Advances in Genomics and Genetics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

SHORT REPORT

Effect of DNA insert length on whole-exome sequencing enrichment efficiency: an observational study

This article was published in the following Dove Press journal:
Advances in Genomics and Genetics

Anna Krasnenko[1,2]
Kirill Tsukanov[1]
Ivan Stetsenko[1]
Olesya Klimchuk[1]
Nikolay Plotnikov[1]
Ekaterina Surkova[1]
Valery Ilinsky[1,3,4]

[1]Genotek Ltd., Moscow, Russia; [2]Pirogov Russian National Research Medical University, Moscow, Russia; [3]Institute of Biomedical Chemistry, Moscow, Russia; [4]Vavilov Institute of General Genetics, Moscow, Russia

Abstract: Whole-exome sequencing (WES) currently allows the identification of the genetic basis of disease for 25%–40% of patients. A key element of WES is high-quality library preparation and target enrichment. In this short report, we examine the critical role of insert size (library portion between the adapter sequences) for enrichment efficiency. Our data can be used to improve WES results when applying the insertion size selection step.

Keywords: NGS, WES, enrichment efficiency, insert size

Introduction

Exome sequencing has revolutionized clinical research and diagnostics.[1,2] In a typical exome sequencing workflow, libraries are constructed from purified DNA, enriched for the exon regions and then sequenced. Targeted enrichment can be useful in a number of situations where particular portions of a whole genome need to be analyzed.

As sequencing and sample preparation technologies develop, the cost of exome sequencing has reduced substantially. However, the preparation of libraries for target enrichment and sequencing is still complex and sensitive.[3] To alleviate these problems, several techniques for optimization of library preparation can be proposed. For example, accurate size selection can boost sequencing efficiency, save money, improve assemblies and even allow sequencing of low-input samples. Typical libraries demonstrate a broad size distribution with average fragment sizes ranging from 10 bp to 1 kb in length. However, the resulting insert size is highly sensitive to initial sample concentration and fragmentation conditions, and the variation of insert sizes is often large.[4] Desired library size is determined by the desired insert size (referring to the library portion between the adapter sequences), because the length of the adaptor sequences is a constant. In turn, optimal insert size is determined by limitations of the next-generation sequencing (NGS) instrumentation and by specific sequencing application.

Standard Illumina® sequencing libraries currently tend to have a fragment size of 100–700 bp for good results. When using Illumina technology, optimal insert size is impacted by the process of cluster generation in which libraries are denatured, diluted and distributed on the two-dimensional surface of the flow cell and then amplified. While shorter products amplify more efficiently than longer products, longer library inserts generate larger, more diffuse clusters than short inserts.[3] In this article, we provide a short technical note on the effect of DNA insert length on the enrichment efficiency and how these data can improve NGS results.

Correspondence: Ekaterina Surkova
Genotek Ltd., Nastavnicheskii Pereulok, 17/1, Moscow 105120, Russia
Tel/fax +7 495 215 1514
Email esurkova@genotek.ru

http://dx.doi.org/10.2147/AGG.S162531

Materials and methods

DNA extraction was performed using QIAamp DNA Mini Kit (Qiagen NV, Venlo, the Netherlands) according to the manufacturer's instructions. The quality of genomic DNA was verified using electrophoresis on agarose gel. At this stage, lack of DNA degradation and RNA contamination were monitored. DNA concentration was measured using a Qubit 3.0 device (Thermo Fisher Scientific, Waltham, MA, USA). DNA libraries were prepared using NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) with adapters for sequencing on the Illumina platform according to the manufacturer's protocol. Double barcoding was performed by polymerase chain reaction (PCR) with a kit of NEBNext Multiplex Oligos for Illumina (Index Primers Set 1). The quality control of obtained DNA libraries was carried out using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). To target the enrichment of the coding regions, the target enrichment system SureSelect XT2 (Agilent Technologies) was used. DNA was sequenced on Illumina MiSeq (prior to enrichment, PE150) and HiSeq 2500 (for exome sequencing) using pair-end 100 bp reads.

Results

To examine the effect of fragment size on enrichment efficiency, we sequenced 71 human DNA libraries on Illumina platform before and after hybridization-based exome enrichment. In our study, insert sizes in DNA libraries ranged from 10 bp to 850 bp.

The proportion of uniquely mapped sequences from the total data obtained provides a metric for enrichment efficiency. Enrichment efficiency is calculated by dividing the number of reads with certain insert length after enrichment by the number of reads with certain insert length before enrichment. For normalization, we took mode of absolute enrichment efficiency as 100% of relative enrichment efficiency.

As shown in Figure 1, insert size crucially impacts enrichment results. The maximum efficiency of enrichment (>90%) is achieved with 250–330 bp insertion length.

