

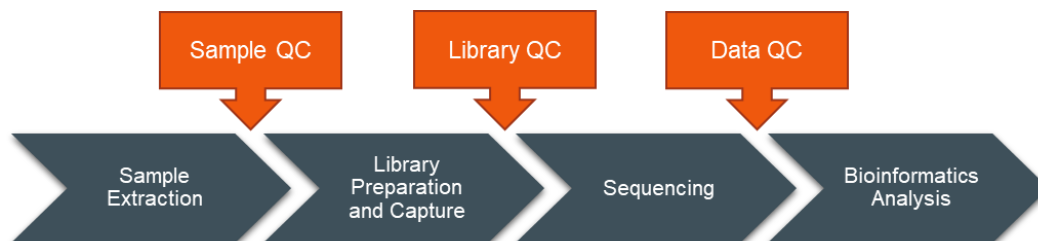
Whole Exome Sequencing Service Report

Contents

1. Library Preparation and Sequencing	3
1.1 DNA Sample QC	3
1.2 Library construction.....	3
1.3 Library QC	4
1.4 Sequencing	4
2. Bioinformatics Analysis Workflow.....	5
3. Project Data Presentation	5
3.1 FastQ files	5
3.2 Data QC.....	6
3.2.1 Data filtering	6
3.2.2 Data QC summary table.....	6
3.3 Mapping statistics.....	7
3.4 Variant calling	7
4. Data Delivery	8
4.1 Glossary of files.....	8
4.2 md5sum check.....	8

1. Library Preparation and Sequencing

Mirxes hWES service workflow includes multiple quality control (QC) checkpoints to ensure the quality of the data generated. An illustration of the workflow is as follows:



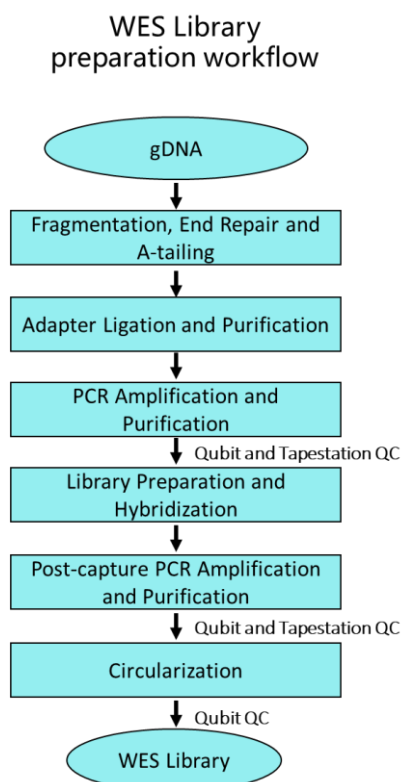
1.1 DNA Sample QC

To guarantee the reliability of sequencing data, all samples need to pass through the following sample quality control (QC) steps before library construction:

- 1) Nanodrop: measures OD260/280 and OD260/230 to check sample purity.
- 2) Qubit Fluorometer: quantifies DNA concentration, which is used to calculate the DNA input for library construction.
- 3) Agilent 4150/4200 TapeStation: checks RNA integrity by measuring DNA Integrity Number (DIN).

1.2 Library construction

The gDNA samples that passed sample QC are processed to generate DNA libraries. The experimental procedures are shown as below:

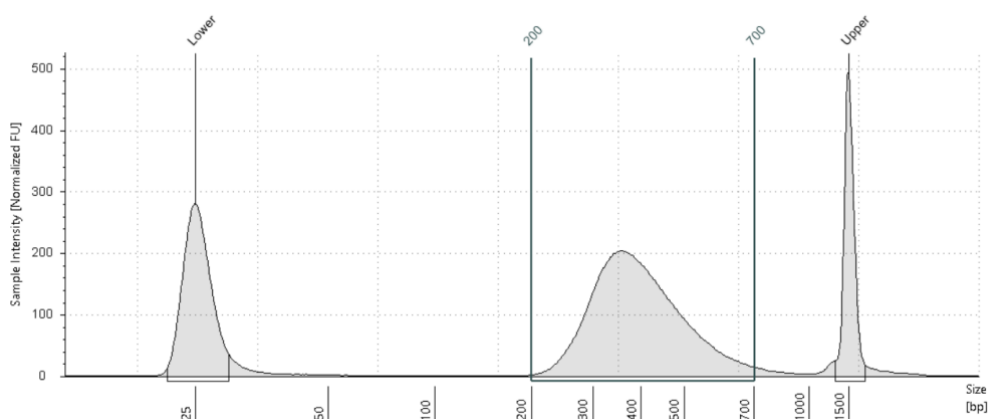


1.3 Library QC

Library concentration was measured by Qubit fluorometer. Libraries with concentrations more than 25ng/μl were considered as passing QC.

