Tech Note

The Best Possible Exome

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. In our analysis, we find 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 6 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 deep-sequenced WGS dataset (average coverage 133x) 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 used a combined database of all known coding sequences from CCDS and all known disease-causing noncoding mutations from HGMD. 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 gold standard, our results can help researchers understand the varying claims by sequencing providers with respect to the strengths and limitations of WES analyses.

For WES analyses, we show data both from a commercially available exome kit as well as from CeGaT's proprietary ExomeXtra®. CeGaT ExomeXtra® is based on Twist's Core exome, adding Twist's RefSeq spike-in to cover further relevant genes and transcripts. It is augmented by adding in (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 which is based on CCDS, Gencode, Ensembl, and RefSeq curated, and (6) regions with pharmacogenetically relevant variants in selected genes.

We evaluate three metrics: Average coverage, evenness of coverage (fold 80), and completeness of coverage (%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 (see 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.

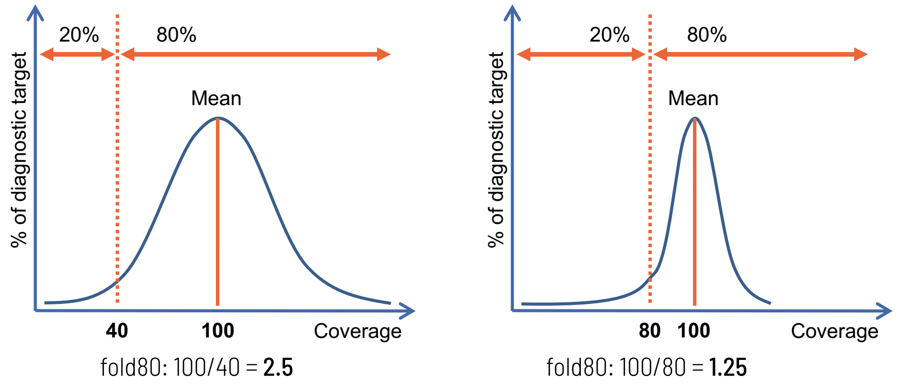

Figure 1: *Evenness of coverage can be evaluated by the fold80 measure which represents the amount of additional sequencing needed to have 80% of all targets covered at the currently observed mean. It is computed as the mean coverage divided by the 20th percentile. Smaller values indicate tighter coverage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

Even without any filtering applied, the WGS dataset does not cover the entire coding sequence and regions of interest (see table 1), and 0.23% of the target region remains uncovered. This is most likely due to limitations of the sequencing technology as reads may be too short to be mapped to ambiguous regions, as well as the reference sequence used, which may represent variable regions in a way incompatible with the data generated from a given sample. Regardless of the reasons, this number provides the upper limit for an optimal exome.

As our data shows, the exome data comes very close to this threshold with 99.5% of the target regions covered, leaving 0.5% uncovered. In addition, 94.4% of known disease-causing non-coding variants are covered in the exome dataset. We also show that a well-balanced exome has a coverage distribution almost as even as a WGS dataset.

Dataset	Read length	Insert size	Total GB sequenced	CDS	DM	Fold80
WGS	2x150bp	272	504	99.77	99.90	1.27
WES	2x100bp	203	15	98.53	92.08	1.37
CEX	2x100bp	191	15	99.48	99.52	1.54

Table 1: *Target regions covered (%) to at least 30x by unfiltered alignments for whole genome (WGS), standard whole-exome (WES), and CeGaT's optimized exome (CEX) data. CDS, coding sequences; DM, disease-causing non-coding mutations (see methods for details); fold80, evenness of coverage (smaller numbers indicate more even coverage).*

For clinical diagnostics, some filters must be applied to ensure mapped reads accurately represent a patient's genome. We apply these filters step by step to show their impact on covered regions (see table 2). First, we remove reads that do not align uniquely to one genomic locus, i.e., there exists more than one alignment with an equal score. Such reads are likely to derive from repetitive regions or pseudogenes. If they are kept, they can lead to distorted allele frequencies, false positive, or missed variants in these regions.

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 that have 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 which are not present in the reference genome. Their poor mapping could contribute 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., 2x100) is largerthan 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 correctly compute coverage. The resulting alignment is a good basis for downstream analyses such as variant calling for clinical evaluation. More information can be found in our Tech Note „Choosing the right read lenght for diagnostic sequencing".

Applying these necessary filtering steps, both WES and WGS data cover a similar percentage of target (96.3%/97.5% and 98.0%, respectively). Some of the advantage of WGS may stem from larger insert sizes and longer read lengths that facilitate mapping in some difficult regions. Another part of the explanation is that not all CDS considered in this analysis are enriched in WES due to differences in the curation of relevant transcripts between exome manufacturers and our own transcript database used here.

