Application Note

Optimizing the Human Whole Exome Sequencing Process

Agilent's streamlined workflow delivers higher precision, less noise and reduced sequencing cost

Authors

Raouf Ben Abdelali,
Laboratoire CERBA, Saint-Ouen L'Aumône, France

Jérôme Audoux,
SeqOne, Montpellier, France

Detlef Trost,
Laboratoire CERBA, Saint-Ouen L'Aumône, France

Anissa Zouaoui,
SeqOne, Montpellier, France

Aicha Boughalem,
Laboratoire CERBA, Saint-Ouen L'Aumône, France

Ramdane Mallek,
Laboratoire CERBA, Saint-Ouen L'Aumône, France

Adrien Jeanniard,
Agilent Technologies

Nicolas Philippe
SeqOne, Montpellier, France

Roubila Meziani,
Agilent Technologies

Jean-Marc Costa,
Laboratoire CERBA, Saint-Ouen L'Aumône, France

Summary

The Laboratoire Cerba (Cerba) has been working with Agilent and SeqOne to optimize its Whole Exome Sequencing (WES) process. The objective of this collaboration was to develop a single, robust method that can accommodate a wide variety of sample types (blood, fresh frozen tissue and FFPE), sample qualities and DNA input amounts (10–200 ng), while providing an efficient pipeline to analyze sequencing data in hereditary and cancer genetics applications. The use of the Agilent SureSelect human all exon V7 combined with the SureSelect^{XT} low input reagent kits and SureSelect XT HS and XT low input enzymatic fragmentation kit meets all of these objectives. This workflow also delivers higher precision, less noise, and reduced sequencing and interpretation costs.

Introduction

Whole Exome Sequencing (WES) has become an essential tool for health care providers and clinical research laboratories using Next-Generation Sequencing (NGS). WES is a powerful tool for the identification of genetic variations involved in human diseases, notably in the detection of point mutations and copy number variations. Although NGS has been in use for over 15 years, library preparation reagents and sequencing technologies are constantly evolving to offer increased performance and flexibility in terms of content, sequencing requirements, and turnaround time.

Created in 1967, Cerba provides private and hospital laboratories worldwide (over 50 countries) with a wide range of specialized clinical pathology tests in oncology, allergology, toxicology, hormonology, infectious and metabolic diseases. They also provide analyses for the diagnosis of hereditary genetic diseases and cancer genomics. With the support of highly qualified experts and sophisticated equipment, Cerba's unique perspectives, built from decades of scientific expertise and innovation, help their stakeholders identify new approaches and anticipate tomorrow's healthcare challenges. Cerba's scientific and medical board continually monitors advances in analytical methods from a scientific and technical perspective to identify and integrate the latest research advances into Cerba's operations. Following the announcement of the latest Agilent human all exon V7, the SureSelect^{XT} low input reagent kits and the SureSelect XT HS and XT low input enzymatic fragmentation kit (DNA shearing, library prep, QC, sequencing, etc.), Cerba decided to benchmark the new solutions to assess their WES performance. To do this benchmarking they compared the new protocols with their historical in-house custom protocol based on the Agilent SureSelect clinical research exome V2.

Materials and methods

To enable a direct comparison of the workflows, the same set of eight genomic DNA samples were prepared using each of the three capture protocols described as follows.

All samples were sequenced on the same instrument with the same protocol. The final optimized protocols were then field tested on a range of samples to ensure the robustness of the workflow. All analyses were performed using the SeqOne genomics interpretation platform.

Library preparation and sequencing

Sample preparation

Eight genomic DNA samples were tested: one Agilent female reference DNA *(part number 5190-8850)*, one Coriell DNA *(NA12878)*, and six constitutional DNA samples extracted from blood (five with intellectual disability and one fetus with polymalformation).

Total blood was collected in EDTA tubes. DNA was extracted using the Qiagen FlexiGene DNA kit, following the manufacturer's protocol. All samples were evaluated on a LabChip GX from Perkin Elmer with the HT DNA High Sensitivity Reagent Kit + HT DNA chip, following the manufacturer's protocol. Quantification was performed on a Qubit 2.0 with the Qubit BR dsDNA assay kit.

Two library preparation methods were used:

- Cerba modified protocol: Cerba used NEBNext dsDNA Fragmentase to shear input DNA and prepared libraries with NEBNext Ultra II DNA Library Prep Kit for Illumina. Libraries were then enriched following the SureSelect^{QXT} target enrichment for Illumina multiplexed sequencing protocol: https://www.agilent.com/cs/library/usermanuals/Public/G9681-90000.pdf.
- SureSelect^{XT} low input protocol coupled with upfront enzymatic shearing using the SureSelect XT HS and XT low input enzymatic fragmentation kit (SureSelect XT HS/LI ENZ): with the protocols outlined in: https:// www.agilent.com/cs/library/usermanuals/public/G9703-90000.pdf and https://www.agilent.com/cs/library/usermanuals/public/G9702-90050.pdf.

