BMC Genomics

RESEARCH

Open Access

Performance comparison of Agilent new SureSelect All Exon v8 probes with v7 probes for exome sequencing

Vera Belova*, Anna Shmitko, Anna Pavlova, Robert Afasizhev, Valery Cheranev, Anastasia Tabanakova, Natalya Ponikarovskaya, Denis Rebrikov and Dmitriy Korostin

Abstract

Exome sequencing is becoming a routine in health care, because it increases the chance of pinpointing the genetic cause of an individual patient's condition and thus making an accurate diagnosis. It is important for facilities providing genetic services to keep track of changes in the technology of exome capture in order to maximize throughput while reducing cost per sample. In this study, we focused on comparing the newly released exome probe set Agilent SureSelect Human All Exon v8 and the previous probe set v7. In preparation for higher throughput of exome sequencing using the DNBSEQ-G400, we evaluated target design, coverage statistics, and variants across these two different exome capture products. Although the target size of the v8 design has not changed much compared to the v7 design (35.24 Mb vs 35.8 Mb), the v8 probe design allows you to call more of SNVs (+ 3.06%) and indels (+ 8.49%) with the same number of raw reads per sample on the common target regions (34.84 Mb). Our results suggest that the new Agilent v8 probe set for exome sequencing yields better data quality than the current Agilent v7 set.

Keywords: Exome sequencing, MGISEQ, BGI, NGS, WES, Agilent SureSelect, Variant calling, Enrichment quality

Introduction

Whole exome sequencing (WES) is widely used in genomic studies as well as genetic tests. Exons (protein coding regions) represent 1–2% of the human genome comprising up to 85% of the known variants significant for diagnostics [1]. At the same time, WES is 3–5 times cheaper than whole genome sequencing [2]. Currently, exome analysis, embracing a set of different characteristics, has proven to be a more efficient diagnostic tool, being especially effective in the area of human clinical genetics [3].

There are several commercial kits for whole exome enrichment. The most known kits are SureSelect (Agilent), TruSeq Capture (Illumina), xGen (IDT), Human Comprehensive Exome (Twist Bioscience), SeqCap EZ (Roche NimbleGen) [4–9]. Enrichment protocols are similar and are based on hybridization of exon sequences with biotinylated DNA or RNA probes with a subsequent capture by streptavidin-covered magnetic beads. Most kits are designed to enrich the libraries for sequencing using the Illumina platform. However, earlier, we managed to adapt the enrichment protocol for sequencing using the MGI platform with the Agilent SureSelect Human All Exon V6 probes that previously showed slightly better performance in exome sequencing in several studies [4, 10–13].

In 2021, Agilent launched an updated enrichment probe set v8 and compared its performance with the other manufacturer [14] but not with the previous version v7. In this study, we focused on comparing the Agilent SureSelect Human All Exon v7 and v8 designed

*Correspondence: verusik.belova@gmail.com

Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Ostrovityanova str. 1, Moscow 117997, Russian Federation

probes. We studied the changes introduced in panel design, metrics of enrichment quality and statistically assessed the efficiency and quality of the variant detection [15].

We prepared 20 libraries, divided them into 2 pools of 10 libraries and performed 2 rounds of enrichment of the pools using the v7 or v8 probes following the RSMU_exome protocol [16]. The sequenced pools were compared using bioinformatics pipeline based on the following characteristics: target regions, the percentages of on-targets, off-targets, and duplicates, as well as depth of coverage of the regions with various GC content.

Materials and methods

Sample Preparation and Sequencing

The libraries were prepared from 20 samples containing 300–600 ng of human genomic DNA taken from 20 patients using MGIEasy Universal DNA Library Prep Set (MGI Tech) following the manufacturer's instructions. DNA fragmentation was performed by sonication with the average fragment length of 250 bp using Covaris S-220. Quality control of the obtained DNA libraries was performed using the High Sensitivity DNA assay with the 2100 Bioanalyzer System (Agilent Technologies).

Previously pooled DNA libraries were enriched following the RSMU_exome protocol [16]. 20 DNA libraries were divided into 2 pools each containing 10 libraries. Each pool was enriched twice with the SureSelect Human All Exon v7 probes and the latest version of the probes SureSelect Human All Exon v8 (Agilent Technologies) for the second time. Finally, we obtained 4 enriched DNA library pools. The concentrations of the prepared libraries were measured using Qubit Flex (Life Technologies) with the dsDNA HS Assay Kit. The quality of the prepared libraries was assessed using Bioanalyzer 2100 with the High Sensitivity DNA kit (Agilent Technologies).

The enriched library pools were further circularised and sequenced by a paired end sequencing using DNBSEQ-G400 with the High-throughput Sequencing Set PE100 following the manufacturer's instructions (MGI Tech) with the average coverage of 100x. We loaded one pool per lane into the patterned flow cells in two different runs. FastQ files were generated using the zebracallV2 software by the manufacturer (MGI Tech).

