occur. For displaying reasons only every other quality value is labeled.

Read trimmer

This preprocessing step, which was implemented as trimmer for I/O streams, allows the trimming of FASTQ entries based on a given read length, nucleotide, or quality value. Read length trimming

causes the truncation of all FASTQ entries after the specified length at the 3' end, whereas the other trimmers are applied at both sites. Nucleotide and quality value based trimming clips all flanking nucleotides and quality values equal to the specified parameters. Figure 2.4 shows an example of using a combination of these two trimmers. In order to allow tracking of all changes, the pipeline logs the number of altered reads in one file and writes all original FASTQ entries of the edited sequences into another file.

```
@GA03_0001:4:1:1068:4935#0/1
NCCACTTCCTAAGNTTCTCTCACACTGAAAGTTAGAAGTGTNTAGGNNNNNNNNNNTCA
+GA03_0001:4:1:1068:4935#0/1
CCBCCC?CCC?88#688888>AA8B?>>ABB>?B>######>############?>?###
```

Input

```
@GA03_0001:4:1:1068:4935#0/1
CCACTTCCTAAGNTTCTCTCACACTGAAAGTTAG
+GA03_0001:4:1:1068:4935#0/1
CBCCC?CCC?88#688888>AA8B?>>ABB>?B>
```

Output

Figure 2.4: FASTQ entry before and after read trimming. A combination of nucleotide and quality trimmer was applied to the input data. 'N' was chosen as nucleotide and '2' as numeric phred value (encoded as '#') for quality trimming.

Read filter

Several read filters, which can be applied in serial, were developed to eliminate short or error prone sequence reads. Similar to read trimmers, all read filters write individual log files containing the number of filtered reads and FASTQ tracking files listing all rejected reads. The following filters are offered:

- **Length filter** - depending on the parameter settings, this filter rejects reads which are either shorter or longer than a specified read length.
- **Quality filter** - this filter parses a read's quality value sequence for the occurrence of a predefined value and rejects all reads containing more than a given amount of this quality value. This threshold is specified either by a universally applied absolute value or by allowing a certain percentage of each read length.
- **Unidentified read call filter** - the filter rejects reads containing more than a specified number of unidentified bases, either absolutely or relatively to each read's length specified.

For PE data, two FASTQ files including the first and second pairs are taken as input files. The filters are applied simultaneously on both files as both reads of a pair must satisfy the criteria to

pass the filter.

Read quality statistics

In order to overview the read quality improvements after trimming and filtering, the edited reads can be characterized again by the same analysis methods as used for raw read quality statistics calculation. Figure 2.5 illustrates the read quality enhancement after fastq trimming and filtering.

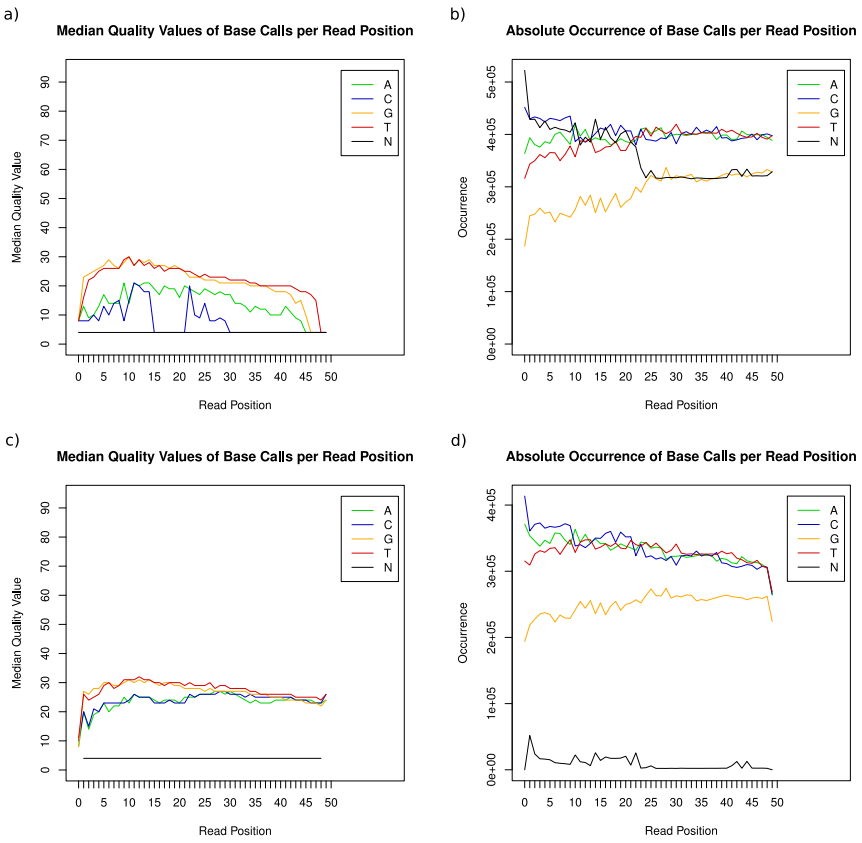

Figure 2.5: Comparisons of base calls before (a, b) and after (c, d) read trimming and filtering. Median base qualities of the raw sequencing run reveal very poor quality at both ends for each nucleotide. Cytosines show additional problems between position 15 and 22 (a). The number of unidentified base calls is shown to be disproportional large (b). Read trimming and filtering greatly enhanced the median base qualities, especially for cytosines and at all 3' ends (c). The number of unidentified base calls decreased to an acceptable size (d).

2.1.2 Sequence alignment

As a prerequisite for detecting variants with NGS, the positions of sequence reads in the reference genome has to be determined by sequence alignment. The developed pipeline supports alignment against the UCSC human genome version hg18 and hg19 of the primary assembly,

including assembled chromosomes, unlocalized sequences, and unplaced sequences. Unlocalized sequences are sequences of an assembly which is associated with a specific chromosome but cannot be ordered or oriented on that chromosome whereas unplaced sequences can not be associated with any chromosome (GRC [2010]).

The developed sequence alignment component takes FASTQ files as input and mainly operates on BAM files (see section 4.6.2). Analysis tools provided by the Genome Analysis Toolkit (described in section 4.7) were integrated into the pipeline for local alignment around DIPs and base quality score recalibration.

BWA alignment

The program BWA (Li and Durbin [2009]) was chosen as short read aligner since it supports alignment in nucleotide and color space, executes gaped alignment in a time efficient way (see section 4.4 for further descriptions), and stores the alignment in SAM format. After alignment, the pipeline directly converts the result files into binary SAM files, also referred to as BAM, to reduce storage usage.

Alignment taggers

The initial alignment does not produce all information required for further downstream analysis. Therefore, alignment taggers were implemented to complement BAM entries with the missing tags.

- **Color space tagger** - BAM files store alignments as a nucleotide sequence, even for reads originating from color encoding NGS platforms. The developed color space tagger adds information about raw color sequence and base calling qualities in form of CS and CQ tags to each BAM entry.
- **Read group tagger** - read group tags are used to group sequence reads to indicate that reads of one group originate from the same DNA sample. This component assigns read group tags derived from the original FASTQ input file name to reads where no read group information is present.

Coordinate sorter

Sequence reads are not sorted in any obvious order after alignment. To ensure time efficient and correct calculations, most analysis tools depend on alignments which are sorted based on coordinate and chromosome order of the reference. This pipeline step prepares the alignment for further processing by ensuring that the correct alignment order is met. As sorting NGS reads

is usually a memory consuming task, the pipeline allows specifying the number of reads stored in RAM to consider the available memory capacities. If PE reads are provided this step will also check and correct any inconsistent SAM PE flag information.

Local realignment around DIPs

The initial alignment of sequence reads may include alignment artifacts due to the suboptimal characteristics of single read alignment algorithms. In contrast to multiple read alignment, single read alignment methods only take data from one individual read instead of the set of all reads into account. The resulting alignment is formally correct but may include false positive SNPs and wrongly aligned DIPs, given the limited information provided. Reads covering the DIP near their 5' or 3' end are particularly likely to be misaligned (example given in figure 2.6).

Multiple local realignment around DIPs corrects alignment artifacts by minimizing the number of mismatching bases across all reads. As multiple local realignment is a time consuming task, only sites likely requiring realignment (specified by a list of already known DIP sites or based on the aligned data itself) are processed. To further improve runtime, the pipeline evenly divides the aligned reads by chromosome sets, executes multiple realignment in parallel, and recomposes the realigned reads to one result BAM file.

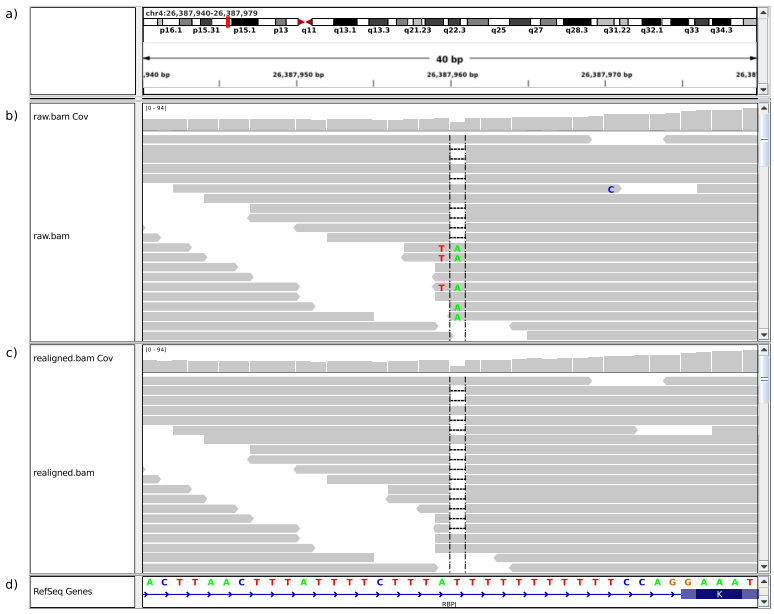

Figure 2.6: Visualization of 40 bp of chromosome 4 in the Integrative Genomics Viewer (IGV, Broad Institute [2010c]) illustrating the effects of local realignment around indels. Position and chromosome ideogram are shown in the upper panel (a) whereas RefSeq gene information is displayed in the lower panel (d). Raw alignment (b) detects eight mismatches in 5 sequence reads between position 26,387,959 and 26,387,960 (shown in red and green) representing false positive SNPs. After local realignment (c), the erroneous mismatches are replaced by a deletion at position 26,387,960.

Base quality score recalibrator

Initial base calling quality calculations introduce bias depending on sequencing cycle and preceding nucleotide (Poplin [2010]). As these quality measurements are used in variant detection methods, a more accurate quality estimation is desired. Therefore, the base quality score recalibrator analyzes and corrects the covariation of the following features of a base call:

- assigned quality value
- position within the sequence read
- preceding and current nucleotide call
- probability of mismatching the reference genome

After performing this recalibration, reports which summarize information about initial and recalibrated base call quality covaritions are generated to facilitate comparison between input and output. Figures 2.7, 2.8, and 2.9 compare certain aspects of initial and recalibrated base calling quality values.

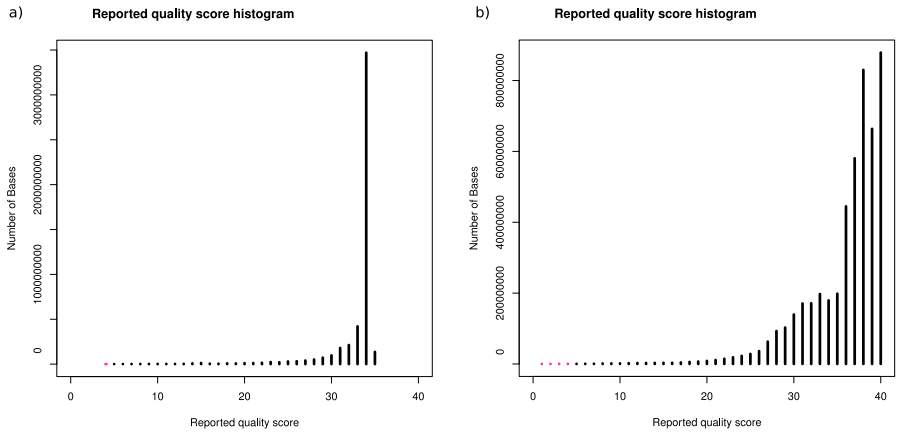

Figure 2.7: Base calling quality histogram before (a) and after (b) recalibration. Recalibrated data shows higher variation in terms of assigned base quality values than initial qualities.

Alignment filters

Several alignment filters were developed to provide mapped, properly paired, non duplicate, and exon spanning reads, as they are required by further downstream analyses. To evaluate filter results, each filter logs the number of rejected and accepted reads and stores filtered and passing BAM entries in two separate files.

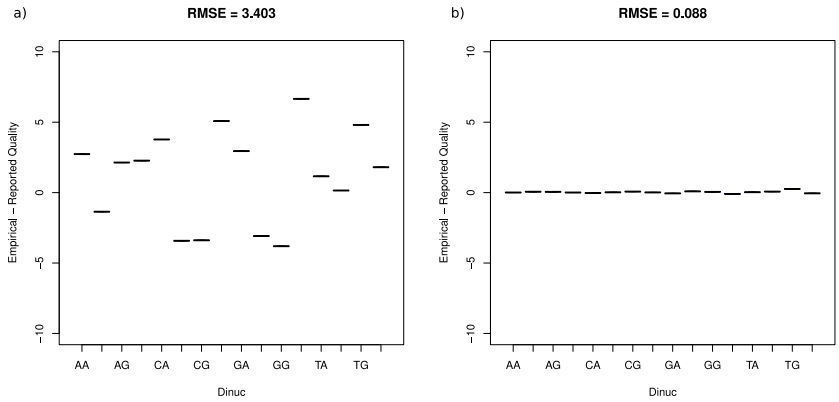

Figure 2.8: Empirical minus reported base calling quality per dinucleotide combination before (a) and after (b) recalibration. Before recalibration, reported quality values greatly differ from expected values which is corrected by recalibration.

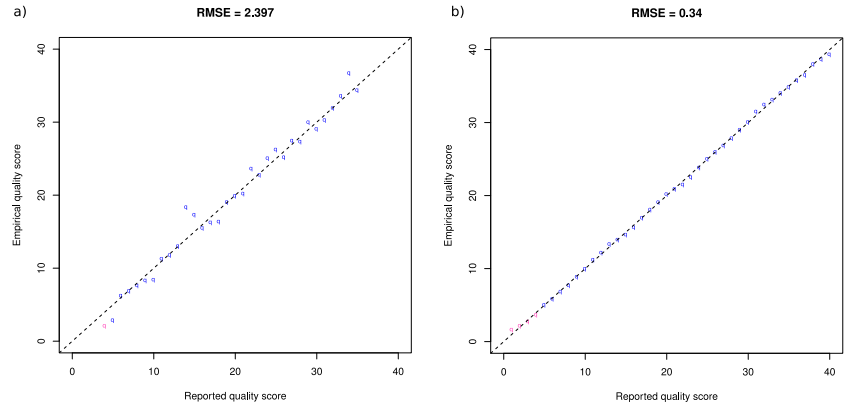

Figure 2.9: Empirical versus reported base calling quality values before (a) and after (b) recalibration.

- **Mapped read filter** - unmapped reads are filtered out at this point in the pipeline.
- **Proper paired filter** - this filter sorts out read pairs which are not within the expected insert size or are aligned in the wrong directions. As SE reads do not contain this information, the filter is only applied on PE data.
- **Duplicate filter** - it is common practice to scan for read duplicates by identifying each read's orientation and 5' core site, where most of the read's bases have been aligned, and comparing all identified characteristics with each other. The pipeline determines the duplicates and filters all reads except the one with the highest sum of base qualities. For PE data, all 5' core sites and orientations of both pairs need to be identical in order to be regarded as duplicates.
- **Exome filter** - exome sequencing aims at the targeted re-sequencing of protein coding re-

gions. As the applied capturing methods may also select DNA fragments originating from non-coding regions, read filtering based on the alignment was implemented. The filter requires a list of exon positions and rejects every read which does not overlap one of these positions. Additionally, capture specificity as defined by Ng et al. [2009] is calculated by the filter.

2.1.3 Alignment statistics

The developed analysis component provides several alignment statistics allowing the user to evaluate the alignment and data quality before variant detection. All steps within this section are applied on reads which passed all precedent filters. If PE library preparation is applied insert size characteristics will be analyzed as well.

BAM statistics

This component provides a quick summary of basic alignment information, including:

- total number of reads
- number of mapped and unmapped reads
- percentages of mapped and unmapped reads with respect to total number of reads
- read coverage with regard to genome size
- mapping quality frequencies

Alignment summary metrics

Additional high level characteristics are reported to support a more fine grained evaluation of the alignment. Therefore, metrics are shown summarized by first read in pair, second read in pair, and aggregated for both reads in pair. Among others, the following information is provided:

- number of reads which solely consist of adenine and/or unidentified bases as they are considered as noise
- median read length
- number and percentage of reads aligned in pairs
- percentage of reads which have been aligned to the forward strand in respect of the total number of aligned reads

- number of high quality aligned reads, i.e. reads with an assigned mapping quality of 20 or higher
- total number of bases of high quality aligned reads
- total number of confidently called bases of high quality aligned reads. A base with calling quality value of 20 or higher is considered to be confidently called.
- the median number of bases mismatching the reference in high quality aligned reads
- the percentage of bases mismatching the reference in high quality aligned reads

Insert size metrics

The pipeline reports information about insert size distribution for PE data by providing the following metrics:

- number of read pairs
- basic statistical measurements including minimum, maximum, mean, standard deviation, and median insert size of all PE reads
- read pair orientation
- bin width of insert size frequencies for histogram plot.

Visualization of the insert size distribution is provided as a histogram in a PDF file (example given in figure 2.10).

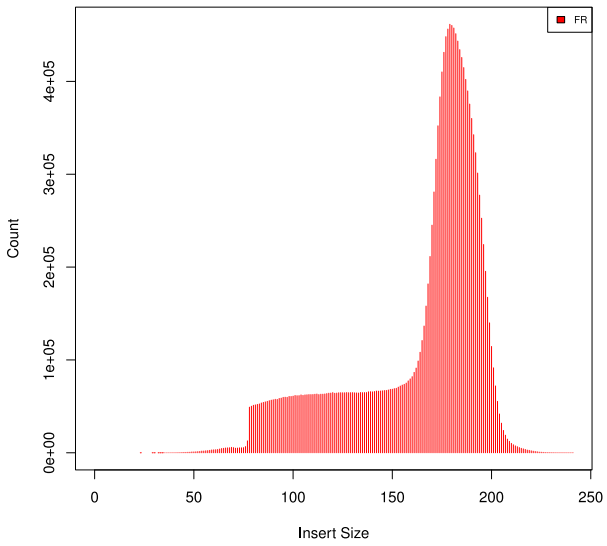

Figure 2.10: Insert size histogram generated by the alignment statistics component. FR indicates that first read in pair is aligned in forward, second read in pair in reverse direction.

2.1.4 Variant detection

The final analysis component deals with the identification of variants while refining all variant calls to improve accuracy. In order to facilitate the search for recessive or dominant causes, the variants are divided by the pipeline into homo- and heterozygous mutations. Variants are detected based on mapped, local realigned, base quality score recalibrated, properly paired, unique, and exon overlapping reads to reduce the number of false positives. DIP calling, SNP identification, and variant score recalibration are realized with analysis components provided by the Genome Analysis Toolkit (see 4.7).

DIP caller

A first set of potential DIPs is detected by combining information including the number of reads covering a DIP site, the number of reference- and DIP supporting reads, read mapping qualities, and mismatch counts. These DIP calls are then filtered based on heuristic cutoffs to remove false positives. The results are reported in form of BED, VCF, and TXT files separately for initial and for filtered DIP sets. To display the results in Genome Browser tracks (see section 4.8), BED files containing basic information about chromosome, start/end positions, number of DIP supporting reads, total number of reads at the site, and the inserted or deleted sequence are generated. VCF files provide further information by comparing consensus supporting and reference supporting reads at the DIP site. The following characteristics are taken into account:

- consensus/reference sequence
- number of supporting consensus DIP reads versus the number of any DIP call at this site, referred to as allele count
- total number of reads at the DIP site
- average number of mismatches per consensus/reference supporting reads
- average mapping quality of consensus/reference supporting reads
- average neighboring base calling quality values from consensus/reference supporting reads
- average neighboring mismatches in consensus/reference supporting reads
- counts of forward- and reverse-aligned consensus/reference supporting reads

The analysis uses RefSeq (Pruitt et al. [2005]) data to annotate DIPs. Variants located within a gene are tagged with the gene's name and meta information about the variant's location (i.e. intron, UTR, coding region), whereas DIPs not overlapping any genes are marked as genomic.

In order to accelerate DIP identification, the pipeline evenly divides the input set and executes DIP identification in parallel.

SNP caller

This analysis step generates a raw set of SNP calls by applying a Bayesian identifier to infer the consensus sequence. Then, potential SNPs are detected by comparing the consensus sequence with the reference genome. Practice has shown that machine artifacts, which appear as a combination of DIP and SNPs, can not be eliminated by local alignment around DIPs (DePristo [2010]). Therefore, DIP masking is applied prior to SNP identification to avoid SNP calling within a user definable window around DIPS.

The program provides an overview by logging the number of visited bases, callable bases, confidently called bases, and actual calls made in a metrics file. Evaluation of the resulting SNP call set is provided by recording dbSNP concordance and transition/transversion ratios of the variants in an evaluation document. Detailed information about each SNP call is given in the generated VCF file which includes total number of reads covering the site, genotype, genotype quality, genotype likelihood provided only for bi-allelic sites, ratio of reference supporting reads to total number of references (referred to as allele balance), number of identified alleles (referred to as allele count), allele frequencies for each allele, root mean square of the mapping qualities of all reads covering the site, number of reads supporting the site with mapping quality zero, dbSNP id if variant is known, and strand bias.

SNP pre-filter

The SNP filter masks ambiguous SNP calls to create an improved SNP call set which is used as training data for variant score recalibration. SNPs near called DIPs, overly clustered SNPs, and SNPs mapping equally well to multiple positions within the reference are considered to be ambiguous. By default, ‘overly clustered’ is defined as having more than 3 SNPs within a 10 bp window.

Variant score recalibrator

Variant score recalibration aims at the improvement of variant scores to provide a more accurate estimate that the detected mutation is an actual true biological variant. Therefore, a subset of the pre-filtered variants is used as training data for clustering, yielding an adaptive error model. HapMap Project and dbSNP data is used to determine true sites within the training set. The resulting error model is used for the recalibration of all raw variants, including the ones which were

rejected in the pre-filtering process. Subsequently, filters are applied on recalibrated variants to detect false positive classified mutations.

Homo-/heterozygous splitter

The final result is represented by two VCF files containing either solely homo- or heterozygous SNP call sets (see figure 2.11 and 2.12 for examples). A developed homo-/heterozygous splitter is applied on variant score recalibrated SNPs and provides additional information about dbSNP concordance and transition/transversion ratios of each call set in separated files.

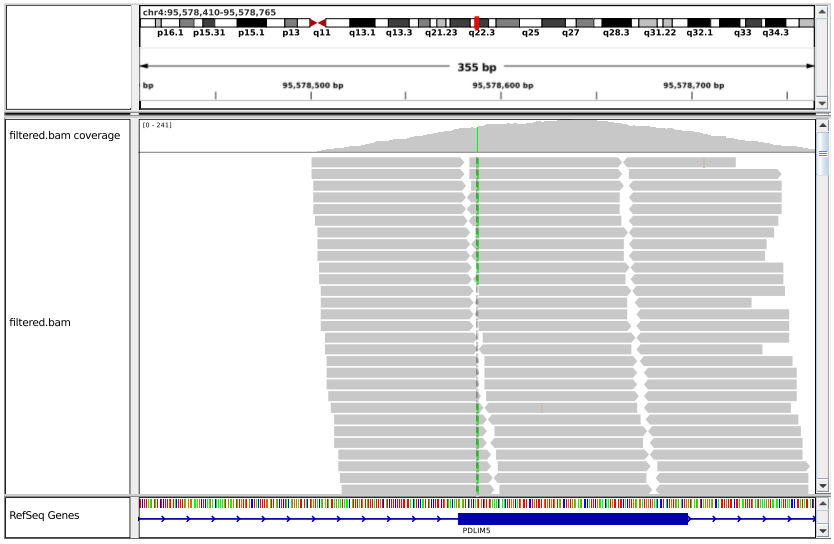

Figure 2.11: Visualization of a homozygous SNP call in IGV. Sequence reads overlapping RefSeq gene PDLIM5 show a homozygous SNP call, displayed in green, in the coverage and sequence track near position 95,578,600. Gray areas indicate matching alignments whereas color codes are used for A (green), C (blue), G (orange), and T (red) mismatches. Homozygous SNPs are displayed by a continuous vertical line in the coverage track.

2.2 Experimental results

The developed pipeline was tested on data sets sequenced with Illumina's Genome Analyzer IIx in combination with PE library preparation. Therefore, two exome samples, which are referred to as S1 and S2, were sequenced with Illumina PE technology resulting in four Illumina 1.3+ FASTQ files. For each sample two files were generated each representing one part of the read pairs. In order to distinguish the files of one sample, first reads are marked by the postscript R1 and second reads by R2. In S1 36,427,610 read pairs were sequenced compared to 36,355,815 read pairs were reported for S2.

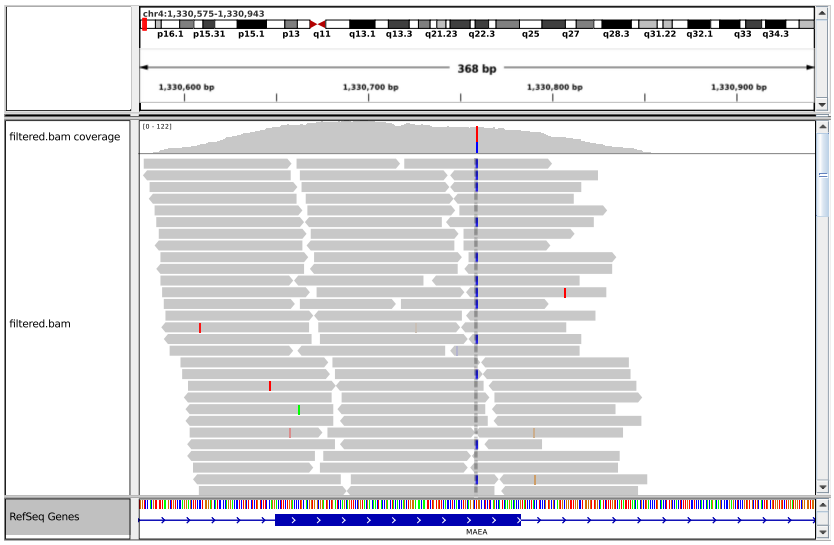

Figure 2.12: Heterozygous SNP call shown in IGV. Parts of the sequence reads show a mutation near position 1,330,750. Since not all reads covering the SNP site contain the mutation, the SNP is identified as homozygous. The coverage track highlights this sort of SNPs by a two colored vertical line (in this case red and blue) which encodes the genotype.

2.2.1 Preprocessing results

The complete set of preprocessing steps was applied by the pipeline. As the input samples were not encoded in standard Sanger format, it was necessary to first apply FASTQ conversion. Read quality statistics processed on the raw data show that each read in each sample is exactly 78 bp long (see figure 2.2). Moreover, a slight decrease in median quality values is reported for each sample. Table 2.1 lists the overall GC content of the samples categorized by first and second reads.

FASTQ input	GC content	AT content	N content
S1_R1	45.93 %	54.03 %	0.04 %
S1_R2	45.79 %	54.09 %	0.12 %
S2_R1	44.94 %	55.02 %	0.04 %
S2_R2	44.84 %	55.03 %	0.13 %

Table 2.1: Table listing overall GC, AT, and N content in percent categorized by sample and first/second read in pair.

The generated heatmaps (see figure 2.13) indicate that the data contains reads where a quality value of '2' is assigned, which is used by Illumina as *Read Segment Quality Control Indicator*. This

indicator is assigned for regions at the start or the end of a read (Mann [2009]). These regions are ignored by further downstream analyses. Therefore, and for enhancing read quality, the quality value '2' and all unidentified base calls were trimmed at both sides of the reads. Between 15 to 30 percent of the reads are affected by trimming (see table 2.2 for a detailed summary).

