Estimating the micro-indel mutation rate in *Plasmodium falciparum* using genomes from mutation accumulation experiments

ABSTRACT

Malaria is a life threatening disease caused by parasites of the *Plasmodium* genus. Infection with *P. falciparum* single handedly accounts for more than 90% of the world's malaria mortality. The global fight against malaria has repeatedly been compromised by drug-resistant *P. falciparum* strains that first emerged in South East Asia. Despite Africa bearing the largest disease load, South East Asia has been the hotspot of the evolution of drug resistance in malaria. *P. falciparum* has a unique genome that is eighty percent AT rich, providing adequate opportunities for mutations. The underlying mutations aid in the selection of drug-resistant strains in an environment where drugs are prevalent. I hypothesized that Asian strains have a higher rate of micro-indel mutations as compared to the African strains of *P. falciparum.* To test the hypothesis, I used data from Claessens et al. (2014) who generated six clone trees of *P. falciparum* strains that belong to four geographically distinct regions. Hamilton et al (2016) used the same data to calculate SNP mutation rate in the six strains. I extended the experiment to calculate the micro-indel mutation rate, which could now be appropriately calculated after the publication of twelve newly assembled PacBio reference genomes for *P. falciparum* strains, which also include the strains that were used for clone tree generation. I created a GATK-based pipeline to detect, for the first time, micro-indels in non-3D7 strains. My analysis suggests that the micro-indel mutation rate of 3D7, the strain from Africa, has a slightly lower mutation rate than Dd2 and W2, the strains with Asian origin.

1. INTRODUCTION

Malaria is a life-threatening disease spread by mosquitoes. According to the WHO, malaria infected around 229 million people worldwide, causing 409000 deaths, most under the age of 5. The disease is caused by five protozoa: *Plasmodium falciparum*, *P. vivax*, *P. malariae*, *P. ovale,* and most recently implicated *P. knowlesi*. Infections with *P. falciparum* account for more than 90% of the world's malaria mortality and therefore remain an important threat to public health on a global scale (Snow, 2015). *P. falciparum* has a complex dixenous life cycle (Fig 1) that occurs in two hosts. Male and female gametocytes mate in a female anopheles mosquito, quickly followed by a round meiosis. The rest of the life cycle, *P falciparum* is haploid, with most mitosis cycles occurring during the intra-erythrocytic stage.

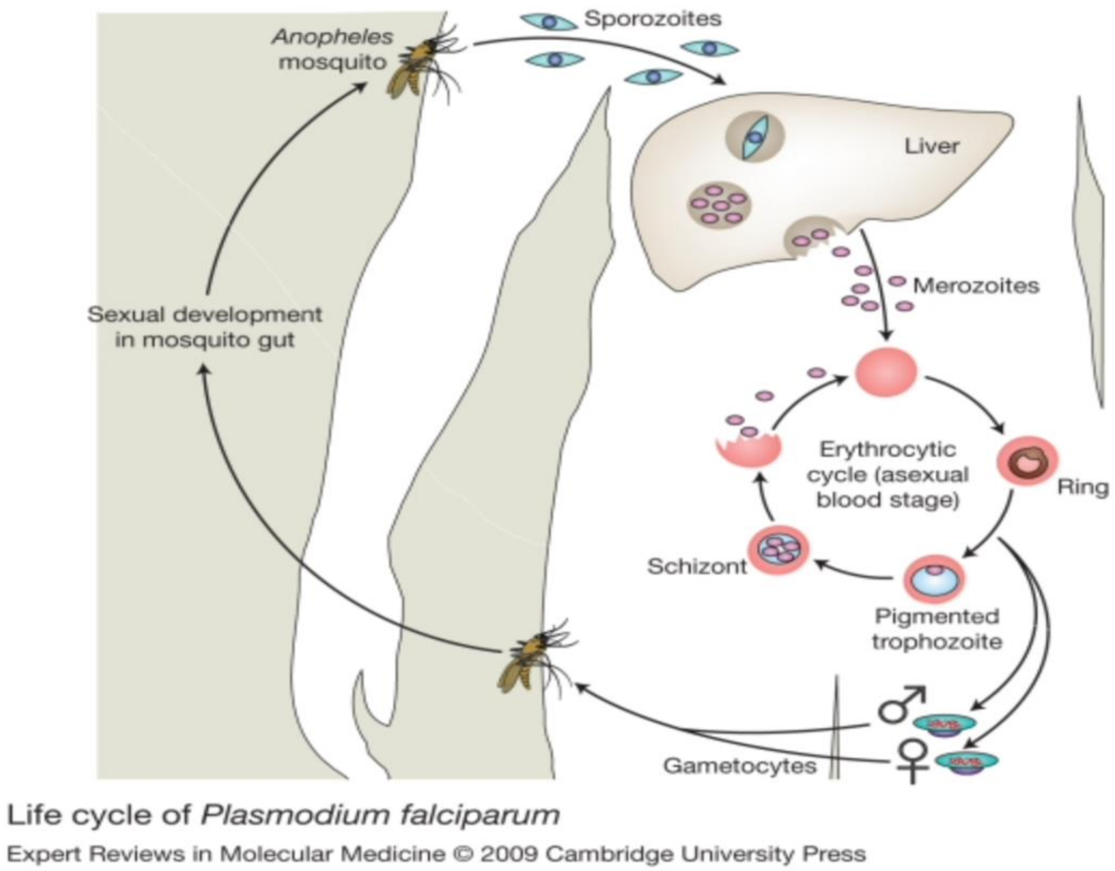

Anopheles mosquito
Sporozoites
Liver
Merozoites
Sexual development in mosquito gut
Erythrocytic cycle (asexual blood stage)
Ring
Schizont
Pigmented trophozoite
♂
♀
Gametocytes
Life cycle of *Plasmodium falciparum*
Expert Reviews in Molecular Medicine © 2009 Cambridge University Press

Fig 1. Life cycle of *Plasmodium falciparum.* When an infected female *Anopheles* mosquito takes a blood meal, sporozoite forms of *P. falciparum* are injected into the human skin. The sporozoites migrate into the bloodstream and then invade liver cells. The parasite grows and divides within liver cells for 8–10 days, then daughter cells called merozoites are released from the liver into the bloodstream, where they rapidly invade erythrocytes. Merozoites subsequently develop into ring-stage, pigmented-trophozoite-stage and schizont-stage parasites within the infected erythrocyte. *P. falciparum*-infected erythrocytes express parasite-derived adhesion molecules on their surface, resulting in sequestration of pigmented-trophozoite and schizont stages in the microvasculature. The asexual intraerythrocytic cycle lasts for 48 hours, and is completed by the formation and release of new merozoites that will re-invade uninfected erythrocytes. It is during this asexual bloodstream cycle that the clinical symptoms of malaria (fever, chills, impaired consciousness, etc.) occur. During the asexual cycle, some of the parasite cells develop into male and female sexual stages called gametocytes that are taken up by feeding female mosquitoes. The gametocytes are fertilized and undergo further development in the mosquito, resulting in the presence of sporozoites in the mosquito salivary glands, ready to infect another human host. (Rowe et al., 2009)

Mutation is defined as an alteration in a DNA sequence. Most mutations occur naturally when DNA fails to replicate accurately. The fidelity of DNA replication is a crucial determinant of genome stability and is central to the evolution of species (Kunkel and Bebenek, 2000). DNA usually replicates with a very high fidelity, with one mismatch nucleotide incorporated once per 10^8 to 10^{10} nucleotides polymerized (Kunkel and Bebenek, 2000). Errors may happen when polymerase enzymes sometimes insert a different nucleotide or too many or too few nucleotides into a sequence. However, most of these replication errors are fixed through various DNA repair processes. Repair enzymes recognize structural imperfections between improperly paired nucleotides, cutting out the wrong ones and putting the right ones in their place. But some mutations make it past the proofreading mechanisms, thus becoming permanent variants (Pray, 2008).

Mutations could be in the form of a single nucleotide variation, known as Single Nucleotide Polymorphisms (SNPs) which are considered the most common forms of nucleotide modifications

(Collins et al., 1998). Other mutations include insertions and deletions of nucleotide bases that occur during a process called slipped strand mispairing or polymerase slippage. These variants, hereafter referred to as ‘micro-indels’, underlie polymorphic variations in short tandem repeats (STRs) that are observed between individuals of a species (Sehn, 2015). Mutations are thus a source of variation with diverse consequences that can be beneficial, adverse or completely neutral. Evolutionary selection pressure then acts on diverse DNA sequences that are generated by the replication errors. Variants within a species provide an opportunity to adapt to changing environmental conditions. The frequency by which mutations occur is at the heart of evolutionary diversification and population viability (Griffiths et al., 2020). The strongest recent evolutionary pressure on the *P. falciparum* population is arguably the usage of antimalarial drugs. First large-scale administration of antimalarials started in the 1950s with Chloroquine, a cheap and then-effective drug (Bronzwaer et al., 2002). And soon, Chloroquine-resistant forms of *P. falciparum* malaria first appeared in Thailand in 1957 (Packard, 2007). They then spread through South and Southeast Asia and by the 1970s were being seen in sub-Saharan Africa and South America. Quinine resistance in *P. falciparum* was first documented in human volunteers in Brazil and in South East Asia in the 1960s (Peters, 1970). Clinical evidence of parasites resistant to mefloquine began to appear in Asia around the time of the drug's general availability in 1985 (Hoffman et al., 1985). After generating parasites resistant to chloroquine, sulfadoxine, pyrimethamine, quinine, and mefloquine, the South East Asian region has now spawned parasites resistant to artemisinins – the world's most potent antimalarial drug. Artemisinin resistance was first reported from Pailin, Western Cambodia, in 2009 (Dondorp et al., 2009).

It is interesting to note that almost all drug resistant alleles were first reported in South East Asia despite Africa carrying the large majority of the disease burden. At the parasite genomic level, *P. falciparum* genomes from Africa and Asia are clearly distinct, with thousands of SNPs being specific to a continent (Manske et al., 2012). One of the most important unanswered questions in anti-malarial drug resistance is why it has repeatedly emerged in South East Asia. One standing hypothesis to explain this phenomenon is that South East Asian parasites display a faster mutation rate than the African *P. falciparum* parasites, in turn promoting the emergence and spread of drug resistance.

