DOI 10.1515/labmed-2013-0016 — J Lab Med 2013; 37(6): 305–315

Molekulargenetische und zytogenetische Diagnostik/ Molecular-Genetic and Cytogenetic Diagnostics

Redaktion: H.-G. Klein

Ina Vogl[a,*], Anna Benet-Pagès[a], Sebastian H. Eck[a], Marius Kuhn, Sebastian Vosberg, Philipp A. Greif, Klaus H. Metzeler, Saskia Biskup, Clemens Müller-Reible and Hanns-Georg Klein

Applications and data analysis of next-generation sequencing

Applikationen und Datenanalyse von Next Generation Sequencing

Abstract: Over the past 6 years, next-generation sequencing (NGS) has been established as a valuable high-throughput method for research in molecular genetics and has successfully been employed in the identification of rare and common genetic variations. Although the high expectations regarding the discovery of new diagnostic targets and an overall reduction of cost have been achieved, technological challenges in instrument handling, robustness of the chemistry, and data analysis need to be overcome. Each workflow and sequencing platform have their particular problems and caveats, which need to be addressed. Regarding NGS, there is a variety of different enrichment methods, sequencing devices, or technologies as well as a multitude of analyzing software products available. In this manuscript, the authors focus on challenges in data analysis when employing different target enrichment methods and the best applications for each of them.

Keywords: bioinformatics; data analysis; molecular genetic diagnostics; next-generation sequencing (NGS).

Zusammenfassung: In den vergangenen 6 Jahren hat sich “next generation sequencing” (NGS) als wichtige Hochdurchsatz-Methode für die molekulargenetische Forschung etabliert und wurde erfolgreich zur Identifikation seltener und häufiger genetischer Varianten eingesetzt. Während die hohen Erwartungen hinsichtlich der Entdeckung neuer diagnostischer Zielstrukturen und einer Senkung der Kosten erreicht wurden, müssen etliche technologische Herausforderungen hinsichtlich Bedienung der Geräte, Robustheit der Chemie und Handhabung der Datenanalyse noch gemeistert werden. Jedes Anreicherungsverfahren und jede Sequenzierplattform haben ihre spezifischen Probleme und Herausforderungen, die man in Betracht ziehen muss. In Bezug auf „next generation sequencing“ gibt es neben einer Anzahl verschiedener Anreicherungsmethoden, Sequenziergeräten und –techniken auch eine Vielzahl von unterschiedlicher Auswertesoftware. In diesem Artikel richten die Autoren ihr Augenmerk vor allem auf die Schwierigkeiten in der Datenanalyse der verschiedenen Anreichungsverfahren und deren bestmögliche Verwendung.

Schlüsselwörter: Bioinformatik; Datenanalyse; molekulargenetische Diagnostik; next generation sequencing (NGS).

[a]Ina Vogl, Anna Benet-Pagès and Sebastian Eck contributed equally to this study.

***Correspondence: Dr. Ina Vogl,** Center for Human Genetics and Laboratory Medicine Dr. Klein, Dr. Rost and Colleagues, Lochamer Str. 29, 82152 Martinsried, Germany, Tel.: +49-89/895578-0, Fax: +49-89/895578-78, E-Mail: ina.vogl@medizinische-genetik.de

Sebastian H. Eck and Hanns-Georg Klein: Center for Human Genetics and Laboratory Diagnostics Dr. Klein, Dr. Rost and Colleagues, Martinsried, Germany

Anna Benet-Pagès: MGZ München, Medizinisch Genetisches Zentrum München, Munich, Germany

Marius Kuhn: Genetikum, Ulm, Germany

Sebastian Vosberg: Clinical Cooperative Group ‘Leukemia’, Helmholtz Zentrum München, German Research Center for Environmental Health, Munich, Germany; and Laboratory of Leukemia Diagnostics, Department of Medicine III, Universität München, Munich, Germany

Philipp A. Greif and Klaus H. Metzeler: Laboratory of Leukemia Diagnostics, Department of Medicine III, Universität München, Munich, Germany

Saskia Biskup: CeGaT GmbH, Tuebingen, Germany

Clemens Müller-Reible: Institut für Humangenetik, Würzburg, Germany

Introduction – basic data analysis caveats

With the emergence of next-generation sequencing (NGS), the analysis of huge amounts of data becomes increasingly important. The quality and the validity of the detected variants vary greatly with the conditions of the library preparation process and the sequencing run. Additionally, data analysis has to be tailored to the specific workflow (amplicon based, enrichment based, whole exome, or genome) and the employed sequencing instruments. Each workflow and sequencing platforms have their particular problems and caveats, which need to be addressed. Platform-specific differences have been extensively discussed in previous publications of our group and others [1, 2]. In this work, we focus on challenges in data analysis when employing different target-enrichment methods.

Alignment

The basic task of sequence alignment is the mapping of millions of short read fragment from about 50 bp to 900 bp to reference sequences that may comprise the complete human genome. Various issues, e.g., variation in base quality, repeat content, or other sequence properties, can influence the correct alignment. Consequently, it is important to choose the appropriate settings and the correct reference sequence for each method.

Tag sequence and read trimming

Nonspecific tag sequences generally have to be removed from the read. Particularly at the end of reads, the tag sequence may pose problems due to its variation in length, possibly leading to false positives in the variant call (Figure 1). Usually, software provide options for finding this kind of sequences, which can be used to trim all kinds of adaptor sequences, tag sequences, and primer sequences.

Furthermore, read trimming after demultiplexing is recommended due to possible lower base calling quality in the first few bases and toward the end of the read. Different types of trimming can be performed based on quality scores, stretches of Ns and a specified number of bases at either the 3′ or 5′ end of the reads.

