White Paper

Streamline whole exome sequencing with a robust, flexible and cost-effective workflow

Markus Storbeck[1], Brian Dugan[2], Dan Heard[2], Cornelia Mechlen[1], Eric Lader[2] and Peter Hahn[1]

[1] QIAGEN GmbH, Hilden, Germany; [2] QIAGEN Sciences Inc. Frederick, MD, USA

Introduction

Comprehensive next-generation sequencing (NGS) approaches such as whole genome sequencing (WGS) and whole exome sequencing (WES) provide us with the context missing from target enrichment technologies like gene panels, which focus on specific disease groups with clear etiology. Although proven highly effective in detecting disease-causing mutations[1,2], targeted gene panels often fail to pinpoint causative variants. Inherited diseases with unknown genetic origins, ambiguous etiology or difficult differential diagnosis require a more comprehensive and unbiased approach to identify disease-causing mutations[3] or to associate genes with a disease[4].

WES is the most cost-effective comprehensive NGS approach as it covers the more disease-significant part of the human genome – the protein-coding regions or exons. Despite targeting the coding regions of approximately 20,000 human genes, the targeted region comprises only about 1–2% of the human genome[5,6]. In comparison to WGS, WES targets meaningfully interpretable regions, which effectively helps to increase sequencing depth and sample throughput while minimizing sequencing cost.

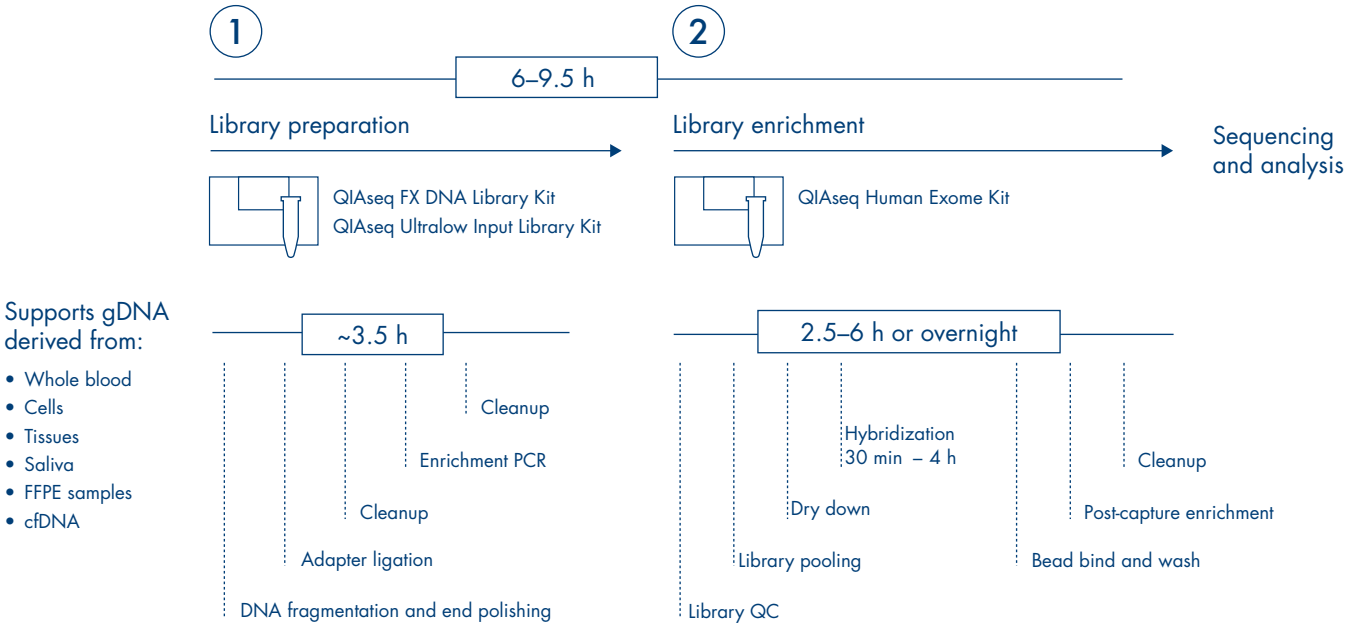
WES workflows are typically complex and time-consuming and the data quality is determined by probe design, capture efficiency and capture specificity. Here we introduce a robust, streamlined and highly flexible workflow (Figure 1) for the enrichment of human coding sequences from indexed whole genome libraries using the new QIAseq® Human Whole Exome Kit on Illumina® instruments. We further show performance metrics of WES from genomic DNA (gDNA) and cell-free DNA (cfDNA) using the QIAseq Human Whole Exome Kit benchmarked against competitor enrichment solutions on the market.

