

Data Analysis Report

Poduval Deepak Balakrishnan

24th September 2020

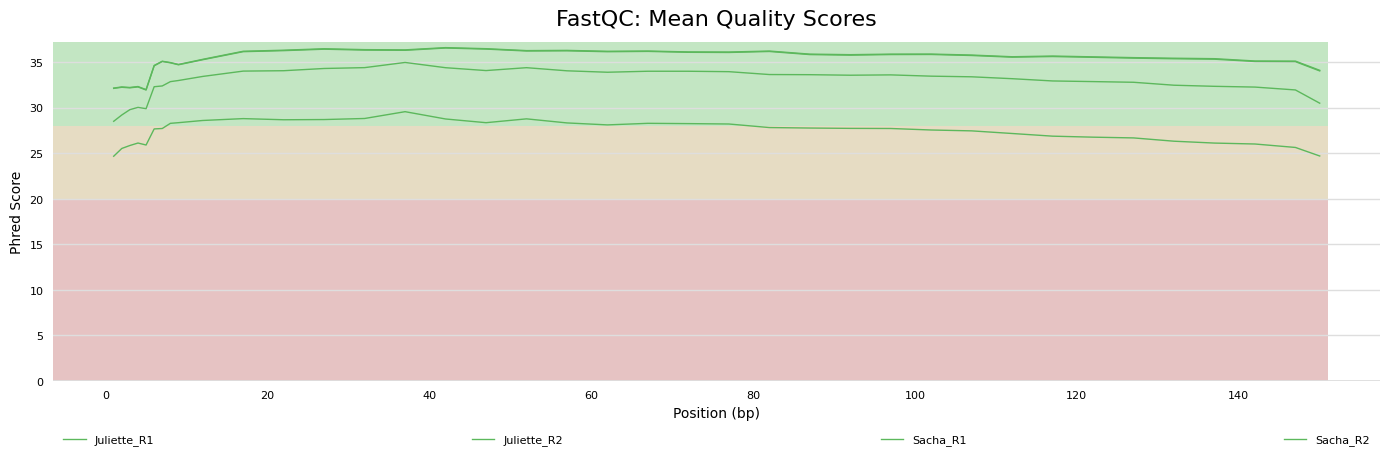
**Sequence Data QC**

The sequence data was generated using Illumina HiSeq. Data quality was checked using FastQC [1] and MultiQC [2] software. The data was checked for base call quality distribution, % bases above Q20, Q30, %GC, and sequencing adapter contamination (Table 1; Figure 1-2). From the Table1, Q20 looks good, but Q30 seems to be lower. From the figure 2, we can see that Juliette sample had read 2 with per sequence qualityt lower than 20.

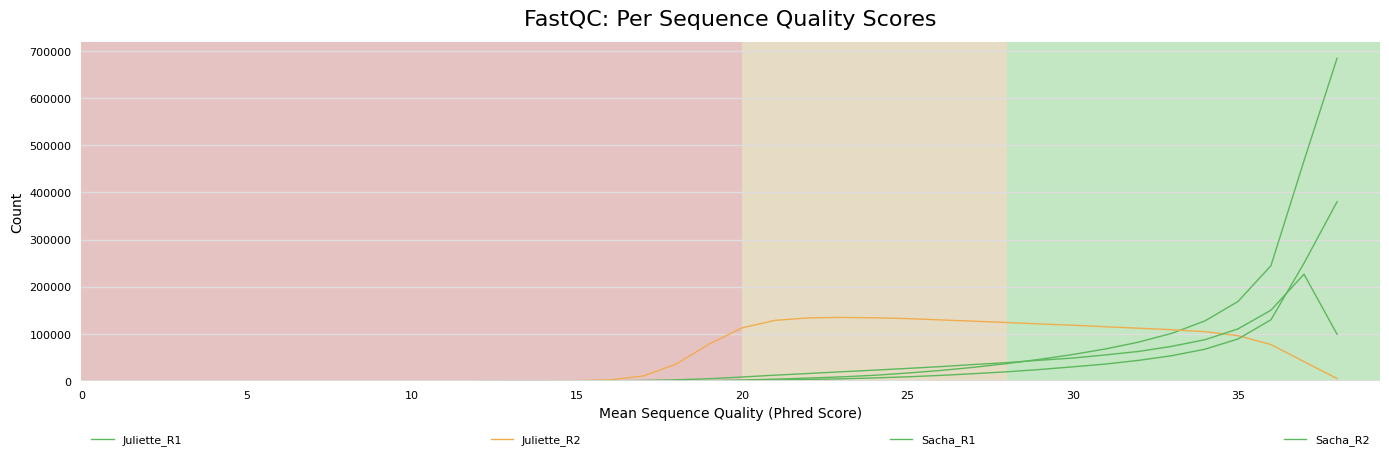
**Table 1.** Summary of Raw sequence data and quality.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample ID | No. of Reads | Read length | GC% | %Q20 | %Q30 |
| Juliette | 4373318 | 151 | 47 | 97.005 | 63.550 |
| Sacha | 2357662 | 151 | 47 | 99.580 | 84.615 |