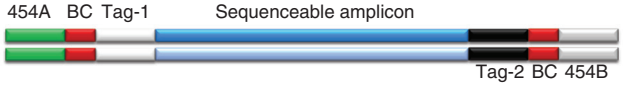

Figure 1 Schematic overview of a ROCHE454 amplicon structure. The 454A and 454B adaptors are usually already trimmed by the 454 software.

Repeat content and homologous regions

In general, the exact target region needs to be validated, and homologous sequences like pseudogenes have to be excluded from further analysis. These regions similar to the region of interest (ROI) can cause false-positive results due to misalignment of sequence reads. This can be avoided either by assay design or at the alignment level. In the case of amplicon-based target enrichment, similar to Sanger sequencing, target-specific primers need to be designed. Discrimination at the alignment level has to be made for enrichment by oligonucleotide hybridization and capturing. This requires software that is able to map all given reads to the entire human genome. Additionally, a local alignment strategy (Smith-Waterman like algorithm [3, 4]) is needed for amplicon-based target enrichment, while a global alignment (Needleman-Wunsch-like algorithm [5]) is required for enrichment by oligonucleotide hybridization and capturing. Moreover, the penalty settings for mismatches and gaps of the algorithm are a necessary tool to precisely discriminate homologous sequences. Another indication for homologous regions are reads that can be mapped equally well to different genomic locations. We recommend a software that is able to flag these reads as multiple matches and ignore them during further analysis.

Illumina paired-end sequencing (Illumina, San Diego, CA, USA [6]) offers an additional feature, which can be exploited to distinguish target gene regions from highly similar regions. Reads whose corresponding paired read could not be mapped or was mapped outside the acceptable insert size parameters are marked as broken pair. This can also be caused by large deletions, insertions, or any genomic rearrangements. These reads should be ignored during subsequent analysis in a diagnostic setting and should not contribute to variant calling (Figure 2).

Variant calling

Variant calling is the next crucial step in the analysis of NGS data. Variants are identified by comparing the aligned short reads to the reference genome. The variants may either be causative for disease, or they may simply represent benign genomic variation without a functional

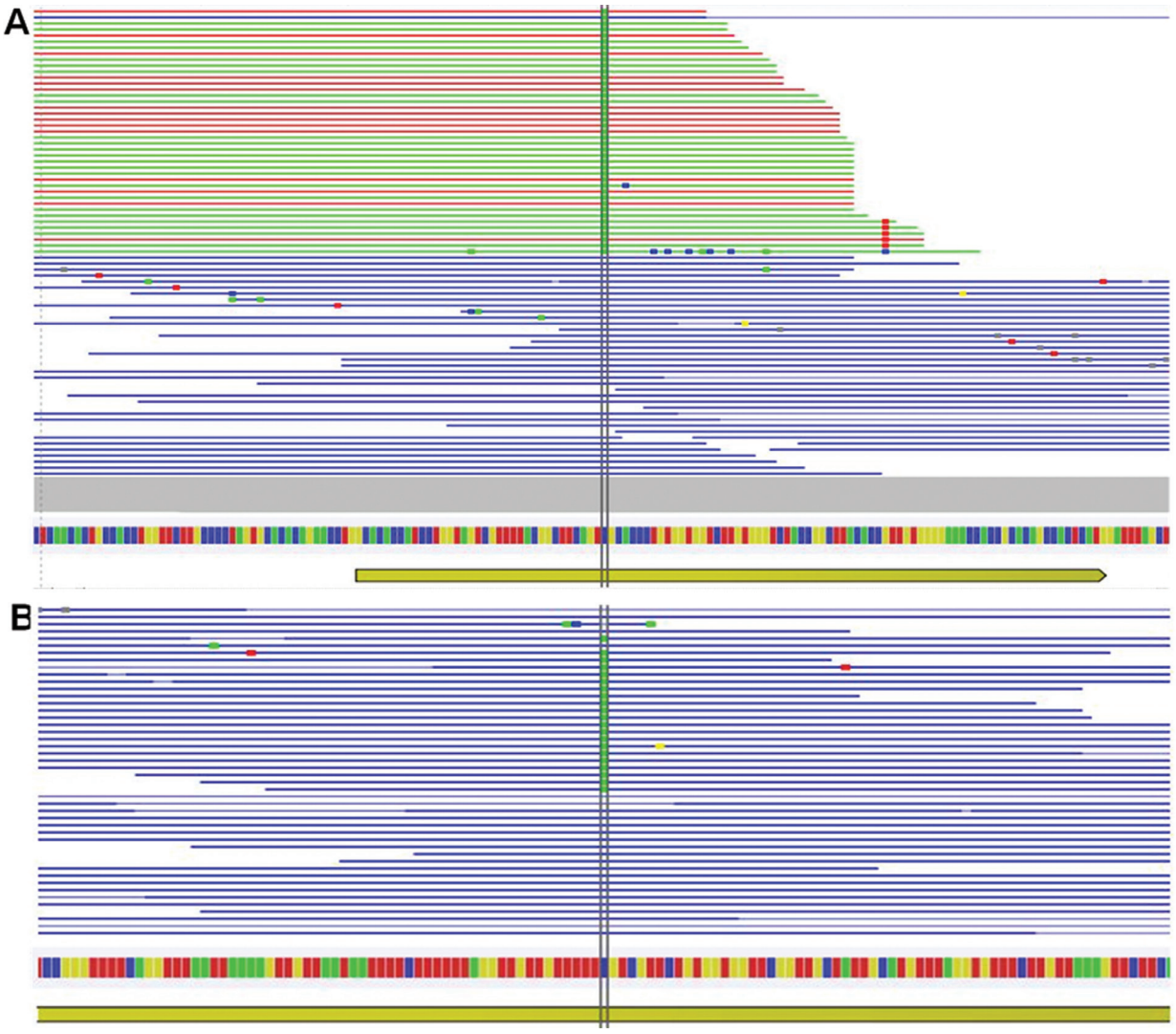

Figure 2 Screenshots of the CLCbio Genomics Workbench.
Reads mapped in pairs are indicated by blue, reads with unmapped paired-end read are indicated by red and green (broken pairs).
(A) A false-positive variant call with the variant allele primarily present in broken reads is highlighted in green between the black lanes.
(B) A validated variant call, with the variant allele primarily present in paired-end reads is highlighted in green.

