Read Length and Nextera® Rapid Capture Exome Data

Nextera Rapid Capture Exome data sets with longer read lengths deliver higher mean coverage and more variant calls.

Highlights

- **Paired-End Data With Exome Sequencing**
  PE Sequencing produces twice the read depth and a higher number of variant calls
- **Longer Reads Produce Higher Coverage**
  Longer read lengths increased mean coverage for all trimmed data sets
- **Higher Coverage Increases True Variant Calls**
  Longer read lengths generated more variant calls with high precision and sensitivity rates

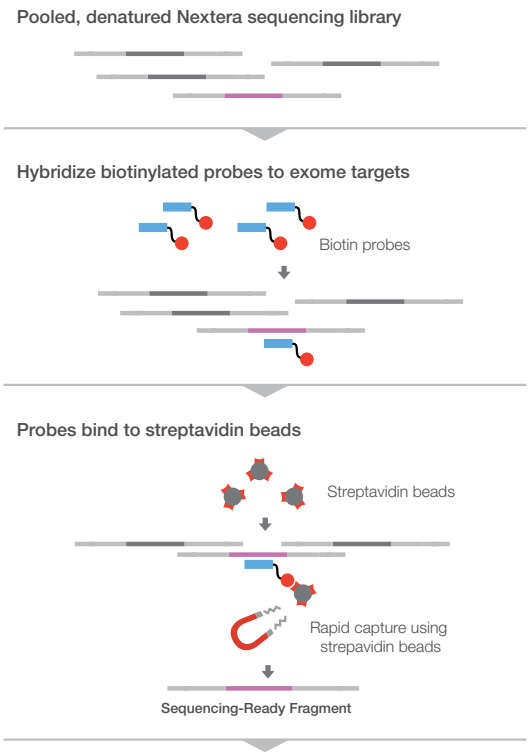
Introduction

The protein coding portion of the human genome—the exome—represents approximately 1% of the human genome. At first glance this seems like a small percentage; however, studies have shown roughly 85% of all mutations associated with Mendelian diseases fall within exome boundaries.[1] This high concentration of disease associated variants within the exome as well as the lower cost compared to whole-genome sequencing (WGS) have led to a steady increase in the use of whole-exome sequencing (WES) over the past several years.[2] The ability to focus on a targeted subset of the genome, and therefore deepen sequencing coverage, significantly increases the power of WES methods to identify, not only causitive variants of common Mendelian genetic diseases, but to detect rare variants associated with complex disorders.[3]

The purpose of this technical note is to provide guidance regarding optimization of sequencing parameters for exome sequencing. We also review the data quality parameters affected by changes in read length and coverage depth. As sequencing depth is the cornerstone of successful variant discovery, any WES method that bolsters coverage levels will also increase experimental accuracy.[4] To that end, Nextera Rapid Capture Exome sequencing, which utilizes a paired-end approach, imparts several key advantages (Figure 1). In addition to producing twice the number of reads for the same amount of hands-on time and effort, sequences aligned as read-pairs enable more accurate read alignment, higher numbers of single nucleotide variant (SNV) calls, and an ability to detect insertions and deletions (indels) that is not possible with single read data.[4] Analysis of differential read-pair spacing also allows removal of PCR duplicates.

Along with the paired-end approach, sequence read length is another means of tuning sensitivity and accuracy. To investigate these relationships, we sequenced and analyzed a well-characterized CEPH trio to assess the impact of read length on mean coverage, coverage uniformity, and variant calling.

Figure 1: Nextera Rapid Capture Exome Workflow and Specifications



Specifications	Nextera Rapid Capture Exome
Target region size	37 Mb
Number of target exons	214,405
Genomic DNA input	50 ng
Hands-on time	5 hours
Total time	1.5 days

The Nextera Rapid Capture Exome Kit provides a fast, simple method for isolating the human exome. The streamlined, workflow combines library preparation and exome enrichment steps, and can be completed in 1.5 days with low DNA sample input.

Methods

Exome Capture and Sequencing

DNA samples for a familial CEPH trio were obtained from the Coriell Institute for Medical Research.[5] The trio, a father (NA12891), mother (NA12892), and daughter (NA12878), is a subset of the CEPH collection of samples, from Utah residents, representing Northern and Western European ancestry. These samples have been sequenced as part of Phase 3 of the HapMap Project[6] and pilot Phase II of the 1000 Genomes Project.[7] In addition, these samples are part of the Illumina Platinum Genomes Project,[8] an effort to provide a gold standard data set for variant calls. Exome libraries were prepared using the Nextera Rapid Capture Exome Kit (Illumina, FC-140-1001). Pooled libraries were loaded on two lanes of two flow cells and sequenced on a HiSeq® 2500 System in rapid-run mode. The samples were sequenced as 2 × 150 bp paired-end reads.

