Vyas et al., Clin Med Biochemistry 2016, 2:1
DOI: 10.4172/2471-2663.1000112

Clinical & Medical Biochemistry

Research Article Open Access

Evaluation of Next Generation Sequencing Platforms for Whole Exome Variant Analysis

Gopi Vyas[1]*, Tanushree Tiwari[2], Aditya Mehta[1,2], Maulik Patel[1,2], Hemant Gupta[2], Arpita Ghosh[2] and Surendra KC[2]

[1]*Gujarat University, Ahmedabad, Gujarat, India*
[2]*Xcelris Labs, Ahmedabad, Gujarat, India*

Abstract

Exome analysis is potentially a cost effective approach for detecting mutations in human body. It is preferred widely over WGS (Whole Genome Sequencing) since it focuses upon the coding and functional variants of the human genome. There are several platforms involved in whole exome sequencing, but the relative study of the same sample on two widely used exome sequencing platforms is taken into consideration here. The results of this study demonstrate the systematic evaluation of two widely used sequencing platforms. However, it reported accuracy of ~98% in terms of SNPs predicted by both the platforms, infact the number of SNPs exclusively falling into exonic region were found to harmonize inspite of the initial difference in raw SNP calling step.

Keywords: Whole exome; Variant analysis; Illumina; Ion torrent; SNP; Human exome; GATK; Next generation sequencing

Introduction

The advancement of Next-Generation Sequencing (NGS) technology facilitates analysis of millions of DNA sequences in a single run. This has led to generation of ample amount of raw data which needs to be analyzed to gain meaningful results. While whole genome sequencing (WGS) provides an optimal solution to variant identification, the combination of data analysis hurdles and the expense of WGS have led to the development and successful adoption of exome sequencing. Exome sequencing provides a more cost effective approach where only the protein coding regions of a genome is utilized to find mutations which are found to be the prime cause of various common human diseases. The human genome comprises of about ~180,000 exons which constitutes 1% of the human genome [1]. The cost of Whole Exome is currently only ~1/10 of the cost of whole genome sequencing [2]. Exome Sequencing involves the target selection using one of several enrichment products; each of which aims to produce a DNA sample where the content is made up of the protein coding and regulatory regions of the genome. However it may not lead to 100% capture of the region of interest but helps in identification of the major mutations in a sample.

The Exome Sequencing Technology is best applied to obtain variants from the DNA sequence. The major genetic variants lie in the protein coding region and are the root cause of various diseases. Also a large number of non synonymous substitutions are predicted to be deleterious. The mutations in the non coding region are found to be weak or have no effects on the phenotypes. Splice sites also represent sequences in which there is high functional variation and are therefore also included in the capture of exomes.Thus exome represents an enriched subset of the human genome that may include highly effective variations [3]. The Next generation technologies have revolutionized genetics and the genomic research. There are many sequencing platforms in the market of which Illumina HiSeq and Ion torrent are the commonly used platforms [4]. Illumina HiSeq is considered to be the current market leader in Next generation sequencing, so to its performance for exome sequencing data was analyzed with other commonly preferred Next generation sequencing platform i.e. Ion torrent [5]. Here we compare the results of Illumina HiSeq to the performance of the Ion Torrent PGM [6]. The pipeline observed for exome data analysis is BWA-GATK pipeline [7]. A common pipeline was used to analyze the two datasets to bring them at a common platform. The accuracy of the alignment has a crucial role in variant detection. Properly aligned reads may lead to fewer errors during SNP calling. Around 3 million SNPs per genome are discovered using whole-genome sequencing because of the larger sequencing target (whole genome sequencing targets about 3 Gb, whereas the typical exome target is about 33 Mb) and around 15,000 to 20,000 variants are discovered per exome, with the variation in this number occurring from different exome target definitions [8]. Thus exome sequencing is a preferred cost effective technique since it focuses on the protein coding regions of the genome**.** Exome sequencing aids medical interpretation for clinical diagnostic purposes, identification of the underlying disease gene mutation and for therapeutic approaches [9].

Materials and Method

Data set

The NA12878 exome datasets were downloaded, SRR292250 (SRA-NCBI) for Illumina and http://ioncommunity.lifetechnologies.com for Ion torrent platform respectively. The Illumina data set was paired end whereas the other was the single end data, both having the common sample ID NA12878. The details of data are provided in, Supplementary Data Table 1 [10].

Raw read quality check and filtration

FASTQC was used to check the quality of raw and filtered fastq files from both the platforms. The Illumina and Ion Torrent reads were filtered by Trimmomatic [11] and PrinSeq-lite [12] respectively. The important parameters used for Ion torrent data were minimum length and quality of 20 whereas for Illumina data minimum length considered was 50 with quality of 20. The filtered reads were used for downstream analysis.

