J Lab Med 2012;36(4):227–239 © 2012 by Walter de Gruyter • Berlin • Boston. DOI 10.1515/labmed-2011-0032

Molekulargenetische und zytogenetische Diagnostik/Molecular-Genetic and Cytogenetic Diagnostics

Redaktion: H.-G. Klein

# Diagnostic applications of next generation sequencing: working towards quality standards

## Diagnostische Anwendung von Next Generation Sequencing: Auf dem Weg zu Qualitätsstandards

**Ina Vogl[1,a,*], Sebastian H. Eck[1,a], Anna Benet-Pagès[2], Philipp A. Greif[3,4], Kaimo Hirv[1], Stefan Kotschote[5], Marius Kuhn[6], Andrea Gehring[7], Carsten Bergmann[8], Hanno Jörn Bolz[8], Manfred Stuhrmann[9], Saskia Biskup[10], Klaus H. Metzeler[4] and Hanns-Georg Klein[1,5]**

[1]Center for Human Genetics and Laboratory Medicine Dr. Klein, Dr. Rost and Colleagues, Martinsried, Germany
[2]MGZ München, Medizinisch Genetisches Zentrum München, Munich, Germany
[3]Clinical Cooperative Group 'Leukemia', Helmholtz Zentrum München, German Research Center for Environmental Health, Munich, Germany
[4]Laboratory of Leukemia Diagnostics, Department of Medicine III, Universität München, Munich, Germany
[5]IMGM Laboratories GmbH, Martinsried, Germany
[6]Genetikum, Ulm, Germany
[7]Institut für Humangenetik, Würzburg, Germany
[8]Bioscientia Zentrum für Humangenetik, Ingelheim, Germany
[9]Institute of Human Genetics, Hannover Medical School, Hannover, Germany
[10]CeGaT GmbH, Tuebingen, Germany

### 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. All major NGS technology companies providing commercially available instruments (Roche 454, Illumina, Life Technologies) have recently marketed bench top sequencing instruments with lower throughput and shorter run times, thereby broadening the applications of NGS and opening the technology to the potential use for clinical diagnostics. 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. To facilitate the implementation of NGS as a routine method in molecular diagnostics, consistent quality standards need to be developed. Here the authors give an overview of the current standards in protocols and workflows and discuss possible approaches to define quality criteria for NGS in molecular genetic diagnostics.

**Keywords:** bioinformatics; genetic variation; Illumina; library preparation; Life Technologies; molecular genetics; molecular genetic diagnostics; next generation sequencing (NGS); rare diseases; Roche.

### 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. Alle größeren NGS-Technologieunternehmen, die bisher kommerziell erhältliche Geräte zu Verfügung stellten (Roche 454, Illumina, Life Technologies), haben vor kurzem auch "benchtop"-Geräte mit geringerem Durchsatz und kürzeren Laufzeiten auf den Markt gebracht, wodurch die Anwendungsgebiete erweitert und der Technologie der mögliche Einsatz in der klinischen Diagnostik eröffnet wurde. 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. Um die Einführung von NGS in die Routinediagnostik zu erleichtern, müssen nachhaltige Qualitätsstandards entwickelt werden. Im Folgenden geben die Autoren einen Überblick über die gegenwärtigen Standards in den Protokollen und Arbeitsabläufen

[a]Ina Vogl and Sebastian Eck contributed equally.

*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

und diskutieren mögliche Ansätze Qualitätskriterien für NGS in der molekulargenetischen Diagnostik zu definieren.

**Schlüsselwörter:** Bioinformatik; genetische Variation; library Herstellung; Illumina; Life Technologies; Molekulargenetik; molekulargenetische Diagnostik; next generation sequencing (NGS); Roche; seltene Erkrankungen.

## Introduction

During the past 6 years, the field of DNA sequencing has dramatically changed by the advent of next generation sequencing (NGS) methods, which rely – in contrast to traditional Sanger sequencing – on massively parallel sequencing of short DNA fragments [1–3]. Reaching an unprecedented throughput, the cost of DNA sequencing decreased by several orders of magnitude. These sequencing techniques offer unique opportunities for molecular diagnostics such as enabling the identification of disease causing mutations in rare and common genetic disorders, cancer diagnosis or rapid pathogen screening of microbial communities. Routine employment of NGS in diagnostics is very promising, yet many challenges remain. Researchers are continuously challenged with new procedures for sample preparation and tremendous increases of data volume requiring new analytical tools for data management. To meet the high-quality requirements for routine application and patient care, quality standards for the application of NGS for molecular genetic diagnostics need to be defined.

An NGS workflow typically consists of the following steps: (i) library preparation, i.e., the preparation of the sample for sequencing, which in particular includes DNA quality control and quantification, shearing of the DNA to fragments of the desired length and quality assessment of the final library. (ii) With the cost of whole genome sequencing still being prohibitive for most applications, the second step comprises the enrichment of specific target regions of interest. Multiple approaches based on either hybridization of oligonucleotide probes complementary to the target region or amplicon amplification are available. (iii) The third step consists of the actual sequencing, whereby each instrument has its specific characteristics and caveats. (iv) The fourth step involves data analysis using bioinformatic techniques to extract information from the massive amounts of raw data. Alignment of the raw reads to a reference sequence, identification of regions with insufficient coverage and variant calling fall in this domain. (v) The final step then consists of the validation of the identified variants using traditional techniques.

Because errors can occur at every single step and consistent quality standards are still missing, the definition of such standards should significantly reduce the risk of generating false data and incorrect data interpretation, thus accelerating the adoption and implementation of NGS techniques for molecular diagnostics. Here we discuss platforms, workflows and data analysis parameters for diagnostic applications and highlight areas where quality criteria need to be defined.

## Library preparation

### DNA quantification

For diagnostic analyses in general, but particularly for NGS procedures, DNA samples of high-quality and purity are required. Hence, the isolation system for genomic DNA should always be the same to generate reproducible results. Ideally, the DNA is prepared in the test performing laboratory according to the accredited standard operating procedure. To determine the quantity of DNA in a sample, a photometric or fluorometric measurement is performed.

DNA concentration can be photometrically measured at a wavelength of 260 nm ($A_{260}$), whereas proteins are quantified at 280 nm ($A_{280}$). The quotient of both values ($A_{260}/A_{280}$) should be in the range of 1.8 and 2.0 to guarantee a high purity of DNA. Thus, values lower than 1.8 indicate a high amount of protein. Owing to the absorbance maximum of carbohydrates at 230 nm, high levels of sugar, salts or organic solvents are detected by a ratio $A_{260}/A_{230}$ over 2.0. DNA samples out of the optimal range (below 1.8 and above 2.0) should be purified with specialized kits.

There are multiple benchtop fluorometer systems available for the quantification of DNA, RNA and protein. Two examples of such instruments are the Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA, USA) and the QuantiFluor system (Promega, Madison, WI, USA). Both systems use intercalating dyes that specifically bind to double-stranded DNA. First, a type curve with a known template has to be generated to calculate the regression formula. The fluorescence from each sample is detected and by the regression formula the amount of molecules per μL is calculated. The software calculates the necessary dilution to achieve the desired concentration. The fluorometric measurement has the advantage that it specifically determines double-stranded DNA content, which is critical because most NGS libraries are prepared exclusively from double-stranded DNA.

A similar concentration and purity of all analyzed DNA samples within one performance is important to guarantee a uniform amount of reads per target (coverage) in the subsequent sequencing reaction.

