Imperial College London

Estimating the number of independent origins of insecticide resistance for inference of recent effective population size of Anopheles gambiae

Xiaoya Zhang

August 2020

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Research at Imperial College London

Formatted in the journal style of Ecology Letters

Submitted for the MRes in Computational Methods in Ecology and Evolution

Abstract

The origin of new mutations within a species is a fundamental process of evolution. However, most of the processes by which mutations occur and species evolve are largely unexplored. In this report, I estimated the number of independent origins and calculated the most recent effective population size of Anopheles gambiae. In 2019, the soft-sweep theory was developed by Khatri and Burt, as a way to estimate the most recent effective population size, which makes use of the number of independent origins (Khatri and Burt, 2019). By implementing the formula developed by Khatri and Burt, the most recent effective population size of Anopheles Gambia, Ne, was developed successfully in this report. A296G on the RDL gene (dieldrin resistance locus) was selected as focal site, which is used for the number of independent origins estimation. In this process, Genealogical Estimation of Variant Age (GEVA) was implemented to calculate haplotype ages and analysis data, and it was developed by McVean's Group in 2019. Input data is 773 mosquito samples distributed in various African countries from Ag1000g Consortium (Ag1000g Consortium, 2014). Then, the obtained haplotype age was used for drawing a dendrogram. Finally, the number of independent origin is determined by the number of intersection points between the haplotype age threshold value and the vertical edge of the tree.

Contents

1. Introduction

In many countries, malaria has been eliminated to some extent through the appropriate use of insecticides. However, the use of pesticides and anti-parasitic drugs has other effects. For example, the rise in insecticide resistance of anopheles mosquitoes, which transmit malaria, threatens malaria control. In Africa, the main vector of Plasmodium falciparum malaria is the blood-sucking mosquito of the Anopheles gambiae. Malaria morbidity and mortality have been significantly reduced as insecticide - based interventions have been used. But these advances are likely to be hampered by rising levels of pesticide resistance and other adaptations of mosquito populations. In conclusion, research of the mutation and evolution of insecticides is important to the exploration of malaria transmission. Particularly, estimation of the time the resistance mutation arose time is crucial aspect of malaria control.

Genomics have been used to address these resistance mutation problems. To better research the spread of resistance between species and geographic location, the genetic background of the non-synonymous alleles found can be analyzed to determine which alleles have undergone recent positive selection. (Clarkson, 2018). Using genome-wide Illumina sequence data of eight African countries from the Anopheles Gambiae 1000 Genome Project (Ag1000G), they provided comprehensive data on genetic variations in the Vgsc gene in mosquito populations. They obtained previously unknown information about anopheles populations: three known resistance alleles and 20 non-synonymous nucleotide substitutions were described. These substitutions occur fairly frequently in one or more populations of Anopheles mosquitoes. They analyzed the genetic backgrounds of the non-synonymous alleles they found to determine which alleles had undergone recent positive selection. The aim is to improve our comprehension of the spread of resistance between species and geographic locations (Clarkson,2018). To assess whether gene drive strategies are vulnerable to resistance arising — resistance arises more easily in larger populations, so it is important to know the most recent effective population size.

The question addressed in this report is to estimate the how many independent origins there are, and to calculate the most recent effective population size for Anopheles Gambiae. To achieve this aim, GEVA can be applied to calculate Haplotype Ages and analyze DNA sequences at rdl loci involved in insecticide resistance for the first step, which is to estimate the number of individual origins. And for the second step, the soft sweep theory proposed by Khatri and Burt can be used to estimate the most recent effective population size (Khatri and Burt, 2019) .

Actual populations are almost impossible to measure, but the effective population size for per generation, Ne, can be estimated, which is the population size required to achieve the same nucleotide diversity as the actual population. There exist several population Genetics methods that can be used to calculate the effective population size (Ne). For example, there exists methods that use linkage disequilibrium, nucleotide diversity and demographic history. Nevertheless, due to the constrains of each of them to some extent, the effective population size of A. gambiae cannot be estimated. For instance, it is very prevalent to use nucleotide diversity for estimating the effective population size. However, due to ancient bottlenecks, the resultant Ne is often skewed (Karasov et al., 2010). In addition, a few advantageous alleles are retained and spread throughout the population when selection pressures occur. As time goes by, their nucleotide differences may not be significant because they are the offspring of ancestors who carried favorable alleles. Method applying linkage disequilibrium studies the occurrence frequency of microsatellite-like loci and their physically linked chance. The method of linkage disequilibrium cannot estimate the large population accurately. These methods take into account stronger genetic drift in small populations (Waples and Do, 2010). For methods that infer demographic history, it is computationally challenging and sensitive to long term changes in populations (Liu & Fu, 2015), which means that they are complex. Instead of tending to size of the most recent effective population, they are better suited to detect more dated population (Khatri and Burt, 2019). In general, methods to estimate the nearest Ne of A. gambiae are limited.

In 2019, Khatri and Burt proposed soft sweeps theory for Ne calculation. This method implements a semi-deterministic forward-time method. The soft sweep theory targets in a different genomic background which have multiple copies of the same mutation. They come from soft sweeps. The backgrounds provide more recent information on the effective population size for per generation than the Ne of nucleoside diversity calculations. And these backgrounds can represent the history of recent population. In addition, the soft sweep theory gives a robust estimate of the most recent effective population size. In estimating the recent Ne value, it makes use of robust estimation on large population size. Khatri and Burt applied the soft sweep theory to the voltage-gated sodium channel (vgsc) gene, which has two mutations in knockdown resistance sites (kdr) that make it resistant towards pyrethroid (Reimer et al., 2008). Soft sweep results showed that the estimate of the most recent Ne was two orders of magnitude larger than the Ne calculated using nucleotide diversity (Khatri and Burt, 2019). Then the simulation data are tested and proved the robustness of the method.