Materials and methods

Input DNA: NA12878 benchmarking DNA was obtained from the Coriell Institute. cfDNA was isolated from plasma of healthy donors using the QIAamp® MinElute™ ccfDNA Midi Kit following the standard protocol. Both gDNA and cfDNA were quantified using the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific).

Library preparation and hybridization capture: The QIAseq FX DNA Library UDI Kit was used to generate indexed whole genome libraries from 50 ng NA12878 gDNA according to handbook protocols. For library generation from 20 ng of cfDNA, the QIAseq Ultralow Input Library Kit was used. Indexed whole genome libraries were quantified using the Agilent® High Sensitivity DNA Kit for the Agilent 2100 Bioanalyzer (Agilent Technologies). Hybridization capture was performed using the QIAseq Human Exome Kit according to handbook protocols, carried out for 4 hours as 8-plex with a total library input of 8 x 400 ng. Competitor library preparation solutions (if provided) and hybridization capture kits were used according to the manufacturer's recommendations. Finalized whole exome libraries were quantified by qPCR using the QIAseq Library Quant Assay Kit on a QIAquant® 384 instrument.

▷



Step 1: Generation of indexed libraries	Step 2: Enrichment of human exonic sequences
Whole genome libraries are prepared in the first step prior to whole exome enrichment. Libraries can be prepared from a variety of DNA analytes as shown above. The QIAseq FX DNA Library Kit enables the processing of gDNA or formalin-compromised DNA in a single-tube reaction for enzymatic DNA fragmentation, end-repair and adapter ligation. Alternatively, the QIAseq Ultralow Input Kit can be used to process physically sheared DNA. The QIAseq Ultralow Input Library Kit enables the generation of libraries from small amounts of cfDNA that are compatible for exome enrichment. These library kits use QIAGEN's unique dual-indexing adapters allowing multiplexed hybridization capture and sequencing.	During whole exome enrichment by hybridization capture, indexed library fragments are bound to biotinylated double-stranded DNA capture probes with highly flexible hybridization times – ranging from 30 minutes to overnight incubation. Bound fragments are immobilized on streptavidin beads and non-targeted fragments are washed away. Enriched library fragments are amplified using a proprietary post-capture amplification mix to generate sequencing-ready libraries.

Figure 1. Workflow principle of the QIAseq Human Exome Kit.

Sequencing: Paired-end sequencing (2 x 150 bp) was performed on an Illumina NextSeq® 550 instrument using the NextSeq 500/550 High Output Kit v2.5 (300 cycles). Demultiplexing and FASTQ file generation were performed on the sequencing instrument.

Data analysis: Sequencing data were analyzed using QIAGEN CLC Genomics Workbench 20 and the Biomedical Genomics Analysis plugin. Data were analyzed using the default QIAGEN CLC workflow for the QIAseq Human Exome Kit. Data obtained from competitor products were analyzed using identical workflow settings while adapting the respective reference genome and target region definitions.

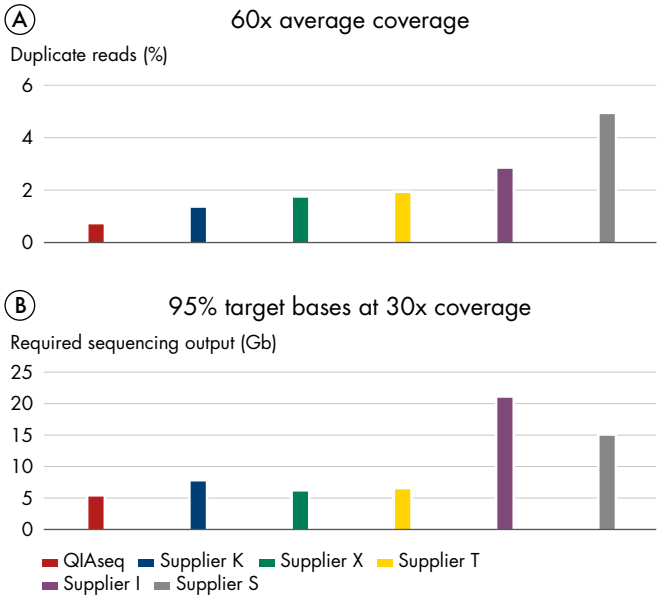
Read sampling: To allow a valid comparison between data obtained from QIAseq Human Exome and competitor products, sequencing reads were randomly sampled from each dataset. This resulted in a predefined average coverage of 60x or 100x, irrespective of the target region size. All data presented in this white paper was obtained through experiments conducted by QIAGEN R&D in Hilden.