### DNA shearing

Shearing of input DNA to smaller fragments of desired length (typically 300–1000 bp) is a key step in the library preparation workflow. The distribution of fragment length should be as uniform as possible. There are several shearing techniques available. These techniques include shearing by nebulization, sonication and enzymatic reaction. The benefits and caveats as well as the preferred sequencing system for each technique are discussed in the following sections.

### Sonication

**Covaris isothermal sonication** Fragmentation of genomic DNA using the Covaris Adaptive Focused Acoustics (AFA) process is regarded as the gold standard. The Covaris

instruments (Covaris, Woburn, MA, USA) work by sending acoustic energy wave packets from a transducer that converges and focuses to a small localized area (Table 1). At this focal point acoustic energy creates cavitation, thereby producing breaks in double-stranded DNA. Specific sample processing tubes (micro/miniTUBES with AFA fiber) and several operating conditions control fragment length (100 bp–5 kb) and distribution. Maximizing the yield of DNA fragments in the desired size (defined by mean peak base pair size) depends on the adjustment of following treatment attributes:

- The sample volume should be adjusted with Tris-EDTA (TE) buffer to 200 µL for miniTUBES and 130 µL for microTUBES, but with an improved protocol volume can be reduced up to 50 µL without any effect on shearing profiles. However, with lower volumes an air space may form in the sample fluid; thus, partitioning the sample which sometimes may result in a broad peak.
- As the DNA fragmenting process is rate limited, fragment size generation is affected by treatment duration, which usually varies from 15 to 900 s depending on the instrument, and acoustic parameters (duty cycle/duty factor, intensity/peak incident power, cycles per burst). The DNA amount can vary from 100 ng to 10 µg and minor adjustments must be set up depending on the amount and type of starting material, concentration, and/or viscosity.
- The bath water is employed to couple acoustic energy to the sample vessel. Daily change of pure distilled or DI water, proper water levels, controlled temperature, and degassing are critical for reproducible results. In a non-degassed water bath dissolved oxygen reduces cavitation and disperses energy, reducing shearing efficiency.

The Covaris AFA process has several advantages over other fragmentation methods. First, DNA is sheared in closed independent vessels reducing sample loss and/or cross-contamination. Second, random break of DNA eliminates the risk of G/C bias, common with enzymatic shearing. Third, the system is automatable for higher throughput, up to 96 samples per run. Fourth, improvements to the shearing protocol in combination with removal of small fragments in subsequent bead-based clean-up steps eliminates the need to size select and extract samples from agarose gels, a critical bottleneck in the overall process. Figure 1 gives an overview over the Covaris AFA Instruments.

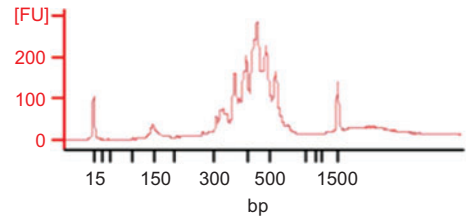


**Figure 1** Example of an amplicon DNA library on a bioanalyzer. The 150-bp peak indicates primer dimer.

**Bioruptor** The bioruptor (Diagenode, Denville, NJ, USA) bath type sonicator generates unfocused acoustic energy, which is dispersed throughout the water bath and absorbed as heat energy. Therefore, ice must be fed to the water bath, and samples allowed cooling down after short bursts of sonication. Alternatively, a water cooling system which allows continuous cooling of the water bath can be connected to the sonicator device. For preparation of sequencing libraries from human samples for subsequent target enrichment it is recommended to use 3–5 µg of double-stranded genomic DNA as starting material. However, smaller amounts of approximately 1 µg might also be sufficient. Fragment size depends on the number of sonication cycles applied [4, 5]. For example, to obtain fragments of human genomic DNA with an average size of 150 bp using the bioruptor sonicator, ultrasound should be applied during three cycles of 15 min each at low power. Recently, a new version of the bioruptor sonicator specifically designed for sequencing library preparation (bioruptor NGS) has been launched. Compared to previous versions, the bioruptor NGS can be operated in a faster and more user friendly manner. Although bath type sonication may result in reduced yield of DNA fragments compared to isothermal sonication, it is a feasible and rather economic approach for sequencing library preparation.

**Nebulization** Nebulization is a simple and cost-effective procedure easy to establish in every laboratory. DNA fragmentation occurs inside a nebulizer, a small plastic device that uses compressed air to atomize liquids, which is connected to a nitrogen or argon air tank, or laboratory compressed air line using appropriate connectors and tubing. The use of a

**Table 1** Overview of Covaris Adaptive Focused Acoustics instruments.

| Product | M-series | S-series | E-series | L-series |
|---|---|---|---|---|
| Key features | Single sample | Single sample | Multisample | Multisample<br>parallel processing |
| Automation compatibility | N/A – personal<br>DNA shearing | Yes – S220R | Yes – E220R | Yes – LE220R |
| Application | DNA shearing | DNA shearing<br>Other | DNA shearing<br>Other | High throughput DNA shearing<br>Other |
| Target peak | 200–3000 bp | 150–5000 bp | 150–5000 bp | 150–5000 bp |
| Sample volume | 50–200 µL | 50 µL–10 mL | 50 µL–10 mL | 50 µL–200 µL/tube |
| Models | M220 | S220™<br>S220X™<br>S220R™ | E220™<br>E220X™<br>E220R™ | LE220™<br>LE220R™<br>L8™ |

regulator that allows accurate control of the pressure from 10 to 60 psi is recommended. Nebulization requires purified DNA (0.5–10 μg) mixed in fragmentation buffer which consists of 80% glycerol and Tris-HCl (pH 8). Nebulization should be performed in an appropriate laminar flow cabinet to prevent contamination of the aerosolized solution and the nebulizer device should be placed on ice during fragmentation. The size of the fragments obtained by nebulization is determined chiefly by the speed at which the DNA solution passes through the atomizer, altering the pressure of the gas blowing through the nebulizer, the viscosity of the solution, and the temperature. Minor changes in these parameters will optimize the size and range of the DNA fragments. Although this is an easy and quick method, it has some disadvantages compared to other DNA shearing techniques. The resulting DNA fragments are usually distributed over a broad range of sizes (200–1000 bp) and size selection is necessary in further steps of the library preparation procedure. In addition, purification of DNA after nebulization is required and DNA recovery is not always optimal if using low amounts of DNA (1 μg or less).

**Enzymatic fragmentation** The Nextera DNA Sample Preparation Kits (Illumina, San Diego, CA, USA) provide a fast and easy workflow, enabling sequencing ready libraries to be generated in 90 min starting with 50 ng genomic DNA or polymerase chain reaction (PCR) amplicons. The Nextera chemistry simultaneously fragments and tags DNA in a single step using a transposon-based proprietary technology. A simple PCR amplification then appends sequencing adapters and sample indices to each fragment. This method generates typical median insert sizes of ~300 bp and is compatible only to Illumina sequencing platforms. Following the addition of two indices to each DNA fragment, up to 96 uniquely indexed samples can be pooled and sequenced together in a single lane on any Illumina sequencer [6].