effect. The standard format for storing and exchanging sequence variation (including SNVs, indels, and larger structural variations) is the so-called variant call format (VCF). The main challenge of variant call is to distinguish true genomic variants from sequencing errors and artifacts. The capability to accurately identify genomic variation is a crucial step in the detection of disease-associated mutations. There are several factors that can interfere with variant detection. First, the presence of short insertions and deletions may lead to false-positive SNV identification, especially if the chosen alignment algorithm is unable to perform a gapped alignment. Second, PCR artifacts introduced during the library preparation may be falsely called as variants. Last, sequencing quality gradually degrades over the read. In consequence, this means that lower-quality bases tend to accumulate toward the end of the read, which may lead to erroneous variant calls [7].

Forward reverse read balance

As a true-positive variant would be expected to be read from both directions, the forward and reverse read balance is a measure of the validity of a certain variant call. If the variant is observed in just one type of the reads, this might be an indication for a sequencing error or an artifact. Paired-end amplicon sequencing is a clear exception

to this, as reads may not overlap and are only sequenced from one direction.

Insertions, deletions, and indels

The detection of insertion, deletions, and indels, where a deletion and an insertion occur at the same position, is more challenging than calling point mutations. The detectable length of these variations strongly depends on the mapping algorithm used because the penalty scores for mismatches, insertions, and deletions vary between them. Some software even made these parameters variable, so the user can influence the mapping. Small events of up to about 25 bp are able to be called on a sequence-dependent manner; larger variants need more complex approaches based on split read alignment or insert size analysis [8]. In a diagnostic setting, we recommend more standardized methods like multiplex ligation-dependent probe amplification (MLPA) or array-comparative genomic hybridization (aCGH) for detecting these larger variants.

Data analysis

Regarding NGS, there is a variety of different enrichment methods, sequencing devices, or technologies as well as a multitude of analyzing software products available. Most applications are compatible with each other, and it is up to the operator to choose them reasonably.

Amplicon-based sequencing (TSCA)

First, a multiplexing approach with the TruSeq Custom Amplicon (TSCA) Kit (Illumina, San Diego, CA, USA) was used to sequence a gene panel of 17 genes known to cause early infantile epileptic encephalopathy (EIEE), combined with sequencing on the Illumina MiSeq.

The "TruSeq Amplicon" approach uses two independent left and right flanking oligonucleotides, which are hybridized to a genomic DNA template enabling polymerase extension and ligation. The flanking oligonucleotides contain universal sequences for step-out PCR and incorporation of universal barcoded Illumina adapters [9]. A total of 370 oligonucleotide pairs for 188 coding exons and bordering intronic regions, totaling approximately 50 kb of cumulative sequence, was easily designed with the Design Studio Web-based tool (Illumina, San Diego, CA, USA). Amplicons of 250 bp length were amplified in a single reaction, and library preparation was finished within 2 days. Five of the 188 total exons failed the probe design (2.6%). Redesign of the failed exons by slightly modifying the chromosomal region did not improve the results. The capture is very scalable, as all steps can be performed in a 96-well PCR plate and can be semiautomated by liquid handling.

A panel of samples with known point mutations, small deletions, and duplications were sequenced in a MiSeq run that produced 5.5 Gb of sequence. Quality control (QC) of sequence data obtained from NGS technologies is extremely important for meaningful downstream analysis. The FASTQC Toolkit [10] was used for quality check and filtering of high-quality data. The mean sequencing quality score (Q-score) was over 30 in 88.2% of the data. The Q-score is a prediction of the probability of an incorrect base call. A higher quality score implies that a base call is more reliable and less likely to be incorrect [11]. For base calls with a Q-score of 30, one base call in 1000 is predicted to be incorrect. We observed a marked drop-off quality (Q<20) in the last 50 to 80 bases of the read in some amplicons within each dataset. Another common artifact of the TruSeq Amplicon technology is the miscalling of the first two bases of the read, thus, increasing the number of false positives. Different end trimming of the reads was applied on each data set. In a diagnostic setting, trimming of low-quality bases at both ends of the reads (mean Q<20) should be performed to achieve analysis of high-quality data. Our analysis confirmed the variant in all but four indels that could not be detected with the standard parameters.

Mapping with the BWA software [12] showed very equal distribution of mapped reads between sample sets. Mean coverage was >500-fold. Overall, 95% of the target sequence was covered at more than 40×. Seven of 188 exons (3.7%) had a coverage between 20× to 40×, and five exons (2.6%) failed amplification in most of the cases. These were very reproducible between samples. The minimal coverage for diagnostic criteria was 40× and Q>20. All fragments which did not achieve these criteria were inspected manually, and an individual decision was taken for subsequent analysis with Sanger sequencing. As base Q-score is often correlated with the complexity of the DNA sequence, manual inspection of the alignment with a visualization tool helps to estimate whether a called variant is a true- or a false- positive call. We observed a high rate of false-positive variants in the range of Q20 to Q30. The major source of false-positive variants arises through either miss-incorporation of bases during PCR amplification. When using a PCR-amplicon approach, all reads align with the same start location based on primer

design and, therefore, extraction of PCR duplicate reads cannot be performed. Designing overlapping PCR amplicons and accepting only consensus variant calls between amplified intervals may help to reduce this problem.

