Research &
Pharma Solutions

Whole Genome Sequencing

Tech Note

Whole Genome Sequencing – Does PCR make a difference?

The entire genetic information of an organism is represented on the genome. Whole genome sequencing (WGS) can give valuable insight into an individual's genetic information and contribute to various research questions. Different steps need to be considered to plan a WGS project or study. One consideration is the choice of library preparation and whether to use a PCR-based or PCR-free protocol.

Sequencing the whole genome is the most comprehensive method to analyze an individual's genetic information. By comparing the individual's genetic information to a reference genome, single nucleotide variants (SNVs), small insertions and deletions (indels), copy number variations (CNVs), and structural variants (SVs) can be studied. Whole genome sequencing can contribute to projects and studies in various areas, such as cancer studies, personalized medicine approaches, discovery of biomarkers, understanding pharmacogenetics, and disease research.

When planning a project, various choices about the project or study design, library preparation, sequencing parameters, and subsequent analyses need to be made. This Tech Note will focus on how PCR amplification can influence whole genome sequencing results. Previous studies claim that the PCR-free approach leads to a more uniform coverage of the sequencing data, especially across complex regions in the genome, such as GC-rich regions. Additionally, the performance in calling different variant types is said to be improved in the PCR-free approach compared to the PCR-based methods [1, 2].

To evaluate the influence of PCR-based library preparation in contrast to the PCR-free approach, we prepared human whole genome samples of the commercially available reference genome HG001. Five HG001 samples were prepared with Illumina's TruSeq DNA Nano preparation kit for the PCR-based approach and five HG001 samples with Illumina's TruSeq DNA PCR-free preparation kit for the PCR-free approach. For each approach, we used three intra-run and three inter-run replicates, where one of the five samples served simultaneously as intra-run and inter-run replicate. All samples were sequenced on the Illumina NovaSeq6000 instrument with a read length of 2 x 150 bp. Analyses were performed using the Illumina DRAGEN Bio-IT Platform and the human reference genome hg19. The results were subsequently compared. Throughout this Tech Note, the samples prepared with the TruSeq DNA Nano kit are highlighted in orange, and the samples prepared with the TruSeq DNA PCR-free kit are highlighted in blue.

Comparing the sequencing metrics

After the alignment, the samples were analyzed and compared regarding the mapping rate, duplication rate, and insert size. The average mapping rate for the PCR-based samples is slightly higher with 85.42% ± 3.04%, compared to the average mapping rate of the PCR-free samples with 81.08% ± 3.94%. With decreased mapping rates, the duplication rates are slightly but not significantly higher for the PCR-free samples (13.80% ± 3.91%) compared to the PCR-based samples (9.64% ± 3.06%). The percentage of unmapped reads is comparable for both approaches: 4.96% ± 0.05% for the PCR-based samples and 5.08% ± 0.07% for the PCR-free samples.