Results and discussion

The QIAseq Human Exome yields highly complex libraries with minimal sequencing requirement

Whole exome sequencing reads were randomly sampled to yield either 60x average coverage or to assure at least 30x coverage of 95% of target bases. Read duplication analysis (Figure 2 A) showed <2% duplication levels for most of the whole exome solutions included in this study. Among these, the QIAseq Human Exome showed the least

percentage of duplicate reads (0.7%) indicating that the whole genome library generated using the QIAseq FX DNA Library Kit was highly complex and that library complexity remained unbiased during hybridization capture. The target enrichment technology used in the QIAseq Human Exome Kit uses double-stranded DNA probes which, in contrast to single-stranded DNA or RNA probes, enables capture of both strands of the input library. This helps generate highly complex libraries even when using small amounts of DNA input.



C

Supplier	ROI size (Mb)
QIAseq	33.94
Supplier K	42.99
Supplier X	34.16
Supplier T	36.7
Supplier I	45.33
Supplier S	49.67

Figure 2. Duplicate reads and required sequencing output. A Duplicate reads were calculated as the percentage of all reads required to obtain an average coverage of 60x. **B** The required sequencing output (in Gb) sufficient to cover 95% of target bases at 30x. **C** ROI sizes in megabases (Mb).

The size of the region of interest (ROI size) and uniformity of coverage determine the amount of sequencing data required to achieve the desired minimal coverage. We determined the amount of sequencing data in gigabases (Gb) required to cover 95% of all target bases at 30x or more. The QIAseq Human Exome required only 5.3 Gb of sequence data to yield 30x coverage for 95% of target bases (Figure 2 B). Expectedly, exome panels with larger target regions required more data. For example, an exome panel (Supplier S), which is 50% larger than QIAseq Human Exome Kit, required approximately 3-fold higher sequencing data output to achieve a similar coverage (Figures 2 B–C). However, if one out of two equal-sized panels requires more sequencing data to achieve a similar coverage, this would likely be due to worse uniformity.

The QIAseq Human Exome delivers comprehensive target coverage with exceptional uniformity

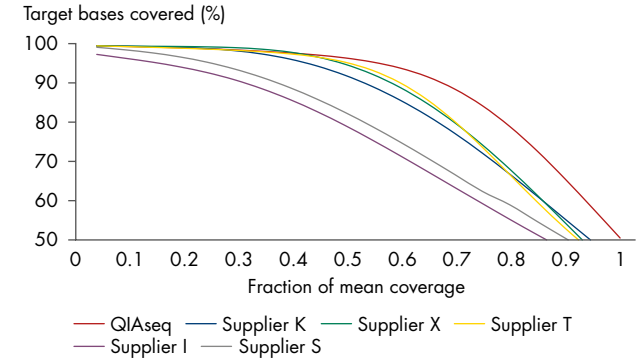
Coverage uniformity describes how much the actual target region coverage fluctuates around its mean value. High coverage uniformity means most target bases are covered at a sequencing depth very similar to the average value, whereas, weak coverage uniformity means there are regions covered either much less or much more than the average. This causes either incomplete target coverage or wasting of sequencing data. Therefore, high coverage uniformity is pivotal for uninterrupted detection of ROIs and minimizing sequencing costs.

The QIAseq Human Exome Kit delivers high coverage uniformity with up to 80% of target bases covered at 0.8-fold of the mean coverage (Figure 3 A). This is reflected in the fold-80 base penalty (<1.3) of the QIAseq Kit (Figure 3 B). Accordingly, the low average coverage of only 60x is sufficient to cover >98% of the target region with at least 20x or ~90% of the target region with at least 40x (Figure 3 C). This enables reliable and comprehensive variant calling throughout the target region.

