Table of Contents
1 Introduction of Nextgene
1.1 Project Workflow of Nextgene
2 Sample Information
3 Method of Project Workflow
4 Bioinformatics Analysis Pipeline
5 Analysis Result
6 Appendices
7 Appendix

Human_WGS_Standard_Analysis_Paired Project Standard Analysis Report

Confidential Report

The PDF format may cause incomplete display of the tables. For information not displayed, please refer to the notes at the bottom of the tables or the report in the Excel format.

1 Background

Human whole genome sequencing enables researchers to catalog the genetic constitution of individuals and capture all the variants present in a single sample. This comprehensive data set provides the most complete picture of an individual's genetic blueprint and underlying disease risk.

Nextgene is one of the first companies to offer human whole genome sequencing services to the public, capable of sequencing up to 30-35X human genomes per year at the lowest cost per genome. We have extensive experience providing whole genome sequencing services on the powerful system, which has been installed and optimized for genomic analysis. Our bioinformatics team has developed a robust, accurate, and efficient pipeline based on our experience with the system, and advanced bioinformatics capabilities. Nextgene is able to expertly meet customer needs for executing large projects with timely turn-around and the highest quality results.

1.1 Advantages of Nextgene

1) State-of-the-art NGS technologies: Nextgene is leader in sequencing capacity using state-of-the-art technology, including the latest generation of Illumina Novaseq 6000 system.

2) Highest data quality: Nextgene guarantees a Q30 score ≥ 85%, surpassing Illumina's official guarantee of 75%.

3) Comprehensive bioinformatics capabilities: Nextgene uses its cutting-edge bioinformatics pipeline and internationally recognized tried-or-class software to ensure accurate, efficient, and robust data analysis.

1.2 Project Workflow of Nextgene

Nextgene's human whole genome sequencing projects are carried out as follows:

Figure 1 Project workflow of nextgene

2 Sample Information

Table 1 Sample information

Project	Sequencer	Sample	Type
2021	Novaseq 6000	20210730DNA	DNA
2021	6000	20210730DNA	DNA

Type: sample type (at normal, T tumor, P patient, D relative)

3 Method of Project Workflow

3.1 Sample Quality Control

Methods of sample quality control refer to QC report.

3.2 Library Construction, Quality Control and Sequencing

The genomic DNA was randomly sheared into short fragments. The sheared fragments were end repaired, A-tailed and further ligated with Illumina paired-end adapters. Suitable fragments were selected for PCR amplification.

Following is the workflow of library construction:

Figure 2 Workflow of library construction

The library was checked with Qubit and real-time PCR for quantification and bioanalyzer for size distribution detection. Quantified libraries will be pooled and sequenced on Illumina platforms, according to effective library concentration and expected data volume.

4 Bioinformatics Analysis Pipeline

Bioinformatics analysis includes:

1) Data quality control: filtering out reads containing adapters or with low quality;

2) Alignment with reference genome, statistics of sequencing depth and coverage;

3) SNP/Indels/SV/CNV calling, annotation and statistics;

4) Somatic SNP/Indels/SV/CNV calling, annotation and statistics (tumor-normal paired samples);

The flowchart below depicts the pipeline of data QC and bioinformatics analysis Nextgene used. Genetic analysis were performed only when tumor-normal paired samples are provided.

Figure 3 Bioinformatics analysis pipeline

5 Analysis Result

5.1 Raw Data

The original Illuminaone image files obtained from Illumina platform are transformed to short reads (Raw data) by base calling. These short reads are recorded in FASTQ by Illumina Basic, which contain sequence information (reads) and corresponding base quality.

Each read in FASTQ format is stored in four lines as follows:

@EAS139:136:FC706VJAAXX:2:2100:10040:193739 1:N:18:ATCACG
ATTTTTATTTTT

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
human_genes
exonic_genes
intronic_genes
upstream_genes
downstream_genes
intergenic
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000
27000
28000
29000
30000
31000
32000
33000
34000
35000
36000
37000
38000
39000

5.5.2 Somatic InDel Detection Result

Table 5.11 The number of somatic InDel in various genomic regions

Sample name	A260
CDS	2
frameshift_deletion	2
frameshift_insertion	0
nonframeshift_deletion	0
nonframeshift_insertion	0
stopgain	0
stoploss	0
unknown	0
intronic	184
UTR3	3
UTR5	0
splicing	0
ncRNA_exonic	0
ncRNA_intronic	33
ncRNA_splicing	0
upstream	3
downstream	4
intergenic	258
Total	487