According to soft sweep theory, the most recent Ne calculation requires two parameters because the species under study are diploids: the number of independent origins, and the mutation rate per generation per base. For the A. gambiae population studied in my report, the given mutation rate has been inferred from the mutation rate of drosophila (Keightly et al., 2014) and is a constant. Thus, only the number of independent origins of the Gambian population is required in order to obtain the effective population size.

2. Aim

The report aims to analyze one resistance mutation in the rdl locus to estimate number of origins. The non-synonymous mutation is on the rdl gene. The focal site is an A to G mutation on position 296, and is used for calculating the number of independent origins, by applying the method GEVA. After that, haplotypes were grouped by adopting Hierarchical Clustering Analysis (HCA) according to age. Then, dendrograms were drawn to show distance relationship between haplotypes. The number of independent origins can be determined from the dendrogram produced by GEVA. At last, the number of independent origins, η, is put into Khatri and Burt formula. Estimation of recent effective population N_e is output of that.

3. Background

The data for this report is from two collections by the Ag1000G Consortium. In 2014, Ag1000G consortium posted 773 sequence data from 8 countries. In 2017, The Ag1000G Consortium collected 1,142 samples of A. gambiae and A. coluzzi from 13 African countries in its second round. *A. gambiae* populations can be seen as a species complex which includes at least 7 species. Members of this species are geographically separated, and in sub-Saharan Africa, migration between populations is also constrained by various ecological conditions (The *Anopheles gambiae* 1000 Genomes Consortium, 2017). Mutations occur in generation to generation. Due to this, various mutations accumulate in the genomes of each subspecies, as well as individuals within the subspecies. Key point is that it is believed that the mutation is under selection which mean we can apply the soft-sweeps theory. In order to research the different genomic backgrounds, one non-synonymous mutation is set as the focal mutations on the *rdl* gene. In *rdl* gene, there are 7 non-synonymous mutations and the focal mutation is A296G. It is corresponding to position 25429236 bp of chromosome 2L of A. gambiae.

4. Material

Sampling mosquitoes

Details of sampling of mosquito population and sequencing method are described by Miles *et al.* From the Anopheles gambiae 1000 Genomes project (Ag1000G) consortium，sequences of *Anopheles gambiae* species complexes are obtained. It includes *Anopheles gambiae* and *Anopheles coluzzi*. The data was got as variant call format (.vcf) which can run in geva directly. The genome sequences of A. gambiae species complexes in Africa in 2014 is shown in phase1. In phase1, 773 mosquitoes are sequenced from 8 African countries. The url link for phase1 data is /production/ag1000g/phase1/AR3.1. *A.* gambiae species complexes in 2017 shown in phase 2 were sequenced in a larger scale. In phase 2, 1142 mosquitoes are sequenced from 13 African countries. The url link for phase2 data is /production/ag1000g/phase2/ar1 for phase 2. In phase 2, there are three NA in the species column, which means in Ag1000g consortium published in 2019 the species sequenced in the corresponding country are not given.

Besides, part of my programs were run on High-Performance Computing system (HPC) of Imperial College London, which is extracting the rdl gene from the whole data part. Since the data needed to be analyzed was very big.

Sample Country	Species	Sample number
Angola	*A.coluzzi*	120
Burkino Faso	*A.coluzzi*	138
	A.gambiae	162
Cameroon	*A.gambiae*	550
Gabon	*A.gambiae*	112
Guinea	*A.gambiae*	62

Guinea-Bissau	*A.gambiae*	92
Kenya	*A.gambiae*	88
Uganda	*A.gambiae*	206

Phase 1| (published in 2017)

Sample Country	Species	Sample number
Uganda	*A.gambiae*	112
Angola	*A.coluzzi*	78
Bukina Faso	*A.coluzzi*	75
	A.gambiae	92
Cote d'Ivorre	*A.coluzzi*	71
Cameroon	*A.gambiae*	297
Mayotte	*A.gambiae*	24
Gabon	*A.gambiae*	69
Ghana	*A.coluzzi*	55
The Gambia	*NA*	65
Guinea	*A.gambiae*	40
Guinea-Bissau	*NA*	91
Kenya	*NA*	48

Phase 2| (published in 2019)

5. Methods

Genealogical Estimation of Variant Age (GEVA) is a method and was used as main methods to calculate the time to the most recent common ancestor (TMRCA) between every pair of the haplotypes for both rdl and vgsc. They output haplotype ages. If the age of one haplotype is one unit older than that of another haplotype, that haplotype is one generation older from the common ancestor than the other haplotype. This report studies and uses GEVA methods, developed by Albers and McVean, in method section, and the results of the GEVA is analyzed in the results section.

5.1) Genealogical Estimation of Variant Age (GEVA)

5.1.1) GEVA Background and algorithm

GEVA is an algorithm for haplotype age estimation. A brief introduction of GEVA is crucial for describing its background, process and output. It was published in 2019 by Albers and McVean. GEVA involves coalescent modeling and infer the time to the most recent common ancestor(TMRCA) between individual genomes. Using data obtained from large-scale sequencing, it is a nonparametric method for estimating the age of origin of single nucleotide polymorphisms at a single site in a target genome. Nonparametric means that the population data does not need to satisfy a basic assumption, such as that the population is normally distributed, or that the method assumes that evolutionary events shape the current population (Albers and McVean, 2019).

Mutation and recombination are two main events. Mutation accumulate in haplotype generation by generation, which increase differences between haplotype pairs. If a recent haplotype is a descendant of the initial haplotype, over many generations, the recent haplotype will exist a lot of mutations compared to the initial one. Recombination break the initial haplotype length in each generation, as it removes some ancestral haplotype segment which leads to the recent haplotype length shorter than the ancestral one.