Captured procedures were tested in three combinations with these two library preparation protocols:

- Cerba modified protocol with SureSelect clinical research exome V2 (CRE V2)
- Cerba modified protocol with SureSelect human all exon V7 (Exome V7)
- SureSelect XT HS/LI ENZ with SureSelect human all exon V7 (Exome V7)

All PCR was performed on the Veriti 96-well Thermal Cycler (Thermo Fisher). The final prepared libraries were checked on LabChip GX (Perkin Elmer) and quantified on a Qubit 2.0 (Invitrogen). Each pool of DNA was clustered using 8.5 pmol and sequenced with 100 bp paired end reads on an Illumina Hiseq1500 in rapid run mode (60 GB).

SureSelect target enrichment probe design

SureSelect clinical research exome V2 (CRE V2) utilizes the SureSelect human all exon V6 as its core design with boosted coverage in disease-associated regions. This enables more comprehensive coverage of highly curated databases, thus facilitating more precise variant calling within these regions. This design consists of targets identified in databases such as the Online Mendelian Inheritance in Man (OMIM), the Human Genome Mutation Database (HGMD) and NCBI's ClinVar, along with additional disease-relevant regions defined by Emory University and the Children's Hospital of Philadelphia. The SureSelect clinical research exome V2 is the only exome kit on the market that comes with a list of included genes and their evidence of disease relevance (target size: 67 Mb).

SureSelect human all exon V7 (exome V7) is a new exome kit that maximizes coverage for a given sequencing depth.

Designed using GRCh38/hg38 genome assembly, the Exome V7 targets the protein coding regions documented in the latest versions of RefSeq, GENCODE, CCDS, and UCSC known genes, including hard-to-capture exons that are omitted from other commercial exome kits. Furthermore, Exome V7 targets all pathogenic variants in the genes included in the ACMG guidelines for secondary findings. A novel primer design algorithm results in an efficient design with a total size of only 48.2 Mb.

All Agilent designs (BED files) can be downloaded freely from Agilent's SureDesign website: https://earray.chem.agilent.com/suredesign/.

Comparison of the three protocols is based on the sequencing depth and the coverage of the human reference genome (RefSeq: NCBI Reference Sequence Database).

Coverage analysis

SeqOne is a genomic interpretation platform that provides a complete, end-to-end solution for genomic analysis in routine clinical environments.

Mapping and duplicate removal

Reads from each sample were aligned to the hg19 human genome using BWA-MEM (0.7.15-r1140) and converted to BAM files using SAMtools (v1.7). Duplicated reads were marked using Picard-tool's MarkDuplicate software. The resulting duplicate metrics were extracted from the MarkDuplicate summary file, while alignment statistics were computed with Picard-tool's Alignment Summary using output BAM files from MarkDuplicate.

For more simplicity and consistency, we aligned all data to hg19. CRE V2 baits are designed based on hg19 but Exome V7 baits are designed based on hg38. We need to note that there are minor differences in global metrics when analysis of V7 is performed based on hg38 versus hg19.

RefSeq intervals

Comparisons of the coverage and depth of different kits and protocols were made on coding exons of RefSeq transcripts. The interval BED file was constructed from the RefGene file downloaded from UCSC: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz. Only exons of mRNA transcripts (identifiers starting with NM_) were selected and only whole or partial exons included in the CDS were retained. Padding of 20 bp was applied to each end of the exons and chromosome Y intervals were removed.

Coverage and depth

Coverage, depth, and on targets metrics were collected with Picard-tool's CollectHSMetrics tool. The tool was run with parameters PROBES_INTERVALS and TARGET_INTERVALS set to the RefSeq interval. The NEAR_DISTANCE parameter was set to 100 to compute on-targets reads with +/-100. Depth and coverage statistics were directly extracted from CollectHSMetrics output file.

Percentage of on-target reads were computed as *ON_PROBES_BASES / PF_BASES_ALIGNED* percentage of +/-100 on-target reads were computed as *ON_PROBES_BASES + OFF_BAIT_BASES) / PF_BASES_ALIGNED.*

Results

SureSelect^{XT} low input protocol coupled with SureSelect XT HS and XT low input enzymatic fragmentation kit with Exome V7 displayed superior performance compared to the Cerba modified protocol with CRE V2 or Exome V7.

Table 1. Global metrics. Average of sequencing metrics for eight libraries (six blood samples, one Agilent female reference DNA and one HapMap DNA as detailed in Materials and Methods) per protocol.