**Gel and capillary electrophoresis** Most NGS technologies require that the DNA fragment sizes in the input library fall within a certain size range (due to, e.g., limitations imposed by the bridge amplification or emulsion PCR processes). Therefore, correct size distribution of the library needs to be ascertained to ensure satisfactory sequencing results. Traditional agarose or polyacrylamide gel electrophoresis can be used especially when one can expect fragments with defined sizes that produce sharp, easily detectable bands (e.g., amplicon-based resequencing). For DNA fragment libraries with a spectrum of DNA segment lengths, specialized instruments (e.g., Agilent 2100 Bioanalyzer and 2200 TapeStation, Agilent, Santa Clara, CA, USA; Bio-Rad Experion, Bio-Rad, Hercules, CA, USA) provide higher sensitivity and better size resolution. These instruments are based on capillary electrophoresis and DNA detection by fluorescent dyes, integrated in a compact cartridge or microfluidics chip. They not only allow precise library size determination but also detection of small contaminants such as primer dimers. This is important because presence of primer dimers in a sequencing library may cause artifacts that severely affect the output of a sequencing run.

**Quantification** The accurate quantification is an important step during NGS library preparation. In the three systems (Roche 454, Roche, Basel, Switzerland; SOLiD and Ion Torrent PGM, both Life Technologies, Carlsbad, CA, USA) which use emulsion PCR (emPCR) for amplification, it is important to apply the correct amount of library molecules to the beads. To avoid mixed reads, there should be one read per bead. For the Illumina platforms it is important to calculate the accurate molarity for the cost-benefit ratio of the run. This means, on the one hand, if the flow cell is loaded to sparse, sequencing yield will decline. On the other hand, if the flow cell is loaded to dense, signal overlap between adjacent clusters reduces overall run yield and quality. There are several different methods to quantitate the library including fluorometric assays and quantitative PCR (qPCR), which is widely considered the most accurate.

*Fluorometric quantification:* DNA libraries can be measured using standard fluorometric methods as previously discussed.

*qPCR:* Sequencing performance depends on the amount of DNA fragments with correctly ligated adaptors and the efficiency of library amplification. Both factors can be measured using real time qPCR. The advantage of this method is that only fragments with both adaptors will be amplified. Fragments in which the ligation of one or both adaptors failed are not quantified, avoiding an overestimation of fragments. It is recommended that the standard library should have a comparable GC-content to the quantified library, because it has been shown that GC-content can influence amplification efficiency in qPCR. As a standard either a plasmid library or an already successfully amplified library can be employed.

## Automation

NGS techniques allow not only the simultaneous sequencing of all known genes associated with a certain disease but also the simultaneous sequencing of tens of patient samples in one sequencing run. It is therefore of great relevance to guarantee identification of each sample in the downstream analysis process, to have greatest possible reproducibility and to avoid any type of cross-contamination [31].

The bottleneck occurs with the sample preparation before sequencing. As previously discussed in detail, current library preparation methods include DNA purification, multiple quality control steps, adaptor ligation and target enrichment, involving several repetitive purification and pipetting steps. Manual methods are not only time consuming but also error prone with considerable variability. By contrast, automated sample preparation enables a higher throughput and an increased reproducibility by improving pipetting accuracy and avoiding sample mix-ups. Examples for automated library preparation are the Agilent Bravo (Agilent) for high

throughput applications and the AB Library Builder™ System (Life Technologies) for low-to-medium throughput.

### Contamination

One of the key advantages of NGS technology is the clonal sequencing of the target DNA. A prerequisite of the clonal sequencing is the amplification of single molecules, which renders the method potentially susceptible to the contamination with extrinsic DNA. Contaminants may be derived from laboratory surfaces or from pipetting devices. In particular, enriched DNA libraries from earlier experiments may give rise to contamination of subsequent sequencing runs. By using a PCR amplicon strategy for DNA library preparation, huge amounts of PCR products are generated in the first PCR. Amplified DNA fragments are diluted more than a 1000-fold before clonal amplification occurs in the second PCR. If even minimal amounts of contaminants are available in the working area, they can function as a template for the clonal amplification and thus be detected by sequencing.

To avoid false positive sequencing results, similar practices and tools as in set-up of a nested PCR must be applied for NGS. At least three laboratory areas must be defined and physically separated: pre-PCR, "clean" post-PCR and post-PCR. The usage of filtered tips and different sets of pipettes is an undisputable requirement. DNA extraction and the first PCR are performed in pre-PCR area. No PCR products are allowed in this part of the laboratory and the movement of samples, equipment and all types of materials from the post-PCR area into the pre-PCR must be avoided. A clonal amplification, e.g., emulsion PCR is prepared in "clean" post-PCR area. The traffic into this area must be reduced to a minimum. If dilution of the DNA library is needed, it should be diluted prior to introducing it into "clean" post-PCR area. For the preparation of clonal amplification in the "clean" post-PCR, we recommend the use of special working cabinets or laminar airflow hoods, which can regularly be decontaminated with UV light.

The experiences in our laboratory with NGS technology impressively underline the importance of very strict compliance with high quality standards. In spite of careful sample handling a cross-contamination was observed in human leucocyte antigen (HLA) typing with NGS technology. In the search for possible contaminants, we identified the source of false positive sequences. The emPCR was obviously contaminated with PCR products from earlier experiments. To our surprise, this experiment was done already 3 months before the contamination occurred. All runs in between have shown no contamination with additional PCR products. Further quality assurance measures were taken afterwards and resolved the problem with contamination in subsequent sequencing runs.

## Enrichment

Even though the raw cost of sequencing dropped significantly with NGS technologies, the cost of sequencing a complete human genome is still prohibitive for most applications. Particularly for clinical diagnostics, only disease relevant genes and regions of interest are to be sequenced. To select these regions, several different approaches, either amplicon-based or based on oligonucleotide hybridization, are commercially available. The individual systems are briefly highlighted in the following sections. Main performance metrics and quality criteria for sequence enrichment are listed in Table 2 (based on [7]).

### Amplicon based enrichment

**Singleplex PCR** Enrichment with singleplex PCR uses amplicon specific primers fused a universal linker tail in a first PCR. These target specific PCR products are pooled and then amplified in a second PCR with primers, consisting of a linker tail, MID-tag (patient specific) and at the end a sequencing adaptor, thereby extending them with the sequences that are required to initiate sequencing and to distinguish reads from the different patients. Because regular distribution of the single fragment amplification efficiency is the key step for a uniform distribution of coverage and to maximize sample size in a single experiment, design and validation of primer sets should be optimized very carefully for best performance of the PCR amplification. Table 3 summarizes a singleplex PCR enrichment protocol successfully used in our laboratory.

Enrichment with singleplex PCR is a fast and simple approach when small numbers of gene-specific amplicons need to be investigated and is best suited for long read sequencing platforms (e.g., Roche GS-FLX, GS Junior system). Commercially available sample preparation approaches can increase throughput, but are less cost efficient for smaller experiments. Singleplex PCR has the advantage over multiplex PCR that one can control and optimize amplification efficiency for each single amplicon; thus, diminishing coverage bias. However, for a large number of samples the number of amplification reactions may increase dramatically and automation is mandatory. In addition, self-designed PCR assays have the advantage that the same set-up as for Sanger sequencing can be maintained, facilitating confirmation of the detected mutations afterwards [8, 9].