The QIAseq Human Exome offers unbiased target enrichment – also for difficult regions

The GC-content of the human exome varies greatly. Hence, to achieve superior coverage uniformity, GC-related biases should be minimized throughout library preparation and hybrid capture. In particular, PCR amplification and probe-to-target hybridization performances may vary with ▷

Ⓐ



Ⓑ

Supplier	Fold-80 base penalty
QIAseq	1.259
Supplier K	1.383
Supplier X	1.324
Supplier T	1.314
Supplier I	1.724
Supplier S	1.648

Ⓒ

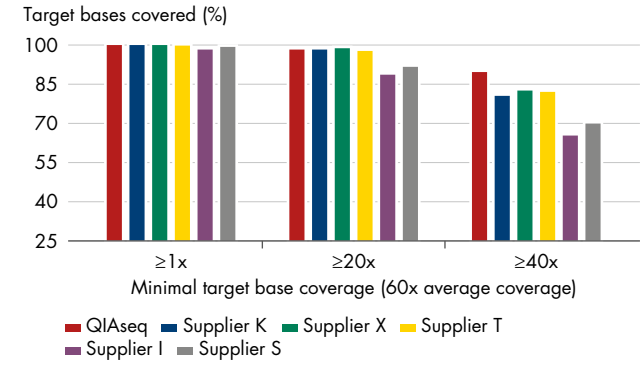


Figure 3. Target region coverage uniformity. A Base coverage plot showing percentage of target bases (y-axis) covered at any given fraction of the average coverage (x-axis). **B** The fold-80 base penalty describes coverage uniformity as a numeric value (smaller is better). A given sequencing output needs to be increased by the fold-80 value to lift 80% of target bases to the current average coverage. **C** Minimal base coverage plot for the QIAseq Human Exome and others at 60x average coverage.

GC content. We analyzed the number of sequencing reads per target versus target GC content (Figure 4). Density plots indicate that the QIAseq Human Exome and the solution from Supplier T show almost unbiased coverage, independent of GC content. Both kits, for the majority of target regions (dark blue areas), show highly homogeneous coverage levels across the ~35–70% GC range. Other suppliers show either a strong generalized variation in coverage (Supplier X), GC-based variation in coverage (Supplier K), or a combination of both (Suppliers I and S). These results largely match our findings on coverage uniformity (Figure 3 A).

These findings underscore the importance of GC-independent target enrichment for strong coverage uniformity as well as the superior capture performance of double-stranded DNA probes (used in the QIAseq Human Exome Kit) over single-stranded probes.

Can these differences be due to variations in the library preparation method? The QIAseq FX DNA Library Kit was used to generate the indexed whole genome libraries for the QIAseq Human Exome workflow as well as in the workflows of Suppliers T, X and K. Thus, differences in GC-dependent coverage come from the respective hybrid capture technology. For Suppliers I and S, library preparation methods proprietary to the supplier were used. Thus, for these solutions, GC-dependent differences in coverage could arise from the library preparation method and/or hybrid capture technology.

Why is GC-content relevant for whole exome sequencing? Insufficient enrichment of target regions with low or very high GC-content may lead to failure in variant detection for a multitude of disease-relevant genes. High GC-content targets can dramatically reduce coverage or cause highly heterogeneous coverage levels. An example is the single-exon *ADAT3* gene (Figure 5), mutations in which are associated with autosomal recessive development of intellectual disability (MIM615286). Due to its average GC-content being far above 70%, probe-to-target hybridization is strongly impaired and may result in loss of coverage (Supplier I) or highly heterogeneous coverage (Supplier S). The QIAseq Human Exome and the solution by Supplier K provided homogeneous coverage throughout the target, very close to the mean coverage of 60x. Even though a single target is not representative of the whole capture panel, this example is in line with the finding that the QIAseq Human Exome provides highly uniform target coverage that is largely independent of GC-content (Figure 4).

