Transmission analysis using 101 SNPs barcode data

18 April, 2024

Data Source

This analysis aims to use **101** SNPs barcode data and **1591** isolates to explore Plasmodium *falciparum* drug resistance profile in the Gambia.

Analysis steps

1. Genetic diversity analysis
    - Allelic richness
    - Mutations / Haplotypes Allele frequencies
2. Transmission dynamics analysis
    - Identity by state
    - Jaccard Similarity Coefficient
    - Sørensen-Dice Coefficient (Dice Similarity)
    - Fst (Fixation Index)
    - Nei's Genetic Distance (Nei 1972)
3. Temporal dynamics analysis
    - Temporal FST (Fixation Index)
    - Haplotype Diversity Over Time
    - Drug Resistance Analysis

Distrubution of samples by Location and Year

This subsection provides a comprehensive analysis of the distribution of **1591** samples collected across **101** loci, categorized by their geographic locations and the year of collection. We aim to present an overview of the overall distribution of samples, followed by a more detailed examination of the sample density in specific locations over various years. This will highlight areas and periods of high sampling activity and identify potential gaps in the data collection efforts.

Distrubution of samples after filtering

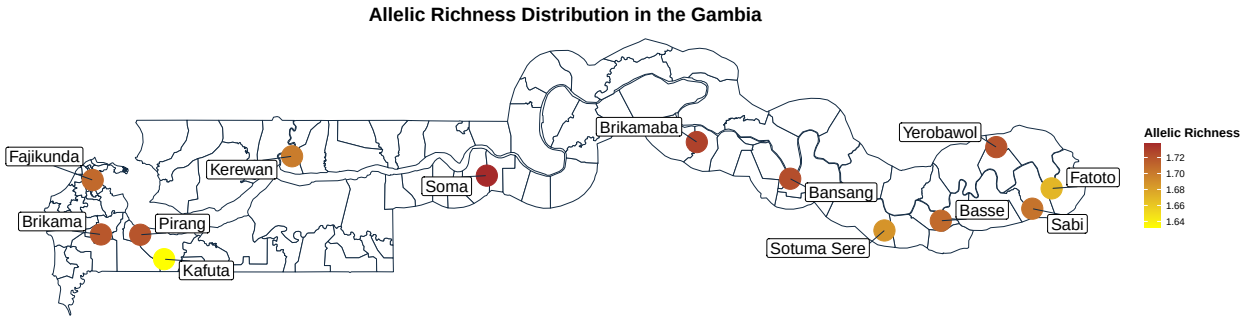
This subsection delves into the distribution of our dataset following the application of rigorous filtering criteria, aimed at refining the quality and relevance of the **1591** samples collected across **101** loci. The filtering process involves the exclusion of samples and loci with a high proportion of missing genotypes (>10%) to ensure the integrity and reliability of subsequent analyses.

After a meticulous filtering process, our dataset now comprises **523** samples out of the original **1591**, distributed across **100** loci, significantly refining our analysis pool. The filtering has not only rationalized the dataset for enhanced accuracy but also minimized potential noise, allowing for a more precise understanding of genetic diversity and other population genetic indices.

Genetic diversity analysis

Assess the genetic diversity of Plasmodium *falciparum* across different villages and time points. This includes calculating measures such as allelic richness, and SNP frequencies.

Allelic richness: Allelic richness is a measure of the number of alleles per locus in a given population, adjusted for the smallest sample size if there are varying sample sizes.