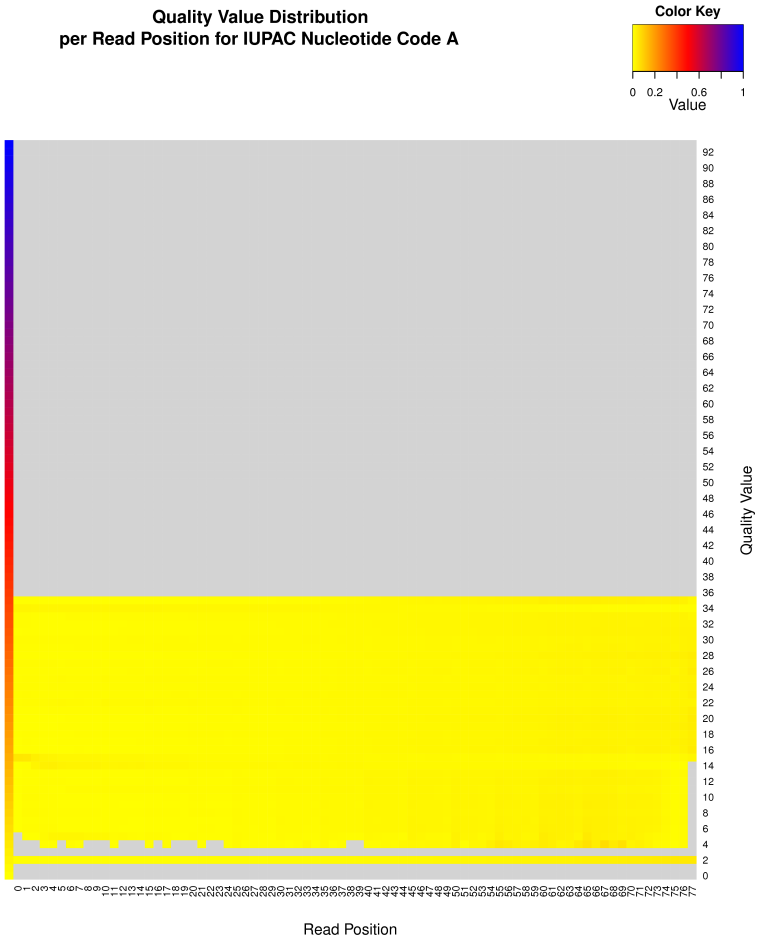

Figure 2.13: Heatmap showing base quality score distribution of raw sequence data of S1_R1 for adenine. Quality values range between two and 35. Light gray areas depict quality values which do not occur in the data.

Read filtering removed read pairs whose reads are shorter than 20 bp or contain more than five percent of unidentified bases in their read sequence. In order to qualify for elimination, it is sufficient that only one of the reads failed the filter. Table 2.3 summarizes the filter results grouped by sample. More than 96 percent of the reads are accepted for further downstream analyses.

FASTQ input	Total number of reads	N & QV Trimmer trimmed	N & QV Trimmer not trimmed
S1_R1	36,427,610	9,854,774 27.05 %	26,572,836 72.95 %
S1_R2	36,427,610	5,869,569 16.11 %	30,558,041 83.89 %
S2_R1	36,355,815	8,738,227 24.04 %	27,617,588 75.96 %
S2_R2	36,355,815	5,104,367 14.04 %	31,251,448 85.96 %

Table 2.2: FASTQ trimming results showing total number of reads in one FASTQ file and the number of trimmed and untrimmed reads as absolute value and percentage. Trimmer was parametrized to trim quality value '2' and IUPAC code 'N'. The table lists statistics for each input file as trimming was applied on each FASTQ file separately.

Sample	Total number of read pairs	Length filter passed	Length filter failed	N filter passed	N filter failed
S1	36,427,610	35,309,589 96.93 %	1,118,021 3.07 %	35,306,212 96.92 %	3,377 0.01 %
S2	36,355,815	35,366,540 97.28 %	989,275 2.72 %	35,362,774 97.27%	3,766 0.01 %

Table 2.3: FASTQ filter results of the applied length and subsequent unidentified base filter (named N filter). The table lists total number of read pairs, number of filtered and passed read pairs as absolute value and percentage. The length filter was parametrized to rejected read pairs with reads shorter than 20 bp whereas the N filter rejected read pairs with more than five percent unidentified bases in one of their reads.

2.2.2 Alignment results

35,306,212 and 35,362,774 read pairs were passed to the alignment program for S1 and S2, respectively. After processing the sequence alignment component, 49.96 (S1) and 46.01 (S2) percent of the alignment input reads are categorized as properly paired, unique, and exon overlapping (see figure 2.14 and tables 2.4 and 2.5 for intermediate filter results), which results in an exome capture specificity of 51.87 (S1) and 46.33 (S2) percent.

Sample	Total number of read pairs	Mapped read filter passed	Mapped read filter failed	Proper paired filter passed	Proper paired filter failed
S1	70,612,424	68,010,027	2,602,397	66,344,318	1,665,709
		96.31 %	3.69 %	93.96 %	2.36%
S2	70,725,548	70,251,286	474,262	69,856,104	395,182
		99.33 %	0.67 %	98.77 %	0.56 %

Table 2.4: Alignment filter results of the applied mapped read and proper paired filter. The table lists the total number of read pairs passed to the sequence alignment program, and number of passed and failed reads per filter. The second row of each sample shows the percentage of accepted and rejected reads in relation to the total number of read pairs passed to the sequence alignment component.

Sample	Total number of read pairs	Duplicate filter passed	Duplicate filter failed	Exome filter passed	Exome filter failed
S1	70,612,424	54,959,346	11,384,972	35,278,875	19,680,471
		77.83 %	16.12 %	49.96 %	27.87 %
S2	70,725,548	53,146,424	16,709,680	32,544,256	20,602,168
		75.14 %	23.63 %	46.01 %	29.13 %

Table 2.5: Alignment filter results of the applied duplicate and exome filter. The table lists the total number of read pairs passed to the sequence alignment program, and number of passed and failed reads per filter. The second row of each sample shows the percentage of accepted and rejected reads in relation to the total number of read pairs passed to the sequence alignment component.

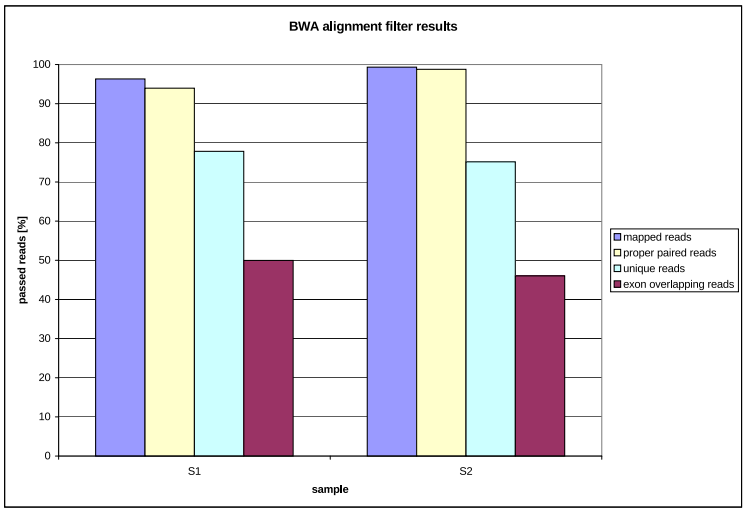

Figure 2.14: Illustration of alignment filter results of all processed filters applied on S1 and S2 in percent. Violet bars show percentage of reads which could be mapped onto the reference, light yellow bars represent the percentage of properly paired reads, turquoise bars encode the percentage of unique reads, and purple bars illustrate the percentage of exon overlapping reads.

2.2.3 Alignment statistics

Alignment statistics were conducted on mapped, properly paired, unique, and exon overlapping reads. Table 2.6 lists selected alignment summary metrics generated by the alignment statistics component. The quality of the passed reads is high, as strand balance is close to optimum and more than 99 percent of the passed sequence reads are high quality assigned reads. Since PE

Sample	Total number of analyzed reads	Mean read length	Strand balance	High quality aligned reads	High quality error rate
S1	35,278,875	74.81	0.48	35,101,928 99.50 %	$1.36 \cdot 10^{-3}$
S2	32,544,256	75.32	0.47	32,388,515 99.52 %	$1.34 \cdot 10^{-3}$

Table 2.6: Alignment summary metrics for sample S1 and S2. Only properly paired, unique, and exon overlapping reads were taken into account. The strand balance metric reports the ratio between the number of reads aligned onto the forward strand and total number of reads. High quality aligned reads describe the number of reads with an assigned mapping quality of 20 or higher. The high quality error rate metric illustrates the percentage of bases that mismatch the reference in high quality reads.

data was processed, additional insert size distribution characteristics were generated. Figure 2.15 illustrates the insert size distributions in a histogram for S1 and S2, whereas table 2.7 lists basic statistics about both samples.

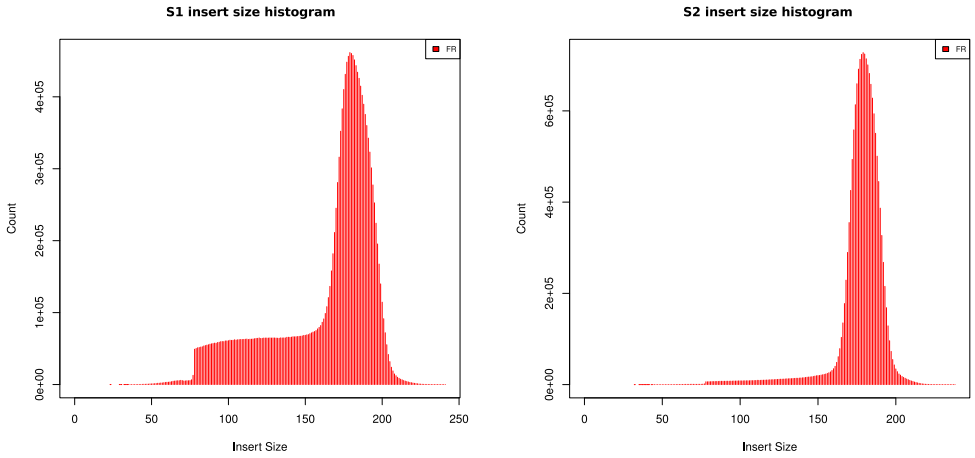

Figure 2.15: Insert size histograms for sample S1 and S2. In S1, the majority of read pairs was between 75 and 200 bp apart, whereas in S2 most reads were aligned within 160 to 200 bps.

Sample	Median insert size	Minimum insert size	Maximum insert size	Mean insert size	Standard deviation
S1	176	23	397	163.55	32.95
S2	180	32	260	177.23	16.18

Table 2.7: Insert size metrics for sample S1 and S2. Table lists median, minimum, maximum, mean, and standard deviation for reported insert sizes.

2.2.4 Variant detection

DIP calling resulted in 1,939 and 2,039 identified DIPs for S1 and S2, respectively. A minimum coverage of six reads at a site was required to be considered in DIP detection. Table 2.8 and figure 2.16 illustrate the DIP calling results in further detail.

Sample	Total number of DIPs	Not in RefSeq	Intron DIPs	Coding DIPs	UTR DIPs	Unknown DIPs
S1	1,939	1,725	173	29	11	1
		88,96 %	8,92 %	1.50 %	0.57 %	0.05 %
S2	2,039	1,823	177	24	13	2
		89.40 %	8.68 %	1.18 %	0.64 %	0.10 %

Table 2.8: Number and categorizations of identified DIPs in S1 and S2. Table lists the total number of identified DIPs, DIPs which did not overlap any RefSeq genes, intron affecting DIPs, DIPs present in RefSeq coding regions, DIPs covering untranslated regions (UTRs), and unknown DIPs. Unknown DIPs are DIPs which did overlap RefSeq genes but no functional information was provided for categorization.

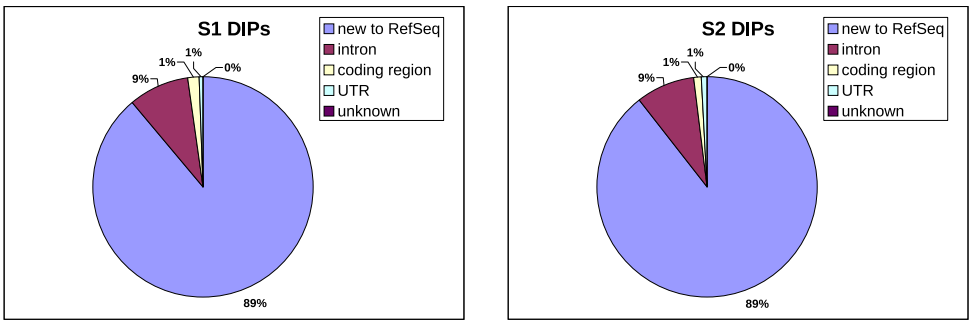

Figure 2.16: Categorization of identified DIPs in S1 and S2. In both samples, the majority of called DIPs was not associated with any RefSeq gene. Among DIPs covering RefSeq genes, DIPs located in intron regions were most prominent, followed by affected coding regions, and DIPs covering UTRs. RefSeq related DIPs were categorized as unknown if no functional information could be provided for the affected region.

SNP calling reported 18,575 (S1) and 18,605 (S2) raw SNPs. 1,410 (S1) and 1,543 (S2) SNPs were marked to be omitted by variant quality score recalibration. Subsequent filtering excluded SNPs with a false discovery rate (FDR) > 0.1, which resulted in 14,583 (S1) and 14,790 (S2) filtered SNPs. In the final pipeline analysis step, both samples were divided into homo- and heterozygous SNPs. Table 2.9 lists the detailed results for new SNPs and all SNPs known to dbSNP. The analysis results show that most detected SNPs were already stored in dbSNP. The percentage of known homozygous and heterozygous SNPs resembled each other, while novel SNPs were primarily labeled heterozygous (illustrated in figure 2.17).

Sample	Novelty status	Homozygous SNPs	Heterozygous SNPs
S1	all	5,696	8,887
	known	5,640	8,362
	new	56	525
S2	all	5,747	9,043
	known	5,686	8,390
	new	61	653

Table 2.9: Homo- and heterozygous SNP calls in S1 and S2. Table lists the total number of homo-/heterozygous SNPs, as well as the number of SNPs already reported in dbSNP, and new SNPs.

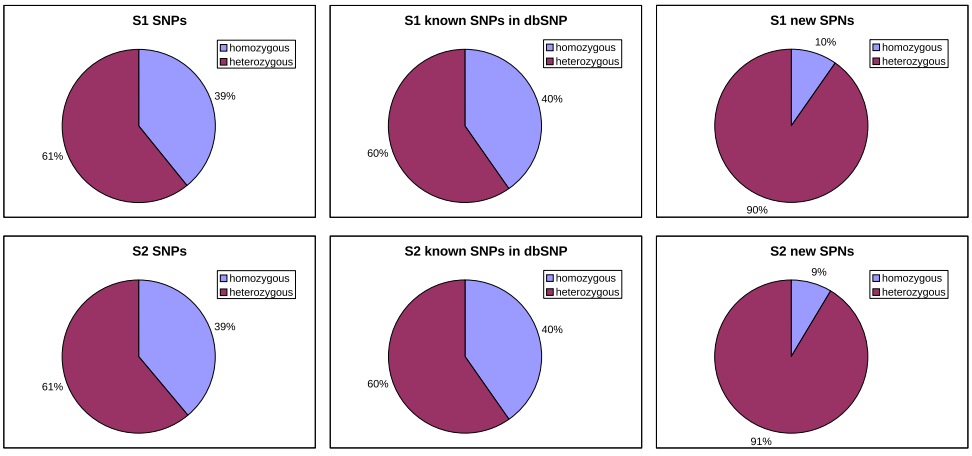

Figure 2.17: Relations between homo- and heterozygous SNPs detected in sample S1 and S2. For each sample, the relations between homo- and heterozygous SNPs for all, reported in dbSNP, and new SNPs are illustrated.

Based on this data, further investigations can be conducted to experimentally verify the authenticity of the identified variants and potential disease driving candidates can be studied in subsequent biological research programs.

In summary, 36,427,610 (S1) and 36,355,815 (S2) read pairs generated by Illumina Genome

Analyzer IIx were analyzed by the implemented pipeline. After read trimming and filtering more than 96 percent of both samples were passed to further downstream analyses. Sequence alignment and subsequent alignment filtering identified 35,278,875 (S1) and 32,544,256 (S2) exon overlapping reads. The aligned reads of both samples showed high quality as more than 99 percent of the exon overlapping reads were assigned a mapping quality of 20 or higher. Subsequent variant detection and filtering resulted in 1,939 (S1) and 2,039 (S2) DIPs and 14,583 (S1) and 14,790 (S2) SNPs of which 61 percent (S1 and S2) were classified as heterozygous.

Chapter 3

Discussion

The effective usage of NGS in modern genetics strongly depends on efficient software solutions which are capable of handling the downstream analyses of generated sequencing data (McPherson [2009]). As the bottleneck shifted from sequence generation to analysis, new and innovative analysis tools are in high demand to conquer the computational challenges of NGS (Flicek [2009]). Implementing a unified software solution that combines these tools (for a detailed list see Olivares [2010]) with necessary HPC approaches greatly facilitates the analysis process and therefore, focus is laid on the development of analysis integrative pipelines.

In this thesis a highly configurable sequencing analysis pipeline for exome studies was developed. The resulting application covers the whole process of exome analysis starting from raw sequence preprocessing, over sequence alignment and alignment statistics, to variant detection. To the best of our knowledge, the developed software is the first freely available exome analysis pipeline providing a streamlined analysis for SE and PE data generated by Illumina or ABI SOLiD^{TM}platforms.

Read preprocessing

The first pipeline module covers read preprocessing which supports the evaluation of raw data quality by providing basic statistics about read length and read quality. Furthermore, the component refines raw data by individually trimming and filtering error prone sequence reads. Several quality control programs were tested for their suitability including TileQC (Dolan and Denver [2008]), PIQA (Martínez-Alcántara et al. [2009]), CANGS (Pandey et al. [2010]), and SolexaQA (Cox et al. [2010]). These programs provide several read processing options but as neither of them supports both, Illumina and ABI SOLiD^{TM}platforms, they did not qualify for incorporation into

the pipeline. Therefore, new preprocessing components were developed which allow the analysis of SE and PE Illumina and ABI SOLiD^{TM}reads. The R component ShortRead (Morgan et al. [2010]) accepts both sequencing platforms but does not handle PE data and emphasizes only Illumina data. The unpublished software FastQC (Barbraham Bioinformatics [2009]) is, besides the work presented here, the only one meeting the requirements for quality statistics but was not available at the time of development.

Sequence alignment

The next pipeline module includes the read alignment component consisting of a mixture of well-established third party programs and newly developed applications. External software was incorporated for short read alignment, local realignment around DIPs, base quality score recalibration, and duplicate removal. The Burrows-Wheeler based aligner BWA (Li and Durbin [2009]) was chosen to be integrated into the pipeline as it supports gapped, time efficient, and quality scored alignment. Furthermore, it is capable of analyzing sequences encoded in nucleotide or color space and handles SE and PE reads. Other alignment programs such as MAQ (Li et al. [2008a]), SOAP (Li et al. [2008b]), Bowtie (Langmead et al. [2009]), and ELAND (Illumina [2010]) were discarded as they were either slower than BWA (MAQ, SOAP) or did not support gapped alignment (Bowtie, ELAND).

Alignment postprocessing

In order to generate accurate read alignments for variant detection, the pipeline refines raw read alignments by applying numerous post alignment steps in the following order: read tagging, local realignment around DIPs, base quality score recalibration, and alignment filtering. This particular sequence of analysis steps was chosen to ensure correct and ideal input for each analysis step.

Local realignment around DIPs corrects misalignments based on the alignment of all reads mapping at the site under investigation. Although this analysis step does not require unmapped reads, they are not yet filtered, because the subsequent base quality score recalibration step relies on this information. The recalibration method considers the probability of mismatching the reference genome to update the base calling quality score. Therefore, the pipeline applies recalibration after realignment.

Subsequent alignment filtering is applied to discard reads which do not originate from exon regions. Mapping and properly-paired filters are applied first since these reads can be easily detected by SAM flags and subsequently reduce the input for computationally more expensive filtering steps. It is a known fact that DNA amplification during library preparation may introduce bias by duplicating DNA fragments. Allowing these duplicate reads in downstream analyses may

lead to wrong conclusions, as the sequence reads derived from technical artifacts rather than real biological data. Therefore, the pipeline applies duplicate filtering. In order to provide information about DNA capture efficiency, exome filtering is executed at the end of the alignment component. A key characteristic of capture efficiency is the capture specificity which is defined by Ng et al. [2009] as the ratio of exon overlapping reads to total number of reads mapping the reference.

The generated proper pair information allows drawing conclusions about the alignment accuracy which can be affected by erroneous library preparation, inconsistencies within the alignment, or structural reallocations. Since this information is only provided by PE reads, PE exome sequencing is preferred to SE approaches. Additionally, SE duplicate removal tends to overestimate the number of true duplicates, especially with increasing coverage. Bainbridge et al. [2010] showed that analysis of PE data with SE approaches nearly quadrupled the amount of detected duplicates. By default, the pipeline performs duplicate removal on SE data, but in order to allow the use of a larger number of reads for variant detection, duplicate removal can be turned off for SE analysis.

Variant detection

Among the several variant callers available, GATK was chosen to be incorporated into the pipeline as it supports several NGS platforms, individual as well as multi sample analyses, and generates VCF 4 output. Moreover, it provides a set of additional SNP analysis tools including SNP quality evaluation, SNP filtering, and standardized downstream recalibration of variant quality scores. Another reason in favor of GATK is that several tools of this suite have proven reliable in large-scale projects like the 1000 Genomes Project and The Cancer Genome Atlas (Broad Institute [2010a]). Other applications tested for SNP detection are CRISP (Bansal [2010]), DiBayes (Applied Biosystems [2010]), SOAPsnp (Li et al. [2009b]), VarScan (Koboldt et al. [2009]), SNPSeeker (Druley et al. [2009]), and SAMtools (Li et al. [2009a]). Some of them are limited to a certain NGS platform (DiBayes, SNPseeker), while others are designed to handle only pooled sequencing data (CRISP). SAMtools and SOAPsnp are equipped for single sample analysis and multi platform support but do not provide further functionality to detect the novelty status of SNP calls. The feature set of VarScan is comparable to GATK as it is designed for pooled and single sample SNP calling, allows analyzing sequence reads derived from nucleotide or color space, and offers SNP annotation. However, since the tool does not provide as many downstream analysis options as GATK, VarScan was not integrated into the pipeline.

The last pipeline module covers variant detection handled by GATK and additional splitters. GATK performs SNP calling based on statistical methods which allow the assignment of scores to

enable the ranking and comparison of SNP calls. Since raw variant calls do include false positives, further SNP refinement is required. This can be achieved by individually filtering the SNP call set based on certain characteristics such as read coverage and quality scores which may cause the drawback of introducing bias. To support a more standardized SNP call refinement, GATK provides a variant score recalibration method which is solely based on covariates obtained by the call set itself. Thereby, bias introduced by unexperienced users can be avoided and comparison between DNA samples is facilitated.

Since the DIP detection tool provided by GATK is based on heuristic cutoffs and not on a statistical model, no DIP quality score is provided and DIP comparison between samples is hampered. More suitable approaches provided by Dindel (Albers et al. [2010]), Pindel (Ye et al. [2009]), and SRMA (Homer and Nelson [2010]) were either not available at the time of development (Dindel, SRMA), or only designed for PE data (Pindel).

Comparison with existing software

Compared to other analysis pipelines and toolkits the developed application incorporates a complete set of necessary analysis steps and offers automatic and efficient execution. The pipeline presented by Eck et al. [2010] supports exome analysis starting with sequence alignment for Illumina data but lacks raw read preprocessing and ABI SOLiD^{TM}support. Galaxy (Goecks et al. [2010]) provides similar analysis functionality as the developed pipeline but does not yet offer a streamlined pipeline for exome sequencing. The Omixon Variant Toolkit (Omixon Webservices [2010]) and GATK provide numerous analysis tools for exome sequencing but are based on already aligned reads. Additionally, the Omixon Variant Toolkit is designed for ABI SOLiD^{TM}data only. Commercial software such as CLCbio GenomicsWorkbench (CLCbio [2010]) and NextGENe (Softgenetics [2010]) offer a similar exome analysis functionality as the developed pipeline but are not freely available to the scientific community. Therefore, the pipeline developed and evaluated in this thesis represents an important contribution to genetic and clinical research.

Outlook

Due to the ongoing development of biological and information technologies, requirements of computational environments continuously change and applications can never be considered finished. The developed pipeline will undergo constant improvements and adaptations to include new findings and software tools for analyzing exome sequencing data.

Today, it is known that library preparation and DNA sequencing may introduce several types of artifacts, including primer sequences and low complexity sequences. Sequence cleaning programs like Seqclean (Chen et al. [2007]) or TagDust (Lassmann et al. [2009]) are capable of iden-

tifying and removing these contaminants. Therefore, the integration of such an analysis tool into the read preprocessing component would enhance raw read quality and increase the percentage of mapped reads.

Currently, heuristic cutoffs are used for DIP identification. As new and improved DIP callers like Dindel are now available, these tools will be tested for compatibility with Illumina and ABI SOLiD™ SE and PE reads and qualified programs will be incorporated into the pipeline.

The current pipeline was designed for analyzing DNA samples originating from only one sample. To gain further insights into hereditary diseases, some exome sequencing approaches analyze samples of several relatives together. Therefore, the application will be extended to be equipped for multi sample analysis.

Conclusion

In conclusion, the implemented analysis software provides an easy to use pipeline for the efficient and well-structured study of exome sequencing data generated by Illumina and ABI SOLiD™ devices. It supports the analysis of both, SE and PE, sequencing data types by implementing different strategies to fit the characteristics of the two different library preparation protocols. The combination of state-of-the-art libraries with newly developed code and HPC approaches allows the application to be capable of meeting the daunting challenges of NGS. Testing the pipeline on biological data obtained by a cooperation with clinical research partners showed that the developed application is tremendously suitable for exome sequencing analysis and generates highly reproducible results.

We strongly believe that the developed pipeline will be of interest not only to the biological community but can also have an impact on upcoming diagnostic applications in clinical settings. As the price for sequencing costs is plummeting, the diagnostic application of exome sequencing for the identification of genetic diseases is on the verge of becoming routine. In this context, fast and robust pipelines for the analysis of exome sequencing data are of utmost importance and will contribute to the adoption of existing and upcoming powerful sequencing technologies in routine clinical applications.

Chapter 4

Methods

4.1 Next-generation sequencing

Prior to the introduction of the first NGS platform, automated Sanger sequencing (Smith et al. [1986], Ansorge et al. [1986]) was the method of choice for DNA sequencing. Its major limitations were the low number of samples which could be analyzed in parallel and its resulting high costs per base. Together with the prospect of a broad field of application, these objectives led to the development of new sequencing technologies. NGS devices provide high throughput and speed at the expense of analyzing only relatively short reads at a lower per base accuracy than automated Sanger sequencing (see table 4.1 for detailed information).

In general, the sequencing process can be grouped into library preparation, sequencing and imaging, and data analysis (Metzker [2010]).

4.1.1 Library preparation

Library preparation describes the preparation of DNA templates prior to sequencing (Roe [2004]). This process includes breaking a DNA sample randomly into smaller fragments, ligating adapter sequences to allow the later use of universal primers, and amplifying the produced DNA templates. Based on different library preparation protocols one can distinguish between single-end (SE), paired-end (PE), and mate-paired (MP) libraries.

Platform	Library/template preparation	Read length [bp]	Run time	Gb per run
Sanger sequencing	bacterial cloning	650	1-3 h	0.0017-0.005
Roche/454 GS FLX Titanium	SE/MP emulsion PCR	400	10 h	0.4-0.6
Illumina HiSeq 2000	SE/PE/MP bridge amplification	35 50 x 50 100 x 100	1.5 d[a] 4 d[b] 8 d[c]	2-35[a] 75-100[b] 150-200[c]
ABI SOLiD™4hq	SE/PE/MP emulsion PCR	75[d] 75 x 35[e] 75 x 75[f]	3 d[g] 12 d[h] 14 d[i]	100[g] N/A N/A

a 35bp SE *b* 50 x 50 PE/MP *c* 100 x 100 PE/MP
d SE module *e* PE module *f* MP module *g* 75bp SE *h* 75 x 35 PE *i* 75 x 75 MP

Table 4.1: Comparison of sequencing platforms, summarizing Roche [2010], Illumina [2010], and ABI [2010]

Single-end library

SE libraries (also referred to as fragment libraries) are created by randomly shearing genomic DNA (gDNA) or complementary DNA (cDNA) into fragments which are less than 1 kb in size.

Paired-end library

PE library preparation resembles the SE protocol except that different sequencing primer (SP) sites are ligated at each end. This is needed, as PE reads are created by sequencing a DNA fragment from both sides sequentially. After analyzing the first read with SP 1 the templates are regenerated and the second read is sequenced by the use of SP 2 (see figure 4.1). This technology allows the creation of read pairs which are between 200 to 500 bp apart in the original sample. This distance is referred to as insert size. Currently only Illumina (4.1.2) and ABI SOLiD™(4.1.2) offer PE protocols (Illumina [2010], ABI [2010]).

Figure 4.1: Schema of PE protocol. Figure adapted from Illumina [2010].

Mate-pair library

In MP protocols, sheared DNA with 2-5 kb in size is labeled at the ends, circularized, and again linearized by cutting the cycles. Only fragments containing the label and therefore both ends of the original DNA fragment are selected and sequenced as described in PE sequencing (see figure 4.2).

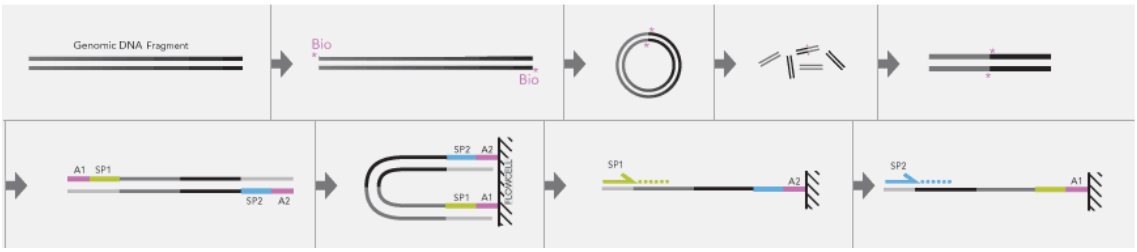

Figure 4.2: MP protocol. Figure adapted from Illumina [2010].