In order to infer the age of TMRCA and get useful information, GEVA includes three coalescent “clock” model, which are got from Bayesian probabilistic

approach. They are mutation clock, recombination and joint clock, which are named after the information they used. Mutation clock means mutational differences are used for infer the age, recombination clock means recombination distance are used during age inference, and joint clock considers both mutation and recombination events.

The approach used in GEVA is in Figure 1. There still presents today a copy of the piece of the ancestral chromosome on which the mutation occurred in the individuals carrying the derived allele (Figure 1). As time pass by, additional mutations accumulate along the haplotypes, whose length is broken by recombination during meiosis in each generation. The information available in whole-genome sequencing data is made full use of to perform comparisons between concordant pairs and discordant pairs.

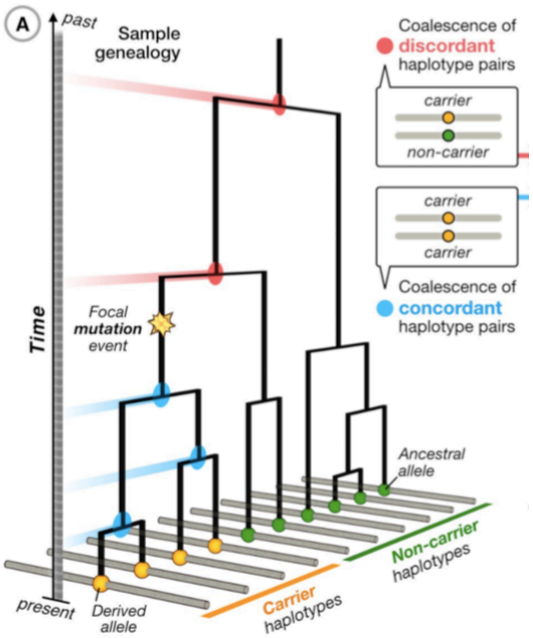


Figure 1|Overview of the GEVA method (Albers and Mc Vean, 2019)

Concordant pairs can be described as two haplotypes, and they both carry the focal mutations. Discordant pairs pairs can be described as two haplotypes. One of them carries the mutation, and the other one is wild type. Concordant pairs coalesce with each other more recently than the time point of the mutation event.

Discordant pairs compare genealogical relationships that are both younger and older, respectively, than the time of mutation. The information from all haplotype pairs is then combined within a composite likelihood framework, so that a derived approximate distribution of age is obtained.

GEVA uses recombination rate and mutation rate for input, and it outputs the shape parameter and the rate parameter, which are necessary for gamma distribution of the TMRCA between two haplotypes. At the same time, GEVA also estimate the shared length. During this, a hidden Markov Model (HMM) is used as probabilistic approach. After running GEVA, with
the shape and rate parameters, a posterior distribution on the TMRCA of given pairs can be obtained. And then allele age can be estimated from the composite posterior distribution, which combines the pairwise TMRCA posteriors available for a given focal variant.

5.1.2) Parameters and formula for TMRCA age calculation

Ne, values of mutation rate, and recombination rate, are three values set as the parameters for GEVA, important to be specified.

Phase 1 data was used as GEVA input. For the reason that GEVA only works on phased data, the phase 2 data obtained from Ag1000g are unphased, which are unacceptable in GEVA. 773 mosquitoes' data sequenced from 8 African countries are kept in the variant call format (.vcf file), and use 0/1 represent their genotypes. Wild type genotype is represented by 0, certain mutation is represented by 1. The rdl gene is from 24863652bp to 25934556bp at chromosome 2L. The rdl focal mutation is A296G locus mutation, and the position is 25429236 bp at chromosome 2L. Aiming to compare the robustness of GEVA, Ne values were set as the parameters for GEVA calculate. Besides, specified mutation rate value and specified recombination rate value are needed in GEVA. Recombination rate is set as 0.5 (cM/Mb), cM is centimorgan and the unit is defined as 1% chance within a million base pairs. The 0.5 value is taken from chromosome 2L recombination map from 2,487,770 bp to 49,364,325 bp region. 0.5 (cM/Mb) has the same meaning with a rate of 0.5×10^{-8} per base pair.

There are two methods for estimating TMRCA. The first one is directly from GEVA and use the mutation clock. It is calculated using the following formula, where α is the shape parameter, and β is the rate parameter:

$$(\alpha/\beta)\times 2\times N_e \qquad (1)$$

Another method is that the physical shared genetic length can be used to calculate age, using mutation clock. It can attest the robustness in age estimation. The formula is below, Where RHS is the chromosomal position of the haplotype break point on the right side of focal mutation, and LHS is the chromosomal position of the haplotype break point on the left side:

$$1 / [(RHS - LHS) \times 2 \times (recombination\ rate)] \qquad (2)$$

5.2) Hierarchical clustering analysis (HCA)

After haplotype ages are constructed successfully, the average linkage method in hierarchical clustering analysis (HCA) is used for haplotype ages clustering. This function performs hierarchical clustering analysis which uses the differences among the haplotype ages that are clustered. Initially, each haplotype age is assigned to one cluster that include itself only. Secondly, the algorithm iterates, adding the two most similar clusters at each stage until there is only one cluster. According to the “the average” clustering method adopted, the distance between the clusters is recalculated at each stage. Many different clustering methods are provided in HCA. Minimum variance method is designed to find compact, spherical clusters. The single link method uses a "friend of a friend" clustering strategy. The complete method finds similar clusters. Other methods can be considered as clustering for methods that have characteristics between single and complete linked methods. Therefore, the complete method is the most suitable for our clustering aim.

Hierarchical clustering analysis algorithm. For example, seven different haplotype ages are shown here, each represented by a different color dot. The top of the timeline represents a long time ago, while the colored points are located closer to a younger age. horizontal edges represent Merge events, while vertical edges represent the distance on the timeline between clusters. Firstly, two points of similar color are grouped together. Secondly, another similar color

is merged into an existing group. Then the different groups merge in the same way.

