Molekulargenetische und zytogenetische Diagnostik/ Molecular-Genetic and Cytogenetic Diagnostics **Redaktion: H.-G. Klein**

Hanns-Georg Klein*, Peter Bauer and Tina Hambuch

Whole genome sequencing (WGS), whole exome sequencing (WES) and clinical exome sequencing (CES) in patient care

Gesamt-Genom-Sequenzierung, Gesamt-Exom-Sequenzierung und klinische Exom-Sequenzierung in der Patientenversorgung

Abstract: Next generation sequencing (NGS, also called Massively Parallel Sequencing) can be performed using a number of different platforms. The general process is very similar across them all: (1) extracted DNA is sheared into fragments (which, in targeted methods can be captured using probes); (2) these fragments are isolated physically on slides (usually called flow cells) or in emulsions and individually amplified, resulting in a library; and (3) the multiple individually amplified fragments are then simultaneously sequenced. After sequencing each fragment individually, the fragments must be re-assembled and the positions called using a series of bioinformatics algorithms. Excellent reviews are available that discuss the technical differences in detail. Recently, the value of NGS for diagnostics in patient care has been widely recognized and its applications include mutation detection in human genetics, molecular pathology and infectious agents as well as HLA typing, RNA sequencing and the detection of cell-free DNA. This paper focuses on applications of three different scales of NGS in human genetics diagnostics and evaluates its status based on our current understanding.

Keywords: Clinical exome sequencing (CES); molecular genetic diagnostics; multi-gene panel sequencing (MGPS); next generation sequencing (NGS); translational genetics; whole exome sequencing (WES); whole genome sequencing (WGS).

Zusammenfassung: Next Generation Sequencing (NGS, auch als massiv-parallele Sequenzierung bezeichnet) kann unter Verwendung verschiedener Geräteplattformen durchgeführt werden. Die grundlegenden Arbeitsschritte sind bei allen Plattformen sehr ähnlich: (1) die extrahierte DNA wird in Fragmente gescheert (welche bei gezielten Anreicherungsverfahren mittels Sonden abgefangen werden können), (2) die Fragmente werden physikalisch auf Objektträgern (üblicherweise als Durchflusszellen bezeichnet) oder in Emulsionen isoliert und individuell amplifiziert, woraus eine sog. Bibliothek entsteht und (3) die verschiedenen individuell amplifizierten Fragmente werden simultan sequenziert. Nach der Sequenzierung müssen die einzelnen Fragmente mit Hilfe von bioinformatischen Algorithmen an ihrer richtigen Position wieder zusammengesetzt werden. Es gibt ausgezeichnete Übersichtsartikel, welche die technischen Unterschiede im Detail beschreiben. Seit kurzem ist der Wert von NGS für die Diagnostik und in der Patientenversorgung allgemein anerkannt. Die Anwendungen umfassen den Nachweis von Mutationen in der Humangenetik, Molekularpathologie und Infektiologie sowie die HLA-Typisierung, RNA-Sequenzierung und die Detektion von zellfreier DNA. Dieser Beitrag konzentriert sich auf Anwendungen von drei verschiedenen Skalen von NGS in der humangenetischen Diagnostik und wertet deren Bedeutung auf der Grundlage unseres derzeitigen Verständnisses.

Schlüsselwörter: Gesamt-Genom-Sequenzierung; Gesamt-Exom-Sequenzierung und klinische

***Corresponding author: Hanns-Georg Klein,** Center for Human Genetics and Laboratory Diagnostics Dr. Klein, Dr. Rost and Colleagues, Lochhamer Str. 29, 82152 Martinsried, Germany, Tel.: +49-89/895578-0, Fax: +49-89/895578-78, E-Mail: klein@medical-genetics.de
Peter Bauer: Institute for Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
Tina Hambuch: Illumina Clinical Service Laboratory, Illumina Inc., San Diego, CA, USA

Exom-Sequenzierung; molekulargenetische Diagnostik; Multi-Gene Panel Sequenzierung (MGPS); translationale Genetik.

DOI 10.1515/labmed-2014-0025
Received July 4, 2014; accepted July 7, 2014

Introduction

The **first step** of each next generation sequencing (NGS) procedure is DNA extraction, which is specific to the sample type (i.e., blood, tissue, body fluids), as is standard for DNA extractions used for other molecular procedures performed in a clinical molecular laboratory. Many commercially available kits can be purchased from several different companies. Once extracted, the DNA should be evaluated for quality in a standardized process of quality control (QC) to assure proper quality and concentration. Although NGS is fairly robust with regard to DNA characteristics, the age, storage and purity of the DNA may affect the outcome of the specific assay and should be optimized accordingly.

The **second step** is the preparation of a library containing all DNA pieces extracted from the sample, followed eventually by the specific enrichment of the regions of interest. Libraries are prepared by randomly shearing high-molecular weight genomic DNA (gDNA) using shearing techniques such as nebulization or sonication. The sheared gDNA then goes through a series of enzymatic reactions to create an adenosine overhang that an oligonucleotide adapter will bind to using DNA ligase. After adaptor ligation, the sample is size-selected by gel electrophoresis, gel extracted and purified. At this point, a genomic library has been created, that can be used directly for sequencing, or from which specific regions can be captured for downstream sequencing.

