Genetic Diagnostics

The Best Possible Exome

Tech Note

The Best Possible Exome

Whole genome sequencing (WGS) seems to be the most complete genomic analysis available. Compared with whole exome sequencing (WES), WGS trades depth of coverage (sensitivity) for breadth of coverage (percent of the genome represented). While WGS is a great tool for research, diagnostic use requires high sensitivity, and the limiting factor is the interpretability of detected variants. Our analysis shows that medically relevant regions are better represented in a well-designed exome.

Whole exome sequencing (WES), the analysis of all known protein-coding sequences in the human genome, has been used for many years in clinical genetic diagnostics. While whole genome sequencing (WGS) is becoming cheaper as sequencing costs decrease, it is still significantly more expensive in terms of sequencing, data processing, and storage costs. At an average coverage of 30x, a WGS dataset contains more than six times the sequencing data of a typical WES analysis at 120x.

On the other hand, WGS provides coverage of nearly the complete genome, while WES is limited to coding regions and proximal regions such as intronic borders and UTR. Furthermore, the enrichment used in WES protocols may result in more uneven coverage, leaving some relevant coding regions with insufficient coverage.

To understand these differences, we have compared a deeply sequenced WGS dataset (350 Gbp raw data, average coverage 100x) with WES data. The analysis has two aims: Firstly, to establish the difference in performance of the two methods in terms of covering diagnostically relevant regions. To this end, we analyze coverage both on coding sequences from the MANE Select database and on disease-causing deep-intronic variants from HGMD and ClinVar. Secondly, to show that most missing regions are not due to the library preparation (WES vs. WGS) but due to issues of mappability and necessary filtering steps during data processing. By using deep WGS as the reference, our results can help researchers understand the varying claims by sequencing providers concerning the strengths and limitations of WES analyses.

For WES analyses, we show data from a commercially available exome kit and from CeGaT's proprietary ExomeXtra® version 6. CeGaT's ExomeXtra® covers coding sequences from multiple transcript databases (MANE, Gencode, Ensembl, RefSeq Curated, CCDS). It is augmented by adding (1) all manually curated coding and non-coding regions from CeGaT's over 20 diagnostic panels covering hundreds of inherited diseases, (2) all pathogenic and likely pathogenic non-coding variants described in HGMD, (3) all pathogenic and likely pathogenic non-coding variants published in ClinVar, (4) the complete mitochondrial genome, (5) remaining coding regions from CeGaT's gene database, (6) regions with pharmacogenetically relevant variants in selected genes, (7) probes for detecting relevant prenatal infections and tumor-driving viruses, and (8) a backbone covering all chromosomes to enable whole-genome CNV detection.

We evaluate three metrics: average coverage, evenness of coverage (fold80), and completeness of coverage (% target covered ≥30x). Evenness of coverage is an important metric to understand how efficient a protocol is, as a more even coverage means that more regions are well covered with a smaller amount of raw data needed. Fold80 measures the width of the coverage distribution (figure 1), its optimal value is 1.0 (all regions covered equally), and good enrichment protocols reach values of 1.1-1.3. Finally, completeness is the most important metric for clinical diagnostics, as incomplete coverage means that some regions cannot be evaluated, leading to reduced sensitivity of the test.