5.3) Calculate the recent effective population size N_e

The recent effective population size Ne was calculated by Khatri and Burt using soft sweeps. The number of independent origins of the A296G mutation was used for estimation (Khatri and Burt, 2019). Mutation rate (μ), population size (N), the chromosomes number (n_S) in the samples and the total mutant frequency at time T x(T) are required parameters for calculating N_e.

Using the number of independent origins calculated from the methods above, η, the recent effective population size can be calculated using the following maximum likelihood function.

$$\overline{\eta}(T) = 2N\mu \ln\left[1 + \frac{n_m}{2N\mu}\right] \qquad (3)$$

$$p(\eta(T)|N, s, \mu) = L\big(N, s, \mu|\eta(T)\big) = \frac{\overline{\eta}(T)^{\eta}}{\eta!} e^{-\overline{\eta}|T|} \qquad (4)$$

Estimate of Ne can be achieved by using this maximum likelihood function, given the independent origins number. There is a maximum point in this function, which corresponds to the most likely Ne. For Ne calculation, the 95% confidence intervals for N_e are e^{-2} units from N_e (Chandradeva,2019).

6. Result

6.1) Allele frequency of A296G mutation

On 25,429,236 bp, a C to G mutation happens, which is non-synonymous. It is from alanine to glycine mutation. A basic analysis of information of A296G mutation was set up for understanding the data being processed. In addition to the A296G mutation, a small part of mosquitoes sampled in the Ag1000g phase 1 and 2 data showed the A296S mutation instead of the A296G mutation. On chromosome 2L a nucleotide changes in 25429235 bp from G to T lead to alanine to serine non-synonymous mutations. In phase2, there were 489 samples has A296G mutations in 1,142 samples. Of these, 179 (15.7%) were homozygous A296G mutations and 310 (27.1%) were heterozygous A296G mutations. In phase 1, 26.1% of the sampled mosquitoes had mutations in A296G (Chandradeva, 2019). In phase 2, there were 66 samples has A296S mutations in 1,142 samples. Among these, there were 39 heterozygous A296S mosquitoes (3.4%) and 27 homozygous mosquitoes (2.4%). Thus, the allele frequency of G genotype was 0.293, while the allele frequency of A296S was 0.041.

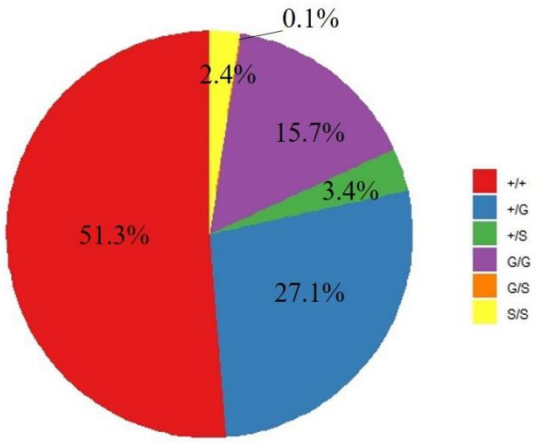


Figure 2| Pie chart showing the percentage of the genotypes, phase 2 in this report. The data is from Ag1000g consortium in 2017.

6.2) GEVA

6.2.1) TMRCA calculation

- **TMRCA calculation using mutation clock**

The results of GEVA are available for the three clock models. Since only one non-synonymous mutation site 25429236 was included, the results of the mutation Clock model were only considered in this study.

GEVA outputs a pairs file, in which containing two parameters, the shape and rate, which are useful for constructing gamma distribution of the TMRCA between two haplotypes. The peak of the gamma distribution represents the most likely haplotype pairs TMRCA. α represents the shape parameter, and β represents the rate parameter, and the TMRCA is calculated by the following formula:

$$(\alpha/\beta)\times 2\times N_e \quad (1)$$

The corresponding histogram was constructed based on this TMRCA formula. As shown in the Figure3, it can be clearly seen that the overall range of TMRCA values is from 81.5 to $4.67*10^5$. The oldest haplotype age calculated was $4.67*10^5$ generations, while the youngest was 81.5 generations. The general trend of the Figure3 is increase to one peak and fall. It is clear that there are only one peak to the age of $3.06*10^3$.

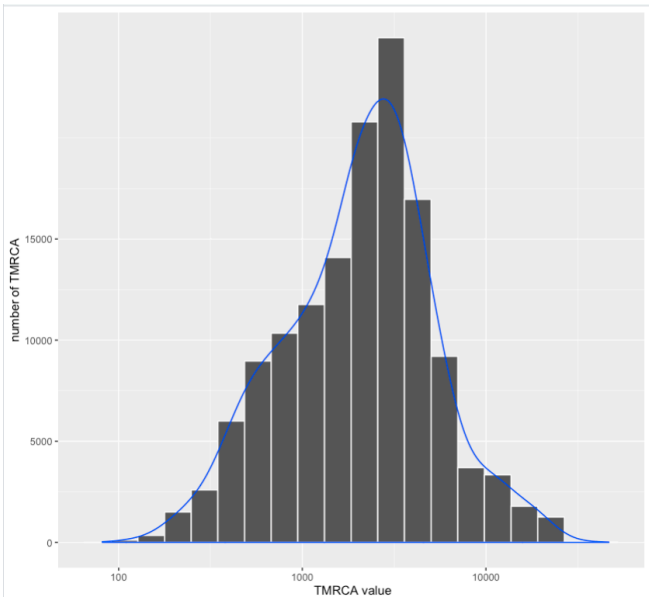


Figure 3| Histograms for TMRCA distribution calculated with mutation clock, using Ne = 10,000.