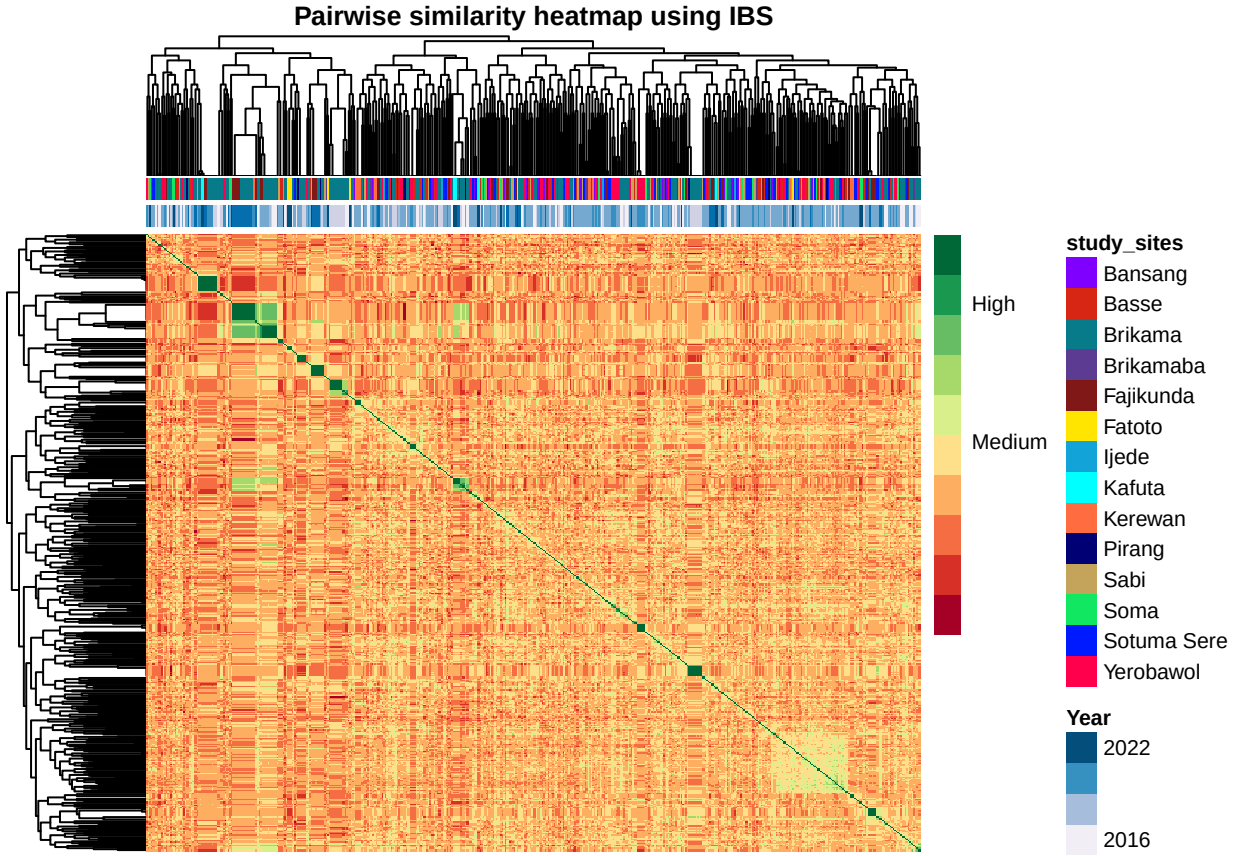


Transmission dynamics analysis

This analysis focus on the identification of genetic clusters of P. *falciparum* using the refined dataset of **523** samples and **100** SNPs derived from post-filtering and methods like network analysis. By applying mathematical models and statistical techniques, we aim to unravel the patterns of genetic exchange between individuals at different locations and over time. This analysis sheds light on the mechanisms driving the spread of genetic variants, contributing to our understanding of population connectivity, dispersal pathways, and barriers to gene flow. Through the integration of geographic and temporal data, we identify hotspots of genetic diversity and areas of significant genetic differentiation, which may suggest historical events or ecological factors influencing transmission dynamics.