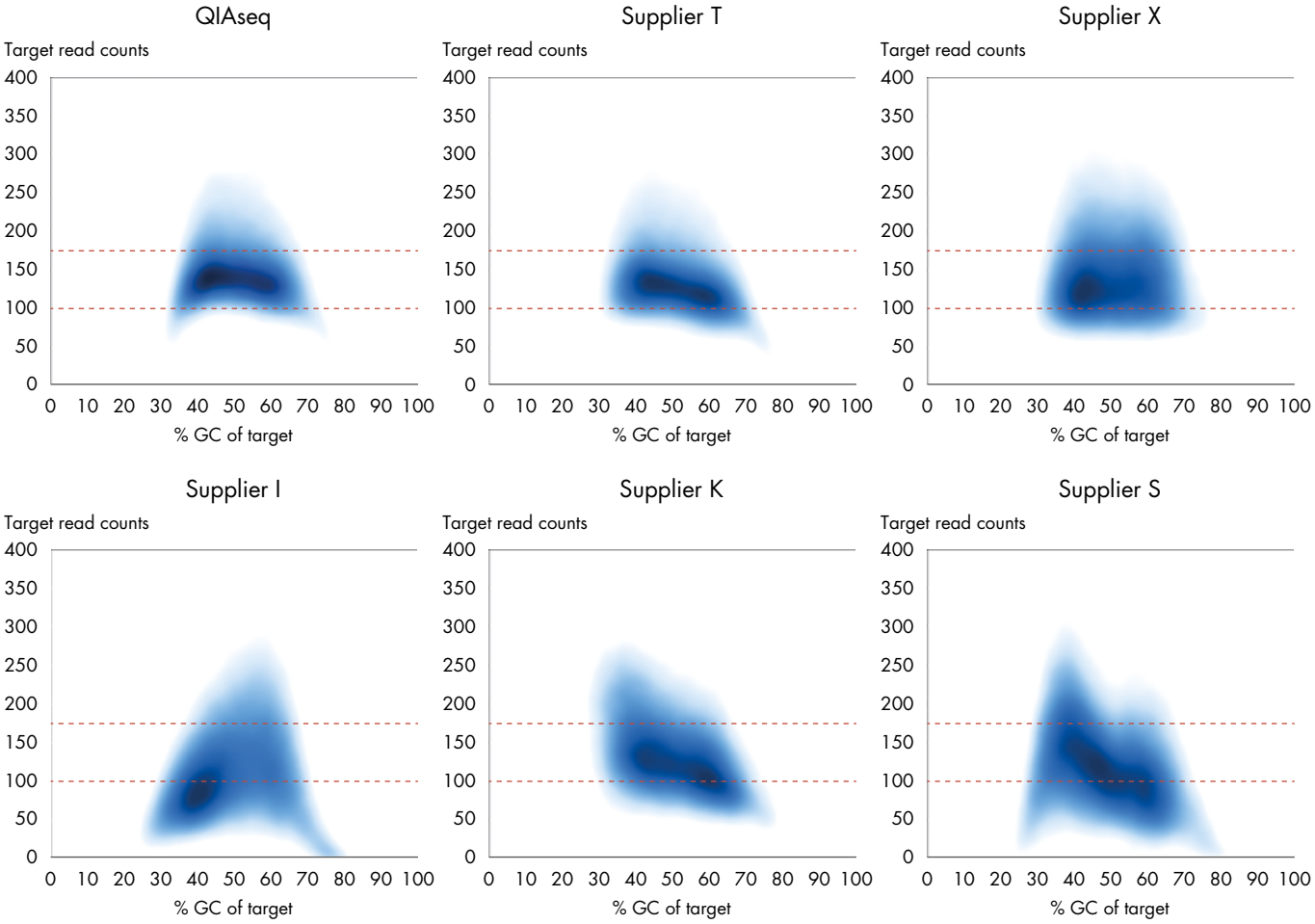


Figure 4. Target coverage vs. GC content. Density scatter-plots of all target regions of the respective exome panel. Figures show on-target read counts in dependence of target GC-content as smoothed color density representations of normal scatterplots. Colors illustrate local densities at each point in a scatterplot, while dark colors indicate a high number of data points. Single values are not displayed. Targets with a read count of zero are excluded.