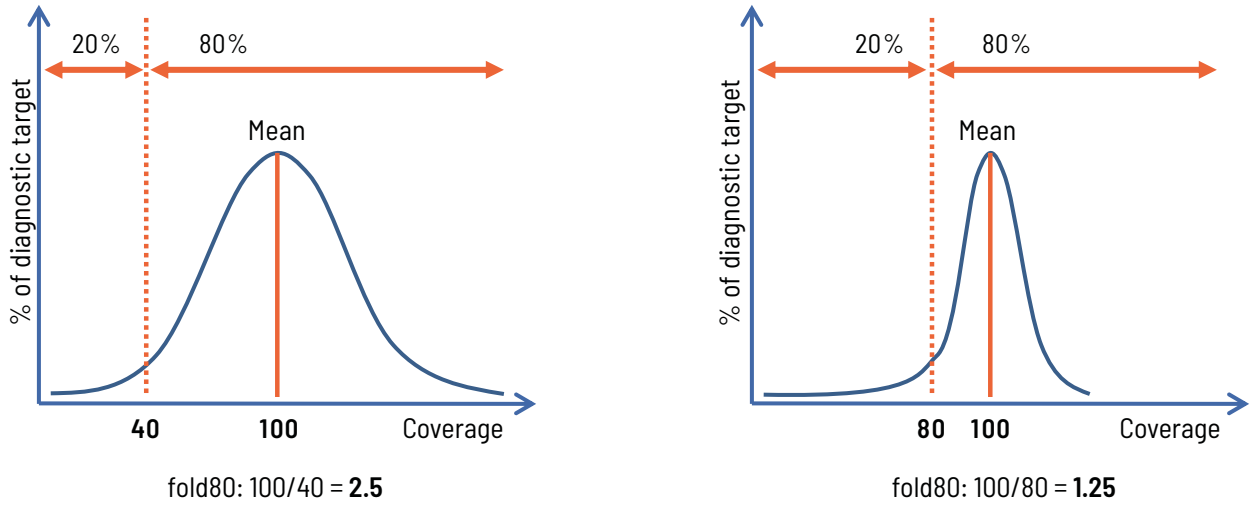


Figure 1: *Evenness of coverage can be evaluated by the fold80 measure which represents the amount of additional sequencing needed to have 80% of all targets covered at the currently observed mean. It is computed as the mean coverage divided by the 20th percentile. Smaller values indicate tighter coverage distributions. Left, large fold80 values correspond to a wide distribution and uneven coverage; Right, small values correspond to a narrow distribution and even coverage.*

In theory, WGS provides coverage of the whole genome, although some regions with extreme GC contents or highly repetitive sequences (e.g., telomeres, centromeres, and long repeat regions) are lost during library preparation. Similarly, WES can theoretically enrich all targeted regions from the input DNA, but some targets show poorer performance in practice. However, even if the whole genome was represented in the sequenced reads, these might not be mappable to the genome reference, leading to gaps in the coverage. For example, reads may be too short to be mapped to ambiguous regions, or the reference genome sequence may represent variable regions in a way incompatible with the data generated from a given sample.

For clinical diagnostics, filters must be applied to ensure mapped reads accurately represent a patient's genome. The first step is to remove reads that do not align uniquely to one genomic locus, i.e., more than one alignment with an equal score exists. Such reads are likely to derive from repetitive regions or pseudogenes. They can lead to distorted allele frequencies, false positives, or missed variants in these regions if kept.

Secondly, we remove duplicate reads, as those can also distort variant allele frequencies and, by falsely increasing the observed coverage, can result in a region appearing to be covered sufficiently while, in truth, all data for this region only stems from a few amplified molecules.

Thirdly, we remove all reads with very low mapping quality (q<15). While these reads are uniquely mapped, they map so poorly to their assigned position that they may represent sequences that are not present in the reference genome. Their poor mapping could contribute to false-positive variant calls, reduce the observed allele frequencies of real variants, or (like duplicates) lead to incorrectly high coverages.

Finally, we remove overlapping paired read ends that occur when the sequencing length (e.g., 2x150) is larger than the insert size. Much like read duplicates, overlapping sequences from the same insert do not contribute additional information and must be only counted once to compute coverage correctly. The resulting alignment provides clinically useable coverage and is a good basis for downstream analyses such as variant calling for clinical evaluation.

Applying these necessary filtering steps, both WES and WGS data cover a similar percentage of the coding targets (98.1% and 97.8%, respectively, table 1), with a small advantage for the WES datasets. While deeply sequenced WGS naturally covers a larger number of deep intronic noncoding disease-causing variants than WES, it is surprising that 40.8% of these can still be analyzed using an off-the-shelf exome kit. CeGaT's ExomeXtra® was designed to enrich these regions and comes closer to WGS performance but does not quite reach it. Some of the remaining differences can be explained by longer WGS read length and insert size, which allows for better mappability.

