Whole Genome Sequencing Report For UK Biobank

2018/2/25

Table of Contents

Results 3
1 Data Production 3
2 Summary Statistics of Alignment 4
3 Data Quality Control 6
Methods 7
1 Whole genome sequencing 7
2 Bioinformatics analysis overview 8
3 Data cleanup 8
4 Mapping and marking duplicates 8
5 Base Quality Score Recalibration (BQSR) 9
6 Web Resources 9
Help 10
FAQs 10
References 10

Results

1 Data Production

To discover genetic variations in this project, we performed whole genome sequencing of 50 DNA sample(s) with averagely 142,649.94 Mb raw bases. After removing low-quality reads we obtained averagely 1,328,520,531 clean reads (132,852.05 Mb). The clean reads of each sample had high Q20 and Q30 , which showed high sequencing quality. The average GC content was 40.30%. All whole genome sequencing data production was summarized in Table1. The base quality scores on clean reads per sample were plotted(Figure1).

Table 1 Summary of whole genome sequencing data (See all)

Samples	Raw reads	Raw bases (Mb)	Clean reads	Clean bases (Mb)	Clean data rate (%)	Clean read Q20 (%)	Clean read Q30 (%)	GC content (%)
1000000001	1,467,700,412	146,770.04	1,387,389,822	138,738.98	94.53	98.73	93.26	40.35
1000000002	1,415,565,816	141,556.58	1,322,096,086	132,209.61	93.40	98.39	92.53	40.24
1000000003	1,392,421,042	139,242.10	1,302,741,102	130,274.11	93.56	98.47	92.72	40.27
1000000004	1,478,831,816	147,883.18	1,380,119,430	138,011.94	93.32	98.57	92.98	40.12
1000000005	1,422,059,414	142,205.94	1,337,337,978	133,733.80	94.04	98.72	93.25	40.16
1000000006	1,419,643,896	141,964.39	1,325,874,350	132,587.43	93.39	98.59	93.09	40.23
1000000007	1,325,310,828	132,531.08	1,240,613,070	124,061.31	93.61	98.45	92.40	40.52
1000000008	1,431,458,870	143,145.89	1,318,673,346	131,867.33	92.12	98.18	92.32	40.09
1000000009	1,398,807,676	139,880.77	1,295,035,552	129,503.56	92.58	98.05	91.94	40.32
1000000010	1,421,611,252	142,161.13	1,303,926,638	130,392.66	91.72	98.17	92.57	40.25
1000000011	1,332,084,442	133,208.44	1,202,734,532	120,273.45	90.29	97.64	91.31	40.25
1000000012	1,475,249,576	147,524.96	1,392,736,566	139,273.66	94.41	98.19	92.30	40.27
1000000013	1,451,411,078	145,141.11	1,357,778,226	135,777.82	93.55	98.52	92.79	40.00
1000000014	1,478,550,746	147,855.07	1,388,976,994	138,897.70	93.94	98.51	92.72	40.40
1000000015	1,393,821,256	139,382.13	1,263,205,406	126,320.54	90.63	98.03	91.85	40.08
1000000016	1,499,508,632	149,950.86	1,410,016,044	141,001.60	94.03	98.50	93.02	40.22
1000000017	1,425,930,522	142,593.05	1,327,182,598	132,718.26	93.07	98.38	92.46	40.22
1000000018	1,494,595,716	149,459.57	1,390,713,516	139,071.35	93.05	98.46	92.89	40.29
1000000019	1,366,876,302	136,687.63	1,232,161,952	123,216.20	90.14	98.25	92.22	40.41
1000000020	1,429,450,876	142,945.09	1,335,037,504	133,503.75	93.40	98.48	92.75	40.36

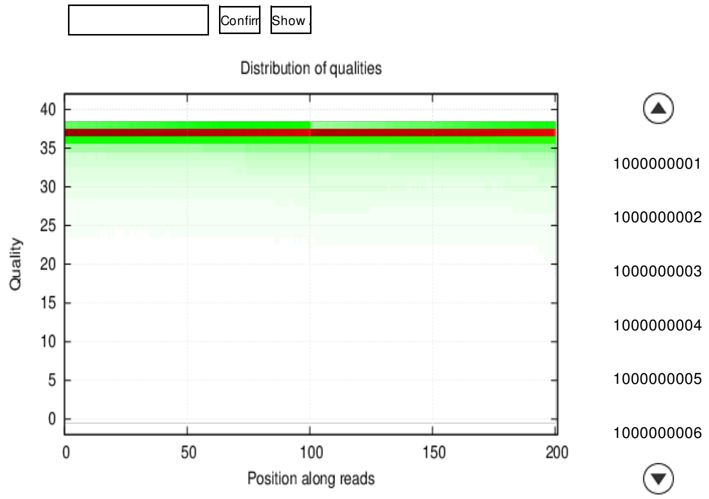

Figure 1 Distribution of base quality scores on clean reads.