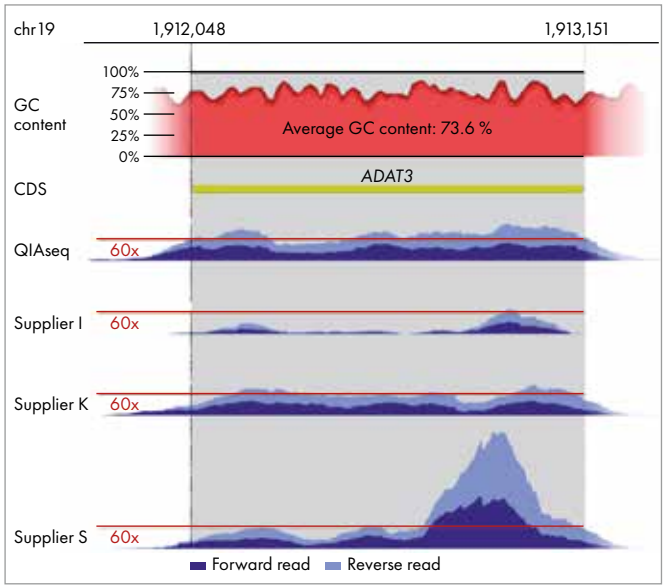


Figure 5. GC-content and coverage of the *ADAT3* gene. Whole exome read mappings with an average coverage of 60x for the respective target region of the QIAseq Human Exome and solutions of other suppliers. The genome browser view shows the *ADAT3* locus on chromosome 19 (CDS, yellow track) and the local GC%-levels (smoothed, 25 bp window). Coverage tracks (blue) represent the local coverage across the *ADAT3* locus. The expected coverage is shown by the 60x markers (red). Coverage levels far below and above the expected coverage indicate poor target enrichment uniformity.

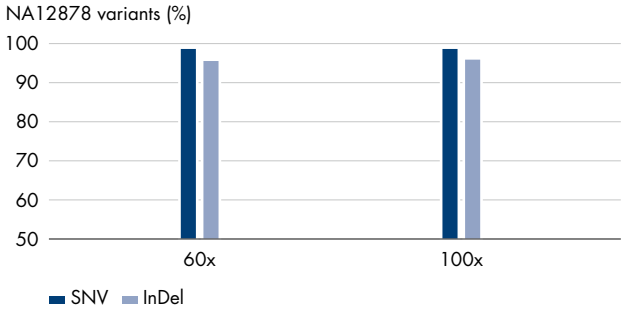
Robust detection of variants: QIAseq Human Exome covers most HGMD® disease-related mutations

To assess variant detection capabilities of the QIAseq Human Exome Kit, we compared the variants called from 60x and 100x read mappings with the high confidence variant list of the Coriell NA12878 control DNA (Figure 6 A). NA12878 variants outside of the target region were not considered. At only 60x average coverage, 98.9% of SNVs and 95.7% of InDels were detected indicating that the QIAseq Human Exome workflow allows robust variant calling throughout the target region. Strikingly, increasing the average coverage to 100x did only slightly enhance variant calling – to 99.0% for SNVs and 96.1% for InDels. Due to the exceptionally high coverage uniformity, average coverage of just 60x is sufficient for comprehensive variant calling. Sequencing at average coverages of 100x or more is not required, making the QIAseq Human Exome Kit a highly cost-efficient solution for human exome sequencing.

The Human Gene Mutation Database (HGMD Professional, QIAGEN Digital Insights), currently lists ~8000 genes and >300,000 disease-causing or disease-related mutations. Whole exome sequencing aims to screen as many disease-related loci as possible. The QIAseq Human Exome target region comprises 95.93% of disease-related HGMD loci (Figure 6 B). Still, due to the efficient design of the kit, the overall target region is reduced to 33.9 Mb. This makes the QIAseq Human Exome Kit one of the most compact whole exome solutions on the market while still covering the major fraction of disease-related HGMD loci. A more comprehensive enrichment of disease-relevant loci is only achieved by solutions with much larger target region size (Suppliers I, K), and requires considerably higher sequencing capacity.

We consider a local coverage of 20x to be reliable enough to call heterozygous germline variants with very high certainty and even lower sequencing depths (15x) have been proposed[7]. Apart from comparing overlap of loci with the target region, we also analyzed which fraction of HGMD loci are sufficiently enriched to allow variant calling at 20x coverage. At 60x average target coverage, 91.5% of disease-relevant HGMD loci were covered ≥20x using the QIAseq Human Exome Kit (Figure 6 C). Only Supplier K, with a 9.1 Mb (27%) larger target region size than QIAseq, covered 1.0% more HGMD loci than the QIAseq Human Exome. Solutions from suppliers with an ROI size similar to QIAseq detected 87.0% (Supplier X) and 86.6% (Supplier T) of HGMD loci. Even exome panels with very large ▷

Ⓐ



Ⓑ

Supplier	ROI size (Mb)	HGMD disease mutations (%)
QIAseq	33.9	95.93
Supplier T	36.7	89.73
Supplier X	34.2	89.62
Supplier I	45.3	98.67
Supplier K	43.0	96.72
Supplier S	49.7	92.52