Bioinformatics pipeline

The quality of the obtained 40 paired fastq files was analysed using FastQC v0.11.9 [17]. Based on the quality metrics, the fastq files were trimmed using Trimmomatic v0.39 [18]. To correctly estimate the enrichment and sequencing quality, all 20 exomes were downsampled to 50 million reads using Picard DownsampleSam v2.22.4 [19]. Reads were aligned to the indexed reference genome GRCh37 using bwa-mem [20]. SAM files were converted into BAM files and sorted using SAMtools v1.9 to check the percentage of the aligned reads [21]. Based on the obtained BAM files, the quality metrics of exome enrichment and sequencing were calculated using Picard v2.22.4, and the number of duplicates was calculated using Picard MarkDuplicates v2.22.4. We performed the quality control analysis with the following bed files: Agilent v7_regions, Agilent v8_regions. Bed files for the GENCODE and RefSeq databases were uploaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1309831311_Di0qVAk2HAMSBFgug0SoMWuDiYQT). Genomic coordinates of unique v7 and v8 regions in the bed files were annotated using the Panther database [22]. Variant calling was performed using bcftools mpileup v1.9.

Statistical analysis

Statistical tests were performed by R (version 4.2) in Rstudio (ver 2022.02.3 Build 492). To estimate a distribution of variables, the Shapiro–Wilk test was used. If we didn't reject H_0 hypothesis, the T-test was performed. Otherwise, Wilcox rank-sum test was used. P-value < 0.05 as the level of statistical significance was used.

Results

Comparison of probe designs

We detected several changes in target design of v8 compared to v7 introduced by the manufacturer. The manufacturer did not alter the probe structure preserving 120 bp biotinylated cRNA probes. The manufacturer claims that coding content was updated according to the database releases (CCDS release 22, GENCODE V31, RefSeq release 95), added the TERT promoter region, but removed non-coding ClinVar Pathogenic variants. The target size of the v8 kit is 35.24 Mb, whereas the target size of the v7 kit is 35.8 Mb, the intersection of the bed files from both kits is 98.42% (34.84 Mb). The percentage of unique target regions is 2.69% (0.96 Mb) and 1.14% (0.4 Mb) for the Agilent v7 and v8 exome, respectively. We compared the v7 and v8 bed files with the bed file containing the coding exons of the GENCODE Genes track (basic subtrack, release V39lift37, Oct 2021) (34.93 Mb). The intersection between v8 and GENCODE v39 was 98.9% (34.07 Mb), the intersection between v7 and GENCODE v39 was 98.2% (34.3 Mb), and 0.29 Mb of the GENCODE v39 regions were absent in both kits. We visualised the overlapping target regions for Agilent v7 exome, Agilent v8 exome, and GENCODE v39 as Venn diagram (Fig. 1) with indicated target sizes using matplotlib-venn library (https://github.com/konstantint/matplotlib-venn).