X-axis is positions along reads. Y-axis is quality value. Each dot in the image represents the quality score of the corresponding position along reads.

2 Summary Statistics of Alignment

Total clean reads per sample were aligned to the human reference genome (GRCh38/HG38) using Burrows-Wheeler Aligner (BWA). On average, 99.98% mapped successfully and 90.70% mapped uniquely. The duplicate reads were removed from total mapped reads, resulting in about 2.89% duplicate rate and 42.16-fold mean sequencing depth on the whole genome excluding gap regions. On average per sequencing individual, 99.10% of the whole genome excluding gap regions were covered by at least 1X coverage and 95.99% had at least 15X coverage(Table2). In addition, the distributions of per-base sequencing depth and cumulative sequencing depth were shown as Figure2 and Figure3, respectively. The insert size distribution of paired sequencing reads was plotted in Figure4.

Table 2 Summary statistics of alignment (See all)

Samples	Clean reads	Clean bases (Mb)	Mapping rate (%)	Unique rate (%)	Duplicate rate (%)	Mismatch rate (%)	Average sequencing depth (X)	Coverage (%)	Coverag at least 4X (%)
1000000001	1,387,389,822	138,738.98	99.98	91.03	2.80	0.22	44.11	98.82	98.14
1000000002	1,322,096,086	132,209.61	99.99	90.79	2.72	0.25	42.07	99.53	98.85
1000000003	1,302,741,102	130,274.11	99.98	90.89	2.62	0.26	41.44	98.86	98.20
1000000004	1,380,119,430	138,011.94	99.98	90.59	3.22	0.23	43.66	98.84	98.18
1000000005	1,337,337,978	133,733.80	99.98	90.34	3.18	0.25	42.34	99.48	98.84
1000000006	1,325,874,350	132,587.43	99.97	90.86	2.90	0.23	42.10	98.83	98.14
1000000007	1,240,613,070	124,061.31	99.95	90.43	2.97	0.28	39.31	99.49	98.84
1000000008	1,318,673,346	131,867.33	99.98	90.05	3.53	0.30	41.56	98.81	98.11
1000000009	1,295,035,552	129,503.56	99.97	91.03	2.81	0.30	41.16	98.80	98.09
1000000010	1,303,926,638	130,392.66	99.99	91.09	2.42	0.31	41.61	99.50	98.83
1000000011	1,202,734,532	120,273.45	99.98	90.10	3.09	0.42	38.06	99.50	98.80
1000000012	1,392,736,566	139,273.66	99.99	91.23	2.66	0.29	44.34	98.83	98.18
1000000013	1,357,778,226	135,777.82	99.99	90.03	3.50	0.27	42.81	99.51	98.92
1000000014	1,388,976,994	138,897.70	99.98	91.34	2.52	0.23	44.29	98.85	98.20
1000000015	1,263,205,406	126,320.54	99.97	90.49	2.98	0.36	40.01	99.52	98.83
1000000016	1,410,016,044	141,001.60	99.99	90.70	2.99	0.30	44.70	98.85	98.21
1000000017	1,327,182,598	132,718.26	99.98	91.27	2.64	0.26	42.25	98.81	98.13
1000000018	1,390,713,516	139,071.35	99.98	91.24	2.68	0.25	44.24	98.85	98.23
1000000019	1,232,161,952	123,216.20	99.98	90.46	3.17	0.27	38.96	98.77	98.03
1000000020	1,335,037,504	133,503.75	99.97	90.37	2.88	0.30	42.34	99.49	98.87