**Multiplicom** The MASTR (Multiplex Amplification of Specific Targets for Resequencing) technology from Multiplicom (Multiplicom N.V., Niel, Belgium) offers a cost-effective DNA target amplification, minimizing hands-on-time. A Multiplexer™ algorithm is used for primer design enabling simultaneous PCR amplification of multiple target sequences, up to 70 amplicons (140 primer pairs), under standard conditions. A simple two-step PCR protocol, without the need of product normalization, allows flexible incorporation of MIDs (molecular barcodes) in each amplified product to unambiguously link each read to the sample it originated from. Some key advantages of the MASTR workflow include the low amount of input DNA (only 20–50 ng) needed per multiplex PCR reaction and standard laboratory equipment (PCR machine and fragment analyzer) is the only requirement. Optional labeling PCR for fragment analysis check on Genescan is also included in the kit. The assay was first designed for use with

**Table 2** General performance measure and suggested quality standards for next generation sequencing [7].

| Performance measure | Suggested quality standards |
|---|---|
| Position and size of region of interest (ROI) | The ROI targeted for sequencing should be defined according to a reference database [e.g., using NCBI RefSeq or consensus coding sequence (CCDS) identifiers]. This is of particular importance if several transcript variants of a gene are known. |
| Fraction of sequencing reads aligned to ROI (on-target percentage, capture efficiency) | Using a set of control samples, a reference range for the percentage of sequence reads aligning to the ROI should be established. Capture efficiency should be assessed for each clinical sample, because variation may point towards poor material quality or technical problems. |
| Average read depth across ROI (coverage) | Using a set of control samples, a reference range for the average read depth typically achieved over the entire ROI should be established. Average read depth should be assessed for each clinical sample. Variation of read depth may point to poor quality of starting material, technical problems with sequence enrichment or sequencing, or may be caused by genomic alterations such as amplifications or deletions. |
| Fraction of ROI covered at sufficient depth and with sufficient quality[a] (distribution of coverage, uniformity) | A set of control samples should be used to define areas within the ROI that are prone to receive below average coverage or yield low quality reads.<br>The overall number of sequencing reads should be adjusted to ensure adequate coverage in such "difficult-to-target" regions. If sufficient coverage of certain areas within the ROI cannot be reliably achieved by sequence enrichment/NGS, alternative strategies need to be established for genotyping these regions.<br>Sequencing results for each sample should be assessed for areas with unexpectedly low coverage, which may point to genomic alterations such as deletions or rearrangements.<br>Diagnostic reports should state whether the entire ROI was sufficiently[a] covered, or whether there were any gaps in coverage. |
| Sensitivity, specificity, accuracy, assessment of allelic bias or dropout | Assessments of sensitivity and specificity are central to the validation of every sequencing-based test. However, some specific considerations apply with regard to sequence enrichment, as it adds opportunity for bias. Sequence variations (both germline variants and somatic mutations) may affect probe or primer binding and reduce capture efficiency. This can lead to reduced or absent representation of one allele (allelic bias or allelic dropout).<br>Sensitivity and specificity of a sequence enrichment/NGS-based test should be established using sets of positive and negative control samples characterized by an established reference method. The positive control set should include mutations at any known mutational hotspots. If type or location of mutations within the ROI are heterogeneous, the positive control set should encompass a representative spectrum of known mutations. The test set should also include any sequence variants which are either relatively common or have special clinical importance.<br>For each clinical sample, known polymorphic sites (e.g., SNPs) within the ROI should be routinely evaluated for evidence of allelic bias. |

[a]"Sufficient" refers to the number and quality of sequencing reads required for confident detection of genetic variants. These parameters closely depend on the algorithms and criteria used for variant calling (see section on single nucleotide variant calling).

the Roche GS-FLX and GS Junior sequencers, but an adapted protocol for "short read sequencing platforms" is now also available.

**Highly multiplexed, amplicon-based systems for target enrichment** Several approaches for target enrichment based on highly multiplexed, sequence-specific generation of amplicons have been developed. Commercially available systems include the Illumina TruSeq Custom Amplicon (TSCA) and Agilent Haloplex systems. These systems can be used to target up to several hundreds of kilobases of sequence, and thus stand in between conventional singleplex or multiplex PCR, and hybridization-based methods (in-solution or array-based hybrid capture) [10]. Primer or probe sequences for these assays are designed by the manufacturers based on the desired target sequences, and reaction conditions are standardized. Thus, these assays require less optimization on the side of the end user than conventional PCR-based approaches, but also offer less possibility for adjustments in the case of difficult-to-target regions. Other advantages in comparison to conventional PCR include the larger size of targetable sequence, and the lower amount of input DNA required due to the higher level of multiplexing. However, multiplexing may also lead to problems with uneven coverage, whereas a carefully optimized PCR setup may provide more balanced coverage over the entire target region [10]. Advantages of highly multiplexed amplicon-based approaches over hybridization-based target enrichment include faster and less laborious library preparation (no DNA shearing, adapter ligation, or overnight hybridization steps required), lower input DNA requirements, and higher capture efficiency especially for small target regions.

The Illumina TSCA system is specifically designed for use with the Illumina MiSeq sequencing platform. It is based on

**Table 3** Example for a singleplex PCR protocol.

| | |
|---|---|
| 1. Singleplex PCR | Primary PCR using primers modified at their 5′ end with a universal linker sequence. A hot-start PCR program, choosing adequate linker tail pairs, and adjusting primer concentration (0.2–1 μM) without affecting amplification efficiency are crucial to avoid primer dimer formation. In addition, a proof-reading polymerase and a maximum of 25 PCR cycles with a touch-down PCR cycling program are recommended. |
| 2. Singleplex PCR evaluation | Agarose gel by electrophoresis. |
| 3. Purification and normalization | SequalPrep Normalization Plate (96-well) Kit (Invitrogen). Automation is recommended to ensure reproducible results. |
| 4. Amplicon pooling | Pooling of all amplicons derived from one individual in a length-weighted equimolar ratio (3 μL for 200bp–250bp products, 3.5 μL for 251bp–300bp, 4 μL for 301bp–350bp, 5.5 μL for 351bp–400bp, 8 μL for 401bp–500bp products, 12 μL for 500–600bp products) improves uniform distribution of coverage. Automation is recommended. |
| 5. Concentration | MinElute PCR purification Kit (Qiagen, Venlo, Netherlands). |
| 6. Secondary PCR | Second PCR using patient specific MID-tagged primers with linker tail and sequencing adaptor. The total number of PCR cycles (primary and secondary PCR) should not exceed 40; thus, the number of secondary PCR cycles will depend on the first PCR. |
| 7. Purification | Magnetic bead PCR purification system should be used. |
| 8. Evaluation | Agilent Bioanalyzer (or similar). Libraries showing primer dimers on the electropherogram might be subjected to a second bead purification. Because longer amplicons amplify less efficiency during emPCR, thus introducing coverage bias, secondary PCR should be repeated for amplicon pools showing irregular amplification patterns (libraries with stronger representation of short amplicons vs. long amplicons). |
| 9. Quantification | PicoGreen. |

hybridization of two oligonucleotide probes to the same strand of unsheared genomic DNA, followed by an extension-ligation reaction and subsequent PCR. Currently, up to 384 amplicons covering up to 96 kb of cumulative target sequence can be amplified in a single reaction tube from 250 ng of input DNA.

The Agilent Haloplex system was developed based on the principle of selector probes [11]. Following digestion of genomic DNA with restriction enzymes, hybridization to probes that are complementary to both ends of individual restriction fragments leads to the formation of circular molecules. These circles are closed by ligation and selectively amplified, using universal priming sequences included in the capture probes. The approach currently can be scaled to target 1–500 kb of sequence from 250 ng of input DNA.

