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

R script documentation

Problem Description:
Identification of the de-novo mutations is crucial to estimate mutation rate in a species.After variant calling using the GenomeAnalysisToolkit (GATK), de-novo mutations need to be filtered out according the Allele Depth (AD) values that are generated for each population at each chromosomal position.

Solution:
The following script was written in R version 4.0.4. The script has two parts
Part 1: This part contains the functions.
Part 2: This part contains the template for calling the functions.

PART 1

1. **ad_table( ) :** This function produces a table containing the alt ratios for each population at every chromosomal position. Alt_ratio is defined as the ratio of alternate AD values to the total AD value of the population.
2. **coverage_table( ) :** This function produces a table containing the coverage for each population at every chromosomal position. Coverage is defined as the total of both the AD value for each population.
3. **combined_ad_and_coverage( ) :** This function combines the output of ad_table() and coverage_table() to output a single table.

For functions 1 to 3,

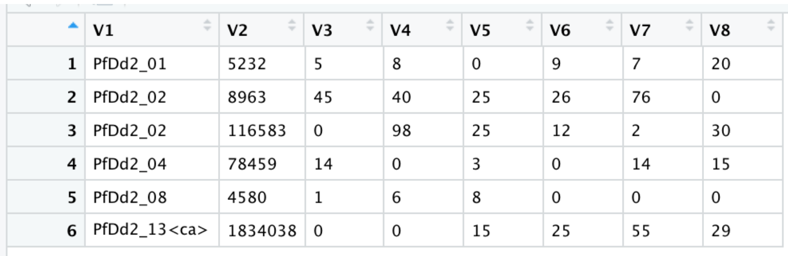
Required Inputs :

- *input_table.df* : A table containing the Allele Depth(AD) values of different populations along   with the chromosome name and position values obtained after the gatk processing pipeline.
- *start_sample_column*: The column from which the values of AD start i.e., excluding the chromosome name and position values.
- *header_table:* The table containing the header for the input_table.df

Output_produced:
A table with alt_ratio and coverage and for each population at each chromosomal population.

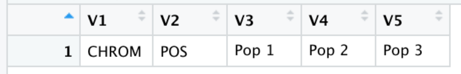
Sample Input: for combined_ad_coverage( )

- *input_table.df*



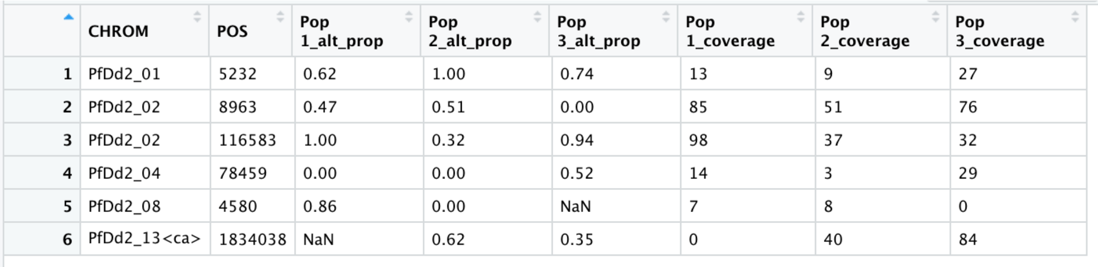
	V1	V2	V3	V4	V5	V6	V7	V8
1	PfDd2_01	5232	5	8	0	9	7	20
2	PfDd2_02	8963	45	40	25	26	76	0
3	PfDd2_02	116583	0	98	25	12	2	30
4	PfDd2_04	78459	14	0	3	0	14	15
5	PfDd2_08	4580	1	6	8	0	0	0
6	PfDd2_13<ca>	1834038	0	0	15	25	55	29

- *start_sample_column = 3*
- *header_table*



	V1	V2	V3	V4	V5
1	CHROM	POS	Pop 1	Pop 2	Pop 3

Sample Output: for combined_ad_coverage( )



	CHROM	POS	Pop 1_alt_prop	Pop 2_alt_prop	Pop 3_alt_prop	Pop 1_coverage	Pop 2_coverage	Pop 3_coverage
1	PfDd2_01	5232	0.62	1.00	0.74	13	9	27
2	PfDd2_02	8963	0.47	0.51	0.00	85	51	76
3	PfDd2_02	116583	1.00	0.32	0.94	98	37	32
4	PfDd2_04	78459	0.00	0.00	0.52	14	3	29
5	PfDd2_08	4580	0.86	0.00	NaN	7	8	0
6	PfDd2_13<ca>	1834038	NaN	0.62	0.35	0	40	84

4. **adding_info_to_tables( ) :** This function adds additional information to the table generated by combined_ad_and_coverage( ) to identify the de-novo mutations. *De-novo mutations are defined as the mutations that are present in only one population at a particular chromosomal locus.*

Required Inputs :

- *new_table.df :* The table generated in the previous function. While analysing the population generation wise we will need to modify the table obtained in the previous function to only contain the alt_ratio and coverage values for the generation being analysed.
- *alt_ratio_min :* The cutoff value of alt_ratio below which the population will be categorised as 'Reference' for that chromosomal population.
- *alt_ratio_max :* The cutoff value of alt_ratio above which the population will be categorised as 'Alternate' for that chromosomal population.
- *start_sample_column :* The column from which the values for each population starts.
- *total_no_of_samples :* The total number of samples in new_table.df.
- *coverage_limit :* The cutoff value of coverage. If a population has a coverage value less than the coverage_limit, it will not be considered for further analysis.