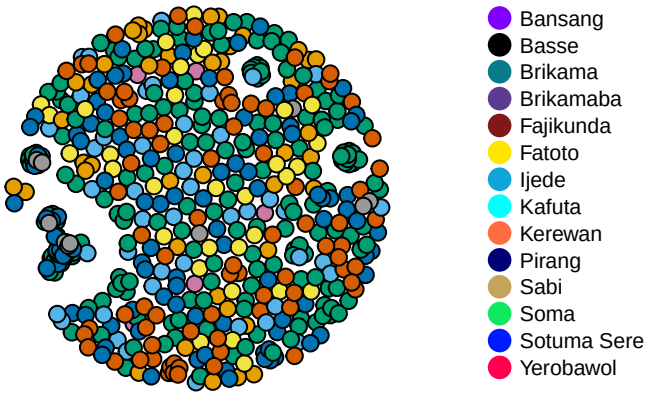
We will create network plot from genetic similarity indices such as IBS, Jaccard Coefficient, Dice Similarity, Fst and Nei's Genetic Distance

Identity by state (IBS): In the field of genetics, identity by state (IBS) refers to the case where two or more genetic sequences are identical because they share a common state, without necessarily deriving from a common ancestor. IBS plays a key role in understanding genetic similarity and diversity within populations, providing information about the genetic structure and relationships between individuals. This contrasts with identity by descent (IBD), where genetic similarities are due to inheritance from a common ancestor. Analyzing IBS across a dataset can illuminate patterns of genetic variation, contributing to the study of population genetics, and identification of potential genetic links in complex traits. IBS helps in quantifying the genetic similarity between pairs of samples based on SNP data, useful for identifying closely related Plasmodium strains or clones.



Network Analysis: Network analysis in genetic studies offers a powerful lens through which to explore the complex relationships and interactions among individuals or populations. By constructing networks that represent these connections, researchers can uncover underlying structures, identify key components or nodes, and understand the dynamics of genetic information flow. This approach facilitates the identification of clusters or communities within genetic data, revealing patterns of similarity, differentiation, or influence that might not be apparent through traditional analytical methods. In the context of this study, applying network analysis to the **523** samples and **100** SNPs can provide insights into the genetic landscape, highlighting potential pathways of gene transmission, genetic bottlenecks, or regions of high genetic diversity. This methodological perspective enriches our understanding of genetic relationships and contributes to a more nuanced view of population genetics.

Let's create Network Plot using IBS metric:



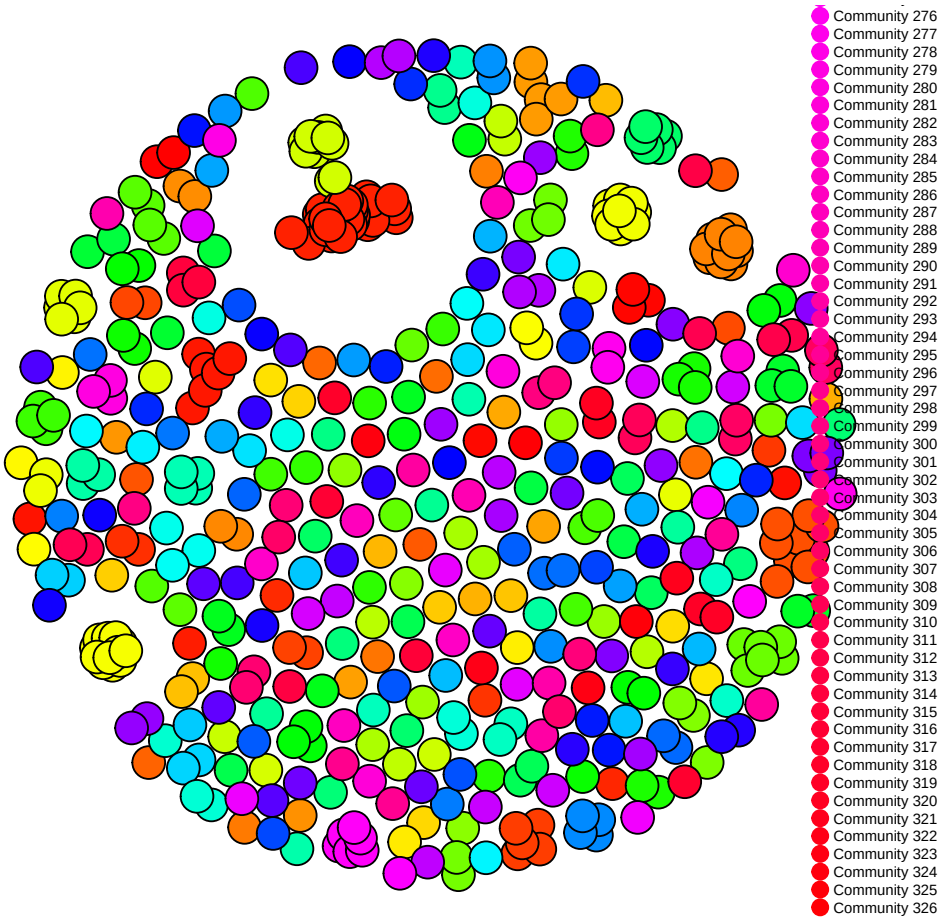
The generated network plot, with a strict threshold of **0.8**, reveals a distinct pattern of genetic connectivity between samples. This high threshold highlights a network characterized by a closely related groups, indicating high genetic similarity within the group, while also highlighting isolated nodes that represent unique genetic signatures. This visualization highlights the presence of significant genetic differentiation and potential barriers to gene flow within the study population.

