Re-analysis of Whole Genome data for *Plasmodium falciparum*

**About Team**

Aug 14, 2021 |*Team Lead - @Sadaf | Technical Support - @vinyjoshi*

Our team - **Collins** - is named after the famous geneticist Francis Sellers Collins [[1]](https://en.wikipedia.org/wiki/Francis_Collins), who led the Human Genome Project. Team-Collins, comprising 19 enthusiastic members passionate about the field of genomics (biostack), gathered from different countries to collaboratively make this task a success.

**Task Objective**

In Stage 1 of HackBio Internship\_2021, our task was to perform analysis on a genomic dataset using various bioinformatic tools of our choice. Malaria is a disease caused by parasitic protozoans in the genus *Plasmodium*, which are transmitted between vertebrate hosts by female mosquitoes in the genus *Anopheles*. The predominant malaria parasite species that infect humans are Plasmodium falciparum (Neafsey et al., 2021) and our team agreed to work on *Plasmodium falciparum* (*P. falciparum*) [[2]](https://en.wikipedia.org/wiki/Plasmodium_falciparum). The objective of our analysis was to evaluate variations in the genome sequence with the reference by [[3]](https://www.ebi.ac.uk/ena/browser/view/PRJNA603776?show=reads) performing a variant call analysis.

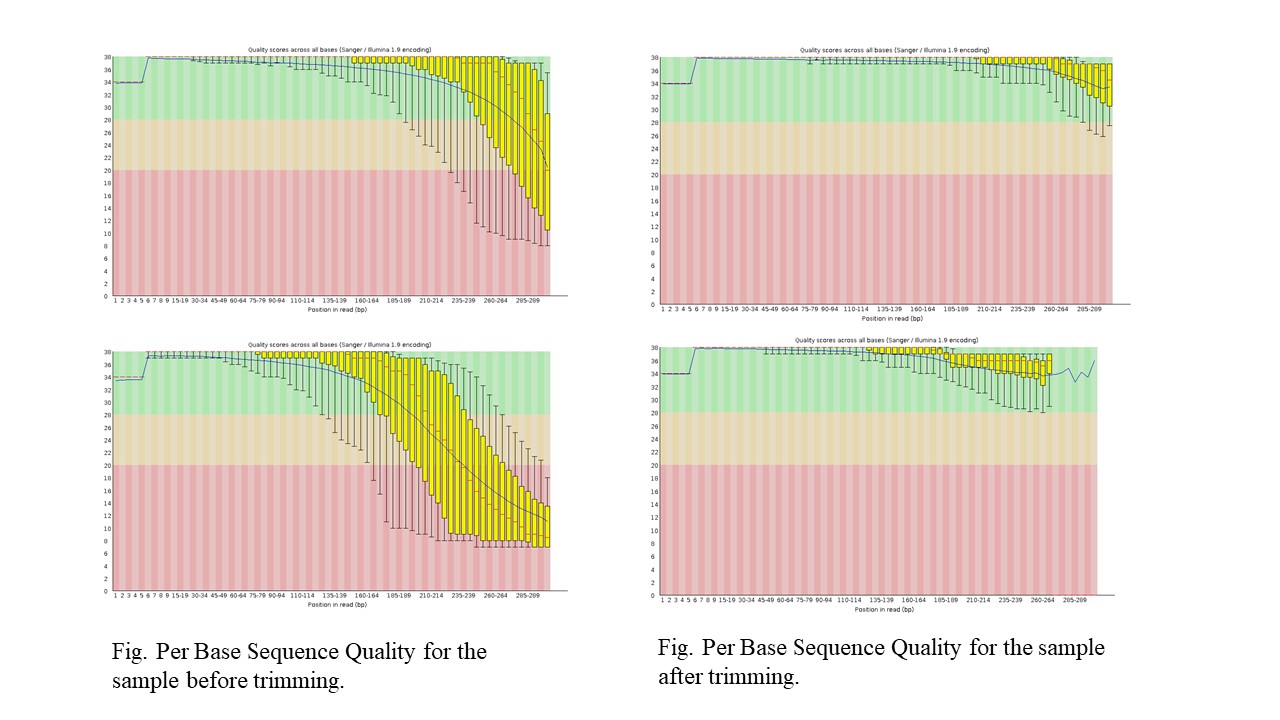