After the initial library preparation step, Sanger sequencing and most NGS technologies use a DNA amplification step to ensure sufficient signal intensity for nucleotide detection. Bacterial cloning of DNA fragments, as used for Sanger sequencing, may incorporate parts of the cloning vector thus introducing artifacts. New technologies avoid this by applying alternative amplification steps or directly sequencing single DNA molecules (Branton et al. [2008], Schadt et al. [2010]). Commonly used amplification procedures are emulsion PCR and bridge amplification.

Emulsion PCR

In emulsion PCR, single-stranded DNA (ssDNA) hybridizes onto oligonucleotide bound beads, ideally leaving one DNA template per bead (see figure 4.3). The beads are part of water-in-oil micro-emulsions which additionally contain all necessary components for PCR. After PCR amplification within these micro-reactors, up to thousands of complementary DNA strands are covalently bound to the beads. The original DNA template strands are washed away and the beads are purified and immobilized for later sequencing (Dressman et al. [2003], Metzker [2010]).

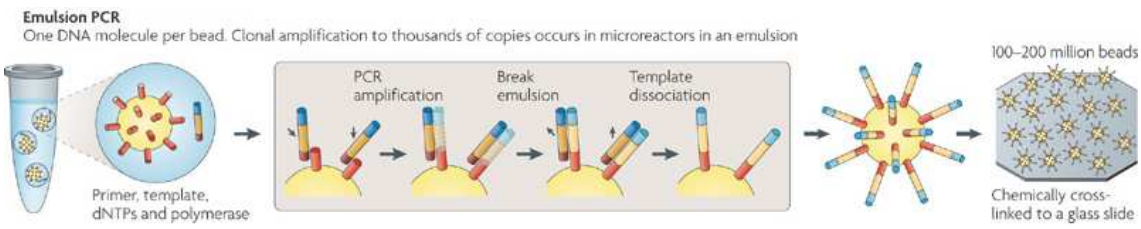

Figure 4.3: Description of emulsion PCR. Figure taken from Metzker [2010].

Bridge amplification

A solid surface which is densely coated with forward and reverse primers is the main device for bridge amplification (see figure 4.4). ssDNA templates anneal randomly to the surface and their complementary strands are built. After denaturation, a washing step removes all original template DNA. The remaining covalently bound complementary strands bind over to nearby reverse primers enabling the creation of yet another replicated strand. The DNA is denaturated again to yield ssDNA and the next cycle of primer annealing and DNA replication is started. Thereby, clusters of DNA are produced all over the solid surface. As a last step prior to sequencing the template DNA is cleaved and washed away. (Adessi et al. [2000], Fedurco et al. [2006], Mardis [2008])

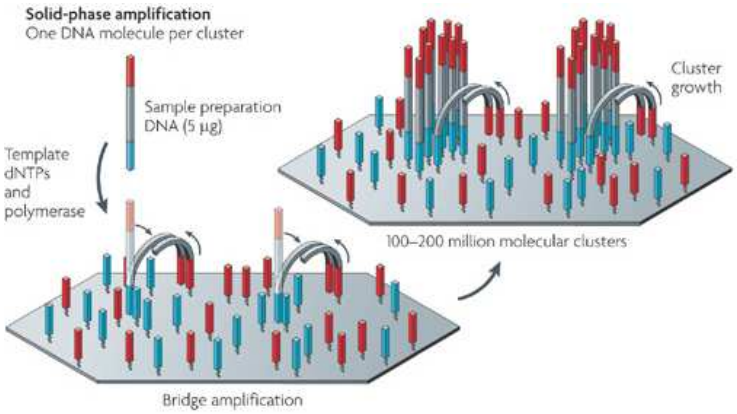

Figure 4.4: Bridge amplification. Figure taken from Metzker [2010].

4.1.2 Sequencing and imaging

The three most widely used NGS platforms up to today are Roche/454, Illumina, and ABI SOLiD™. All three technologies implement a sequencing-by-synthesis approach in which the synthesis of a complementary DNA strand is used to determine the DNA sequence. A second shared feature is the use of a template amplification step to increase signal intensity for nucleotide identification. Third generation sequencing (TGS) platforms are capable of analyzing single DNA molecules but are out of the scope of this thesis and will not be further discussed. An review about TGS technologies is given in Schadt et al. [2010] and Munroe and Harris [2010].

Roche/454

In 2005 Roche/454 introduced the first commercially available NGS device. The platform analyzes DNA by the use of an alternative sequencing technology known as pyrosequencing (Ronaghi et al. [1996], Ronaghi et al. [1998]) where nucleotides are detected based on the release of pyrophos-

phate. After the amplification of target DNA with emulsion PCR, beads, pyrosequencing enzymes, and pyrosequencing sulfates are loaded into a pico titer plate device, which places each bead into an addressable position within the plate. After the sequencing primer has been annealed the first sequencing cycle is started. Each cycle, nucleotides of one type (either dATP, dCTP, dGTP, or dTTP) are added to the plate. When incorporated, the nucleotide releases a pyrophosphate thereby triggering a series of downstream reactions. The use of luciferase in these reactions causes emission of a light signal which is proportional to the amount of integrated nucleotides. This signal is detected by a CCD camera and image information is stored for further processing. The remaining nucleotides are washed away and the next type of nucleotides is added to the plate. (Margulies et al. [2005], Ansorge [2009], Roche [2010])

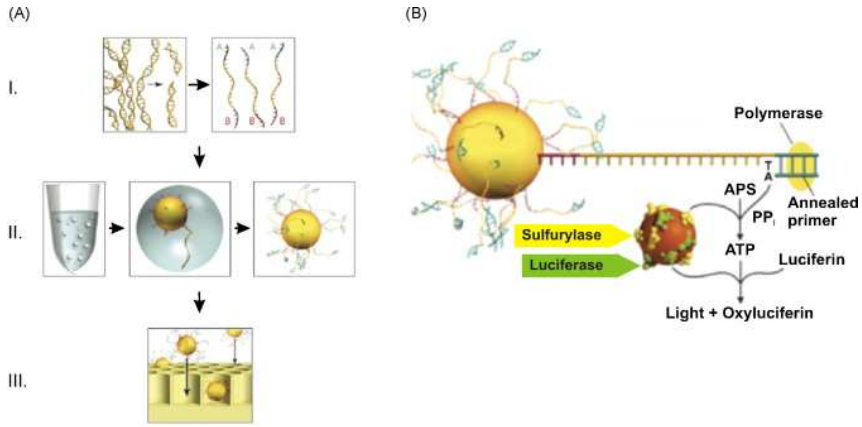

Figure 4.5: Roche/454 sequencing workflow. (A) Library preparation. (I) DNA is randomly sheared and universal primer sites are ligated. (II) Emulsion PCR results in beads with thousands of identical template DNAs. (III) Beads and pyrosequencing molecules are loaded into a pico titer plate. (B) Illustration of a pyrosequencing reaction including the enzymes DNA polymerase, DNA primer, ATP sulfurylase, and luciferase, and the substrates adenosine 5' phosphosulfate (APS) and luciferin. Figure taken from Ansorge [2009].

Illumina

The Illumina (formerly known as Solexa) sequencing technology analyzes different DNA samples in parallel by the use of bridge amplification and dye-terminated nucleotides. After primer hybridization, nucleotides which are labeled by different fluorescent dyes are added to the slide. In each sequencing cycle, only one nucleotide is incorporated into the complementary strand due to its attached terminator group. A washing step removes all remaining free nucleotides before the newly added base is identified. Finally, the terminator and dye group are cleaved off, and the sequencing cycle is restarted. (see figure 4.6) (Bentley et al. [2008], Metzker [2010]). is

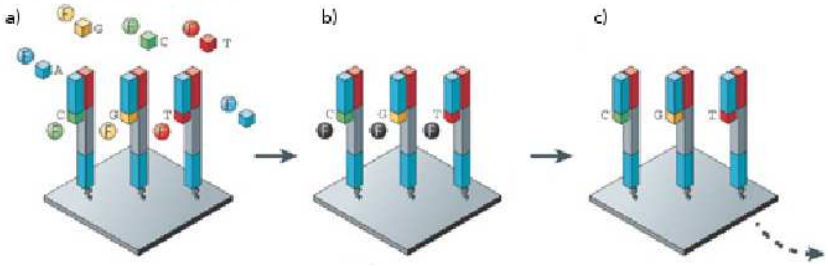

Figure 4.6: Illumina sequencing schema. a) After the sequencing primer is hybridized to the primer site, fluorescently labeled, reversibly terminated nucleotides are incorporated into the complementary strand. b) The remaining nucleotides are washed away and the fluorescence signal identifying the base is recorded. c) The fluorescent label and terminator group are removed and a new cycle of sequencing is started. Figure adapted from Metzker [2010].

ABI SOLiD™

The SOLiD™system is based on ligating fluorescently labeled dinucleotide probes to the DNA template under investigation (figure 4.7)(Tomkinson et al. [2006], Landegren et al. [1988]). After emulsion PCR, beads are covalently bound to a glass slide and universal sequencing primers, ligases, and a pool of labeled dinucleotide probes are added to the glass slide. The DNA sequence is determined by recording the color code, representing the first two bases of the dinucleotide, in several cycles of DNA ligation and cycles of primer reset. Figure 4.8 describes the sequencing process in further detail. Due to the two color encoding each base is determined independently two times. Therefore, the distinction between true SNPs and sequencing errors is possible.

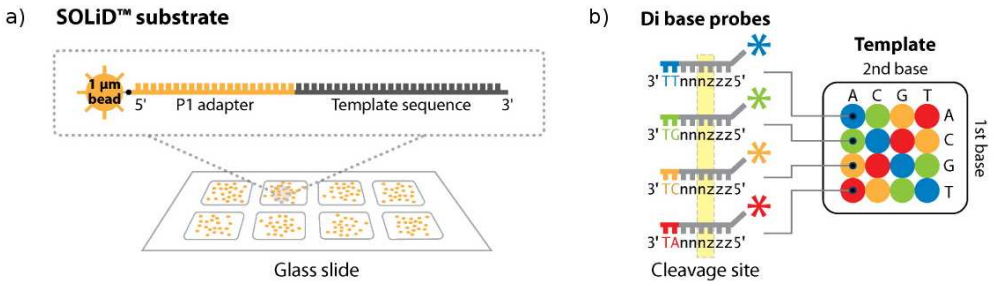

Figure 4.7: ABI SOLiD™sequencing chemistry. a) The SOLiD™substrate consists of an emulsion PCR bead, its covalently bound primer site (two sites for MP and PE), and the DNA template to be sequenced. b) SOLiD™uses '1,2-probes' (a version of a dinucleotide probe) where the first and second nucleotides are analyzed. The remaining six bases consist of either degenerated or universal bases (Metzker [2010]). Each dye represents 4 of 16 possible dinucleotide sequences. Figure adapted from Mardis [2008].

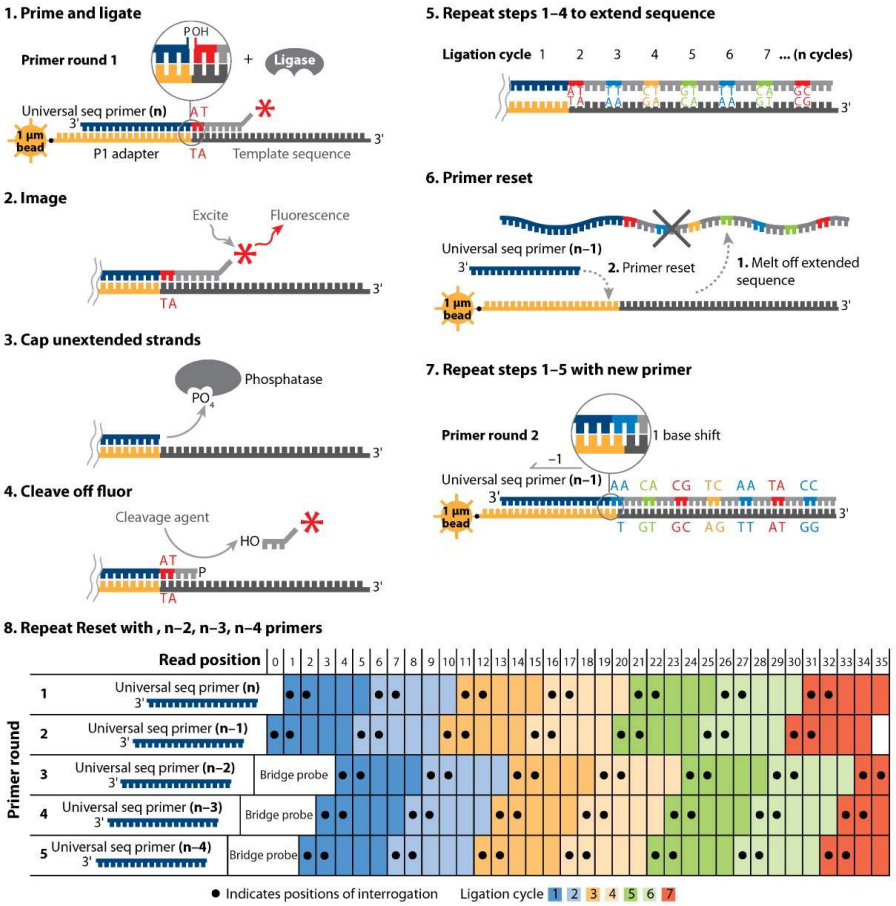

Figure 4.8: ABI SOLiD^{TM}sequencing cycle. The complementary dinucleotide hybridizes to the already primer-bound template sequence and is ligated (1). After fluorescence is measured (2) un-extended strands are capped (3) and the dye is cleaved off (4) leaving a free 5' phosphate group available for further reactions. This process is repeated for several cycles until the required read length is achieved (5). The synthesized strand is removed, a new primer with a one-base offset is hybridized (6) and the ligation cycles are repeated (7). This primer reset process is repeated for five rounds providing a dual measurement of each base (8) (ABI [2010]). Figure adapted from Mardis [2008].

4.2 Sequencing applications

The combination of different types of sample input and library preparations allows for the analysis of numerous sequencing applications such as whole genome sequencing, ChIP-Seq, metagenomics, targeted re-sequencing (e.g. exome sequencing), RNA-Seq, Methyl-Seq, and others. For a brief review on these methods see Wold and Myers [2008]. As this thesis focuses on the analysis of exome sequencing data, this application is described in further detail.

4.2.1 Exome sequencing

Exome sequencing describes the process of targeted re-sequencing of a genome's entire protein coding regions with the goal to identify mutations. Sample input is created by using exome capturing arrays or capturing libraries (NimbleGen [2010], Agilent [2010], Gnirke et al. [2009]) which isolate and enrich the DNA templates to be analyzed (see figure 4.9). After sequencing, quality estimations allow the evaluation of each analyzed base and sequence alignment is used to map the relatively short reads onto a reference genome. In order to gain first insights into library preparation and sequencing efficiency, filtering steps are required to determine the percentage of sequence reads which do not originate from protein coding regions or could not be aligned at all. Subsequently, variant detection algorithms obtain a set of genome positions where the analyzed sample differs significantly from the reference. Since these call sets contain numerous non-biologically based variations, further filtering steps are applied to increase the number of true biological variants.

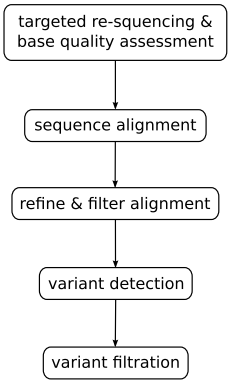

Figure 4.9: Illustration of exome sequencing workflow. After targeted re-sequencing of protein coding regions, base calling and quality assessment methods are automatically applied. Sequence alignment of the resulting reads followed by refinement and filtration enable variant detection. In order to enhance the called variant set, false positive calls are filtered out, too.

4.3 Base/color calling quality assessment

In order to provide accuracy estimates, every NGS platform assigns phred-like quality values for each base or color call. Initially introduced by the base-calling program Phred (Ewing and Green [1998], Ewing et al. [1998]), this quality measurement links error probabilities logarithmically to a base or color call. It is defined as

$$q_{phred} = -10 \cdot \log_{10}(p) \tag{4.1}$$

where p is the estimated error probability for that call. Roche/454 and ABI SOLiD™directly adopted this definition (Illumina [2010], Roche [2010]) whereas Illumina analysis pipelines prior to

version 1.3 applied the odds ratio instead of p:

$$q_{Ill1.3-} = -10 \cdot \log_{10}(\frac{p}{1-p}) \tag{4.2}$$

It is straightforward to show that the two scores commute as:

$$q_{phred} = 10 \cdot \log_{10}(10^{\frac{q_{Ill1.3}}{10}} + 1) \text{ and } q_{Ill1.3} = 10 \cdot \log_{10}(10^{\frac{q_{phred}}{10}} - 1) \tag{4.3}$$

Later Illumina switched to the original phred scoring schema (see formula 4.1).

4.4 Alignment

Alignment can be described as

> the process of determining the most likely source within the genome sequence for the observed DNA sequence read, given the knowledge of which species the sequence has come from. (Flicek and Birney [2009])

Traditional alignment programs for Sanger sequencing such as BLAST (Altschul et al. [1990]) or BLAT (Kent [2002]) do not scale well with NGS reads in terms of memory usage, processing time, and mapping accuracy. Therefore, several new alignment tools, especially designed for mapping large amount of short reads, were developed. The algorithms are able to handle NGS specific sequence read error profiles, PE reads, color space, gapped alignment, and reads originating from repetitive regions. (Trapnell and Salzberg [2009], Flicek [2009])

Generally, all short read alignment programs use a two step procedure to map a sequence. First, heuristic techniques identify a small subset of the reference genome where the read is most likely to align. Then, a slower and more accurate alignment algorithm is used to determine the exact position of the sequence read. For the latter, suitable algorithms are discussed in Batzoglou [2005].

Short read aligners can be divided in Burrows-Wheeler transform (BWT, Burrows and Wheeler [1994]) and hash table based algorithms. Hash table based aligners either index the read sequences and search through the reference or vice versa. Read indexing algorithms require little memory but may be inefficient for aligning a small amount of reads whereas reference indexing methods have a large memory footprint. BWT based aligners use a reversible compression algorithm to build a reference index suffix tree and then search within this suffix tree for possible alignments. In contrast to hash table based methods, when using the BWT index only a fraction of time is needed for whole genome sequence alignment. (Flicek and Birney [2009])

In order to handle ambiguity or lack of accuracy in alignments, Li et al. [2008a] introduced the concept of *mapping qualities* which are a measure for the confidence that a read actually originated from the position it was aligned to by the mapping program. They consider a read alignment as an estimate of the true alignment and calculated the mapping quality Q_s by phred scaling the alignment's error probability P:

$$Q_s = -10 \cdot \log_{10}(P\{read\ is\ wrongly\ mapped\}) \tag{4.4}$$

Consequently, the mapping quality is set to 0 for reads which map equally well to multiple positions in the genome. It is common practice to apply mapping qualities to 255 to indicate that mapping quality is not available (Li et al. [2009]). As PE reads combine information of both DNA fragment sides their mapping qualities Q_p are calculated as $Q_p = Q_{s1} + Q_{s2}$. This applies only if both alignments are consistent, i.e. if the insert size and alignment direction is correct. If alignments do not add up both reads will be treated as SE regarding their mapping quality score calculations.

A detailed discussion about several available short read aligners is given by Li and Homer [2010]. A recently established, public online repository (Division for Bioinformatics [2010]) provides an overview about many currently available NGS alignment programs and lists for each tool short description, authors, publications, implementation language, supported file formats, and further information.

4.5 Variant detection

The main goal of exome sequencing is detecting variations from the reference genome to determine genes associated with rare or common disorders (Ng et al. [2010]). Generally, SNPs are determined by the comparison of an assembled consensus sequence, which represents the most likely genotype based on the analyzed sequence reads, with its reference genome. Various SNP and DIP callers were implemented to identify sites which differ statistically significant from the reference genome. Simple variant detection approaches apply fixed filters which are based on the percentage of reads containing the same non-reference base call. More advanced methods include Bayesian identifiers in combination with prior genotype probabilities to infer the genotype and detect variants. Most Bayesian methods differ regarding their estimated prior genotype probabilities. Some SNP callers define their prior probabilities considering transition and transversion rates or even nucleotide substitution patterns based on previous studies on NCBI dbSNPs (Zhao and Boerwinkle [2002]) (Li et al. [2008a], Li et al. [2009b], Martin et al. [2010]).

It is advisable to consider different quality indices such as base and mapping quality as poor

data quality affects SNP calling accuracy. Phred scaled quality scores for consensus and variant quality estimation describe the probability that the genotype call is incorrect or that an inferred variant is in fact identical to the reference (Li [2008]).

In order to distinguish biological variants from variants caused by sequencing errors, several pre- and post filtering steps must be applied. A common approach to reduce false positive calls is to ignore SNP calls around DIPs and clusters of SNPs, as practice has shown that misalignments around these sites are commonly error prone (Broad Institute [2010a]). Additional post filtering must be individually adjusted for each data set which may introduce further bias. In order to ensure comparability between results, a more standardized way of variant filtration was recently introduced in the Genome Analysis Toolkit (see section 4.7). It recalibrates variant scores based on trained data provided by dbSNP (Wheeler et al. [2007]), the HapMap Project (Consortium [2003]), and optionally the 1000 Genomes Project (1000 Genomes [2008]). The Genome Analysis Toolkit also offers statistics about transition/transversion rate, HapMap concordance, and dbSNP concordance, which allows drawing conclusions about the quality of called SNP sets.

4.6 File formats

Several file formats were established for handling data in sequencing and NGS projects (see table 4.2 for an overview of the most commonly used formats and their specifications). As NGS technologies generate large amount of data, new file formats were introduced, especially designed for efficient data processing.

4.6.1 FASTQ format

The FASTQ format is a text-based file format for storing sequence read data. For each read the nucleotide sequence as well as assigned quality values are listed. The format is closely related to the FASTA sequence file format (Pearson and Lipman [1988]) and is seen as a de facto standard for storing NGS data. Similar to the FASTA format, FASTQ lacks an explicit definition which led to the introduction of several incompatible variants. Cock et al. [2010] addressed this issue and defined the Open Bioinformatics Foundation (OBF [2010]) consensus of the FASTQ format with four different line types as follows:

```
@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)
```

Figure 4.10: FASTQ format definition

File format	Description	Publication/Specification
FASTA	stores nucleotide an protein sequences	Pearson and Lipman [1988]
FASTQ	stores nucleotide sequences and their quality values	Cock et al. [2010] Cock et al. [2010]
GFF	*General Feature Format*; exchange format for feature description within sequences	WTSI [2000]
GTF	*Gene Transfer Format*; based on GFF with a structure providing a separate definition and format name	WUSTL
BED	*Browser Extensible Data*; defines data displayed in the UCSC Genome Browser annotation track	UCSC [2010]
WIG	aka *WIGGLE*; format to display continuous-valued data in a track format	UCSC [2010]
BIGBED	*Big Browser Extensible Data*; compressed, binary indexed BED	Kent et al. [2010]
BIGWIG	*Big WIGGLE*; compressed, binary indexed WIG	Kent et al. [2010]
ROD	*Reference Ordered Data*; file format for storing data ordered by references	McKenna et al. [2010]
VCF	*Variant Call Format*; stores sequence variants	1000 Genomes [2010a] 1000 Genomes [2010b]

Table 4.2: List of common file formats used in NGS.

'@' indicates the beginning of the title line with unlimited length. It often contains a read identifier and may include any additional information like read length or read position within the plate. The second line type stores the nucleotide sequence which may be line wrapped. There are no restrictions regarding the allowed characters in this line type. Still, Cock et al. [2010] recommend to use only upper case IUPAC single letter codes for (ambiguous) DNA or RNA (IUPAC [1970]) in sequence lines. '+' at the start of a line signals the end of sequence lines and optionally repeats the title line. The subsequent lines contain the read's nucleotide quality scores in ASCII encoding. It is crucial to notice that both line delimiters - '@' and '+' - can occur anywhere in the quality lines which complicates file parsing. To enable simple line parsing most programs avoid line wrapping in their output routines.

FASTQ variants

Currently, three different FASTQ variants are known: standard Sanger, Illumina prior to version 1.3, and Illumina after version 1.3. Illumina applied in its early version different base quality calculations (see section 4.3) and introduced different ASCII offsets. Table 4.3 lists type and range of the quality score as well as ASCII offset and range as handled in the different FASTQ variants.

FASTQ variant	Quality score		ASCII	
	type	range	offset	range
Sanger FASTQ	phred	0 - 93	33	33 - 126
Illumina 1.3- FASTQ	Illumina 1.3-	-5 - 62	64	59 - 126
Illumina 1.3+ FASTQ	phred	0 - 62	64	64 - 126

Table 4.3: FASTQ file format variants including ASCII encoding and quality calculation.

Since all FASTQ variants share valid ASCII ranges and contain no header information it is not always possible to determine the used variant solely based on the file content.

4.6.2 Sequence Alignment/Map format

The Sequence Alignment/Map (SAM) format (Li et al. [2009]) was designed to store nucleotide alignments in a generic way. It aims at supporting several sequencing platforms, allowing short and long reads up to 128 Mb, saving SE as well as PE information, and including various types of alignments. The tab delimited text format can be divided into a header and an alignment section. Header lines can be identified by the '@' character at the start of each line and contain, among others, information about the reference against the reads have been aligned, and read groups. In the alignment section, a tab delimited line describes a read alignment and contains required information about read name, read flags, reference sequence, alignment position, mapping quality, extended CIGAR, reference sequence of the paired read, position of the paired read, inferred insert size within the read pair, read sequence, and read quality. The extended CIGAR string characterizes the type of alignment operation including clipped, spliced, multi-part, and padded alignment. Allowed operations are match/mismatch (M), insertion (I), deletion (D), skipped base (N), soft clipping (S), hard clipping (H), and padding (P). Additional optional fields allow the documentation of less important or program specific data. Color space read information is also described in the optional fields. To accelerate parsing and ease data processing, the binary file format Binary Alignment/Map (BAM) equivalent to SAM was introduced (Li et al. [2009a]).

SAMtools

SAMtools describes the sum of software packages designed for parsing and manipulating alignments in the SAM/BAM format available in C, C++, Java, Perl, Python, Ruby, and Common Lisp. They provide several utilities for format conversions, alignment sorting, alignment merging, file indexing, PCR duplicate removal, generating alignments in a per-position format, and many more. (Li et al. [2009])

4.6.3 VCF format

The variant call format (VCF) is a text based file format designed for storing the most prevalent types of sequence variations - including SNPs, DIPs and larger structural variants - together with rich annotations in a standardized way (Danecek et al. [2010]). It is divided into a header and a body section where each header line is identified by a leading '#'. The header stores mandatory information about the file format version and body content. Optional header lines contain metadata about annotations in the VCF body section. Commonly used annotations include genotype likelihoods, dbSNP membership, ancestral allele, read depth, and mapping quality (Danecek et al. [2010], 1000 Genomes [2010a], 1000 Genomes [2010b]).

VCFtools

The freely available software library VCFtools provides general Perl and Python APIs and supports VCF format validation, migration, annotation, file comparison, basic statistics, merging, and creation of intersections and complements (Danecek et al. [2010]).

4.7 Genome Analysis Toolkit

The Genome Analysis Toolkit (GATK) is a structured Java cross-platform API specifically designed for the development of efficient and robust analysis tools for already base called and aligned NGS data (McKenna et al. [2010]). To ensure a quick and stable parallel processing of the data, GATK uses the functional programming philosophy Map/Reduce (Dean and Ghemawat [2008]) and requires its input to be sorted in the chromosome order of the analyzed genome (referred to as reference ordered). Besides providing an API for support the development of user-specific analysis tools, GATK already contains a set of data access patterns and tools for generally needed tasks including recalibration of base quality scores, local realignment, genotyping, and variant call detection (Broad Institute [2010a]). Currently, the GATK is used in large-scale sequencing projects

like the 1000 Genomes Project (1000 Genomes [2008]) and The Cancer Genome Atlas (TCGA [2005]).

4.8 Genome visualization

The introduction of NGS technologies was accompanied by the generation of data in unprecedented amounts and speed. Although many NGS analysis tasks are now accomplished by automated processes, their resulting data still requires visual inspection and interpretation by researchers. In order to ease exploration and interpretation of NGS data, qualitative and quantitative abstraction is of utmost importance. Therefore, tools, especially designed for displaying large amount of data, ranging from simple stand alone software to complex integrated software packages, were developed (Nielsen et al. [2010], Evanko [2010]).

The presentation of the reference genome and their mapped sequence reads represented as letter strings allow the inspection of sequence alignments. As displaying each and every single read can impede the structured representation of data, a simplified stacked visualization is offered where only variances between reference and reads are highlighted. To reveal the exact read sequence a zooming mechanism is used for the detailed representation of NGS data. The inclusion of additional information from external sources may facilitate the interpretation of NGS data. Therefore, genome browsers incorporate biological annotations such as gene expression, and genotype variation within a graphical interface (Nielsen et al. [2010], Cline and Kent [2009]).

A detailed list of viewers can be found in Olivares [2010] and Nielsen et al. [2010].