Library quality was analysed by Agilent 4150/4200 TapeStation. A high-quality library should present a narrow peak with library size around 350 – 425 bp.

Here is a representative graph from Tapestation:

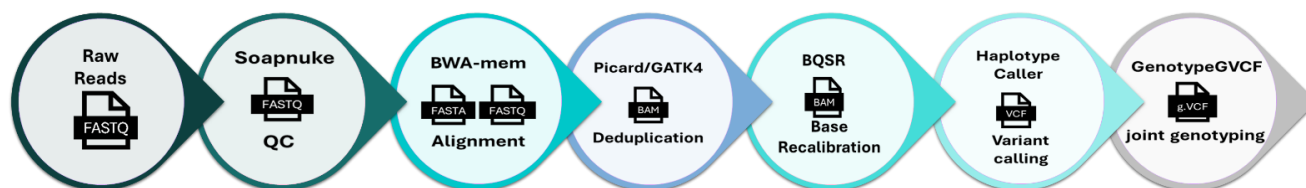


1.4 Sequencing

Libraries were sequenced on MGI DNBSEQ-T7 Sequencer.

2. Bioinformatics Analysis Workflow

The figure below summarizes the bioinformatics analysis workflow used for WES:



3. Project Data Presentation

3.1 FastQ files

The FastQ file is a text-based format for storing base calling information generated from the sequencer. An example of a single FastQ file record:

```
@E100026000L1C001R00100000007/1
ACTCTCGAGCACAGGATGCTGCAGGAGGAGAAGAGGCTTCGCACAGCCTATCTGCGTACAGAAGACTCTG
+
GFFFGGGEGGGFGGGFFGFGGFEFDGFEFFFFGGFFFFGFFBFGFFFFGFFGFFD!FFGFFFFGFGGG
```

Each record in a FastQ file consists of four lines:

Line1: Read sequence identifier. Begins with '@' character

Line2: Shows the nucleotide sequence of a single read

Line3: Quality score identifier and is always a "+" sign

Line4: Base quality scores for each nucleotide in the sequence shown in line 2

The number of records in a FastQ file equals the number of reads generated during a sequencing run. There are two FastQ files generated for paired-end sequencing run. The file names have this naming convention:

File Name Prefix	Description
LAB_ACCESSION_R1.fq.gz	R1 = File containing forward reads
LAB_ACCESSION_R2.fq.gz	R2 = File containing reverse reads

The FastQ files delivered are clean FastQ files after the data filtering process is performed as described in section 3.2.1.

3.2 Data QC

3.2.1 Data filtering

Raw data is filtered by removing adapter sequences and low-quality reads so that downstream analysis is performed on clean reads. The data is processed as follows:

SOAPnuke (PMID: 29220494) was used for quality control and pre-processing of FastQ files.

The parameters used for this step are as below:

Parameter	Description	Parameters Applied
-l / --lowQual	Low-quality threshold	12
-q / --qualRate	Low-quality rate threshold	0.5
-n / --nRate	N rate threshold	0.1
-f / --adapter1	Adapter sequence	AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA
-r / --adapter2	Adapter sequence	AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG

3.2.2 Data QC summary table

The table below shows data QC for 3 representative samples. A summarized table for “Total Data Output” and “Q30” is provided in the “Data Delivery Form”. The full list of data QC is provided separately in a file named **“WES_QC_Stats_PO7000077352_05_04_2023.xlsx”**.

External ID	Lab Accession Number	Total Data Output (Gb)	Clean Q30 (%)
ABC01	NA24631	8.92	91.65
ABC02	NA24694	8.37	91.04
ABC03	NA24695	7.22	91.81

- (1) External ID: the names or ID of samples provided by customer
- (2) Lab accession number: internal accession number assigned by Mirxes Genomics Lab
- (3) Total data output (Gb): total data output after filtering
- (4) Clean Q30 (%): It refers to the proportion of base number with Phred Quality Score more than 30 (error rate less than 0.1%) after filtering

3.3 Mapping statistics

Trimmed reads were aligned against the hg38 human reference genome (hg38.fa) with BWA-MEM algorithm. Picard was used as a duplicate marking program. GATK-BQSR was used for recalibration. The table below shows mapping statistics for 3 representative samples. The full alignment details can be found in a separately attached excel file "[WES_QC_Stats_PO7000077352_05_04_2023.xlsx](#)".