Variant calls are first annotated with snpEff [9] and intersected with dbSNP [13], Exome Sequencing Project (ESP) [14], COSMIC data [15], and our in-house mutation database. In addition, we use features from the Alamut-HT software (Interactive Biosoftware, Rouen, France) to help with data interpretation and predict the mutation effects on splice sites and protein function.

Probe ligation-based DNA fragment enrichment (HaloPlex™)

The HaloPlex™ Target Enrichment Kit provided by Agilent Technologies (Agilent Technologies, CA, USA) was combined with sequencing on the Illumina MiSeq and data analysis with SeqNext of JSI medical systems (JSI medical systems, Kippenheim, Germany).

The HaloPlex™ Target Enrichment System is based on probe hybridization and circularization of enzymatically digested DNA. In contrast to other hybridization-based methods, each probe is an oligonucleotide designed to hybridize to both ends of a targeted DNA restriction fragment, so that the fragments are able to form circular DNA molecules. A PCR step enables the enrichment of perfectly ligated and barcoded fragments plus an incorporation of standard Illumina paired-end sequencing motifs. The complete workflow to capture a target region from 1 to 500 kb takes about 1 day and capturing a target region from 500 kb to 5 Mb about 1.5 days. This procedure results in millions of fragments that are all constructed in the following order: Illumina sequencing motif – barcode sequence – universal primer sequence – region of interest – universal primer sequence – Illumina sequencing motif [16].

With this setup, it is achievable to produce about 4 Gb of sequencing data with a 2×150 paired end run on the MiSeq. On average, there are 80% of the data above a quality value of Q30, which demonstrates a good quality compared to competing devices. With the MiSeq Reporter, it is possible to generate demultiplexed raw data files (fastq format), which can be loaded directly into the SeqNext software. Alternatively, the data can be aligned to the human genome and, afterwards, imported into the SeqNext Software.

Before the data is loaded into the program, the operator is able to make some preliminary settings. For example, poor quality data can be excluded from the analysis by specifying a quality score threshold. This depends on the quality score graph of each MiSeq run. The adapters can be trimmed by entering the universal primer sequence in the corresponding field; bases between the universal primer sequences are retained for further analysis. Additional bases at the fragment ends that do not match the reference can also be removed automatically by a user input.

Main preferences, which affect mutation calling and the sorting of variants to distinct or not distinct ("other") variations, can be done at the beginning of the setting tab. The user may enter an absolute coverage value if this coverage level has to be reached in both sequencing directions together or separately. Positions with an absolute coverage below the entered value are not called. Furthermore, it is possible to specify a ratio read direction. Afterwards, the user has to select a minimal coverage line. Positions with coverage below the entered value will get a warning for this region. Variations that pass the main filter settings will be classified into distinct or not distinct variations depending on variant frequency and user-defined thresholds. This enables the user to exclude mosaicism. Finally, the user can specify in which cases a variation in a homopolymeric region should be listed, in the distinct tab or in the homopolymer tab.

In this setting, we used an absolute coverage of five, a minimum coverage line of 20, and a minimal variant frequency of 15%. The minimum base quality threshold was set at 20, yet an average base quality of the individual variant could not be displayed because this option was not available in the software. The SeqNext software annotates the detected variants with HGVS Nomenclature, even historical nomenclature, dbSNP rs-identifiers, and a custom-made database (editable and correlated with in-house Sanger data).

The resulting variants were filtered and screened against previously detected variants, dbSNP, 1000 Genomes [17], and HGMD® professional (Biobase, Wolfenbuettel, Germany).

Enrichment by oligonucleotide-based sequencing (TSCE)

For the (TSCE) library preparation, one standard workflow was established and validated. Data analysis is performed with the CLC Genomics Workbench (CLCbio, Aarhus, Denmark) and custom-developed Perl scripts (Figure 3).

Using the Illumina Design Studio tool, a total of 5860 probes were designed for a total of 5365 target regions. Success rate of the probes was estimated at ≥95%. The library preparation followed the Nextera® Enrichment

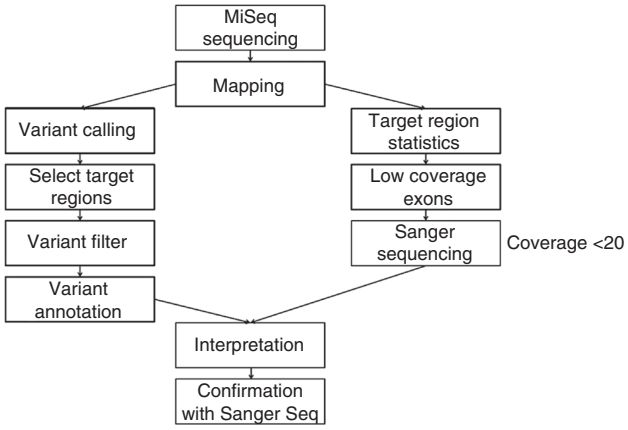

Figure 3 Analysis pipeline, schematic overview.
Analysis steps are performed with the CLC Genomics Workbench and custom developed Perl scripts. Exons with one or more low coverage bases (<20) are reanalyzed by Sanger sequencing.

Sample Preparation Guide and was finished within 2.5 days. All steps are performed in a 96-well plate.