The **third step** is the sequencing reaction that can be carried out using a variety of different methods, which have been extensively reviewed [1–4]. The method currently most widely being used is sequencing by synthesis, which requires the DNA library to be denatured and loaded on to a glass flow cell where it hybridizes to a lawn of fixed oligonucleotides that are complementary to the adaptors. The single-stranded, bound library fragments are then extended, and the free ends hybridize to the neighboring lawn of complementary oligonucleotides. This "bridge" is then grown into a cluster through a series of polymerase chain reaction (PCR) amplifications. In this way, a cluster of approximately 2000 clonally amplified molecules are formed; across a single lane of a flow cell, there can be over 37 million individual, amplified clusters. After amplification, the clusters are denatured to free the 5′ ends, and primers initiate a sequencing reaction whereby individually labeled nucleotides competitively bind to the template. After the laser excites the fluorescent dye on the nucleotide, images are captured for each dye. The process is stepwise, enabling each added nucleotide to be measured. This process is repeated between 36 and 120 times. At this point, the opposite end of the fragment is also sequenced. To do this, the newly synthesized strand is removed by denaturation, the 3′ strand is unblocked and new double-stranded DNA clusters are generated, as they were initially in the cluster station, by bridge PCR. This time, the forward strands are removed, and the opposite strand is primed for sequencing and sequenced in the same manner. In this way, each fragment is sequenced from opposite ends (Figure 1).

The **fourth step** of an NGS analysis is the data management requiring a series of downstream analyses that are summarized under the term "Bioinformatics". The accuracy of calls is dependent on the combination of the quality and complexity of the substrate being analyzed, the chemistry, the platform analytics, and the combined downstream bioinformatic algorithms applied to a sequencing run. Each NGS specific application (diploid sequencing, cancer sequencing, etc.) may differ in the algorithms and filters utilized during data analysis. Algorithms that are used in the process of alignment and variant calling may be optimized to different types of variants (e.g., insertion-deletion events or single nucleotide substitutions), or may perform unequally in different regions of the genome; these types of performance differences may or may not be coupled to the platform that is used, the capture methods employed, or other parameters. Therefore, the accuracy of NGS is reliant on a very specific combination of factors, including the platform, and DNA.

The bioinformatics process typically involves three stages: (1) assessment of individual signal strength for each base called, (2) mapping of the fragments to a reference sequence, and (3) identification of variants as well as quality of the consensus calls. The signal strength is assessed against the noise for quality evaluation in a method similar to the approach used in generating PHred scores used in Sanger sequencing. After the millions of individual sequence fragments have been read, they must be mapped to the correct locations in the genome in a process called mapping or alignment. Alignment methods are variable in stringency and optimized for specific types of activities, such as identifying unique regions or reconstructing translocation events.