Community Detection Methods When we analyze networks, it is important to discover communities. Communities are an important property of many networks in which a particular network may have multiple communities such that nodes inside a community are densely connected. These methods leverage algorithms to cluster nodes (e.g., individuals, or populations) based on patterns of similarity or connectivity, revealing underlying modular structures that might correspond to functional units, genetic similarities, or evolutionary relationships. It can be also used with machine learning to detect communities with similar properties. Moreover, since people tend to group with others similar to them, it helps us to identify users with a high number of connections and observe the depth of their reach within a network.The community detection algorithms (e.g., cluster_louvain) can help identify clusters within the network that represent closely related

groups of samples. By applying these techniques, researchers can uncover hidden patterns of organization within genetic data, offering insights into the dynamics of genetic variation, interaction, and evolution. This approach not only enhances our understanding of genetic architecture but also guides targeted research and conservation efforts by identifying critical components within the genetic network.

1. **Louvain algorithm**

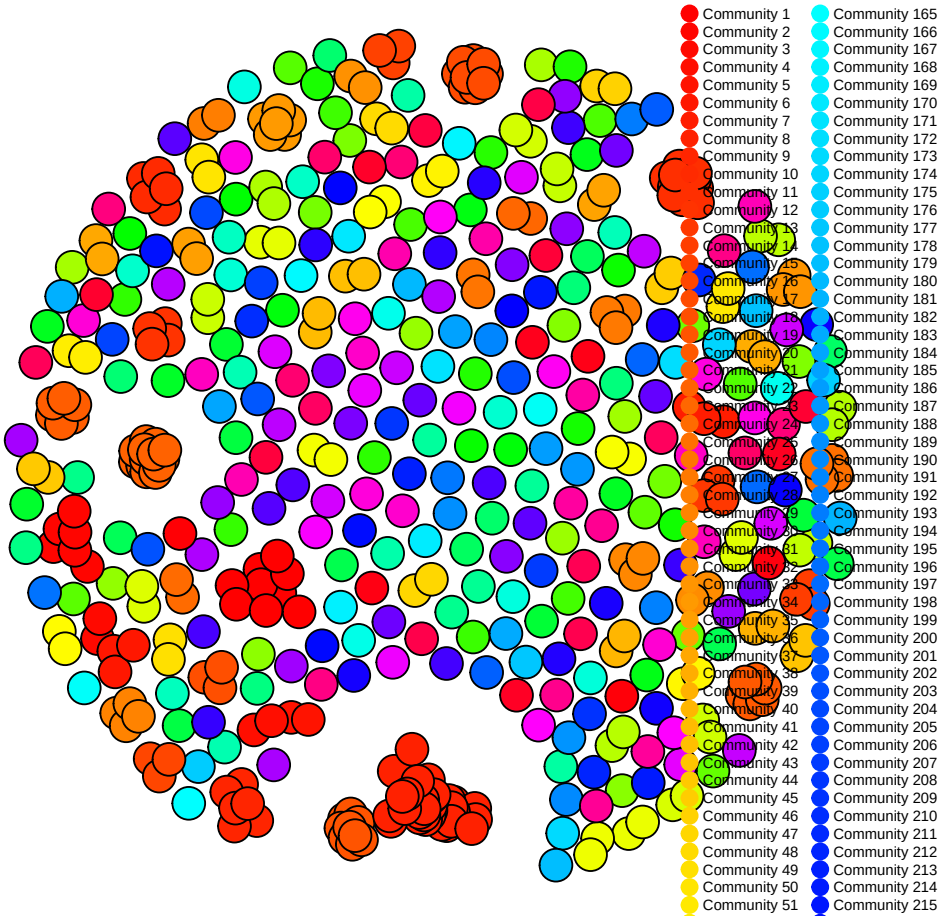
The Louvain algorithm presents itself as a very efficient and widely adopted method for detecting communities within complex networks. It's a simple algorithm that can quickly find clusters with high modularity in large networks. The so-called modularity measures the density of connection within clusters compared to the density of connections between clusters (Blondel 2008). Its iterative, hierarchical approach allows the algorithm to discover structures at different scales, from small, tightly knit groups to larger, less connected communities. The simplicity and scalability of the algorithm makes it particularly useful for analyzing large sets of genetic data, providing insight into the underlying structure and organization of genetic information, thereby facilitating deeper understanding. in-depth look at genetic relationships and evolutionary models. It is used as an objective function to be maximized for some community detection techniques and takes on values between -1 and 1.