The mapping was performed with the global alignment algorithm. Two to three percent of the reads could not be mapped to the human genomic reference assembly (hg19), while 6–11% of the mapped reads were so-called broken reads and, thus, ignored for variant calling. The coverage was ≥20× in 93.52–96.73% of the regions. Of the exons, 105 of 4018 (2.6%) failed over all samples. The probes fishing these regions need to be redesigned. We defined all bases with a coverage level of 20× or more suitable for variant calling. In conclusion, we determined that an average coverage level of 80–120× is optimal for variant calling. While an increase in coverage to ~200× does not yield significant improvements in the number of bases covered at 20× or more, a drop in average coverage to ~40× results in significantly less bases covered at 20×.

Validation experiments with a set of samples with known mutations were used to adjust the final mapping parameters.

For variant detection, the probabilistic variant caller implemented in the CLCbio Genomics Workbench is used. The caller works with a Bayesian model and a maximum likelihood algorithm to calculate the probabilities of candidate variants. All potential mutations and variants of unknown significance are confirmed by Sanger sequencing. All variants above a Q-Value of 27 have been confirmed by Sanger sequencing, while no variant with a Q-Value below 24 could be confirmed. Variants with a Q-Value between 25 and 27 are in a twilight zone. In these cases, the quality of the reference base and the surrounding bases and the region characteristics (e.g., homopolymer region) are decisive to estimate whether it is a true variant or a false-positive call (Figure 4). The variant call is annotated with GTF-Files for exon numbering, coding region change, and gene information. The Genome Trax™ module of Biobase can be used in the CLCbio software for annotating all variants with additional information like HGMD®, COSMIC database [15], dbNSFP [18], PGX, GWAS data, Online Mendelian Inheritance in Man (OMIM®, Baltimore, MD, USA) information, and experimentally verified transcription factor-binding sites (TFBS). On the last step, the information from dbSNP common flagged and the entire dbSNP is added. This allows a very specific filtering for any individual indication. For instance, in the case of dominant disorders, the dbSNP common will be filtered away, while for recessive disorders, the filtering of the known SNPs has to be adjusted to the incidence rate of the analyzed disease. This list of variants will be inspected for the interpretation of the result. The procedure is as follows: if a variant is found, which is already reported in literature and even in the HGMD® database or any other disease-related database (e.g., Leiden Open Variation Database (LOVD [19]), then, the variant is confirmed with Sanger sequencing. If a variant with unknown significance is found, in silico prediction tools like Pholyphen2 [20], MutationTaster [21], SIFT [22], etc., will be used to estimate the pathogenicity, and all putative disease-causing mutations are confirmed by Sanger sequencing. If no

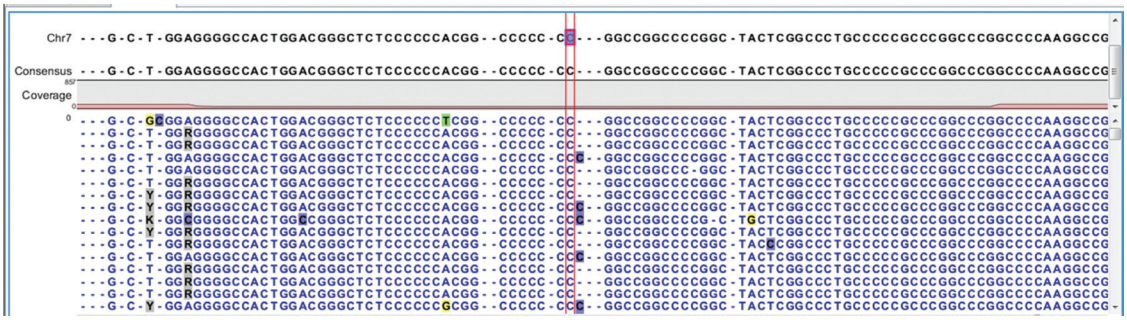

Figure 4 True insertion of a Cytosin in a homopolymer region.

causative variant is found, all regions, which are below 20× coverage, need to be sequenced by Sanger.

Whole-exome-based sequencing (WES)

For whole-exome sequencing data analysis, several mapping, variant calling, and annotation workflows are applied. The filtering strategy mainly consists of two stages, a variant-based and a gene annotation-based filtering. The first filtering step includes a genotype and a population frequency filter. The genotype filter removes all variants that are not in accordance with the clinical history and the pedigree information. The population frequency filter then removes all frequent variants based on the global allele frequency (GAF) score from the 1000 Genomes Project and an in-house allele frequency database. In this second filtering step, all remaining variants are linked to expert-curated and literature data mining-based disease annotations (Figure 5). With this strategy and MESH disease ontologies, it is possible to link causative variants to the clinical diagnoses. Finally, the data has to be visualized to be interpreted, validated, and reported to the requesting physician. Important parameters to follow-up on a variant are covered over this particular variant, genotype quality (GQ), protein effect, the gene and exon(s) affected, or the publications supporting the gene-disease link and segregation analysis within the family.

Taken together, pedigree analyses in combination with clinical information, background population frequencies, and disease annotations provide a powerful information basis to find the needle in the haystack.

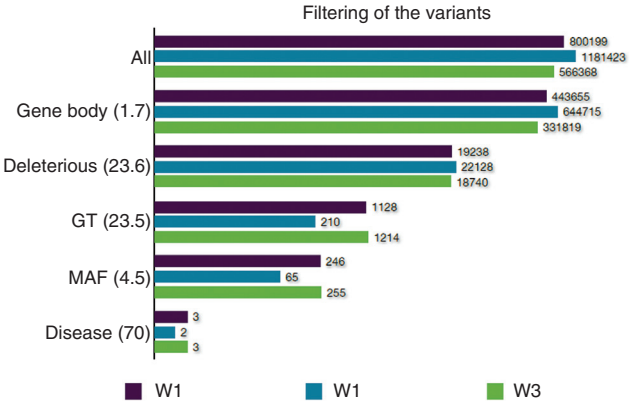

Figure 5 Filtering of variants in three different datasets (log scale). For each filtering step the number of remaining variants and the filter factor is given. The filters are (i) gene body (variants that overlap with a transcript), (ii) deleterious (variants that alter the protein sequence or hit a canonical splice-site), (iii) GT (genotype filter derived from the medical report), (iv) MAF (1000 genomes background filter), and (v) disease (filter for the primary disease and MESH parents from the medical report).