We used the percentage of aligned reads instead of real enrichment efficiency. The amount of uniquely aligned reads may depend on multiple factors such as aligner used, reference genome, number of mismatches allowed, soft clipping and hard clipping. But those factors can only proportionally

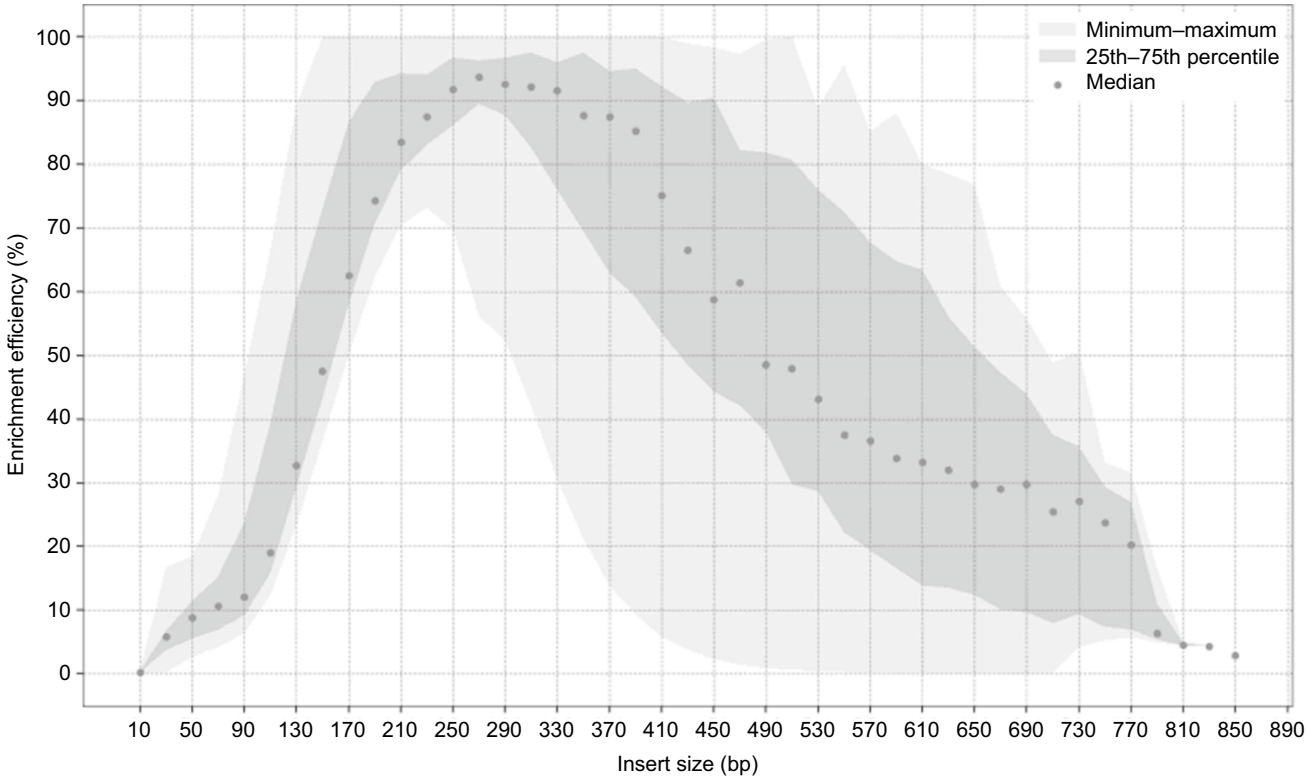

Figure 1 Impact of insert length on enrichment efficiency.
Note: Values were calculated for 71 human DNA samples.

Table 1 Exome sequencing coverage

	Target coverage at 1×
Minimum	19.4
Maximum	75.7
Mean	43.0
0.25 quantile	31.3
0.50 quantile (median)	43.9
0.75 quantile	50.7

increase or decrease the number of all reads, but they will not influence distribution of insert size.

Table 1 shows average exome sequencing coverage. Sequencing depth cannot affect enrichment efficiency – sequencing coverage depends only on sequencing setup and is independent of insert length distribution.

Discussion

In summary, our results indicate that 250–330 bp DNA fragments demonstrate the highest enrichment efficiency. For exome sequencing, about 80% of human exomes on each chromosome are <200 bp in length.[5] Given these data, the insert size of 250–300 bp is the optimal length for whole-exome sequencing. Therefore, the determination of size selection is an important step for effective enrichment and subsequent sequencing. Narrowing of distribution profile of the length of fragments significantly increases sequencing efficiency. This is especially important if several samples are pooled in a single run, because fragment length distribution affects their relative enrichment efficiency and final representation in sequencing results. These results should help guide experimental design and can be used as a metric for comparison of DNA library quantification methods.

Conclusion

Examination of whole-exome sequencing enrichment efficiency revealed 250–330 bp DNA inserts as most appropriate for improving results in our study. Our study demonstrates that size selection is an important step for effective sequencing.

Disclosure

The authors are employed by Genotek Ltd. The authors report no other conflicts of interest in this work.

References

1. Baldridge D, Heeley J, Vineyard M, et al. The exome clinic and the role of medical genetics expertise in the interpretation of exome sequencing results. *Genet Med*. 2017;19(9):1040–1048.
2. Ku CS, Cooper DN, Patrinos GP. The rise and rise of exome sequencing. *Public Health Genomics*. 2016;19(6):315–324.
3. Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 2014;56(2):61–64, 66, 68.
4. Turner FS. Assessment of insert sizes and adapter content in fastq data from nexteraxt libraries. *Front Genet*. 2014;30(5):5.
5. Sakharkar MK, Chow VT, Kangueane P. Distributions of exons and introns in the human genome. *In Silico Biol*. 2004;4(4):387–393.