Project 05 WGS Analysis Report

Author: Tianze (Vincent) Luo

Table of Contents

0. Preparations

- Interactive resource allocation: `srun -M teach --account=hugen2072-2025s -N 1 -c 24 -t 4:00:00 --pty bash` (max 24 cores)
  - `scancel <jobid>`
  - `scontrol show job <jobid>`
- Reads: `p5_1.fastq.gz`, `p5_2.fastq.gz`
- Reference sequence: `p5/Homo_sapiens_assembly38.fasta`

1. Read QC

1. Examine the files. Are they truly `fastq` files? Are the sequences and the base-calling quality score data in the same file? Are there paired reads in the files?

   > I have checked the head and tail for two read files. They are truly `fastq` files with both sequences and base-quality-score data included in the same file. Based on the head and tail, all the reads are paired.

2. What did I do here? Does it look like it's a single-read experiment or a paired-read experiment? Are the pairs together in the same file or in separate files?

   > `grep -c` was used to count the number of 1s and 2s in the corresponding paired-end read file. The number of reads with '/1' and '/2' are the same (n=377479648=total_line_number/4). Thus, it looks like this is a paired-end read experiment. The pairs are in separate files.

3. Does everything appear to be as it should? Are there any extra lines at the top? Is there any missing or truncated lines at the bottom? If the experiment has paired reads, do the read names match through the file? I.e, if readXXX/1 is the first read of the _1 file, is readXXX/2 the first read of the _2 file, and if readYYY/1 is the second read of the _1 file, is readYYY/2 the second read of the _2 file? That matching of read names should persist all the way to the end of the file.

> Everything appears to be as it should. There is no extra line at the top, and nothing truncated at the bottom. The line number of both files is a multiple of 4, indicating their integrity as `fastq` files.

> As indicated using `diff -s` in the code, the experiment is paired-end and the read names match through the files.

4. How many reads does each file have? Remember to take into account how many lines there are per read when figuring this value out. For paired reads, the _1 file and the _2 file should have the same number of reads before qc.

> 377479648 reads.

5. What is good and what is bad about the reads?

> The fastqc report before trimming showed that: for `p5_1.fastq.gz`, we have 377479648 reads with sequence length of 150. The average per base sequence quality is high; very clear Per tile sequence quality (no bubbles or other technical errors); avg per sequence quality peaks at 29; well-matched %A-T & %C-G; no N content; all with Sequence Length = 150; mostly have 1 duplicate; no overrepresented sequences. The only bad about it is that Per sequence GC content showed a bit sharp peak, but %GC=41, which is good for WGS.

> `p5_2.fastq.gz` is almost the same as its paired-end reads `p5_1.fastq.gz`. One more bad (but not severe) thing about it is that all its reads have worse qualities at the end of the reads, as indicated by lower Per base sequence quality plot and some lighter color at the end part of the Per tile sequence quality.

6. Post cutadapt, check the trimmed-reads fastqc results

> `1_trimmed.fastq.gz`: Total Sequences = 377194499 and Sequence length = 10-150; Quality similar to itself before `cutadapt`.

> `2_trimmed.fastq.gz`: Total Sequences = 377194499 and Sequence length = 10-150; Quality for bases at the end fixed. Others are similar to itself before `cutadapt`.

-------- End of script `WGS_pipeline_part1.slurm.sh` --------

2. Alignment

`{1,2}_trimmed.fastq.gz` -> (`.sam` {bwa} -> `.bam` ->) `sorted.bam`

3. Alignment Quality Control

`sorted.bam`

- -> `dupsmarked.bam` {gatk MarkDuplicatesSpark}
- -> `dupsmarked_cleaned.bam` {gatk BaseRecalibrator + gatk ApplyBQSR}
    - Produce alignment stats files

Review the output files from samtools.

1. How many total reads passed qc in your pipeline?

   > 757534121 QC-passed reads

   > `luo_submissionC_alignment_statistics.out`

2. How many were mapped and what percentage of qc-passed reads is that?

   > 756067316 mapped (99.81%)

   > `luo_submissionC_alignment_statistics.out`

3. What is the total qc-passed read depth and what is the mapped read depth?

   > No QC-failed reads --> total qc-passed read depth = mapped read depth; The averaged mapped read depth is **33.5823**.

```
#       calculate Avg.Mapped.ReadDepth
sum_perBaseMappedDepth=$(awk '{sum += $3} END{print sum}'
$dir_alignmentQC/depth_statistics.out)
n_bases=$(wc -l $dir_alignmentQC/depth_statistics.out)
avg_mappedDepth=$(awk -v sum="$sum_perBaseMappedDepth" -v
n="$n_bases" 'BEGIN{ print sum / n }')
echo "-------- Avg.Mapped.ReadDepth = $avg_mappedDepth --------"

#-------- Avg.Mapped.ReadDepth = 33.5823 --------
```

4. Genotyping/Variant-Calling

`dupsmarked_cleaned.bam`

- -{gatk HaplotypeCaller}-> `genotypes.g.vcf.gz`
- -{gatk GenotypeGVCFs}-> `genotypes.vcf.gz`

Notes about {gatk HaplotypeCaller}

- 61,645,801 read(s) filtered by: MappingQualityReadFilter
- 65,233,097 read(s) filtered by: NotDuplicateReadFilter
- 126,878,898 total reads filtered out of 757,459,485 reads processed

5. Genotype QC

`genotypes.vcf.gz`

- -{gatk MakeSitesOnlyVcf + gatk VariantRecalibrator + gatk ApplyVQSR}-> `final.vcf.gz`

- Produce variant calling metrics using {gatk CollectVariantCallingMetrics}
  - `genotype_metrics.variant_calling_detail_metrics`
  - `genotype_metrics.variant_calling_summary_metrics` (VCF-wise metrics)

Check `variant_calling_summary_metrics` —> VCF-wise

Q1: How many snps were called? How many indels?

> TOTAL_SNPS = 3,266,424
>
> - NUM_IN_DB_SNP = 3,231,938
> - NOVEL_SNPS = 34,486

> TOTAL_INDELS = 756,811
>
> - NUM_IN_DB_SNP_INDELS = 723,573
> - NOVEL_INDELS = 33,238

Q2: What is the Ts/Tv (transition/transversion) ratio for this call set?

> DBSNP_TITV = 2.073715

> NOVEL_TITV = 1.263306

Q3: What is the indel (insertion/deletion) ratio for this call set?

> DBSNP_INS_DEL_RATIO = 0.984485

> NOVEL_INS_DEL_RATIO = 1.075818

Q4: Are these statistics consistent with what we would expect for sequencing? Do you think these reads are from whole-exome or whole-genome sequencing

> From this sequencing experiment, we got a known SNP TI/TV Ratio of ~2.07, a known InDel ratio of ~0.98, and a total of ~4.0M (= 3,266,424 + 756,811) variants per sample. These metrics **MATCH** with what we would expect for **whole-genome sequencing data**.

-------- End of script `WGS_pipeline_part2_24cores.slurm.sh` --------