Applications

The use of NGS technologies has been mainly limited to research facilities, but new sequencing devices with a lower throughput and a high quality of sequencing data enables now the application of these technologies in genetic diagnostics. By using NGS, it is possible to sequence several genes simultaneously, in a cost-effective and time-saving manner. Especially multigene panel diagnostic is a very powerful tool to identify the causative mutation for heritable diseases that are genetically heterogeneous and can only be assigned to a group of diseases because of their clinical variability, for example, muscular disorders or neuropathies. A variety of enrichment methods is available on the market to design custom gene panels. Each laboratory should decide depending not only on the disease/gene panel characteristics but also on the laboratory facilities, which is the best approach. In this section, we will discuss the experiences with the amplicon-based TruSeq and HaloPlex systems, the custom enrichment-based approach for larger gene panels (TruSeq custom enrichment, TSCE), and whole-exome sequencing.

Amplicon based/HaloPlex™

TruSeq Amplicon and HaloPlex™ products have several advantages in sample preparation compared to other enrichment methods. Both methods combine the speed of PCR with the sensitivity of hybridization providing a robust solution for targeting smaller capture regions in <2 days. Another positive aspect is that standard laboratory devices and no additional laboratory equipment are needed. Furthermore, the operator is able to follow always the same laboratory protocol, so it is possible to combine different gene panels in one enrichment step. Robotics is not necessary to complete the sample preparation, but protocols can be easily automated. Our experiences showed that automation of the protocols not only increases reproducibility between samples but also enables higher throughput, and it is less time-consuming and avoids sample mix-up [2].

We recommend using these approaches for small gene panels. Our design for the TSCA kit contains 50 kb of cumulative sequence, but new versions of the TSCA kit can cover up to 650 kb with 1536 amplicons per reaction.

However, this is a multiplex approach, and raising the number of oligonucleotid primer pairs might decrease the overall efficiency of PCR amplification. The HaloPlex™ has two different protocols for 1 to 500 kb and 500 kb to 5 Mb. So far, we tested different panel sizes with 170-, 217-, and 615-kb target region.

Both kits, TruSeq Amplicon and HaloPlex™, support multiplexing of up to 96 different samples. In combination with a MiSeq sequencer, we can parallel sequence several patients in a 2×150-bp run in a single standard flow cell (28 tiles). Recently, Illumina has launched the Nano (eight tiles) and Micro (two tiles) flow cells, which differ in output and number of tiles imaged. This offers the user more flexibility when planning the number of samples to be sequenced in a run. The number of sequencing cycles when using an amplicon approach will depend on the amplicon length. In contrast to shot-gut library sequencing, which accepts different sequencing read lengths, the number of cycles when sequencing amplicon products has to be maintained for the same gene panel unless the length of the amplicon is changed in a new amplicon design. Using HaloPlex™, it is possible to decide during the order process between a read length of 100, 150, and 250 bps.

Coverage variability in PCR enrichment approaches is often observed in regions with an elevated number of SNPs, deletions, and high GC content due to bias on failed oligonucleotide design or PCR amplification. A disadvantage of the amplicon-based technology is the difficulty in improving enrichment of low-coverage regions because primer design and PCR amplification is strongly limited by the complexity of the DNA sequence.

One last advantage of the HaloPlex™-based enrichment approach has to be mentioned at this point. With HaloPlex™, each target is covered by multiple amplicons. If an unknown variation occurs in a restriction site, it may affect only a few fragments and not every fragment in this region. This minimizes allelic dropouts, and the variation can be called correctly due to the filter setting of the analyzing software.

Overall, the TruSeq and HaloPlex™ PCR-based enrichment methods are mainly recommended for diagnostic screening of small gene panels. Small gene panels can be designed for diseases with moderate heterogeneity or for phenotypes that can be well characterized within a subpanel of genes in high heterogeneous diseases. Maintaining a small number of genes in the design facilitates data analysis in terms of speed and complexity, as well as the interpretation of the variant calls. Further analysis of low-coverage/low-quality regions with a second method (Sanger sequencing) in addition to the number of variants to be confirmed will be moderate, and accurate analysis of the complete region can be implemented within a few days.

Enrichment based

The most widespread application of target enrichment by oligonucleotide hybridization followed by NGS is targeted diagnostics using gene panels. Using this approach, all genes with known or implicated contribution to a specific indication are simultaneously enriched and sequenced. In comparison to Sanger sequencing, a lot of cost and labor can be saved, yet, the same diagnostic standard may be retained by choosing appropriate quality parameters. Target sequences that fail to reach these quality thresholds have to be resequenced by an alternative method [2, 23].

In comparison to amplicon-based enrichment methods, hybridization-based methods employ multiple oligonucleotide probes, which are complementary to the genomic target regions to capture and enrich all regions of interest. Prior to enrichment, the genome is randomly fragmented, either enzymatically or by sonication. As most of the target regions are covered by more than one capture probe, this enables an even coverage, while minimizing PCR artifacts and potentially compensating for single enrichment probe failures. Enrichment efficiency is influenced by the underlying sequence properties like GC and repeat content [24, 25]. In contrast to amplicon-based approaches, the enrichment is quantitative, resulting in increased off-target effects. Employing this approach, we generate 60–75% reads that are on target.