Output produced:

A table with additional information for each locus such as

- *Reference:* The number of populations that are categorised as containing the reference allele at that chromosomal population.
- *Alternate:* The number of populations that are categorised as containing the alternate allele at that chromosomal population.
- *Mixed:* The number of populations where the alt_ratio is between the alt_ratio_min and alt_ratio_max are classified as 'Mixed'.
- *No_info:* The number of populations about which we do not have relevant information.
- *Check_coverage:* If only one population at a locus is 'Alternate' and its coverage value is greater than the coverage_limit value, then this column contains the coverage value of the 'Alternate' population.
- *Sample_name:* The population name which de-novo* at that chromosomal population

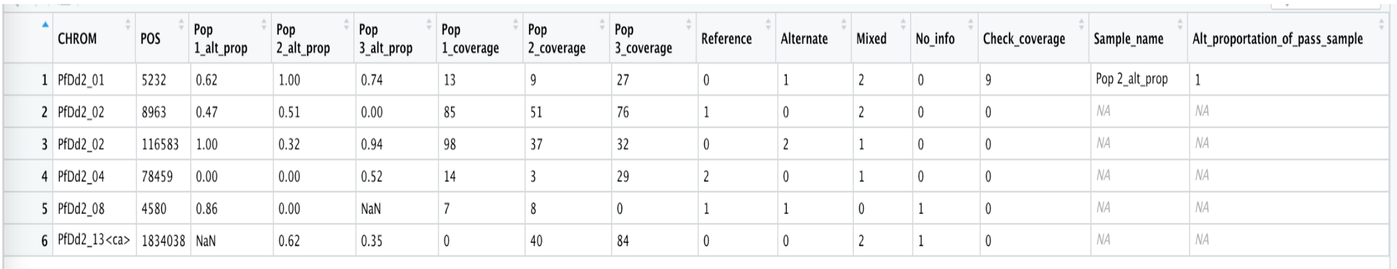
- *Alt_proportion_of_pass_sample:* Alt_proportion of the population that has a de-novo mutation at that locus.

**Note: If only one population at a locus is 'Alternate' and its coverage value is greater than the coverage_limit value, then that population is said to have a de-novo mutation at that chromosomal position.*

Sample Input:

- *new_table.df = sample output table from* combined_ad_coverage( ) function
- *alt_ratio_min = 0.2*
- *alt_ratio_max = 0.8*
- *start_sample_column = 3*
- *total_no_of_samples = 3*
- *coverage_limit = 8*

Sample Output:



	CHROM	POS	Pop 1_alt_prop	Pop 2_alt_prop	Pop 3_alt_prop	Pop 1_coverage	Pop 2_coverage	Pop 3_coverage	Reference	Alternate	Mixed	No_info	Check_coverage	Sample_name	Alt_proportation_of_pass_sample
1	PfDd2_01	5232	0.62	1.00	0.74	13	9	27	0	1	2	0	9	Pop 2_alt_prop	1
2	PfDd2_02	8963	0.47	0.51	0.00	85	51	76	1	0	2	0	0	NA	NA
3	PfDd2_02	116583	1.00	0.32	0.94	98	37	32	0	2	1	0	0	NA	NA
4	PfDd2_04	78459	0.00	0.00	0.52	14	3	29	2	0	1	0	0	NA	NA
5	PfDd2_08	4580	0.86	0.00	NaN	7	8	0	1	1	0	1	0	NA	NA
6	PfDd2_13<ca>	1834038	NaN	0.62	0.35	0	40	84	0	0	2	1	0	NA	NA

5. **ompg_filtering( ):** This function filters de-novo mutations from the table generated by adding_info_to_tables( ) function . This function filters rows where the value of 'Alternate' is 1 and the check_coverage column has a non zero value.

Required Inputs :

- testing.df: The table from the previous function that is to be filtered for de-novo mutations.

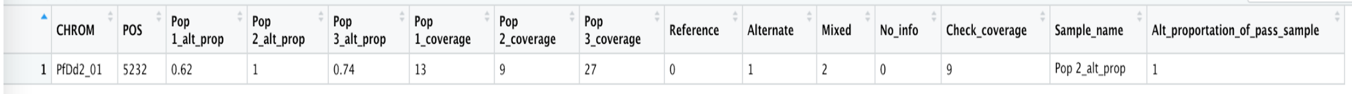
Output produced:

A table containing the de-novo mutations.

Sample input:

- *testing.df :* sample output table from adding_info_to_tables( ) function

Sample output:



	CHROM	POS	Pop 1_alt_prop	Pop 2_alt_prop	Pop 3_alt_prop	Pop 1_coverage	Pop 2_coverage	Pop 3_coverage	Reference	Alternate	Mixed	No_info	Check_coverage	Sample_name	Alt_proportation_of_pass_sample
1	PfDd2_01	5232	0.62	1	0.74	13	9	27	0	1	2	0	9	Pop 2_alt_prop	1

PART 2 : This section lays out the template on how to call the functions defined in Part 1 of the script.