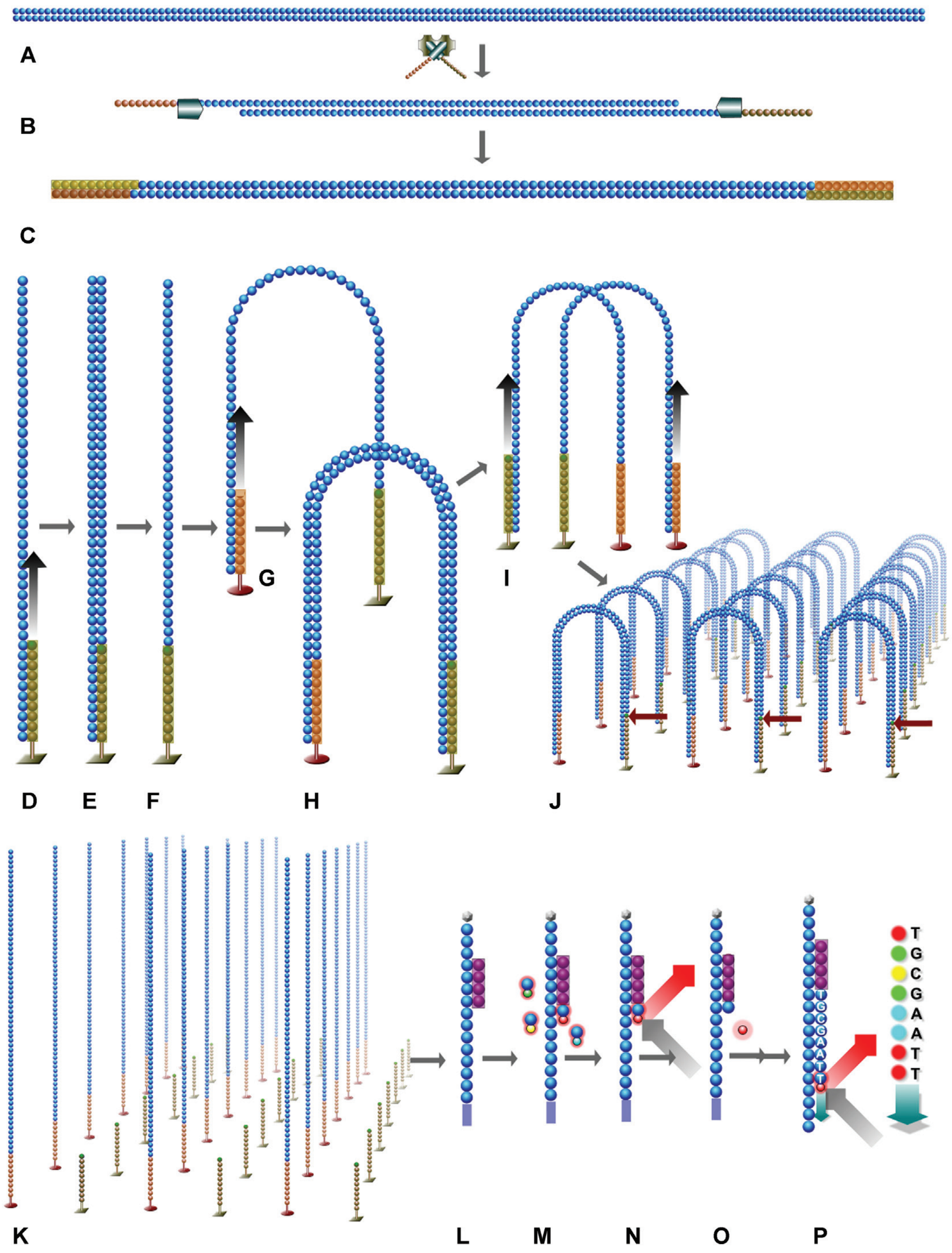

Figure 1 Sequencing by synthesis.
Library prep: (A) genomic DNA incubated with Nextera transposase, (B) DNA is fragmented and fragments are tagged with Illumina adapter sequences, (C) limited cycle PCR fills the gaps to create a functional library. Bridge amplification: (D) denatured library is bound to oligonucleotides attached to flow cell surface, (E) bound library is 3′-extended by polymerase, (F) hybridized template is denatured and washed off leaving a copy of the template bound to surface, (G) DNA is allowed to loop over and hybridize to adjacent flow cell oligonucleotide, (H) DNA polymerase creates a double-stranded molecule, (I) DNA loop is denatured and allowed to hybridize to flow cell oligonucleotides and extended again, (J) the process is repeated until a cluster of approximately 1000–5000 copies are generated, (K) the cluster is linearized by denaturing followed by cleavage of the reverse strand (flow cell oligos contain a cleavable base) and 5′ ends are blocked, (L) sequencing primer is hybridized. Sequencing: (M) DNA polymerase adds the correct reversibly terminated fluorescent labeled nucleotide from a pool of all four nucleotides, (N) flow cell is imaged to capture the color of each cluster, (O) the reversible terminator and fluorescent dye are cleaved off, (P) the sequencing by synthesis cycle is repeated up to 300 times to read additional bases with the color of each cluster being converted to a base call (courtesy of Abizar Lakdawalla, Illumina Inc., San Diego, CA, USA).

Typically, fragments are aligned to a reference. Differences between the reference genome and the sample genome can be mapped and evaluated. Components such as the number of independent sampling events (represented by the different fragments), the quality of the calls and how uniquely and well a specific fragment could be aligned to a specific region are then assigned a quality score, which can be used to gauge the confidence of that specific call.

Whole genome sequencing (WGS)

So-called whole genome sequencing (WGS) is most often used in the assessment of a rare disease of suspected genetic etiology where symptoms may be overlapping or non-specific and first tier testing has been inconclusive [5–7] or where WGS presents the fastest possible aid for differential diagnostic evaluation [8]. The clinical assumption driving this testing is that the condition is caused by a single gene (monogenic or Mendelian conditions), and that genome sequencing will provide the most cost and time-effective assessment of the possible genes involved. The primary intention of clinical testing is not gene discovery; however, as with microarray testing, variants may be identified in genes for which the function is not yet established, but only suspected or perhaps completely undefined. In such cases, if those variants are thought to be likely causative, additional testing may be required and ideally clinical laboratories should have plans for how to make such recommendations to physicians who have ordered the test.

Technical aspects

WGS typically provides coverage for approximately 95% of the genome; however, there is a challenge in that most of the genome is not currently well understood, and therefore clinical associations for many regions are speculative at this time. That said, the genome does provide coverage for approximately 5% more of the coding region than capture-based approaches, and this results in the inclusion of approximately 200 additional genes. Furthermore, because there is no capture of fragments involved, there is more evenness in coverage and less ascertainment bias, which can be useful in detecting more complex variants, such as copy number or structural variants.

From a technical perspective, it is important to explicitly define the test and intended use to clarify the capabilities and deliverables of the assay. First, WGS, and in particular clinical WGS, is not representative of every base position of the entire genome nor can it detect all types of sequence variants that might be present in a whole genome. Several types of variants, such as copy number and structural variants can be detected, but with differing degrees of accuracy relative to single nucleotide variants, and these differences in technical accuracy relative to variant type should be established and described. In particular, for clinical WGS, thresholds or statistical algorithms can be used to determine whether each variant call meets strict quality metrics that are used to ensure that when calls are made in a clinical context that they meet a minimum threshold of accuracy.