Claessens et al. (2014) used an experimental evolution approach to systematically characterize the different types of mutations and the rate at which they occur in different strains of *P. falciparum* during the asexual intra-erythrocytic cycle (Fig 1) within the human host. They generated six ‘clone trees’ from culture-adapted *P. falciparum* strains from four geographically distinct regions (Table 1) . A clone tree involves regular cloning of parasites every 20 to 30 cycles of replication. One clonal population was arbitrarily selected for the next round of cloning to produce the next ‘generation’. Six clone trees resulted in a total of 284 clonal populations, each one of them was whole genome sequenced from PCR-free libraries. Hamilton et al. (2016) also used this dataset to calculate the Single Nucleotide Polymorphism (SNP) mutation rate (2.45 × 10^{10} substitution per nucleotide per erythrocyte life cycle) which is relatively constant across strains (1.64 × 10^{10} substitution per nucleotide per erythrocyte life cycle in KH-02 to 3.20 × 10^{10} substitution per nucleotide per erythrocyte life cycle in in Dd2) . Hamilton et al. (2016) calculated the micro-indel mutation rate for the 3D7 strain and found that the micro-indel mutation rate was almost 10 times higher (21.1×10^{10} micro-indels per nucleotide per erythrocyte life cycle). The calculation of the micro-indel mutation rate had to be limited to that single strain, 3D7, from which the "reference genome" had been derived. However, since then 12 new reference genomes of *P. falciparum* strains, including all 6 that were used for building clone trees, have been assembled from PacBio long reads data and have recently been published by Otto (2018). Hence, we should now be able to estimate the micro-indel mutation rate in the six strains with the clone trees, by using the appropriate reference genome.

To extend the experiment to calculate the micro-indel mutation rate of the six *P. falciparum* strains, I first created a pipeline, using the Genome Analysis toolkit (GATK) (Poplin *et al.* 2017) best practices pipeline. The micro-indels discovered by Hamilton et al. (2014) acted as a control to test the validity of the pipeline. With this unique dataset, I addressed the following biological questions, 1. Is the micro-indel mutation rate similar across *P. falciparum* strains? 2. If not, is there a correlation between the mutation rate and the geographical origin of the strain? Could different mutation rates of *P. falciparum* strains from South-East Asia and Africa explain the higher predisposition of South-East Asian strains to evolve drug resistance? 4. What types of micro-indels, in terms of length, AT-richness, chromosomal location etc., can we detect in each clone tree?

Improved understanding of the process of how and why anti-malarial drug resistance emerges in South East Asia could provide critical information in developing strategies to prevent the spread of the current wave of artemisinin resistance.

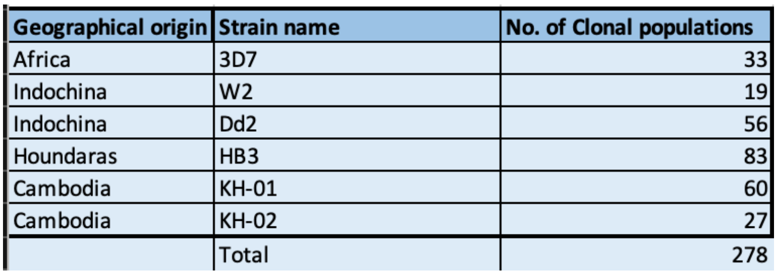
2. MATERIALS AND METHODS

Parasite *in vitro* culture and clone tree generation

Prior to my internship, Claessens et al.(2014) and Hamilton et al.(2016) cultured six distinct *P. falciparum* strains (3D7, W2, Dd2, HB3, KH-01, KH-02) and obtained specific clone trees for each of these strains.

All *P. falciparum* strains were cultured in human O+ erythrocytes. A limiting dilution of one culture led to the random sampling of individually infected erythrocytes. This individual parasite was expanded asexually, leading to a homogeneous clone representative of the original infecting parasite. Whole genome sequencing with Illumina HiSeq using 100 bp paired-end reads was performed when sufficient DNA was available. This cloning by limiting dilution was performed to generate multiple individual clones, hence all the resulting clonal populations belong to the same generation. A clone tree was finally obtained by repeating this process on one or multiple clones of each generation. Fig 2 shows the clone trees for the Dd2 strain obtained by Claessens et al. and Hamilton et al. The clone trees of the six strains of *P. falciparum* contained a total of 284 populations. More details of this method are available in Claessens et al. (2014). One advantage of this approach is that it reduces the impact of selection, thus approximating the molecular mutation rate (Barrick and Lenski, 2013).

The fastq paired-end reads for the 284 clonal populations are accessible on the ENA server (accession numbers were provided in Supplementary Table S2 from Hamilton et al., (2016)). To analyse patterns of *de novo* mutations, I used the data from the six distinct (Table 1) mutation accumulation lines reported by Claessens et al. (2014).



Geographical origin	Strain name	No. of Clonal populations
Africa	3D7	33
Indochina	W2	19
Indochina	Dd2	56
Houndaras	HB3	83
Cambodia	KH-01	60
Cambodia	KH-02	27
	Total	278

Table 1. The six strains of P. falciparum, their respective geographical origins and number of clonal populations generated for each strain. Claessens et al. (2016) generated data for 284 clonal populations, however 5 strains of 3D7 (3D7_1d, 3D7_2a, 3D7_2b, 3D7_2c and 3D7_2d) had a low coverage (only 70% of the genome was covered by more than 10 reads) and one strain of Dd2 (Dd2_(m)2a) produced an error during processing and hence these 6 samples were not analysed. It is to be noted that W2 is a clone derived from the Dd2 strain, hence these two genomes are virtually identical.

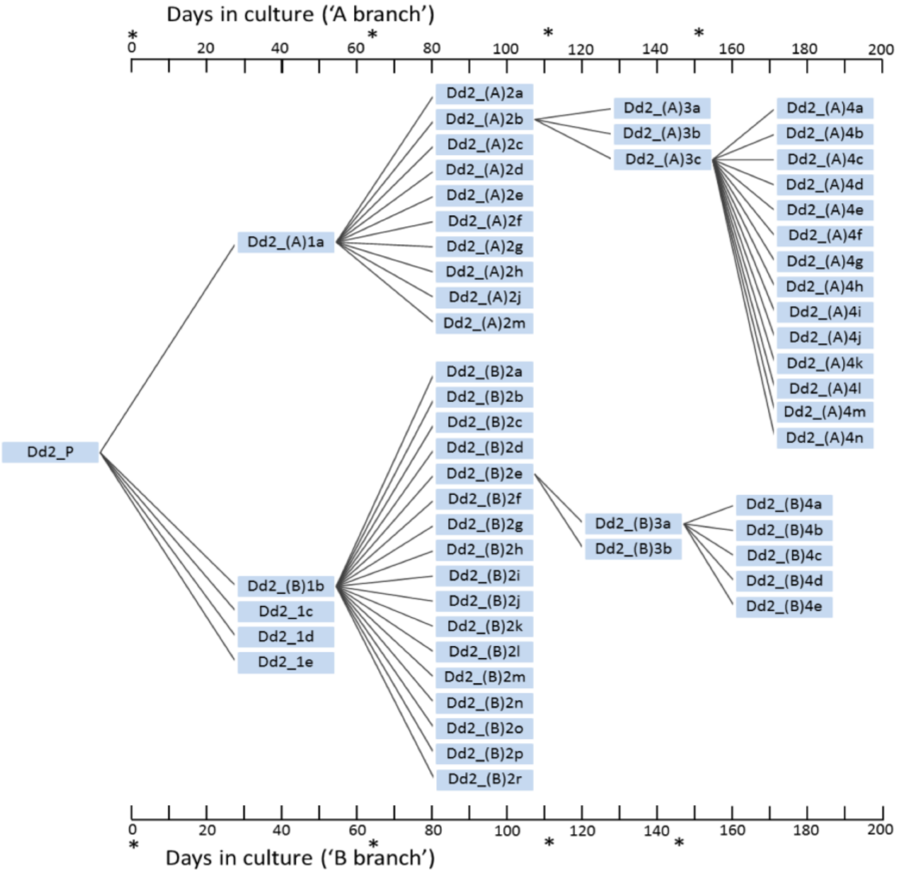


Fig 2. Generating the Dd2 clone tree: samples Dd2_(A)1a and Dd2_(B)1b were further clonally diluted for three generations forming two 'branches', referred to as the A and B branches. Each box indicates a whole-genome sequenced clone. The figures of the remaining 5 strains of *P. falciparum* are provided in the Appendix (Fig 1). Asterisks on the x-axes indicate when clonal dilutions were performed (Source Antoine Claessens).