Confirr Show

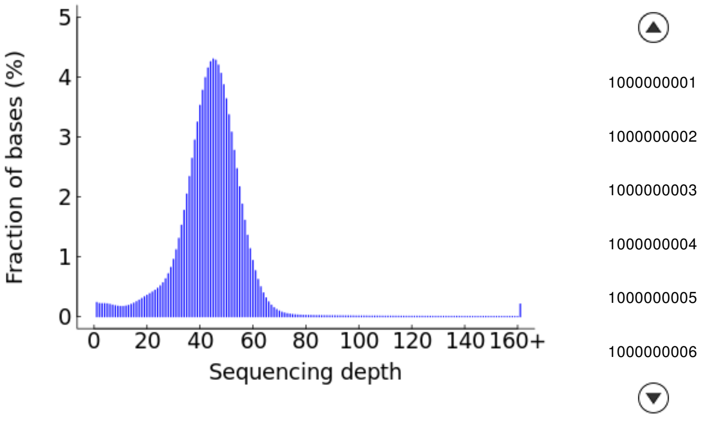

Figure 2 The distribution of per-base sequencing depth on the whole genome.

X-axis denotes sequencing depth, while Y-axis indicates the percentage of the whole genome excluding gap regions under a given sequencing depth.

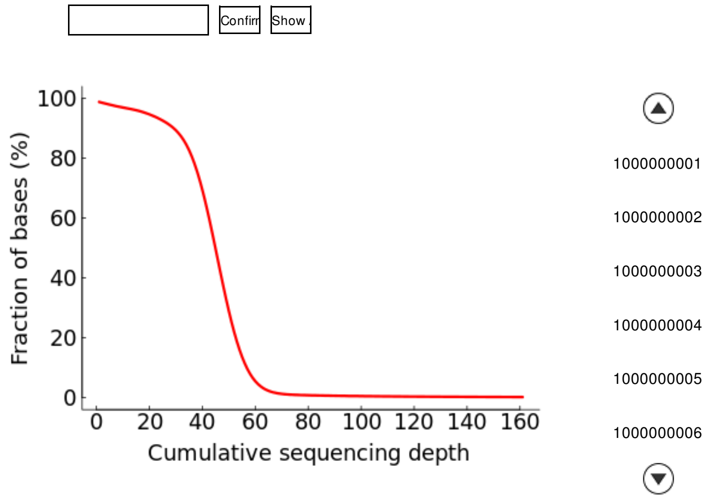

Figure 3 Cumulative depth distribution on the whole genome.

X-axis denotes sequencing depth, and Y-axis indicates the fraction of the whole genome excluding gap regions that achieves at or above a given sequencing depth.

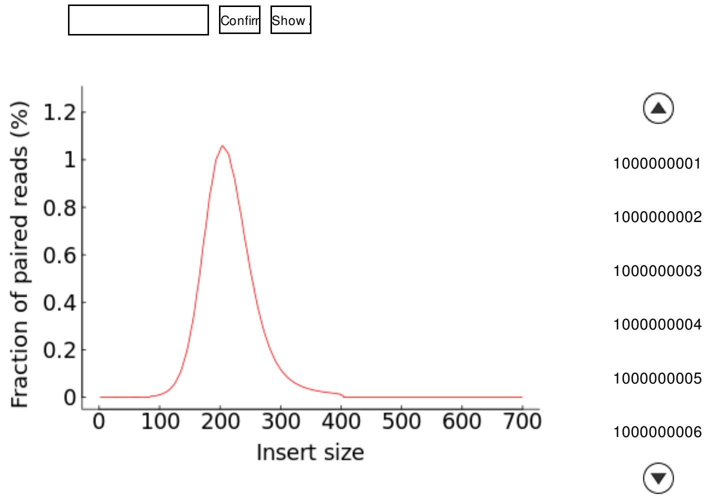

Figure 4 Insert size distribution of paired reads.

X-axis denotes insert size of paired reads, and Y-axis shows the fraction of paired reads with a given insert size.

3 Data Quality Control

The strict data quality control (QC) was performed in the whole analysis pipeline for the clean data , the mapping data, the variant calling, etc. Several quality control items for each sample were checked in Table3, where 'Y' showed PASS and 'N' showed FAIL. If some criteria were not met, measures such as re-sequencing or other effective methods would be carried out to improve the data quality and ensure qualified sequencing data.