Protocol/Average metrics	Average # of reads (million)	Average % of aligned reads	Average % on target +/-100	Average of median depth (X) (RefSeq)
Cerba modified protocol with Exome CRE V2	88	98.9%	93.5%	74.13
Cerba modified protocol with Exome V7	82	99.1%	95.5%	78.63
SureSelect XT HS/LI ENZ with Exome V7	87	99.5%	97%	84.63

All three runs (described in Table 1) have similar sequencing output, with 82 to 88 million reads per sample and a consistent average aligned reads percentage among protocols. The average of all eight libraries for each sequencing metric and the results are shown in Table 1. The median depth increased when Exome V7 is used instead of the CRE V2 (78.63X to 74.13X, respectively) due to the size of the target (48.2 Mb versus 67 Mb). The percentage of on target +/- 100 also increased (95.1 to 95.5%, respectively). The new SureSelect human all exon V7 targets the protein-coding regions from the latest version of RefSeq, including hard-to-capture exons. The novel probe design algorithm results in an efficient, streamlined design, which enables the detection of these regions. When comparing SureSelect XT HS/LI ENZ to the Cerba-modified protocol, using Exome V7, capture performance (on target +/- 100) improved even further from 95.5 to 97.1% while median coverage depth increased from 78.63X to 84.63X (Table 1). For all samples, the SureSelect XT HS/LI ENZ protocol with Exome V7 shows an improvement in coverage of RefSeq coding exons compared to the Cerba modified protocol with both Exome CRE V2 and Exome V7 at 10X and 30X (Figure 1a and 1b). The base coverage of the SureSelect XT HS/LI ENZ protocol is more consistent across replicates compared to the Cerba modified protocol. More strikingly, the 30X base coverage of the SureSelect XT HS/LI ENZ protocol is increased by 5% compared to the Cerba modified protocol with Exome V7 and by 7% compared to the Cerba modified protocol with CRE V2 (Figure 2). Higher coverage of the SureSelect XT HS/LI ENZ protocol is also evident at 40X and 50X (Figure 2).

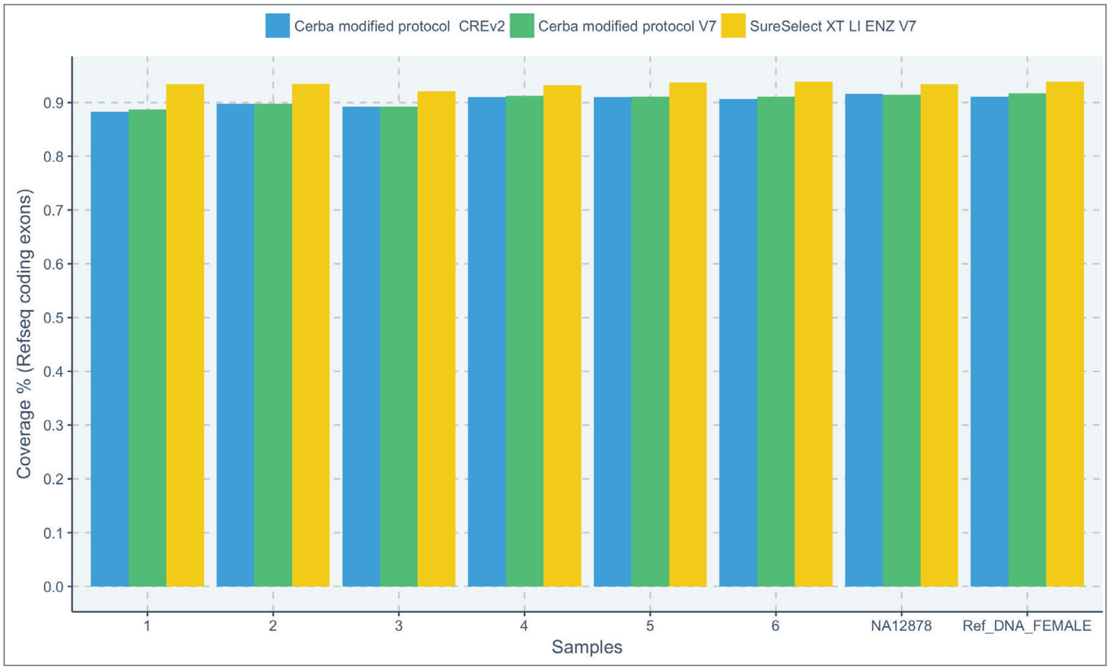

Figure 1a. 10X base coverage of RefSeq coding exons per sample. Six blood samples (numbered 1-6), one Coriell DNA (NA12878), and one Agilent female reference DNA (part number 5190-8850) were tested in three sequencing runs. SureSelect XT HS/LI ENZ with Exome V7 (yellow) shows the improvement in coverage of RefSeq coding exons compared to the Cerba modified protocol with both Exome CRE V2 (blue) and Exome V7 (green) at 10X.