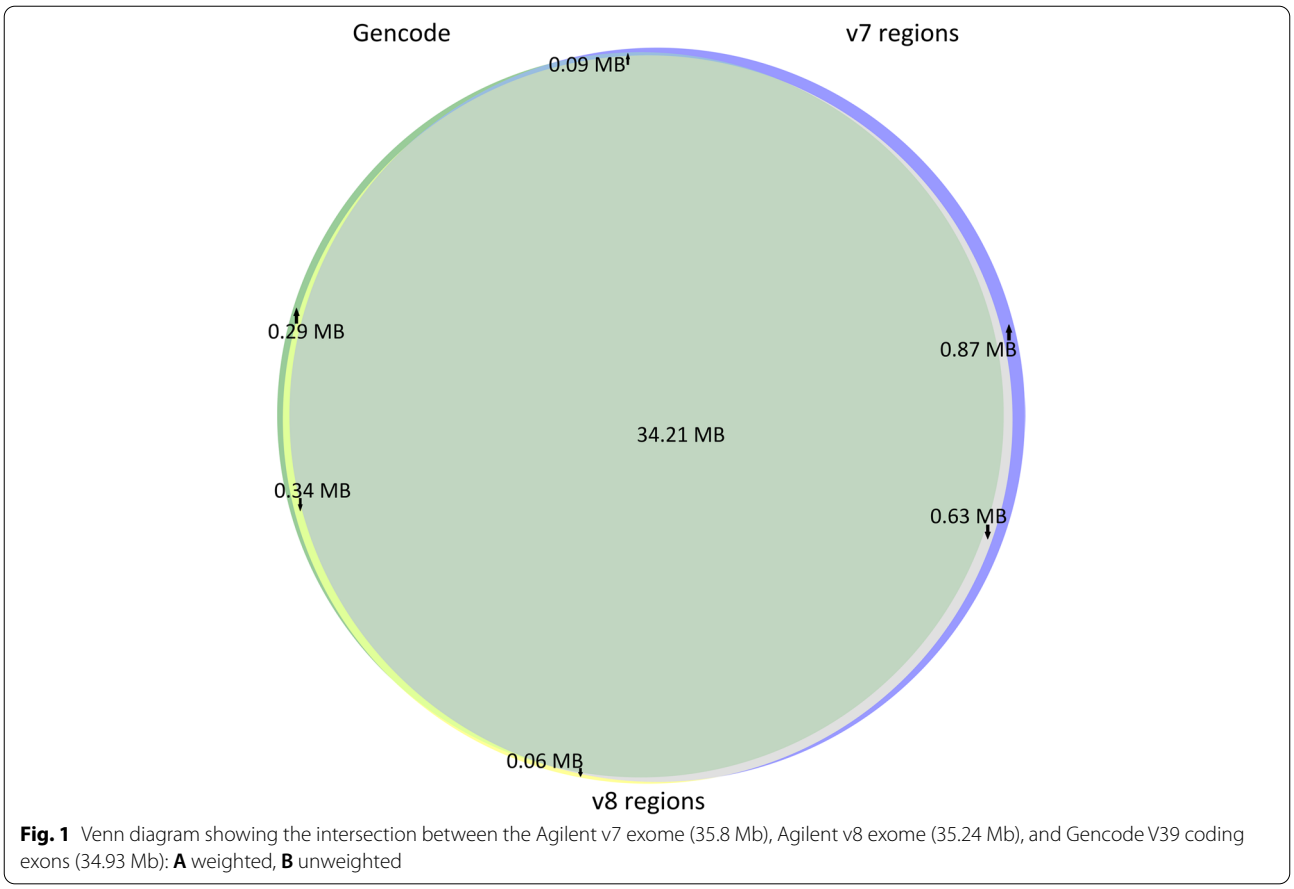

Fig. 1 Venn diagram showing the intersection between the Agilent v7 exome (35.8 Mb), Agilent v8 exome (35.24 Mb), and Gencode V39 coding exons (34.93 Mb): **A** weighted, **B** unweighted

We collected precise information (chromosomal coordinates, Gene ID, an annotation) on target regions from the v7 (0.96 Mb) and v8 kits (0.4 Mb) which is provided in the Supplementary Table 1. The figure S1 analysing the distribution of lengths of the changed fragments (Supplementary Table 1) demonstrates that most altered positions are short (less than several dozens of base pairs) which means that the manufacturer adjusted design of certain probes using the previous version of the targets. We analysed those fragments that were longer than 30 bp as we were interested in detecting unique fragments for the v7 and v8 kits in the current version of the bed file of the GENCODE v39 database. Our analysis also included the bed file containing the coordinates of all exons plus 20 bases at each end from the RefSeq ALL database (Source data version: NCBI Homo sapiens 109.20211119 (2021–11-23)). The unique sequences of the v8 target fit better into the current GENCODE v39 database than those of the v7 target. The intersection of the unique regions of the current GENCODE v39 with v8 was 0.33 Mb and with v7 was 0.09 Mb. The intersection of the RefSeq bed with a larger size (95.47 Mb) which includes exons + -20 bp from all curated and predicted genes with v8 was 0.23 Mb and with v7 0.26 Mb. We visualised the overlapping unique target regions for Agilent v7 exome, Agilent v8 exome, and GENCODE v39 (Fig. 2A), and NCBI RefSeq exons (Fig. 2B) as Venn diagrams. Together, these results suggest that v8 target was updated with current information of exonic variants.

Enrichment quality

To assess the enrichment quality, the obtained data (raw reads) for 40 exomes (20 samples enriched by the v7 or v8 probes) were downsampled to 50 M reads. The coverage statistics were calculated using Picard, and metrics were averaged for the samples from each v7 and v8 pool. The results obtained for each downsampled sample in the pool are shown in the Supplementary Table 2. We detected no significant differences in the number of on-target ($W = 217$, p-value $= 0.66$) and aligned reads ($T = 1.23$, p-value $= 0{,}26$), but detected in off-target ($W = 279$, p-value $= 0.03$) reads and the percentage of duplicates ($T = -3.76$, p-value $= 0.00066$) (Fig. 3). Mean + SD target coverage was similar for both kits and was $56.38x \pm 1.18$ and $56.88x \pm 1.32$ for v7 and v8, respectively. However, median target coverage values for two kits were different and equal to $53.4\times$ and $48.6\times$ for