### Hybridization based enrichment

**NimbleGen** The NimbleGen (Roche NimbleGen, Madison, WI, USA) enrichment protocol is optimized for the Roche 454 System. In contrast to other suppliers such as Agilent or Illumina, which use a web design studio, the design of the baits is performed directly by NimbleGen. This enrichment method employs 55–105 mer DNA baits, which are designed overlapping to cover the target region [12]. There are currently two different kits available which offer two different quantities of baits (385 K or 2.1 M), resulting in a target region up to 5 Mb or up to 50 Mb, respectively; 500 ng of genomic DNA is used as input. An on-target ratio of 66% of the reads should be achieved. Reads are defined as "on-target" if they share at least 1 bp true overlap with the targeted region during alignment. Not all reads are expected to map to the target region. Off-target reads can result from unspecific hybridization of the baits to different genomic locations. This is especially problematic, if the underlying genomic context shares great similarity with additional genomic regions (i.e., repeat stretches, pseudogenes or regulatory motifs). These regions may lower the target enrichment efficiency and specificity. Coverage per gene is uniformly distributed (Figure 2). The amount of mappable reads varies with different alignment settings. More stringent mapping parameters result in more reads being sorted out during alignment. In an experiment performed in our laboratories the NimbleGen assay achieved high specificity (Table 4).

**Agilent** The SureSelect Target Enrichment (Agilent) workflow is a solution-based system utilizing 120-mer biotinylated RNA baits to capture regions of interest. Starting from genomic DNA, a shearing step produces small fragments that are then coupled to specific adaptors and indexes or barcodes. The sample is then hybridized with biotinylated RNA baits that are further selected by using magnetic streptavidin beads. The company offers different enrichment kits for whole exome [50 Mb, if desired also V4 including untranslated regions (UTR)]. The different versions are optimized with respect to underrepresented regions in previous versions plus additional balancing of baits. Several other kits such as the X chromosome or the coding kinome are available. A customized design is possible through access of the web interface "eArray".

**GC-content bias** A general problem of hybridization based enrichment systems is the influence of target sequence GC-content. Hybridization efficiency declines with higher levels of GC-content [12]. The first coding exons of genes are generally known to be GC-rich [13]. Therefore, coverage and capture rate of the first coding exons are highly variable (Figure 3).

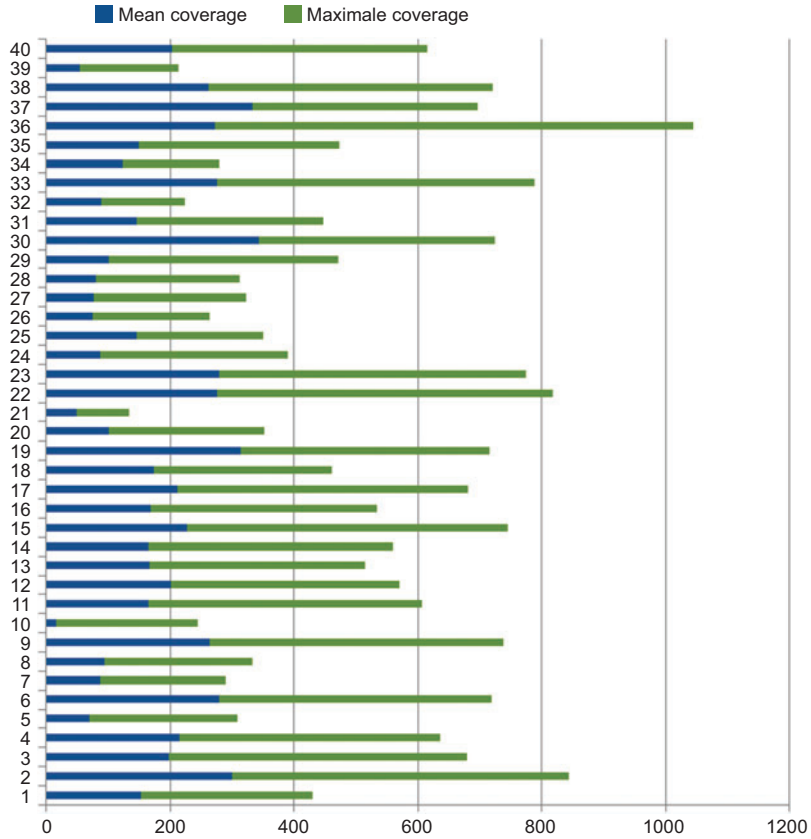


**Figure 2** Coverage distribution of an assay containing 40 genes.
Green bars indicate the maximum coverage, blue bars denote the mean coverage of the gene.

## Quality assessment: instruments

### Roche 454 sequencing

The Roche 454 technology uses a combination of pyrosequencing and emPCR. The capacity of the GS-FLX platform is around 350–700 Mb per 10 h run, enabling the sequencing of amplicons and larger gene panels. By contrast, the GS-FLX is not economical in sequencing whole exomes or complete genomes. Targets captured by the NimbleGen/Roche enrichment system (on array or in solution) are compatible with both 454 FLX and the benchtop version, the 454 GS Junior.

**Table 4** Overview of an assay containing 40 genes enriched with NimbleGen.

| Reference | Total mapped reads | Mapped reads in targeted region | Specificity |
|---|---|---|---|
| chr1 | 62.553 | 48.286 | 0.772 |
| chr2 | 2.417 | 1.212 | 0.501 |
| chr3 | 17.479 | 10.596 | 0.606 |
| chr4 | 25.053 | 19.888 | 0.794 |
| chr6 | 14.865 | 11.389 | 0.766 |
| chr7 | 6.173 | 4.607 | 0.746 |
| chr10 | 7.871 | 4.441 | 0.564 |
| chr11 | 20.404 | 9.856 | 0.483 |
| chr12 | 28.736 | 16.102 | 0.560 |
| chr14 | 14.29 | 8.503 | 0.595 |
| chr15 | 6.813 | 3.988 | 0.585 |
| chr17 | 9.138 | 3.243 | 0.355 |
| chr18 | 14.972 | 10.869 | 0.726 |
| chr19 | 1.993 | 1.521 | 0.763 |
| chr20 | 3.592 | 2.511 | 0.699 |
| chr21 | 4.415 | 1.164 | 0.264 |

During emPCR, the enriched DNA is clonally amplified on specific beads which are then individually placed into millions of reaction wells on the GS-FLX picotiter plate (PTP), each containing sequencing enzymes. Successive flows with one of the four deoxynucleotide triphosphates (dNTPs) result in incorporation by synthesis and consecutive localized luminescence, which is recorded by an integrated CCD camera [1]. Roche recommends an enrichment of the beads in the range of 5%–20% to achieve a successful sequencing run; 90% of the Raw Wells should be KeyPass Wells. At least 50% of the KeyPass Wells should be passed filter Wells.