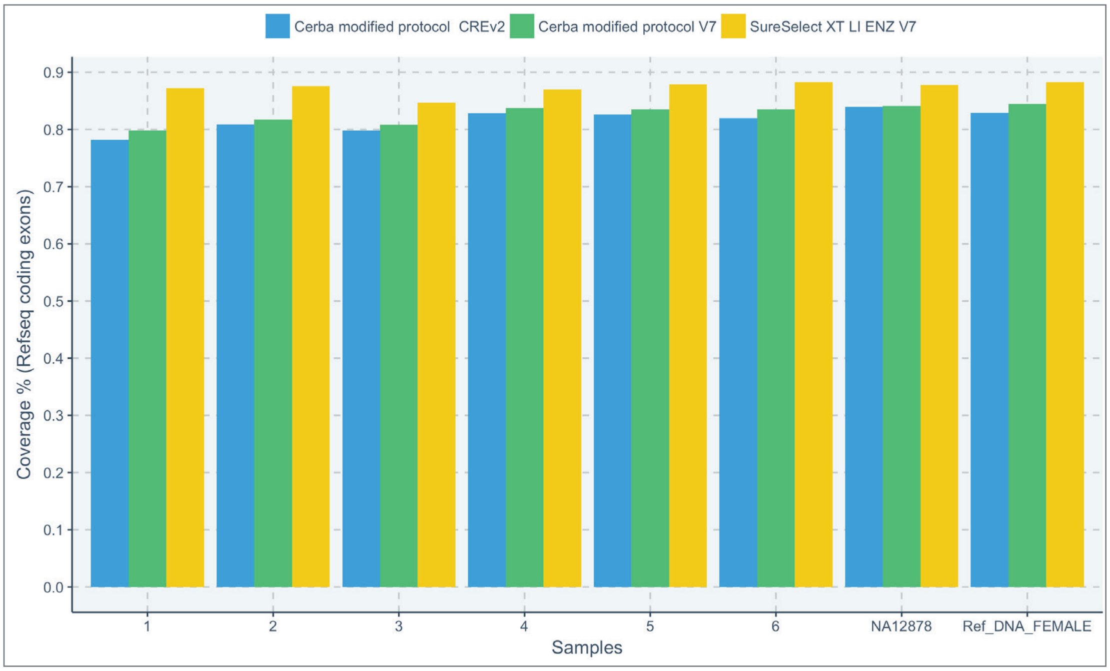

Figure 1b. 30X base coverage of RefSeq coding exons per sample. Six blood samples (numbered 1-6), one Coriell DNA (NA12878), and one Agilent female reference DNA (part number 5190-8850) were tested in three sequencing runs. SureSelect XT HS/LI ENZ with Exome V7 (yellow) shows improvement in coverage of RefSeq coding exons compared to the Cerba modified protocol with both Exome CRE V2 (blue) and Exome V7 (green) at 30X.