- **TMRCA calculation from the shared haplotype length**

In addition, there is a simple method that can be used to attest the robustness in age estimation of shared haplotype pairs. It directly uses the physical shared length for age calculation. The shorter the shared haplotype length is, the older the TMRCA is from current:

$$1 / [(RHS - LHS) \times 2 \times (\text{recombination rate})] \qquad (2)$$

The corresponding histogram was constructed based on this TMRCA formula. As shown in the Figure 4 below, it can be clearly seen that the overall range of TMRCA values is from 800.8 to $3.33*10^7$. The oldest haplotype age calculated was $3.33*10^7$ generations, while the youngest was 800.8 generations. The general trend of the Figure 4 is first increase to one peak and decline, then to rise to the second peak and finally fall. It is clear that there are two peaks and a trough to the age of $6*10^4$, so it can be divided into two populations using this as a threshold. One population ranges from 800.8 to $6*10^4$, and the other from $6*10^4$ to $3.33*10^7$.

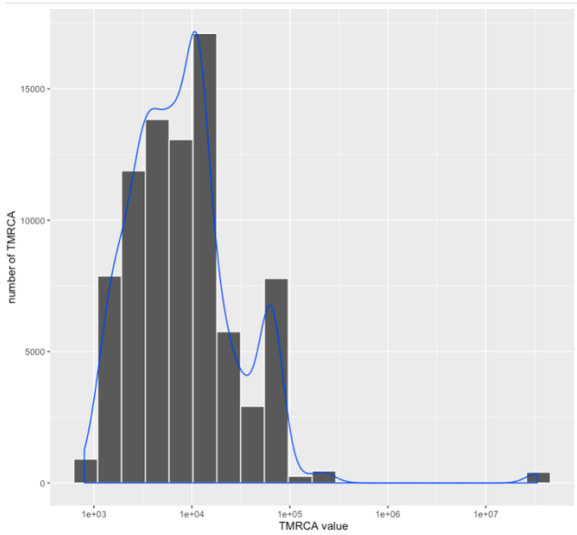


Figure 4| Histograms for TMRCA distribution calculated with mutation clock and the shared haplotype length, using Ne = 10,000.

6.2.2) GEVA performed on N_e = 10,000

GEVA was performed on one effective population size Ne 10000, using mosquito data from phase1 based on A296G mutation. And the mutation rate was 3×10^{-9}(Khatri and Burt, 2019). The input of GEVA is rdl gene information from 773 samples vcf file of Ag1000 Genomes project, including chromosome name, position, mutation information, genotype of each individual and so on. By executing the program as described above, two result files are created, which are pairs file and sites file. The pairs file contains the results of all pairwise analyses conducted to estimate the ages contained in the sites file. Pairs file is based on mutation clock model, recombination clock model and joint clock model. It includes some parameters for each pairs, which are Segment LHS, Segment RHS, shared, shape, rate and so on.

As a result, a total of 1617,900 haplotype pairs was obtained. Since there were three clock models implied by GEVA, the sample number accounted was 539,300 in total. However, GEVA only analyzed concordant and discordant pairs, meaning that no wild-type homozygous individuals were considered. As discussed in the discussion section, if GEVA also considers wild-type integration

pairs, in theory, since mosquitoes are diploid, there are 773×2 = 1546 haplotypes. Then, after sieving the repeats, there should be n×(n-1)/2 haplotype pairs, not including the haplotype pairs compared to themselves. Therefore, if the wild-type homozygous individual is included, the number should be equal to 1,194,285. The shape and rate parameters are used to obtain a posterior distribution on the TMRCA of given pairs. Allele Age is estimated from the composite posterior distribution.

6.2.3) TMRCA distribution –finding threshold

- The reason why a threshold is important: Assuming that, in top Figure 5 below, the red tips are haplotypes with the resistance mutation, black tips are haplotypes with the wild type mutation. The branches have also been coloured accordingly. So it is easy to pick out the independent origins. And in bottom Figure 5, in haplotype age distribution, there are two peaks and one intersection point. The two peaks mean different populations and the pink vertical line represents for pairwise haplotype age, which is threshold. Therefore, it is very crucial to find the threshold.

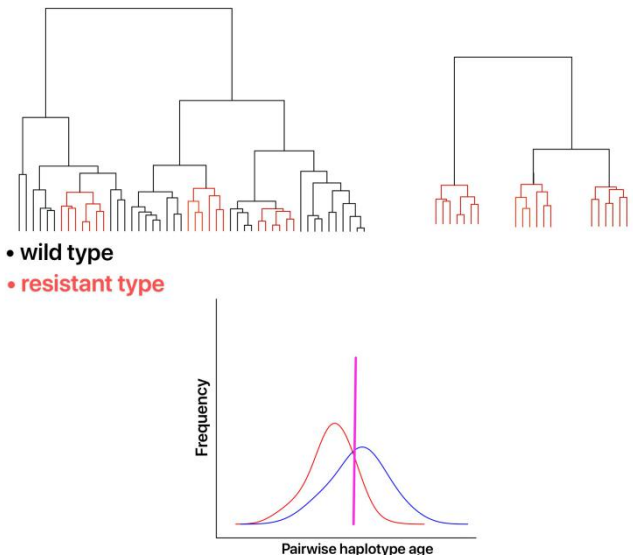
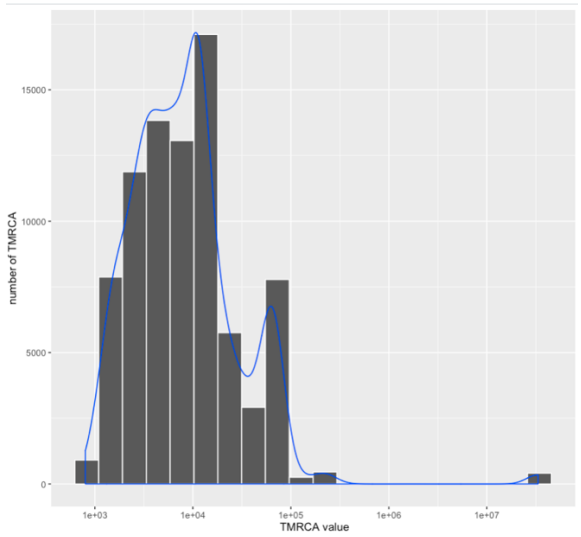