Community 276
Community 277
Community 278
Community 279
Community 280
Community 281
Community 282
Community 283
Community 284
Community 285
Community 286
Community 287
Community 288
Community 289
Community 290
Community 291
Community 292
Community 293
Community 294
Community 295
Community 296
Community 297
Community 298
Community 299
Community 300
Community 301
Community 302
Community 303
Community 304
Community 305
Community 306
Community 307
Community 308
Community 309
Community 310
Community 311
Community 312
Community 313
Community 314
Community 315
Community 316
Community 317
Community 318
Community 319
Community 320
Community 321
Community 322
Community 323
Community 324
Community 325
Community 326

2. **Walktrap algorithm**.

The walktrap algorithm is used to identify communities in large networks via random walks. These random walks are then used to compute distances between nodes. Then, nodes are assigned into groups with small and large community distances via bottom-up hierarchical clustering.

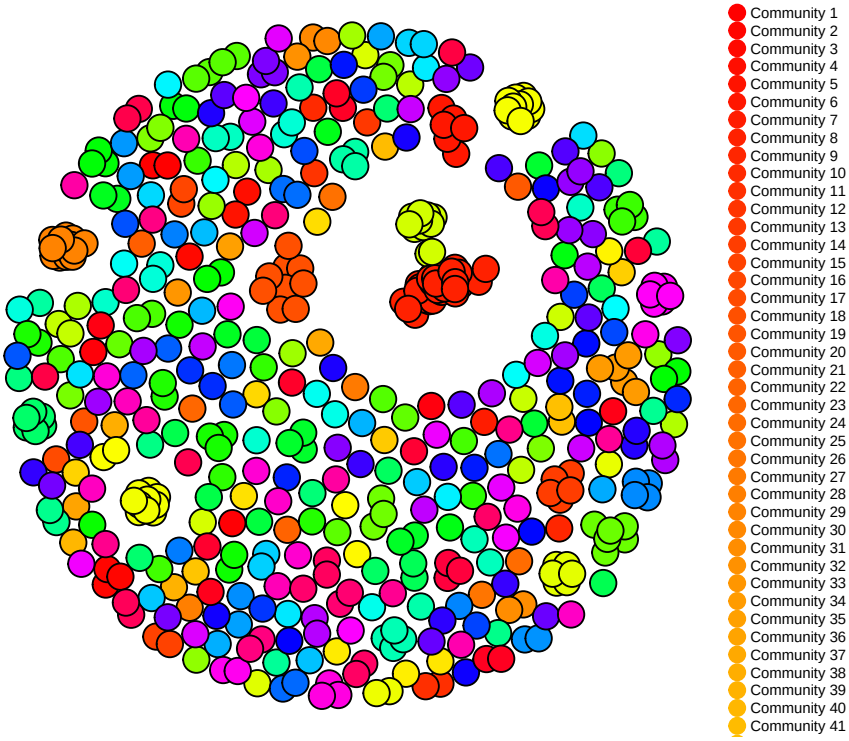
It is important to mention that this algorithm considers only one community per node, which in some cases may be not right.