Generating a suitable pipeline for micro-indel discovery

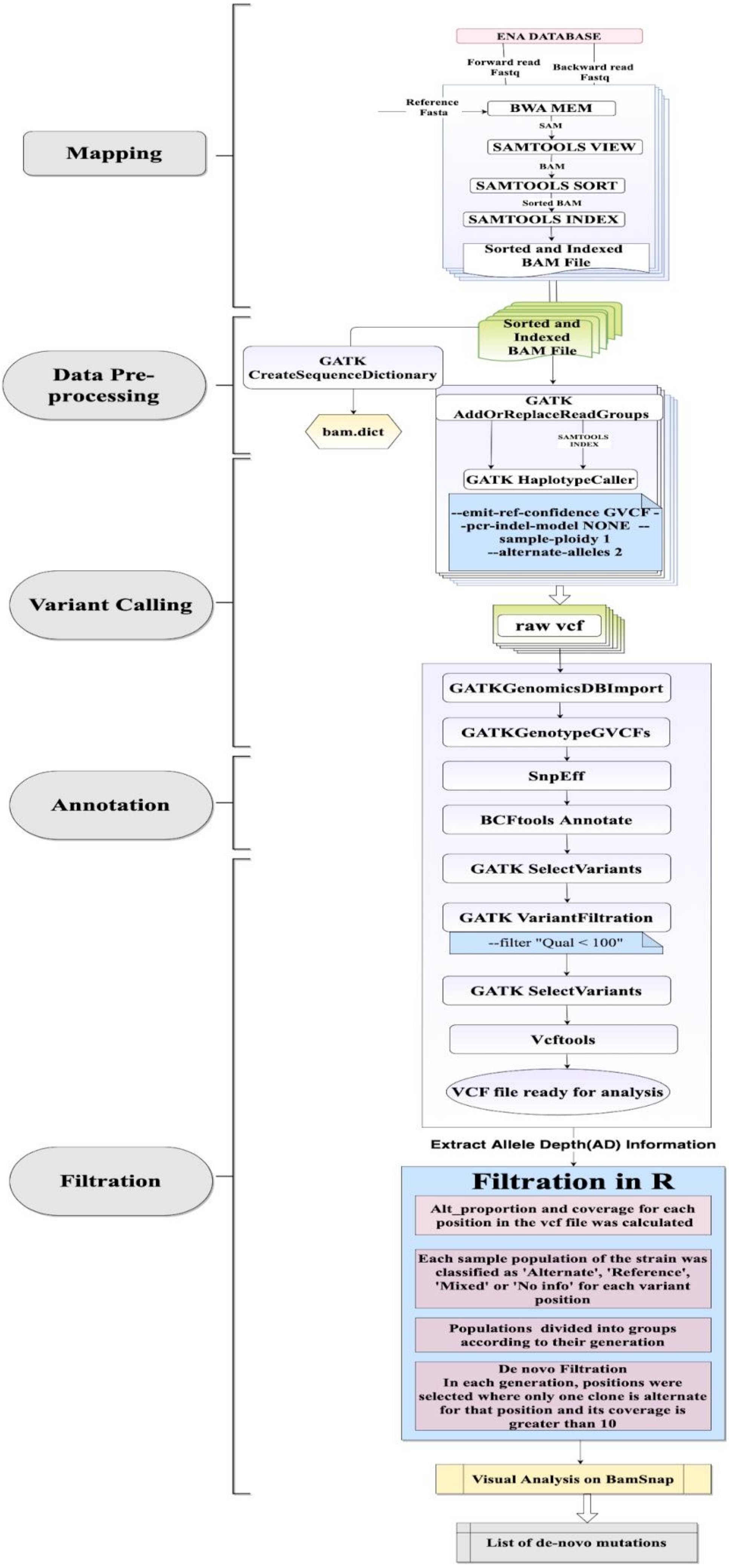

Mapping
Data Pre-processing
Variant Calling
Annotation
Filtration
ENA DATABASE
Forward read Fastq
Backward read Fastq
Reference Fasta
BWA MEM
SAM
SAMTOOLS VIEW
BAM
SAMTOOLS SORT
Sorted BAM
SAMTOOLS INDEX
Sorted and Indexed BAM File
Sorted and Indexed BAM File
GATK CreateSequenceDictionary
bam.dict
GATK AddOrReplaceReadGroups
SAMTOOLS INDEX
GATK HaplotypeCaller
--emit-ref-confidence GVCF --pcr-indel-model NONE --sample-ploidy 1 --alternate-alleles 2
raw vcf
GATKGenomicsDBImport
GATKGenotypeGVCFs
SnpEff
BCFtools Annotate
GATK SelectVariants
GATK VariantFiltration
--filter "Qual < 100"
GATK SelectVariants
Vcftools
VCF file ready for analysis
Extract Allele Depth(AD) Information
Filtration in R
Alt_proportion and coverage for each position in the vcf file was calculated
Each sample population of the strain was classified as 'Alternate', 'Reference', 'Mixed' or 'No info' for each variant position
Populations divided into groups according to their generation
De novo Filtration
In each generation, positions were selected where only one clone is alternate for that position and its coverage is greater than 10
Visual Analysis on BamSnap
List of de-novo mutations

Fig 3. A schematic representation of the pipeline developed for Variant discovery in haploid *P. falciparum* .

- **Mapping the genomes to the reference**

The forward and backward fastq reads for each clonal population were mapped on to its respective *P. falciparum* reference genome using Burrows-Wheeler Aligner (BWA) software (version 0.7.17) (Li and Durbin, 2009). The resulting SAM file was converted to a BAM file and sorted using Samtools (version 1.1) (Li et al., 2009).

- **Data Preprocessing**

The resulting sorted BAM files were pre-processed using Genome Analysis Toolkit (GATK) software (version 4.1.1) (Van der Auwera et al., 2013) best practices. This step involves data cleanup operations to correct for technical biases and make the data suitable for analysis. I used tools listed below during this step.

1. GATK CreateSequenceDictionary

This tool was used to create a sequence dictionary file from a reference sequence provided in FASTA format, which is required by many processing and analysis tools further in the pipeline. The output file contains a header but no SAMRecords, and the header contains only sequence records (Van der Auwera et al., 2013).

2. GATK AddOrReplaceReadGroups

Many tools such as Picard (version 2.5.0) ("Picard Toolkit." 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute) or GATK require or assume the presence of at least one RG (Read Group) tag, to which each read can be assigned . This tool enables the user to assign all the reads in the input BAM to a single new read-group (Van der Auwera et al., 2013).

- **Variant Calling**

Variant micro-indels were called on the analysis ready BAMs to generate gvcf files containing information about all the variants. The tools listed below were used in this step.

1. GATK HaplotypeCaller

The HaplotypeCaller calls SNPs and micro-indels simultaneously via local de novo assembly of haplotypes in an active region. It means that if the program encounters a region showing signs of variation, it discards the existing mapping information and completely reassembles the reads in that region. HaplotypeCaller runs per-sample to generate an intermediate file called a GVCF for scalable variant calling in DNA sequence data (Van der Auwera et al., 2013).

2. GATK GenomicsDBImport

This step consists in consolidating the contents of GVCF files across multiple samples in order to improve scalability and speed for the next step, joint genotyping. This step produces a datastore instead of a GVCF file.(Van der Auwera et al., 2013). At this step all the variants called by the HaplotypeCaller in all the clonal populations of a clone tree were consolidated together.

3. GATK GenotypeGVCFs

At this step, we use the database generated in the previous step and pass them all together to the joint genotyping tool, GenotypeGVCFs which produces a set of joint-called SNP and micro-indel calls ready for filtering. This cohort-wide analysis enables sensitive detection of variants even at difficult sites (Van der Auwera et al., 2013). The --max-alternate-allele 2 parameter was used at this step to only select for biallelic variants.

- **Annotation**

The VCF file was annotated using the SnpEff (version 4.3) (Cingolani et al., 2012) and Bcftools (version 1.10.2) (Heng Li, 2011). The SnpEff databases *Plasmodium falciparum_3d7*, *Plasmodium falciparum_dd2, Plasmodium falciparum_hb3* were used for annotation. The SnpEff databases for KH-01 and KH-02 strains were not available. A new SnpEff database can be created from a general feature format (GFF) file. However, due to time constraints, I could not create the new SnpEff databases for the KH-01 and KH-02 strains. The genomic regions were annotated using the tab delimited text file defining RegionType from Miles et al. (2016) for the 3D7 strains.

- **Filtering**

The resulting VCF file was filtered using GATK SelectVariants to subset micro-indels and GATK VariantFiltration to filter the subset using annotations in order to remove false positives and generate micro-indels having a high quality score (QUAL >100).

- **Filtering for De-novo mutations**

Allele depth (AD) scores were extracted from the filtered vcf and were further analysed with R version 3.6.1. (R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL https://www.R-project.org/). For each loci in the filtered vcf, the number of alternate reads divided by the total number of reads, that I called *alt_proportion*, was calculated for each sample. If the *alt_proportion* was greater than 0.6, the sample was annotated as ‘alternate’. If the alt_proportion was less than 0.2, the sample was annotated as ‘reference’. If the alt_proportion was between 0.2 and 0.6, the sample was annotated as ‘mixed’. A de novo mutation was defined as an micro-indel found in a clone annotated as alternate but whose parental clone is annotated as reference. When the clone has been cloned out even further, we expect to find these de novo mutations in all subsequent subclones, e.g., a de novo mutation found in Dd2_(A)2b will not be detected in (A)1a but will be found in all (A)3 and (A)4 generations (Fig 2b). To filter out false positive hits, I selected the loci that had only one ‘alternate’ sample in its respective generation since a true micro-indel would likely not appear independently in multiple samples at the same time in a population. I also filtered out variants with less than 10 reads of coverage. The resulting list of variants were visualized through BamSnap (Version 0.2.17) (Kwon et al., 2021) and only the seemingly true de novo mutations were selected.

The complete analysis was performed on IRD itrop HPC (South Green Platform) at IRD Montpellier.

Calculation of the micro-indel mutation rate

After visualisation of putative variants on BamSnap and obtaining a list of seemingly true micro-indels for each generation of the clone trees, I calculated the micro-indel mutation rate in each clone tree, using the same methodology as described by Claessens et al. (2014). In this case, mutation rate is defined as the sum of total number of micro-indels across of the clones in the respective dilution generation divided by the sum of all the clones in the dilution generation per life cycle per nucleotide (equation 1).

$$\mu = \frac{\left\{ \Sigma i /_{\Sigma c} \right\}}{L * G} \qquad \text{(-equation 1)}$$

Where, μ is the mutation rate, Σi is the number of micro-indels across all the clones in that dilution generation, Σc is the sum of all the clones for that dilution generation, *L is* the total number of life cycles between the respective clonal 'generations' and *G* is the Genome size of *P. falciparum.* The data generated from the first generation of each clone tree could not be used for the calculation of the mutation rate because these in vitro strains of *P. falciparum* have been culturing in the lab for a long period of time and may contain mutations that have been accumulated over time and hence the determination of the de novo mutations would not have been accurate. The micro-indels/per erythrocytic life cycle (ELC) were weighted by the size of the clone tree in a 'generation' and then were added together to calculate the micro-indel mutation rate per ELC per nucleotide for all the clonal samples of a particular strain.