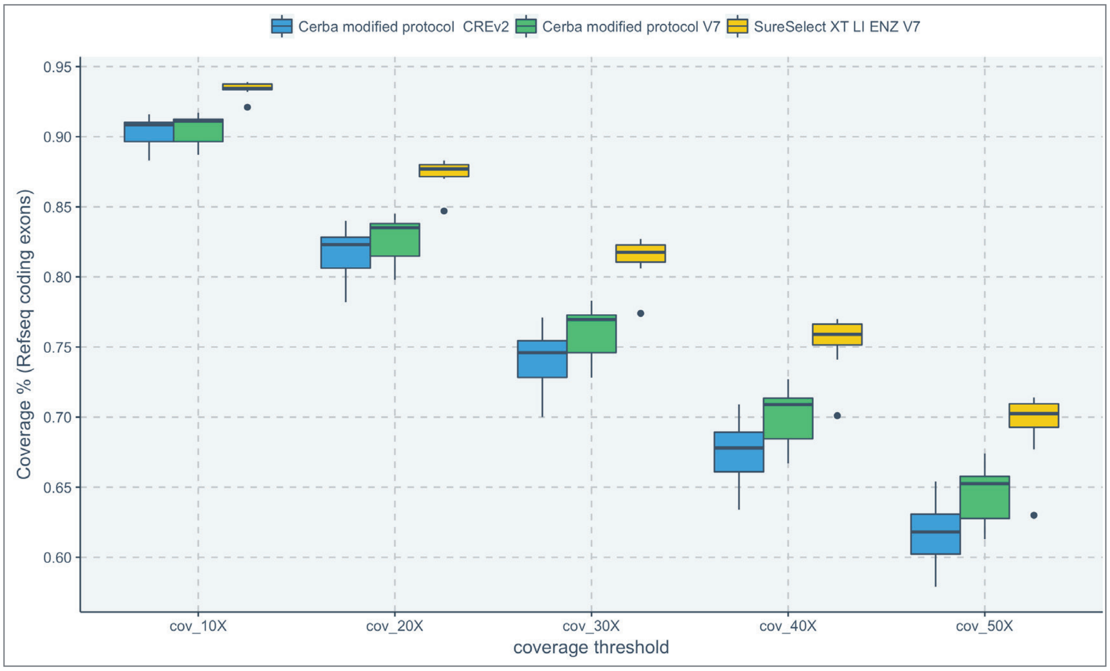

Figure 2. Box plot of coverage of RefSeq coding exons for each protocol (all eight libraries detailed in the materials and methods are averaged for each sequencing metric). The base coverage of the SureSelect XT HS/LI ENZ workflow (yellow) is more consistent across replicates compared to the Cerba modified protocols (green and blue). More strikingly, 30X base coverage of the SureSelect XT HS/LI ENZ protocol increased by 5% compared to the Cerba modified protocol with Exome V7 and by 7% compared to the Cerba modified protocol with CRE V2. Higher coverage of the SureSelect XT HS/LI ENZ protocol is also evident at 40X and 50X for all samples.

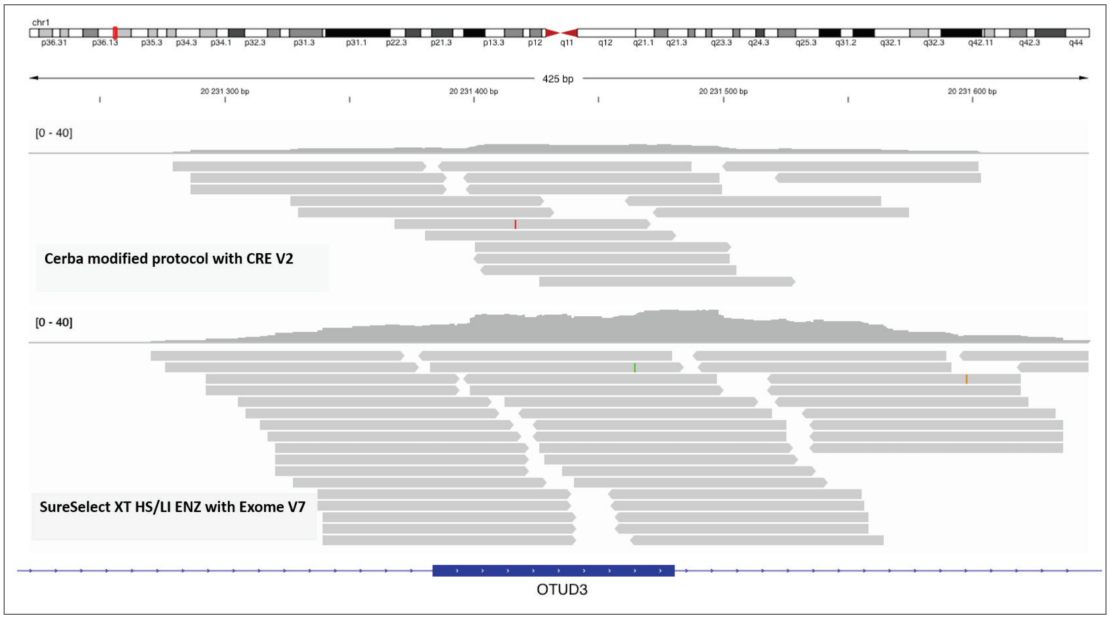

Figure 3a. Read depth coverage of OTUD3 exon 6/8. The results shown are from blood sample 1. SureSelect XT HS/LI ENZ with Exome V7 shows better coverage uniformity compared to Cerba modified protocol using CRE V2.