3. **Infomap algorithm**.

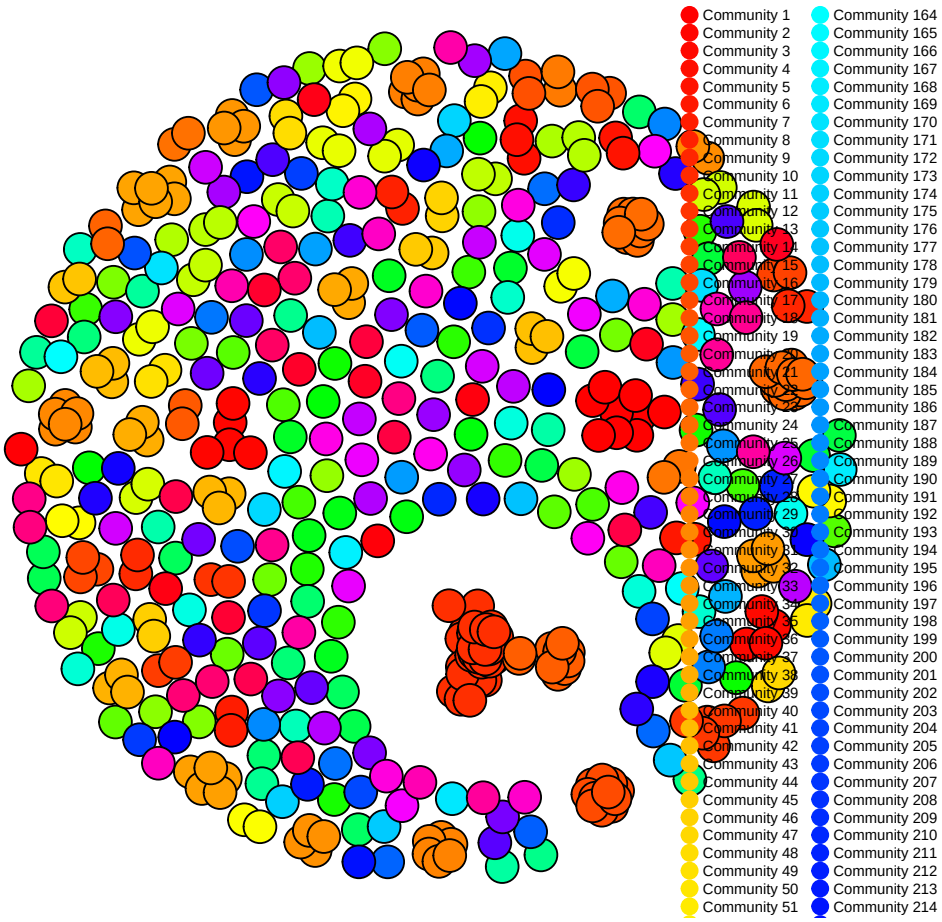
The Infomap method was first introduced by Rosvall and Bergstrom (2008). The procedure of the algorithm is in the core identical to the procedure of Blondel. The algorithm repeats the two described phases until an objective function is optimized. However, as an objective function to be optimized, Infomap does not use modularity but the so-called map equation.