Table 3 Data quality control for samples （See all)

Samples	Clean read Q20 (%)	Clean read Q30 (%)	GC content (%)	Mapping rate (%)	Duplicate rate (%)	Mismatch rate (%)	Average sequencing depth (X)	Coverage (%)	Genome coverage >15X (%)	UK co rat
1000000001	Y(98.73)	Y(93.26)	Y(40.35)	Y(99.98)	Y(2.80)	Y(0.22)	Y(44.11)	Y(98.82)	Y(96.05)	Y(9
1000000002	Y(98.39)	Y(92.53)	Y(40.24)	Y(99.99)	Y(2.72)	Y(0.25)	Y(42.07)	Y(99.53)	Y(96.10)	Y(9
1000000003	Y(98.47)	Y(92.72)	Y(40.27)	Y(99.98)	Y(2.62)	Y(0.26)	Y(41.44)	Y(98.86)	Y(95.89)	Y(9
1000000004	Y(98.57)	Y(92.98)	Y(40.12)	Y(99.98)	Y(3.22)	Y(0.23)	Y(43.66)	Y(98.84)	Y(95.95)	Y(9
1000000005	Y(98.72)	Y(93.25)	Y(40.16)	Y(99.98)	Y(3.18)	Y(0.25)	Y(42.34)	Y(99.48)	Y(96.23)	Y(9
1000000006	Y(98.59)	Y(93.09)	Y(40.23)	Y(99.97)	Y(2.90)	Y(0.23)	Y(42.10)	Y(98.83)	Y(95.84)	Y(9
1000000007	Y(98.45)	Y(92.40)	Y(40.52)	Y(99.95)	Y(2.97)	Y(0.28)	Y(39.31)	Y(99.49)	Y(95.88)	Y(1
1000000008	Y(98.18)	Y(92.32)	Y(40.09)	Y(99.98)	Y(3.53)	Y(0.30)	Y(41.56)	Y(98.81)	Y(95.70)	Y(9
1000000009	Y(98.05)	Y(91.94)	Y(40.32)	Y(99.97)	Y(2.81)	Y(0.30)	Y(41.16)	Y(98.80)	Y(95.80)	Y(1
1000000010	Y(98.17)	Y(92.57)	Y(40.25)	Y(99.99)	Y(2.42)	Y(0.31)	Y(41.61)	Y(99.50)	Y(96.05)	Y(9
1000000011	Y(97.64)	Y(91.31)	Y(40.25)	Y(99.98)	Y(3.09)	Y(0.42)	Y(38.06)	Y(99.50)	Y(95.45)	Y(9
1000000012	Y(98.19)	Y(92.30)	Y(40.27)	Y(99.99)	Y(2.66)	Y(0.29)	Y(44.34)	Y(98.83)	Y(95.96)	Y(1
1000000013	Y(98.52)	Y(92.79)	Y(40.00)	Y(99.99)	Y(3.50)	Y(0.27)	Y(42.81)	Y(99.51)	Y(96.38)	Y(9
1000000014	Y(98.51)	Y(92.72)	Y(40.40)	Y(99.98)	Y(2.52)	Y(0.23)	Y(44.29)	Y(98.85)	Y(96.02)	Y(9
1000000015	Y(98.03)	Y(91.85)	Y(40.08)	Y(99.97)	Y(2.98)	Y(0.36)	Y(40.01)	Y(99.52)	Y(95.73)	Y(9
1000000016	Y(98.50)	Y(93.02)	Y(40.22)	Y(99.99)	Y(2.99)	Y(0.30)	Y(44.70)	Y(98.85)	Y(96.06)	Y(9
1000000017	Y(98.38)	Y(92.46)	Y(40.22)	Y(99.98)	Y(2.64)	Y(0.26)	Y(42.25)	Y(98.81)	Y(95.92)	Y(9
1000000018	Y(98.46)	Y(92.89)	Y(40.29)	Y(99.98)	Y(2.68)	Y(0.25)	Y(44.24)	Y(98.85)	Y(96.06)	Y(9
1000000019	Y(98.25)	Y(92.22)	Y(40.41)	Y(99.98)	Y(3.17)	Y(0.27)	Y(38.96)	Y(98.77)	Y(95.52)	Y(9
1000000020	Y(98.48)	Y(92.75)	Y(40.36)	Y(99.97)	Y(2.88)	Y(0.30)	Y(42.34)	Y(99.49)	Y(96.23)	Y(1

● Methods

1 Whole genome sequencing

The qualified genomic DNA sample was randomly fragmented by Covaris technology and the fragment of 350bp was obtained after fragment selection. The end repair of DNA fragments was performed and an "A" base was added at the 3'-end of each strand. Adapters were then ligated to both ends of the end repaired/dA tailed DNA fragments, then amplification by ligation-mediated PCR (LM- PCR), then single strand separation and cyclization. The rolling circle amplification (RCA) was performed to produce DNA Nanoballs (DNBs). The qualified DNBs were loaded into the patterned nanoarrays and pair-end read were read through on the BGISEQ-500 platform and high-throughput sequencing are performed for each library to ensure that each sample meet the average sequencing coverage requirement. Sequencing-derived raw image files were processed by BGISEQ-500 basecalling Software for base-calling with default parameters and the sequence data of each individual is generated as paired-end reads, which is defined as "raw data" and stored in FASTQ format.