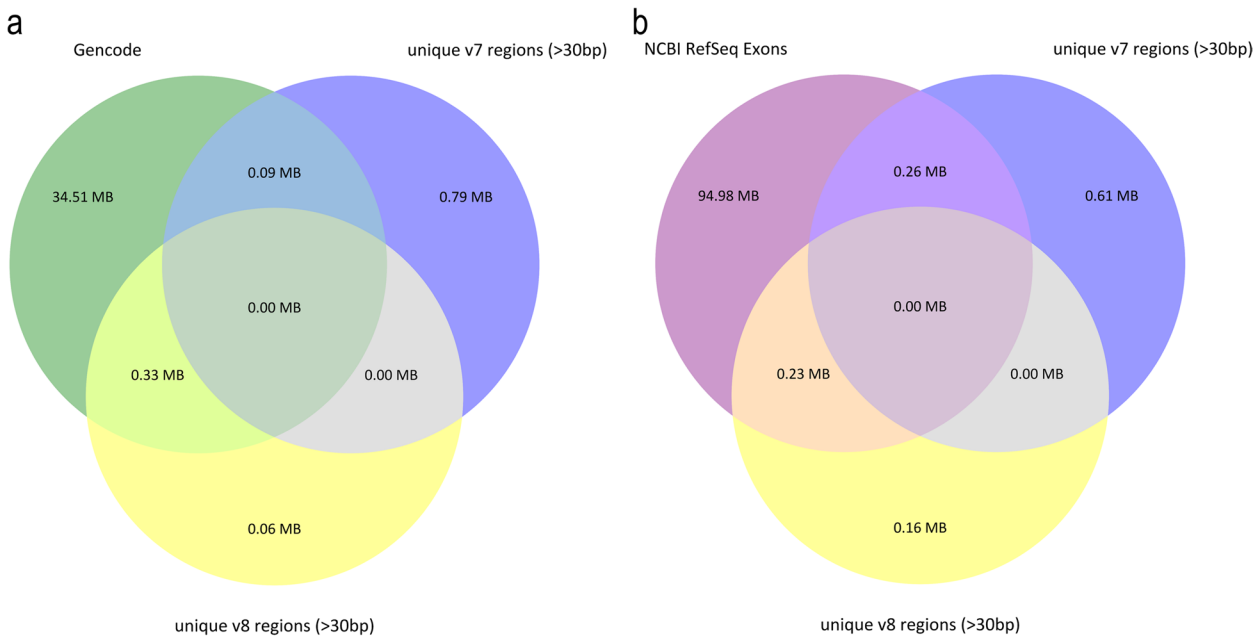

Fig. 2 Venn diagram showing the intersection of unique regions longer than 30 bp of the Agilent v7 exome (0.88 Mb) and Agilent v8 exome (0.39 Mb) with: **A** Gencode V39 coding exons (34.93 Mb); **B** NCBI RefSeq ALL exons + -20 bp (95.47 Mb)