The map equation exploits the duality between finding cluster structures in networks and minimizing the description length of the motion of a so-called random walk (Bohlin 2014). This random walker randomly moves from object to object in the network. The more the connection of an object is weighted, the more likely the random walker will use that connection to reach the next object. The goal is to form clusters in which the random walker stays as long as possible, i.e., the weights of the connections within the cluster should take on greater values than the weights of the connections between objects of different clusters. The map equation code structure is designed to compress the descriptive length of the random walk when the random walker lasts for extended periods of time in certain regions of the network.



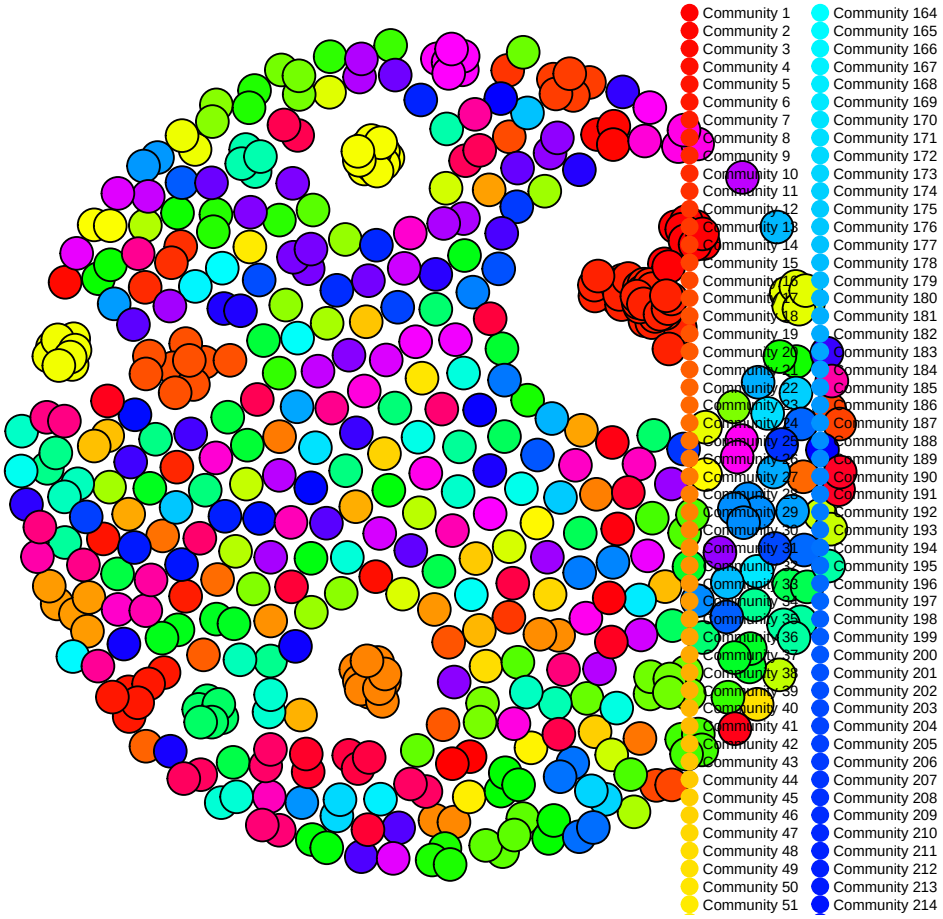
4. **Greedy modularity algorithm**.

The greedy modularity optimization is another method which helps us detect communities by iteratively including and removing nodes to maximize the modularity.



5. **Leiden algorithm**.