***Corresponding authors:** Gopi Vyas, Gujarat University, Ahmedabad, Gujarat, Tel: +919974535144; E-mail: gopi19.vyas@gmail.com

Received December 15, 2015; **Accepted** February 08, 2016; **Published** February 16, 2016

Citation: Vyas G, Tiwari T, Mehta A, Patel M, Gupta H, et al. (2016) Evaluation of Next Generation Sequencing Platforms for Whole Exome Variant Analysis. Clin Med Biochemistry 2: 112. doi:10.4172/2471-2663.1000112

Mapping and SNP calling

The alignment was done with BWA i.e. Burrows Wheeler Aligner. The Reference file was indexed with bwa - bwtsw [13] before the reads are aligned against hg19 reference genome. This alignment resulted with the same file which was converted to bam using same tools followed by its sorting. In order to avoid over-representation of specific sequence amplified during PCR, Remove Duplicates option was set to true using GATK. This helped in removal of duplicates from the output bam file. The reads were aligned again using Realignment Target Creator, for determining the suspicious intervals across the genome [14].

Quality score recalibration

The qualities in the fastq file are computed using a model. There are discrepancies between the model and empirical error rate. The error rate was computed by identifying all the sites where there were mismatch between the reads and reference. Thus these quantified covariances were used to recalibrate the base qualities to more accurate qualities. The recalibrated bam file can be further used for SNP calling step.

Variant calling and validation step

The unified genotype option was used to call all the aligned reads in the BAM files together and produce a VCF file with sites and genotypes for all samples; -glm argument either uses SNP or INDEL as a parameter to call SNPs or INDELs respectively. BOTH represents that both SNPs and INDELs are called together [15]. The BED file was required since to restrict the analysis only to the exome. However the reads may even align to the non-targeted regions, but the coverage may be low and no reliable variant calls are made in that case. The exome capture technology provided with BED files includes the list of regions that are targeted during the exome capture. The exome capture for Illumina data was performed using NimbleGen EZ Exome SeqCap (SeqCap_EZ_exome bed) whereas Ion torrent data used 318 chips and was enriched with TargetSeq Exome Enrichment Kit (TargetSeq Exome_hg19 bed file). Hence their respective BED files were used [10,16]. Variant Recalibrator was used to assign a well calibrated probability to each variant in a call set. It creates a Gaussian Mixture model by looking at annotation values over a high quality subset of input call set (HapMap, 1000G, dbSNP) and then evaluates the input variants.

Annotation

Functional annotation divides variants into synonymous variants (those that do not change the amino acid sequence), missense variants (those that introduce an amino acid change), and loss-of-function variants (those that prematurely truncate proteins and those disrupting protein splicing). Some studies further divide variants into different classes on the basis of the predicted effects of the protein [8]. The false positives variants, variants are annotated with effect prediction algorithms. ANNOVAR provides a wide variety of different annotation techniques, organized in the categories gene-based, region-based and filter-based annotation. It annotates single nucleotide variants (SNVs) and insertions/deletions, such as examining their functional consequence on genes, reporting functional importance scores, finding variants in conserved regions, or identifying variants reported in the 1000 Genomes Project and dbSNPs [16]. ANNOVAR can utilize annotation databases from the UCSC Genome Browser or any annotation data set conforming to Generic Feature Format version 3.The tool depends on several databases, which need to be downloaded individually. This approach ensures that the correct database version is used and the download of large unnecessary datasets is avoided [17].

The count of synonymous (10,778-Illumina and 9,908-Ion Torrent) and no synonymous (9,676-Illumina and 9,156-Ion Torrent) SNPs act in accordance with our analyzed data. 99.2% of nonsynonymous exonic SNPs were found in Illumina data and 95.8% non synonymous exonic SNPs were found in Ion torrent. Similarly, 99.3% of synonymous SNPs from Illumina data were obtained whereas in case of Ion Torrent it was 98.39%. Thus on the whole the SNPs found in Illumina data had greater concordance as compared to Ion Torrent [18].

Results

A total of 84,855,147 paired end reads from Illumina and 29,281,484 single end reads of Ion Torrent platforms were filtered resulting in 56,952,387 and 21,182,754 PE and SE reads respectively. These high quality reads were mapped on respective chromosomes using Qualimap (Supplementary Data Table 2) [19], it was seen that Ion torrent had more percent of mapped reads as compared to Illumina but at the same time the mapping coverage (Supplementary Data Tables 3 and 4), mapping quality of Illumina was considerably greater than that of Ion torrent data. The SNPs and INDELS were identified using GATK pipeline. Ion Torrent data was found to have more number of Variant Calls i.e. 4, 21,319 as compared to Illumina's 42,922. (Supplementary Data Table 5). However on selecting Snips from total variant call made, Ion Torrent produced more number of INDELs than SNPs where as Illumina produced an acceptable number of variant calls and SNPs individually. By the number of variant calls made, ion torrent was having higher false positives as compared to Illumina. The SNPs were filtered with depth ≥ 5 and reads with mapping quality ≥ 30. Only sites with QUAL ≥ 50 were considered as potentially variable sites. This stringent filtering fetched around 36,175 and 21,330 SNPs in Illumina and Ion Torrent data respectively (Supplementary Data Table 6). A total of 59% SNPs were unique to Illumina data and 18% to Ion torrent. Similarly both the platforms showed quite a few Common SNPs i.e. around 53.3% SNPs were found common in Illumina whereas 90.45% were found to be common in Ion Torrent out of the total SNPs.