Diagnostic aspects

When deciding to offer WGS, the laboratory should consider the needs of the clinical population being served, and offer specific support regarding what the test can and cannot be used for, and the degree to which the clinical laboratory is able to support the wide range of potential clinical questions for which WGS might be employed. In the Illumina Clinical Services Laboratory, approximately 2.9 billion bases are called to quality thresholds that correspond to sensitivity over 97% and specificity over 99%. The specific accuracy and distribution of those calls is displayed in Table 1 and Figure 1. However, it should be noted that this is the accuracy for detection of single nucleotide variants. Insertion or deletion events (plus or minus up to seven base pairs) can be detected with an accuracy of 80%, and larger insertion or deletion events are detected with less accuracy. Each type of variant and regional characteristic of a genome should be evaluated in a series of

Table 1 Differences in sensitivity and specificity according to depth of coverage.[a]

Depth of coverage	Sensitivity	Specificity
30×	>99.9%	>99.99%
10×	98.0%	99.99%

[a]As molecules are independently sequenced, a diploid sample requires multiple independent sampling events to detect both alleles. Different depths of coverage, or sampling events, will result in differing probabilities of detecting a variant. This table reflects the ability to detect a variant when present (sensitivity) and the probability of correctly identifying the nucleotide(s) present (specificity) using whole genome sequencing in the Illumina Clinical Services Laboratory (ICSL). These data were generated using multiple, orthologously established heterozygous and homozygous loci across the genome.

validation experiments so that the specific confidence in the accuracy of a call can be reported, as additional or supplementary testing may be appropriate in some situations (Figure 2).

Interpretation

Once calls are made, one must identify which variants, across the 3–4 million variable positions, including an average 9600 amino acid changing positions and 73 premature termination positions (internal data from WGS analyses), are likely to be clinically relevant. Given such a large amount of information, a thoughtful plan must exist for how to identify and evaluate the information that is most likely to be relevant and informative to the clinical questions being considered. Typically, this process involves information gathering (annotation), information assessment (interpretation) and integration of all the resulting information into a report that addresses the clinical question(s) for which the test was ordered. Currently, multiple tools are available to support each of these steps, but the processes are still mostly manual and require extensive evaluation by a trained team.

A genome can potentially be used to assess many different questions regarding an individual's health, predisposition, carrier status and propensity for drug reactions. Genome sequencing can also be used to identify somatic or mosaic variants. The opportunities for clinical applications of genome sequencing are rather numerous and adoption of the technology has grown significantly. However, infrastructure, technical refinement and outreach to physicians will be necessary to enable WGS to become a tool that can be used broadly to address all the potential clinical questions that are possible.

Whole exome sequencing (WES)

In human genetics, whole exome sequencing (WES) was one of the first kitted applications for NGS. From the beginning, this technology was used to identify ultra-rare Mendelian disorders by sequencing informative families with distinct but unresolved phenotypes [5]. In this respect, recessive conditions, preferentially observed in children from consanguineous parents, propelled the breakthrough of this application, because research analyses yielded hundreds of new disease genes for almost all clinical disciplines [9–11]. Furthermore, WES in trios with intellectual disability established the importance of de novo mutations as a new diagnostic paradigm. Over the years, WES technology has greatly improved in terms of uniformity of the data sets, robustness of the laboratory process and advancements in the filtering and interpretation of individual variants. Nonetheless, the technology is still not able to deliver >90%–95% of all coding sequences. Moreover, the test is not really quantitative and therefore only crudely accounts for gene dosage alterations in the human exome. Therefore, WES must be regarded as a very powerful screening test but exclusion of suspected diagnoses has to be validated thoroughly (and is usually not possible). Although almost everybody is convinced that WES will be a short technological era, ending as soon as WGS will have shown to produce better data at virtually no extra costs, for the moment, it represents the mainstream application and is setting standards for any technology beyond.

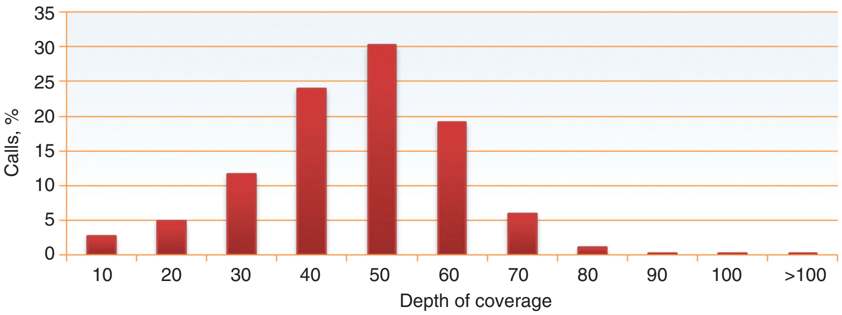

Figure 2 Distribution of depths of coverage for whole genome sequencing.
This graph depicts the distribution of depths (independent sampling events, see Table 1 for description) achieved with an average of at least 30-fold depth of coverage. Because whole genome sequencing does not involve pull-down or focused amplification events, the distribution is consistently narrower around the average than more focused approaches. For this reason, average depths of coverage may differ across different methodologies and still result in similar accuracies. The minimum call made in the Illumina Clinical Services Laboratory is 10-fold, which corresponds to sensitivity and specificity of 98% and 99.9% (see Table 1).

Technical aspects

WES enrichment strategies involve either hybridization-based capturing using commercially available capture probes or amplification-dependent enrichment with kitted primer pools for approximately 50–70 Mb of coding human sequences (CCDS) and additional relevant adjacent regions such as exon-intron boundaries, 5′- and 3′-untranslated regions and hundreds of non-coding regions such as microRNA binding sites, etc. Usually, enrichment protocols are accomplished within 1–2 working days and result in 50%–80% target specific DNA enrichment (i.e., after sequencing these moieties turn out to map back to human exons, whereas 20%–50% of the sequencing reads are still unspecifically derived from non-coding human sequences). Standard protocols warrant 5–10 Gb raw data sequencing to create a robust (>60×) mean coverage for the target region. Not surprisingly, some difficult sequences within the human exome are not represented by enough reads. This vertical coverage is an important measure to define the maximum sensitivity of the assay because any region that is not covered >20 independent sequencing reads might contain heterozygous variants that are just not detected. There is growing consensus that diagnostic WES analyses should mention this vertical coverage at 20× coverage, which is normally in between 85% and 92%, the latter for analyses with more than average total sequencing reads.