2 Bioinformatics analysis overview

Figure1 showed the data flow for the whole genome sequencing analysis.

The bioinformatics analysis began with the sequencing data (raw data from the BGISEQ machine). First, the clean data was produced by data filtering on raw data. All clean data of each sample was mapped to the human reference genome (GRCh38/HG38). Burrows-Wheeler Aligner (BWA)\[1\]\[2\]software was used to do the alignment. To ensure accurate variant calling, we followed recommended Best Practices for variant analysis with the Genome Analysis Toolkit(GATK, https://www.broadinstitute.org/gatk/guide/best-practices). Local realignment around InDels and base quality score recalibration were performed using GATK\[3\]\[4\], with duplicate reads removed by Picard tools\[5\]. The sequencing depth and coverage for each individual were calculated based on the alignments.

In addition, the strict data analysis quality control system(QC) in the whole pipeline was built to guarantee qualified sequencing data.

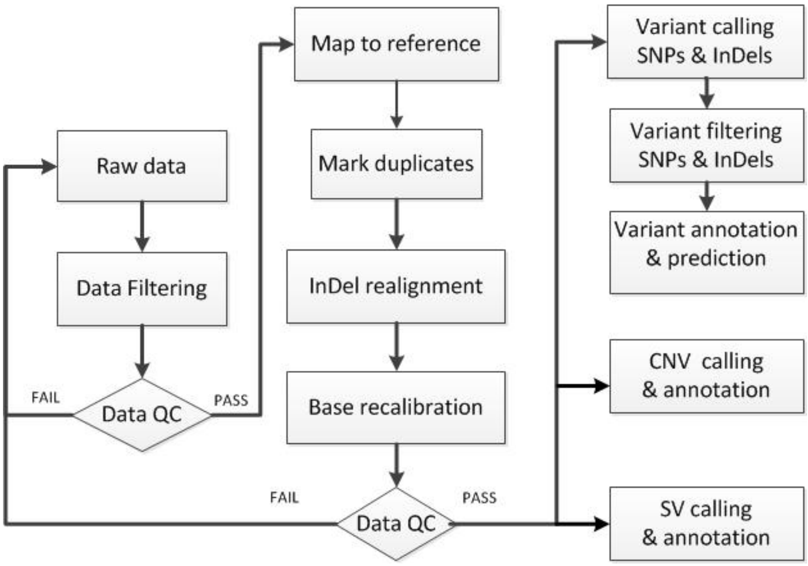

Figure 1 The whole genome sequencing analysis pipeline.

3 Data cleanup

In order to decrease noise of sequencing data, data filtering was done firstly, which included: (1) Removing reads containing sequencing adapter; (2) Removing reads whose low-quality base ratio (base quality less than or equal to 5) is more than 50%; (3) Removing reads whose unknown base ('N' base) ratio is more than 10%. Statistical analysis of data and downstream bioinformatics analysis were performed on this filtered, high-quality data, referred to as the " clean data ".

4 Mapping and marking duplicates

All clean reads were aligned to the human reference genome (GRCh38/HG38) using Burrows-Wheeler Aligner (BWA V0.7.15). We did mapping for each lane separately and also add the read group identifier, which by lane, into the alignment files. Here we used BWA-MEM method. Below are the BWA commands used for the alignments:

```
bwa mem -M -Y -R -t 16 'read_group_tag' hg38.fasta read1.fq.gz read2.fq.gz |
samtools view -Sb - > aligned_reads. BAM
```

Here the 'read_group_tag' need to be provided, e.g., '@RG\tID:GroupID\tSM:SampleID\tPL:illumina\tLB:libraryID'.