External ID	Lab Accession Number	Mapping Rate (%)	PE mapping Rate (%)	Duplication Rate (%)	Average depth
ABC01	NA24631	99.61	99.25	3.35	114.54
ABC02	NA24694	99.59	99.19	3.14	106.02
ABC03	NA24695	99.65	99.33	2.70	93.27

- (1) External ID: the names or ID of samples provided by customer
- (2) Lab accession number: internal accession number assigned by Mirxes Genomics Lab
- (3) Mapping rate: Number of mapped reads over total reads
- (4) PE mapping rate: Number of mapped paired reads over total reads
- (5) Duplication rate: Number of duplicate reads over mapped reads
- (6) Average depth: Average depth over the reference after removing duplicates

3.4 Variant calling

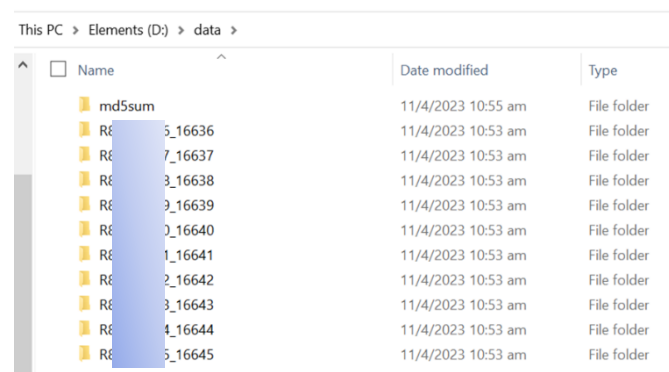
Germline variant calling:

GATK-haplotypeCaller was used for germline SNP/InDel calling. The deliverable from this will be a vcf file (as shown in section 4).

GATK-GenotypeGVCFs was used for joint genotyping calling. The deliverable from this will be a g.vcf file (as shown in section 4).

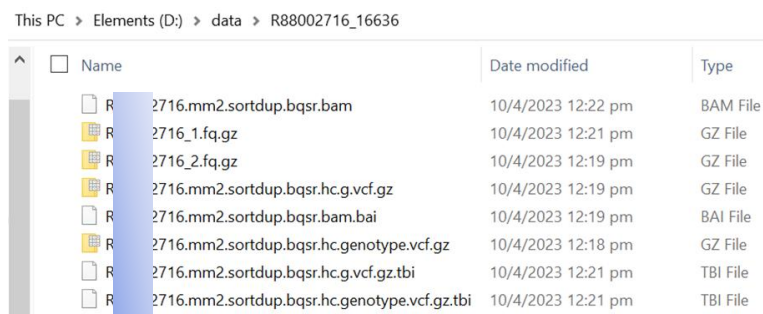
4. Data Delivery

The data is provided in an external hard drive under the data directory. A sample representation is shown below:



Name	Date modified	Type
md5sum	11/4/2023 10:55 am	File folder
R88002716_16636	11/4/2023 10:53 am	File folder
R88002716_16637	11/4/2023 10:53 am	File folder
R88002716_16638	11/4/2023 10:53 am	File folder
R88002716_16639	11/4/2023 10:53 am	File folder
R88002716_16640	11/4/2023 10:53 am	File folder
R88002716_16641	11/4/2023 10:53 am	File folder
R88002716_16642	11/4/2023 10:53 am	File folder
R88002716_16643	11/4/2023 10:53 am	File folder
R88002716_16644	11/4/2023 10:53 am	File folder
R88002716_16645	11/4/2023 10:53 am	File folder

Inside each of the directory, we will have multiple file types (as shown in the example below):



Name	Date modified	Type
R88002716.mm2.sortdup.bqsr.bam	10/4/2023 12:22 pm	BAM File
R88002716_1.fq.gz	10/4/2023 12:21 pm	GZ File
R88002716_2.fq.gz	10/4/2023 12:19 pm	GZ File
R88002716.mm2.sortdup.bqsr.hc.g.vcf.gz	10/4/2023 12:19 pm	GZ File
R88002716.mm2.sortdup.bqsr.bam.bai	10/4/2023 12:19 pm	BAI File
R88002716.mm2.sortdup.bqsr.hc.genotype.vcf.gz	10/4/2023 12:18 pm	GZ File
R88002716.mm2.sortdup.bqsr.hc.g.vcf.gz.tbi	10/4/2023 12:21 pm	TBI File
R88002716.mm2.sortdup.bqsr.hc.genotype.vcf.gz.tbi	10/4/2023 12:21 pm	TBI File

4.1 Glossary of files

The explanation of each file is as shown below:

File type	Description
*_1/2.fq.gz	fastq files read 1/2
*.bwa.sortdup.bqsr.bam	recalibrated bam file
*.bam.bai	index file for recalibrated bam
*.hc4.genotype.vcf.gz	variant call file from haplotypeCaller
*.hc4.genotype.vcf.gz.tbi	index of vcf file output from haplotypeCaller
*.hc4.g.vcf.gz	variant call file from GenotypeGVCFs
*.hc4.g.vcf.gz.tbi	index of g.vcf file output from GenotypeGVCFs

4.2 md5sum check

The data directory contains a sub-directory named as md5sum which contains the md5sum hash for all samples in the directory.

User can verify the md5sums using the following command on the terminal:

```
openssl md5 -binary <filename> | base64
```

Please contact genomics_techsupport@mirxes.com if you need further assistance.