With this method, modest size region starting from approximately 500 kb to 25 Mb of sequence can be enriched (TSCE), rendering the method most useful in the diagnostics of very heterogeneous disorders with a variety of possible candidate genes, for example, arrythmogenic cardiac disorders or hearing impairments. Using combinations of different MIDs, multiplexing for up to 96 samples/run is possible.

Coverage variability in enrichment-based approaches is observed in regions with a high GC content or in a homologous region due to the probe-binding affinities or of the probes-nonspecific binding, respectively. Thus, it is difficult to increase enrichment in regions of low coverage, which is a definite disadvantage of the TSCE technology or any other enrichment-based approach. The advantage is that there are no taq sequences to be trimmed, leading to fewer discrepancies at the mapping and variant calling.

In conclusion, the TruSeq Custom Enrichment method is mainly recommended for diagnostic screening of large gene sets. Large gene panels can be designed for diseases with high heterogeneity or for phenotypes difficult to assign clinically. Especially for these large gene panels, time and cost can be saved in comparison to Sanger sequencing. Further analysis of low-coverage regions with a second method (Sanger sequencing) in addition to the number of potential disease-causing variants to be confirmed will be moderate.

Whole exome based

For whole-exome enrichment, an in-solution hybridization-based technology is used. The target region can vary between 40 and up to 70 Mb depending on exons, flanking regions, and UTRs to be included. With every whole-exome enrichment, a small percentage (2–3%) of the target region is not covered sufficiently for reliable SNV calling due to several reasons (regions of high homology, GC-rich regions, etc.). Usually, a mean coverage of 150–200× is recommended on the Illumina HiSeq. Increasing the number of reads per exome does not necessarily help to target regions with low coverage as this is due to reduced efficiency in the enrichment and not in the sequencing process.

Diagnostic reporting

Diagnostically reporting of the NGS-based result has to take additional criteria into account in comparison to the Sanger-based reports. Technical details such as the target region (exons, flanking regions, UTRs), the enrichment assay, the sequencing platform, the analysis software, and quality settings should be specified in the method section. Furthermore, the target regions achieving the criteria have to be declared. Regions not passing the criteria are analyzed by another method if necessary. Exon deletion and duplication has to be excluded as well as intronic regulatory SNPs. It has to be mentioned that the effect of unknown synonymous SNPs with regard to exonic splicing enhancer (ESE) motifs is uncertain and not predictable.

There is strong preference to only report known or likely pathogenic variants as no universally accepted guidelines exist for reporting of all detected variants of unknown significance and incidental findings. By gene panel diagnostic, the amount of unknown variants and incidental findings is reduced in comparison to whole-exome sequencing. Reporting is an ongoing discussion, and the reports should follow the general international standards of the UK Clinical Molecular Genetics Society (CMGS) and the Swiss Society of Medical Genetics (SSMG).

Outlook – regulatory SNPs and CNVs

Regulatory SNPs and exonic splicing enhancer (ESEs)

There are many regulatory elements in the genome, which are difficult to predict. Thus, they have not been sequenced in a Sanger testing. But with the coming of the NGS and different enrichment methods, it is possible to sequence larger UTR and intronic regions. There are a few algorithms to predict mutations in splice sites and regulatory elements like Human Splicing Finder, Transfac® (Biobase), and MatInspector (Genomatix). The additional analysis of these regions causes a huge extension of the data analysis time and an expertise to judge the result of these programs.

Copy number variations (CNVs)

In addition to the identification of point mutations and small insertion or deletion events in patients' DNA samples, NGS allows also to detect gene copy number variants (CNVs). The resolution for the detection of CNVs ranges from a single base pair up to nonoverlapping windows of several hundred base pairs. Challenging due to the experimental variability, a more coarsened resolution will result in more robust candidate regions, which are less sensitive against outliers and, moreover, easier to interpret. The detection windows may be evenly distributed across the genome with a fixed window size, or, as array comparative genomic hybridization (aCGH) profiling is currently the gold standard for genetic diagnosis of CNVs [26], the partitioning of the genome may be based on the chromosomal coordinates of the microarray probes. This approach additionally allows direct comparison of both methods [27]. In the case of targeted sequencing (e.g., exomes, gene panels), the chromosomal coordinates of either the region of interest or the baits used for enrichment are preferential limiters of detection windows. Figure 6 shows the principle of CNV detection using a read