Down-Sampling

To determine the impact of read lengths on enrichment performance metrics, the sequence data from each CEPH trio sample was trimmed down to generate 101 bp and 76 bp, paired-end data sets. Down-sampled subsets of the three CEPH trio exomes were generated by random selection using the Picard DownsampleSam[9] script and then included in separate BAM files.

Alignment and Variant Calling

Analysis of exome data was performed using the Enrichment workflow within the HiSeq Analysis Software (HAS).[10] Reads were aligned to the human hg19 genome[11] using the Burrows-Wheeler Aligner (BWA).[12] PCR duplicates as well as non-uniquely mapped reads were filtered out to ensure accurate alignment and variant calling rates. Similarly, to account for misaligned reads around indels leading to incorrect indel zygosity calls and false positive variant calls, indel realignment was performed using the Genome Analysis Tool Kit (GATK) software.[13,14] Variant calling and indel realignments were performed with GATK using default parameters. In addition, effects of varying read length and depth of coverage on alignment, variant calling, and enrichment performance were analyzed using a combination of Picard CalculateHsMetrics[15] and metrics generated by the Enrichment workflow within HAS.

Comparison to Orthogonal Data

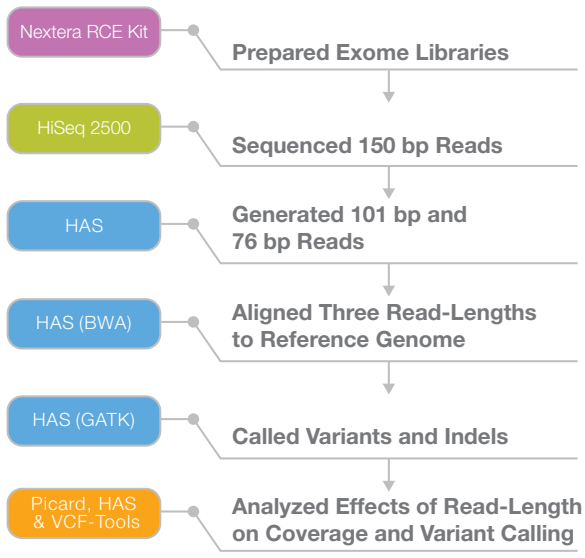
A subset of variant calls from this study were compared to orthogonal data from the National Institute of Standards and Technology (NIST) Genome in a Bottle Consortium.[16] All comparisons were performed using the VCFtools suite of tools[17] where "precision" is computed as the ratio of [# of True Positive Calls/(# of True Positive Calls + # of False *Positive* Calls)] and "recall" is computed as the ratio of [# of True Positive Calls/(# of True Positive Calls + # of False *Negative* Calls)]. All plots were generated using standard packages in R.[18]

Results

Effect of Read Length on Mean Coverage and Uniformity

Following our bioinformatic workflow (Figure 2) we assessed the impact of read length on mean coverage in the down-sampled and original data sets (Figure 3). For all down-sampled data (50 M, 75 M, 100 M, 150 M, 200 M, 300 M, 400 M) and original data (480 M),

Figure 2: Sequencing and Data Analysis Workflow



Nextera Rapid Capture Exome libraries were pooled and sequenced on a HiSeq 2500 System as 2 × 150 bp reads. Reads were trimmed to 101 bp and 76 bp with HiSeq Analysis Software (HAS) and aligned to the reference genome with the Burrows-Wheeler Aligner (BWA). Variant calling and indel realignments were performed with Genome Analysis Tool Kit (GATK). Effect of read length on mean coverage, variant calling, and enrichment performance were analyzed with Picard CalculateHsMetrics14 and HAS.