Figure 5| Hypothetical dendrogram and haplotype age distribution

- The blue curve is drawn to show the distribution more clearly. Thus, thresholds of consistent wild homozygous age are easy to determine. As shown in Figure4, the age distribution is calculated and the threshold is successfully obtained. It is clear that there are two peaks and a trough to the age of $4*10^4$, so it can be divided into two populations using this as a threshold. One population ranges from 800.8 to $4*10^4$, and the other from $4*10^4$ to $3.33*10^7$.



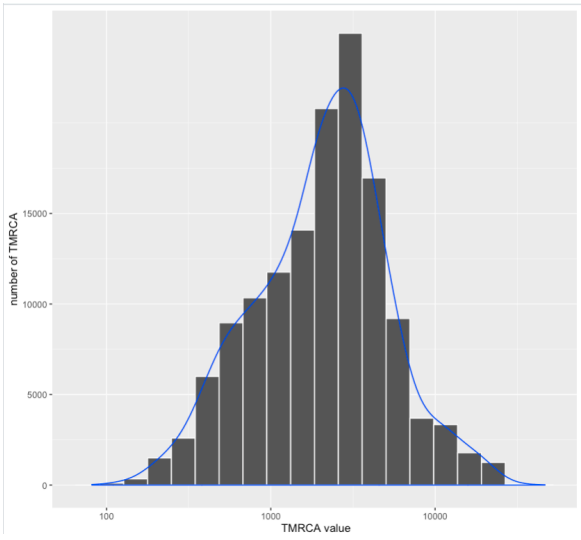


Figure 3 and Figure 4

6.2.4) Dendrogram

GEVA only analyzed consistent and non-consistent haplotype pairs, so the analysis of wild-type homozygotes was ignored. In order to solve this defect, we extracted the TMRCA of the identical haplotype pair after the interchange between the genotype of the mutation site and the reference site as the result of the homozygote of the wild type. The results of the non-consistent haplotype pairs after the swap were the same as those before the swap. On this basis, we obtained TMRCA of mutant homozygous and wild-type homozygous homozygous pairs and TMRCA of non-uniform haplotype, and constructed the full matrix based on TMRCA.

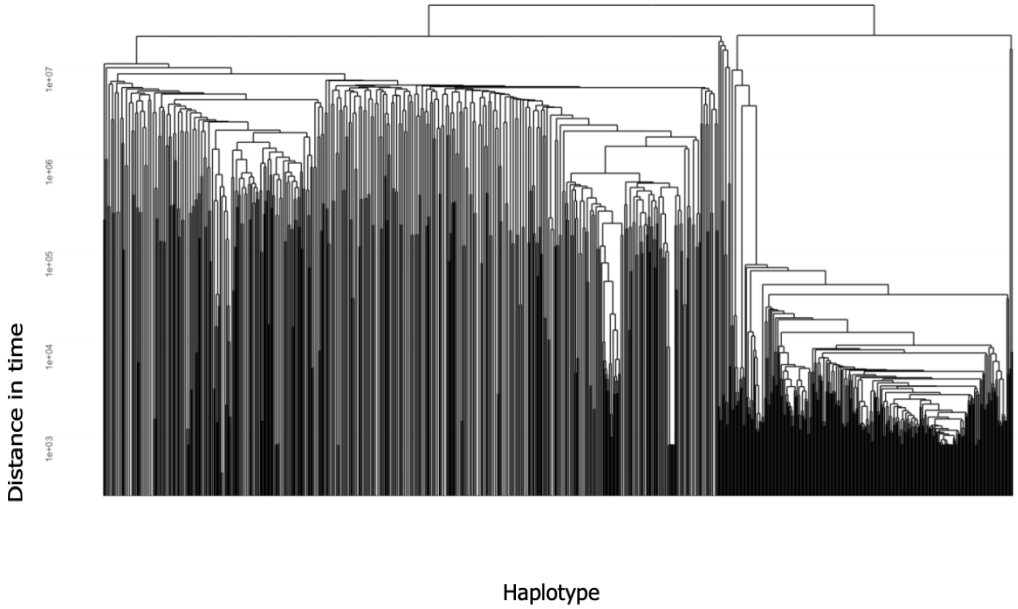


Figure 6| Dendrogram for TMRCA distribution (full tree), using the shared haplotype length, Ne = 10,000. This full tree uses TMRCA of R*R, R*W, and W*W (R represents for resistant type while W means wild type, R*R means TMRCA between resistant haplotypes are concordant). As is shown in figure above, the y-axis is using log scale. It's not evenly distributed on the Y-axis. The closer to the top of the Y-axis, the longer the distance. The closer to the bottom of the Y-axis, the shorter the distance.