3. RESULTS

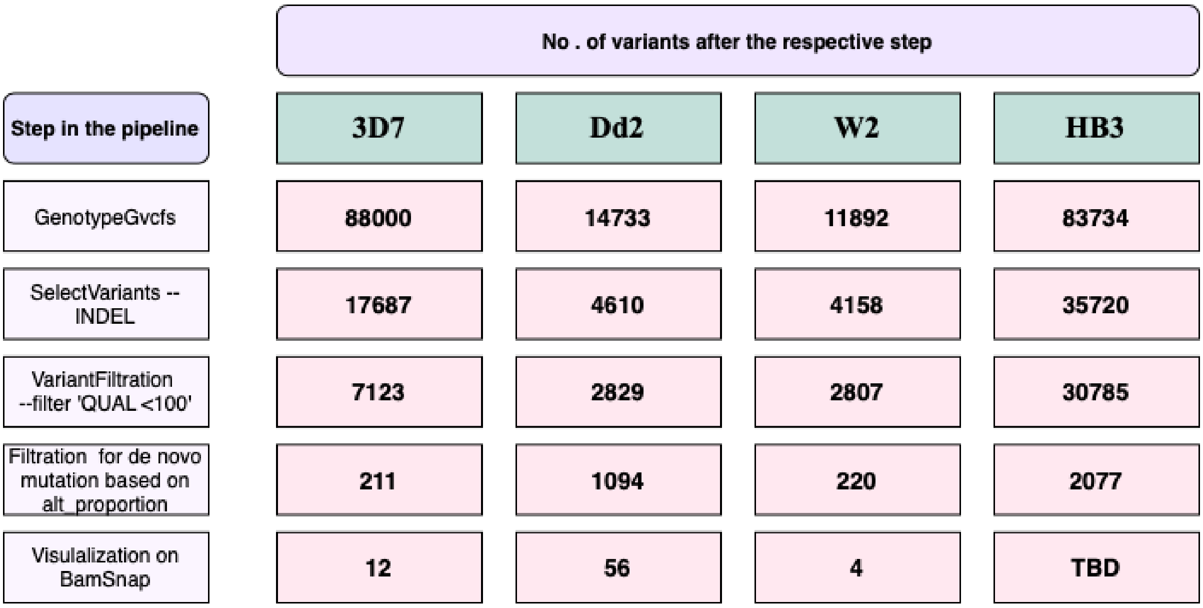
A pipeline suited for variant discovery in the haploid P. falciparum genome.

The fastq files from each sample of a clone tree were mapped against its respective reference genome. After preprocessing the data, variants were called using the GATK HaplotypeCaller with the sample ploidy =1. After the joint genotyping of all samples of a clone tree, micro-indels with a quality score greater than 100 were filtered and were further filtered in R on the basis of allele depth of each sample at a particular position to identify the de novo micro-indels. Prior to testing the GATK pipeline, I first tested out a SAMtools based pipeline for calling SNPs and micro-indels. It detected all the SNPs but only five percent of the true micro-indels that were reported in the 3D7 strain of *P. falciparum* by Hamilton et al (2016). I did not attempt to optimize that pipeline.

The pipeline (Fig 3), especially suited for the *P. falciparum* analysis was developed using multiple GATK tools, based on the MalariaGEN (Pearson et al., 2019) approach. Hamilton et al (2016) used tools such as 'RealignTargetCreator', 'Micro-indelRealigner' and 'UnifiedGenotyper,' with ploidy = 1, to call micro-indels in all realigned BAM. However, since then GATK has replaced the above mentioned tools with a more sophisticated tool known as the 'HaplotypeCaller' for variant calling.

UnifiedGenotyper was a position based variant caller that called SNPs and micro-indels on a per-locus basis whereas HaplotypeCaller discards the alignment information around a position where it suspects a variant call variants via local reassembly of the reads in the region that has evidence for the presence of variation. This has a high impact on calling micro-indels in highly repetitive regions. UnifiedGenotyper is not recommended anymore. Hamilton et al (2016) used CombineGVCFs for combining the variants from all the sub clonal populations in each clone tree. CombineGVCFs is quite inefficient and typically requires a lot of memory (GATK- How to consolidate GVCFs 2021). Hence, I use the more efficient GenomicsDBImport for combining GVCFs.

For the 3D7 clone tree, after filtering the variants in R, 211 putative micro-indels were found that were further visually inspected using BamSnap to identify the true *de novo* micro micro-indels. 12 true *de novo* micro micro-indels were identified in the second and 3 'generations' of the 3D7 clone tree. The number of variants selected after the application of each filter can be seen in Fig 4. The false positives were obvious sequencing/mapping errors, most of which were located in long homopolymer repeats that are typical of this AT-rich genome. An even greater number of false positives were identified for the Dd2 and W2 genomes. This might be explained by the fact that the assembly of the Dd2 reference genome is of lower quality than 3D7's (Otto 2018). . Many false positives were an actual alternation from the reference genome but were not *de novo* in nature, as shown in Fig 5 and 6.



	No . of variants after the respective step			
Step in the pipeline	3D7	Dd2	W2	HB3
GenotypeGvcfs	88000	14733	11892	83734
SelectVariants --INDEL	17687	4610	4158	35720
VariantFiltration --filter 'QUAL<100'	7123	2829	2807	30785
Filtration for de novo mutation based on alt_proportion	211	1094	220	2077
Visulalization on BamSnap	12	56	4	TBD

Fig 4. The number of micro-indel variants at the relevant steps in the pipeline identified in the four strains. The number of variants after the VariantFiltration step exclude the number of variants that were found in the first generation of each respective clone tree.

Compared to the Hamilton dataset, the current pipeline was able to detect 11 out of 11(Hamilton et al. (2016) reported SNPS and 98 out of 142 micro-indels (70 %) in the core genome (excluding the micro-indels found in sample 3D7_1b and 3D7_(o)2g that were not mapped correctly and resulted in an empty BAM file). Hamilton et al reported a total of 16 SNPS and 164 micro-indels in 3D7, out of which 11 and 150 were found in the core genome respectively. I used annotations from Miles et al. (2016) to annotate the type of region where a variant was found and only selected for variants that were present inside the core genome. A similar conclusion to detect the efficiency of the pipeline was not possible for the other strains because Hamilton et al. (2016) mapped the non- 3D7 strains to the 3D7 reference genome before calling the variants, whereas I used the newly assembled PacBio reference genomes of the non- 3D7 strains.

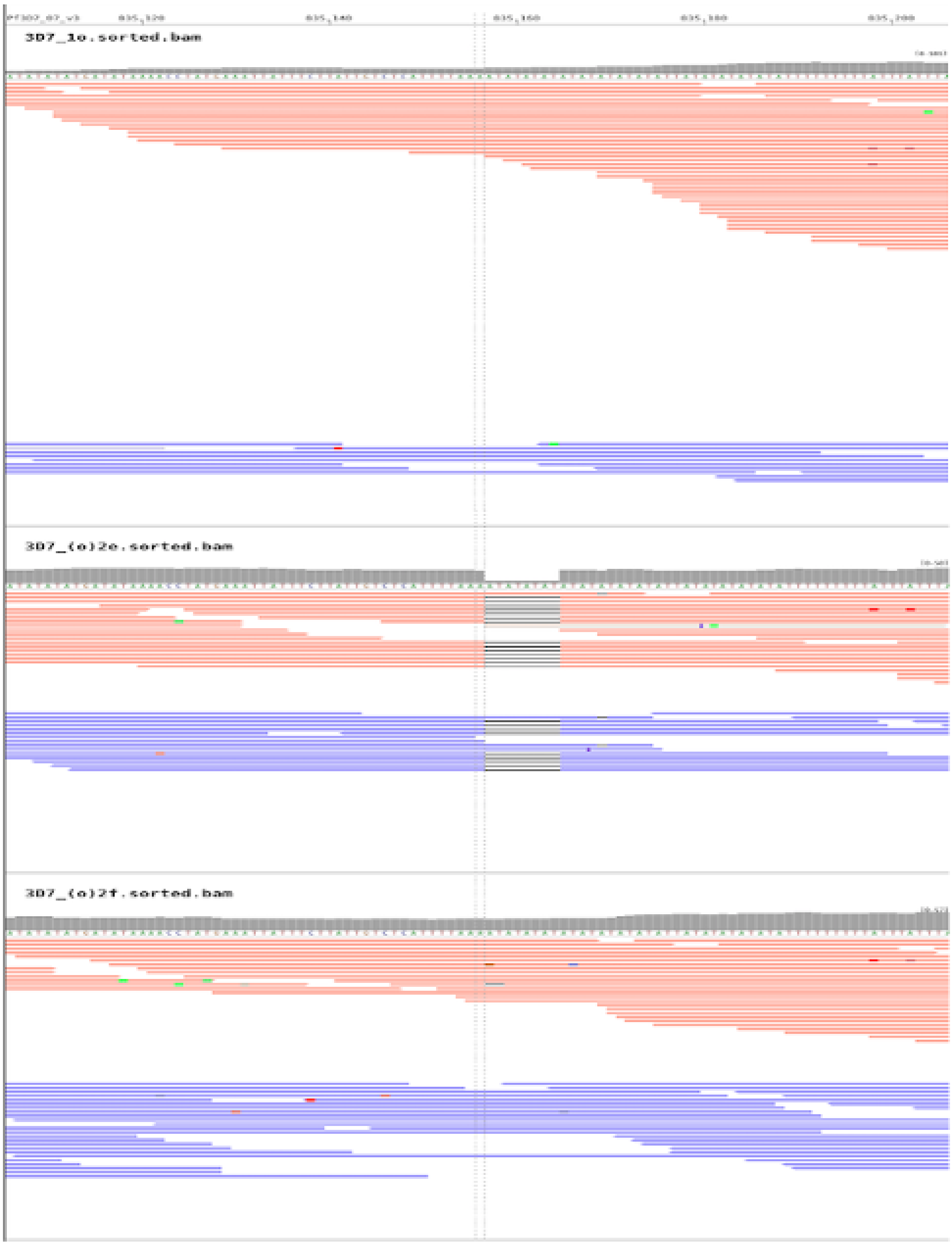
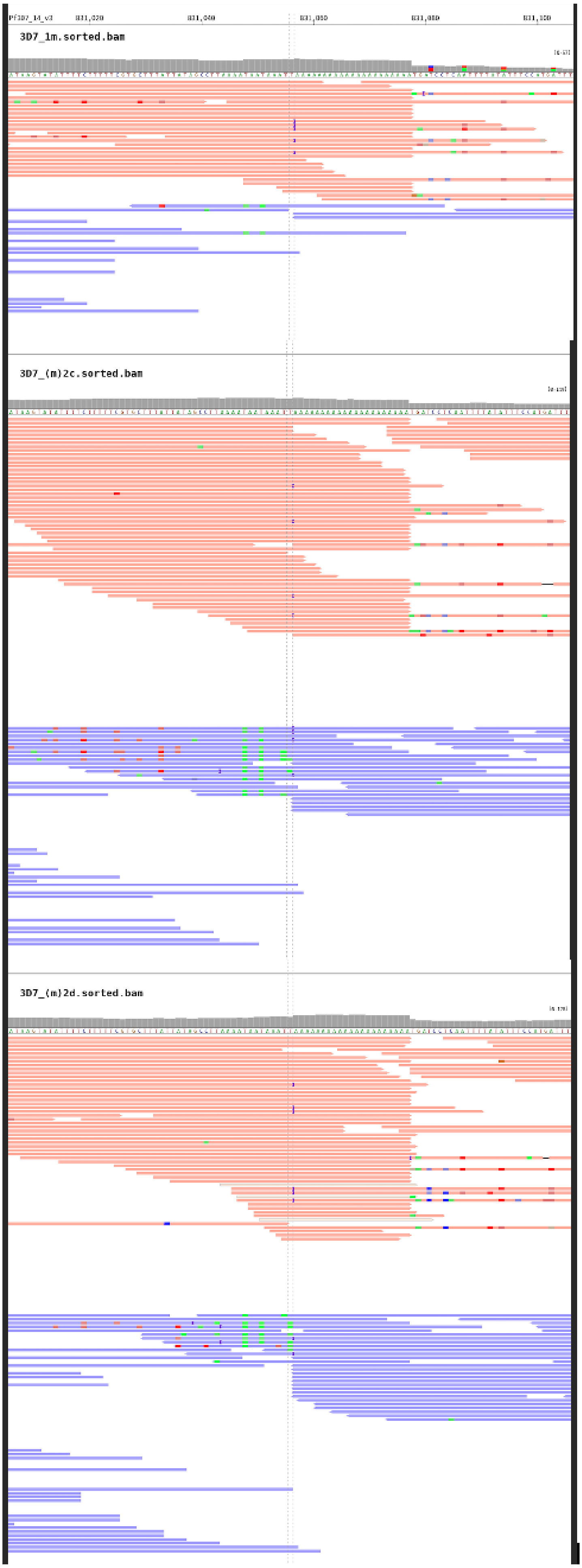