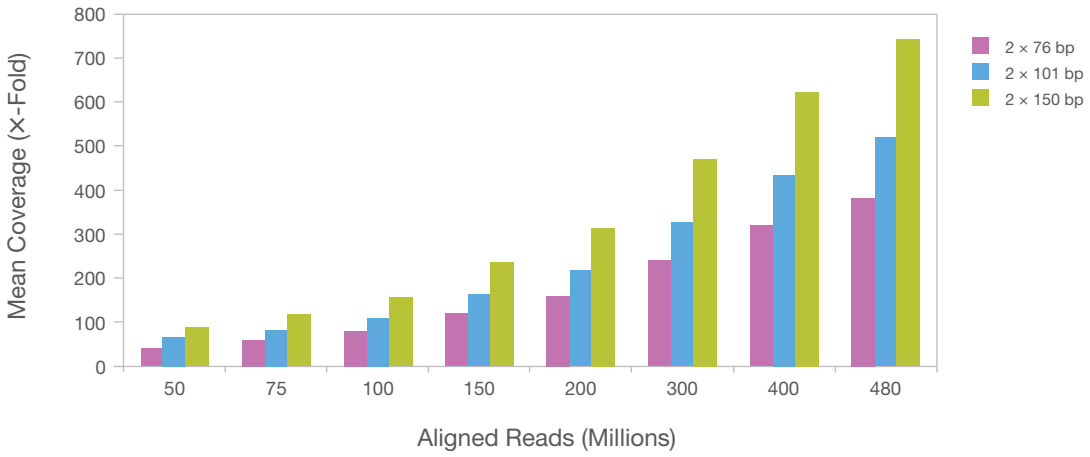
we observed the same trend: mean coverage increases with longer read lengths. Additionally, our data show that the increase in mean coverage with longer read length is magnified the deeper a sample is sequenced. For example, the increase in coverage between 75 bp and 150 bp reads—in the 50 M read data set—is *approximately 50x*; whereas the increase in coverage—in the 400 M read data set—is *approximately 360x* (Figure 3).

Next, we examined the effect of varying read length on coverage uniformity, where uniformity, at a defined depth, is calculated as [# of Target Bases Sequenced/(# of Total Target Bases × 100)]. As expected in any sequencing study, we found an inverse relationship between mean coverage and coverage uniformity. However, our data show that the decrease in coverage uniformity as mean coverage increases is smaller for longer read lengths (Figure 4, Δ). This indicates that longer read length, in addition to greater sequencing depth, can be used to modulate the loss of coverage uniformity in sequencing studies where greater mean coverage is required.

Effect of Read Length on Variant Identification

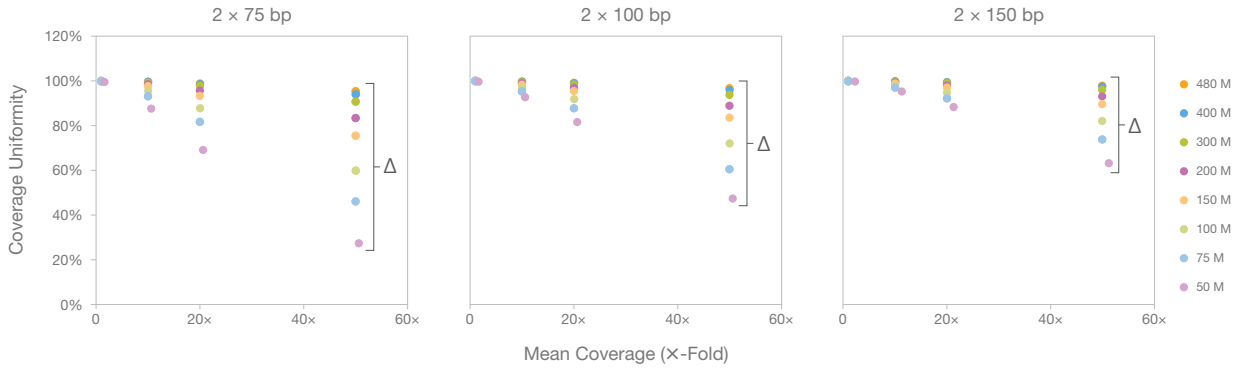
For the CEPH exome samples included in this study, we performed SNP and indel calling at various read lengths and coverage depths. SNVs identified in the trio samples show greater than 98% concordance with variants in dbSNP (version 135) for all read lengths and depths tested. The indels, at all read lengths and depths, showed

Figure 3: Effect of Read Length on Mean Coverage



Mean coverage was calculated for all downsampled data sets (50 M—480 M) at each read length. Mean coverages were calculated as the average of the trio.

Figure 4: Coverage Uniformity and Mean Coverage for Down Sampled Runs



Coverage comparison of data from the three read lengths and the down sampled runs have been plotted. The X axis is the mean coverage for bases within the target region and the Y axis is coverage uniformity. Uniformity, at a defined depth, is calculated as [# of Target Bases Sequenced/(# of Total Target Bases × 100)]. The average coverage uniformity of the CEPH trio is shown here.