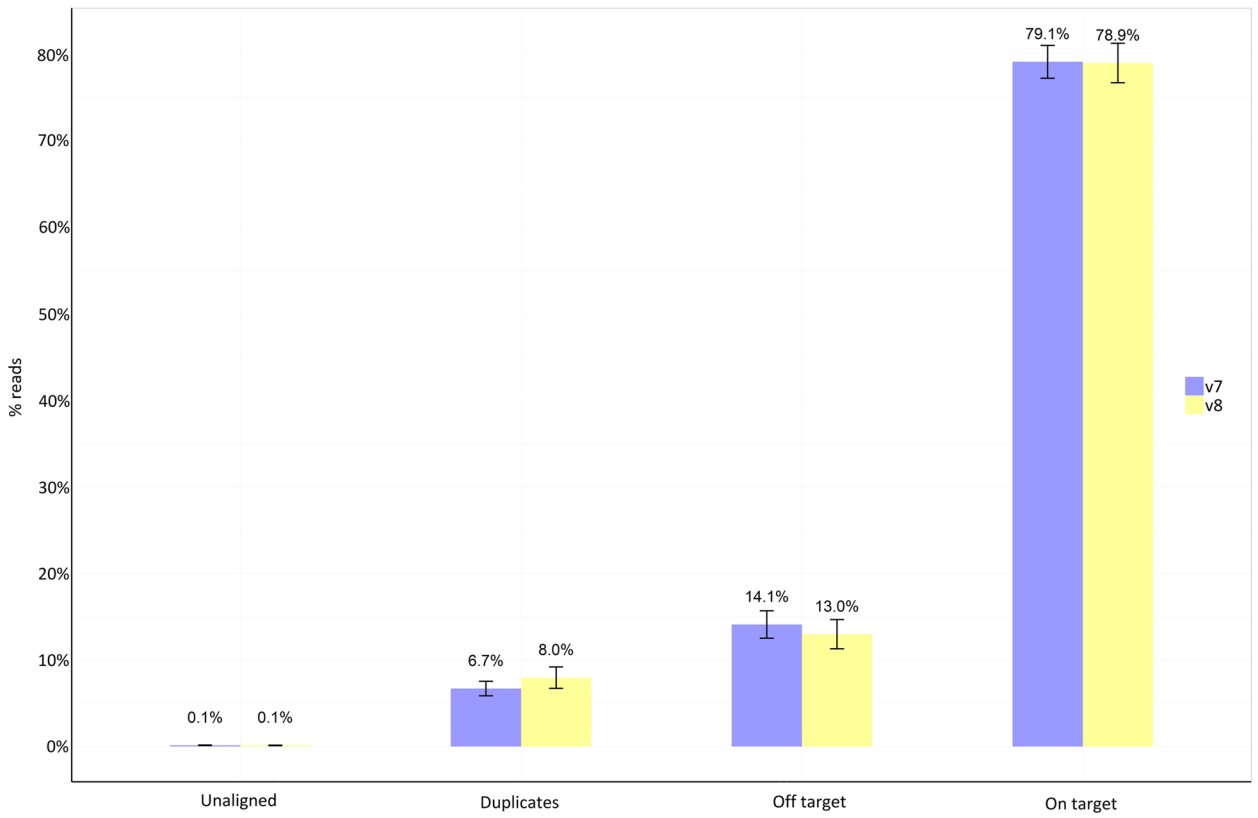

Fig. 3 Barplots show the average values for on-targets, off-targets, duplicates, and unaligned reads in the samples from the v7 and v8 exome pools (downsampled results)

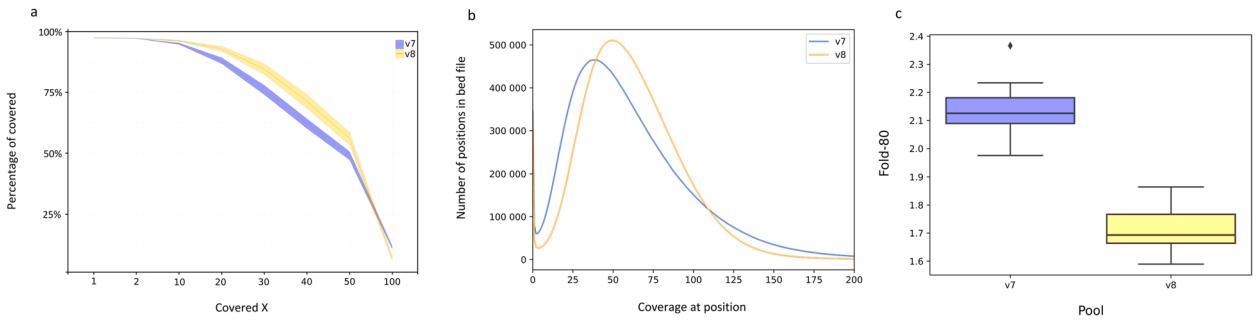

Fig. 4 **A**, **B**, **C** Performance of exome protocol in terms of coverage quality in downsampled samples: **A** Dependence of coverage quality of the v7 and v8 target regions from depth (mean ± SD values); **B** Graph showing the number of positions (bp, y-axis) vs. the coverage (x-axis); **C** Box plot showing the Fold-80 metric for the samples from the v7 and v8 pools

v8 and v7, respectively. Therefore, we suggest a higher coverage uniformity for v8 target.

The average values of metrics in the pools which reflect the target region coverage quality in the v8 kit are higher than in the v7 kit. The percentage of the target regions with ≥ 10 × coverage is 96.28 ± 0.0024% and 95.08 ± 0.0036% for v8 and v7 (T = -12.31, df = 38, *p*-value = 7.88e-14), respectively. At the same time, the percentage of the target regions with 20 × coverage in v7 is 5% less than in v8 (88.07 ± 0.013% for v7 vs. 92.96 ± 0.01% for v8, T = -13.198, df = 38, *p*-value = 1.97e-15) indicating that v8 has higher enrichment quality. The 40 × on-target coverage is in the range of 57–66% (mean ± SD = 62 ± 0.02%) for the v7 kit and is 9% less than that of the v8 kit (which is in the range of 65%-77%, mean ± SD = 71 ± 0.03%, T = -11.76, df = 38, *p*-value = 1.14e-13). At the same time, the distribution of v8 is closer to the normal distribution (Fig. 4B), there are fewer overcovered (≥ 80 × coverage) or undercovered positions (the inflection point is shown in Fig. 4A) which allows obtaining the sufficient coverage for more positions using fewer data.

The FOLD_80 parameter which reflects the coverage uniformity in the v8 pool samples (mean ± SD = 1.72 ± 0.076) is better than that of the v7 pool samples (mean ± SD = 2.13 ± 0.094) (Fig. 4C). The closer the value is to 1, the fewer rounds of sequencing a sample requires to obtain 80% of the targeted bases with the original mean coverage.