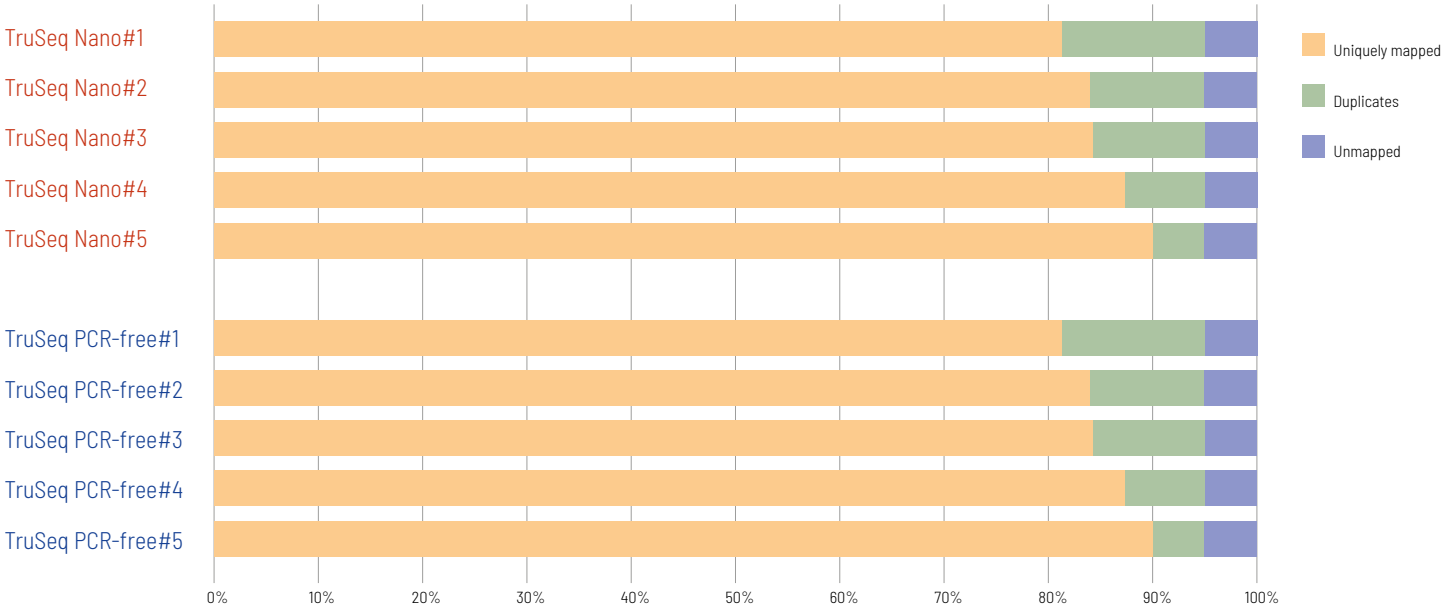


Figure 1 | Sequencing metrics. Proportions of uniquely mapped (green), unmapped (light blue), and duplicate reads (yellow).

Analyzing the uniformity of coverage

Previous studies claim that the PCR-free approach leads to a more uniform coverage of the sequencing data. The coverage uniformity is defined as the percentage of covered bases with at least 80% of the average coverage. Thus, we first have a look at the average coverage. The average coverage, also called sequencing depth, for the PCR-free protocol is slightly higher than for the PCR-based approach (37.16 ± 5.74 vs. 32.36 ± 5.75) and can be explained by a larger number of total reads sequenced for two of the replicates (see figure 2A). The coverage uniformity is high and does not differ significantly between protocols (figure 2B). Thus, the PCR-free approach is not essential for a uniform coverage.

Determining the GC content

The GC content is often discussed in the context of PCR-based or PCR-free methods. The average GC content in percent is given in figure 3A. Figure 3B shows the per-sequence GC content of every sample. No marked differences are visible between the PCR-based approach using the TruSeq DNA Nano kit (orange) and the TruSeq DNA PCR-free kit (blue). Thus, the PCR-free approach does not outperform the PCR-based approach with respect to the GC content of the sequencing output.

A)

Sample Name	Sequencing Depth	M Input Reads
TruSeq Nano#1	31.5 x	769.0
TruSeq Nano#2	29.7 x	701.5
TruSeq Nano#3	30.5 x	715.2
TruSeq Nano#4	43.4 x	985.2
TruSeq Nano#5	26.7 x	586.0
TruSeq PCR-free#1	33.5 x	807.8
TruSeq PCR-free#2	43.5 x	1034.0
TruSeq PCR-free#3	59.3 x	1473.4
TruSeq PCR-free#4	15.2 x	356.4
TruSeq PCR-free#5	34.3 x	930.8

B)

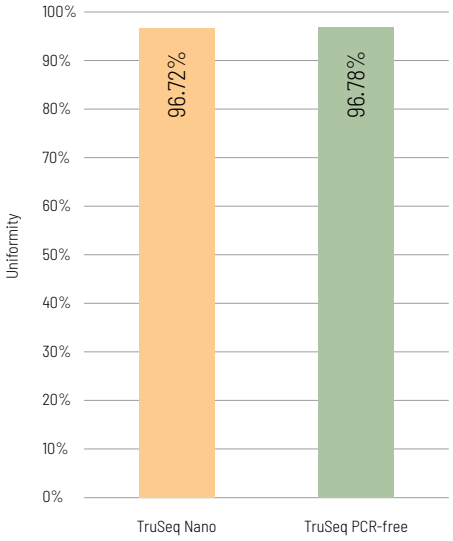


Figure 2 | Coverage-related statistics. A) The average coverage (also called sequencing depth) and the amount of input reads in million. B) The average uniformity of coverage for TruSeq Nano and TruSeq PCR-free.