Bioinformatic analysis pipelines are rather similar to genomic variant calling and annotation pipelines except for the technical problems in identifying small and large insertions/deletions. The technical explanation for a reduced sensitivity for these variants is the relatively uneven distribution of sequencing reads over different exons. In particular, complex genomic structures, repeat elements and GC-rich sequences are usually poorly covered and therefore overrepresented in low-confidence exonic regions.

Diagnostic aspects

Applying WES in a clinical setting mandates the laboratory to validate all processing steps (library preparation, enrichment, sequencing, bioinformatics, reporting). In particular, the demonstration that the processed data are correct is not easy to prove. Partly, participation in external quality assessment (such as the NGS scheme of EMQN) or sequencing of standard DNAs that have been validated for most of their variants (such as the "genome in a bottle" individuals at the Coriell DNA repository, [12]) will approximate diagnostic parameters such as sensitivity, specificity, positive predictive value and negative predictive value. Especially the latter – negative predictive value – is barely defined with a test that "only" covers 85%–92% of the anticipated target regions. Furthermore, only part of the human exome is interpretable for diagnostic purposes because nobody is able to deduct pathogenic variant effects in genes that have not been linked to the patient's phenotype before. In this respect, almost all laboratories started to curate disease-specific gene lists being used to filter the exomic variant set down to a smaller list with potentially disease-relevant variants. Moreover, this phenotype-specific filtering greatly reduces the risk for uncovering any unsolicited finding, for example, the discovery of predisposing tumor genes in families asking for a diagnosis of familial deafness. Although the identification of actionable genetic variants in genes not related to the phenotype of the patient might be advantageous for individuals, this procedure has to be explicitly explained during the pre-test informed consent information and opt-in or opt-out models for these actionable variants should have been established before any testing is publically offered.

Interpretation

Because genomic variation is very common for the outbred human population, anybody interpreting the variant list of a human exome is facing approximately 25,000 variants in every dataset. Almost all of these variants are likely benign and only represent the phenotypic diversity of the human race. Nonetheless, the identification process for those variants that are potentially linked to human traits and disease phenotypes is therefore a tremendous challenge with a high risk to misinterpret both: benign variants as potentially disease-causing and vice versa. Usually, common variants as identified in the 1000 genomes project with frequencies >1% are regarded as almost always benign. Thus, 99% of the variation can be filtered by only looking at rare variants in WES datasets. Thereafter, laboratories usually exclude coding variants without any published evidence and moderate to weak evolutionary conservation. Both additional filter categories are ambiguous because published evidence might be misleading or simply overlooked, and in silico prediction based on evolutionary conservation has a moderate diagnostic precision (<80%). In this respect, a universal binning of "variants of unclear significance – VUS" has been proposed and widely accepted by sequencing laboratories ([13], see also Table 2). Although clear observations

Table 2 Variants of unclear significance (VUS) classification (according to Plon et al. [13]).

Nomenclature	Description	Error probability
VUS1	Benign without clinical significance	<0.1%
VUS2	Very likely without clinical significance	<0.5%
VUS3	Unknown significance	
VUS4	Likely pathogenic	<5%
VUS5	Almost definitively pathogenic	<0.1%

can be assigned without much ambiguity, for many missense variants neither the literature nor the modeling will allow a decisive classification. Thus, these types of VUS are usually classified as VUS3 (Table 2).

In practice, laboratories applying robust filtering for variant qualities and decent gene panels for diagnostic request in WES data sets still face an abundance of potentially relevant variants: 0–2 deleterious mutations related to the disease phenotype and another 4–9 VUS [13] might be expected according to Yang et al. [9]. In addition, 0–1 medically actionable mutation, 0–1 autosomal recessive carrier status and 0–4 relevant pharmacogenomics variant is uncovered. Having this stated for the phenotype-associated and actionable genes, another 1–3 deleterious mutations in unrelated diseases and 17–41 VUS in unrelated genes as well as 17–25 deleterious variants in genes with no known disease associated are detected. This mere mass of potentially relevant information makes the reporting process and genetic counseling an extremely difficult job, unraveling a new bottleneck of the technology: interpretation and communication. This tremendous challenge can only be answered by assembling multidisciplinary teams with strong expertise in genomic medicine.