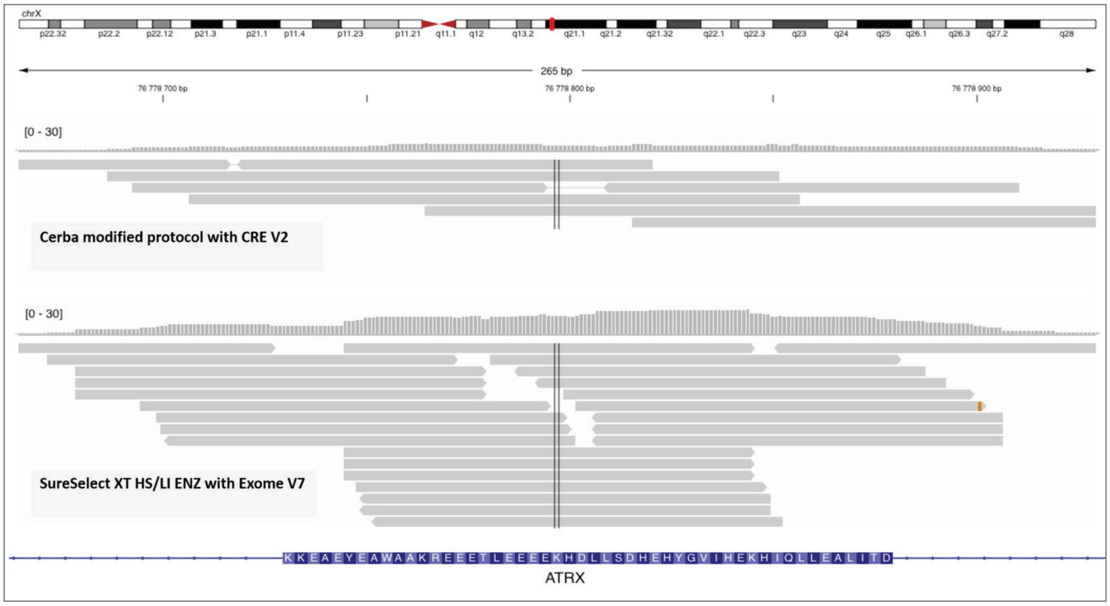
chrX
265 bp
76 778 700 bp
76 778 800 bp
76 778 900 bp
[0 - 30]
Cerba modified protocol with CRE V2
[0 - 30]
SureSelect XT HS/LI ENZ with Exome V7
K K E A E Y E A W A A K R E E E T L E E E E K H D L L S D H E H Y G V I H E K H I Q L L E A L I T D
ATRX

Figure 3b. Coverage of ATRX exon 30/34. The results shown are from blood sample 1. SureSelect XT HS/LI ENZ with Exome V7 shows better coverage uniformity compared to Cerba modified protocol using CRE V2.

At Laboratoire Cerba, WES is used to identify genetic variations involved in constitutional diseases like intellectual deficiency. Thus, we compared several important genes for constitutional diseases (OTUD3 and ATRX) in CRE V2 and Exome V7. We observed better coverage and uniformity for SureSelect XT HS/LI ENZ with Exome V7. Two examples are illustrated in (Figure 3a and 3b).

SureSelect^{XT} low input protocol coupled with SureSelect XT HS and XT low input enzymatic fragmentation kit workflow is compatible with 10 ng and 200 ng DNA input.

To evaluate the impact of the initial DNA input amount on the performance of SureSelect XT HS/LI ENZ with Exome V7, we prepared three libraries at 10 ng and three libraries at 200 ng input (one blood sample and one Agilent female reference DNA) (Figure 4).