Unsurprisingly, WGS covers a larger number of noncoding disease-causing mutations even though the difference is not major (90.7% for WES vs 99.6% for WGS) and almost disappears when using an optimized diagnostic exome as CeGaT's ExomeXtra® (99.0% vs 99.6%), the remaining difference being explained by longer WGS read length and insert size.

Dataset	Average Coverage			CDS			DM		
	WGS.	WES	CeGaT	WGS.	WES	CeGaT	WGS.	WES	CeGaT
raw	133	167	174	99.77	98.53	99.48	99.90	92.08	99.52
uniquely mapping	129	160	163	98.08	96.79	97.72	99.65	91.81	99.29
no duplicates	121	133	132	98.05	96.53	97.60	99.63	91.09	99.19
high quality	121	133	132	98.02	96.50	97.58	99.63	91.08	99.19
non-overlapping	111	118	111	98.00	96.31	97.50	99.62	90.73	99.02

Table 2: *Impact of different alignment filtering steps on target regions covered (%) to at least 30x by whole-genome sequencing, an off-the-shelf exome, and CeGaT ExomeXtra® . CDS, coding sequences; DM, disease-causing non-coding mutations (see methods for details).*

By the current state of clinical knowledge, most known disease-causing mutations are within coding regions (89%) or adjacent splice regions, while novel variants at intergenic or deep-intronic positions can rarely be clinically assessed. Customized exome enrichments as well as disease-specific panel designs can include deep-intronic, UTR, and intergenic regions to increase sensitivity for known disease-causing mutations.

Our conclusion from this data is that the vast majority of underrepresented regions are due to 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 current WES kits cover the vast majority of regions with disease-causing mutations.

Methods: *All analyses started from high-quality DNA. Exomes were enriched using Twist Core+RefSeq exome or the CeGaT ExomeXtra® , respectively. Sequencing of both WGS and WES samples was performed using Illumina NovaSeq 6000 instruments. Read length was 2x150 for WGS and 2x100 for WES.*

Reads were aligned using bwa mem (0.7.17-r1188) against the hg19 reference with masked pseudoautosomal regions on chromosome Y. Alignment, deduplication and filtering as well as coverage calculations were performed using proprietary software. For the coding targets, a curated database based on CCDS (r22), Gencode (31), Ensembl (98) and RefSeq curated (10.2019) was used, containing 198,427 regions spanning 34.3 Mb.

The set of disease-causing ("DM") variants was extracted from HGMD (release 19.3, 169,593 variants), and filtered against the CDS set to retain only non-coding variants, resulting in 17,782 variants. Data shown are means of 3 different samples.

Conclusion:

Even an extraordinary deeply sequenced genome (133x) does not significantly outperform CeGaT ExomeXtra® in terms of covering medically relevant regions.

A Fair Comparison

In this analysis, we use a deep WGS dataset (133x) as benchmark and show that commercially available exomes and CeGaT 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 ExomeXtra® vastly outperforms WGS with respect to diagnostic targets covered (table 3).

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 to both datasets.

Again we can show that CeGaT 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.

At the same time, thousands of irrelevant variant calls usually obtained by WGS analyses are avoided, improving diagnostic speed and accuracy.

	WGS	CeGaT Exome Xtra
Average diagnostic coverage	~30x	~110x
Total GB sequenced	95	15
Coding sequences covered >20x	87.7%	97.8%
Disease-causing non-coding mutations covered >20x	89.8%	99.4%
Coding sequences covered >30x	27.1%	97.5%
Disease-causing non-coding mutations covered >30x	27.2%	99.0%

Table 3: *The numbers refer to high quality reads (uniquely mapping, removal of duplicates and overlapping reads).*

CeGaT

About Us

CeGaT GmbH is a leading global provider of genetic diagnostics and mutation-related disease analyses. The company combines its next-generation sequencing (NGS) process and analysis pipelines with its medical expertise – dedicated to identifying the genetic cause of disease and supporting patient management.

Genetic mutations can trigger a wide range of diseases, from epilepsy to Parkinson's. Through the use of NGS, it is possible to analyze all genes associated with a disease phenotype simultaneously – both fast and effectively. An interdisciplinary team of scientists and physicians evaluates the data and summarizes the findings in a comprehensive medical report. All services are performed in-house.

CeGaT, founded in 2009 and based in Tübingen, Germany, is accredited according to CAP, CLIA and DIN EN ISO 15189:2014.

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

Phone: +49707156544-333
Fax: +49707156544-56
Email: sales@cegat.com

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