A)

Sample Name	GC
TruSeq Nano#1	41%
TruSeq Nano#2	41%
TruSeq Nano#3	41%
TruSeq Nano#4	41%
TruSeq Nano#5	41%
TruSeq PCR-free#1	41%
TruSeq PCR-free#2	41%
TruSeq PCR-free#3	41%
TruSeq PCR-free#4	41%
TruSeq PCR-free#5	41%

B)

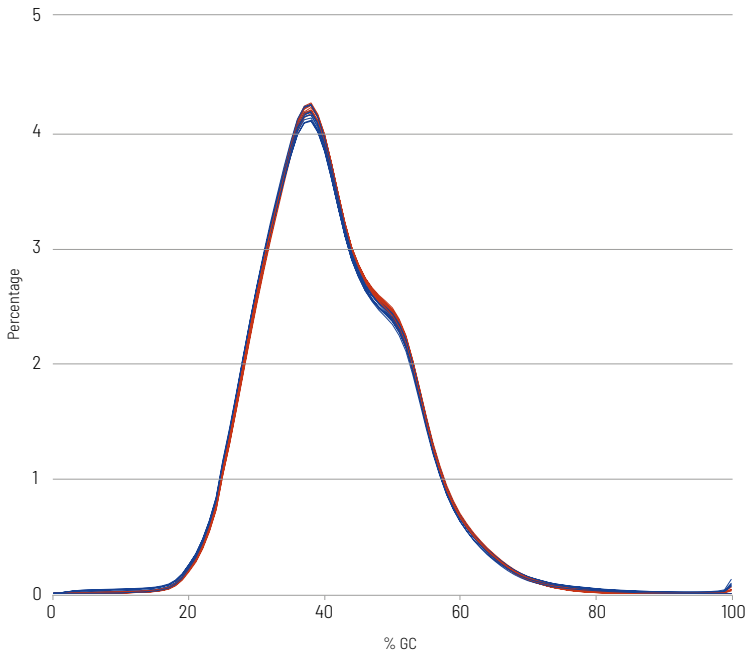


Figure 3 | GC content. A) shows the average GC content in percent for every sample. **B)** shows the per-sequence GC content.

Examining the variant calling performance

Calling different variant types with PCR-free methods is said to be improved compared to the PCR-based methods. Thus, we also examined the variant calling performance of the PCR-free and PCR-based methods. We used the commercially available reference genome HG001 as samples for both approaches. This genome in a bottle is well characterized and allows for calculating performance metrics for variant calling such as recall and precision.

The recall, also called sensitivity, is calculated based on the true positive (TP) and false negative (FN) calls:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The precision, also called specificity, is calculated based on the true positive (TP) and false positive (FP) calls:

$$\text{Precision} = \frac{TP}{TP + FP}$$

We determined the recall and precision for single nucleotide variants (SNVs) and small insertions and deletions (indels).

Recall and precision for SNVs are very high with very little variation. The values are comparable between protocols (see figure 4A). For indels, recall and precision are lower than for SNVs. The PCR-free protocol outperforms the PCR-based approach in recall and precision (figure 4B). However, both protocols produce data suitable for variant calling with very high sensitivity and specificity for both SNV and indel calling.

Conclusion

The PCR-free approach is not essential for a uniform coverage and does not outperform the PCR-based approach with respect to the GC content of the sequencing output for human samples. Both protocols produce data suitable for variant calling with very high sensitivity and specificity for both SNV and indel calling. Thus, the PCR-free and the PCR-based approach are both suitable for high-quality analyses of human samples. For samples with very high or low GC contents, a PCR-based approach might be beneficial. However, the advantage of the PCR-based approach is that a smaller amount of input material is required for the library preparation.