Most known disease-causing variants are within coding regions (85%) or close to CDS borders (10%), but more and more relevant intergenic or deep-intronic variants are described. At the same time, novel variants at intergenic or deep-intronic positions can rarely be clinically assessed. Customized exome enrichments can include deep-intronic, UTR, and intergenic regions, focusing only on known disease-causing mutations.

Our conclusion from this data is that most underrepresented regions are affected by the necessary filtering of mapped reads and might benefit from the better mappability of much longer reads beyond what current second-generation sequencing technologies allow. For most genetic diagnostic questions, WES is as good as WGS at a greatly reduced price, as optimized WES kits such as CeGaT's ExomeXtra® cover the vast majority of regions with disease-causing variants.

Table 1: *Sequencing parameters and target regions covered (%) to at least 30x by filtered alignments for whole genome (WGS), standard whole exome (WES), and CeGaT's optimized exome (CEX) data. CDS, coding sequences; DI, deep-intronic disease-causing variants (see methods for details); fold80, evenness of coverage (smaller numbers indicate more even coverage).*

Dataset	Read length	Insert size	Total GB sequenced	CDS	DI	Fold80
WGS	2x150bp	272	350	97.5	98.8	1.27
WES	2x100bp	260	12	98.0	45.4	1.43
CEX	2x100bp	245	17.5	98.1	97.8	1.31

Methods

All analyses started from high-quality DNA. Exomes were enriched using Twist Core+RefSeq exome or CeGaT's ExomeXtra® version 6, respectively. Sequencing of both WGS and WES samples was performed using Illumina instruments. Read length was 2x150 for WGS and 2x100 for WES. Reads were aligned using bwa-mem2 (2.1, WES samples) or Illumina DRAGEN (4.2.4, WGS samples) against the GRCh38 reference with masked pseudoautosomal regions on chromosome Y. Deduplication and filtering as well as coverage calculations were performed using proprietary software. For the coding targets, the MANE Select database (v1.3) was used, containing 191,616 regions spanning 33.4 Mbp. The set of disease-causing deep-intronic (DI) variants was extracted from HGMD Professional (release 24.2) and ClinVar (downloaded 2024-07-15), and filtered to retain only non-coding variants more than 30 bp distant from any known CDS, resulting in 5,796 variants. The data shown are means of three different samples.

A Fair Comparison

In this analysis, we used a deeply sequenced WGS dataset (100x) as a benchmark and show that commercially available exomes and CeGaT's ExomeXtra® come close to providing the same coverage of diagnostically relevant regions.

Due to cost constraints, diagnostic WGS is typically performed at an average coverage of 30x. A fair comparison of current approaches needs to take this into account. Using the same standard as before, 30x diagnostic on-target coverage, we show that CeGaT's ExomeXtra® outperforms WGS with respect to diagnostic targets covered (table 2). Obviously, a WGS dataset with an average coverage of 30x cannot provide 30x coverage on all targets, as coverage is not perfectly uniform. We thus also applied a relaxed threshold of 20x for comparison.

Again, we can show that CeGaT's ExomeXtra® covers more relevant regions at higher coverage and thus delivers higher sensitivity than 30x WGS. This boosts solution rates and even allows us to detect mosaicism, which is systematically missed in most WGS analyses. Note that the advantage shown by WGS in terms of covering deep-intronic disease-causing variants (as shown in table 1) is lost when considering a more realistic amount of raw data, with 53.3% covered at 30x, compared to 97.8% for CeGaT ExomeXtra®. At the same time, thousands of irrelevant variant calls usually obtained by WGS analyses are avoided, improving diagnostic speed and accuracy. Including a whole genome backbone brings CNV detection from CeGaT ExomeXtra® to the same level as array CGH or WGS analyses.