**Analysis**

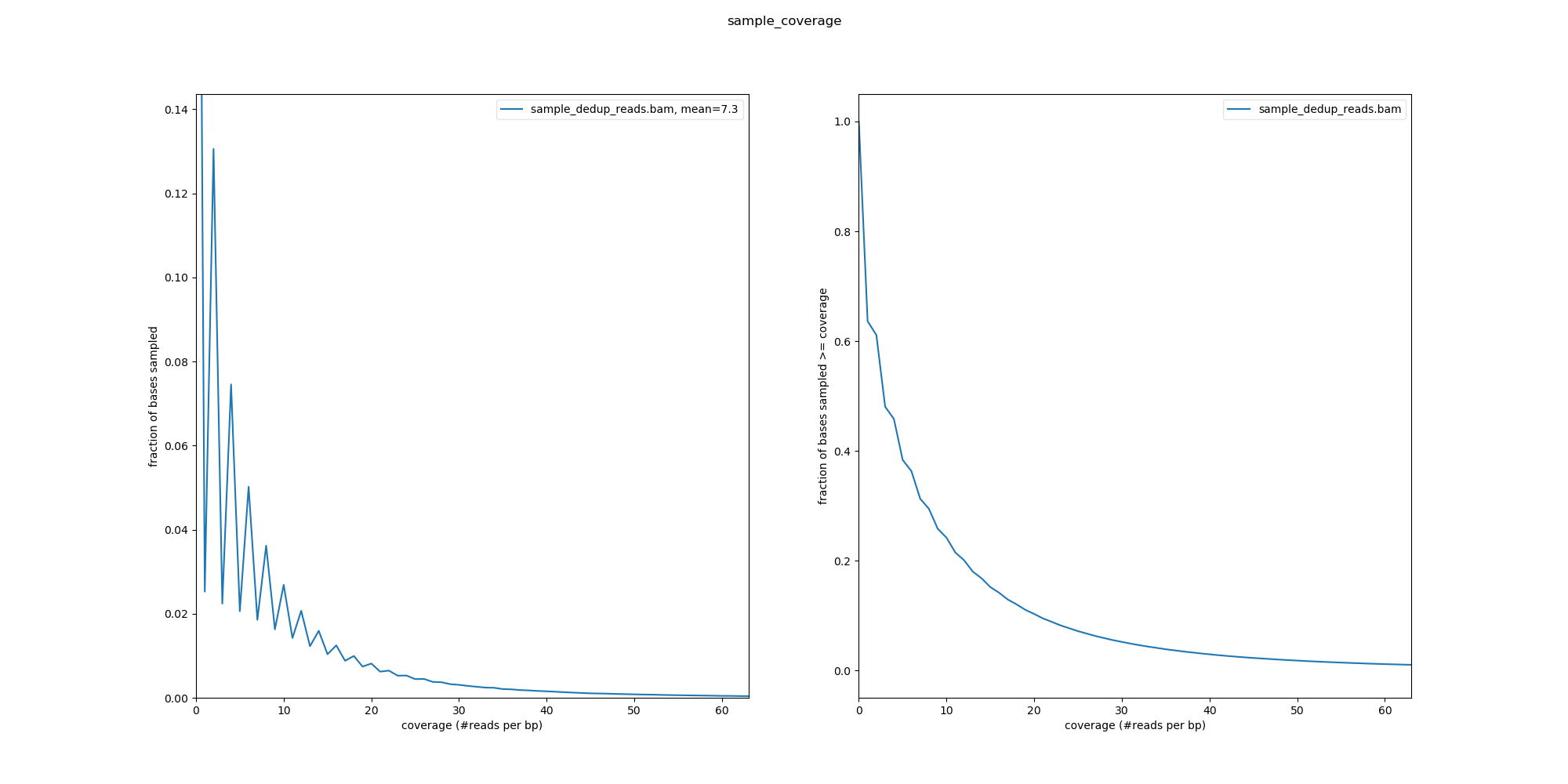
We started our analysis by performing a quality assessment on the dataset generating a FastQC report and then filtering out poor-quality reads to re-assess the quality to get a consolidated MultiQC report from all the samples. Following this, indexing of the reference genome that we retrieved from the PlamoDB database was done, which enables locating possible alignment sites for query sequences in a genome, rapidly saving time during alignment. Later, reads were mapped to the reference genome, and we got a SAM file. Further, this SAM file output was sorted by various SAM attributes generating a BAM output file, which was then executed for the AddOrReplaceReadGroups command to align all the reads to a single new read-group. Next was to validate our BAM file to a specific format to ensure that no errors appear later due to improper formatting. Subsequently, any duplicate reads were identified and characterized as being derived from a single piece of DNA. Followed by base recalibration, variant calling was executed using Haplotype caller to obtain a raw VCF file, for our sample. Variant annotation was then performed on the vcf file to annotate the dataset and identify genetic variation from the ‘reference’ genome of P.falciparum.