Clinical exome sequencing (CES)

Clinical exome sequencing (CES) refers to an application of NGS that focuses on genes, in which mutations have been found to be associated with disease and reported in the Human Mutation Database® [14]. This subset of the exome currently contains approximately 5000 genes (25% of the exome) and is continuously expanding. In 2011, Ambry Genetics (Aliso Viejo, CA, USA) was the first CLIA laboratory to introduce a “Clinical Diagnostic Exome” using NGS technology. Subsequently, Illumina Inc. (San Diego, CA, USA) has developed a kit for CES, containing 2800 genes in its first version that has recently been expanded to 4813 genes using the Nextera enrichment method (TruSight™ One). The main advantages of the CES approach are (1) cost efficacy by focusing on clinically characterized genes, allowing trio analyses and improved data quality (deep coverage), thereby reducing data analysis, interpretation and turn-around time; (2) the avoidance of generating large numbers of VUS, thereby limiting the complexity of genetic counseling; and (3) the option to carry out the analyses on a benchtop scale instrument such as the Illumina MiSeq. For a growing number of Mendelian diseases, biochemical (i.e., mass spectrometry) or immunological (i.e., flow cytometry) test panels have now become available, which may facilitate the interpretation of DNA findings by functional validation (translational genetics).

Technical aspects

The workflow for the Illumina TruSight Exome Kit involves the preparation of an indexed, pooled library from as little as 50 ng of DNA followed by target enrichment using the one step Nextera technology (TruSight rapid capture). Depending on the number of samples, the enriched pool sample is sequenced by synthesis on a flow cell of either an Illumina MiSeq or a HiSeq instrument. The entire procedure may take <3 days. The Nextera enrichment technology employs 80mer capture probes, targeted to the center of each exon and allows an average insert size of the library of 500 bp. This insert size provides additional information on most of the clinically relevant adjacent splice sites. More detailed technical information is provided on the Illumina homepage [15]. The sequencing data are first analyzed using an on-instrument software (i.e., MiSeq reporter software). After the alignment to reference DNA and creation of the appropriate file format, a variety of software tools can be applied for functional computing such as the CLCbio genomics workbench (CLCbio, Aarhus, Denmark, [16]).

Diagnostic aspects

From the viewpoint of a diagnostic laboratory, the scope of a CES analysis for routine diagnostics in human genetics is still very broad, time-consuming and a big leap from the multi-gene panel sequencing (MGPS), which just arrived in patient care. Regarding coverage, data quality, cost and avoidance of unsolicited findings or VUS, MGPS is presently the better choice, particularly if the disease entity is clinically well defined. Yet, the CES approach seems to be an extremely useful diagnostic tool for complex clinical syndromes such as developmental delay, intellectual disability or multiple congenital malformations. The genetic reasons for approximately 50% of cases are still not being resolved, although array comparative genomic hybridization (CGH) and improved fluorescence in situ hybridization (FISH) and sequencing techniques are applied [17]. Because complex disorders may involve several hundred genes, a CES analysis is the method of choice, if classic karyotyping, array CGH and conventional DNA sequence analysis remain inconclusive. Owing to the still large amount of data generated by CES analysis, it is recommended to analyze these cases as trios (i.e., healthy parents – affected child) thereby excluding non-causative variants in accordance with the proposed mode of inheritance [18–20]. It may also be advisable to specifically blind out unsolicited genetic information such as genes associated with late manifesting diseases or cancer. Professional genetic counseling is important for the patient and the family to opt in or opt out for certain genetic information. Owing to the capability of CES to readily detect de novo dominant mutations, it is expected that CES analyses will improve the sensitivity of genetic diagnostics in developmental disorders by 20%–30% [19, 20].

Unfortunately, the reimbursement situation for NGS diagnostics of rare diseases is very unclear, hampering the progress in this field. In Germany, diagnostic sequencing has been restricted to the Sanger method as of October 2013. Currently, the health insurance companies are being asked by quotation to cover the cost.

Interpretation

The CES data set (ca. 12 Mb targeted) is in comparison with WES (ca. 50 Mb) and WGS (ca. 3000 Mb) significantly smaller (Figure 2) and has the advantage of dealing with known disease associations. The workflow for a classic CES trio data analysis includes variant calling, quality control (i.e., reads on target, base quality), masking of undesired genes (i.e., cancer, late-manifesting disease genes), variant annotation (i.e., dbSNP, EVS, HGMD®, OMIM®) and quality filtering. Depending on the proposed model of inheritance, numerous options are available for filtering and detection of variants of interest (see Figure 3, [21]). Candidate variant interpretation is supported by public data bases from clinical resources (HGMD, OMIM), functional analysis data bases (GO, KEGG) and in silico modeling programs (Mutation Taster, SIFT, Polyphen, [22]). The interpretational work is concluded by reporting clearly deleterious variants that follow a reasonable inheritance model and are consistent with the patient's phenotype or potentially pathogenic VUS. The report can be further elaborated by additional biochemical analyses or imaging.

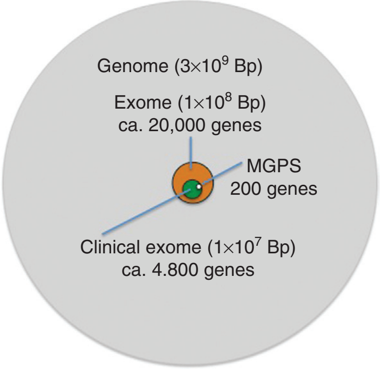

Figure 3 Illustration of the size of the whole genome, the whole exome and the clinical exome.
Multi-gene panels are shown for comparison.

Discussion and summary

In summary, three different approaches to large-scale genomic analysis for diagnostic purposes are discussed. The main differences include the size of genomic regions interrogated (Figure 3), the cost efficacy, the general work load and the spectrum of detectable genetic variants.