A major advantage of the 454 technology is the comparably long read length that may reach up to 450 bp (up to 1000 bp with the recently released GS-FLX+ system, respectively). The long reads are ideal for de novo sequencing of small genomes – an application that is primarily beneficial for microbiology. Human genetic diagnostics may take advantage of nucleotide haplotype information over a range of hundreds of base pairs which can eliminate the need of segregation analysis for two recessive mutations. Moreover, long

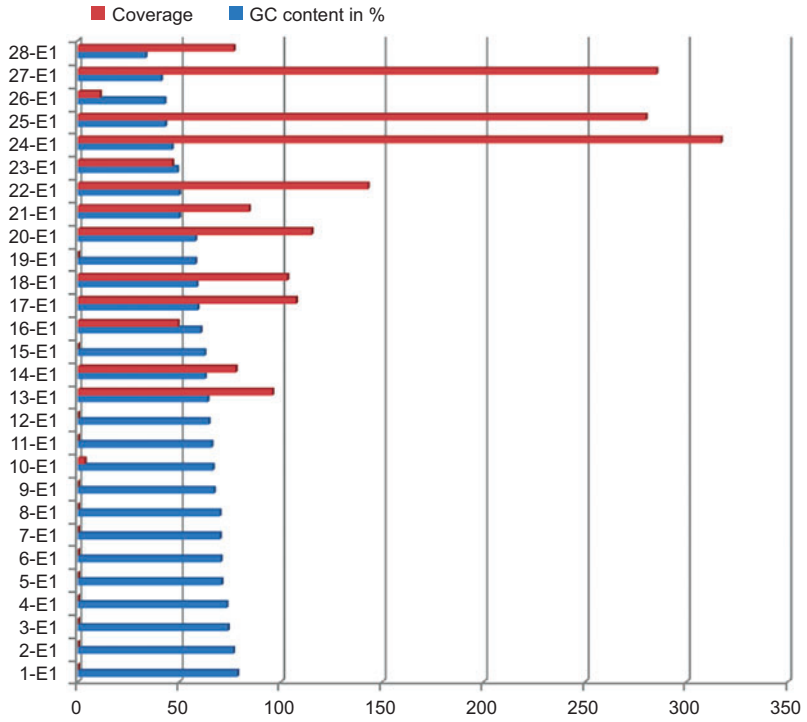


**Figure 3** Coverage in relation to GC-content (in %) of the coding first exons of an assay containing 40 genes.

reads facilitate alignment in "difficult" regions with repetitive sequences. The main issue of this technology includes systematic errors in homopolymeric regions, for more detailed discussion see ref. [14] and section on False positives.

## SOLiD

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is an NGS technology developed by Life Technologies and has been commercially available since 2008. The current version, the SOLiD 5500, is capable of producing up to 180 GB of sequence during a 10-day run. A library of DNA fragments is prepared from the sample to be sequenced, which is then used to generate clonal bead populations. These fragments are bound to magnetic beads such that only a single species of fragment is present at each bead. In a comparable manner to 454 sequencing, emPCR is employed to amplify the fragments. The resulting PCR products attached to the beads are then covalently bound to a glass slide. In the sequencing reaction, a set of fluorescence labeled di-base probes compete for ligation to the sequencing primer. Specificity of the di-base probe is achieved by interrogating every first and second base in each ligation reaction. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length. The color system is redundant as four distinct colors are used to detect the 16 different dinucleotide combinations. This leads to the term "color-space sequencing".

The manufacturer claims that, as every sequence position is read out twice, the SOLiD system produces very high quality data with a per-base accuracy of >99.94% and a consensus accuracy of 99.999% at 15× coverage. However, head-to-head comparisons with other NGS platforms failed to demonstrate superior accuracy for the SOLiD technology [15, 16]. One study evaluated the SOLiD, Illumina and 454 platforms for single nucleotide polymorphism (SNP) genotyping in human samples, using Sanger sequencing and microarrays as the reference. In this study, overall variant call accuracy actually was slightly lower with the SOLiD than with the Illumina platform. Of note, SOLiD seems to have higher sensitivity for variant detection in regions with low coverage. With the SOLiD system, similar to the Illumina platform, sequencing errors predominantly seem to occur towards the 3′ ends of reads [3]. It is important to recognize that the evolution of sequencing instruments and chemistry is extremely rapid, and therefore these comparisons performed 3 years ago may not adequately reflect contemporary system performance. Regarding the situation in early 2012, the read length of the SOLiD system is limited at 75 bp for single-end reads, whereas 150 bp or longer reads can be achieved with the Illumina platform. The shorter read length of the SOLiD platform may represent a disadvantage for certain applications such as amplicon resequencing. One issue specific to the SOLiD platform is that calling multiple adjacent SNPs is computationally more challenging, due to the "color space" representation of sequence data [17].

## Illumina sequencing-by synthesis (SBS)

SBS technology, first introduced in 2006 by Solexa, is now marketed by Illumina [2]. Sequencing takes place on the solid surface of a transparent "flow cell", where clusters of identical

DNA molecules are initially generated by a PCR-like process called bridge amplification. The cyclic sequencing process then involves strand elongation by incorporation of fluorescently labeled nucleotides and optical imaging. Subsequent cleavage of the fluorescent dye together with a "reversible terminator" enables further strand elongation in the following cycle. Currently, the Genome Analyzer (GA) IIx and HiSeq series instruments are capable of producing up to 95 Gb and 600 Gb, respectively, of sequence output per run. Whereas these instruments are primarily used for large-scale research projects including whole exome or whole genome sequencing, the MiSeq benchtop sequencer (introduced in 2011) is marketed for applications such as targeted resequencing of disease gene panels.

Several studies have investigated the technology-specific error profile of SBS sequencing technology. Overall, these studies show that base substitution errors are more frequent than insertion/deletion type errors, and that the error rate increases in later cycles [14, 18, 19]. Errors may preferentially occur in specific sequence contexts, and therefore errors may be strand specific (i.e., occur only in reads mapping to either the forward or the reverse strand of the genome). Although the Illumina sequencing chemistry has undergone improvements over time, knowledge of such technology-specific error profiles is important to identify false positive variant calls caused by sequencing artifacts. In addition, a recent study compared the MiSeq system with two other "benchtop" sequencers (Roche 454 GS Junior and Life Technologies Ion Torrent PGM) and concluded that the MiSeq offered the lowest error rate of the three instruments, with a substitution error rate of 0.1 per 100 bases and <0.001 indel errors per 100 bases for 2×150 bp paired-end reads [14]. Of note, this study looked at de novo sequencing of a bacterial genome, and it is unclear if the same error rates apply to other applications such as targeted resequencing of human genes. A study evaluating the HiSeq and GAIIx platforms using plant and viral genomes showed comparable error rates, with substitution rates of 0.11%–0.16% for the HiSeq and 0.28% for the older GAIIx instrument, and indel rates of $<3\times10^{-5}$ [19]. Illumina has stated some basic quality criteria for a successful MiSeq instrument run, these include, for a 2×150 bp run:

- Phred-scale quality scores of at least 30 for >75% of all bases for a 2×150 bp run.
- Cluster densities of 50–1300 K/mm$^2$, ideally 800 K/mm$^2$.
- Fluorescent signal intensities of >200 units for all four bases throughout the run.
- Sequence yield >1 GB.

These criteria probably represent conservative estimates of instrument performance, and the results observed in practice will depend on the specific assay in use. Therefore, users should carefully monitor run parameters over time to detect suboptimal instrument runs, even if they fall within the above specifications. Additionally, we recommend inclusion of a control library containing a known sequence (e.g., the phiX viral genome) with every instrument run. This allows direct measuring of the sequencing error rate for each run and helps to detect variations in sequencing quality.