Note:
(1) Sample name: Sample ID.
(2) CDS: The number of somatic InDels in coding region.
frameshift_deletion: A deletion of one or more nucleotides that cause frameshift changes in protein coding sequence.
frameshift_insertion: An insertion of one or more nucleotides that cause frameshift changes in protein coding sequence.
nonframeshift_deletion: A deletion that does not cause frameshift changes.
nonframeshift_insertion: An insertion that does not cause frameshift changes.
stopgain: An insertion or a deletion that leads to the creation of stop codon at the variant site.
stoploss: An insertion or a deletion that leads to the elimination of stop codon at the variant site.
unknown: Unknown function (due to various errors in the gene structure definition in the database file).
(3) intronic: The number of somatic InDels in intronic region.
(4) UTR3: The number of somatic InDels in 3'UTR region.
(5) UTR5: The number of somatic InDels in 5'UTR region.
(6) splicing: The number of somatic InDels within 2-bp of a splicing junction.
(7) ncRNA_exonic: The number of somatic InDels in exonic region of non-coding RNAs.
(8) ncRNA_intronic: The number of somatic InDels in intronic region of non-coding RNAs.
(9) ncRNA_splicing: The number of somatic InDels within 2-bp of a splicing junction of non-coding RNAs.
(10) upstream: The number of somatic InDels within 1kb away from transcription start site.
(11) downstream: The number of somatic InDels within 1kb away from transcription end site.
(12) intergenic: The number of somatic InDels in intergenic region.
(13) Total: The total number of somatic InDels.

5.5.3 Somatic SV Detection Result

Statistics of detected somatic SVs are listed below:

Table 5.12 Statistics of detected somatic SVs

Sample name	DUP	BND	INV	DEL
A260	50	679	135	100

Note:
(1) Sample name: Sample ID.
(2) DUP: (Tandem duplication) The number of tandem duplications.
(3) INV: (Inversions) The number of inversions.
(4) DEL: (Deletion) The number of deletions.
(5) BND: (Translocation) The number of translocations.

Figure 5.12 Number of different types of somatic SV in each sample

The x-axis represents the samples, and the y-axis represents the number of each type of SV.

5.5.4 Somatic CNV Detection Result

Statistics of detected somatic CNVs are listed below:

Table 5.13 Statistics of detected somatic CNVs

Sample name	gain_count	gain_size	loss_count	loss_size	total_count	total_size
A260	234	646408000	46	555251000	280	1201659000

Note:
(1) Sample name: Sample ID.
(2) gain_count: The number of gains.
(3) gain_size: The total size of gains.
(4) loss_count: The number of losses.
(5) loss_size: The total size of losses.
(6) total_count: The total number of CNVs.
(7) total_size: The total size of CNVs.

Figure 5.13 The size of genomic regions affected by somatic CNVs in each sample

The x-axis represents the samples, and the y-axis represents the total size of genomic regions affected by gains or losses (Mb).

6 References

[1] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754-1760. (BWA_MEM)

[2] Kent W J, Sugnet C W, Furey T S, et al. The human genome browser at UCSC[J]. Genome research, 2002, 12(6): 996-1006. (UCSC)

[3] Picard: http://sourceforge.net/projects/picard/ . (Picard)

[4] DePristo M A, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data[J]. Nature genetics, 2011, 43(5): 491-498. (GATK)

[5] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079. (Samtools)

[6] Sherry S T, Ward M H, Kholodov M, et al. dbSNP: The NCBI database of genetic variation[J]. Nucleic acids research, 2001, 29(1): 308-311. (dbSNP)

[7] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data[J]. Nucleic acids research, 2010, 38(16): e164. (ANNOVAR)

[8] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes[J]. Nature, 2012, 491(7422): 56-65. (1000g)

[9] Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders[J]. Nucleic acids research, 2005, 33(suppl 1): D514-D517. (OMIM)

[10] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource[J]. Nucleic acids research, 2004, 32(suppl 1): D258-D261. (GO)

[11] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30. (KEGG PATHWAY)

[12] Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples[J]. Nat Biotechnology, 2013.doi:10.1038/nbt.2514. (muTect)

[13] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs[J]. Bioinformatics, 2012, 28(14): 1811-1817. (Strelka)

[14] Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis[J]. Bioinformatics, 2012, 28: i333-i339. (DELLY)

[15] Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: A tool for assessing copy number and allelic content using next generation sequencing data[J]. Bioinformatics, 2012, 28(3):423-425. PubMed PMID: 22155870. (Control-FREEC)

7 Appendix

The software used in the whole genome sequencing data analysis are listed below:

Main software used in the whole genome sequencing data analysis

Analytical content	Software	Comments	Version
Alignment	BWA	Map the sequencing reads to the reference genome, and output the alignment file in the bam format	v0.7.17
	Samtools	Sort the bam file	v1.8
	Picard	Merge all bam files from the same sample and mark the duplicated reads	v2.18.9
SNP/InDel detection	GATK	Detect and filter SNPs/InDels	v4.0
SV detection	DELLY	Detect SVs	v0.7.8
CNV detection	control-FREEC	Detect CNVs	v11.4
Somatic SNP/InDel detection	MuTect/Strelka	Detect and filter Somatic SNPs/InDels	muTect:v1.1.4; Strelka:v2.9.4
Somatic SV detection	DELLY	Detect somatic SVs	v0.7.8
Somatic CNV detection	Control-FREEC	Detect somatic CNVs	v11.4
Annotation	ANNOVAR	Annotate variants	2015Dec14