GC content

The AT_DROPOUT metric is 2 times lower for the exomes enriched with the v8 kit (v7 mean = 29.23%, v8 mean = 15.92%, W = 360, *p*-value = 2.88e-06). GC_DROPOUT does not differ between the kits (v7 mean = 12.9%, v8 mean = 13.09%, T = -0.352, df = 38, *p*-value = 0.72). Both AT_DROPOUT and GC_DROPOUT metrics indicate the percentage of misaligned reads that correlate with low (%-GC is < 50%) or high (%-GC is > 50%) GC content, respectively. Figure 5 demonstrates that the v8 probes (Fig. 5B) provide slightly more uniform coverage of regions with the GC content in the range of 40–60% (the red zone in the Fig. 5 shows a high density of regions with similar mean coverage and GC-content). However, this value is high in the v7 probes as well (Fig. 5A). We visualized the distribution of %-GC content of exonic regions with different coverage in the Figure S2 (Supplementary table 2). The density curves of %-GC were identical for v7 and v8 samples and correlated with previous results for Agilent of Wang et al. [23]. For low-covered exonic regions (< 10 × or < 20x) we observed no drastic curve shift on the graph towards high or low %-GC both on v7 and v8 samples.

SNV and INDEL calling comparison

Furthermore, we estimated the calling quality by calculating the number of single nucleotide variants (SNVs) and small insertions and deletions (indels) detected by different kits with the equal number of raw reads per sample. Table 1 shows the average result of calling for the v7 and v8 pools filtered by the quality of the entire bed files (results for each sample are provided in Supplementary Table 3). The following filters were used for calling: cut-off for the variants with the coverage depth exceeding 13 reads (DP > 13) and a parameter QUAL > 30. The mean + SD numbers of SNVs and indels obtained were 25,736 ± 380 and 743 ± 22 for the v7 exomes and 25,558 ± 362 and 699 ± 18 for v8. A higher amount of called variants for the v7 kit can be accounted for a larger size of the target design. As different kits provide bed files of different sizes, we compared variant calling in the overlapping target regions of the v7 and v8 kits. This approach enables a correct comparison of two probe designs. Using the same target (bed v7 cross

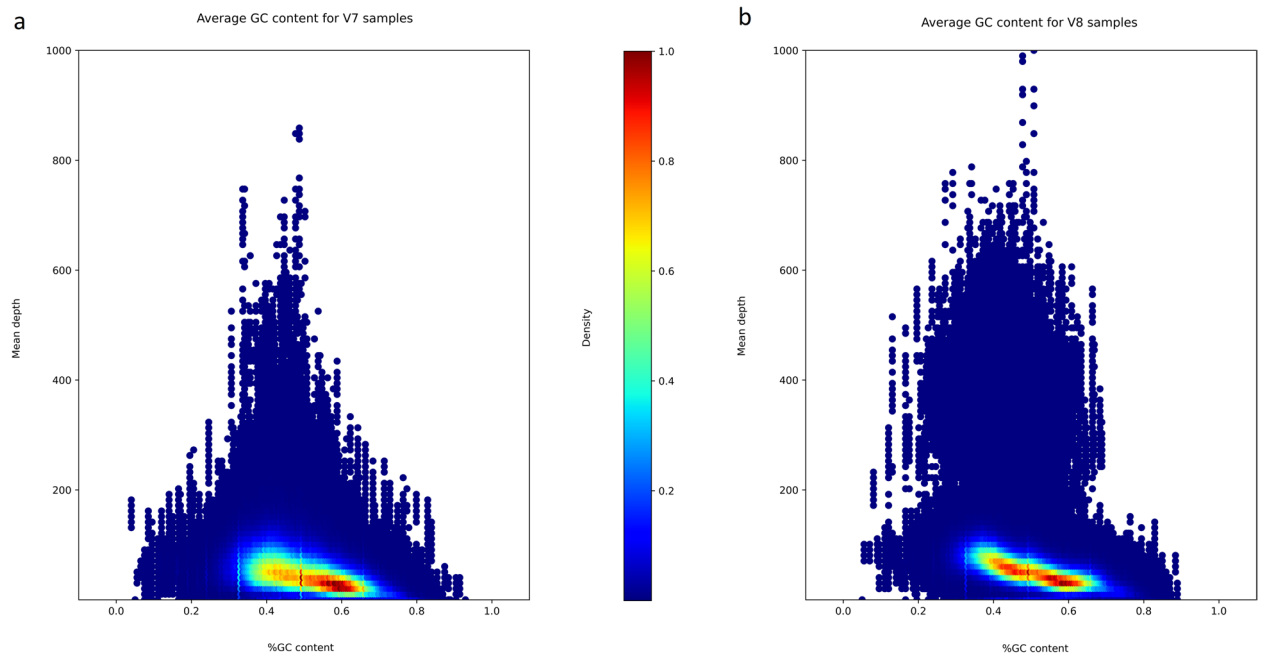

Fig. 5 Density plot showing Mean Depth vs. %GC content for: **a** Agilent v7; **b** Agilent v8. Density plot showing %GC content vs. mean depth. The data in this plot was collected by merging all samples from the V7 and V8 pools. Density estimation was performed using 2D plots. More specifically, we chose data points in a fixed rectangle (%GC content ∈ [0;1], mean depth ∈ [0;1000]) and split it into evenly spaced 200 × 100 grid and counted the data points in each cell of the grid. Finally, we normalised the grid to the range of [0,1] and plotted it using "jet" colormap from matplotlib library