Fig 5. A png generated from BamSnap that shows deletion of AATATATAT sequence that was replaced by A in 3D7(o)2e clone at the 835156 position of the Pf3D7_07_v3 chromosome. This is an example of *de novo* mutation because no such mutation was found in the parental clone of 3D7_(o)2e, i.e., 3D7_1o, nor was such a mutation found in other clonal populations of the same generation as 3D7_(o)2f. The red and blue colour represent the forward and the backward reads respectively.

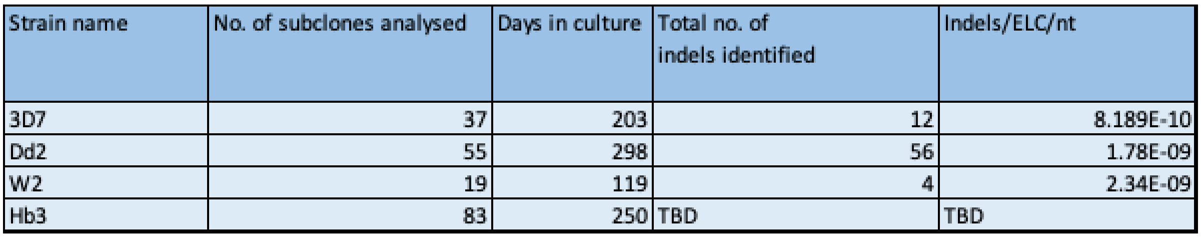


Pf3D7_14_v3
831,020
831,040
831,060
831,080
831,100
3D7_1m.sorted.bam
3D7_(m)2c.sorted.bam
3D7_(m)2d.sorted.bam

Fig 6. A png generated from BamSnap that shows addition of A nucleotide at 831056 position of the Pf3D7_14_v3 chromosome. This nucleotide was absent in the reference genome. This is not a *de novo* mutation as it is present in 3D7_1m and all its progeny. The insertion in two of the progenies 3D7_(m)2c clone and 3D7(m)2d is visible in the picture above.

Micro-indel Mutation rates in four strains of *P. falciparum*

The micro-indel mutation rate was calculated according to the formula given in equation 1.



Strain name	No. of subclones analysed	Days in culture	Total no. of indels identified	Indels/ELC/nt
3D7	37	203	12	8.189E-10
Dd2	55	298	56	1.78E-09
W2	19	119	4	2.34E-09
Hb3	83	250	TBD	TBD

Table 2. The micro-indel mutation rate in four *P. falciparum* strains .The micro-indels/per erythrocytic life cycle (ELC) were weighted by the size of the clone tree in a ‘generation’ and then were added together to calculate the micro-indel mutation rate per ELC per nucleotide for all the clonal samples of a particular strain.

4. DISCUSSION

- ## Variant Discovery pipeline

Recent advancements in Next-Generation Sequencing techniques have allowed a high-throughput detection of a vast number of variations in a fairly cost-efficient manner. However, there still are inconsistencies and debates about how this ‘big data’ should be processed and analysed. To accurately extract biologically relevant information from genomics data, choosing appropriate tools, knowing how to best utilize them and interpreting the results correctly is crucial. Almost all publicly available algorithms and tools focus on a single aspect of the entire process and do not provide a workflow that can aid the researcher from start to finish. GATK addressed the issue to a certain extent and provided ‘Best Practices’. GATK ‘Best Practices’ tried to provide step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing data. However, it is important to emphasize that even though ‘GATK Best Practices’ provide guidance regarding experimental design, quality control and pipeline implementation options for variant discovery and analysis, the pipeline heavily depends on many factors such as sequencing technology and the hardware infrastructure that are at our disposal, so I had to create a pipeline tailored specifically for my analysis.

Creating a pipeline tailored specifically for one’s research becomes all the more challenging when working with haploid organisms like *P. falciparum* because most of the tools used for variant discovery assume that a sample is diploid by default. Tools are typically not optimised for non-diploid organisms. For example, several variant annotations (Inbreeding Coefficient, StrandOddRatios etc.) are not appropriate for use with non-diploid cases. The lack of these annotations makes it difficult to filter out the false positives without visual inspection. Hence, my first challenge was creating a working pipeline for variant discovery and analysis.

I generated the pipeline to detect variants in the *P. falciparum* genome, for any strain with an available reference genome. The list of true variants from Hamilton et al. (2016) was used to check the efficiency of the pipeline for the 3D7 strain. My pipeline detects seventy percent of the variants detected by Hamilton et al (2016). The discrepancy in the number of true hits detected could be due to the following reasons: 1. Hamilton et al (2016) used a different version of GATK (version 3.3.0) that was running the 'UnifiedGenotyper', a completely different algorithm from the current 'Haplotype Caller'. 2. I discarded the telomeric and sub telomeric region during my analysis for 3D7 and only selected for the core region of the genome. Considering the core genome, my pipeline has an efficiency of 70% when compared to the Hamilton et al (2016). 3. Importantly, the Hamilton dataset was generated by feeding 'known sites' to the GATK algorithm. The list of known sites was generated with experimental *P. falciparum* genetic crosses, using Mendelian error rates as an indicator of genotypic accuracy by Miles et al (2016). This dataset, generated with the 3D7 reference genome, acts as training set for the GATK algorithms like BaseRecalibrator and VariantRecalibrator to improve the efficiency of variant detection in BAM files mapped to the 3D7 reference genome only, hence they could not be used for our non-3D7 strains analysis. Finally, resources like the Region type annotation to annotate the core region of the genome are present specifically for the 3D7 genome, making it harder to analyse other strains. Due to time constraints, I was unable to build the SnpEff databases for the KH-01 and KH-02 strains and hence could not perform the analysis for these two strains.

- **micro-indels and the AT rich genome of *P. falciparum***

P. falciparum has a unique genome composition with an exceptionally high AT content compared to other Plasmodium species and eukaryotes in general. The *P. falciparum* genome is nearly 70% AT rich in coding regions and ~90% AT rich in non-coding regions. This extremely high bias leads to a high abundance of Low Complexity Regions, which in turn render the sequencing, mapping and variant calling in that organism all the more complicated. Previous analysis of the clone tree samples revealed a significant excess of G:C to A:T transitions compared to other types of nucleotide substitution, which would naturally cause AT content to equilibrate close to the level seen across the *P. falciparum* reference genome (80.6% AT). Such AT-rich repetitive sequences can then cause DNA polymerase slippages and unequal crossing over events, as tandem repeats are prone to slipped- strand mispairing during DNA replication (Li et al., 2002 and Lovett 2004) and are associated with micro-indel mutations (Montgomery, 2013). I calculated the micro-indel mutation rate in 3 strains of the *P. falciparum* species which were found to be 8.19E-10, 1.782E-09, 2.34E-09 per erythrocytic life cycle per nucleotide for 3D7, Dd2 and W2 respectively. This equals to a ratio of 0.25, 0.23 and 0.5 SNP per micro-indels for 3D7 , Dd2 and W2 respectively. The high rate of micro-indel mutation in *P. falciparum* contrasts to other species, for example in *E. coli* nucleotide substitutions are 10x *more* frequent than micro-indels (Lee et al., 2012) .