Larger target regions yield higher numbers of potentially disease-causing variants, but also many more VUS and unsolicited findings. Although cost efficacy and the general workload for enrichment steps are generally not scalable which means that a small gene panel and WES virtually impose compared costs and workload, this equality is not true for data analysis and interpretation. At least there, the sequencing costs, data analysis efforts, data storage, data interpretation and complexity of genetic counseling correlate with the size of the target regions. In this respect, a diagnostic laboratory is always trying to reduce interpretation complexities by looking up phenotype-specific gene lists first, regardless of which sequencing data set has been produced. By omitting enrichment biases, WGS allows the detection of structural variants, copy number variants, non-coding variants and coding variants in previously not annotated genes. Among the enrichment-based NGS methods, WES seems preferentially useful in a research setting, because it allows the de novo detection of disease-associated mutations in candidate genes, whereas – owing to its focus on disease-associated genes – CES is more suitable for diagnostic purposes. A technical summary of diagnostic applications and data analysis of NGS has recently been published by Vogl et al. [22]. Figure 4 summarizes the bioinformatic and interpretational workflow for WGS, WES and CES and obviously illustrates that a diagnostic laboratory has to implement a very complex workflow for bioinformatics and interpretation. High surveillance has to be made to apply correct

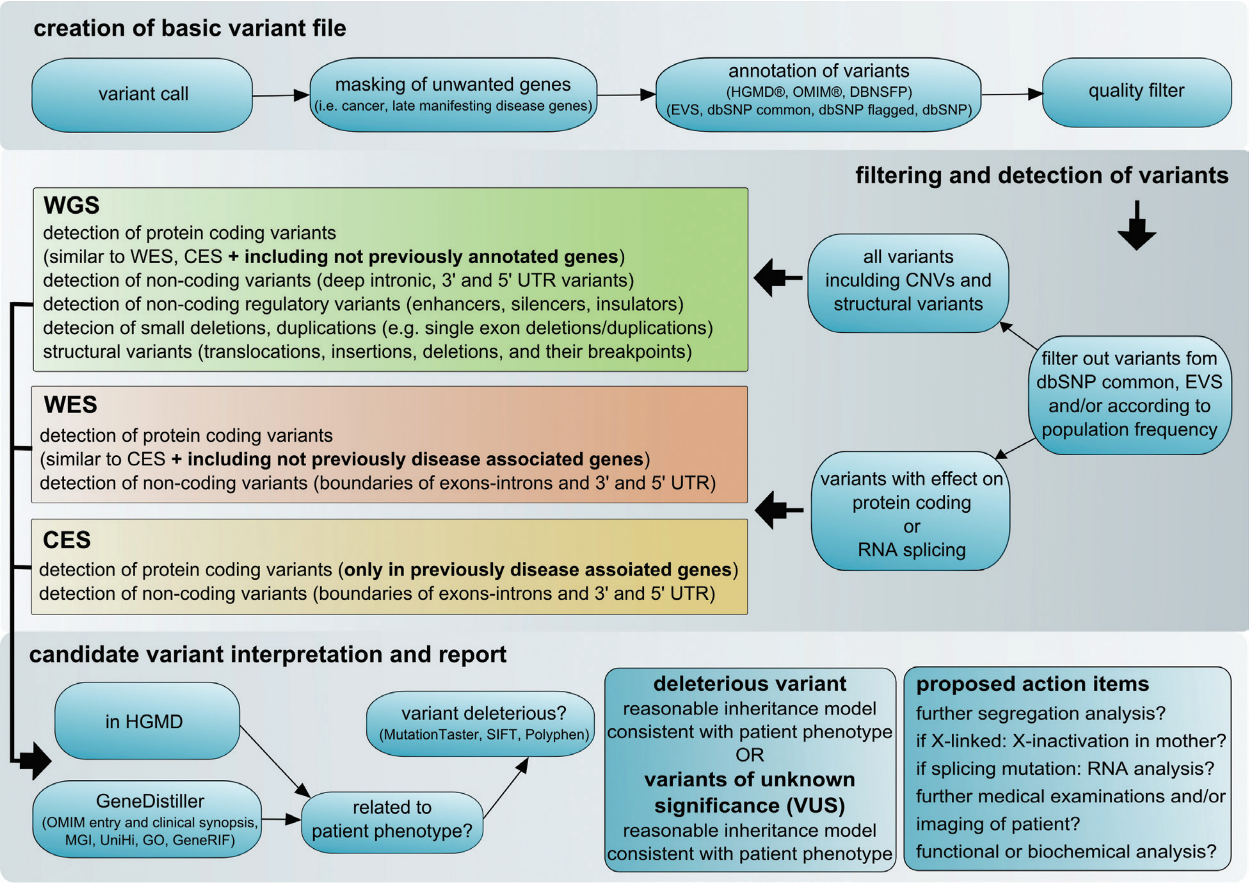
creation of basic variant file
variant call
masking of unwanted genes (i.e. cancer, late manifesting disease genes)
annotation of variants (HGMD®, OMIM®, DBNSFP) (EVS, dbSNP common, dbSNP flagged, dbSNP)
quality filter
filtering and detection of variants
WGS
detection of protein coding variants
(similar to WES, CES + including not previously annotated genes)
detection of non-coding variants (deep intronic, 3' and 5' UTR variants)
detection of non-coding regulatory variants (enhancers, silencers, insulators)
detecion of small deletions, duplications (e.g. single exon deletions/duplications)
structural variants (translocations, insertions, deletions, and their breakpoints)
all variants inculding CNVs and structural variants
filter out variants fom dbSNP common, EVS and/or according to population frequency
WES
detection of protein coding variants
(similar to CES + including not previously disease associated genes)
detection of non-coding variants (boundaries of exons-introns and 3' and 5' UTR)
variants with effect on protein coding or RNA splicing
CES
detection of protein coding variants (only in previously disease assoiated genes)
detection of non-coding variants (boundaries of exons-introns and 3' and 5' UTR)
candidate variant interpretation and report
in HGMD
GeneDistiller (OMIM entry and clinical synopsis, MGI, UniHi, GO, GeneRIF)
related to patient phenotype?
variant deleterious? (MutationTaster, SIFT, Polyphen)
deleterious variant
reasonable inheritance model consistent with patient phenotype
OR
variants of unknown significance (VUS)
reasonable inheritance model consistent with patient phenotype
proposed action items
further segregation analysis?
if X-linked: X-inactivation in mother?
if splicing mutation: RNA analysis?
further medical examinations and/or imaging of patient?
functional or biochemical analysis?