## Bioinformatics

### Alignment

A key point in the analysis of NGS data is sequence alignment, where millions of short reads have to be aligned to a reference sequence in a reasonable time. A variety of different alignment algorithms and strategies have been implemented since NGS has been available [20]. Each alignment algorithm has certain advantages and disadvantages and is tailored to a specific instrument for which it was primarily designed. The choice which program performs best is left to individual user's preferences; however, two important aspects need to be considered. The differences and implications when using a local vs. a global alignment algorithm and the concept of mapping uniqueness, particularly how the chosen algorithm handles non-uniquely mapped reads. The handling of these reads can significantly influence variant calling results.

Global alignment attempts to align the read over the full span of bases, whereas local alignment algorithms trim mismatching bases at the ends of reads. In turn, local alignments typically lead to a greater number of aligned reads and thus more produced data, but may also introduce certain mapping biases. By contrast, global alignments maximize alignment quality at the cost of a greater portion of reads that cannot be aligned.

Mapping uniqueness addresses the problem that a given read can have more than one match to the reference sequence. The multiple hits are a result of underlying sequence similarity in the reference genome, for example, in repeat regions or pseudogenes. There are several ways to define mapping uniqueness, one of the most widely adopted being the following. A read is uniquely mapped if its best hit contains less mismatches to the reference sequence than its second best hit. There are currently two alternatives in dealing with such reads: (i) discard all non-uniquely mapped reads or (ii) randomly choose a position to align each read but flag them as non-unique. Flagged reads are then ignored during variant calling and thus do not contribute to variant calls. The second option has the advantage that even though the reads are non-uniquely mapped, they still can provide valuable information in overall coverage and repeat content of the sequenced region.

### Single nucleotide variant calling

Per base coverage is an important criterion to assess the reliability of a variant call at any given genomic position. For use in clinical, a minimum coverage should be defined, which is required to determine the underlying genotype with confidence. Variants below the minimum coverage threshold need to be confirmed by an independent technology. Although the coverage required for diagnostic applications depend on the employed sequencing platform, some general considerations apply.

First of all, the read coverage of NGS runs is not uniformly distributed [21]. Coverage for all instruments is reduced in

regions with an extreme AT- or GC-rich background. The coverage fluctuation may require a higher "average" coverage to assure that each base has the previously defined "minimal" coverage for variant calling. For example, to reach a probability of >99% accuracy for each allele at a heterozygous position at least twice, a coverage of 12-fold or more is needed [assuming probability (sequencing allele A)=probability (sequencing allele B)=0.5; see supplementary Eq. (1) in the supplementary data which accompanies this article at http://www.degruyter.com/view/j/labm.2012.36.issue-4/issue-files/labm.2012.36.issue-4.xml]. The presence of both alleles in at least two independent reads is considered the baseline criterion for a heterozygous variant call.

In addition, coverage distribution of the two alleles in the case of a heterozygous position is typically biased towards the reference allele. Variant alleles are treated as mismatches to the reference sequence during the alignment step. Therefore, these reads are more difficult to align on average. A portion of variant reads will be dropped during alignment, particularly if sequencing errors coincide with genuine variants in these reads. Further quality criteria for a reliable variant call include: (i) the average quality of the variant bases, (ii) indication of the variant alleles by reads from both the forward and the reverse strand, and (iii) the number of variant calls at the surrounding region. Although the definite numbers have to be individually adjusted to the specific applied NGS technology, we propose the following general thresholds for NGS in a diagnostic set-up. (i) To achieve a sensitivity of more than 99% variant detection, a coverage of 30-fold is required. (ii) The variant allele should be sequenced by at least two independent reads from either strand. In the case of a heterozygote variant, the ratio of variant reads/reference reads should be in the range of 0.1–0.9. (iii) Regarding quality values, we propose a minimum of Q30 average base quality at the variant position; however, it has to be considered that adjustments to the quality value threshold comprise a trade-off between sensitivity and specificity. Higher quality values result in a lower false positive rate, whereas the false negative rate possibly increases as some genuine variants can be filtered out. (iv) In the case where three or more variants coincide in any window of 10 bp, all variants should be filtered out. Variants appearing this clustered are often an indication of problems such as pseudogenes or repeats in the underlying genomic context. An example of an identified mutation in the *KRAS* and some of the quality criteria discussed can be seen in Figure 4.

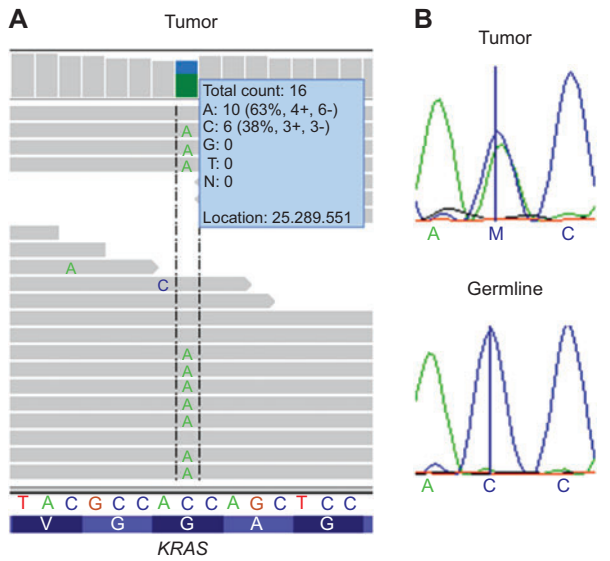


**Figure 4** Point mutation in the *KRAS* proto-oncogene identified by exome sequencing of an acute promyelocytic leukemia sample (APL) [22].
(A) Exome data set of the leukemia sample is displayed using the integrative genomics viewer [23]. Vertical gray bars represent the read depth at each position of the reference sequence. Horizontal gray bars symbolize the 76-bp reads aligned to the reference sequence. The frequency of 63% of the mutant nucleotide A in the diagnostic leukemia sample indicates a heterozygous point mutation causing an amino acid substitution (G12V). (B) Capillary sequencing of the leukemia sample confirmed the heterozygous mutation (upper panel), whereas absence of the mutation in the germline control sample from the same patient (lower panel) indicates somatic status.

## Small insertion and deletion calling

The identification of short insertions and deletion (Indels) are inherently more difficult to detect than single nucleotide variation. Because NGS alignment algorithms are built primarily for speed to handle the massive amount of raw data, they tolerate only a certain number of mismatches to the reference sequence. The gaps introduced by Indels render the reads with this type of variation increasingly more difficult to align with a growing Indel length. The quality criteria above are generally also applicable to Indel calling; however, Indels with a length exceeding ~20 nucleotides will be difficult to detect. Paired-end reads generated by the Illumina platforms can help to circumvent this problem. As two reads from the ends of the same DNA fragment are generated, initially unaligned reads are submitted to a second, more sensitive but slower alignment algorithm. This second round of alignment is not performed on the whole genome, but the already aligned paired read is used as "anchor" so that only the immediate vicinity of this read is used as reference (Supplementary Figure 1).