- **Fitness cost of Drug resistance, mutation rate and Multiplicity of Infection (MOI)**

Each new mutation in an individual can increase its fitness, decrease it, or has no effect on it. The mutation rate can itself evolve, because it is subject to genetic change in the genome that encodes DNA replication and repair systems (Kunkel and Bebenek 2000). Many studies have documented the evolution of increased mutation rates (Mao et al.,1997 ;Sniegowski et al., 1997; Giraud et al., 2001; Notley-McRobb et al., 2002;Pal et al., 2007; Swings et al., 2017), which can evolve in certain conditions. For example, after a recent environmental change that creates opportunities for novel adaptations and new beneficial mutations (Desai et al., 2011; Sniegowski et al., 2000), a cell with a mutator allele is more likely to produce large-effect beneficial mutations than a cell with a wild-type mutation rate. Thanks to their improved fitness, cell lineages with newly acquired beneficial alleles (and their linked mutator alleles) can increase in frequency in the population. Thus, hypermutation can readily evolve when mutator alleles hitchhike to fixation with beneficial mutations (Chao and cox, 1983; Gentille et al, 2011; Giraud et al., 2011; Cox and Gibson 1974). However, most new mutations are thought to be effectively neutral or deleterious, and only a small fraction are beneficial in a given environment (Eyre-Walker and Keightley 2007). Following a similar principle, the *P. falciparum* strains facing a high drug pressure in South East Asia should have more opportunities to give rise to beneficial mutations that may aid in promoting drug resistance and a greater antigenic variation. However, some drug resistance mutations in the genome may be associated with fitness costs to the parasite. For example, using *in vitro* growth competition assays, the chloroquine-resistant 7G8 strain of *P. falciparum* was outcompeted by the chloroquine-sensitive D10 strain (Hayward et al., 2005). Studies utilizing cultured malarial parasites, animal models and samples collected from infected individuals have generally shown that resistance-mediating polymorphisms lead to malarial parasites that are out competed by wild type both in-vitro and in-vivo alike and wild type ends up replacing the resistant type when drug pressure diminishes. For example, in Malawi in the early 1990's, chloroquine treatment was abolished due to widespread resistance among the *P. falciparum* population, and sulphadoxine-pyrimethamine was used instead. By 2005, the estimated chloroquine efficacy had returned to 99% (Laufer et al, 2006), demonstrating that a lack of drug pressure, associated with high recombination rates, led the wild-type parasites to completely outcompete the chloroquine-resistant ones.

Although most micro-indels occur in intergenic regions and are presumably neutral, about 30% of the ~5400 *P. falciparum* genes contain highly repeated sequences (mainly 'AAT' codons coding for Asparagine) that are prone to such micro-indels (Murlidharan and Goldberg, 2013). Should such mutation lead to a frameshift (if the insertion or deletion is not a multiple of 3) it would likely disrupt the open reading frame and prevent translation of the protein, leading to dramatic consequences to the cell. Furthermore, it was recently shown that, under drug pressure, some *P. falciparum* parasites develop resistance by increasing the copy number of specific genes. These copy number variants are always flanked by AT tracks and thus key to the development of resistance (Guler et al., 2013). Thus , as postulated elsewhere, Rathod et al. (1997), Trotta et al. (2004) and Castellini et al. (2011), it is plausible that the mutation rate is linked to how fast drug resistance is acquired. In our very limited dataset, the 3D7 strain, originally isolated in Africa, shows a ~2 fold reduced micro-indel mutation rate compared to Dd2 and W2 strains that have their origins in Indochina, a region where, historically, antimalarial drug resistance has first been detected. More strains are of course needed before we could potentially identify a trend linked with the continent of origin.

- **Other hypotheses as to why drug resistance is first detected in South East Asia**

Malaria and the fight against it are distinctly different in various parts of the world. In Asia, most P. falciparum infections get medically treated, resulting in severe pressure for the parasite to develop resistance. On the other hand, in most African countries, the vast majority of infections are not treated, due increased host immunity or poverty (WHO report). Furthermore, in Africa, the majority of infections have multiple distinct strains of *P. falciparum* circulating in the blood of the infected individual. However, in Asia the infections are usually from a single strain. (Singh and Sharma 2016).Lower mixed strain infections in South East Asia may allow even less-fit parasites to be transmitted to the next host due to reduced level of intra-host competition. In contrast, the higher mixed strain infection rate in Africa may drive more intense intra-host competition, and may therefore reduce the probability of transmission of less-fit parasites (Singh and Sharma, 2016). Thus, fitness-reducing mutations including drug-resistance mutations might have a higher chance of spreading in SEA compared to Africa in patients not taking drugs. Since Africa has a higher rate of asymptomatic infections as well as untreated patients, this would also result in higher competition between drug resistant and drug sensitive clones in the absence of drugs, further decreasing the spread of drug resistance mutations with a fitness cost.

Mixed strain infection by *P. falciparum* has recently been demonstrated to lead to within-host competition in patients (Bushman et al., 2016), the possible mechanisms of which might include strain-transcending immunity, resource competition (e.g., RBCs) or direct interference between strains (Bruce et al, 2002; Matcalf et al., 2011; Raberg et al., 2006). Within-host competition would also lead to higher rate of recombination between gametes of different genotypes and efficient removal of deleterious mutations in Africa.

More in-depth analysis is required to answer questions like if these micro-indel mutations are abundant in coding regions (Hamilton et al., 2016), if these micro-indels occur in multiples of 3 and cause frameshift mutations leading to truncated proteins, and test if there are specific hotspots of micro-indel mutations in the *P. falciparum* genome.

One of the main challenges facing malaria elimination, is the incredible capacity of the malaria parasite to adapt to a changing environment. Uncovering the mutations that lead the parasite to adapt to the changes in its environment would definitely bring us a step closer to eliminating the disease that still kills nearly half a million people each year.

ACKNOWLEDGEMENT

I would like to thank Antoine Claessens, my research supervisors and Marc-Antoine Guery, for their guidance, encouragement and useful critiques of this research work. I would also like to thank Yannis Michalakis and Sandrine Maurice and Renaud Vitalis, for their advice and assistance in keeping my progress on schedule, for reviewing the intermediate report and providing valuable feedback.

REFERENCE

- **Barrick J.E.,** Lenski R.E. (2013) Genome dynamics during experimental evolution. Nat. Rev. Genet. 14:827–839.
- **Bronzwaer S.**, Cars O., Buchholz U., Molstad S., Goettsch W., Veldhuijzen I. K. et al. (2002). A European Study on the Relationship between Antimicrobial Use and Antimicrobial Resistance. Emerging Infectious Diseases. 8(3):278–82.
- **Bruce, M.C.**, Day, K.P.(2002). Cross-species regulation of malaria parasitaemia in the human host. *Curr Opin Microbiol.* 5(4):431–437. 10.1016/S1369-5274(02)00348-X
- **Bushman, M.,** Morton, L., Duah, N., et al. (2016). Within-host competition and drug resistance in the human malaria parasite *Plasmodium falciparum*. *Proc Biol Sci.* 283(1826):20153038. 10.1098/rspb.2015.3038
- **Castellini, M.A.,** Buguliskis, J.S., Casta, L.J., Butz, C.E., Clark, A.B., Kunkel, T.A and Taraschi,T.F. (2011) Malaria drug resistance is associated with defective DNA mismatch repair. *Mol. Biochem. Parasitol.* **177**: 143–147.
- **Cingolani, P.,** Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, Austin.6(2):80-92. PMID: 22728672
- **Claessens, A.,** Hamilton, W.L., Kekre, M., Otto, T.D., Faizullabhoy, A., Rayner, J.C., et al. (2014) Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. PLoS Genet 10(12): e1004812. https://doi.org/10.1371/journal.pgen.1004812
- **Collins, F.S.,** Brooks, L.D. and Chakravarti, A. (1998).A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. Genome Res. 8: 1229-1231. doi 10.1101/gr.8.12.1229.
- **Chao, L.** and Cox, E.C. (1983). Competition between high and low mutating strains of Escherichia coli. Evolution. 37: 125–134. pmid:28568016
- **Cox, E.C.** and Gibson, T.C. (1974) Selection for high mutation rates in chemostats. Genetics. 77: 169–184. pmid:4603159
- **Desai, M.M.** and Fisher, D.S. (2011) The balance between mutators and nonmutators in asexual populations. Genetics.188: 997–1014. pmid:21652523
- **Dondorp, A.M.,** Nosten, F., Yi, P., Das, D., Phyo, A.P., Tarning, J., Lwin, K.M., Ariey, F., Hanpithakpong, W., Lee, S.J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S.S., Yeung, S., Singhasivanon, P., Day, N.P., Lindegardh, N., Socheat, D., White, N.J. (2009). Artemisinin resistance in Plasmodium falciparum malaria. N Engl J Med. 361:455–467.
- **Eyre-Walker, A.** and Keightley, P.D. (2007) The distribution of fitness effects of new mutations. Nat Rev Genet. 8: 610–618. pmid:17637733

- **Fairhurst, R.,** Nayyar, G., Breman, J., Hallett, R., Vennerstrom, J., Duong, S., Ringwald, P., Wellems, T., Plowe, C., Dondorp, A. (2012). Artemisinin-resistant malaria: research challenges, opportunities, and public health implications. The American Journal of Tropical Medicine and Hygiene. 87: 231–241.
- **Garcia L.S.** (2010) Malaria. Clin Lab Med. 1:93-129.
- **Giraud, A.,** Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., et al. (2001). Costs and benefits of high mutation rates: Adaptive evolution of bacteria in the mouse gut. Science. 291: 2606–2608. pmid:11283373
- **Giraud, A.**, Radman, M., Matic, I., Taddei, F. (2001). The rise and fall of mutator bacteria. Curr Opin Microbiol. 24: 582–585. pmid:11587936
- **Griffiths, A.J.F.**, Miller, J.H., Suzuki. D.T., et al. (2000) An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; Sources of variation. Available from : https://www.ncbi.nlm.nih.gov/books/NBK22012/
- **Guler, J. L.,** Freeman, D. L., Ahyong, V., Patrapuvich, R., White, J., Gujjar, R., Phillips, M. A., DeRisi, J., & Rathod, P. K. (2013). Asexual Populations of the Human Malaria Parasite, Plasmodium falciparum, Use a Two-Step Genomic Strategy to Acquire Accurate, Beneficial DNA Amplifications. *PLoS Pathogens*, *9*(5), e1003375. https://doi.org/10.1371/journal.ppat.100337
- **Hamilton, W.,** Claessens, A., Otto, T., Kekre, M., Fairhurst, R., et al. (2016). Extreme mutation bias and high AT content in Plasmodium falciparum. Nucleic Acids Research, Oxford University Press pp.gkw1259. doi:10.1093/nar/gkw1259⟩. ⟨hal-01989279⟩
- **Heng, Li.** (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993, https://doi.org/10.1093/bioinformatics/btr509.
- **Hoffman S.L.**, Rustama, D., Dimpudus, A.J., Punjabi, N.H., Campbell, J.R., Oetomo, H.S., Marwoto, H.A., Harun, S., Sukri, N., Heizmann, P. (1985). RII and RIII type resistance of Plasmodium falciparum to combination of mefloquine and sulfadoxine/pyrimethamine in Indonesia. Lancet 2(8463):1039-1040
- **Kunkel, T. A.,** & Bebenek, K. (2000). DNA Replication Fidelity. Annual Review of Biochemistry, 69(1): 497–529. doi:10.1146/annurev.biochem.69.1.497
- **Kwon, M.**, Lee, S., Berselli, M., Chu, C., & Park, P. J. (2021). BamSnap: A lightweight viewer for sequencing reads In BAM files. Bioinformatics. doi:10.1093/bioinformatics/btaa1101
- **Laufer, M. K.**, Thesing, P. C., Eddington, N. D., Masonga, R., Dzinjalamala, F. K., Takala, S. L., . . . Plowe, C. V. (2006). Return of Chloroquine Antimalarial efficacy in Malawi. *New England Journal of Medicine. 355*(19): 1959-1966. doi:10.1056/nejmoa062032
- **Lee, H.,** Popodi, E., Tang, H., and Foster, P. L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proceedings of the National Academy of Sciences. 109(41): E2774–E2783. https://doi.org/10.1073/pnas.1210309109
- **Li, H.,** Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25(16): 2078-9
- **Li, H.,** and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25:1754-1760. [PMID: 19451168]
- **Li,Y.-C.**, Korol,A.B., Fahima,T., Beiles,a and Nevo,E. (2002). Microsatellites: genomic distribution, putativa functions, and mutational mechanism: a review. *Mol. Ecol.* **11**:253–256
- **Lovett,S.T.** (2004) Encoded errors: Mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* 52:1243–1253.
- **Manske, M.,** Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O'Brien, J., Djimde, A., Doumbo, O., Zongo, I., Ouedraogo, J.-B., Michon, P., Mueller, I., Siba, P., Nzila, A., Borrmann, S., Kiara, S. M., Marsh, K., Jiang, H., … Kwiatkowski, D. P. (2012). Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. 487(7407):375–379. https://doi.org/10.1038/nature11174
- **Mao E.F.,** Lane, L., Lee, J., Miller, J.H.(1997). Proliferation of mutators in a cell population. J Bacteriol.179: 417–422. pmid:8990293
- **Montgomery,S.B.**, Goode,D.L., Kvikstad,E., Albers,C.a, Zhang,Z.D., Mu,X.J., Ananda,G., Howie,B., Karczewski,K.J., Smith,K.S. *et al.* (2013). The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.*, **23**:749–761.