| **STEPS** | **TOOLS** |
| --- | --- |
| EC2 Instance account created for team members across different countries to collaborate on a common workspace. | AWS Management Console [[4]](https://signin.aws.amazon.com/signin?redirect_uri=https%3A%2F%2Fus-east-2.console.aws.amazon.com%2Fec2%2Fv2%2Fhome%3Fregion%3Dus-east-2%26state%3DhashArgs%2523ConnectToInstance%253AinstanceId%253Di-021bcaf5660658f26%26isauthcode%3Dtrue&client_id=arn%3Aaws%3Aiam%3A%3A015428540659%3Auser%2Fec2&forceMobileApp=0&code_challenge=qnlu9saBiYWIGEubdFjrwSNpBrI9Pcdj9VhLFdJYxU4&code_challenge_method=SHA-256) |
| Sample and Reference genomic datasets were collected for *P. falciparum*.  **Accession No. :** [**PRJNA603776**](https://www.ebi.ac.uk/ena/browser/view/PRJNA603776?show=reads)  **Sample ID:** [**SRR12740661**](https://plasmodb.org/plasmo/app/downloads/Current_Release/Pfalciparum3D7/) **[5]** | European Nucleotide Archive (ENA)/ PlasmoDB |
|
| Quality checks on raw data | FastQC [6] |
|
| Trimming out poor-quality reads. | Trimmomatic [7] |
|
| Re-assessing the quality on FastQC. | MultiQC [8] |
| Indexing the reference genome, and aligning the reads. | Burrows-Wheeler Aligner (BWA) [9] |
|
| Sorting of SAM Files  Add or Replace Read Groups  Validation of BAM files  Mark Duplicates  Building BAM Index | Picard [10] |
|
| Base Recalibration | GATK-Base Recalibrator [10] |
| Annotation and Filtration of Raw Variants | SamTools [11]  SnpEff [12] |

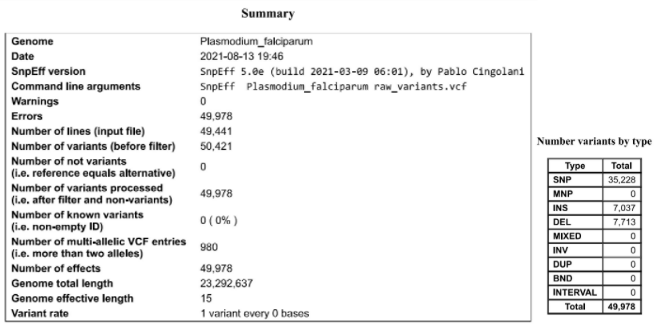
**Result and Discussion**

The aim of our analysis was to observe any genetic variation in the datasets from the reference genome we acquired. After executing the workflow, we were able to generate our final variant annotation report. The coverage for our sample was found to be low. According to the snpEff results, a total of 49,978 variants were present, attributed mainly to Single Nucleotide Polymorphisms, insertion, and deletion mutations. However, we did not observe any known variants in our analysis.









**Conclusion**

The completion of this analysis is attributed to the dedicated contribution of the majority of our team members. We demonstrated the known variability on the *P. falciparum* genome using various bioinformatic analyses on pre-existing data and did not find any known variants. We intend to gain more insights into the mutability of the protozoan by including more samples in future.

**Supplementary Material**

The analytical reports, along with the video presentation can be found in the link below -

<https://drive.google.com/drive/folders/17c4NujWtEIVH6f0TcPQXa6sY121Vtq6J>

**References**

1. <https://en.wikipedia.org/wiki/Francis_Collins>
2. <https://en.wikipedia.org/wiki/Plasmodium_falciparum>
3. <https://www.ebi.ac.uk/ena/browser/view/PRJNA603776?show=reads>
4. <https://aws.amazon.com/>
5. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ Jr, Treatman C, Wang H. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009 Jan;37(Database issue):D539-43. doi: 10.1093/nar/gkn814. Epub 2008 Oct 28. PMID: 18957442; PMCID: PMC2686598.
6. FastQC - Andrews, S. FASTQC. A quality control tool for high throughput sequence data. (2010).
7. Trimmomatic - Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
8. MultiQC - Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.
9. BWA - Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]
10. GATK - Van der Auwera GA & O'Connor BD. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media
11. Samtools - Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.
12. Snpeff - P. Cingolani *et al.*, “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff,” *Fly*, vol. 6, no. 2, pp. 80–92, Apr. 2012, doi: 10.4161/fly.19695

**Contribution**

| **CONTRIBUTION** | **SLACK USERNAME** |
| --- | --- |
| **Workflow and Conceptualization** | @Sadaf, @Mosopefoluwa |
| **Software Installation** | @vinyjoshi |
| **Data Collection** | @Sadaf |
| **Data Analysis** | @Sadaf, @vinyjoshi, @Rajarshi\_Mondal, @Nasrin, @Nuiter, @ZubairAlam, @Micholufemi, @Aditi, @Jeeel193 @Adedoyinsoye, @Mosopefoluwa, @Preeti |
| **Synopsis** | @Preeti, @Micholufemi, @Mosopefoluwa |
| **Video Presentation** | @IshreshthA, @Shabnam, @Jeeel193, @-anirudh1009-, @vinyjoshi, |
| **Tweet** | @Grace, @Nuiter, @bodeekerin |