Table 2: *Average coverage and target regions covered (%) to at least 20x/30x by filtered alignments for whole genome (WGS), standard whole exome (WES), and CeGaT's optimized exome (CEX) data.*

	WES	WGS	ExomeXtra®
Total GB sequenced	12	100	17.5
Insert size	260	370	245
Average diagnostic coverage	~120x	~30x	~115x
≥20x coverage			
Coding sequences	98.2	91.4	98.3
Deep-intronic disease-causing variants	50.0	89.4	98.7
≥30x coverage			
Coding sequences	98.0	54.9	98.1
Deep-intronic disease-causing variants	45.4	53.3	97.8

Conclusion

Whole exome sequencing with CeGaT's ExomeXtra® outperforms WGS and standard whole exome in terms of coverage of medically relevant regions.

Want to Know More?

For more information visit our website
www.cegat.com/single-exome-diagnostics

All Achievements of Genetic Testing Integrated into Our Exome Diagnostics

Thus far, we have covered details on how we optimized our wet-lab approach, our enrichment, and how these improvements benefit the clinical outcome. However, this is only the first of many steps involved in achieving comprehensive genetic diagnostics. The next steps are crucial to gather all the many pieces that need to be put together to solve the diagnostic puzzle.

Bioinformatic analysis is the second step in genetic diagnostics. The aim is to extract as much clinically relevant information as possible from the sequencing data, while maximizing the reduction of variants for manual interpretation. In other words, we collect all necessary puzzle pieces. By constant optimization of our in-house bioinformatic pipeline we ensure a stringent but complete list of potential causative variants is prepared for our diagnosticians. Comparative exome diagnostics, especially Trio ExomeXtra®, offers additional insights that help to identify causative variants in a patient. For Trio ExomeXtra®, we account for variants in genes with reduced penetrance, variable expressivity, and also imprinting effects (IVERP). In trio analysis, we additionally include the detection of uniparental hetero- and isodisomies (UPD), while for singleton cases, isodisomies are detected.

Additional efforts, for all CeGaT ExomeXtra® diagnostics include checking for relevant SNV/CNV combinations, detecting mosaic variants (as low as 5% NAF), and screening for repeat expansions related to reported phenotypes.

Medical interpretation, the third and final step, is crucial to ensure that all the insights gathered through our superior enrichment and bioinformatics are combined with the best human know-how. Our diagnostic team consists of experts with many years of clinical experience, including medical doctors specializing in human genetics. This interdisciplinary team thoroughly investigates the relevant findings related to the patient's phenotype, using the most recent literature for data interpretation, and evaluating the variants in accordance with ACMG guidelines.