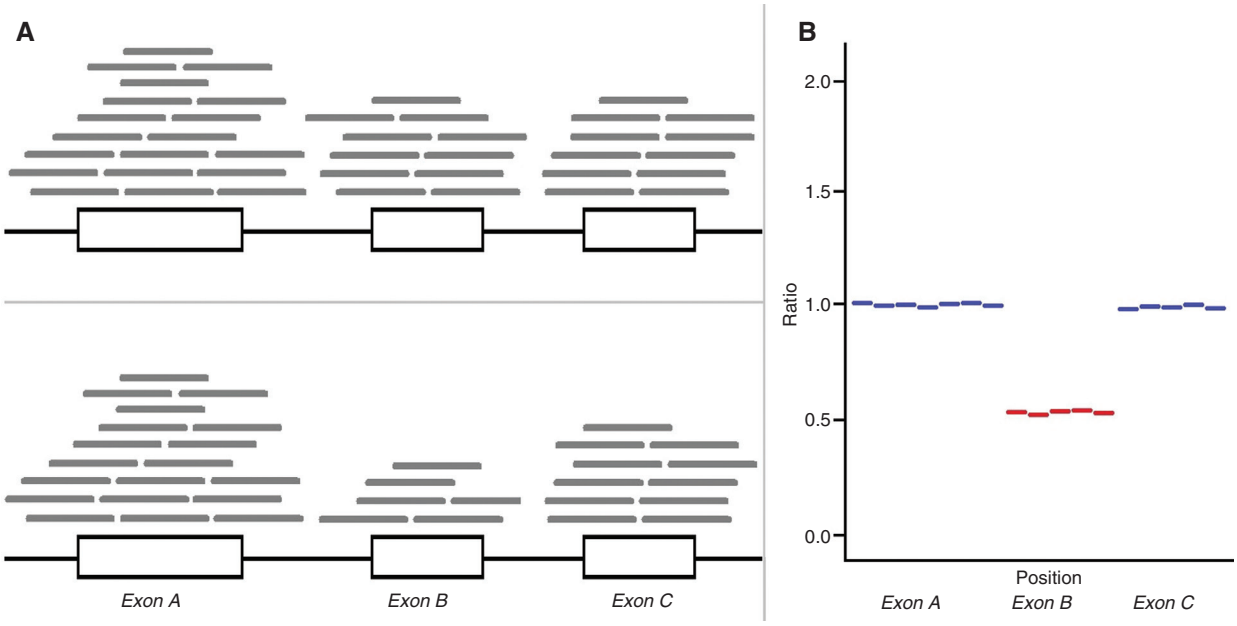

Figure 6 Example of a CNA detected on NGS data in a tumor/normal sample pair.
(A) Visualization of sequence reads mapped to exon sites and resulting read depth. Upper panel indicates normal sample, lower panel indicates tumor sample with copy number loss of exon B by ~50%. Gray horizontal bars, sequence reads; black horizontal line, intronic/intergenic sites; black boxes, exon regions. (B) Example of de novo detection of coverage ratios on the same exons. Ratio of ~0.5 indicates 1 somatic copy loss of the corresponding region. Blue horizontal bars, equivalent coverage in both samples; red horizontal bars, somatic read depth loss; x-axis, genomic position; y-axis, coverage ratios between tumor and normal sample.

depth-based approach comparing a tumor sample and the corresponding germline sample to identify regions of altered gene copies. In a rather homogenous cell population, possible somatic gene copy changes are two-copy loss, one-copy loss, one-copy gain, and two-copy gain, resulting in hypothetical coverage ratios of 0.0, 0.5, 1.5, and 2.0.

Owing to the variability of the enrichment efficiency, the detection of CNVs using targeted sequencing, as well as calling somatic copy number alterations (CNAs), implicitly requires a germline control of the same patient for comparison. The distribution of the detection windows must be equal in both samples to match the read depth of the test and the control sample and to compute a ratio on the coverage. Simple fold change approaches are the easiest way to detect CNVs, but are neither capable to handle regions of zero coverage nor to detect mono-allelic deletions. This problem may be overcome using mathematical modeling approaches, e.g., a linear regression model [28]. Further detection strategies are based on (i) the size of the chromosomal region that is spanned by paired-end reads, (ii) separate mapping of both ends of a read to detect insertions/deletions resulting in CNVs, and (iii) de novo assembly of reads to small pieces of contigs [29].

To further enhance the reliability of in silico determined CNVs using detection tools based on NGS, putative CNVs may be correlated to well-established routine methods like fluorescence in situ hybridization (FISH) or multiplex ligation-dependent probe amplification (MLPA). Anyway, it was published recently that calling CNVs using Illumina NGS is comparable in performance to aCGH [27], supporting its possible supersession in the future as calling mutations on the sequence level (point mutations, insertions, deletions, indels) and on a dosage level (CNVs) at once will notably reduce time effort and costs in clinical diagnostics.

Conflict of interest statement

Authors' conflict of interest disclosure: The authors stated that there are no conflicts of interest regarding the publication of this article.
Research funding: P.A.G. was supported by Deutsche Krebshilfe grant (109031).
Employment or leadership: None declared.
Honorarium: P.A.G. recieved honorarium from Illumina, Inc.

Received June 17, 2013; accepted July 22, 2013; previously published online August 29, 2013

References

1. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012;30:434–9.
2. Vogl I, Eck Sebastian H, Benet-Pagès A, Greif Philipp A, Hirv K, Kotschote S, et al. Diagnostic applications of next generation sequencing: working towards quality standards. J Lab Med 2012;36:227–39.
3. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–7.
4. Smith TF, Waterman MS, Fitch WM. Comparative biosequence metrics. J Mol Evol 1981;18:38–46.
5. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53.
6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9.
7. Dolled-Filhart MP, Lee M Jr., Ou-Yang CW, Haraksingh RR, Lin JC. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. Sci World J 2013;2013:730210.
8. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet 2011;12:363–76.
9. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80–92.
10. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS One 2012;7:e30619.
11. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 2008;9:128.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25: 1754–60.
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
14. Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, Nalls MA, et al. Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am J Hum Genet 2012;91:794–808.
15. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91:355–8.
16. Akhras MS, Unemo M, Thiyagarajan S, Nyren P, Davis RW, Fire AZ, et al. Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications. PloS one 2007;2:e915.
17. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
18. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
19. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 2011;32:557–63.
20. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
21. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;7:575–6.
22. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
23. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 2012;30:1033–6.
24. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 2011;29:908–14.
25. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:182–9.
26. Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, Veltman JA. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res 2007;14:1–11.
27. Hayes JL, Tzika A, Thygesen H, Berri S, Wood HM, Hewitt S, et al. Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation. Genomics 2013. Available online 15 April 2013.
28. Rigaill GJ, Cadot S, Kluin RJ, Xue Z, Bernards R, Majewski IJ, et al. A regression model for estimating DNA copy number applied to capture sequencing data. Bioinformatics 2012;28:2357–65.
29. Xi R, Lee S, Park PJ. A survey of copy-number variation detection tools based on high-throughput sequencing data. Curr Protoc Hum Genet 2012;Chapter 7:Unit7 19.