## Detection of tumor-specific variants

For the systematic detection of tumor-specific sequence variants on a genome-, exome- or transcriptome-wide level, it is inevitable to compare tumor and germline control samples from the same individual [22, 24–26]. Although these approaches are only about to enter the diagnostic routine and still require sophisticated filtering of variants, there are certain general considerations which need to be taken into account when dealing with variant detection in tumor samples. Any tumor sample is a mix of tumor and normal cells (e.g., stromal cells, infiltrating white blood cells). Therefore, the threshold for calling a somatic variant needs to be adjusted to the percentage of tumor cells in the sample [27]. Because any tumor

results from clonal expansion of a cell undergoing malignant transformation, different subclones may arise. Whereas certain mutations are common to all tumor cells, others are uniquely associated with the evolution of subclones [28, 29]. Hence, it is very challenging to define a universal threshold for the allele frequency of a tumor-specific variant call, especially if the amount of tumor cells in the sample is not known or mutations in subclones of the tumor might have diagnostic implications. In general, it might be useful to set a lower cut-off level for the allele frequency and a higher minimum coverage when looking for somatic variants in tumor samples in comparison to the detection of germline variants in non-tumor samples.

### False positives

Even though stringent quality criteria are applied during variant calling, some systematic errors of the different sequencing systems remain and must be considered. The most well-known systematic error type is the problem of determining the exact length of a homopolymer stretch in 454 and Ion Torrent sequencing. These systems use flow-based sequencing approaches which mean that multiple identical nucleotides are incorporated in a single flow. The emitted signal (fluorescence signal in the case of 454 or pH shift in the case of Ion Torrent) is in theory proportional to the number of incorporated nucleotides. However, in practice signal strength does not increase linear with growing homopolymer length, resulting in exceeding difficulty to call the correct number of bases with increasing homopolymer length [1, 14, 30, 31].

In addition to these error types, general error profiles are non-randomly distributed. In a given read sequencing errors tend to accumulate towards the read end. With growing read length fluorescent signals may gradually decay making it increasingly difficult to determine the accurate nucleotide. Furthermore, in the case of the Illumina cluster based sequencing phasing problems increase with growing read length. Dephasing is the term used to describe single molecules which either had a failed nucleotide incorporation in any sequencing cycle and are thus "lagging" behind in their respective cluster or exhibited the incorporation of multiple nucleotides in a single sequencing cycle (and are now "ahead" in the cluster). These out-of-phase molecules will continue to emit a false signal in all following sequencing cycles [2].

In summary, knowledge of the individual strengths and weaknesses of the employed sequencing system have to be taken into account, when data are analyzed and interpreted. Moreover, a variant calling algorithm exploiting the specific error models of the instruments should be used.

## Validation

### Sanger sequencing

From our perspective, Sanger sequencing is currently still the gold standard for validating variants detected by NGS. Worldwide, there is more than 20 years of experience with this technology, although the sensitivity for the detection of minorities against a wild type background by the Sanger method is substantially lower compared to NGS (see also below). Because, even in a highly automated process, cross-contaminations with another patient's DNA cannot be fully excluded, it currently remains important to confirm all identified pathogenic aberrations concerning the germline by Sanger sequencing.

There are two approaches to validate the results of NGS with Sanger sequencing. First, aberrations detected in a patient's sample by Sanger sequencing can be resequenced with an NGS method. Thus, identical results in both procedures can serve to validate the data. The second approach to validate NGS data is resequencing of the entire region by Sanger and to match both data sets. Either way, we propose that a minimum of 10 independent samples should be completely analyzed in parallel for all diagnostically relevant genes to complete the validation process. The validation of large gene panels with up to 100 genes and more is very costly and labor intensive and may undergo at least a limited validation process by resequencing the detected variations. The limitations in the validation process should be documented in the "methods" section in a medical report.

Regarding allele frequency, Sanger sequencing is limited in the detection of low frequency variants (<10% variant frequency). In the case of certain disorders, in particular cancer, where low frequency, tumor-specific variants are expected, different validation techniques need to be applied. A detection rate as low as ~5% can be achieved using first generation pyrosequencing, whereas using allele-specific PCR and real time PCR theoretically offers a detection rate of a single molecule in a background of $10^6$ molecules.

## Outlook

NGS has the potential to revolutionize genetic diagnostics. It allows the study of larger regions of the genome for disease causing mutations, an approach that would be too cost- and labor-intensive with traditional methods. By using NGS, the diagnostic spectrum will be expanded from Mendelian diseases to polygenic disorders, which require the simultaneous study of several different loci. Currently, a multitude of different platforms are available, yet only a few independent studies comparing different approaches have been carried out so far. With the increasing use of NGS in molecular diagnostics, consented quality criteria need to be further elaborated. These criteria will depend on the specific assay, the sequencing platform and the clinical application. Still, considerable technical challenges remain to be mastered. For example, Indels and structural variations (large deletion/insertions, translocation) are difficult to detect. For the analysis of disorders, which frequently exhibit these types of mutations, other analytical approaches may be more suitable. It should also be mentioned that the cost benefit can only be fully exploited if the enormous capacities of the sequencing instruments are completely utilized. With the increasing size of gene panels studied by NGS, the time and manpower required for validation and

interpretation of variants needs to be considered. Nevertheless, we expect that further technical developments and studies address these issues and that NGS will continue to become an essential tool not only for research but also for molecular 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 Deuteche Krebshilfe grant (109031).
**Employment or leadership:** None declared.
**Honorarium:** P.A.G. received honorarium from Illumina, Inc.

## References

1. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872–6.
2. 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.
3. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26:1135–45.
4. Gregory BD, O'Malley RC, Lister R, Urich MA, Tonti-Filippini J, Chen H, et al. A link RNA metabolism and silencing affecting Arabidopsis development. Dev Cell 2008;14:854–66.
5. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008;133:523–36.
6. Adey A, Morrison HG, Asan X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 2010;11:R119.
7. Mertes F, Elsharawy A, Sauer S, van Helvoort JM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 2011;10:374–86.
8. De Leeneer K, De Schrijver J, Clement L, Baetens M, Lefever S, De Keulenaer S, et al. Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. PLoS One 2011;6:e25531.
9. Jiang Q, Turner T, Sosa MX, Rakha A, Arnold S, Chakravarti A. Rapid and efficient human mutation detection using a bench-top next-generation DNA sequencer. Hum Mutat 2012;33:281–9.
10. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010;7:111–8.
11. Johansson H, Isaksson M, Sorqvist EF, Roos F, Stenberg J, Sjoblom T, et al. Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res 2011;39:e8.
12. Clark MJ, Chen R, Lam HY, Karczewski KJ, Euskirchen G, Butte AJ, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 2011;29:908–14.
13. Kalari KR, Casavant M, Bair TB, Keen HL, Comeron JM, Casavant TL, et al. First exons and introns – a survey of GC content and gene structure in the human genome. Silico Biol 2006;6:237–42.
14. 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.
15. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 2009;10:R32.
16. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, et al. Whole exome capture in solution with 3 Gbp of data. Genome Biol 2010;11:R62.
17. Mardis ER. A decade's perspective on DNA sequencing technology. Nature 2011;470:198–203.
18. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 2011;39:e90.
19. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 2011;12:R112.
20. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 2010;11:473–83.
21. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008;36:e105.
22. Greif PA, Eck SH, Konstandin NP, Benet-Pages A, Ksienzyk B, Dufour A, et al. Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. Leukemia 2011;25:821–7.
23. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24–6.
24. Greif PA, Yaghmaie M, Konstandin NP, Ksienzyk B, Alimoghaddam K, Ghavamzadeh A, et al. Somatic mutations in acute promyelocytic leukemia (APL) identified by exome sequencing. Leukemia 2011;25:1519–22.
25. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24.
26. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010;11:685–96.
27. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76.
28. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012;481:506–10.
29. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366:883–92.
30. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011;475:348–52.
31. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol 2011;12:R1.