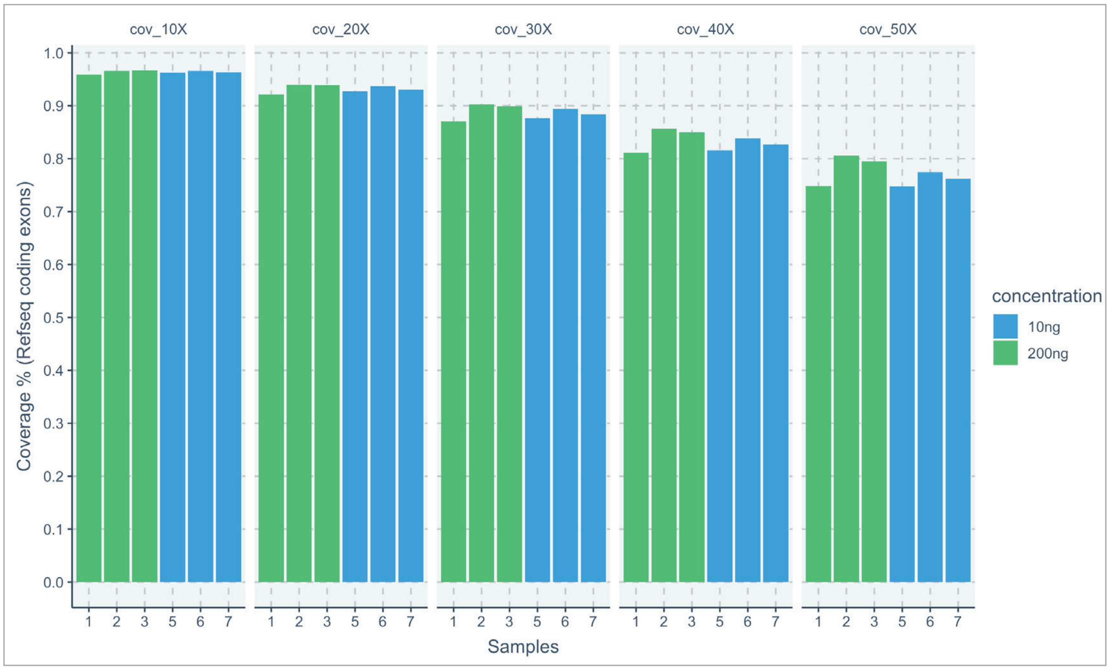

Figure 4. Coverage of RefSeq coding exons using 10 ng (blue) and 200 ng (green) DNA as input with SureSelect XT HS/LI ENZ protocol with Exome V7. One blood sample (numbered 1, 2, and 5), and one Agilent female reference DNA (part number 5190-8850) (numbered 3, 6, and 7) were tested in a single sequencing run. The base coverage (~95% at 10x to ~75% at 50x) is not significantly different between the libraries prepared with 10 ng and 200 ng of DNA input.

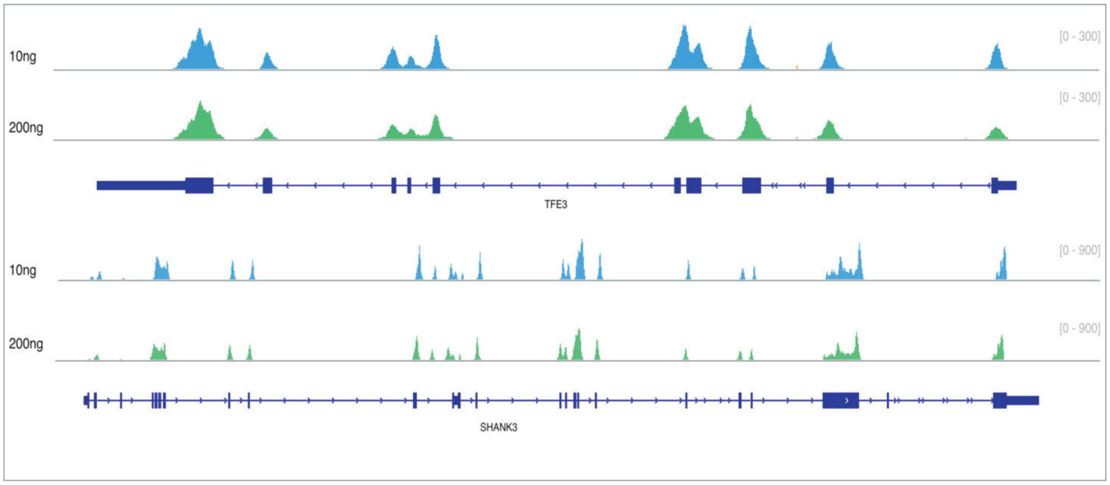

Figure 5. Read coverage on two example genes (TFE3 and SHANK3) using 10 ng (blue) and 200 ng (green) DNA as input with SureSelect XT HS/LI ENZ protocol with Exome V7. One blood sample was used as DNA input. Two example genes, TFE3 and SHANK3, show similar coverage uniformity and read depth with 10 ng (blue) and 200 ng (green) DNA input. TFE3 scale is 300 bp and SHANK3 scale is 900 bp as shown on the right in the graph.

Sequencing depth varied between 95 and 115 million reads across the six libraries tested. After sequencing, we plotted the base coverage in Figure 4. The coverage results (~95% at 10X to ~75% at 50X) don't show significant differences between the libraries prepared with 10 ng and 200 ng of DNA, suggesting that as low as 10 ng of DNA input can be used to provide similar coverage (Figure 4). Furthermore, the uniformity and read depth are similar between 10 ng and 200 ng of DNA input, as shown in Figure 5 with example genes SHANK3 and TFE3.

SureSelect^{XT} low input protocol coupled with SureSelect XT HS and XT low input enzymatic fragmentation kit workflow is compatible with the Agilent NGS automation platforms.

To evaluate the impact of automation on the performance of SureSelect XT HS/LI ENZ workflow, we prepared seven libraries with 200 ng input on the Agilent bravo NGS (option A, part number G5541A) for pre-PCR steps and Agilent bravo NGS workstation (option B, part number G5522A) for post-PCR steps (same DNA samples as detailed in materials and methods).