Ⓒ

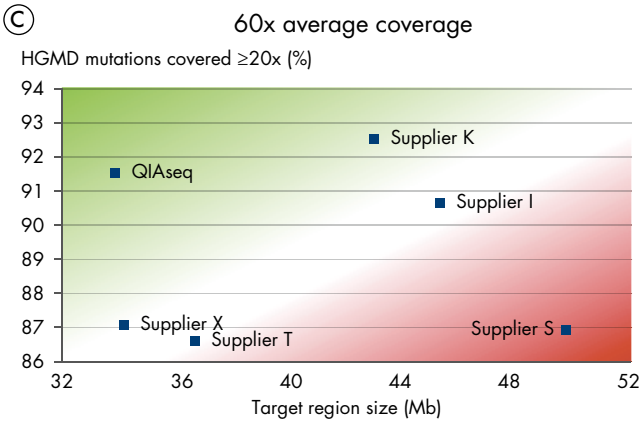


Figure 6. Detection of variants by exome sequencing. A Identification of expected SNVs and InDels in Coriell NA12878 control DNA at 60x and 100x average exome coverage using the QIAseq Human Exome Kit. **B** Comparison of target region sizes (ROI size) and fraction of disease-causing mutations listed in HGMD (2020.4) that overlap with the region of interest. **C** Percentage of disease-related HGMD loci with an actual coverage ≥20x (at 60x average coverage). Green areas indicate detectability of large fractions of HGMD variants despite small overall target region size. Red areas indicate large target region sizes and low detectability of disease-related HGMD variants.

ROI sizes of 45 Mb or more (Suppliers I, S) detected fewer (≥20x) HGMD loci than the QIAseq Human Exome. Thus, the QIAseq Human Exome Kit, with its efficient design, compact target region and superior coverage uniformity, detects the majority of disease-relevant loci with very low sequencing effort (60x) at greatly reduced sequencing cost per sample.

The QIAseq Human Exome allows unrestricted target enrichment from small amounts of cfDNA

The QIAseq Human Exome Kit is a universal solution that enables target enrichment not only from high-quality gDNA but also from challenging samples such as formalin-compromised DNA and cfDNA. Here we used 20 ng cfDNA from plasma to generate indexed whole genome libraries using the QIAseq Ultralow Input Library Kit followed by hybridization capture using the QIAseq Human Exome Kit and 60x sequencing (Figure 7).

Despite the reduced input amount and the challenging nature of cfDNA, the resulting data were of adequate quality and comparable to data obtained from gDNA. Given the decreased input, duplicate reads increased moderately to 2.9%. Also, the sequencing effort required to cover 95% of target bases increased moderately (~35%) to 7.5 Gb at 30x and 5.0 Gb at 20x. The base coverage and uniformity of 60x cfDNA exomes were only slightly reduced compared to gDNA; cfDNA libraries still achieved 96.6% of target bases covered at ≥20x (98.1% for gDNA) and a fold-80 base penalty of 1.39 (1.26 for gDNA). ▷