The total variants found (Supplementary data I) only in the exonic region was 20,857 and 99.19% were already present in the dbSNP in case of Illumina. Similarly, for Ion Torrent total exonic variants were 19,469 and 97.1% were found in dbSNP. However, the total exonic SNP calls made by both the platforms were found nearly comparable to the ~12,500 variants that affect the protein coding portion of an individual's genome [20]. 168 exonic SNPs in Illumina and 563 SNPs in Ion Torrent were not reported in dbSNPs and can be expected as novel SNPs (Supplementary Data Tables 7-9).

Conclusion

The Sequencing coverage and the percent mapping reads were relatively more in case of Ion torrent data. Whereas the mean mapping coverage and the mapping quality was found more in Illumina reads. This gives the Illumina reads a slight upper hand of the two. Apart from this, the mapping coverage of 24x in case of Illumina suggested that larger part of exonic region was covered by the Illumina reads.

The Variant calling statistics gave a picture where both the platforms produced nearly same number of filtered SNPs. However there was a large difference observed in the raw variants called, where Ion torrent had dramatically large numbers of variants (SNP+INDELs) called. Further these SNPs were divided into different classes on the basis of the predicted effects of the protein. Here the count of synonymous and nonsynonymous SNPs found in the exonic region was relatively similar for both the platforms. A research showing that an average of non synonymous SNPs usually predicted in human

exomes were also found correlating with our data. Moreover around 99.19% and 97.1% SNPs called by Illumina and Ion torrent respectively showed concordance with the SNPs present in dbSNP. Out of the total exonic SNPs; 168 SNPs in Illumina and 563 SNPs in Ion Torrent were not reported in dbSNP and hence can be expected to be novel SNPs. Moreover the large number of novel SNPs reported by Ion Torrent also suggests that Ion Torrent might have more number of false positives. Thus in future the SNPs that show no concordance with dbSNP may be used to validate as novel findings and the non synonymous SNPs and leaves a scope of future research since it leads to a protein change. Exome Sequencing technology has its application in both the medicine and research since it identifies the functional variation. It also paves way for personalized medicine as sequencing the exome would suggest the underlying genetic etiology of an individual. Our comprehensive evaluation of exome data and tools may assist in selecting a suitable platform for Exome Sequence analysis and the Variant calling pipeline as well. Thus according to the requirement of the researcher the specific platform may be used considering the type of data and the objective.

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: 272-2766.
2. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, et al. (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13.
3. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.
4. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.
5. Hang G, Wang J, Yang J, Li W, Deng Y, et al. (2015) Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. BMC Genomics 16: 581.
6. Bao R, Huang L, Andrade J, Tan W, Kibbe WA, et al. (2014) Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform. 13: 67-82.
7. De Ligt J, Boone PM, Pfundt R, Vissers LE, de Leeuw N, et al. (2014) Platform comparison of detecting copy number variants with microarrays and whole-exome sequencing. Genom Data 2: 144-146.
8. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics 30: 2114-2120.
9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
10. Eric Vallabh Minikel (2012) Exome Sequencing pipeline using GATK. cureFFI.org
11. http://www.ncbi.nlm.nih.gov/sra/?term=SRR292250
12. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12: 227.
13. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
14. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863-864.
15. Wiley GB, Kelly JA, Gaffney PM (2014) Use of next-generation DNA sequencing to analyze genetic variants in rheumatic disease. Arthritis Res Ther 16: 490.
16. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, et al. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15: 256-278.
17. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, et al. (2012) Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28: 2678-2679.
18. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, et al. (2008) Genetic variation in an individual human exome. PloS Genet 4: e1000160.
19. Dolled-Filhart MP, Lee M Jr, Ou-Yang CW, Haraksingh RR, Lin JC (2013) Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. ScientificWorldJournal 2013: 730210.
20. http://ioncommunity.lifetechnologies.com/docs/DOC-2659

Citation: Vyas G, Tiwari T, Mehta A, Patel M, Gupta H, et al. (2016) Evaluation of Next Generation Sequencing Platforms for Whole Exome Variant Analysis. Clin Med Biochemistry 2: 112. doi:10.4172/2471-2663.1000112