4.9 Pipeline concept

The developed exome pipeline allows authorized users the parallel or serial execution of a fixed, pre-defined sequence of steps on a cluster. A properties file, containing all cluster and user data, provides all necessary information (i.e. URL, user name, and password) for user authentication and authorization which is checked at the beginning of each pipeline run. Additional parameters allow input specification and fined tuned configuration of each pipeline step. Before a particular step is submitted to a queue, the information required for its execution is determined by the program. The pipeline then transfers all missing files to the task's folder on the cluster. If any data is already stored on the cluster (due to previous analysis steps) it can be referenced for further processing steps and paths to these files is automatically resolved. Finally, the appropriate command for analysis step execution is generated. On the workstation, the pipeline checks every few seconds the status of all started analysis tasks and fetches the output files including exit status

information once the commands have completed. Then the next pipeline step will be started. The pipeline automatically checks at each start if there are already finished steps reported to avoid unnecessary computations in case of pipeline abortions and restarts. It is possible but not advisable to define computationally inexpensive tasks to be run directly on the workstation in order to avoid additional I/O on the cluster.

4.10 IT infrastructure

Sophisticated hard- and software systems are required to conquer the computationally expensive tasks of NGS data analysis. This section and figure 4.11 describe the IT infrastructure used in this thesis.

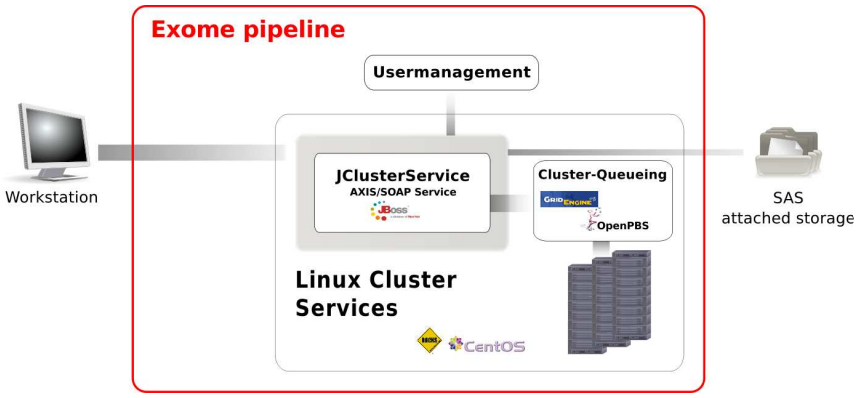

Figure 4.11: Two Sun Fire™X4600 M2 Servers (32 CPUs, 160 GB RAM in total) attached via GBit Ethernet interconnect to SAS storage with 16 TB (extendable to max. 256 TB) are the basis for all HPC calculations. CentOS (CentOS [2010]) and Rocks cluster distribution (Rocks cluster [2010]) were used as operating systems of choice for both servers. The pipeline software is based on the Java Platform, Enterprise Edition (JEE) three-tier architecture and uses JClusterService API for communication with the HPC cluster. Data is secured by the detached user management system.

Hardware infrastructure

To provide an efficient system for analyzing and visualizing exome sequencing data, a 64 bit computing cluster, consisting of the following components, was introduced:

- two Sun Fire™X4600 M2 Servers each with
- four Quad-Core AMD Opteron™8356, 2.3 GHz, 80 GB RAM and
- Serial attached SCSI (SAS) storage of 16 TB (extendable up to 256 TB) connected via
- GBit Ethernet interconnect

Software infrastructure

The developed pipeline is attached to a file based user management system in order to ensure that only authorized users access the high-performance computing cluster. All required input and output data is transferred to the HPC cluster using the JClusterService API (Stocker et al. [2009]). The Oracle Grid Engine was chosen as cluster-queuing system, which internally handles task scheduling and resource management between different analysis jobs.

Appendix A

Bibliography

Article references

[Adessi et al., 2000] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, and E. Kawashima. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res*, 28(20):E87, Oct 2000.

[Albers et al., 2010] Cornelis A Albers, Gerton Lunter, Daniel G Macarthur, Gilean McVean, Willem H Ouwehand, and Richard Durbin. Dindel: Accurate indel calls from short-read data. *Genome Res*, Oct 2010. doi: 10.1101/gr.112326.110. URL `http://dx.doi.org/10.1101/gr.112326.110`.

[Altschul et al., 1990] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990. doi: 10.1006/jmbi.1990.9999. URL `http://dx.doi.org/10.1006/jmbi.1990.9999`.

[Ansorge et al., 1986] W. Ansorge, B. S. Sproat, J. Stegemann, and C. Schwager. A non-radioactive automated method for DNA sequence determination. *J Biochem Biophys Methods*, 13(6):315–323, Dec 1986.

[Ansorge et al., 1987] W. Ansorge, B. Sproat, J. Stegemann, C. Schwager, and M. Zenke. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res*, 15(11):4593–4602, Jun 1987.

[Ansorge, 2009] Wilhelm J Ansorge. Next-generation DNA sequencing techniques. *N Biotechnol*, 25(4):195–203, Apr 2009. doi: 10.1016/j.nbt.2008.12.009. URL `http://dx.doi.org/10.1016/j.nbt.2008.12.009`.

[Bainbridge et al., 2010] Matthew N Bainbridge, Min Wang, Daniel L Burgess, Christie Kovar, Matthew J Rodesch, Mark D'Ascenzo, Jacob Kitzman, Yuan-Qing Wu, Irene Newsham, Todd A Richmond, Jeffrey A Jeddeloh, Donna Muzny, Thomas J Albert, and Richard A Gibbs. Whole exome capture in solution with 3 Gbp of data. *Genome Biol*, 11(6):R62, 2010. doi: 10.1186/gb-2010-11-6-r62. URL `http://dx.doi.org/10.1186/gb-2010-11-6-r62`.

[Bansal, 2010] Vikas Bansal. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, 26(12):i318–i324, Jun 2010. doi: 10.1093/bioinformatics/btq214. URL `http://dx.doi.org/10.1093/bioinformatics/btq214`.

[Batzoglou, 2005] Serafim Batzoglou. The many faces of sequence alignment. *Brief Bioinform*, 6 (1):6–22, Mar 2005.

[Bentley et al., 2008] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R. Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall A Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo A Baybayan, Vincent A Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John A Bridgham, Rob C Brown, Andrew A Brown, Dale H Buermann, Abass A Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R. Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W. Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip A Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc A Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer A Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark A Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D. Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie A Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008. doi: 10.1038/nature07517. URL `http://dx.doi.org/10.1038/nature07517`.

[Botstein and Risch, 2003] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 33 Suppl:228–237, Mar 2003. doi: 10.1038/ng1090. URL `http://dx.doi.org/10.1038/ng1090`.

[Branton et al., 2008] Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean Ling, Carlos H Mastrangelo, Amit Meller, John S Oliver, Yuriy V Pershin, J. Michael Ramsey, Robert Riehn, Gautam V Soni, Vincent Tabard-Cossa, Meni Wanunu, Matthew Wiggin, and Jeffery A Schloss. The potential and challenges of nanopore sequencing. *Nat Biotechnol*, 26(10):1146–1153, Oct 2008. doi: 10.1038/nbt.1495. URL `http://dx.doi.org/10.1038/nbt.1495`.

[Burrows and Wheeler, 1994] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical Report 124, 1994.

[Chen et al., 2007] Yi-An Chen, Chang-Chun Lin, Chin-Di Wang, Huan-Bin Wu, and Pei-Ing Hwang. An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics*, 8:416, 2007. doi: 10.1186/1471-2164-8-416. URL http://dx.doi.org/10.1186/1471-2164-8-416.

[Cline and Kent, 2009] Melissa S Cline and W. James Kent. Understanding genome browsing. *Nat Biotechnol*, 27(2):153–155, Feb 2009. doi: 10.1038/nbt0209-153. URL http://dx.doi.org/10.1038/nbt0209-153.

[Cock et al., 2010] Peter J A Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, 38(6):1767–1771, Apr 2010. doi: 10.1093/nar/gkp1137. URL http://dx.doi.org/10.1093/nar/gkp1137.

[Consortium et al., 2010] 1000 Genomes Project Consortium, Richard M Durbin, Gonçalo R Abecasis, David L Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010. doi: 10.1038/nature09534. URL http://dx.doi.org/10.1038/nature09534.

[Consortium, 2003] International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, Dec 2003.

[Cox et al., 2010] Murray P Cox, Daniel A Peterson, and Patrick J Biggs. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11:485, 2010. doi: 10.1186/1471-2105-11-485. URL http://dx.doi.org/10.1186/1471-2105-11-485.

[Dean and Ghemawat, 2008] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/1327452.1327492.

[Dolan and Denver, 2008] Peter C Dolan and Dee R Denver. TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics*, 9:250, 2008. doi: 10.1186/1471-2105-9-250. URL http://dx.doi.org/10.1186/1471-2105-9-250.

[Dressman et al., 2003] Devin Dressman, Hai Yan, Giovanni Traverso, Kenneth W Kinzler, and Bert Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A*, 100(15):8817–8822, Jul 2003. doi: 10.1073/pnas.1133470100. URL http://dx.doi.org/10.1073/pnas.1133470100.

[Druley et al., 2009] Todd E Druley, Francesco L M Vallania, Daniel J Wegner, Katherine E Varley, Olivia L Knowles, Jacqueline A Bonds, Sarah W Robison, Scott W Doniger, Aaron Hamvas, F. Sessions Cole, Justin C Fay, and Robi D Mitra. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods*, 6(4):263–265, Apr 2009. doi: 10.1038/nmeth.1307. URL http://dx.doi.org/10.1038/nmeth.1307.

[Eck et al., 2010] S. H. Eck, E. Graf, A. Benet-Pagès, T. Meitinger, and T. Strom. Analysis Pipeline for Exome Sequencing Data. *Genome Informatics*, 2010.

[Evanko, 2010] Daniel Evanko. Supplement on visualizing biological data. *Nat Methods*, 7(3 Suppl):S1, Mar 2010.

[Ewing and Green, 1998] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8(3):186–194, Mar 1998.

[Ewing et al., 1998] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3):175–185, Mar 1998.

[Fedurco et al., 2006] Milan Fedurco, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res*, 34(3):e22, 2006. doi: 10.1093/nar/gnj023. URL http://dx.doi.org/10.1093/nar/gnj023.

[Flicek, 2009] Paul Flicek. The need for speed. *Genome Biol*, 10(3):212, 2009. doi: 10.1186/gb-2009-10-3-212. URL http://dx.doi.org/10.1186/gb-2009-10-3-212.

[Flicek and Birney, 2009] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nat Methods*, 6(11 Suppl):S6–S12, Nov 2009. doi: 10.1038/nmeth.1376. URL http://dx.doi.org/10.1038/nmeth.1376.

[Gnirke et al., 2009] Andreas Gnirke, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, Carsten Russ, Stacey Gabriel, David B Jaffe, Eric S Lander, and Chad Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, 27(2):182–189, Feb 2009. doi: 10.1038/nbt.1523. URL http://dx.doi.org/10.1038/nbt.1523.

[Goecks et al., 2010] Jeremy Goecks, Anton Nekrutenko, James Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010. doi: 10.1186/gb-2010-11-8-r86. URL http://dx.doi.org/10.1186/gb-2010-11-8-r86.

[Hedges et al., 2009] Dale J Hedges, Dale Hedges, Dan Burges, Eric Powell, Cherylyn Almonte, Jia Huang, Stuart Young, Benjamin Boese, Mike Schmidt, Margaret A Pericak-Vance, Eden Martin, Xinmin Zhang, Timothy T Harkins, and Stephan ZÃ$\frac{1}{4}$chner. Exome sequencing of a multigenerational human pedigree. *PLoS One*, 4(12):e8232, 2009. doi: 10.1371/journal.pone.0008232. URL http://dx.doi.org/10.1371/journal.pone.0008232.

[Homer and Nelson, 2010] Nils Homer and Stanley F Nelson. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol*, 11(10):R99, Oct 2010. doi: 10.1186/gb-2010-11-10-r99. URL http://dx.doi.org/10.1186/gb-2010-11-10-r99.

[Horner et al., 2010] David Stephen Horner, Giulio Pavesi, Tiziana Castrignanò, Paolo D'Onorio De Meo, Sabino Liuni, Michael Sammeth, Ernesto Picardi, and Graziano Pesole. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform*, 11(2):181–197, Mar 2010. doi: 10.1093/bib/bbp046. URL http://dx.doi.org/10.1093/bib/bbp046.

[IUPAC, 1970] IUPAC. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. *Eur J Biochem*, 15(2):203–208, Aug 1970.

[Jasny and Roberts, 2003] Barbara R. Jasny and Leslie Roberts. Building on the DNA Revolution. *Science*, 300(5617):277–, 2003. doi: 10.1126/science.300.5617.277. URL http://www.sciencemag.org.

[Kent et al., 2010] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, Sep 2010. doi: 10.1093/bioinformatics/btq351. URL http://dx.doi.org/10.1093/bioinformatics/btq351.

[Kent, 2002] W. James Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4): 656–664, Apr 2002. doi: 10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002. URL http://dx.doi.org/10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002.

[Koboldt et al., 2009] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, Sep 2009. doi: 10.1093/bioinformatics/btp373. URL http://dx.doi.org/10.1093/bioinformatics/btp373.

[Koboldt et al., 2010] Daniel C Koboldt, Li Ding, Elaine R Mardis, and Richard K Wilson. Challenges of sequencing human genomes. *Brief Bioinform*, 11(5):484–498, Sep 2010. doi: 10.1093/bib/bbq016. URL http://dx.doi.org/10.1093/bib/bbq016.

[Landegren et al., 1988] U. Landegren, R. Kaiser, J. Sanders, and L. Hood. A ligase-mediated gene detection technique. *Science*, 241(4869):1077–1080, Aug 1988.

[Lander et al., 2001] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H.

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

[Langmead et al., 2009] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009. doi: 10.1186/gb-2009-10-3-r25. URL http://dx.doi.org/10.1186/gb-2009-10-3-r25.

[Lassmann et al., 2009] Timo Lassmann, Yoshihide Hayashizaki, and Carsten O Daub. TagDust–a program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, 25(21):2839–2840, Nov 2009. doi: 10.1093/bioinformatics/btp527. URL http://dx.doi.org/10.1093/bioinformatics/btp527.

[Li and Durbin, 2009] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009. doi: 10.1093/bioinformatics/btp324. URL http://dx.doi.org/10.1093/bioinformatics/btp324.

[Li and Homer, 2010] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, 11(5):473–483, Sep 2010. doi: 10.1093/bib/bbq015. URL http://dx.doi.org/10.1093/bib/bbq015.

[Li et al., 2008] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008a. doi: 10.1101/gr.078212.108. URL http://dx.doi.org/10.1101/gr.078212.108.

[Li et al., 2009] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009a.

[Li et al., 2008] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, Mar 2008b. doi: 10.1093/bioinformatics/btn025. URL http://dx.doi.org/10.1093/bioinformatics/btn025.

[Li et al., 2009] Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. SNP detection for massively parallel whole-genome resequencing. *Genome Res*, 19(6):1124–1132, Jun 2009b. doi: 10.1101/gr.088013.108. URL http://dx.doi.org/10.1101/gr.088013.108.

[Li et al., 2009] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009c. doi: 10.1093/bioinformatics/btp336. URL http://dx.doi.org/10.1093/bioinformatics/btp336.

[Mardis, 2006] Elaine R Mardis. Anticipating the 1,000 dollar genome. *Genome Biol*, 7(7):112, 2006. doi: 10.1186/gb-2006-7-7-112. URL http://dx.doi.org/10.1186/gb-2006-7-7-112.

[Mardis, 2008] Elaine R Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008. doi: 10.1146/annurev.genom.9.081307.164359. URL http://dx.doi.org/10.1146/annurev.genom.9.081307.164359.

[Margulies et al., 2005] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P

Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005. doi: 10.1038/nature03959. URL http://dx.doi.org/10.1038/nature03959.

[Martin et al., 2010] E. R. Martin, D. D. Kinnamon, M. A. Schmidt, E. H. Powell, S. Zuchner, and R. W. Morris. SeqEM: An adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, Sep 2010. doi: 10.1093/bioinformatics/btq526. URL http://dx.doi.org/10.1093/bioinformatics/btq526.

[Martínez-Alcántara et al., 2009] A. Martínez-Alcántara, E. Ballesteros, C. Feng, M. Rojas, H. Koshinsky, V. Y. Fofanov, P. Havlak, and Y. Fofanov. PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics*, 25(18):2438–2439, Sep 2009. doi: 10.1093/bioinformatics/btp429. URL http://dx.doi.org/10.1093/bioinformatics/btp429.

[McKenna et al., 2010] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, Sep 2010. doi: 10.1101/gr.107524.110. URL http://dx.doi.org/10.1101/gr.107524.110.

[McPherson, 2009] John D McPherson. Next-generation gap. *Nat Methods*, 6(11 Suppl):S2–S5, Nov 2009. doi: 10.1038/nmeth.f.268. URL http://dx.doi.org/10.1038/nmeth.f.268.

[Metzker, 2010] Michael L Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, Jan 2010. doi: 10.1038/nrg2626. URL http://dx.doi.org/10.1038/nrg2626.

[Munroe and Harris, 2010] David J Munroe and Timothy J R Harris. Third-generation sequencing fireworks at Marco Island. *Nat Biotechnol*, 28(5):426–428, May 2010. doi: 10.1038/nbt0510-426. URL http://dx.doi.org/10.1038/nbt0510-426.

[Ng et al., 2008] Pauline C Ng, Samuel Levy, Jiaqi Huang, Timothy B Stockwell, Brian P Walenz, Kelvin Li, Nelson Axelrod, Dana A Busam, Robert L Strausberg, and J. Craig Venter. Genetic variation in an individual human exome. *PLoS Genet*, 4(8):e1000160, 2008. doi: 10.1371/journal.pgen.1000160. URL http://dx.doi.org/10.1371/journal.pgen.1000160.

[Ng et al., 2009] Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E Eichler, Michael Bamshad, Deborah A Nickerson, and Jay Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, Sep 2009. doi: 10.1038/nature08250. URL http://dx.doi.org/10.1038/nature08250.

[Ng et al., 2010] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–35, Jan 2010. doi: 10.1038/ng.499. URL http://dx.doi.org/10.1038/ng.499.

[Nielsen et al., 2010] Cydney B Nielsen, Michael Cantor, Inna Dubchak, David Gordon, and Ting Wang. Visualizing genomes: techniques and challenges. *Nat Methods*, 7(3 Suppl):S5–S15, Mar 2010. doi: 10.1038/nmeth.1422. URL http://dx.doi.org/10.1038/nmeth.1422.

[Pandey et al., 2010] Ram Vinay Pandey, Viola Nolte, and Christian Schlötterer. CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. *BMC Res Notes*, 3:3, 2010. doi: 10.1186/1756-0500-3-3. URL http://dx.doi.org/10.1186/1756-0500-3-3.

[Pearson and Lipman, 1988] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448, Apr 1988.

[Pruitt et al., 2005] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):D501–D504, Jan 2005. doi: 10.1093/nar/gki025. URL http://dx.doi.org/10.1093/nar/gki025.

[Roe, 2004] Bruce A Roe. Shotgun library construction for DNA sequencing. *Methods Mol Biol*, 255:171–187, 2004. doi: 10.1385/1-59259-752-1:171. URL http://dx.doi.org/10.1385/1-59259-752-1:171.

[Ronaghi et al., 1996] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 242(1):84–89, Nov 1996. doi: 10.1006/abio.1996.0432. URL http://dx.doi.org/10.1006/abio.1996.0432.

[Ronaghi et al., 1998] M. Ronaghi, M. Uhlén, and P. Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363, 365, Jul 1998.

[Sanger et al., 1977] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, Dec 1977.

[Schadt et al., 2010] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Hum Mol Genet*, 19(R2):R227–R240, Oct 2010. doi: 10.1093/hmg/ddq416. URL http://dx.doi.org/10.1093/hmg/ddq416.

[Service, 2006] Robert F Service. Gene sequencing. The race for the $1000 genome. *Science*, 311(5767):1544–1546, Mar 2006. doi: 10.1126/science.311.5767.1544. URL http://dx.doi.org/10.1126/science.311.5767.1544.

[Smith et al., 1986] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986. doi: 10.1038/321674a0. URL http://dx.doi.org/10.1038/321674a0.

[Stocker et al., 2009] Gernot Stocker, Maria Fischer, Dietmar Rieder, Gabriela Bindea, Simon Kainz, Michael Oberstolz, James G McNally, and Zlatko Trajanoski. iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis. *BMC Bioinformatics*, 10:390, 2009. doi: 10.1186/1471-2105-10-390. URL http://dx.doi.org/10.1186/1471-2105-10-390.

[Tomkinson et al., 2006] Alan E Tomkinson, Sangeetha Vijayakumar, John M Pascal, and Tom Ellenberger. DNA ligases: structure, reaction mechanism, and function. *Chem Rev*, 106(2):687–699, Feb 2006. doi: 10.1021/cr040498d. URL http://dx.doi.org/10.1021/cr040498d.

[Trapnell and Salzberg, 2009] Cole Trapnell and Steven L Salzberg. How to map billions of short reads onto genomes. *Nat Biotechnol*, 27(5):455–457, May 2009. doi: 10.1038/nbt0509-455. URL http://dx.doi.org/10.1038/nbt0509-455.

[Venter et al., 2001] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M.

Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001. doi: 10.1126/science.1058040. URL `http://dx.doi.org/10.1126/science.1058040`.

[Wheeler et al., 2007] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Yuri Kapustin, Oleg Khovayko, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Vadim Miller, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Steven T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35(Database issue):D5–12, Jan 2007. doi: 10.1093/nar/gkl1031. URL `http://dx.doi.org/10.1093/nar/gkl1031`.

[Wheeler et al., 2008] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, Oleg Khovayko, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, Vadim Miller, James Ostell, Kim D Pruitt, Gregory D Schuler, Martin Shumway, Edwin Sequeira, Steven T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center

for Biotechnology Information. *Nucleic Acids Res*, 36(Database issue):D13–D21, Jan 2008. doi: 10.1093/nar/gkm1000. URL http://dx.doi.org/10.1093/nar/gkm1000.

[Wold and Myers, 2008] Barbara Wold and Richard M Myers. Sequence census methods for functional genomics. *Nat Methods*, 5(1):19–21, Jan 2008. doi: 10.1038/nmeth1157. URL http://dx.doi.org/10.1038/nmeth1157.

[Ye et al., 2009] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, Nov 2009. doi: 10.1093/bioinformatics/btp394. URL http://dx.doi.org/10.1093/bioinformatics/btp394.

[Zhao and Boerwinkle, 2002] Zhongming Zhao and Eric Boerwinkle. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*, 12(11):1679–1686, Nov 2002. doi: 10.1101/gr.287302. URL http://dx.doi.org/10.1101/gr.287302.

Book references and talks

[DePristo, 2010] SNP calling. Genome Sequencing and Analysis, Broad Institute of Harvard and MIT NGS workshop I, February 2010. URL http://www.broadinstitute.org/files/shared/mpg/nextgen2010/nextgen_garimella.pdf.

[Mann, 2009] Illumina Quality Scores. Illumina Customer Facing, 2009.

[Poplin, 2010] Base quality score recalibration. Genome Sequencing and Analysis, Broad Institute of Harvard and MIT NGS workshop I, February 2010. URL http://www.broadinstitute.org/files/shared/mpg/nextgen2010/nextgen_poplin.pdf.

[R Development Core Team, 2010] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org. ISBN 3-900051-07-0.

Specifications

[1000 Genomes, 2010] Variant Call Format (VCF) version 4.0. Last visited on 2010-10-21, 2010a. URL http://www.1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcf4.0.

[1000 Genomes, 2010] Variant Call Format (VCF) version 4.0 - encoding structural variants. Last visited on 2010-10-21, 2010b. URL http://www.1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcf_4.0_sv.

[Li et al., 2009] Sequence Alignment/Map (SAM) Format. Last visited on 2010-10-21, 2009. URL http://samtools.sourceforge.net/SAM1.pdf.

[WTSI, 2000] GFF (General Feature Format) specifications document. Last visited on 2010-10-21, 2000. URL http://www.sanger.ac.uk/resources/software/gff/spec.html.

[WUSTL,] GTF2.2 format (Revised Ensembl GTF). Last visited on 2010-10-21. URL http://mblab.wustl.edu/GTF22.html.

Web link references

[1000 Genomes, 2008] 1000 Genomes - A Deep Catalog of Human Genetic Variatio. Last visited on 2010-11-10, 2008. URL http://www.1000genomes.org/.

[ABI, 2010] ABI SOLiD™. Last visited on 2010-11-10, 2010. URL http://www.appliedbiosystems.com/.

[Agilent, 2010] Agilent Technologies. Last visited on 2010-11-10, 2010. URL http://www.agilent.com/.

[Applied Biosystems, 2010] DiBayes. Last visited on 2010-11-10, 2010. URL http://solidsoftwaretools.com/gf/project/dibayes/.

[Barbraham Bioinformatics, 2009] FastQC: A quality control application for FastQ data. Last visited on 2010-11-10, 2009. URL http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

[Broad Institute, 2010] The Genome Analysis Toolkit. Last visited on 2010-11-10, 2010a. URL http://www.broadinstitute.org/gsa/wiki/index.php/Main_Page.

[Broad Institute, 2010] Integrative Genomics Viewer. Last visited on 2010-11-10, 2010c. URL http://www.broadinstitute.org/igv/.

[Carlson, 2009] The Bio-Economist. Last visited on 2010-11-10, September 2009. URL http://www.synthesis.cc/2009/09/the-bio-economist.html.

[CentOS, 2010] CentOS The Community ENTerprise Operating System. Last visited on 2010-11-10, 2010. URL http://www.centos.org/.

[CLCbio, 2010] CLCbio Genomics Workbench. Last visited on 2010-11-10, 2010. URL http://www.clcbio.com/.

[Danecek et al., 2010] The Variant Call Format and VCFtools. Last visited on 2010-11-10, 2010. URL http://vcftools.sourceforge.net/.

[Division for Bioinformatics, 2010] NGSlib. Last visited on 2010-11-10, 2010. URL http://ngslib.i-med.ac.at/.

[ENA, 2010] Annual Report 2009. Last visited on 2010-11-10, 2010. URL http://www.ebi.ac.uk/enateam/.

[GRC, 2010] Genome Reference Consortium. Last visited on 2010-11-10, 2010. URL http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.

[Illumina, 2010] Illumina Inc. Last visited on 2010-11-10, 2010. URL http://www.illumina.com/.

[International HapMap Project, 2006] About the International HapMap Project. Last visited on 2010-11-10, 2006. URL http://hapmap.ncbi.nlm.nih.gov/abouthapmap.html.

[Li, 2008] Maq: Mapping and Assembly with Qualities. Last visited on 2010-11-10, 2008. URL http://maq.sourceforge.net/.

[Li et al., 2009] SAMtools. Last visited on 2010-11-10, 2009. URL http://samtools.sourceforge.net/.

[Morgan et al., 2010] ShortRead. Last visited on 2010-11-10, 2010. URL http://www.bioconductor.org/packages/2.3/bioc/html/ShortRead.html.

[NimbleGen, 2010] NimbleGen. Last visited on 2010-11-10, 2010. URL http://www.nimblegen.com/.

[OBF, 2010] Open Bioinformatics Foundation. Last visited on 2010-11-10, 2010. URL http://www.open-bio.org/.

[Olivares, 2010] SEQanswers the next generation sequencing community. Last visited on 2010-11-10, 2010. URL http://seqanswers.com/.

[Omixon Webservices, 2010] Omixon Variant Toolkit. Last visited on 2010-11-10, 2010. URL http://www.omixon.com/omixon/.

[Roche, 2010] Roche 454. Last visited on 2010-11-10, 2010. URL http://www.454.com/.

[Rocks cluster, 2010] Rocks. Last visited on 2010-11-10, 2010. URL http://www.rocksclusters.org/.

[Softgenetics, 2010] NextGENe. Last visited on 2010-11-10, 2010. URL http://softgenetics.com/NextGENe.html.

[TCGA, 2005] The Cancer Genome Atlas. Last visited on 2010-11-10, 2005. URL http://cancergenome.nih.gov/.

[UCSC, 2010] UCSC Genome Browser. Last visited on 2010-11-10, 2010. URL http://genome.ucsc.edu/.

List of Figures

List of Tables

Appendix B

Glossary

3’ end	DNA end with a terminal hydroxyl group
5’ end	DNA end with a terminal phosphate group
allele	any of two or more variants of a gene’s DNA sequence
API	Application Programming Interface
BAM	Binary Alignment/Map
BED	Browser Extensible Data
CCDS	Consensus CDS
cDNA	complementary DNA, synthesized from a mature mRNA template
CDS	CoDing Sequence
consensus	idealized sequence reflecting the most common base of multiple sequence reads at each position of the genome
CPU	Central Processing Unit
dATP	deoxyAdenosine TriPhosphate
dCTP	deoxyCytidine TriPhosphate
dGTP	deoxyGuanosine TriPhosphate
DIP	Deletion/Insertion Polymorphism
DNA	DeoxyriboNucleic Acid
dTTP	deoxyThymidine TriPhosphate
ENA	European Nucleotide Archive
exome	complete set of all protein-coding regions of a genome
forward strand	DNA strand in direction of 5’ to 3’ end
gDNA	genomic DNA

genotype	entire set of genes in a cell, an organism, or an individual
GFF	General Feature Format
GRC	Genome Reference Consortium
GTF	Gene Transfer Format
haplotype	set of alleles closely linked on a chromosome that are transmitted together
HapMap	Haplotype Map of the human genome
hard clip	truncated read bases
HPC	High-Performance Computing
IUPAC	International Union of Pure and Applied Chemistry
JEE	Java Platform, Enterprise Edition
kb	kilo base
Mb	mega base
MP	Mate Pair
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
PE	Paired End
phenotype	physical trait or feature of an organism, as determined by a particular genotype
RAM	Random-Access Memory
reverse strand	DNA strand in direction of 3’ to 5’ end
SAM	Sequence Alignment/Map
SAS	Serial Attached SCSI
SCSI	Small Computer System Interface
SE	Single End
sequence read	sequencing machine output of an analyzed DNA fragment
SNP	Single Nucleotide Polymorphism
soft clip	alignment ignored read bases
SP	Sequencing Primer
SRA	Sequence Read Archive
transition	substitution of a purine for another purine (adenine to guanine or vice versa) or of a pyrimidine for another pyrimidine (cytosine to thymine or vice versa)
transversion	substitution of a purine for a pyrimidine or vice versa
USCS	University of California, Santa Cruz
UTR	UnTranslated Region
WGS	Whole Genome Sequencing

WTSI	Wellcome Trust Sanger Institute
WUSTL	Washington University in ST. Louis

Appendix C

Acknowledgments

This work was supported by the GEN-AU project Bioinformatics Integration Network (BIN) of the Austrian Ministry for Science and Research. I would like to thank my supervisor, Zlatko Trajanoski, for his encouragement, support, vision, and belief in me.