The pre-and post-capture yields are similar between samples (data not shown). For the seven libraries tested, the sequencing depth varied between 102 and 110 million reads. One sample (NA12878) had a lower number of reads at 64 million, likely due to a quantification error during the pooling process. The base coverage is highly reproducible (Figure 6). For all seven samples, the base coverage was at least 96% at 10X, and more than 80% at 50X (except for the NA12878 which is due to lower raw read depth), which are concordant with manually–prepared libraries shown in Figure 4. These results show that Agilent NGS automation platforms can generate high-quality libraries with SureSelect XT HS/LI ENZ workflow.

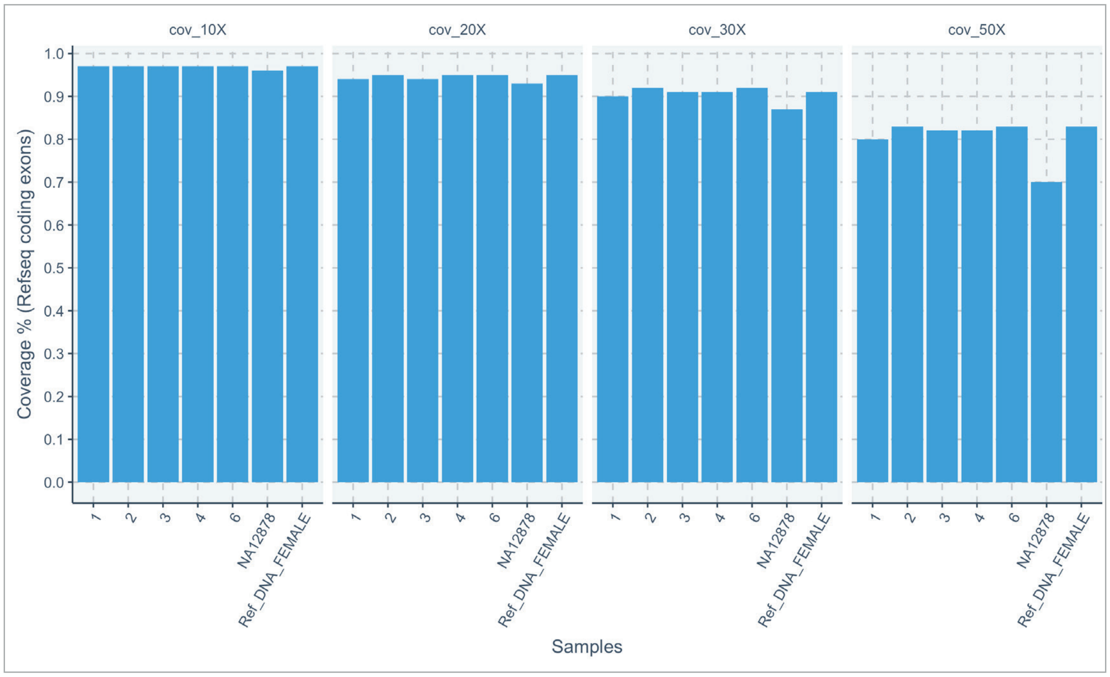

Figure 6. Coverage of RefSeq coding exons using SureSelect XT HS/LI ENZ with Exome V7 automated on the Agilent bravo NGS and Agilent bravo NGS workstation. Five blood samples (samples 1, 2, 3, 4 and 6), one Coriell (NA12878), and one Agilent female reference DNA (part number 5190-8850) were sequenced in a single run. The SureSelect XT HS/LI ENZ protocol with Exome V7 was automated on the bravo platforms and showed base coverage of RefSeq coding exons at 10X, 20X, 30X, and 50X comparable to that from manual preparation. (See Figure 4.)

Conclusion

The SureSelect^{XT} low input reagent kits and SureSelect XT HS and XT low input enzymatic fragmentation kit form a fast, streamlined, highly reproducible, and easily automatable workflow for both exome and custom panel applications. Sample inputs as low as 10 ng can be successfully used. Although not shown here, our preliminary data with FFPE samples also showed great sequencing performance. The use of the SureSelect human all exon V7, combined with the SureSelect^{XT} low input reagent kits and SureSelect XT HS and XT low input enzymatic fragmentation kit provides excellent coverage, uniformity, and read depth of the targeted regions allowing better sequencing results and economy. This workflow is compatible with different sample types and DNA qualities (blood, fresh frozen tissue, and FFPE). One key benefit of this approach is that a single workflow can be implemented in the laboratory for both constitutional and cancer applications. The implementation of the Agilent NGS automation platforms provides an automation solution to perform high-throughput DNA sample preparation and target enrichment.

www.agilent.com

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

PR7000-2017

Printed in the USA, March 27, 2019
5994-0667EN

Agilent
Trusted Answers