- **Miles A.,** Iqbal Z., Vauterin P., Pearson R., Campino S., Theron M., Gould K., Mead D., Drury E., Brien J.O. et al. (2016). Micro-indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome Res. 26:1288–1299.
- **Muralidharan, V.,** & Goldberg, D. E. (2013). Asparagine Repeats in Plasmodium falciparum Proteins: Good for Nothing? *PLoS Pathogens*. *9*(8): e1003488. https://doi.org/10.1371/journal.ppat.1003488
- **Notley-McRobb, L**., Seeto, S., Ferenci, T. (2002). Enrichment and elimination of mutY mutators in Escherichia coli populations. Genetics. 162: 1055–1062. pmid:12454055
- **Notley-McRobb, L.,** King, T., Ferenci, T. (2002). rpoS mutations and loss of general stress resistance in Escherichia coli populations as a consequence of conflict between competing stress responses. J Bacteriol. 184: 806–811. pmid:11790751
- **Otto, T.D.,** Böhme, U., Sanders, M.J et al. (2018). Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres [version 1; peer review: 3 approved]. Wellcome Open Res, 3:52 (https://doi.org/10.12688/wellcomeopenres.14571.1)
- **Packard, R.M**. (2007). The making of a tropical disease: a short history of malaria. Baltimore: Johns Hopkins University Press, 2007.
- **Pearson, R. D.,** Amato, R., & Kwiatkowski, D. P. (2019). An open dataset of Plasmodium falciparum genome variation in 7,000 Worldwide samples. doi:10.1101/824730
- **Pal, C.,** Maciá, M.D., Oliver, A., Schachar, I., Buckling, A. (2007). Coevolution with viruses drives the evolution of bacterial mutation rates. Nature. 450: 1079–1081. pmid:18059461
- **Peters, W.** (1970) . Chemotherapy and Drug Resistance in Malaria, 1st ed . London: Academic Press.
- **Poplin, R.,** Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., . . . Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. doi:10.1101/201178
- **Pray, L.** (2008) DNA Replication and Causes of Mutation. Nature Education, 1(1):214.exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. EMBO J, 17:5418–26.
- **Raberg, L.,** de Roode, J.C., Bell, A.S., et al. (2006). The role of immune-mediated apparent competition in genetically diverse malaria infections. *Am Nat.* 168(1):41–53. 10.1086/505160
- **Rathod,P.K.,** McErlean,T. and Lee,P.C. (1997) Variations in frequencies of drug resistance in Plasmodium falciparum. *Proc. Natl. Acad. Sci. U.S.A.* **94**: 9389–9393.
- **Rowe, J.,** Claessens, A., Corrigan, R., & Arman, M. (2009). Adhesion of Plasmodium falciparum-infected erythrocytes to human cells: Molecular mechanisms and therapeutic implications. *Expert Reviews in Molecular Medicine, 11,* E16. doi:10.1017/S1462399409001082
- **Sehn, J. K.** (2015). Chapter 9 - Insertions and Deletions (Micro-indels). In Clinical Genomics (pp. 129-150). doi https://doi.org/10.1016/B978-0-12-404748-8.00009-5
- **Singh, G. P.,** & Sharma, A. (2016). South-East Asian strains of Plasmodium falciparum display higher ratio of Non-synonymous TO SYNONYMOUS polymorphisms compared to African strains. *F1000Research, 5*, 1964. doi:10.12688/f1000research.9372.2
- **Sniegowski P.D.,** Gerrish, P.J., Lenski, R.E. (1997) Evolution of high mutation rates in experimental populations of E. coli. Nature. 387: 703–705. pmid:9192894
- **Sniegowski, P.D.**, Gerrish, P.J., Johnson, T., Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. BioEssays. 22: 1057–1066. pmid:11084621
- **Snow, R.W.** (2015) Global malaria eradication and the importance of Plasmodium falciparum epidemiology in Africa. BMC Med., 13:23.
- **Swings, T.**, Van den Bergh, B., Wuyts, S., Oeyen, E., Voordeckers, K., Verstrepen, K.J., et al. (2017). Adaptive tuning of mutation rates allows fast response to lethal stress in Escherichia coli. eLife. 6: e22939. pmid:28460660
- **Takala-Harrison,** S., & Laufer, M. K. (2015). Antimalarial drug resistance in Africa: key lessons for the future. Annals of the New York Academy of Sciences. 1342(1): 62–67. https://doi.org/10.1111/nyas.12766
- **Trotta,R.F.**, Brown,M.L., Terrell,J.C. and Geyer,J.A. (2004). Defective DNA repair as a potential mechanism for the rapid development of drug resistance in Plasmodium falciparum. *Biochemistry*. **43**:4885–4891.
- **Ugur Sezerman, O.**, Ulgen, E., Seymen, N., & Melis Durasi, I. (2019). Bioinformatics Workflows for Genomic Variant Discovery, Interpretation and Prioritization. In Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations.

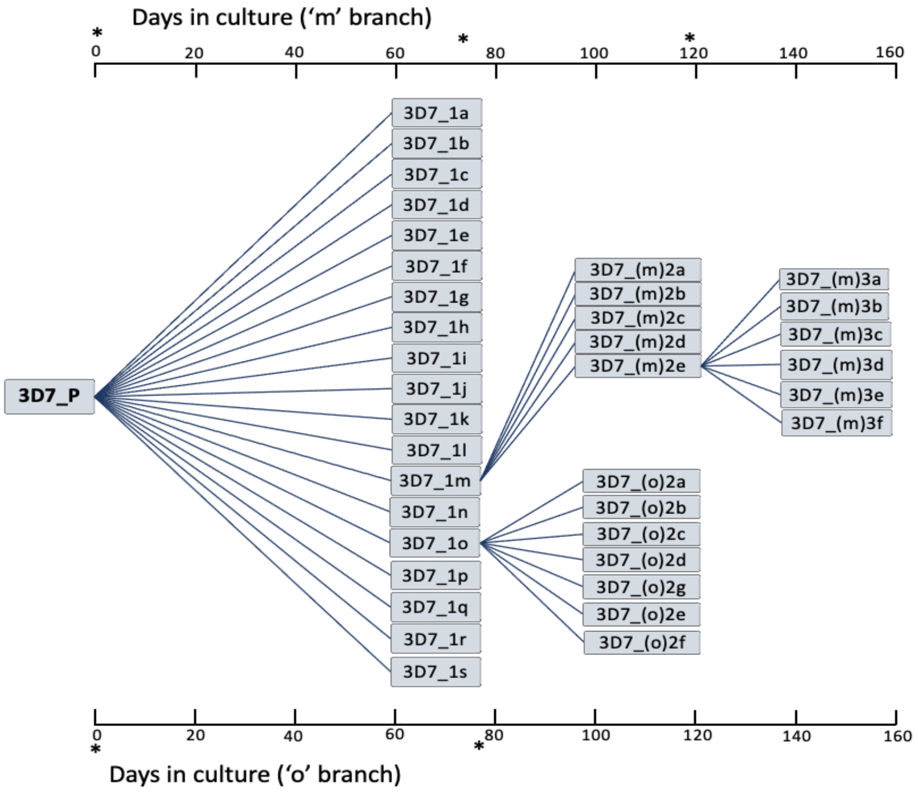
IntechOpen. https://doi.org/10.5772/intechopen.85524

- **Van der Auwera**, **G.A.,** Carneiro, M., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K., Altshuler, D., Gabriel, S., DePristo, M. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics, 43:11.10.1-11.10.33.
- **White N.J. (2004).** Antimalarial drug resistance. J Clin Invest.113:1084–1092.