**Figure 1.** Average per base quality

****

**Figure 2.** Histogram per Sequence Quality Scores.



**Quality Trimming**

The sequence data was processed using fastp [3] to remove adapter sequences and low quality bases.

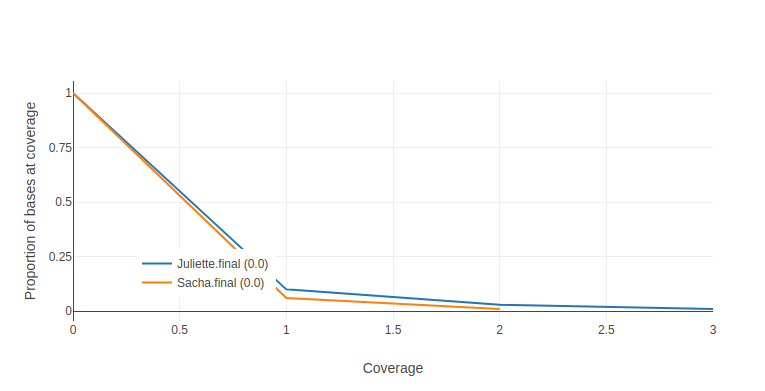
**Genome Alignment and Structural Variant Calling**

Quality trimmed reads were mapped to Human (GRCh38) genome, using bwa-mem algorithm with default parameters [4] (Table 3).The PCR duplicates were marked and removed using sambamba [5]. The depth was calculated using mosdepth [6] program (figure 3) as well as Qualimap [7] (Figure 4). Variants were called using freeBayes v1.0.2 with default parameters [8] and filtered using vcftools with minimum quality of 20 [9].

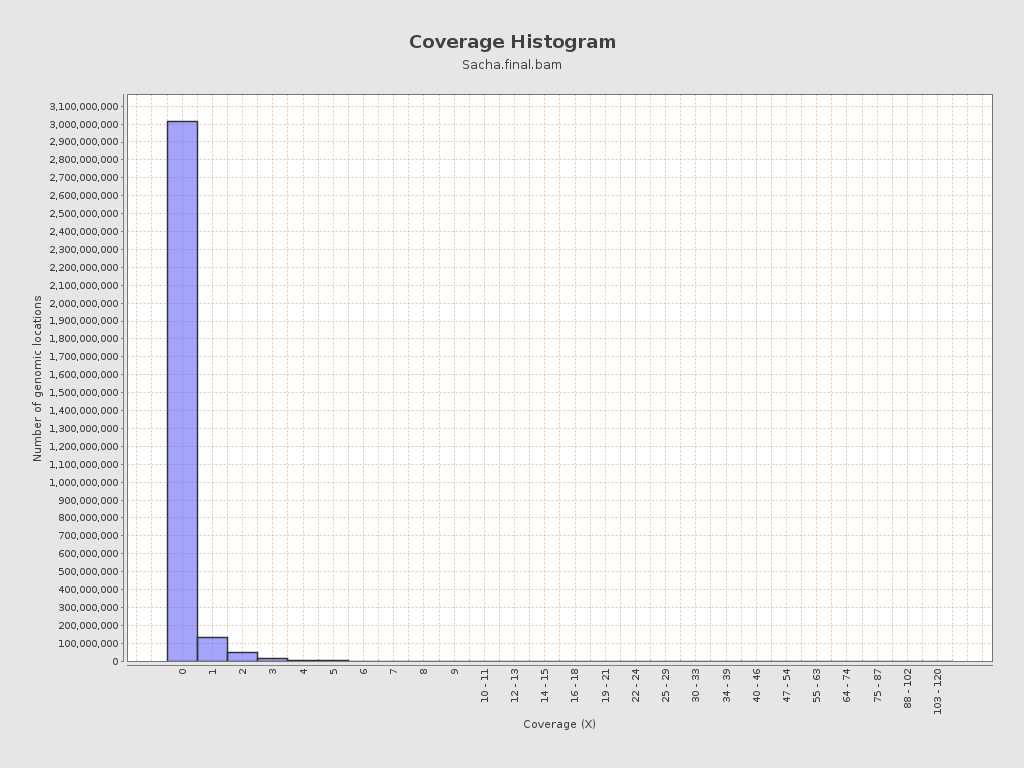
**Table 3.** Alignment summary statistics.