Further thanks go to the members of the Institute for Genomics and Bioinformatics and my colleges at the Section for Bioinformatics for their fruitful discussions, help, and friendship. I want to express my gratitude to Gernot Stocker for his valuable input, assistance, and constant support.

I would also like to acknowledge Michael Speicher of the Institute of Human Genetics for providing experimental data.

I am indebted to my family for their unfailing support and encouragement.

Maria Fischer

Graz, Austria, November 2010

Appendix D

Publications

Stocker G, **Fischer M**, Rieder D, Bindea GL, Kainz S, Oberstolz M, McNally J, Trajanoski Z.: iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis. BMC Bioinformatics 2009, 10:390 PMID: 19941647

Geigl JB, Obenauf AC, Waldispuehl-Geigl J, Hoffmann EM, Auer M, Hörmann M, **Fischer M**, Trajanoski Z, Schenk MA, Baumbusch LO, Speicher MR.: Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. Nucleic Acids Res. 2009, 37 PMID: 19541849

BMC Bioinformatics

Software

iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis

Gernot Stocker[1], Maria Fischer[1], Dietmar Rieder[1], Gabriela Bindea[1], Simon Kainz[1], Michael Oberstolz[1], James G McNally[2] and Zlatko Trajanoski*[1]

Address: [1]Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria and [2]Center for Cancer Research Core Imaging Facility, Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, 41 Library Drive, Bethesda, MD 20892, USA

Email: Gernot Stocker - gernot.stocker@tugraz.at; Maria Fischer - maria.fischer@tugraz.at; Dietmar Rieder - dietmar.rieder@tugraz.at; Gabriela Bindea - gabriela.bindea@tugraz.at; Simon Kainz - simon.kainz@tugraz.at; Michael Oberstolz - michael.oberstolz@student.tugraz.at; James G McNally - mcnallyj@dce41.nci.nih.gov; Zlatko Trajanoski* - zlatko.trajanoski@tugraz.at

* Corresponding author

Published: 26 November 2009

BMC Bioinformatics 2009, **10**:390 doi:10.1186/1471-2105-10-390

Received: 1 June 2009
Accepted: 26 November 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/390

Abstract

Background: In recent years, the genome biology community has expended considerable effort to confront the challenges of managing heterogeneous data in a structured and organized way and developed laboratory information management systems (LIMS) for both raw and processed data. On the other hand, electronic notebooks were developed to record and manage scientific data, and facilitate data-sharing. Software which enables both, management of large datasets and digital recording of laboratory procedures would serve a real need in laboratories using medium and high-throughput techniques.

Results: We have developed iLAP (Laboratory data management, Analysis, and Protocol development), a workflow-driven information management system specifically designed to create and manage experimental protocols, and to analyze and share laboratory data. The system combines experimental protocol development, wizard-based data acquisition, and high-throughput data analysis into a single, integrated system. We demonstrate the power and the flexibility of the platform using a microscopy case study based on a combinatorial multiple fluorescence in situ hybridization (m-FISH) protocol and 3D-image reconstruction. iLAP is freely available under the open source license AGPL from http://genome.tugraz.at/iLAP/.

Conclusion: iLAP is a flexible and versatile information management system, which has the potential to close the gap between electronic notebooks and LIMS and can therefore be of great value for a broad scientific community.

Background

The development of novel large-scale technologies has considerably changed the way biologists perform experiments. Genome biology experiments do not only generate a wealth of data, but they often rely on sophisticated laboratory protocols comprising hundreds of individual

steps. For example, the protocol for chromatin immunoprecipitation on a microarray (Chip-chip) has 90 steps, uses over 30 reagents and 10 different devices [1]. Even adopting an established protocol for large-scale studies represents a daunting challenge for the majority of the labs. The development of novel laboratory protocols and/or the optimization of existing ones is still more distressing, since this requires systematic changes of many parameters, conditions, and reagents. Such changes are becoming increasingly difficult to trace using paper lab books. A further complication for most protocols is that many laboratory instruments are used, which generate electronic data stored in an unstructured way at disparate locations. Therefore, protocol data files are seldom or never linked to notes in lab books and can be barely shared within or across labs. Finally, once the experimental large-scale data have been generated, they must be analyzed using various software tools, then stored and made available for other users. Thus, it is apparent that software support for current biological research - be it genomic or performed in a more traditional way - is urgently needed and inevitable.

In recent years, the genome biology community has expended considerable effort to confront the challenges of managing heterogeneous data in a structured and organized way and as a result developed information management systems for both raw and processed data. Laboratory information management systems (LIMS) have been implemented for handling data entry from robotic systems and tracking samples [2,3] as well as data management systems for processed data including microarrays [4,5], proteomics data [6-8], and microscopy data [9]. The latter systems support community standards like FUGE[10,11], MIAME [12], MIAPE [13], or MISFISHIE [14] and have proven invaluable in a state-of-the-art laboratory. In general, these sophisticated systems are able to manage and analyze data generated for only a single type or a limited number of instruments, and were designed for only a specific type of molecule.

On the other hand, commercial as well as open source electronic notebooks [15-19] were developed to record and manage scientific data, and facilitate data-sharing. The influences encouraging the use of electronic notebooks are twofold [16]. First, much of the data that needs to be recorded in a laboratory notebook is generated electronically. Transcribing data manually into a paper notebook is error-prone, and in many cases, for example, analytical data (spectra, chromatograms, photographs, etc.), transcription of the data is not possible. Second, the incorporation of high-throughput technologies into the research process has resulted in an increased volume of electronic data that need to be transcribed. As opposed to LIMS, which captures highly structured data through rigid user interfaces with standard report formats, electronic notebooks contain unstructured data and have flexible user interfaces.

Software which enables both, management of large datasets and recording of laboratory procedures, would serve a real need in laboratories using medium and high-throughput techniques. To the best of our knowledge, there is no software system available, which supports tedious protocol development in an intuitive way, links the plethora of generated files to the appropriate laboratory steps and integrates further analysis tools. We have therefore developed iLAP, a workflow-driven information management system for protocol development and data management. The system combines experimental protocol development, wizard-based data acquisition, and high-throughput data analysis into a single, integrated system. We demonstrate the power and the flexibility of the platform using a microscopy case study based on combinatorial multiple fluorescence in situ hybridization (m-FISH) protocol and 3D-image reconstruction.

Implementation

Workflow-driven software design

The design of a software platform that supports the development of protocols and data management in an experimental context has to be based on and directed by the laboratory workflow. The laboratory workflow can be divided into four principal steps: 1) project definition phase, 2) experimental design and data acquisition phase, 3) data analysis and processing phase and 4) data retrieval phase (Figure 1).

Project definition phase

A scientific project starts with a hypothesis and the choice of methods required to address a specific biological question. Already during this initial phase it is crucial to define the question as specifically as possible and to capture the information in a digital form. Documents collected during the literature research should be collated with the evolving project definition for later review or for sharing with other researchers. All files collected in this period should be attached to the defined projects and experiments in the software.

Experimental design and data acquisition

Following the establishment of a hypothesis and based on preliminary experiments, the detailed design of the biological experiments is then initiated. Usually, the experimental work follows already established standard operating procedures, which have to be modified and optimized for the specific biological experiment. These protocols are defined as a sequence of protocol steps. However, well-established protocols must be kept flexible in a way that particular conditions can be changed. The

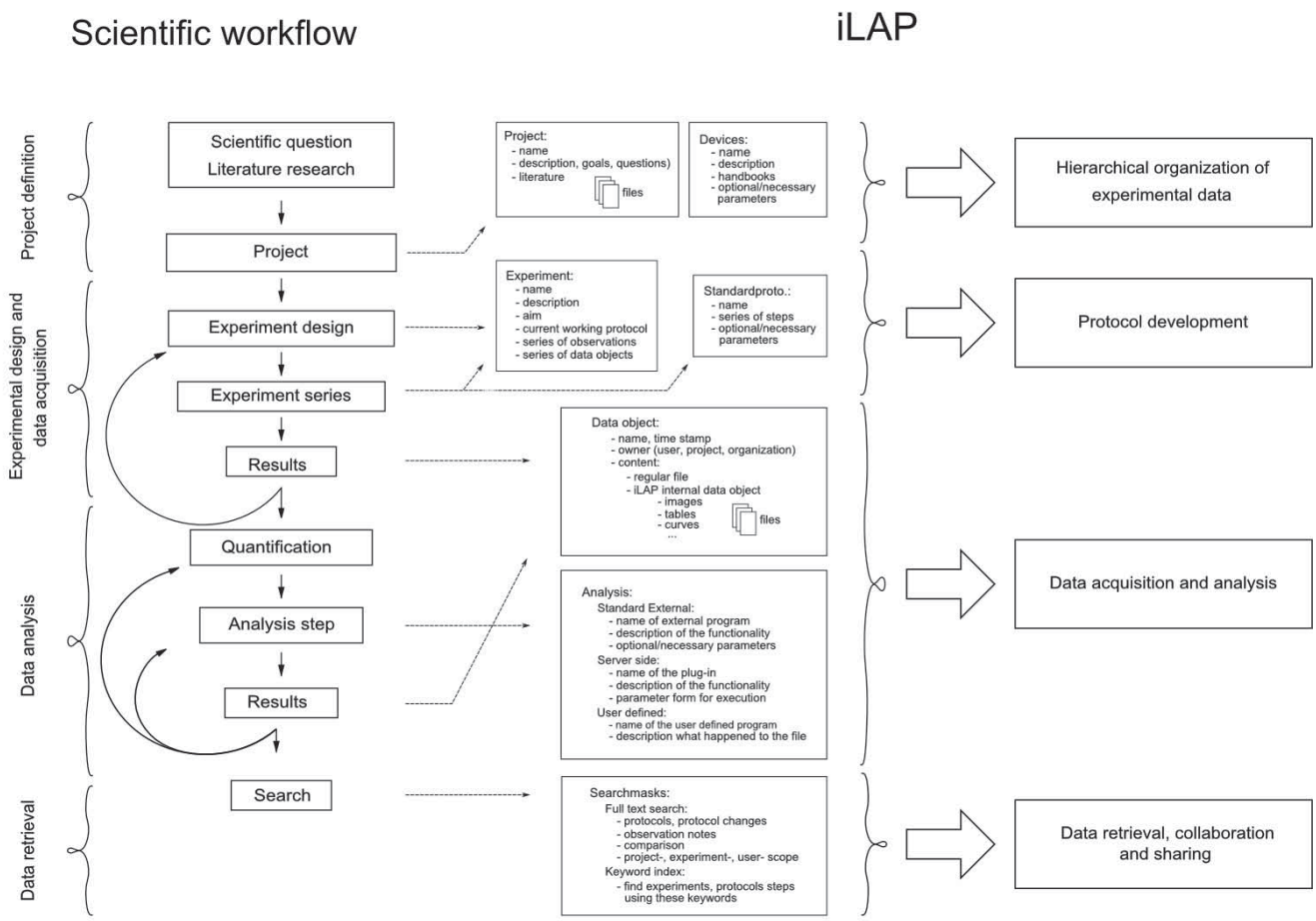

Figure 1

Mapping of the laboratory workflow onto iLAP features. The software design of iLAP is inspired by a typical laboratory workflow in life sciences and offers software assistance during the process. The figure illustrates on the left panel the scientific workflow separated into four phases: project definition, data acquisition and analysis, and data retrieval. The right panel shows the main functionalities offered by iLAP.

typically changing parameters of standard protocol steps (e.g. fixation times, temperature changes etc.) are important to record as they are used to improve the experimental reproducibility.

Equipped with a collection of standard operating procedures, an experiment can be initiated and the data generated. In general, data acquisition comprises not only files but also observations of interest, which might be relevant for the interpretation of the results. Most often these observations disappear in paper notebooks and are not accessible in a digital form. Hence, these experimental notes should be stored and attached to the originating protocol step, experiment or project.

Data analysis and processing

After storing the raw result files, additional analysis and post-processing steps must be performed to obtain processed data for subsequent analysis. In order to extract information and to combine it in a statistically meaningful manner, multiple data sets have to be acquired. The software workflow should enable also the inclusion of external analytical steps, so that files resulting from external analysis software can be assigned to their original raw data files. Finally, the data files generated at the analysis stage should be connected to the raw data, allowing connection of the data files with the originating experimental context.

Data retrieval

By following the experimental workflow, all experimental data e.g. different files, protocols, notes etc. should be organized in a chronological and project-oriented way and continuously registered during their acquisition. An additional advantage should be the ability to search and retrieve the data. Researchers frequently have to search

through notebooks to find previously uninterpretable observations. Subsequently, as the project develops, the researchers gain a different perspective and recognize that prior observations could lead to new discoveries. Therefore, the software should offer easy to use interfaces that allow searches through observation notes, projects- and experiment descriptions.

Software Architecture

iLAP is a multi-tier client-server application and can be subdivided into different functional modules which interact as self-contained units according to their defined responsibilities (see Figure 2).

Presentation tier

The presentation tier within iLAP is formed by a Web interface, using Tapestry [20] as the model view controller and an Axis Web service [21], which allows programming access to parts of the application logic. Thus, on the client side, a user requires an Internet connection and a recent Web browser with Java Applet support, available for almost every platform. In order to provide a simple, consistent but also attractive Web interface, iLAP follows usability guidelines described in [22,23] and uses Web 2.0 technologies for dynamic content generation.

Business tier and runtime environment

The business tier is realized as view-independent application logic, which stores and retrieves datasets by communicating with the persistence layer. The internal management of files is also handled from a central service component, which persists the meta-information for acquired files to the database, and stores the file content in a file-system-based data hierarchy. The business layer

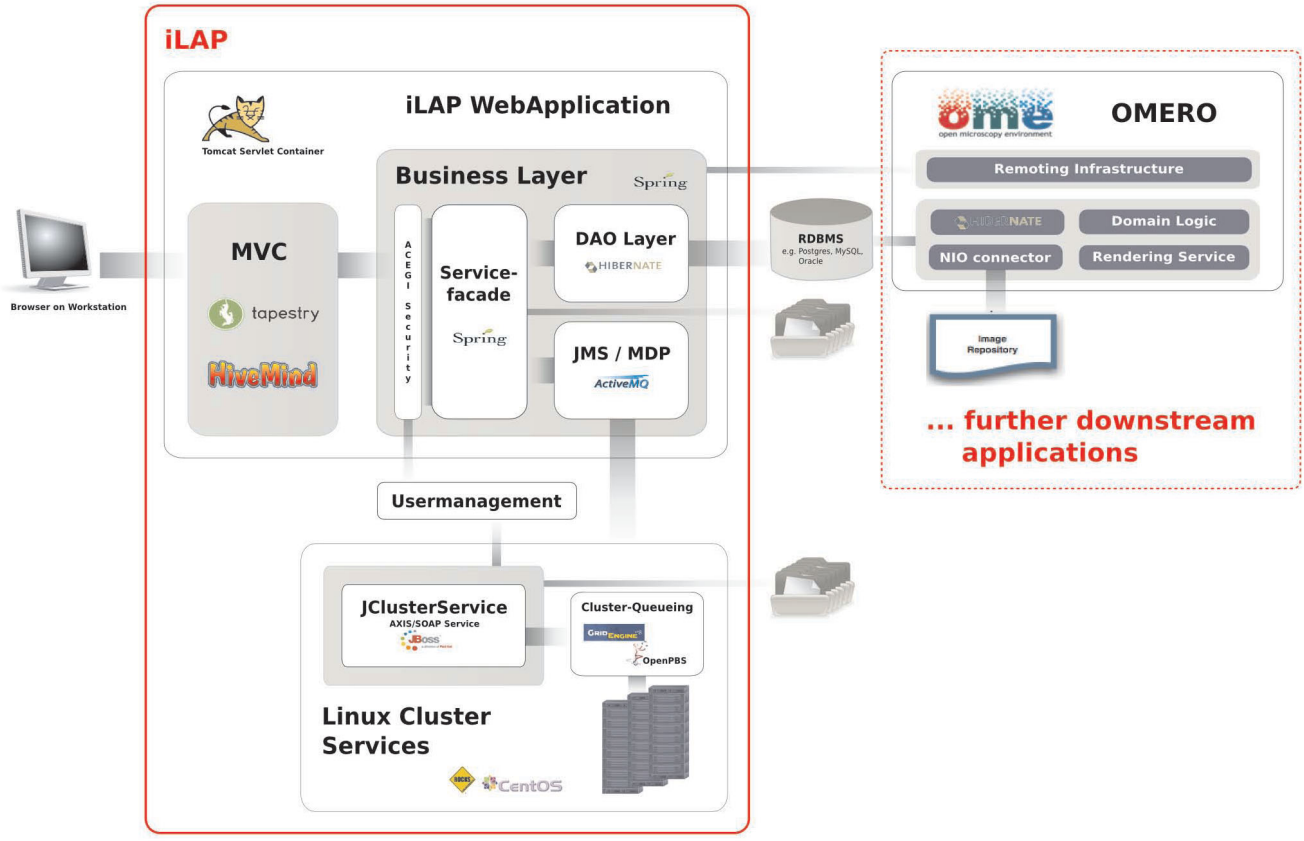

Figure 2
Software Architecture. iLAP features a typical three-tier architecture and can hence be divided into a presentation tier, business tier and a persistence tier (from left to right). The presentation tier is formed by a graphical user interface, accessed using a web browser. The following business layer is protected by a security layer, which enforces user authentication and authorization. After access is granted, the security layer passes the user requests to the business layer, which is mainly responsible for guiding the user through the laboratory workflow. This layer also coordinates all background tasks like automatic surveying of analysis jobs on a computing cluster or synchronizing/exchanging data with further downstream applications. (e.g. OMERO (open microscopy environment) image server). Finally, the persistence layer interacts with the relational database.

also holds asynchronous services for application-internal JMS messaging and for integration of external computing resources like high-performance computing clusters. All services of this layer are implemented as Spring [24] beans, for which the Spring-internal interceptor classes provide transactional integrity.

The business tier and the persistence tier are bound by the Spring J2EE lightweight container, which manages the component-object life cycle. Furthermore, the Spring context is transparently integrated into the Servlet context of Tapestry using the HiveMind [25] container backend. This is realized by using the automatic dependency injection functionality of HiveMind which avoids integrative glue code for lookups into the Spring container. Since iLAP uses Spring instead of EJB related components, the deployment of the application only requires a standard conformed Servlet container. Therefore, the Servlet container Tomcat [26] is used, which offers not only Servlet functionality but J2EE infrastructure services [27] such as centrally configured data-sources and transaction management realized with the open source library JOTM [28]. This makes the deployment of iLAP on different servers easier, because machine-specific settings for different production environments are kept outside the application configuration.

External programming interfaces

The SOAP Web service interface for external programmatic access is realized by combining the Web service framework Axis with corresponding iLAP components. The Web service operates as an external access point for Java Applets within the Web application, as well as for external analysis and processing applications such as ImageJ.

Model driven development

In order to reduce coding and to increase the long term maintainability, the model driven development environment AndroMDA [29] is used to generate components of the persistence layer and recurrent parts from the above mentioned business layer. AndroMDA accomplishes this by translating an annotated UML-model into a JEE-platform-specific implementation using Hibernate and Spring as base technology. Due to the flexibility of AndroMDA, application external services, such as the user management system, have a clean integration in the model. Dependencies of internal service components on such externally defined services are cleanly managed by its build system.

By changing the build parameters in the AndroMDA configuration, it is also possible to support different relational database management systems. This is because platform specific code with the same functionality is generated for data retrieval. Furthermore, technology lock-in regarding the implementation of the service layers was also addressed by using AndroMDA, as the implementation of the service facade can be switched during the build process from Spring based components to distributed Enterprise Java Beans. At present, iLAP is operating on one local machine and, providing the usage scenarios do not demand it, this architectural configuration will remain. However, chosen technologies are known to work on Web server farms and crucial distribution of the application among server nodes is transparently performed by the chosen technologies.

Asynchronous data processing

The asynchronous handling of business processes is realized in iLAP with message-driven Plain Old Java Objects (POJOs). Hence, application tasks, such as the generation of image previews, can be performed asynchronously. If performed immediately, these would unnecessarily block the responsiveness of the Web front-end. iLAP delegates tasks via JMS messages to back-end services, which perform the necessary processing actions in the background.

These back-end services are also UML-modelled components and receive messages handled by the JMS provider ActiveMQ. If back-end tasks consume too many calculation resources, the separation of Web front-end and JMS message receiving services can be realized by copying the applications onto two different servers and changing the Spring JMS configuration.

For the smooth integration of external computing resources like the high-performance computing cluster or special compute nodes with limited software licenses the JClusterService is used. JClusterService is a separately developed J2EE application which enables a programmer to run generic applications on a remote execution host or high-performance computing cluster. Every application which offers a command line interface can be easily integrated by defining a service definition in XML format and accessing it via a SOAP-based programming interface from any Java-application. The execution of the integrated application is carried out either by using the internal JMS-queuing system for single host installations or by using the open source queuing systems like Sun Grid Engine (Sun Microsystems) or OpenPBS/Torque.

Results

Functional overview

The functionality offered by the iLAP web interface can be described by four components: 1) hierarchical organization of the experimental data, 2) protocol development, 3) data acquisition and analysis, and 4) data retrieval and data sharing (Figure 1). iLAP specific terms are summarized in Table 1.

Hierarchical organization of experimental data

This part of the user interface covers the project definition phase of the experimental workflow. The definition of projects and experiments consists solely in inserting the required descriptive parameters via a Web form. In doing so, a hierarchical structure with projects, sub-projects and experiments is created and displayed in the iLAP overview. The hierarchy (Figure 3) and other screen shots can be found in the iLAP user manual (Additional file 1). This overview is the starting point of iLAP, from which almost every activity can be initiated. By navigating through the tree, an information box appears alongside. This box details information about the current node in the tree and the operations which can be performed on the database managed object represented by the node. Already in this early stage, files derived from literature research can be uploaded to projects and experiments, and ongoing observations can be stored using the general note dialog. If multiple files must be associated with projects and experiments, a Java Applet can be used to upload the files to the generated project/experiment structure. iLAP can manage every file independent of their file type, and can thus be considered as a generic document management system. File types only need to be considered for subsequent processing and data extraction.

Protocol development

When starting experimental work, the iLAP facility manager should define commonly used standard protocols using the protocol development masks. Therefore, a sequence of steps must be defined which describes the typical ongoing experiment in detail. Dynamic protocol parameters, which may be adapted for protocol optimization during the experiment, can be associated with the pre-defined steps. These parameters can be either numerical values, descriptive text or predefined enumeration types, all of which can be preset by default values and marked with appropriate units. In order to force the acquisition of critical parameters in the data acquisition wizard, parameters can be marked as required. According to our experience and the experience of other users, it is helpful to define small and reusable standard protocol units, which can be used as building blocks during the experiment-specific protocol assembly. Automatic internal versioning takes care of changes in standard protocols so that dependent protocols used in previous experiments remain unaffected.

Equipped with a collection of standard protocols, an experiment can be initiated and should be defined at the beginning of the workflow. The name of each experiment, its general description and specific aims, must be provided in order to be able to distinguish between different experiments. The detailed experiment procedure is defined by its current working protocol which can be composed step by step or by reusing existing current working protocols from already performed experiments. If the experiment is following a standard protocol, the current working protocol should be created by simply copying the predefined standard protocol steps and parameter definitions. In order to consider also the concurrent nature of simultaneously executed steps the experimenter should be able to define different sub-branches (e.g. cells are treated with different drugs in order to study their response)

Table 1: iLAP Terminology:

iLAP specific terms	Description
Project	Logical unit which can be structured hierarchically and holds experiments, notes and other files (e.g. derived from literature research).
Experiment	Logical unit which corresponds to one biological experiment and holds a current working protocol, experiment specific documentation files, parameter values, raw files, notes, and analysis steps.
Standard protocol	Frequently used and well established protocol template also known as standard operating procedures (SOP).
Current working protocol	Sequence of protocol steps for a specific experiment which holds raw files, notes and experiment specific parameter values.
Protocol step	One single step in a protocol which is defined by a name, description, and a list of definable parameters. A sequence of protocol steps defines a protocol.
Step group	Protocol step which groups multiple protocol steps to a logical unit. It can be used as a step container for sequentially executed protocol steps or within split steps.
Split step	Protocol step which can contain multiple (step groups) which have to be executed concurrently.
Protocol step parameter	Changing parameters which are associated with a step and can hold either textual or numerical values as well as a selection from a predefined value list (enumeration).
Note	Notes are textual descriptions which are intended to be used for documenting abnormal observations at almost anywhere within iLAP.
Raw file	Raw files are files which are produced by laboratory instruments and are not processed by any analysis step captured within iLAP.
Analysis step	Description of a processing step which manipulates, analyzes or processes a raw file, and generates processed files which are linked to the original raw file. Analysis steps can be either external e.g. using external software or internal using iLAP-internal analysis modules.
Analysis step parameter	Parameters and values used during the analysis step.

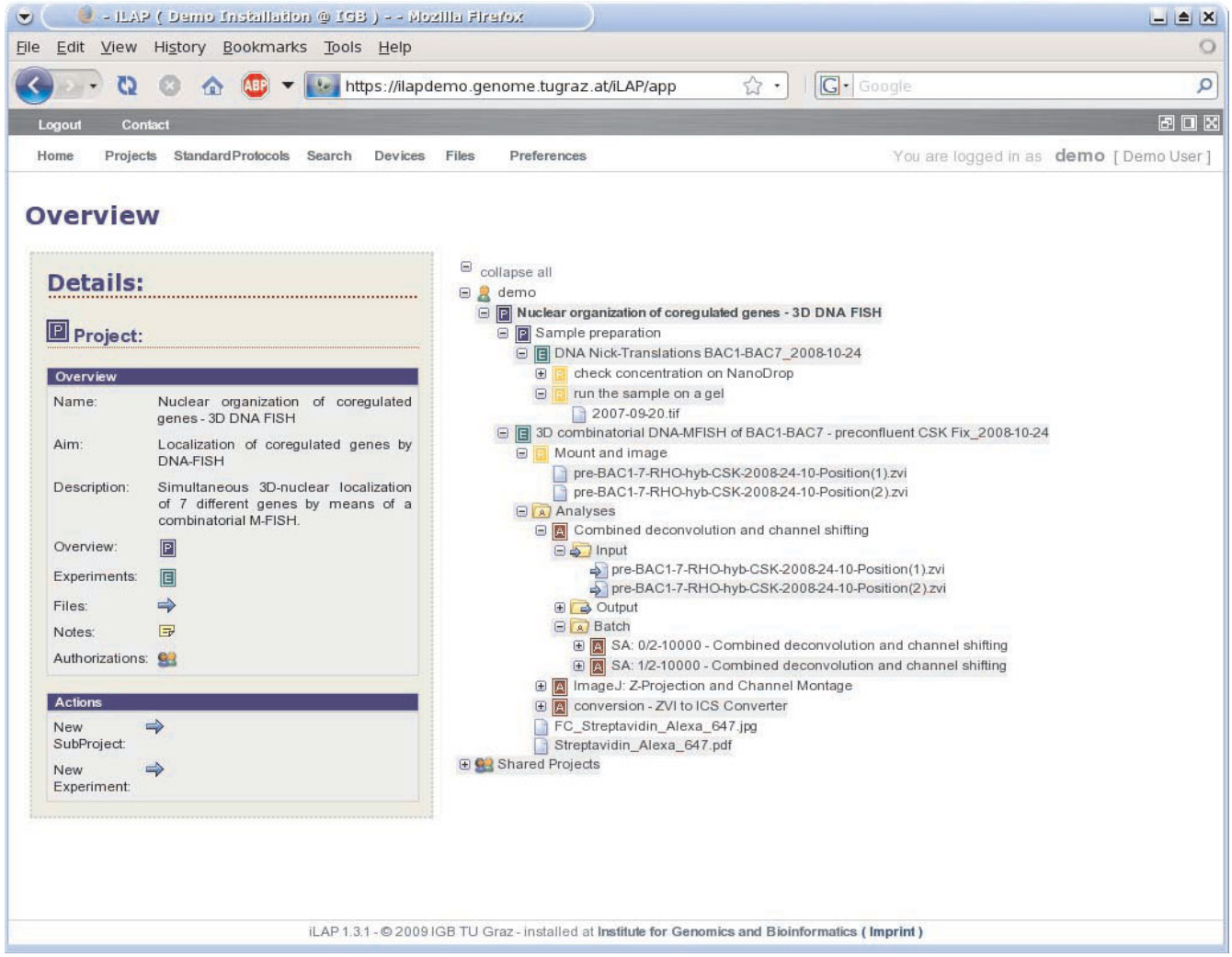

Figure 3

Hierarchical organization of data in iLAP overview. The continuous use of iLAP inherently leads to structured recording of experiments, conserving the complete experimental context of data records throughout the history of the research project. In doing so, a hierarchical structure with projects, sub-projects and experiments is created and can be displayed in this iLAP overview tree. The P-icons in the tree stand for projects and sub-projects, the E-icon for experiments and the A-icon for analysis steps. Files attached to protocol steps are considered as raw files and are therefore collected under the step container visualized with the R-icon. The consistent association of color schemes to logical units like projects, experiments, etc. can be directly recognized in this overview. By clicking on one of the tree icons on the left hand a detailed overview appears about the selected item. Also actions like creation of new projects etc. can be directly initiated using the quick-links in the "Actions" section of "Details".

named split steps. These split steps lead to different branches of the experimental workflow called step groups which are separately handled during the data acquisition phase.