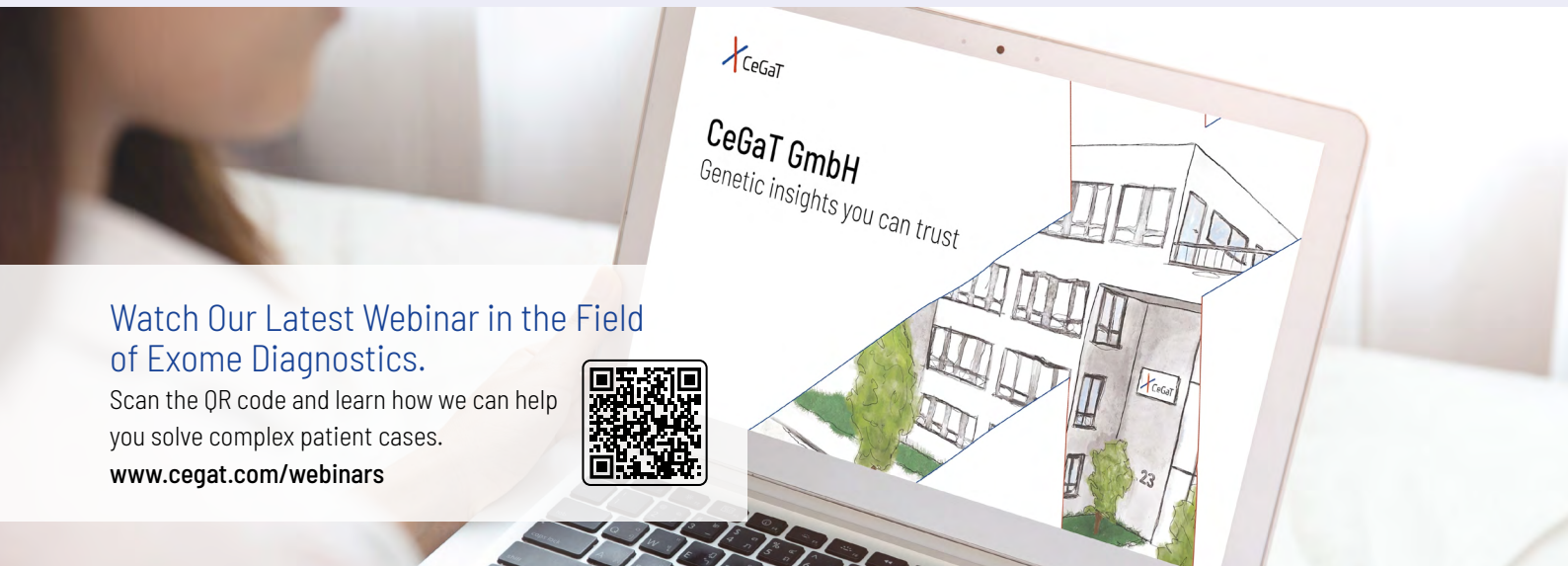
The reported variants are described from a molecular genetics and clinical perspective, including addressing the disease association as well as its genetic relevance in respect to patient's family. The variant classification is visualized using an ACMG/ACGS classification table, which provides a transparent overview on the criteria used and points assigned.

The puzzle pieces are put together and the puzzle is solved. The final report is revised by medical doctors and genetic experts to maximize clinical utility. When applicable, further recommendations are also provided, such as further testing, and potential therapeutic approaches based on the reported causative variants.

In conclusion, CeGaT ExomeXtra® Diagnostics goes way beyond regular exome diagnostics approaches and offers higher sensitivity than genome-based diagnostics. It is optimized for patients with complex, heterogeneous, and unspecific symptoms and offers optimal diagnostics for prenatal cases. In-house design and processing from beginning to end ensures the highest standard of quality at every step of the way, delivering the best service to identify the genetic cause of the diseases for our patients.

Reliable Support at Every Step of the Process

Our diagnostic support team is ready to assist you with any questions you may have and to discuss the best possible approach for the genetic investigation of your patients. Please don't hesitate to reach out to us, either via email at **diagnostic-support@cegat.com** or via phone at **+49 7071 56544-220**. We are looking forward to hearing from you!



Watch Our Latest Webinar in the Field of Exome Diagnostics.

Scan the QR code and learn how we can help you solve complex patient cases.
www.cegat.com/webinars

About Us

CeGaT is a global provider of genetic analyses for a wide range of medical, research, and pharmaceutical applications.

Founded in 2009 in Tübingen, Germany, the company combines state-of-the-art sequencing technology with medical expertise – with the aim of identifying the genetic causes of diseases and supporting patient care. For researchers and pharmaceutical companies, CeGaT offers a broad portfolio of sequencing services and tumor analyses. CeGaT generates the data basis for clinical studies and medical innovations and drives science forward with its own insights.

The owner-managed company stands for independence, comprehensive personal customer service, and outstanding quality. CeGaT's laboratory is accredited according to CAP/CLIA, DIN EN ISO 15189, and DIN EN ISO/IEC 17025 and thus meets the highest international standards. To obtain first-class results, all processes are carried out in-house under scientific and medical supervision.

CeGaT GmbH
Genetic Diagnostics
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-220
Fax: +49 7071 56544-56
Email: diagnostic-support@cegat.com
Web: www.cegat.com/diagnostics

DAkkS Deutsche Akkreditierungsstelle D-ML-13206-01-00

Accredited by DAkkS according to DIN EN ISO 15189:2014

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS

CLIA CERTIFIED ID: 99D2130225

2024/11