Table 1 Average (mean ± SD) results of variant calling of SNV and indels for the samples from the v7 and v8 pools using their own target (bed v7, bed v8) and target intersection (bed v7 vs. v8) filtered by DP > 13 and QUAL > 30

Design	Variant type	Count on target v7 (35.7 Mb)	Count on target v8 (35.1 Mb)	Count on target V7 cross v8 (34.84 Mb)
V7 pool	SNV	25 736 ± 380	-	24 374 ± 347
	indel	743 ± 22	-	612 ± 18
V8 pool	SNV	-	25 558 ± 362	25 120 ± 355 **(+3.06%)**
	indel	-	699 ± 18	664 ± 17 **(+8.49%)**

v8 = 34.84 Mb), we calculated the average (mean ± SD) variant numbers. The number of SNVs and indels for the samples from the v8 pool were 3.06% (T = -9.3, df = 38, *p*-value = 2.61e-11) and 8.49% (T = -6.71, df = 38, *p*-value = 6.03e-08) higher than that of the v7 pool, respectively (Table 1). Then we performed intersection over union for variant calling results on "v7 cross v8" bed file and after that evaluated the quality of unique variants for v7 and v8 samples. The mean + SD coverage of unique SNVs and indels obtained were 74.5 ± 6.7 and 53.7 ± 3.3 for the v7 exomes and 60.5 ± 3.8 and 55.5 ± 3.4 for v8. The mean + SD QUAL value of unique SNVs and indels were 84.9 ± 7.8 and 128.2 ± 8.7 for the v7 exomes and 86.9 ± 6.1 and 132.3 ± 5.3 for v8.Together, these results shows that we obtained more unique variants without loss of quality using v8 probes.

Discussion

Overall, 2.76% of the target was excluded from v7, while 1.15% of the target was included into v8. Based on our data, we believe that no dramatic changes in probes significant for the clinical potential were added. Most modifications probably lie in changing the approach powered by machine learning to probe design. Most changed fragments in certain regions are several base pair long thus implying that manufacturer only adjusted certain target regions. However, some targets were quite long (dozens to thousands of base pairs) which indicates the functional changes as well. Changes affecting longer fragments arise from the updated information in the current versions of the databases. For instance, the transcript of the largest fragment in the NACA gene (3330 bp) that was excluded from the

v8 target undergoes splicing and is characterised as tsl5 – no single transcript supports the model structure (ENST00000454682.6 NACA-203). The manufacturer excluded all regions now considered to be not protein coding from the target and included certain regions that were unknown when the v7 probes were designed. This can be proved by analysing the intersection between the unique regions of both kits and the latest releases of databases in the same way we analysed Gencode V39 release (Oct 2021).

The major problem of WES is a non-uniform coverage of target regions resulting from the sensitive hybridization reaction of probes with the target fragments of DNA libraries. The introduced changes in the v8 probe design markedly improved the enrichment quality. The v8 probes with the same sizes of raw data per sample provided higher coverage of the larger percent of target regions. We noted that the degree of inadequate coverage of a particular region based on its AT content was better in case of the v8 version. We wondered if variant calling detected the same SNVs and indels in the samples obtained with the v7 and v8 kits. Indeed, the samples enriched with the v8 probes allowed for obtaining more useful data than the v7 probes due to new probe design and higher enrichment quality (uniform coverage of target fragments).

Noteworthy, the presented calculations were performed according to our in-lab gDNA standards. We aimed at estimating relative statistical metrics rather than absolute metrics as it is more correct to analyse them similarly to GIAB or Platinum Genomes. We intended to reveal the advantages that could be gained by an NGS facility performing exome sequencing if it switched to a new version of an enrichment kit.

Therefore, novel probe design Agilent all-exon v8 provides enough advantages as compared to the previous version of the kit and can be recommended as an advanced, more efficient generation of sequencing kits.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08825-w.

Additional file 1: Table of unique v8 and v7 regions.

Additional file 2: Picard statistics.

Additional file 3: Variant calling statistics.

Additional file 4: Figure S1. Distribution of fragments lengths.

Acknowledgements
Not applicable.

Authors' contributions
VB – Conceptualization, Methodology, Investigation, Validation, Writing – Original Draft Preparation; ASh – Methodology, Investigation, Writing – Original Draft Preparation; AP, RA and VCh – Formal Analysis, Methodology, Software, Visualization; AT, NP – Investigation; DR – Resources and Funding Acquisition; DK – Conceptualization, Project Administration, Methodology, Supervision, Writing – Review & Editing. The author(s) read and approved the final manuscript.