Once the protocol design phase is completed and all necessary protocol steps with their parameters are defined the researchers should be able to generate a printout of the current working protocol with which the experiment can be performed at the lab bench.

Data acquisition and analysis

After having finished all experimental work and having created raw data files with different laboratory instruments the data acquisition within iLAP should be performed. By going through the early defined current working protocol steps, generated raw data files, used protocol parameter values and observation notes must be entered. Wizard-based input masks (wizard), which are derived from the defined current protocol steps, assist the experimenters during this work. On every step the user has

to fill in the value fields for required parameters and can attach files and notes to each of the steps. During the creation of the working protocol, it is important to name those steps to which files are attached in a descriptive way. Files that are directly connected to experimental steps are considered as raw files and are protected against deletion. Note, files can be linked to the protocol steps anywhere in iLAP, i.e. also before and after the data acquisition.

For this data association, the iLAP workflow offers also the possibility to transfer all generated files to a central repository and associate automatically files with their generating protocol step at once, using a Java Applet. All the internal linkages to protocol steps, experiments or projects are performed automatically without the need of any user interference. As the files are attached to a protocol and an experiment, the overall context is preserved and the likelihood of reproducibility of the same conditions is increased. Within iLAP experimental notes are stored and attached to the originating protocol step, experiment or project and are retrievable using a keyword based search mask

Data analysis

The analysis steps are recorded in iLAP by either reusing existing analysis templates or describing new analysis steps applied to the previously uploaded raw data files. Additional analysis tools can be developed in Java as described in the iLAP user manual (Additional file 1). According to the file type, internally implemented analysis steps or the description of externally performed analysis steps are associated with the raw data files. Result files from analysis programs together with the used parameters can be easily attached to analysis definitions. As an example, a server analysis tool was implemented for deconvolving three dimensional image stacks, executed on a remote high-performance computing cluster using the JClusterService (see methods).

Integration of external programs

A proof of concept about external access of programs using the iLAP application programming interface was shown by the implementation of a plugin for the widely used image processing software ImageJ [30,31]. This Java plugin enables ImageJ to transfer the image files directly from iLAP to the client machine. This functionality appears as a regular dialog in the graphical user interface of ImageJ, and allows upload of result files back into iLAP in a transparent manner.

Automatic post processing tool chain

Background tasks like the generation of previews are performed using the internal post-processing tool chain which is started asynchronously as soon as the files are associated with the originating experiment in iLAP. According to the detected file type, multiple post-processor steps are executed and results are automatically stored back into the database. This flexible system approach is also used to automatically inform and synchronize further downstream applications like OMERO [9] image server from the Open Microscopy Environment project. Therefore, iLAP is able to transfer files - transparently for the user - to a server where a comparable project/dataset structure is created.

Data retrieval and information sharing

The use of the described data acquisition features inherently leads to structured recording of experiments, conserving the complete experimental context of data records throughout the history of research projects. It is often necessary to go back to already completed experiments and to search through old notes. Therefore, iLAP offers search masks which allow keyword based searching in the recorded projects, experiments and notes. These results are often discussed with collaboration partners to gain different opinions on the same raw data.

In order to allow direct collaboration between scientists iLAP is embedded into a central user management system [4] which offers multiple levels of access control to projects and their associated experimental data. The sharing of projects can be done on a per-user basis or on an institutional basis. For small or local single-user installations, the fully featured user management system can be replaced by a file-based user management which still offers the same functionalities from the sharing point of view, but lacks institute-wide functionalities (Additional file 2). This is only possible because iLAP keeps the source of user accounts separated from the internal access control to enable easy integration of additional local or institution wide user management systems.

Since sophisticated protocols are crucial for successful experiments iLAP-users can export their protocols not only in PDF format (Additional file 3) but also in an exchangeable XML format (Additional file 4 and 5). In that way scientists can directly pass over their optimized protocols to partners who do not share the data using iLAP internally but need to get the protocol information transferred. The same XML files can be also used on a broader basis for protocol exchange using central eScience platforms like MyExperiments [32]. This platform aims for an increased reuse and repurpose of commonly shared workflows achieving at the same time reduced time-to-experiment and avoiding reinvention. Ongoing standardization efforts regarding the XML format like FuGE [10,11] are currently not supported but could be integrated in future versions of iLAP.

Case Study

In order to test the functionality of the system, we used a high-throughput microscopy study. The focus of this study was on the three dimensional nuclear localization of a group of seven genes. This required the development of a combinatorial multicolor fluorescence in situ hybridization (m-FISH) protocol. This protocol enables simultaneous detection and visualization of all seven genes by using a combination of three different fluorescent labels. The elaboration and optimization of m-FISH required many different protocol steps and parameters. Thus it was crucial to keep a record of any parameter and procedure changes during the process of protocol development. These changes were directly connected with data produced in the lab (e.g. concentration of the FISH probes, probe labeling efficiencies etc.) and the resulting imaging data. In the final combinatorial m-FISH protocol, 70 steps and 139 different parameters were present. Using this protocol we conducted 10 experiments and produced 1,441 multicolor 3D-Image stacks of which 984 were subsequently corrected for color shifts and processed by 3D-deconvolution performing 100 iterations of the maximum likelihood estimation algorithm available with the Huygens Deconvolution Software (Scientific Volume Imaging - SVI http://www.svi.nl). These image processing steps were realized as batch analysis in iLAP, which delegated the compute intensive procedure to a high-performance computing cluster and then stored all processed image stacks in the analysis container of the corresponding experiments. Afterwards FISH signals were detected and analyzed using a custom image analysis procedure which was realized as a Matlab (MathWorks Inc.) extension of Imaris (Bitplane Inc.) using the Imaris-XT programming interface. This extension automatically recorded FISH signal coordinates, signal to signal distances, the nuclear volume and several additional parameters of each imaged nucleus. These externally generated data files were transferred back into iLAP and stored in the context of the corresponding experiment as an external analysis step. A summary of the data acquisition and analysis is shown in Figure 4.

During the course of the study we observed several clear advantages of the iLAP system over a lab-book in paper form, which was maintained in parallel. The first and most valuable feature of iLAP is the direct connection between protocol steps and data files which cannot be realized using a paper lab book. A second notable advantage of the iLAP system was that lab-tasks that were performed in parallel or in overlapping time-frames could also be stored as such, whereas in the traditional lab book all tasks performed in the lab were written sequentially which implied a break-up of connected protocols. A third advantage was that iLAP allowed for rapid searching and finding of experiments, protocols and desired terms, which required only a few mouse clicks as opposite to the cumbersome search using a paper notebook. Moreover, iLAP enabled easy collaboration functionality, data backup or parameter completeness checks.

Conclusion

We have developed a unique information management system specifically designed to support the creation and management of experimental protocols, and to analyze and share laboratory data. The design of the software was guided by the laboratory workflow and resulted in four unified components accessible through a web interface. The first component allows the hierarchical organization of the experimental data, which is organized in a generic document management system. The second component focuses on protocol development using templates of standard operating procedures. Next, the data acquisition and analysis component offers the possibility to transfer the generated files to a central repository and to associate the files with the corresponding protocol steps. Additionally, external data analysis programs can be integrated and executed on a remote high-performance computing cluster. The last component enables collaboration and data sharing between scientists using iLAP on a user or institutional level as well as protocol transfer with external users.

Although designed in an experimental context for high-throughput protocols like microarray studies of gene expression, DNA-protein binding, proteomics experiments, or high-content image-based screening studies, iLAP has also proven to be valuable in low- and medium-throughput experiments. For example, protocols for qPCR analysis of gene expression using 96 and 384-well formats -a widely used technique- can be easily developed and can contribute significantly to establishment of robust assays. Moreover, since the workflow-oriented concept of iLAP offers the flexibility of a more general scientific data management system it is not limited to a special laboratory protocol, instrument, or type of molecule. For example, its application for next-generation sequencing is straightforward since similar requirements on the computational environment (increasing amount of data, integration of analysis tools, or use of high-performance computing infrastructure) have to be met.

In summary, we have developed a flexible and versatile information management system, which has the potential to close the gap between electronic notebooks and LIMS and can therefore be of great value for a broader community. Extensive tests in our and other labs have shown that the benefits of better information access and data sharing immediately result in reduced time spent managing information, increased productivity, better tracking and oversight of research, and enhanced data quality.

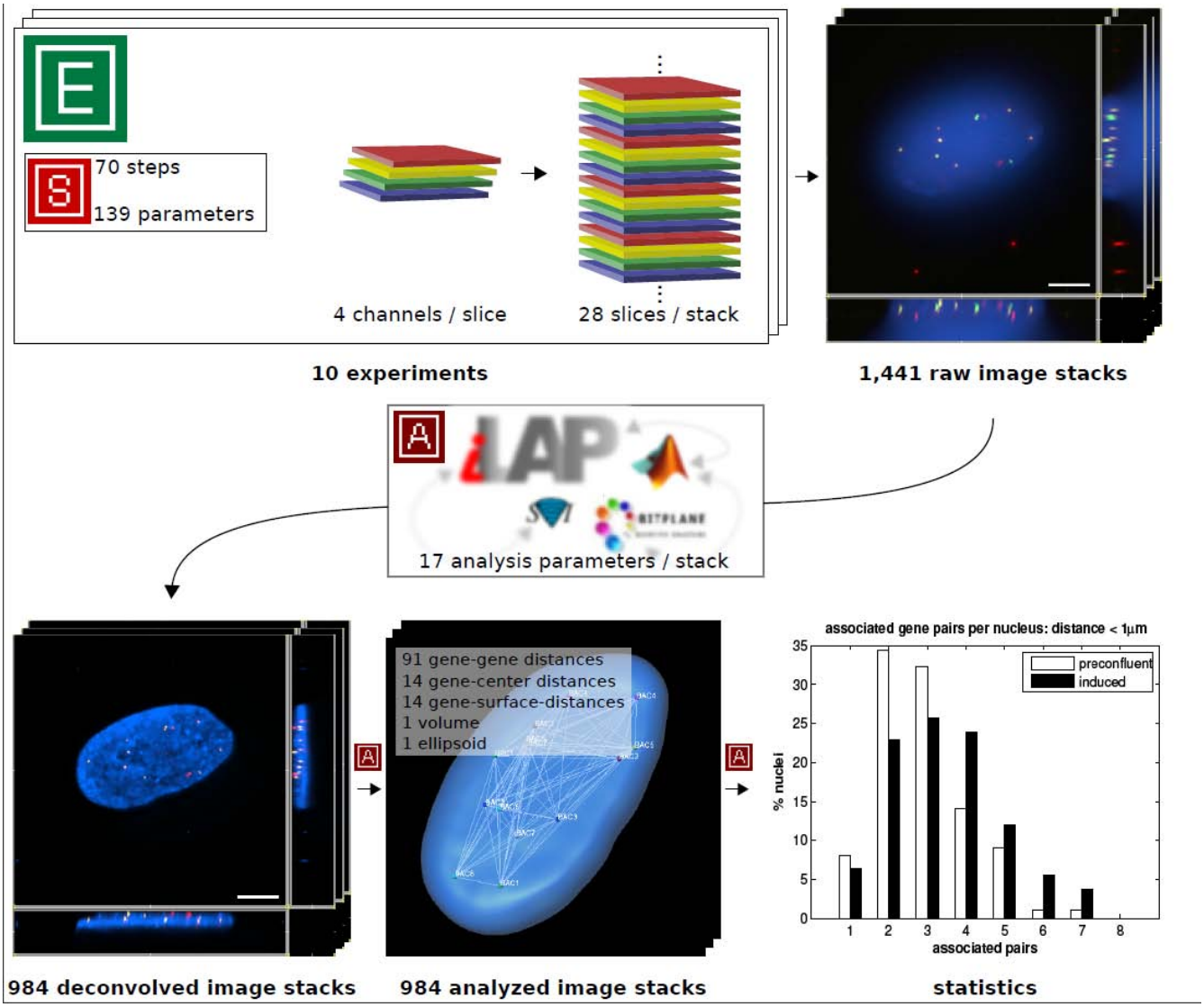

Figure 4

Case study summary. The functionality of iLAP was tested in a high-throughput microscopy study. The figure illustrates a summary of the data acquisition and data analysis performed. In 10 experiments a protocol consisting of 70 steps with 139 different parameters was used to generate three-dimensional multicolor image stacks. Each of the 1,441 raw image stacks consisted of 28 optical sections (slices) where each slice was recorded in 4 different channels. The raw image stacks were stored in the iLAP system and thereby connected with the corresponding experiments and protocols. By utilizing the integrated analysis functionality of iLAP the 984 raw images processed by the Huygens 3D-deconvolution package and analyzed by an external semiautomatic procedure implemented in Matlab and Imaris-XT. The analytical pipeline produced data for 121 different distance measurements of each single image. The resulting images and data were then stored in their experimental context within the iLAP system.

Availability and requirements

In order to reach a broader audience of users we have implemented a Java-based installer application, which is guiding an inexperienced computer user through the installation process (see Additional file 2). The basic installer package of iLAP has been tested on most common operating systems for which a Java Virtual Machine Version 1.5 or higher is available, e.g. Unix-based systems (Linux, Solaris, etc.), MacOS and Windows and can be downloaded from http://genome.tugraz.at/iLAP/. In addition to the requirement of a Java VM, a PostgreSQL database must be either locally installed or accessible via network. PostgreSQL comes with an easy-to-use installation wizard, so the complete installation should not be a significant entry level barrier. For further information about installation, please read the installation instruc-

tions from the download web site and in case of problems please contact the developers under iLAP@genome.tugraz.at. For initial testing purposes, please see also our test environment http://ilapdemo.genome.tugraz.at.

Regarding hardware requirements, the most critical issue is disk space for large data files. These are stored in a directory hierarchy where the base directory must be specified during the installation process. The requirements regarding processor performance and memory depend on the user basis, but PC or server hardware with 2 GB of RAM should be sufficient for most installations.

The production environment for our central in-house installation consists of a 4-processor AMD-system X4600 from Sun Microsystems, with 16 GB of RAM which is connected to an 8TB SAN storage. For computational intensive tasks, iLAP delegates the calculations to a 48-node high-performance computing cluster using the JClusterService interface.

Abbreviations

JavaEE: Java Enterprise Edition platform; MDA: Model Driven Architecture; OMERO: Open Microscopy Environment Remote Objects; SOAP: Simple Object Access Protocol; FuGE: Data standard for Functional Genomic Experiment.

Authors' contributions

The conceptual idea for iLAP goes back to GS, JGM and ZT and was elaborated by GS in the course of the GENAU-mobility/NIH-visiting-scientist program. GS and MF performed the implementation of the main software modules including persistence-, business- and web tier. SK implemented the data retrieval functionality and worked also on the integration of OMERO. GB together with GS was responsible for the Java-Applet-based file transfer functionality which was additionally extended to work as an ImageJ-Plugin. Archiving functionality and the XML export for general experiment protocol sharing was implemented by MO. DR contributed with extremely useful comments about conceptual ideas, their practical application and their usability. DR's constant input derived from permanent testing under real work conditions lead to major improvements in functionality, usability and responsiveness. The project was coordinated by GS.

Additional material

Additional file 1

***iLAP user manual**. The iLAP user manual contains a detailed description of the user interface including screen shots.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-390-S1.PDF]

Additional file 2

***iLAP installation and administration manual**. The iLAP installation and administration manual contains a detailed description of the installation process for all supported platforms including screen shots.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-390-S2.PDF]

Additional file 3

***m-FISH protocol in PDF format**. This file contains the combinatorial multiple fluorescence in situ hybridization (m-FISH) protocol in PDF format used in the case study section.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-390-S3.PDF]

Additional file 4

***m-FISH protocol in XML format**. This file contains the combinatorial multiple fluorescence in situ hybridization (m-FISH) protocol in XML format used in the case study section for protocol exchange.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-390-S4.XML]

Additional file 5

***Document Type Definition for the XML protocol format**. This file contains the Document Type Definition for the XML format used for protocol exchange created in collaboration with Will Moore from the OMERO.editor project.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-390-S5.DTD]

Acknowledgements

The authors thank the staff of the Institute for Genomics and Bioinformatics for valuable comments and contributions. Special thanks go to Hillary Mueller and Tatiana Karpova from the Laboratory of Receptor Biology and Gene Expression, National Cancer Institute. Special thanks also go to Will Moore for the ongoing successful collaboration regarding the common protocol exchange format. During the large scale analysis on the high-performance computing cluster we were supported by Scientific Volume Imaging http://www.svi.nl with a cluster license of Huygens Deconvolution Software. This work was supported by the Austrian Ministry for Science and Research, GEN-AU Project Bioinformatics Integration Network (BIN).

References

1. Acevedo LG, Iniguez AL, Holster HL, Zhang X, Green R, Farnham PJ: **Genome-scale ChIP-chip analysis using 10,000 human cells.** *Biotechniques* 2007, **43:**791-797.
2. Piggee C: **LIMS and the art of MS proteomics.** *Anal Chem* 2008, **80:**4801-4806.
3. Haquin S, Oeuillet E, Pajon A, Harris M, Jones AT, van Tilbeurgh H, *et al.*: **Data management in structural genomics: an overview.** *Methods Mol Biol* 2008, **426:**49-79.
4. Maurer M, Molidor R, Sturn A, Hartler J, Hackl H, Stocker G, *et al.*: **MARS: microarray analysis, retrieval, and storage system.** *BMC Bioinformatics* 2005, **6:**101.
5. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3:**SOFTWARE0003.
6. Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, Korner E, *et al.*: **MASPECTRAS: a platform for management and analy-**

sis of proteomics LC-MS/MS data. *BMC Bioinformatics* 2007, **8:**197.
7. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *J Proteome Res* 2004, **3:**1234-1242.
8. Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, Hussey P, *et al.*: **Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments.** *J Proteome Res* 2006, **5:**112-121.
9. Moore J, Allan C, Burel JM, Loranger B, MacDonald D, Monk J, *et al.*: **Open tools for storage and management of quantitative image data.** *Methods Cell Biol* 2008, **85:**555-570.
10. Jones AR, Pizarro A, Spellman P, Miller M: **FuGE: Functional Genomics Experiment Object Model.** *OMICS* 2006, **10:**179-184.
11. Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, *et al.*: **The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics.** *Nat Biotechnol* 2007, **25:**1127-1133.
12. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, *et al.*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29:**365-371.
13. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, *et al.*: **The minimum information about a proteomics experiment (MIAPE).** *Nat Biotechnol* 2007, **25:**887-893.
14. Deutsch EW, Ball CA, Berman JJ, Bova GS, Brazma A, Bumgarner RE, *et al.*: **Minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE).** *Nat Biotechnol* 2008, **26:**305-312.
15. Drake DJ: **ELN implementation challenges.** *Drug Discov Today* 2007, **12:**647-649.
16. Taylor KT: **The status of electronic laboratory notebooks for chemistry and biology.** *Curr Opin Drug Discov Devel* 2006, **9:**348-353.
17. Butler D: **Electronic notebooks: a new leaf.** *Nature* 2005, **436:**20-21.
18. Kihlen M: **Electronic lab notebooks - do they work in reality?** *Drug Discov Today* 2005, **10:**1205-1207.
19. Bradley J-C, Samuel B: **SMIRP-A Systems Approach to Laboratory Automation.** *Journal of the Association for Laboratory Automation* 2004, **5:**48-53.
20. Apache Software Foundation: **Tapestry web frame work.** 2009 [http://tapestry.apache.org/].
21. Apache Software Foundation: **Java implementation of the SOAP ("Simple Object Access Protocol").** 2009 [http://ws.apache.org/axis/].
22. Krug S: *Don't make me think! A Common Sense Apporach to Web Usability* Indianapolis, Indiana, USA: New Riders Publishing; 2000.
23. Johnson J: *Web Bloopers: 60 Common Web Design Mistakes and How to Avoid Them* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 2003.
24. SpringSource: **Spring lightweight application container.** 2009 [http://www.springsource.org/].
25. Apache Software Foundation: **Services and configuration microkernel.** 2009 [http://hivemind.apache.org/].
26. Apache Software Foundation: **Apache servlet container.** 2009 [http://tomcat.apache.org/].
27. Johnson R, Hoeller J: *Expert One-on-One J2EE Development without EJB. Wrox* 2004.
28. **OW2 Consortium** *Java Open Transaction Manager (JOTM)* 2009 [http://jotm.ow2.org/xwiki/bin/view/Main/WebHome?].
29. Bohlen M: **AndroMDA.** 2009 [http://www.andromda.org/].
30. Rasband WS: **ImageJ.** 2009 [http://rsb.info.nih.gov/ij/].
31. Abramoff MD, Magelhaes PJ, Ram SJ: **Image Processing with ImageJ.** *Biophotonics International* 2004, **11:**36-42.
32. Roure D, Goble C, Bhagat J, Cruickshank D, Goderis A, Michaelides D, *et al.*: **myExperiment: Defining the Social Virtual Research Environment.** :182-189.

Published online 18 June 2009

Nucleic Acids Research, 2009, Vol. 37, No. 15 ***e105***
doi:10.1093/nar/gkp526

Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays

Jochen B. Geigl[1], Anna C. Obenauf[1], Julie Waldispuehl-Geigl[1], Eva M. Hoffmann[1], Martina Auer[1], Martina Hörmann[2], Maria Fischer[3], Zlatko Trajanoski[3], Michael A. Schenk[2], Lars O. Baumbusch[4,5,6] and Michael R. Speicher[1,*]

[1]Institute of Human Genetics, Medical University of Graz, Harrachgasse 21/8, A-8010 Graz, [2]Das Kinderwunsch-Institut Schenk GmbH, Am Sendergrund 11, A-8143 Dobl, [3]Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14/V, 8010 Graz, Austria, [4]Department of Genetics, Institute for Cancer Research, [5]Department of Pathology, Norwegian Radium Hospital, Oslo University Hospital, 0310 Oslo and [6]Biomedical Research Group, Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, 0316 Oslo, Norway

Received February 13, 2009; Revised June 2, 2009; Accepted June 2, 2009

ABSTRACT

Clinical DNA is often available in limited quantities requiring whole-genome amplification for subsequent genome-wide assessment of copy-number variation (CNV) by array-CGH. In pre-implantation diagnosis and analysis of micrometastases, even merely single cells are available for analysis. However, procedures allowing high-resolution analyses of CNVs from single cells well below resolution limits of conventional cytogenetics are lacking. Here, we applied amplification products of single cells and of cell pools (5 or 10 cells) from patients with developmental delay, cancer cell lines and polar bodies to various oligo tiling array platforms with a median probe spacing as high as 65 bp. Our high-resolution analyses reveal that the low amounts of template DNA do not result in a completely unbiased whole genome amplification but that stochastic amplification artifacts, which become more obvious on array platforms with tiling path resolution, cause significant noise. We implemented a new evaluation algorithm specifically for the identification of small gains and losses in such very noisy ratio profiles. Our data suggest that when assessed with sufficiently sensitive methods high-resolution oligo-arrays allow a reliable identification of CNVs as small as 500 kb in cell pools (5 or 10 cells), and of 2.6–3.0 Mb in single cells.

INTRODUCTION

Many clinical applications, such as pre-implantation and non-invasive prenatal diagnosis, would benefit from the ability to characterize the entire genome of individual single cells by high resolution. Furthermore, in specific cancer research applications, such as the investigation of disseminated tumor cells (micrometastases) in bone marrow or circulating tumor cells in blood, often only single cells or very small cell numbers are available for analyses. The same applies to precancerous lesions, such as cells with dysplasia or early adenomas. In addition, due to the discovery that the genome of all humans has copy-number variations (CNVs) (1–3) and that these may contribute to phenotype variability and disease susceptibility (4), screening of whole genomes for CNVs represents one of the most fascinating areas in human genetics at present. More recently, evidence was reported that CNVs may arise in somatic cells resulting in somatic CNV mosaicism in differentiated human tissues (5,6). The prospect that the presence of some CNVs may be limited to confined somatic areas and their potential impact on physiological processes further fuels the need for reliable CNV screening approaches in small cell amounts.

*To whom correspondence should be addressed. Tel: +43 316 380 4110; Fax: +43 316 380 9605; Email: michael.speicher@medunigraz.at

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Comparative genomic hybridization (CGH) allows scanning of the whole genome for CNVs. However, CGH is usually performed with DNA extracted from thousands of cells and thus measures the average copy number of a large population of cells. Accordingly, CGH is sensitive to CNV heterogeneity within the cell population. Without preceding special, unbiased whole genome amplification, CGH is not amenable to single cell or few cell analyses.

Recently, first results were published describing the hybridization of single cell amplification products to various array platforms. Initial studies reported resolution limits of entire chromosomes (7) or of 34 Mb at best (8) and thus failed to demonstrate a significant improvement compared with conventional methodologies. By using the GenomePlex library technology for DNA amplification (GenomePlex Single Cell Whole Genome Amplification Kit, Sigma-Aldrich), we reported that copy number changes as small as 8.3 Mb in single cells can be detected reliably (9). Another group employed a 3.000 BAC array and achieved the detection of about 60% of gains, losses and interspersed normal regions 'smaller than 20 Mb' (10). Therefore, to the best of our knowledge, even the most advanced published single cell array-CGH technologies have resolution limits which represent only a slight improvement as compared to conventional CGH on metaphase spreads.

These earlier studies do not offer a detailed map of how robust a genome with CNVs is represented when whole genome amplification products are applied to oligo tiling arrays. To this end, we specifically selected clinical samples from some individuals in which previous analyses had revealed defined deletions on chromosome 22. We performed analyses on oligo tiling array platforms, which possess the highest density of oligonucleotides at present, i.e. NimbleGen's Chromosome 22 Tiling array (HG18 CHR22 FT) covering 385.210 oligos resulting in a median probe spacing of 65 bp and to a custom made chromosome 22 array (Agilent) with 241.700 oligo probes and a median probe spacing of 104 bp. In addition, we employed the NimbleGen Whole Genome Tiling Array (HG18 WG Tiling 2.1M CGH v2.0D) consisting of 2.1-million oligo probes, resulting in a median probe spacing of 1169 bp. During the evaluation of these cells, we noted that standard array CGH-evaluation programs are not suited for the evaluation of single cell amplification products and we therefore developed a new algorithm. In order to test the robustness of this algorithm and to start to address specific biological questions, we analyzed single cells from two cancer cell lines and polar bodies on a 244K whole genome array (Agilent).

As reported previously multiple displacement amplification with Φ29 polymerase results in different amplification of regions in relation to the GC content (11). The same applies to a linker adaptor whole genome amplification approach (12), because when these amplification products were hybridized to a BAC array GC rich regions on chromosome 19 had to be excluded from analysis (10). As we did not observe any nucleotide related amplification bias when applying the GenomePlex library technology to a tiling BAC array (9), we applied this amplification method to all experiments described here.

MATERIALS AND METHODS

Samples from clinical cases, cancer cell lines and polar bodies

We used cells from two probands (P1 and P2) with mental retardation and dysmorphic features in whom previous analyses performed on the whole genome 44K Agilent array had shown deletions on chromosome 22 with sizes of 2.8 Mb (P1) and 3 Mb and 1.2 Mb (both P2), respectively. Furthermore, we prepared new cells from the stable female renal cell carcinoma cell line 769P, because we are very familiar with this cell line from previous analyses (9) and the colorectal cancer cell line HT29, which is known to be chromosomally instable (13). For polar body analyses oocyte collection and processing were done according to standard protocols.

Isolation of single cells and whole genome amplification

Cultured cells were centrifuged at 700 *g* for 10 min, re-suspended in 1×PBS and transferred onto a polyethylene-naphthalate (PEN) membrane covered microscope slide (Zeiss, Austria) by cyto-centrifugation at 120 *g* for 3 min. After removing the supernatant, slides were air dried at room temperature overnight. Isolation of single cells and cell pools was carried out using a laser microdissection and pressure catapulting system (LMPC; P.A.L.M., Zeiss, Austria). Single cells and cells pools were randomly selected and directly catapulted into the cap of a 200 µl Eppendorf tube containing 10 µl digestion mix.

We performed whole genome amplification of the single cells and cell pools according to our recently published protocol (9,14). In brief, we employed the GenomePlex Single Cell Whole Genome Amplification Kit (#WGA4; Sigma-Aldrich, Germany) according to the manufacturer's instructions with some modifications. In a final volume of 10 µl, the specimens were centrifuged at 20.800 g for 10 min at 4°C. After cell lysis and Proteinase K digest, the DNA was fragmented and libraries were prepared. Amplification was performed by adding 7.5 µl of 10× Amplification Master Mix, 51 µl of nuclease-free water and 1.5 µl Titanium Taq DNA Polymerase (#639208; Takara Bio Europe/Clontech, France). Samples were amplified using an initial denaturation of 95°C for 3 min followed by 25 cycles, each consisting of a denaturation step at 94°C for 30 s and an annealing/extension step at 65°C for 5 min. After purification using the GenElute PCR Clean-up Kit (#NA1020; Sigma-Aldrich, UK), DNA concentration was determined by a Nanodrop spectrophotometer. Amplified DNA was stored at −20°C.