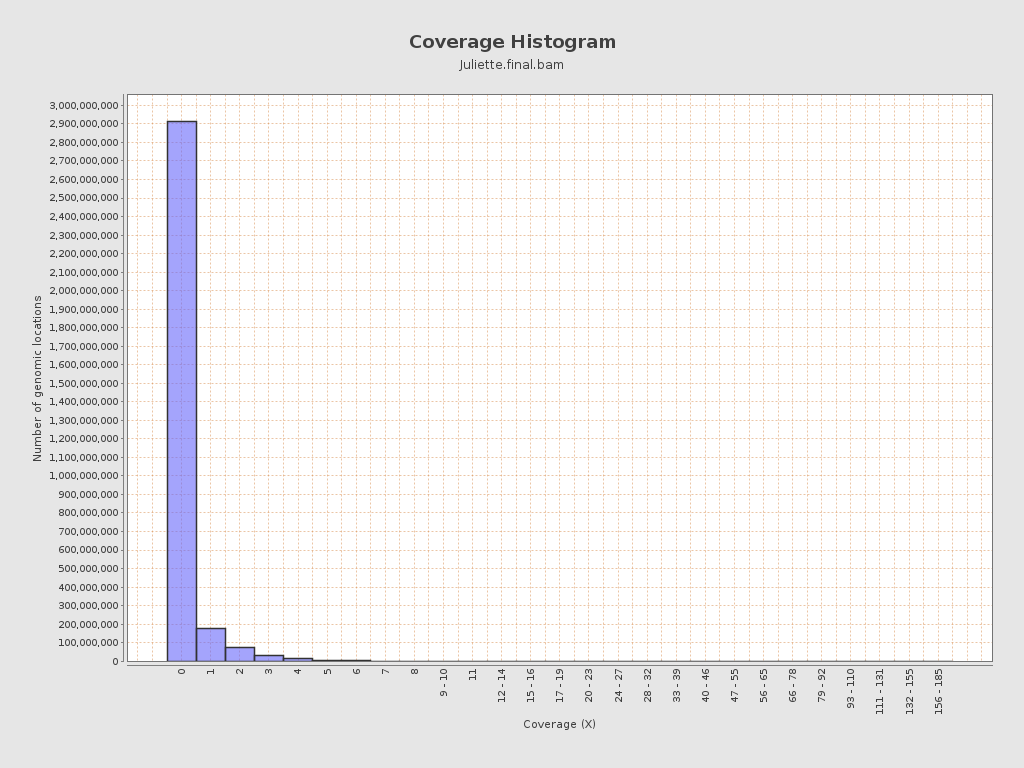
|  |  |  |
| --- | --- | --- |
| **Strains** | Juliette | Sacha |
| Total reads | 4310119 | 2339181 |
| Mapped Reads | 4309063 | 2338849 |
| % mapped reads | 99.98 | 99.99 |

**Figure 3. Coverage Histogram from Mosdepth**

****

**Figure 4. Coverage Histogram from Qualimap**

****



This is how far I could manage to do in the time available. Next step is to use SnpEff and SnpSift to annotate the variants [10] with respect suitable databases such as dbSNP and Clinvar to answer the questions. I have also used Pindel [11] to predict structural variations.

**Reference**

1. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 24 May 2017.
2. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
3. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
5. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.
6. Brent S Pedersen, Aaron R Quinlan, Mosdepth: quick coverage calculation for genomes and exomes, Bioinformatics, Volume 34, Issue 5, 01 March 2018, Pages 867–868, <https://doi.org/10.1093/bioinformatics/btx699>
7. Konstantin Okonechnikov, Ana Conesa, Fernando García-Alcalde, Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data, Bioinformatics, Volume 32, Issue 2, 15 January 2016, Pages 292–294, <https://doi.org/10.1093/bioinformatics/btv566>
8. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-bio]. 2012. http://arxiv.org/abs/1207.3907. Accessed 24 May 2017.
9. [*The Variant Call Format and VCFtools*](http://dx.doi.org/10.1093/bioinformatics/btr330), Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011
10. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6:80–92.
11. Kai Ye, Marcel H. Schulz, Quan Long, Rolf Apweiler, Zemin Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads, Bioinformatics, Volume 25, Issue 21, 1 November 2009, Pages 2865–2871, <https://doi.org/10.1093/bioinformatics/btp394>