Funding
This work was supported by grant №075–15-2019–1789 from the Ministry of Science and Higher Education of the Russian Federation allocated to the Center for Precision Genome Editing and Genetic Technologies for Biomedicine.

Availability of data and materials
All 40 exome sequences were deposited into the NCBI open-access sequence read archive (SRA) in fastq.gz format under BioProject ID PRJNA832525: Agilent_v7_vs_v8_RSMU (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA832525/).

Declarations

Ethics approval and consent to participate
This study conformed to the principles of the Declaration of Helsinki. The appropriate institutional review board approval for this study was obtained from the Ethics Committee at the Pirogov Russian National Research Medical University. All patients provided informed consent for sample collection, subsequent analysis, and publication thereof and manuscript does not contain any individual person's data.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 April 2022 Accepted: 5 August 2022
Published online: 12 August 2022

References

1. Choi M, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci. 2009;106(45):19096–101.
2. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med. 2018;20:1122–30.
3. Fridman H, Bormans C, Einhorn M, Au D, Bormans A, Porat Y, Sanchez LF, Manning B, Levy-Lahad E, Behar DM. Performance comparison: exome sequencing as a single test replacing Sanger sequencing. Mol Genet Genomics. 2021;296(3):653–63. https://doi.org/10.1007/s00438-021-01772-3 (Epub 2021 Mar 11 PMID: 33694043).
4. García-García G, Baux D, Faugère V, et al. Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep. 2016;6:20948. https://doi.org/10.1038/srep20948.
5. Pengelly RJ, Ward D, Hunt D, et al. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. Sci Rep. 2020;10:3235. https://doi.org/10.1038/s41598-020-60215-y.
6. Shohdy, K.S., Bareja, R., Sigouros, M. et al. Functional comparison of exome capture-based methods for transcriptomic profiling of formalin-fixed paraffin-embedded tumors. npj Genom. Med. 6, 66 (2021). https://doi.org/10.1038/s41525-021-00231-7
7. Díaz-de Usera A, Lorenzo-Salazar JM, Rubio-Rodríguez LA, Muñoz-Barrera A, Guillen-Guio B, Marcelino-Rodríguez I, García-Olivares V, Mendoza-Alvarez A, Corrales A, Íñigo-Campos A, González-Montelongo R, Flores C. Evaluation of Whole-Exome Enrichment Solutions: Lessons from the High-End of the Short-Read Sequencing Scale. J Clin Med. 2020;9(11):3656. https://doi.org/10.3390/jcm9113656.
8. Barbitoff YA, Polev DE, Glotov AS, et al. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. Sci Rep. 2020;10:2057. https://doi.org/10.1038/s41598-020-59026-y.
9. Comparison of Whole Exome Capture Products – Coverage & Quality vs Cost. B Marosy, J Gearhart, B Craig, KF Doheny. Center for Inherited

Disease, Johns Hopkins Genomics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore. https://cidr.jhmi.edu/xtras/shared/documents/ASHG2018_ExomeComparison_FINAL.PDF.
10. Chung J, Son DS, Jeon HJ, et al. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. Sci Rep. 2016;6:26732. https://doi.org/10.1038/srep26732.
11. Shigemizu D, Momozawa Y, Abe T, et al. Performance comparison of four commercial human whole-exome capture platforms. Sci Rep. 2015;5:12742. https://doi.org/10.1038/srep12742.
12. Bonfiglio S, Vanni I, Rossella V, et al. Performance comparison of two commercial human whole-exome capture systems on formalin-fixed paraffin-embedded lung adenocarcinoma samples. BMC Cancer. 2016;16:692. https://doi.org/10.1186/s12885-016-2720-4.
13. Díaz-de Usera A, et al. Evaluation of whole-exome enrichment solutions: lessons from the high-end of the short-read sequencing scale. J Clin Med. 2020;9(11):3656.
14. SureSelect Human All Exon V8 Datasheet: High Performance Exome Built on Advanced and Proven Technology. 21 Jul 2021. URL: https://www.agilent.com/cs/library/datasheets/public/V5-datasheet-exome-v8-5994-3154EN-agilent.pdf
15. Kumaran M, Subramanian U, Devarajan B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics. 2019;20:342. https://doi.org/10.1186/s12859-019-2928-9.
16. Belova V, et al. System analysis of the sequencing quality of human whole exome samples on BGI NGS platform. Sci Rep. 2022;12:609.
17. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2017.
18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
19. Broad Institute GitHub: Picard. URL: https://broadinstitute.github.io/picard/
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
21. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
22. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol. 2009;563:123–40. https://doi.org/10.1007/978-1-60761-175-2_7.
23. Wang Q, Shashikant CS, Jensen M, et al. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. Sci Rep. 2017;7:885. https://doi.org/10.1038/s41598-017-01005-x.

Publisher's Note