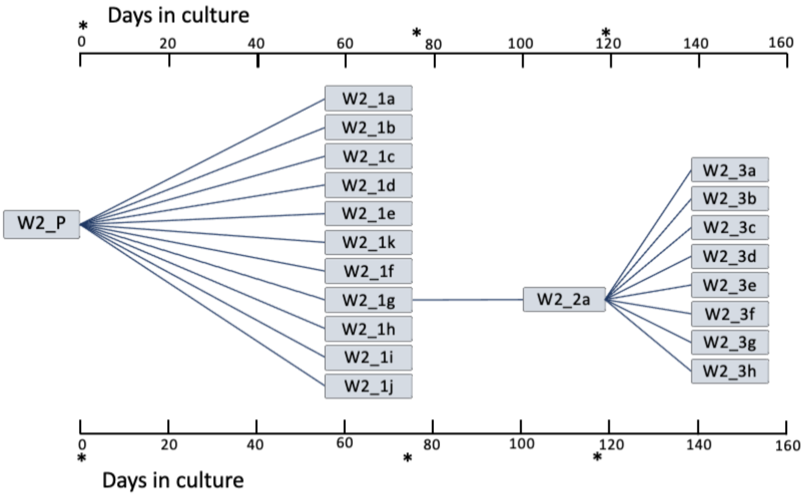
- **WHO, 2014.** Status Report on Artemisinin Resistance: September-2014. Geneva, World Health Organization

APPENDIX

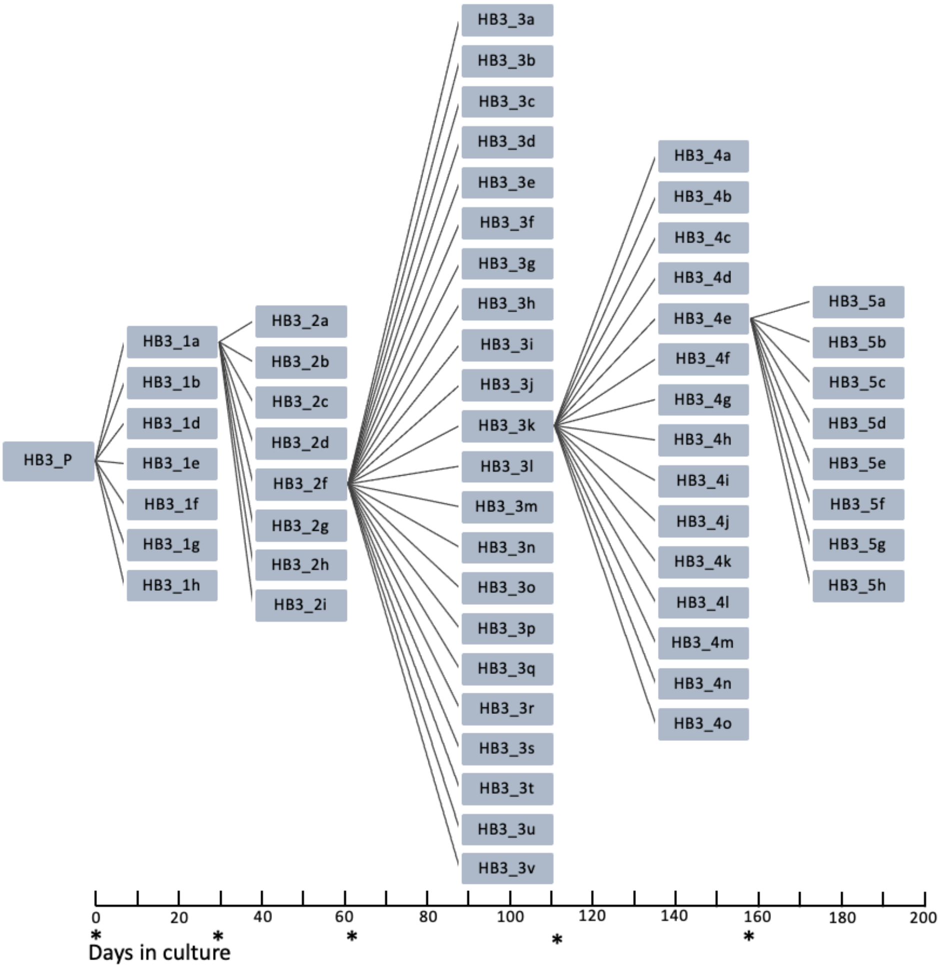
A. Clone tree for 3D7



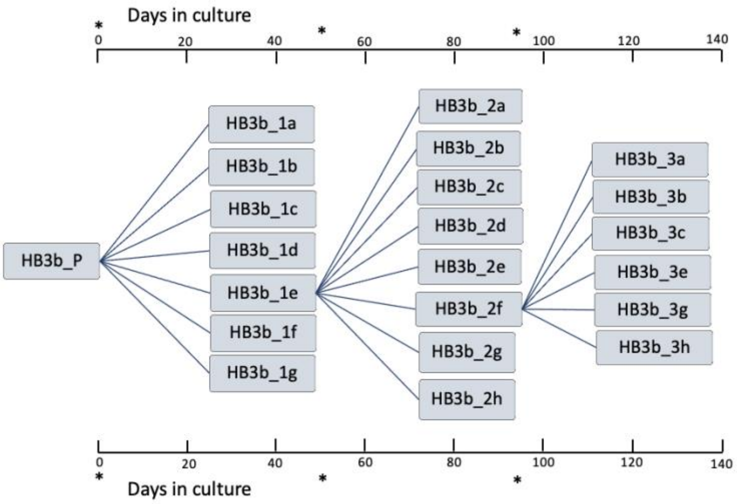
B. Clone tree for W2



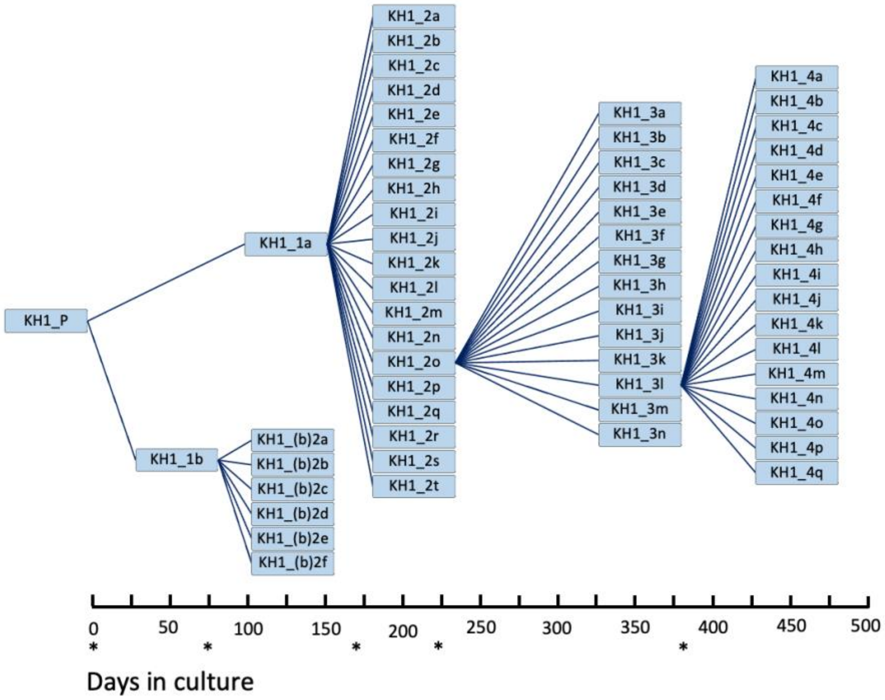
C. Clone tree for HB3



D. Clone tree for HB3b



E. Clone Tree for KH-01



F. Clone Tree for KH-02

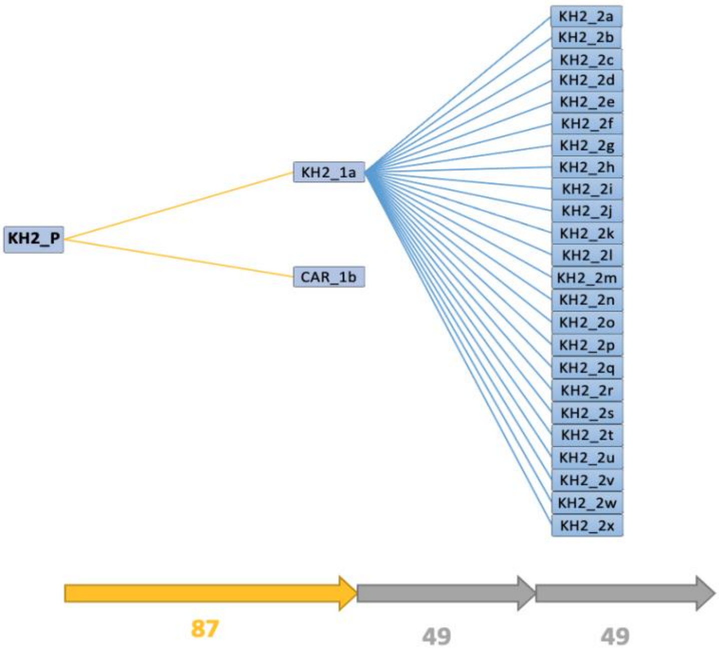
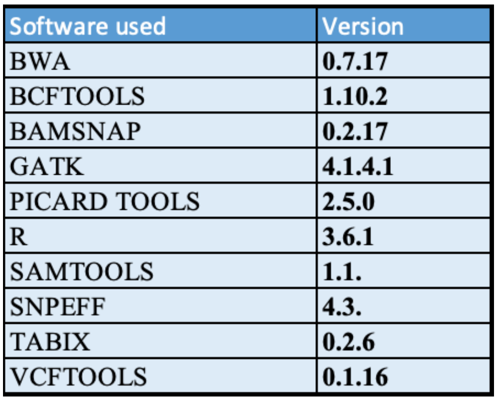


Fig1. Generating clone trees. (A) The 3D7 clone tree. Sample 3D7_1o was thawed and clonally diluted following whole genome sequence analysis. (B) The W2 clone tree.(C) and (D) show two HB3 clone trees started independently. Unlike all other clone trees, HB3b (C) was initiated from a subclone of HB3. (E) The KH-01 clone tree (F) The KH-02 Clone tree. Asterisks on the x-axes indicate when clonal dilutions were performed. (Source- Antoine Claessens)



Software used	Version
BWA	**0.7.17**
BCFTOOLS	**1.10.2**
BAMSNAP	**0.2.17**
GATK	**4.1.4.1**
PICARD TOOLS	**2.5.0**
R	**3.6.1**
SAMTOOLS	**1.1.**
SNPEFF	**4.3.**
TABIX	**0.2.6**
VCFTOOLS	**0.1.16**

Table 1. The version of the software used during the analysis.