Figure 4 Summary of the bioinformatics and interpretational workflow in WGS, WES and CES (courtesy of S. Dölken and I. Vogl, Center for Human Genetics and Laboratory Diagnostics, Martinsried, Germany).

Table 3 Key features of different scales of Next Generation Sequencing (NGS) approaches in comparison to conventional step diagnostics using Sanger sequencing.[a]

Variables	Cost	Analytical sensitivity	Diagnostic sensitivity	Interpretability	Visible variants	Avoidance of incidental findings	Analytical specificity
	Coverage, Q30 bases	P (true positive)	Quality of clinical examination, specificity of phenotype	VUS, genotype-phenotype correlation	SNVs, MNVs, CNVs, non-coding, structural variation		P (true negative)
WGS	$$$	+	+++	+	+++++	+	+
WES	$$	++	++	++	+++	++	++
CES (Trio)	$($)	+++	+++	+++	+++	+++	++++
MGPS	$	++++	+	+++	+++	++++	+++
Conventional diagnostics	$$$$	+++++	+	++++	++++	+++++	++++

[a]CES, clinical exome sequencing; CNV, copy number variant; MGPS, multi-gene panel sequencing; MNV, multi-nucleotide number variant; SNV, single-nucleotide number variant; VUS, variants of unclear significance; WES, whole exome sequencing; WGS, whole genome sequencing.

filtering and data basing for true positives and false positives. Moreover, every yet-unclassified potentially relevant variant might warrant in-depth follow-up analyses in the family but often even prompt the clinicians to collect additional clinical evidence by further clinical, biochemical or radiological analyses. Table 3 gives a synopsis of the key features of currently applied NGS methods and might serve as a basis to guide NGS diagnostics. Nonetheless, as technology evolves so fast, an expert update on current diagnostic algorithms is always advisable prior to requesting NGS diagnostics.

Conflict of interest statement

Author contributions: All authors have accepted responsibility for the entire content of this submitted manuscript and approved submission. All authors contributed equally to this work.

Research funding: P.B. was partly supported by an EU research grant (NeurOmics; grant agreement 2012-305121).

Employment or leadership: T.H. is an employee of Illumina Inc. (San Diego, CA, USA).

Honorarium: T.H. is an employee of Illumina Inc. (San Diego, CA, USA), P.B. received speaker honoraria from Illumina Inc. (San Diego, CA, USA).

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Mardis E. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 2013;6:287–303.
2. Tucker T, Marra M, Friedmann JM. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet 2009;85:142–54.
3. Voelkerding KV, Dames S, Durtschi JD. Next generation sequencing for clinical diagnostics – principles and application to targeted resequencing for hypertrophic cardiomyopathy. J Mol Diagn 2010;12:539–51.
4. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, et al. Next generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn 2011;11:333–43.
5. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 2010;42:30–5.
6. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature 2014; 511:344–7.
7. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 2011;13:255–62.
8. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med 2012;4:154ra35.
9. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. N Engl J Med 2013;369:1502–11.
10. Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 2011;478:57–63.
11. Novarino G, Fenstermaker AG, Zaki MS, Hofree M, Silhavy JL, Heiberg AD, et al. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. Science 2014;343:506–11.
12. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 2014;32:246–51.
13. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 2008;29:1282–91.
14. The Human Gene Mutation Database. Available at: www.hgmd.org. Accessed 23 June, 2014.
15. Illumina. Products/TruSight One Sequencing Panel. Available at: www.illumina.com/products/trusight-one-sequencing-panel.ilmn. Accessed 23 June, 2014.
16. CLCbio. CLC Genomics Workbench. Available at: www.clcbio.com/products/clc-genomics-workbench. Accessed 23 June, 2014.
17. Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, Becker C, et al. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet 2006;140:2063–74.
18. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, et al. A de novo paradigm for mental retardation. Nat Genet 2010;42:1109–12.
19. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe nonsyndromic sporadic intellectual disability: an exome sequencing study. Lancet 2012;380:1674–82.
20. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 2012;367:1921–9.
21. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 2012;20:490–7.
22. Vogl I, Benet-Pagès A, Eck SH, Kuhn M, Vosberg S, Greif PA, et al. Applications and data analysis of next-generation sequencing. J Lab Med 2013;37:305–15.