The quality of the amplification was evaluated using a multiplex PCR approach (15) and samples with four bands on an agarose gel were selected for further array-CGH analysis.

Array-comparative genomic hybridization (array-CGH)

We carried out array-CGH using various oligonucleotide microarray platforms as outlined in the text. For the analysis of amplified DNA samples, reference DNA amplified with the same protocol as described above was used.

Agilent platform. Samples were labeled with the Bioprime Array CGH Genomic Labeling System (#18095-12, Invitrogen, Carlsberg, CA) according to the manufacturer's instructions. Briefly, 500 ng test DNA and reference DNA were differentially labeled with dCTP-Cy5 or dCTP-Cy3 (#PA53021 and #PA55021, GE Healthcare, Piscataway, NJ). Slides were scanned using a microarray scanner (#G2505B; Agilent Technologies, Santa Clara, CA).

NimbleGen platform. Hybridizations on the 2.1 M whole genome array (HG18 WG Tiling 2.1M CGH v2.0D) and the chromosome 22 specific 385K array (HG18 CHR22 FT, both Roche NimbleGen Systems, Reykjavik, Iceland) were performed at service from Roche NimbleGen.

Array-CGH evaluation platform

Data normalization and calculation of ratio values were conducted employing NimbleGen's NimbleScan software package and the Feature Extraction software 9.1 from Agilent Technologies, respectively. The algorithm developed for this study focuses on detecting which ratio values differ significantly [two times standard deviation (SD)] from the ratio profile's mean and should therefore be considered as over- or underrepresented. The concept of the algorithm includes the employment of running means with different window sizes and analyses at progressively greater levels of smoothing and then combining these analyses.

The algorithm is implemented in 'R' (version 2.7.0) (16) and addresses three specific issues (i.e. location of windows, window size and threshold selection), which have a significant impact on the identification of very small CNVs in noisy CGH-profiles.

Positioning of windows. Consecutive data points are combined and their mean ratio values are presented in graphs of array-CGH results. The algorithm iterates through the profile by changing window positions, employing a sliding window approach.

The positioning of such windows may have an impact on the ability to detect small CNVs: the scheme in Supplementary Figure 1a illustrates a heterozygous deletion (black), the windows (red) used for the calculation of mean ratio values, and their calculated ratio profiles (blue). In the example on the left side, one window (light red) is located directly inside the deletion, thus the mean ratio value characterizing this region will reflect the actual DNA loss. In addition, the size of the deletion is shown correctly in the ratio profile. On the right side of Supplementary Figure 1a, the windows are positioned in such a way that two windows cover deleted and undeleted regions by half. As a result, these two windows are assigned mean ratio values generated in equal parts from balanced and lost regions. Therefore, the decrease of the ratio value will be lower and the region displayed in the profile (i.e. the size of the two windows) will be larger than the actual deletion.

Taking this into account, the algorithm calculates the mean ratio value for each window and assigns it only to the center of the respective window (Supplementary Figure 1b, blue dots).

As a consequence, CNVs do not appear with a sharp transition border at the location of breakpoints but as a more or less steep slope. For example, Supplementary Figure 1c shows the ratio profiles of the non-amplified DNA (upper panel) in comparison with the averaged ratio profile of the 10-cell pool (lower panel) obtained with DNA of proband P2. The 10-cell pool ratio was generated with a window size of 5.000 oligos (corresponding to 325 kb). Iterative calculations were made with windows of the same size, each moved by 1000 oligos. Note that the three largest CNVs (i.e. deletions with sizes of 3 and 1.2 Mb, and duplication of 532 kb) have already been correctly identified and are therefore shown in green and red, respectively. However, the ratio profile of the 10-cell pool shows no sharp change of the ratio values at the breakpoints.

Window size and threshold selection. The mean ratio value is calculated for each window based on the ratio values it contains. Assuming that a window's ratio values are distributed normally, we estimate the SD by considering the outmost value that is within $\pm 34.1\%$ of the mean. In our previous single cell experiments performed on BAC-arrays, we defined thresholds as ± 1.5 times the SD (9). Due to the higher noise on oligo-arrays as compared to BAC-arrays, thresholds had to be defined more stringently as ± 2 times the SD.

Importantly, when testing calculations with various window sizes we noted that different regions may be called over- or underrepresented. Supplementary Figure 2a and b illustrate again two calculations of the 10-cell pool of proband P2. Both calculations were made with fixed window sizes of 500 oligos (corresponding to 32.5 kb) (Supplementary Figure 2a) and 2.500 oligos (162.5 kb) (Supplementary Figure 2b). In each case, the mean ratio for the entire window and not only the center position is shown. When using the 500 oligo size windows, many of the respective mean ratio values at the chromosomal locations of the three largest CNV regions are above or below the thresholds and are therefore displayed in green and red. However, within these regions there are also many windows which are neither significantly increased nor decreased (black colored regions), and are therefore impeding the distinct identification of CNVs. On the other hand, there are no false positive calls. When using larger windows, e.g. 2.500 oligos, there are more regions within the three largest CNVs which are significantly increased or decreased (Supplementary Figure 2b). However, also some false positive regions are now identified which were not observed with the 500 oligo windows.

These data suggest that a simple increase of the window size alone may not be efficient for improvements of CNV

identification. At the same time the observation that different window sizes identify different regions as over- or underrepresented suggests that real CNVs should show specific patterns if the calculations are repeated with various window sizes. Furthermore, these patterns should enable to distinguish between false positive calls and real existing CNVs as illustrated in Supplementary Figure 2c. Panel (1) shows four different calculations, each with a different window size and threshold, as a different SD exists for every calculation. If a window shows a significantly increased or decreased mean ratio value, the mean position of that window will be displayed above or below the respective region of the ratio profile [panel (2)]. Depending on the window size it will be labeled with a different color and distance to the *X*-axis. The thus generated color bar code facilitates the estimation of the size of a CNV because the smaller the CNV the less color bars will be generated [panel (3); compare for example Figure 2a and b]. For more detailed size estimations the algorithm generates a table with all localizations of significant calls which allows the estimation of the CNV size very accurately.

A correctly identified CNV should show the smallest sized windows and also larger windows (depending on the size of the CNV) which have been determined as significant gains or losses [panel (2)]. Other bar code patterns should not occur as they suggest that regions identified as decreased or increased are more likely to be artifacts [panel (4)]: an example of this would be that no gains and losses are identified using the smallest windows but noted at larger window sizes [panel (4), left; for further examples see Supplementary Figures 6c and 7]. Due to the noisy CGH-pattern our algorithm does not require all windows to be detected as CNVs; although the majority of windows of a given size should be identified as gained or lost. Windows detected as CNVs should be continuous, thus no gap between the identification of two different window sizes should occur [panel (4), center]. A single call at any window size, except the smallest window size, is certainly an artifact [panel (4), right]. Therefore the pattern of identified regions with significant deviations from the mean ratio value can help to distinguish between true and false positives. This iterative color bar code generation avoids that a user has to adjust the window size for an individual experiment, therefore preventing the introduction of user bias.

The only user-defined option to interfere with the data representation is the selection which of the ratio profiles should be shown in the center.

RESULTS

Cells from clinical cases (probands P1 and P2) and establishment of their CNV status

We used cells from two probands (P1 and P2). Previous analyses performed on the whole genome 44K Agilent array had shown deletions on chromosome 22 with sizes of 2.8 Mb (P1) and 3 Mb and 1.2 Mb (both P2), respectively. When hybridizing non-amplified DNA to the NimbleGen Chromosome 22 Tiling array, we observed additional CNVs below the resolution limits of the 44K Agilent array. Proband P1 had an additional duplication of 272 kb (Figure 1a), whereas in proband P2 one additional deletion (size: 2.5 kb) and five duplications of various sizes (532, 335, 296, 255 and 85 kb) (Figure 1b) were observed. These additional CNVs, which had been unknown to us when we designed the experiments, turned out to be very useful for the estimation of resolution limits.

For each proband, we prepared cell pools, each consisting of 5 and 10 cells. In addition, we prepared one single cell from P2 and three different single cells from P1. Cell isolation by laser microdissection and subsequent hybridization were performed as previously (14). All experiments were conducted on the NimbleGen Chromosome 22 Tiling array (HG18 CHR22 FT), all amplification products of proband P2 were hybridized to the Agilent custom-made chromosome 22 array and the samples of proband P1 were additionally hybridized to the Whole Genome Tiling Array (HG18 WG Tiling 2.1M CGH v2.0D).

Evaluation of CNVs of probands P1 and P2 in noisy ratios in whole genome amplification products

As expected from our previous experience (9), amplification products yielded significantly noisier ratio profiles on the oligo-arrays than non-amplified DNA did. SDs are a reliable estimate of this noise (9). On the NimbleGen arrays the SDs of non-amplified DNA were in the range of about 0.3, whereas for amplified single-cell or cell-pool material they increased to values ranging from 0.45 to 0.7 (Table 1). By contrast, the SDs on the Agilent arrays were generally lower, i.e. about 0.1 for non-amplified DNA and 0.35–0.66 for amplification products (Table 1). When trying to evaluate these noisy ratios with currently used CGH-programs, such as those available on CGHweb (http://compbio.med.harvard.edu/CGHweb; e.g. CBS, CGHseg, cghFLasso), CNVs were not detected and/or the rate of false positive calls was high (data not shown). This reflects that present CGH programs are not designed for the evaluation of noisy ratio profiles.

We therefore developed a new CGH evaluation algorithm. New features of this algorithm include that the entire evaluation is conducted in an automated way without user interaction in order to avoid that selection of thresholds or sliding window sizes are influenced by user bias. The algorithm iteratively calculates values above or below thresholds for various window sizes, analyses the data at progressively greater levels of smoothing and then combines the data. These calculations result in a pattern distribution of regions identified as imbalanced, which allows to distinguish between artifacts and real imbalances and also to estimate the size of CNVs (details in 'Materials and Methods' section).

In a first step, we reevaluated the array-CGH profiles of the non-amplified DNA, shown in Figure 1, with this algorithm. As expected, all previously observed gains and losses could be identified again (Figure 2a and b). In addition, we evaluated the DNA of proband P2 on the custom-made Agilent Chromosome 22 Tiling array

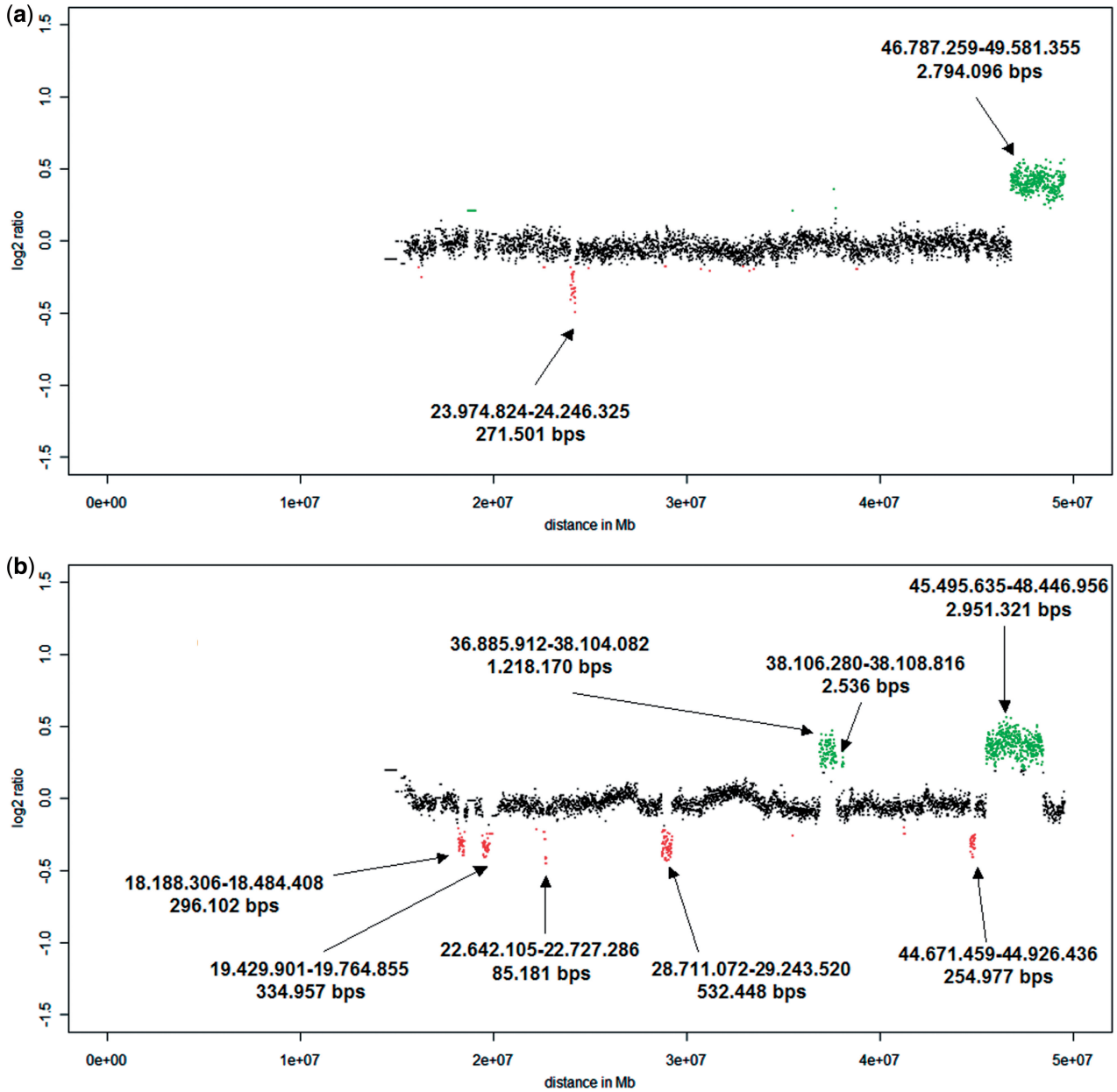

Figure 1. Ratio profiles of non-amplified DNA of probands P1 (**a**) and P2 (**b**) on the NimbleGen Chromosome 22 Tiling array. The calculation of these ratio profiles was based on a classical approach, using a window size of 100 adjacent oligos (corresponding to 6.5 kb) thresholds were simply determined as ±2 times SD. On the NimbleGen arrays losses are illustrated in green above the *X*-axis, whereas gains are shown in red below the *X*-axis. The sizes of observed CNVs are displayed at the respective locations.

(Supplementary Figure 3), yielding an almost identical ratio profile as on the NimbleGen array.

Results obtained with cell samples from proband P2

Analyses of amplification products obtained with 5- and 10-cell pools. The NimbleGen Chromosome 22 Tiling array comprises 385.210 oligonucleotides and has a median probe spacing of 65 bp. On this array platform, we detected the three largest CNVs of 3, 1.2 Mb and 532 kb with ease (Figure 3a and b) with both amplification products of the cell pools (5 or 10 cells). However, smaller CNVs could not be identified.

The custom-made Agilent array consists of 241.700 oligo probes with a median probe spacing of 104 bp. When applying the 5- and 10-cell pools to this array platform, we identified only the 3 Mb deletion in each case, but no other CNVs (Supplementary Figures 4a and b).

These results suggest that probe density on the array platform may have an important impact on resolution limits. Thus, depending on the array platform resolution limits for the CNV, detection in cell pools consisting of 5–10 cells are in the range of about 500 kb.

Analyses of amplification products obtained with a single cell. As expected, noise of the single cell amplification

Table 1. Summary of the standard deviations determined for each experiment on the various array-platforms

Proband	Sample	NimbleGen		Agilent
		Chromosome 22 array	Whole genome array	Chromosome 22 array
P1	Non-amplified DNA	0.29	0.29	ND
	Pool 10 cells	0.45	0.50	ND
	Pool 5 cells	0.42	0.68	ND
	Single cell #1	0.59	0.75	ND
	Single cell #2	0.80	0.89	ND
	Single cell #3	0.66	1.05	ND
P2	Non-amplified DNA	0.30	ND	0.11
	Pool 10 cells	0.51	ND	0.30
	Pool 5 cells	0.59	ND	0.35
	Single cell #1	0.87	ND	0.66

ND: Not done.

products was increased, which is also reflected in the SD (Table 1), and resulted in a poorer resolution. On the NimbleGen Chromosome 22 Tiling array, we clearly detected the 3 Mb-deletion, whereas smaller CNVs could not be identified (Figure 4). Similarly, this deletion was also detected on the Agilent Chromosome 22 Tiling array (Supplementary Figure 5).

These results suggest that CNVs in single cells with a size of 3 Mb can be detected on appropriate array platforms.

Results obtained with cell samples from proband P1

Analyses of amplification products obtained with 5- and 10-cell pools. In general, the hybridization patterns with the cell samples from proband P1 appeared to be noisier on the NimbleGen Chromosome 22 Tiling array as compared to proband P2. This is not reflected in the SDs (Table 1), which may be due to the fact that the SDs of proband P2 are increased as a result of the unexpected large number of CNVs on chromosome 22. When applying our evaluation algorithm, this increased noise is reflected in the multiple regions above the threshold, which could only be identified with small window sizes (Figure 5). In both cell pools (5 or 10 cells), we detected the deletion of 2.8 Mb, but not the duplication of 271 kb (Figure 5a and b). However, the 10-cell pool also identified a 650 kb large deletion at position 21 Mb (Figure 5a). As shown below, when the same amplification product was hybridized to another array platform, i.e. the NimbleGen Whole Genome Array, this deletion was not visible suggesting that this copy number change is a false positive result and was probably caused by a hybridization artifact rather than by an amplification artifact.

For proband P1, we could also compare the ratio profiles of the NimbleGen Chromosome 22 Tiling array with the NimbleGen Whole Genome Tiling Array. On the latter array, chromosome 22 is represented with 26.718 clones, corresponding to median probe spacing of 937 bp. We first compared the ratio profiles obtained with non-amplified DNA on both array platforms and found that these were nearly identical (Supplementary Figure 6a). With the amplification products of the 5- and 10-cell pools, we again detected the 2.8 Mb deletion in each case (Supplementary Figure 6b and c).

In this case, there were no significant resolution differences between the two array-platforms. In fact, the hybridization patterns on the whole genome tiling array appeared to be less noisy as compared to the chromosome 22 tiling array (compare Figure 5a and b with Supplementary Figure 6b and c). In summary, our results suggest that resolution limits for the CNV detection in cell pools consisting of 5–10 cells are in the range of ~500 kb.

Analyses of amplification products obtained with single cells. We hybridized three different single cell amplification products from proband P1 to the NimbleGen Chromosome 22 Tiling array. However, only in one of the three single cells ('Single cell #1') of proband P1, we were able to identify the 2.8 Mb deletion (Figure 6).

When repeating the single cell analyses of cells on NimbleGen's Whole Genome Tiling Array, we made the same observation, i.e. we discovered the 2.8 Mb-deletion only with the same amplification product from the cell which had allowed us to identify the deletion on the chromosome 22 tiling array (Supplementary Figure 7).

In order to get a more detailed insight whether CNVs with the size of 2.8–3.0 Mb are only borderline-detectable, we also evaluated well-known landmarks on the X-chromosome for hybridizations performed on the NimbleGen 2.1 M Whole Genome Tiling Array. The X-chromosome is represented by 106.458 oligos on this array. Proband P1 is male and the hybridization was carried out with female reference DNA. Due to the different sexes of proband and reference DNA certain landmarks regions on the X-chromosome should show a balanced profile, whereas other regions should show decreased ratio values. The balanced regions include the first pseudoautosomal region (PAR1; size: 2.6 Mb) at chromosome Xp22.3, the XY homology region (XY-HR; size: 4 Mb) at chromosome Xq21.3, and the second pseudoautosomal region (PAR2; size: 320 kb) at chromosome Xq28 (Supplementary Figure 8a). This expected hybridization pattern was indeed observed with non-amplified DNA (Figure 7a). Moreover, both the PAR1 and XY-HR were reliably detected in the cell pool hybridizations (Figure 7b and c) and even in all three single cells (Figure7d and Supplementary Figure 8b and c).

Analysis of single cells from two cancer cell lines. In order to further examine how reliably our new algorithm works, we tested single cells from two cancer cell lines on a 244K whole-genome array (Agilent). The first cell line was the female renal cell carcinoma cell line 769P. This cell line is chromosomally very stable as shown by our own previous analyses (9) and by other extensive studies employing M-FISH and array-CGH (17,18). Therefore, we expected that all analyzed cells should show an almost identical CGH-profile. The second cell line was colorectal cancer cell line HT29, which has a good level of chromosomal instability (CIN) with a highly reproducible modal chromosome number (13). Therefore, in this case, we estimated

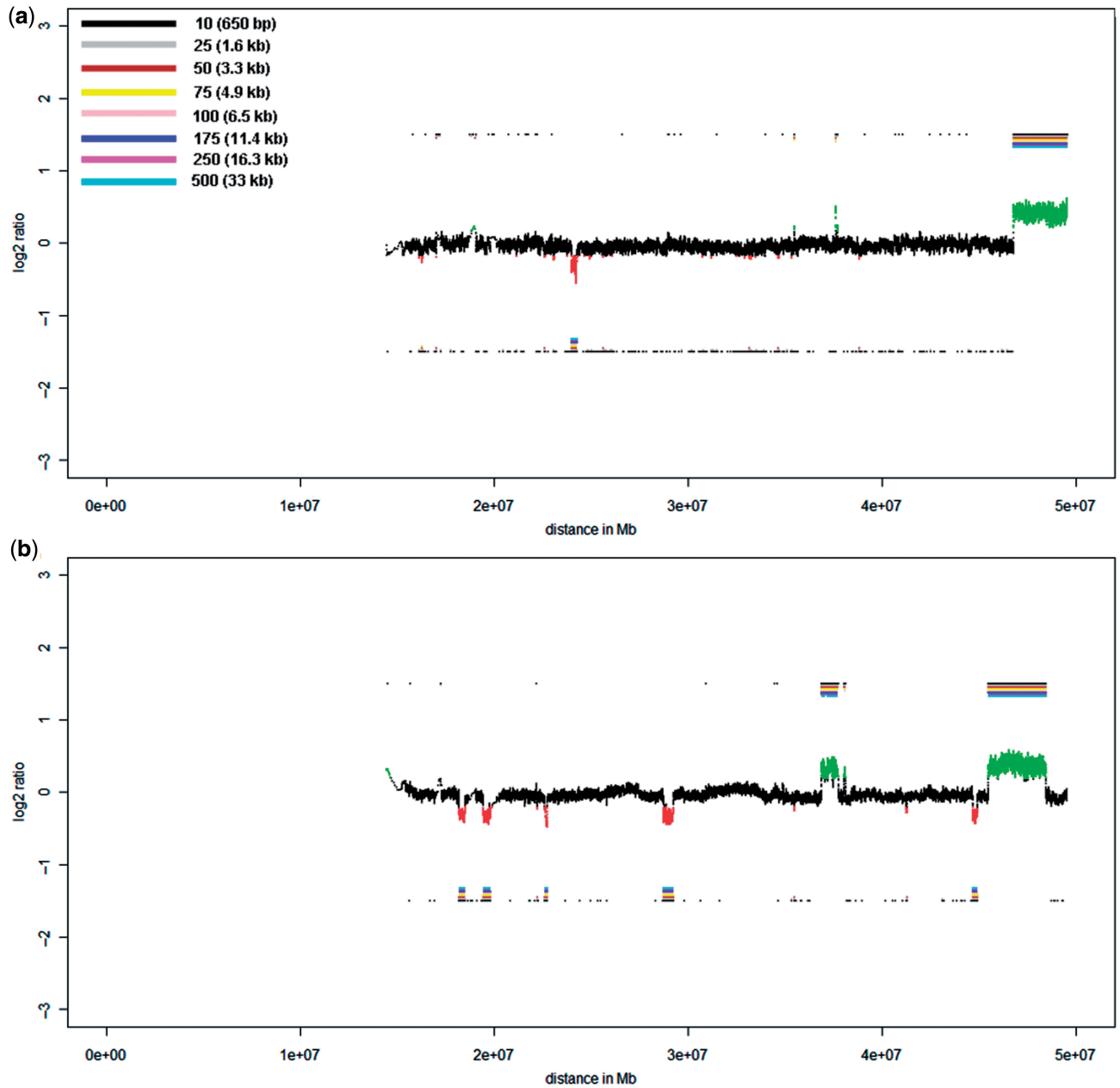

Figure 2. This figure displays the same ratio profiles as in Figure 1a and b, i.e. the ratio profiles of probands P1 (**a**) and P2 (**b**), now calculated with the algorithm described in this manuscript. The center profile is based on calculations with window sizes of 100 adjacent oligos (corresponding to 6.5 kb). A color bar code presents the window size (each in adjacent oligos and the respective physical size) for which calculations have been conducted. In the case of non-amplified DNA we selected very small window sizes, in the other cases with whole genome amplification products the window sizes were larger.

that these cells could show some cell-to-cell variation. Thus, in addition to testing our algorithm's robustness, we could also address the phenomenon of CIN, which is frequently observed in cancer and which is characterized by cell-to-cell variability (19).

In cell line 796P areas of copy number change identified by hybridization of non-amplified DNA could also be detected with the single cell products. To test the reproducibility of the algorithm we compared the ratio profile of non-amplified DNA (Supplementary Figure 9a) with four single cells which met our described quality criteria. For example, chromosome 1 harbors the equivalent of a single copy deletion on the p-arm covering a region of ~30 Mb and the equivalent of a single copy gain on the q-arm of ~90 Mb (9). 769P also has a small single copy deletion on chromosome 9 of 6.3 Mb (genomic position 16.7–23.0 Mb) (9). These regions of copy number change were easily identified in single-cell amplified material and non-amplified DNA (Supplementary Figure 9b). We indeed always discovered the same numerical aberrations and, notably, the ~6.3 Mb deletion on chromosome 9p was detected in each cell.

Cell line HT29 is near triploid and, according to previous publications, shows relative excess of chromosome arms

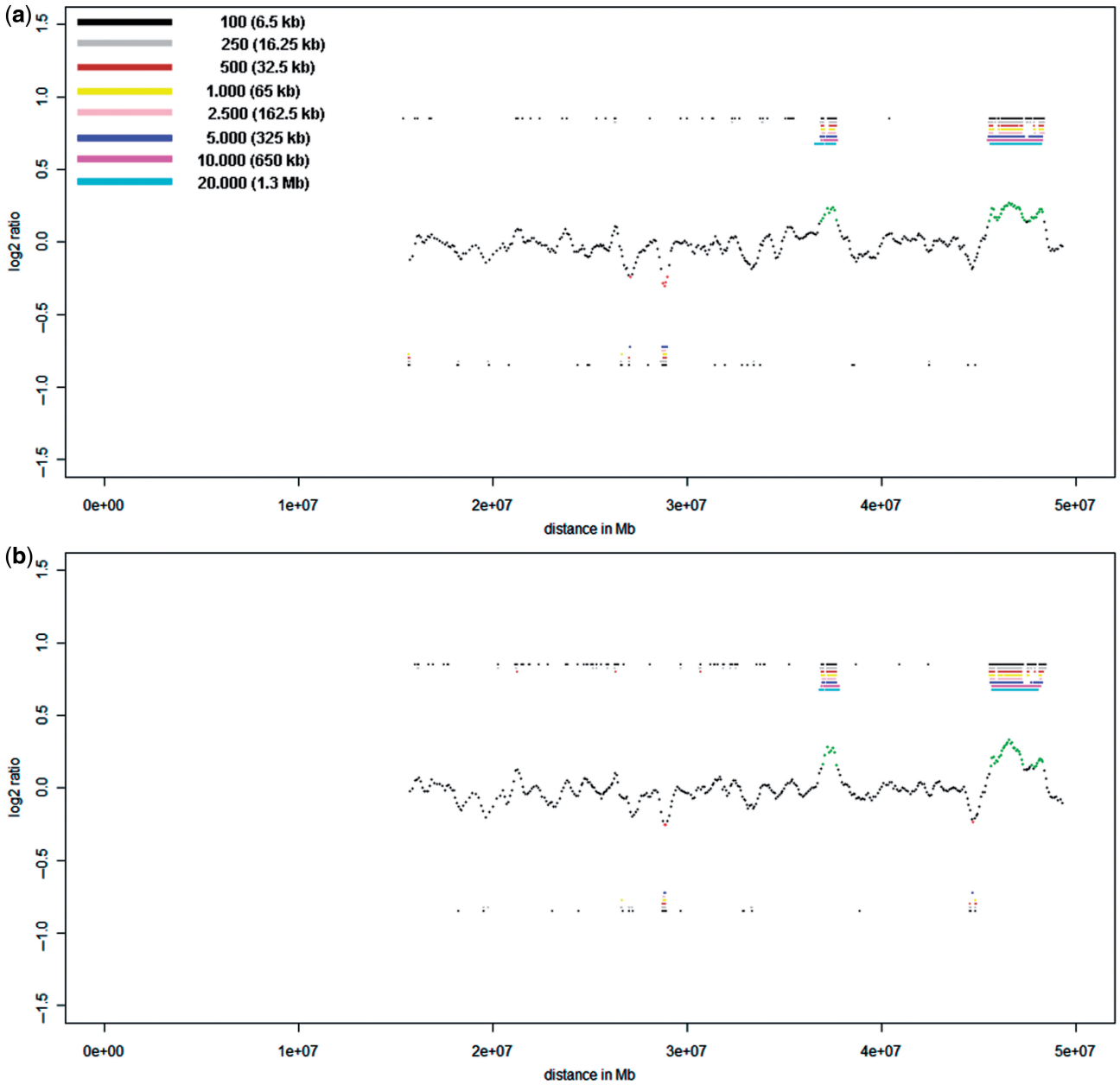

Figure 3. Cell-pool results obtained for proband P2 on the NimbleGen Chromosome 22 Tiling array. (**a**) Evaluation of the 10-cell pool on the NimbleGen Chromosome 22 Tiling array. The profile shown in the center was obtained with a window size of 5.000 oligos (corresponding to 325 kb). The two largest CNVs show bar codes from black to cyan, demonstrating that the size of the CNVs is in the range of 1.3 Mb or larger (actual sizes: 3 and 1.2 Mb, respectively; compare Figure 1b). In contrast, the largest duplication has a bar code ranging only from black to blue, showing that the size of this CNV is somewhere between 325 and 650 kb (the actual size is 532 kb, Figure 1b). To the left side of this duplication another region at position 26.5 Mb appears to be potentially duplicated. However, the calls are not uninterrupted from black to blue, as there is no pink bar revealing that this CNV call is likely to be an artifact [compare panel (4) in Supplementary Figure 2c]. (**b**) Hybridization of the 5-cell pool from proband P2 on the NimbleGen Chromosome 22 Tiling array resulted in a CNV recognition pattern similar to that of the 10-cell pool. The algorithm shows the presence of the 255 kb large duplication at position of about 44.7–44.8 (compare Figure 1b), however, the larger 296 and 335 kb duplications were not identified.