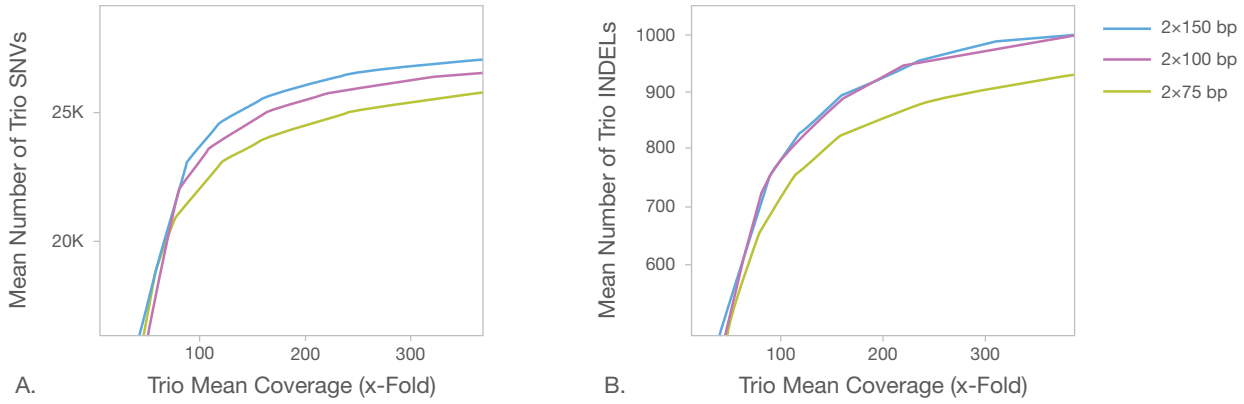
a concordance rate between 70–80%.

In all cases, the number of variants called was positively correlated with mean coverage (Figure 5A and 5B). The 2 × 76 bp data at 20× mean coverage enabled identification of 6781 SNV calls and 178 indel calls. At 400× coverage, 25805 SNVs and 927 indels were identified. The 2 × 150 bp reads enabled the identification of a maximum of 27529 SNVs and 1038 indels. At the same mean coverage, longer reads generated more SNVs than shorter reads (Figure 5A). However, with indels, for the same coverage depth, the number of indels identified from 2 × 76 bp to 2 × 101 bp reads increased by approximately 100 indels, while those identified from 2 × 101 bp and 2 × 150 bp reads are nearly identical (Figure 5B).

To further validate the SNV and Indel calls, we calculated the Mendelian conflict rate between CEPH trio family members. The conflict rate is based on comparison to expected Mendelian inheritance patterns and is an established proxy for the false positive rate. As the number of SNV calls increased with longer read lengths, we also observed a slight increase in false positive rate. The largest increase in false positive rate occurred from 101 to 150 base pairs. However, for indels, as the read length increased, we found a consistent decrease in false positive rate with the largest drop in false positives occurring between 76 and 101 base pairs (Figure 5C).

It bears noting that the actual false positive rates (0.4–0.45%) are comparable to the low false positive rate observed with whole-genome sequencing data[19] and do not represent a significant barrier to long read sequencing. Read lengths longer than 101 bps offer clear advantages for sequencing requiring high mean coverage

Figure 5: SNVs and Indels, Mean Coverage and Varying Read Length



(A) Average number of SNV's for the CEPH trio have been plotted against the average mean coverage. (B) Average number of indel's for the CEPH trio have been plotted against the average target region coverage.

Table 1: Mendelian Conflict Rate, Precision, and Sensitivity

	SNP Quality Data			Indel Quality Data		
	2 × 76 bp	2 × 101 bp	2 × 150 bp	2 × 76 bp	2 × 101 bp	2 × 150 bp
Mendelian Trio Conflic Rate	0.40%	0.41%	0.45%	5.92%	4.97%	4.74%
Precision (Compared to NIST)	91.6%	90.4%	89.4%	57.4%	55.3%	52.7%
Sensitivity (Compared to NIST)	94.7%	97.1%	98.3%	76.0%	79.1%	77.4%

Mendelian trio conflict rate for each of the three read length data sets is tabulated. Variant calls from the read length data were compared against variant calls generated by variant calls NIST (from the same trio: NA12892, NA12891, and NA12878). Precision is calculated as the ratio of [# of True Positive Calls/(# of True Positive Calls + # of False Positive Calls)] and the recall rate (sensitivity) is calculated as the ratio of [# of True Positive Calls/(# of True Positive Calls + # of False Negative Calls)]. The average precision and sensitivity for the trio is reported.

levels such as tumor-normal studies or identification of chromosomal rearrangements.