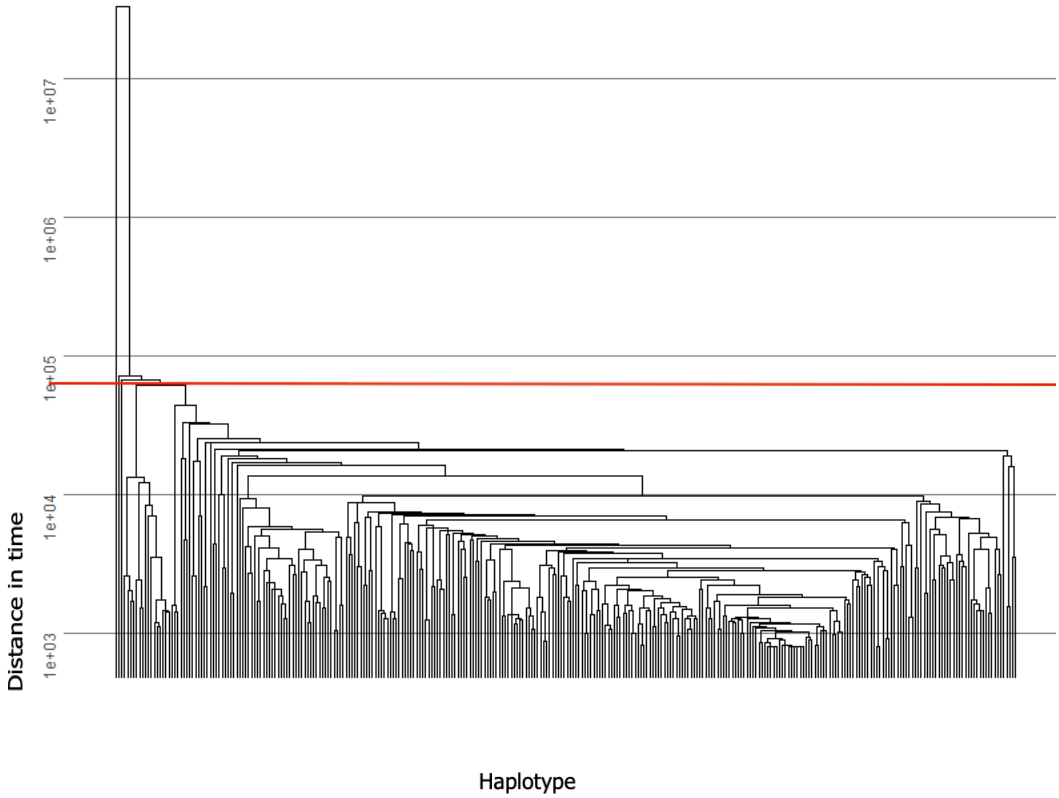


Figure 7| Dendrogram for TMRCA distribution (R*R), using the shared haplotype length, Ne = 10,000. This full tree uses TMRCA of R*R. As is shown in figure above, the y-axis is using log scale. It's not evenly distributed on the Y-axis. The closer to the top of the Y-axis, the longer the distance. The closer to the bottom of the Y-axis, the shorter the distance. Then, the red line represents the age threshold which is 8,0000 generations. The number of intersections between the red line and the R*R tree is the number of independent origin. As is shown in this figure, there are 4 independent origins.

6.2.5) The recent effective population size Ne calculation

The recent effective population size can be calculated using the number of independent origins, by implementing formula (3) and (4) in this report. Since the result is the number of chromosomes in the population, and mosquitoes are diploid, the calculated Ne should be divided by two. Then, to get the number of mutants, the population size was multiplied by the mutant allele frequency. The mutant allele frequency is 0.33319. And then finally when you plug that into the formula, the highest point of the function is the most likely nearest effective

population size. The independent origin number obtained is 4, the Ne calculated from formula (3) and (4) is 7.91 × 107 (95% confidence interval: 1.92 × 107, 2.22 × 108).

7. Discussion and Future Perspectives

7.1) Project result

In general, the main findings of this thesis is Ne of A. gambiae was identified, using the haplotype ages as a reference. Ne is calculated from the number of independent origins of focal mutation A296G within rdl gene loci. The number of independent origins is 4, which is based on TMRCA between concordant resistant haplotypes (R*R). Thus, after implementing 4 into the formula obtained by Khatri and Burt, the result is 7.91 × 107 (95% confidence interval: 1.92 × 107, 2.22 × 108). These result is quite consistent with Yao and Chandradeva. Yao's result is Ne is 1.622 × 107 (95% confidence interval: 3.87 × 106, 1.03 × 108) by an independent origin of 1, and Ne is 7.91×107 (95% confidence interval: 1.92 × 107, 2.22 × 108) by an independent origin of 4. Chandradeva's result are 2 independent origins gives a Ne of 2.20 × 107 (95% CI.: 2.80 × 106, 1.2 × 108), or 5 independent origins gives a Ne of 7.08 × 107 (95% confidence interval: 2.00 × 107, 1.73 × 108).

7.2) Result improvement and Future Perspectives

The number of independent origins based on R*R is obtained successfully. However, due to the limitation of time, that based on full tree is not be calculated, which contains R*R, R*W, and W*W. Although the dendrogram for TMRCA distribution has been created using the shared haplotype length, Ne = 10,000, as is shown in figure 6. The number of independent origins cannot be obtained by this way. Because when a horizontal line is drawn on the full tree dendrogram, the results show many independent origins that are uncountable. It implies that the method for counting the independent origin number on a full tree should be improved in the future.

Moreover, a TMRCA distribution calculated from the shared haplotype length has been created, which is based on full matrix (Figure 8). It is also necessary for finding threshold and estimating independent origins using full tree. The process for generating it is: After exchanging the genotypes of the mutation site and the reference site, a full matrix based on TMRCA was constructed, which includes

R*R, R*W, and W*W. The matrix contains a total of 773*(773-1)/2 haplotype pairs. The TMRCA range of consistent mutant homozygous haplotype pairs based on mutation Clock model is 800.8~3.33*107, with an average TMRCA of 1.82*105. The TMRCA range of homozygous haplotype pairs was 800.8~5.1*107, and the average TMRCA was 6.49*106. The TMRCA range of non-uniform haplotype pairs was 887.16~5.2*107, and the average TMRCA was 4.72*107. Obviously, the mean TMRCA of non-consistent haplotype pairs was significantly higher than that of consistent wild homozygous haplotype pairs and homozygous mutant homozygous haplotype pairs. And the mean TMRCA of homozygous mutant haplotype was significantly lower than that of homozygous haplotype pair of wild homozygous mutant. The same mutant homozygous haplotype had the smallest range of TMRCA variation, while the same wild type homozygous haplotype pair and non-uniform haplotype pair had a larger range of TMRCA variation. This result indicates that the homozygous mutant has a smaller generation number between haplotypes, while the homozygous wild type has a larger generation number between haplotypes and non-uniform haplotypes. These results are consistent with the theory. In general, the Full TMRCA distribution should be important for future analysis on finding independent origin number on a full tree.