Picard-tools(v2.9.0)[5] was used to sort the SAM files by coordinate and converted them to BAM files.

```
java -jar picard-tools-2.9.0/SortSam.jar I=aligned_reads. BAM
O=aligned_reads.sorted. BAM  SORT_ORDER=coordinate
```

The same DNA molecules can be sequenced several times during the sequencing process. The resulting duplicate reads are not informative and should not be counted as additional evidence for or against a putative variant. We used Picard tools(v2.9.0)[5] to mark the duplicate reads , which were ignored in downstream analysis.

```
java -jar picard-tools-2.9.0/MarkDuplicates.jar \
   I=aligned_reads.sorted. BAM  \
   O=aligned_reads.sorted.dedup. BAM  METRICS_FILE=metrics.txt \
   CREATE_INDEX=truealigned_reads. BAM
```

Samtools(v1.3) was used to coverte BAM to CRAM files.

```
samtools -C -T hg38.fasta -o aligned_reads.cram aligned_reads. BAM
```

5 Base Quality Score Recalibration (BQSR)

The variant calling method heavily relied on the base quality scores in each sequence read. Various sources of systematic error from sequencing machines leaded to over or under-estimated base quality scores. So the BQSR step was necessary to get more accurate base qualities, which in turn improved the accuracy of variant calls. The following commands were used to do this step.

```
java -jar GenomeAnalysisTK.jar -T BaseRecalibrator \
   -R gatk_ref/hg38.fasta \
   -I aligned_reads.sorted.dedup.realigned. BAM  \
   -knownSites dbsnp_141.hg38. VCF  \
   -knownSites Mills_and_1000G_gold_standard. InDels .hg38. VCF  \
   -knownSites 1000G_phase1. InDels .hg38. VCF  \
   -o recal.table

java -jar GenomeAnalysisTK.jar -T PrintReads \
   -R gatk_ref/hg38.fasta \
   -I aligned_reads.sorted.dedup.realigned. BAM  \
   -BQSR recal.table -o aligned_reads.sorted.dedup.realigned.recal. BAM
```

6 Web Resources

The URLs for data presented herein and data format details are as follows:

UCSC build HG38, http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips

RefGene database, http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz

dbSNP, http://www.ncbi.nlm.nih.gov/snp

GATK database, ftp://ftp.broadinstitute.org/gsapubftp-anonymous/bundle/

1000 Genomes Project database, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release

SAM / BAM /CRAM file format, Sequence Alignment/Map Format Specification http://samtools.github.io/hts-specs/SAM

VCF format, http://www.1000genomes.org/wiki/analysis/vcf4.1

● Help

● FAQs

How to view BAM files in Microsoft Windows ?

Create index of BAM using Picard tools, named *.bai. Then open it with IGV.

Why do we use BWA-MEM?

BWA-MEM is designed for longer sequences ranged from 70bp to 1Mbp and split alignment. It is also the latest and is generally recommended for high-quality queries as it is faster and more accurate. The performance evaluation of BWA-MEM can be seen in the related paper (Li H. 2013, arXiv) http://arxiv.org/pdf/1303.3997v2.pdf.

What's the fragment length range of small InDel in exome and whole genome re-sequencing?

For small InDel , the range is from 1 to 50bp.

Base quality is not completely true, do we take this situation into consideration in variant calling?

Yes. Base Quality Score Recalibration step was used to correct raw base quality scores before calling variants by GATK package.

Do we use UnifiedGenotyper or HaplotypeCaller to call variants in GATK v3.3.0 in the pipeline?

Use HaplotypeCaller. The HaplotypeCaller is a more recent and sophisticated tool than the UnifiedGenotyper. Its ability to call SNPs is equivalent to that of the UnifiedGenotyper, its ability to call InDels is far superior, and it is now capable of calling non-diploid samples.

Compared to whole genome re-sequencing, exome sequencing only for the exon regions of DNA can be more simple, economical and efficient. Why should we select whole genome re-sequencing? And what's the sense?

The large structure variations and the mutations in non-exome region can be calling by whole genome sequencing, so that we will have a more comprehensive understanding of genome.

What is the purpose of checking genotype barcode of 21 SNP sites in data quality control step?

Genotype barcode of 21 SNP sites should be checked with sequencing data calls. We generate genotype barcode of 21 SNP sites by Sequenom for all the samples to track the identity of samples during the sequencing process.

● References

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

[2] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics, 26:589-595.

[3] DePristo MA. et al. (2011)A framework for variation discovery and genotyping using next generation DNA sequencing data. Nature genetics 43, 491-498.

[4] McKenna,A. et al. (2010)The Genome Analysis Toolkit: a MapReduce framework for analyzing next generation DNA sequencing data. Genome Research 20,1297-1303.

[5] Picard Tools (http://broadinstitute.github.io/picard/).

Technical Support E-mail:info@bgitechsolutions.com

Website: www.bgitechsolutions.com