8q, 13q, 19q and 20q, relative deficiency of 8p, 14q, 17p, 18q and 21q, and pronounced intermetaphase variation (13). To the best of our knowledge, no high-resolution array-CGH profile of this cell line has been published yet. However, a partial high-resolution profile is available on the Agilent web-page (http://www.servicexs.com/blobs/Agilent/Agilent_CGH_brochure.pdf). Our array-CGH profile obtained with non-amplified DNA was consistent with previously published numerical aberrations (13) and with gains and losses described on the aforementioned Agilent web-page (Supplementary Figure 10a). This cell line also harbors two small homozygous deletions on 16p (size: 1.29 Mb; genomic position 6.0–7.3 Mb) and on 20p (size: 1.81 Mb; genomic position 14.2–16.0 Mb).

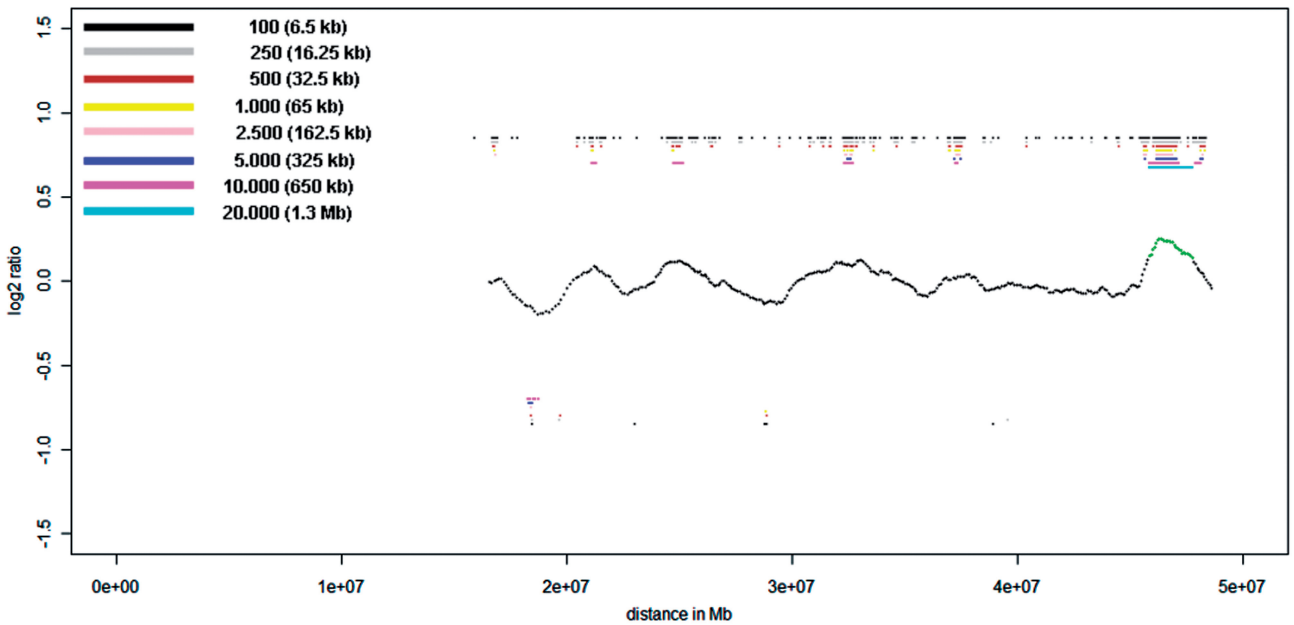

Figure 4. Chromosome 22 profile for proband P2 obtained with a single cell amplification product on the NimbleGen Chromosome 22 Tiling array. Beside the 3 Mb deletion, the bar code pattern displays a possible presence of two smaller deletions at positions 34 and 38 Mb with sizes between 650 kb and 1.3 Mb. The deletion at position 38 Mb corresponds to the location of the real existing 1.2 Mb deletion. However, the second putative deletion at position 34 Mb is false positive, demonstrating that CNVs with a size of <2 Mb cannot be reliably detected in a single cell. Here the center profile was obtained with a 20.000 oligo sliding window (1.3 Mb).

The aforementioned larger numerical changes were easily observed in all four different single cells shown in Supplementary Figure 10b–e. Interestingly, we could even unequivocally identify the small deletions on 16p and 20p in three cells (16p deletion: Supplementary Figure 10c–e; 20p deletion: Supplementary Figure 10b, d and e). In the other cells the ratios at the respective regions were decreased, yet they did not exceed the threshold. Thus, it may be especially easy to detect very small (<2 Mb) homozygous deletions in single cell amplification products.

As expected from the previously reported intermetaphase variation (13), we also observed some alterations not present in all cells. The best example is the deletion of the distal part of 6q. This deletion is easily visible with non-amplified DNA, however, the decrease of the ratio values is not as pronounced as e.g. for 3p or the distal region of 4q (Supplementary Figure 10a), suggesting that this numerical change may be present as mosaic. In fact, in two (Supplementary Figure 10c and d) of the four analyzed single cells, we observed a balanced ratio profile for the entire chromosome 6. In one cell (Supplementary Figure 10b) there was no gain of 18p, which was otherwise visible in all other cells and also in cells from non-amplified DNA. Furthermore, in another cell we observed a large, balanced region within an area on chromosome 7, which was overrepresented in all other cells (Supplementary Figure 10c). This suggests that CIN is in this cell line not only caused by whole-chromosome changes but also by structural rearrangements resulting in segmental aneuploidies. Applying single-cell array-CGH, we had previously made similar observations with the colorectal cell line HCT116 (9).

Analysis of polar bodies

Polar bodies represent an interesting model as chromosomal gains and losses observed in the first and second polar body should complement one another to a large extent. For example, a gain of a certain chromosome in the first polar body leaves two options for this chromosome for the second polar body: first the same chromosome could be lost, indicating a balanced status for this chromosome in the oocyte, or it could be balanced, indicating a loss of this chromosome in the oocyte. However, the gain of a certain chromosome should never be observed in both the first and the second polar body and the same applies for the loss of a chromosome. In preimplantation genetic diagnosis, we focus on polar bodies as Austrian legislation prohibits the analyses of blastomeres.

By now, we have analyzed by CGH 231 polar bodies, including 170 matching first and polar bodies demonstrating that our approach is highly reliable even for the analyses of haploid genomes (manuscript in preparation). Here we present an particularly interesting pair of first and second polar bodies showing complementary gains and losses for chromosomes 1, 9, 10, 13, 18, 20 and 21 (Supplementary Figure 11a and b). However, the first polar body had in addition a gain of chromosome 14 (Supplementary Figure 11a), whereas the second polar body had additional gains of chromosomes 16 and17 and losses of chromosomes 2, 3, 4, 6, 7, 11 and 15 (Supplementary Figure 11b). Thus, the corresponding oocyte should be unbalanced.

Inspection of the ratio profiles revealed another interesting phenomenon: in each polar body ratio, values were at four different levels. For example, in the first polar body (Supplementary Figure 11a), we observed chromosomes

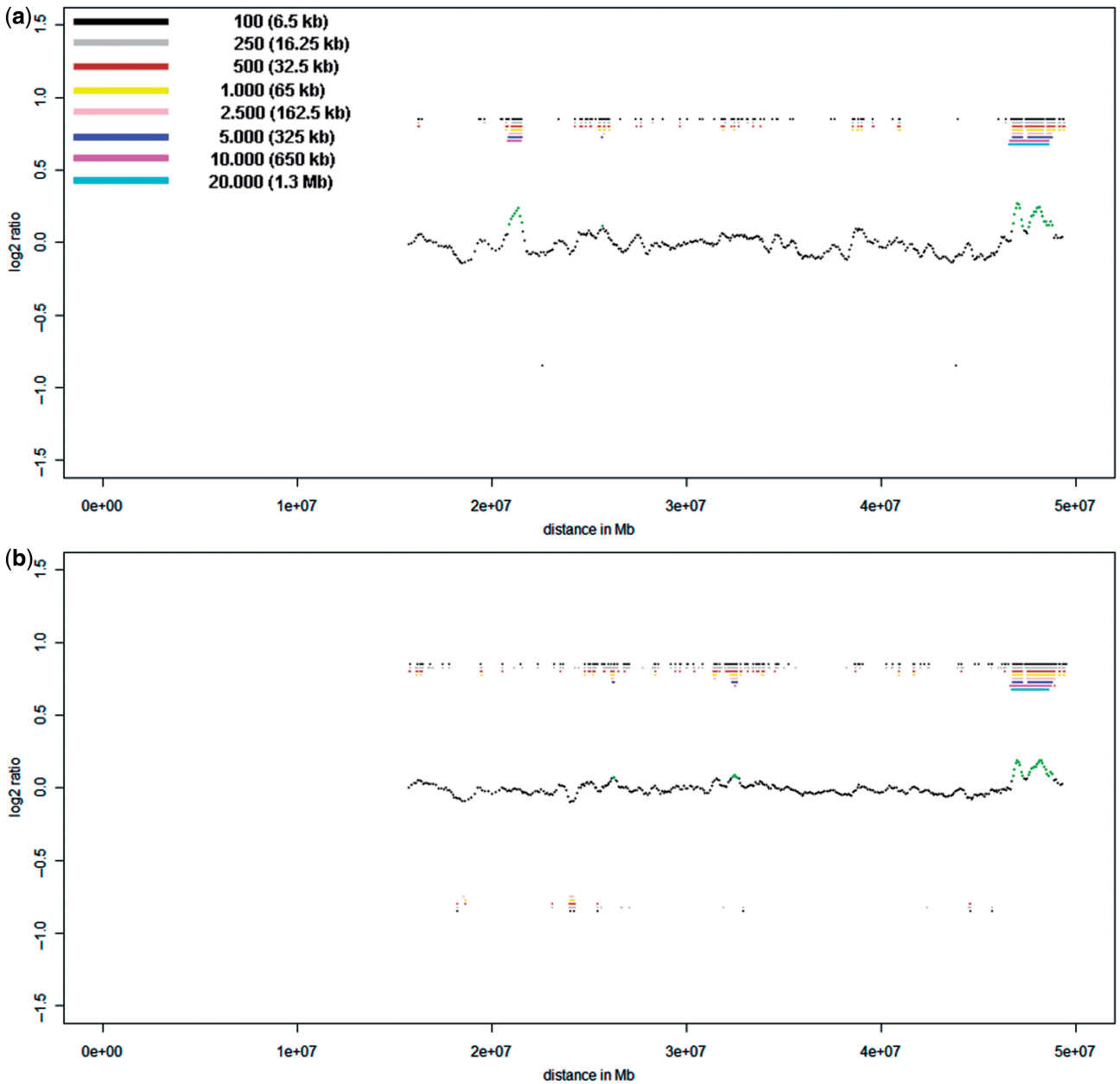

Figure 5. Cell-pool results obtained for proband P1 on the NimbleGen Chromosome 22 Tiling array. (**a**) Hybridization of the 10-cell pool clearly identified the 2.8 Mb-deletion. The algorithm also identified another deletion with a size of about 650 kb at position 21 Mb. This deletion is likely to be an artifact (compare Supplementary Figure 6b and details in text). (**b**) The 5-cell pool of proband P1 also allowed precise identification of the 2.8 Mb-deletion. In addition, at positions 27 and 32 Mb, the algorithm shows the possible presence of two further deletions, each with a size below the 500 kb limit for reliable CNV identification in cell pools. At position 23–24 Mb some bar codes reveal a duplication, which in fact corresponds to the real 272 kb duplication. In both cases the center profile was obtained with a sliding window of 5.000 oligos (325 kb).

with average ratio profiles of about 1 (i.e. chromosomes 10, 14 and 19), 0 (i.e. chromosomes 2, 3, 4, 6, 7, 11, 15, 22), −0.3 (i.e. chromosomes 5, 8, 12, 16, 17), and −1.5 (i.e. chromosomes 1, 9, 13, 18, 20, 21). These different ratio levels are indicated on the right side of each figure ('1–4'; Supplementary Figure 11). If the two meiotic divisions proceed without any errors, the first polar body should receive 23 chromosomes, each consisting of two chromatids, whereas the second polar body should get 23 chromosomes, each consisting of one chromatid. The four different levels of ratio values we observed in this and other (manuscript in preparation) polar body pairs most likely reflects that meiotic segregation errors even during meiosis I may involve not only chromosomes but also single chromatids. This pair of polar bodies and results from other polar bodies (our unpublished data) demonstrate that high rates of chromosome segregation errors may occur during female meiosis.

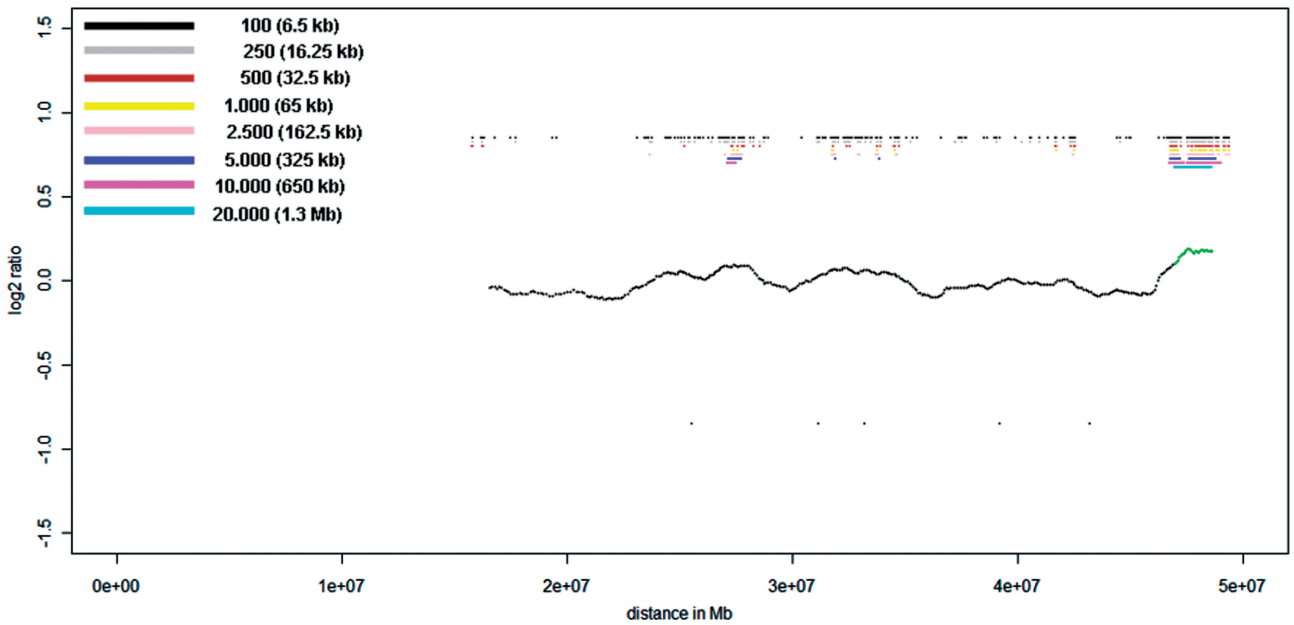

Figure 6. Identification of the 2.8 Mb deletion in a single cell ('#1') of proband P1 on the NimbleGen Chromosome 22 Tiling array. The center profile was generated using a 20.000 oligo sliding window (1.3 Mb).

DISCUSSION

In this study, we evaluated the performance of amplification products from cell pools or single cells on oligo tiling path arrays. Our results suggest that the use of arrays with a sufficient density of oligos allows the reliable detection of CNVs with a size of 3 Mb (P2) or 4 Mb (size of XY-HR). However, below 3 Mb, the detection of CNVs in single cells becomes critical as we missed a deletion of 2.8 Mb in 2 of 3 cells, whereas we identified the PAR1 region of 2.6 Mb on the X-chromosome in all of these three analyzed single cells. This indicates that reliable detection of CNVs with a size range of below 3 Mb is already at the resolution limit of present protocols for single cell analysis. In contrast, both robustness and resolution increase if only 5 or 10 cells are being analyzed, as we were able to identify CNVs as small as 500 kb in such cell pools.

Confirming our previous observations (9) our results again demonstrate that CGH-profiles from single cells or from a few cells are significantly noisier than those from non-amplified DNA. Amplification of the entire genome of a single cell most likely includes multiple stochastic amplification events due to the low amount of template DNA. Thus, while whole genome amplification products appear to be 'unbiased' at low resolution, e.g. if hybridized to metaphase chromosomes [as shown for example by (12) or (20)], variant amplification becomes more obvious on oligo tiling arrays and affects the detection sensitivity of small CNVs.

This requires particularly sensitive methods for data interpretation. Currently available array-CGH programs have been developed for the evaluation of ratio profiles with limited noise, which are usually achieved when non-amplified DNA is used.

In previous experiments when we (9) or others (8,10) tested the performance of amplified DNA on array-platforms, the standard procedure involved a comparison of ratio profiles obtained with amplified DNA versus a baseline profile usually generated with non-amplified DNA. Resolution is then estimated based on the concordance between the two ratio profiles. During these comparisons users will presumably adjust parameters, such as window smoothing or thresholds, until the best correlation between the profiles is achieved. However, since whole genome amplification of single cells or few cells involves a number of stochastic events, CGH-evaluation parameters, which may be optimal for a particular single cell experiment, may be less suited in the next experiment. Accordingly, lacking the option of a comparison with a baseline-ratio profile, the user will not know which parameters are optimal for a most sensitive CNV identification. In fact, in most scenarios performing single cell/few cell analyses reliable baseline profiles are not available for comparison, because otherwise there would be no need for an elaborate single cell analysis. Accordingly, our tests with various standard array-CGH programs revealed in fact that these had not been developed for noisy CGH-patterns and therefore they are not suited for the identification of very small changes in extremely noisy CGH ratio patterns.

For these reasons we developed a new algorithm with the specific aim of detecting small CNVs in very noisy ratio profiles. For the aforementioned reasons the algorithm excludes user interaction. Instead, ratios are iteratively calculated at progressively greater levels of smoothing and the analyses are then combined. This generates a pattern of regions gained or lost. Based on such a pattern the algorithm determines regions of significant ratio deviation. Thus, the main advantages over and differences from other CGH-programs include (i) no user interaction and thus avoidance of user bias; (ii) identification of small CNVs in noisy ratio profiles; (iii) distinction

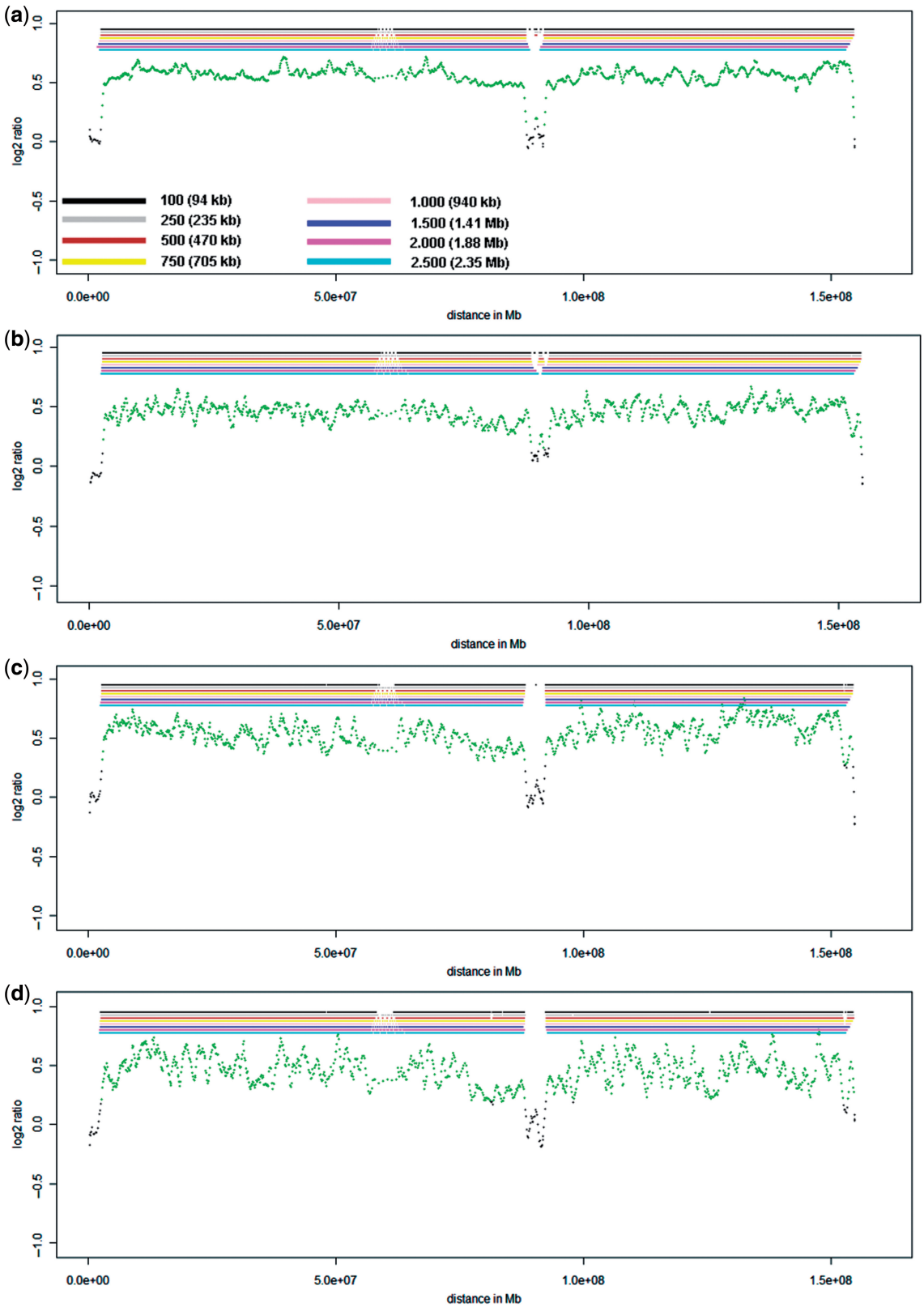

Figure 7. Ratio profiles of the X-chromosome. (**a**) Evaluation of the X-chromosome with non-amplified DNA. All X-chromosome landmark regions, i.e. PAR1, PAR2 and the XY-homology region (compare Supplementary Figure 8a) are identified. (**b**) X-chromosome evaluation of the 10-cell pool, which results in a similar ratio profile as obtained with the non-amplified DNA. (**c**) X-chromosome evaluation of the 5-cell pool, again with a similar ratio profile. (**d**) X-chromosome evaluation of the single cell '#1' from proband P1. For this cell the deletion on chromosome 22 was also identified.

between real CNVs and artifacts; and (iv) reliable size estimates for CNVs based on color coding and tables listing positions of over- and underrepresented regions.

Our comparisons of the ratio profiles between different chromosome 22 tiling array platforms and other oligo tiling arrays suggest that probe density on the array may have an important impact on the resolution limits. Furthermore, as demonstrated in our cell pool experiments, stochastic amplification artifacts are already reduced if only 5 or 10 cells are amplified, resulting in a drastic improvement of both robustness and resolution. This will pave the way for the establishment of detailed CNV-maps from small cell numbers. In addition, we demonstrated that specific biological questions can now be addressed with unprecedented resolution such as CIN in biological samples including cancer cells or polar bodies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Mag. Maria Langer-Winter for critically reading the manuscript.

FUNDING

European Commission (DISMAL project, contract no. LSHC-CT-2005-018911 and GENINCA, contract no. HEALTH-F2-2008-202230), the FWF Austrian Science Fund (P19455), the Österreichische Nationalbank (12569) and the Austrian Ministry for Science and Research (Project GEN-AU BIN); PhD-Program Molecular Medicine of the Medical University of Graz (to A.C.O.). Funding for open access charge: Medical University of Graz.

Conflict of interest statement. None declared.

REFERENCES

1. 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**, 949–951.
2. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
3. Tuzun,E., Sharp,A.J., Bailey,J.A., Kaul,R., Morrison,V.A., Pertz,L.M., Haugen,E., Hayden,H., Albertson,D., Pinkel,D. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
4. 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. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
5. Bruder,C.E., Piotrowski,A., Gijsbers,A.A., Andersson,R., Erickson,S., de Stahl,T.D., Menzel,U., Sandgren,J., von Tell,D., Poplawski,A. *et al.* (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.*, **82**, 763–771.
6. Piotrowski,A., Bruder,C.E., Andersson,R., de Stahl,T.D., Menzel,U., Sandgren,J., Poplawski,A., von Tell,D., Crasto,C., Bogdan,A. *et al.* (2008) Somatic mosaicism for copy number variation in differentiated human tissues. *Hum. Mutat.*, **29**, 1118–1124.
7. Hu,D.G., Webb,G. and Hussey,N. (2004) Aneuploidy detection in single cells using DNA array-based comparative genomic hybridization. *Mol. Hum. Reprod.*, **10**, 283–289.
8. Le Caignec,C., Spits,C., Sermon,K., De Rycke,M., Thienpont,B., Debrock,S., Staessen,C., Moreau,Y., Fryns,J.P., Van Steirteghem,A. *et al.* (2006) Single-cell chromosomal imbalances detection by array CGH. *Nucleic Acids Res.*, **34**, e68.
9. Fiegler,H., Geigl,J.B., Langer,S., Rigler,D., Porter,K., Unger,K., Carter,N.P. and Speicher,M.R. (2007) High resolution array-CGH analysis of single cells. *Nucleic Acids Res.*, **35**, e15.
10. Fuhrmann,C., Schmidt-Kittler,O., Stoecklein,N.H., Petat-Dutter,K., Vay,C., Bockler,K., Reinhardt,R., Ragg,T. and Klein,C.A. (2008) High-resolution array comparative genomic hybridization of single micrometastatic tumor cells. *Nucleic Acids Res.*, **36**, e39.
11. Lage,J.M., Leamon,J.H., Pejovic,T., Hamann,S., Lacey,M., Dillon,D., Segraves,R., Vossbrinck,B., Gonzalez,A., Pinkel,D. *et al.* (2003) Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res.*, **13**, 294–307.
12. Klein,C.A., Schmidt-Kittler,O., Schardt,J.A., Pantel,K., Speicher,M.R. and Riethmuller,G. (1999) Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc. Natl Acad. Sci. USA*, **96**, 4494–4499.
13. Abdel-Rahman,W.M., Katsura,K., Rens,W., Gorman,P.A., Sheer,D., Bicknell,D., Bodmer,W.F., Arends,M.J., Wyllie,A.H. and Edwards,P.A. (2001) Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement. *Proc. Natl Acad. Sci. USA*, **98**, 2538–2543.
14. Geigl,J.B. and Speicher,M.R. (2007) Single-cell isolation from cell suspensions and whole genome amplification from single cells to provide templates for CGH analysis. *Nat. Protoc.*, **2**, 3173–3184.
15. van Beers,E.H., Joosse,S.A., Ligtenberg,M.J., Fles,R., Hogervorst,F.B., Verhoef,S. and Nederlof,P.M. (2006) A multiplex PCR predictor for aCGH success of FFPE samples. *Br. J. Cancer*, **94**, 333–337.
16. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
17. Fiegler,H., Redon,R., Andrews,D., Scott,C., Andrews,R., Carder,C., Clark,R., Dovey,O., Ellis,P., Feuk,L. *et al.* (2006) Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.*, **16**, 1566–1574.
18. Fiegler,H., Carr,P., Douglas,E.J., Burford,D.C., Hunt,S., Scott,C.E., Smith,J., Vetrie,D., Gorman,P., Tomlinson,I.P. *et al.* (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, **36**, 361–374.
19. Geigl,J.B., Obenauf,A.C., Schwarzbraun,T. and Speicher,M.R. (2008) Defining 'chromosomal instability'. *Trends Genet.*, **24**, 64–69.
20. Gangnus,R., Langer,S., Breit,E., Pantel,K. and Speicher,M.R. (2004) Genomic profiling of viable and proliferative micrometastatic cells from early-stage breast cancer patients. *Clin. Cancer Res.*, **10**, 3457–3464.