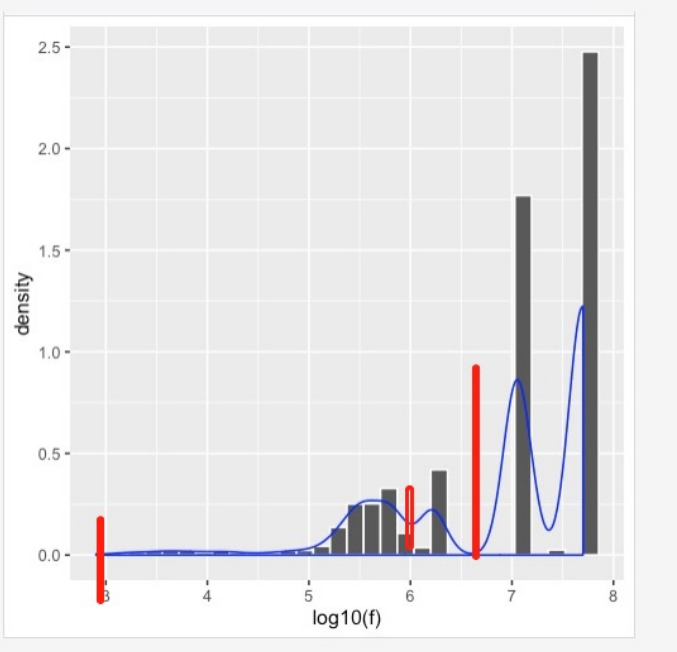


Figure 7| Full TMRCA distribution calculated from the shared haplotype length

8. Conclusion

In this report, the independent origins number of A. gambiae was estimated and the recent effective population size (Ne) was calculated using the soft-sweep theory was developed by Khatri and Burt. First, GEVA was implemented for TMRCA distribution (histogram) using haplotype ages of the focal site A296G within rdl gene. Secondly, the independent origins number was obtained as 4. After that, an age threshold is got from histogram plotted. Then, the independent origins number was used for drawing a dendrogram. This is based on haplotype ages and this gives a more accurate estimation than pairwise genetic distance, since haplotype ages considers recombination. Finally, the recent effective population size was calculated by implementing the formula developed by Khatri and Burt. The recent effective population size result is 7.91 × 107 (95% confidence interval: 1.92 × 107, 2.22 × 108).

Acknowledgements

I would like to express my greatest gratitude here to my supervisors, Dr. Bhavin Khatri and Dr. Vassiliki Koufopanou, for their guidance throughout the project and tireless help on the problems occurred, especially the during the hard days of GEVA running. I also would like to thank Prof. Austin Burt for his kindly assistance and first guidance. Many thanks to Rosary Yao for providing information that I need. Finally, I want to thank all those who helped me during this research project.

9. Reference

Khatri, B.S., and Burt, A. (2019). Robust estimation of recent effectivepopulation size from number of independent origins in soft sweeps.*Molecularbiology and evolution,* 36(9):2040–2052.

Hemingway, J. (2016). Averting a malaria disaster: will insecticide resistance derail malaria control? *Lancet* 387, 1785–1788

Bhatt, S. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature* 526, 207–211.

Clarkson, C.S., Miles, A., Harding, N.J., Weetman, D. *et al.* (2018). The genetic architecture of target-site resistance to pyrethroid insecticides in the African malaria vectors Anopheles gambiae and Anopheles coluzzii.. doi:10.1101/323980

Karasov, T., Messer, P.W. and Petrov, D.A. (2010). Evidence that adaptation in drosophila is not limited by mutation at single sites.*PLoSGenet*, 6(6):e1000924.

Waples, R.S., and Do, C. (2010). "Linkage Disequilibrium Estimates of Contemporary N E Using Highly Variable Genetic Markers: A Largely Untapped Resource for Applied Conservation and Evolution." *Evolutionary Applications*. 3.3: 244-62. Web.

Liu, X. and Fu, Y.X. (2015).Exploring population size changes using SNP frequency spectra. *Nature genetics*.[Online]47(5),555–559.Available from:doi:10.1038/ng.3254

Reimer, Lisa, Fondjo, Etienne, Patchoké, Salomon, Diallo, Brehima, Lee, Yoosook, Ng, Arash, Ndjemai, Hamadou M, Atangana, Jean, Traore, Sekou F, Lanzaro, Gregory, and Cornel, Anthony J (2008). "Relationship Between kdr Mutation and Resistance to Pyrethroid and DDT Insecticides in Natural Populations of Anopheles Gambiae." *Journal of Medical Entomology*. 45.2: 260-66. Web.

Keightley, P.D., Ness, R.W., Halligan, D.L. and Haddrill, P.R., (2014). Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. Genetics.. doi:10.1534/genetics.113.158758

Cordes, D., Haughton, V., Carew, J.D., Arfanakis, K., and Maravilla, K. (2002). Hierarchical clustering to measure connectivity in fmriresting-state data.Magnetic resonance imaging, 20(4):305–317.

Albers, P.K. and Mcvean, G. (2020). "Dating Genomic Variants and Shared Ancestry in Population-scale Sequencing Data." *PLoS Biology* 18.1: E3000586. Web.