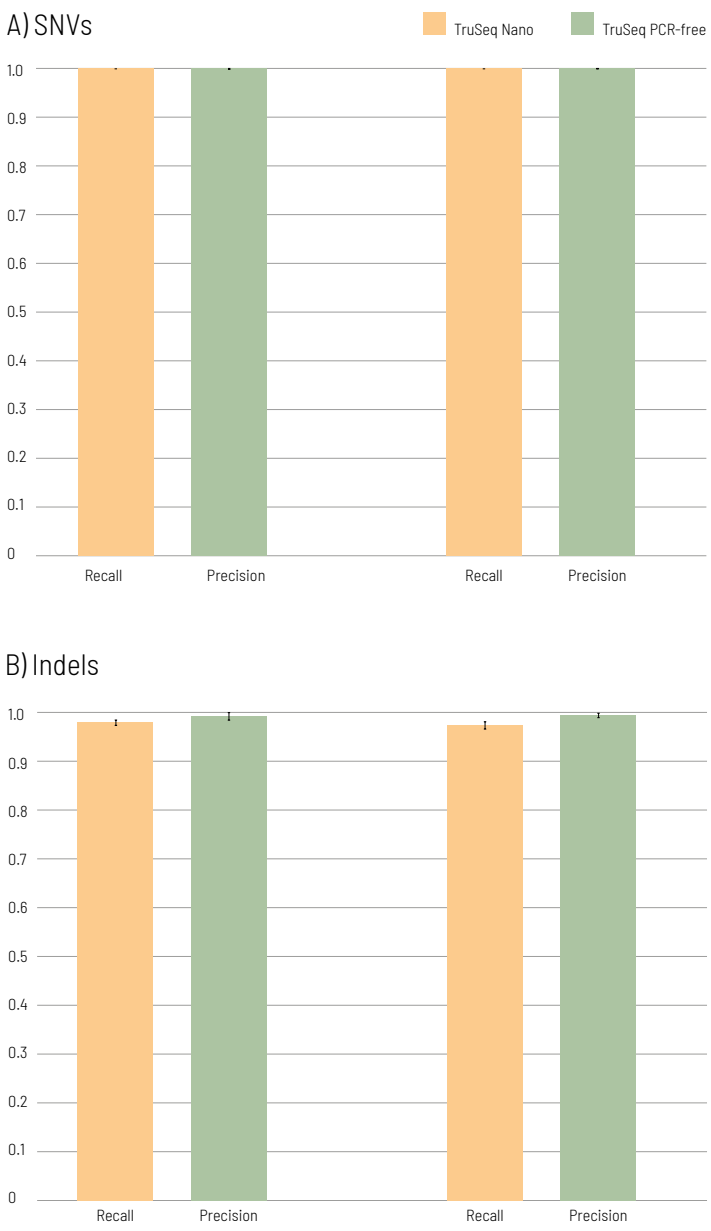


Figure 4 | Variant calling performance. A) Recall and precision for single nucleotide variant (SNV) calling. B) Recall and Precision for indel calling.

References

1. Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 2015 May;25(5):736-49. doi: 10.1101/gr.185892.114. Epub 2015 Mar 30. PMID: 25823460; PMCID: PMC4417121.

2. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009 Apr;6(4):291-5. doi: 10.1038/nmeth.1311. Epub 2009 Mar 15. PMID: 19287394; PMCID: PMC2664327


Sequencing
complete
illumina

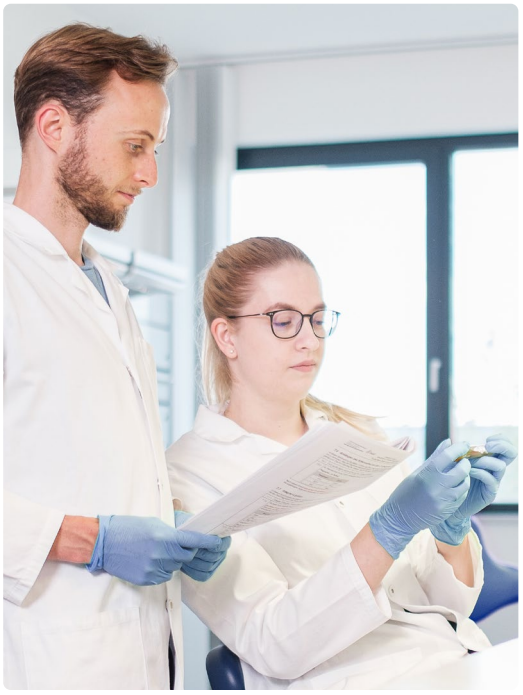
About Us

CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bio-informaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.

For more details please visit
www.cegat.com/rps



CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS
CLIA CERTIFIED ID: 99D2130225

DAkkS Deutsche Akkreditierungsstelle D-PL-13206-01-00
Accredited by DAkkS according to DIN EN ISO/IEC 17025:2018

2024/07