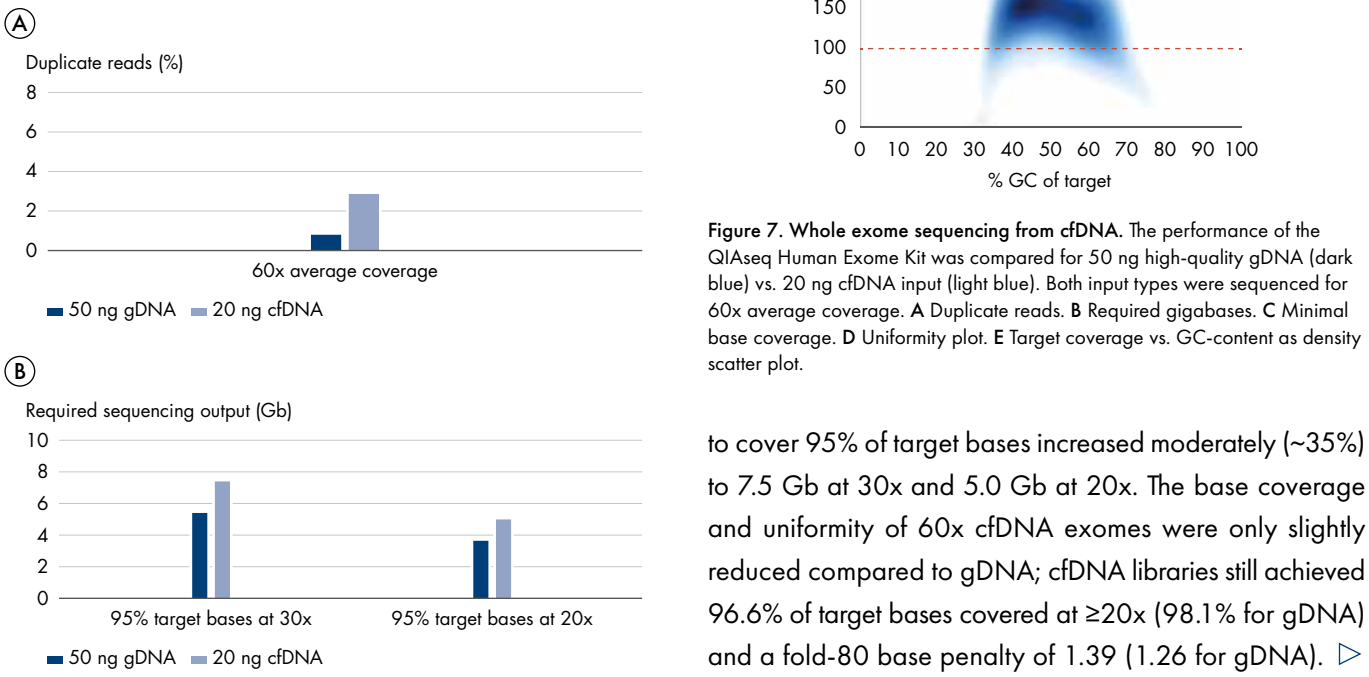


Figure 7. Whole exome sequencing from cfDNA. The performance of the QIAseq Human Exome Kit was compared for 50 ng high-quality gDNA (dark blue) vs. 20 ng cfDNA input (light blue). Both input types were sequenced for 60x average coverage. **A** Duplicate reads. **B** Required gigabases. **C** Minimal base coverage. **D** Uniformity plot. **E** Target coverage vs. GC-content as density scatter plot.

This indicates a similarly efficient and uniform target enrichment from cfDNA libraries compared to gDNA libraries.

GC-dependent coverage of cfDNA libraries was highly similar to gDNA (Figure 4; QIAseq) with only minimal drops at extremely low or extremely high GC levels (Figure 7 E). Our data indicate that hybrid-capture-based enrichment of human exonic regions is feasible from cfDNA. Given higher sequencing depth and sufficient cfDNA input, the QIAseq Human Exome Kit delivers excellent data quality that may enable cfDNA-based applications like the detection of somatic mutations at moderately low allelic frequencies.

Conclusions

In this white paper, we compared various performance metrics to benchmark the new QIAseq Human Exome Kit for gDNA and cfDNA libraries against currently available competitor hybridization capture solutions. We successfully generated high-quality whole exome libraries from gDNA using the QIAseq FX DNA Library UDI Kit and the QIAseq Human Exome Kit. QIAseq libraries were highly complex with best-in-class coverage uniformity and were almost free of GC bias, which allowed comprehensive variant calling even at 60x average coverage. Our focused target region design includes up to 95.9% of disease-related HGMD loci – all within one of the most compact whole exome panels in the market. The streamlined workflow – including variable hybridization time and reduced sequencing requirements of less than 6 Gb per exome with 95% target bases covered at ≥30x depth – saves time and lowers sequencing cost. Taken together, these features make the QIAseq Human Exome kit a highly efficient screening tool for human inherited disease and the discovery of novel disease-causing variants.

References

1. Sun, Y. et al. (2019). BMC Med. Genomics. **12**(1):76.
2. Cortese, A. et al. (2020). Neurology. **94**(1):e51–e61.
3. Karakaya, M. et al. (2018). Hum. Mutat. **39**(9):1284–1298.
4. Li, J. et al. (2019). Gene. **700**:168–175.
5. Ng, S.B. et al. (2009). Nature. **461**(7261):272–6.
6. Petersen, B.S., Fredrich, B., Hoeppner, M.P., Ellinghaus, D., and Franke, A. (2017). BMC Genet. **18**(1):14.
7. Song, K., Li, L. and Zhang, G. (2016). Sci. Rep. **6**:35736.

To learn more about simplifying your exome sequencing workflow, visit:
www.qiagen.com/exome-sequencing

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at www.qiagen.com or can be requested from QIAGEN Technical Services or your local distributor.