Effect of Read Length on Precision and Sensitivity

Comparison to NIST "Genome in a Bottle" Repository

To further evaluate the accuracy of the variants identified in this study, we compared them to variants in an orthogonal study. The NIST repository integrates a total of 12 NA12878 datasets from five sequencing platforms and curates variants after filtering those with low coverage, discordant genotypes, segmental duplications, structural variants and regions with evidence of bias. A total of 25,165 high-confidence SNVs and 722 indels from GiB were compared to the NA12878 variants from all three read length data sets. For SNVs, overall high precision rates, between 89—91%, were observed and longer read lengths were correlated to slightly lower precision rates (Table 1).

This observation was accompanied by a corresponding increase in the recall rate (also known as sensitivity). This indicates an overall increase in the ability to detect rare variants, with increasing read length. As with the false positive rate, we found the greatest improvement in data quality (measured here as increasing sensitivity) between the 76 bp and 101 base pair read lengths. Similarly, for indels, we find that the 101 bp read length marks a reasonable tradeoff between precision and sensitivity—particularly when the time and cost required to generate 2 × 150 bp reads as opposed to 2 × 101 bp reads are considerred (Table 1).

Discussion

While the simplest and most cost effective way to maximize coverage depth is to utilize paired end sequencing, read length can also be used to optimize coverage depth. Our data show the Nextera Rapid Capture Exome workflow generates high-quality variants at all three

tested read lengths, but that higher read lengths generate greater numbers of variants with similar precision and sensitivity.

Different kinds of sequencing studies require different sequencing depths and read lengths. Furthermore, the specific variants under investigation (eg. SNV, indel, large chromosomal rearrngements) will also be a factor in choosing read length. Here we found the number of SNVs identified increased with read length up to 150 bps. With indels, however, the number of variants identified did not increase significantly beyond 101 bps (Figure 5, Figure 6). Given the results of this study, we recommend *a minimum* 100× coverage and a 2 × 101 bp reads length for SNV and indel identification for Nextera Rapid Capture Exome sequencing.

Learn More

To download all sequence data from this study, visit
https://basespace.illumina.com/s/xs2HUjmXiRCk
https://basespace.illumina.com/s/utak9TSxfp7F

To learn more about Nextera Rapid Capture Exome kits visit
www.illumina.com/products/nextera-rapid-capture-exome-kits.ilmn

References

1. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–276.
2. Wang Z, Liu X, Yang BZ, and Gelernter J. (2013) The role and challenges of exome sequencing in studies of human diseases. Front Genet 4: 160.
3. Lupski JR, Gonzaga-Jauregui C, Yang Y, Bainbridge MN, Jhangiani S, et al. (2013) Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. Genome Med 5: 57.
4. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218): 53–9.
5. Coriell Institute for Medical Research (http://ccr.coriell.org/sections/Collections/NIGMS/CEPHResources.aspx?PgId=525&coll=GM).
6. He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, et al. (2009) Geographical affinities of the HapMap samples. PLoS One 4(3):e4684.
7. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422): 56–65.
8. Illumina Platinum Genomes Project (http://www.illumina.com/platinumgenomes).
9. Picard DownsampleSam (http://picard.sourceforge.net/command-line-overview.shtml#DownsampleSam).
10. HiSeq Analysis Software (http://support.illumina.com/sequencing/sequencing_software/hiseq-analysis-software.ilmn).
11. GRCh37, hg19, Feb. 2009 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/).
12. Burrows-Wheeler-Aligner (http://bio-bwa.sourceforge.net/).
13. Genome Analysis Toolkit (http://www.broadinstitute.org/gatk/).
14. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5): 491–498.
15. Picard CalculateHsMetrics (http://picard.sourceforge.net/command-line-overview.shtml#CalculateHsMetrics).
16. National Institute of Stantards and Technology (NIST) Genome in a Bottle Consortium (http://genomeinabottle.org).
17. VCFtools (http://vcftools.sourceforge.net/).
18. R Development Core Team. (2011) R: A language and environment for statistical computing. ISBN: 3–900051–07–0.
19. IGN Genome Sequencing Data Quality (http://www.illumina.com/services/whole-genome-sequencing-services/ign-data-quality.ilmn).

Illumina •.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • www.illumina.com

FOR RESEARCH USE ONLY

 Pub. No. 770-2013-027 Current as of 20 December 2013

illumina®