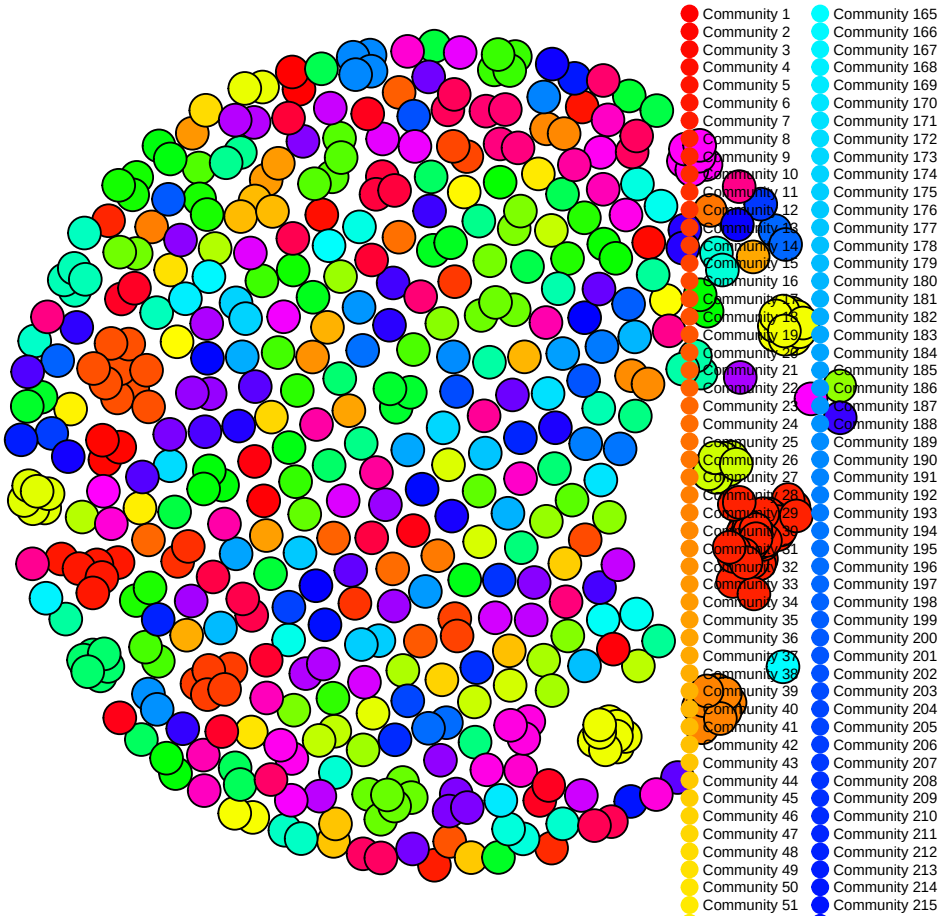
The Leiden algorithm is similar to the Louvain algorithm, cluster_louvain, but it is faster and yields higher quality solutions. It can optimize both modularity and the Constant Potts Model, which does not suffer from the resolution-limit



6. **Community detection based on label propagation**.

This algorithm implements the label propagation-based community detection algorithm described by Raghavan, Albert and Kumara. This version extends the original method by the ability to take edge weights into consideration and also by allowing some labels to be fixed.

Weights are taken into account as follows: when the new label of node i is determined, the algorithm iterates over all edges incident on node i and calculate the total weight of edges leading to other nodes with label 0, 1, 2, . . . , k - 1 (where k is the number of possible labels). The new label of node i will then be the label whose edges (among the ones incident on node i) have the highest total weight.



Network analysis can complement phylogenetic analysis by providing insights into the potential transmission dynamics and relationships that are not strictly hierarchical and might involve recombination events or horizontal gene transfer.

Jaccard similarity coefficient: Measures the similarity between sample sets, defined as the size of the intersection divided by the size of the union of the sample sets. Primarily used for presence-absence data.

Sørensen-Dice coefficient (Dice similarity): The Dice coefficient is a measure of the similarity between two sets, A and B. The coefficient ranges from 0 to 1, where 1 indicates that the two sets are identical, and 0 indicates that the two sets have no overlap. The Dice coefficient has several advantages over other similarity metrics. It is particularly useful for imbalanced datasets, where one set may be much larger than the other.

A high Dice coefficient value indicates a high level of similarity between the predicted and ground truth masks, meaning that the segmentation model or algorithm is performing well. Conversely, a low Dice coefficient value indicates poor segmentation performance.

Fst (Fixation Index): Measures genetic differentiation among populations based on genetic polymorphism data. It ranges from 0 (no differentiation) to 1 (complete differentiation). Fst is widely used in population genetics to assess genetic structure.

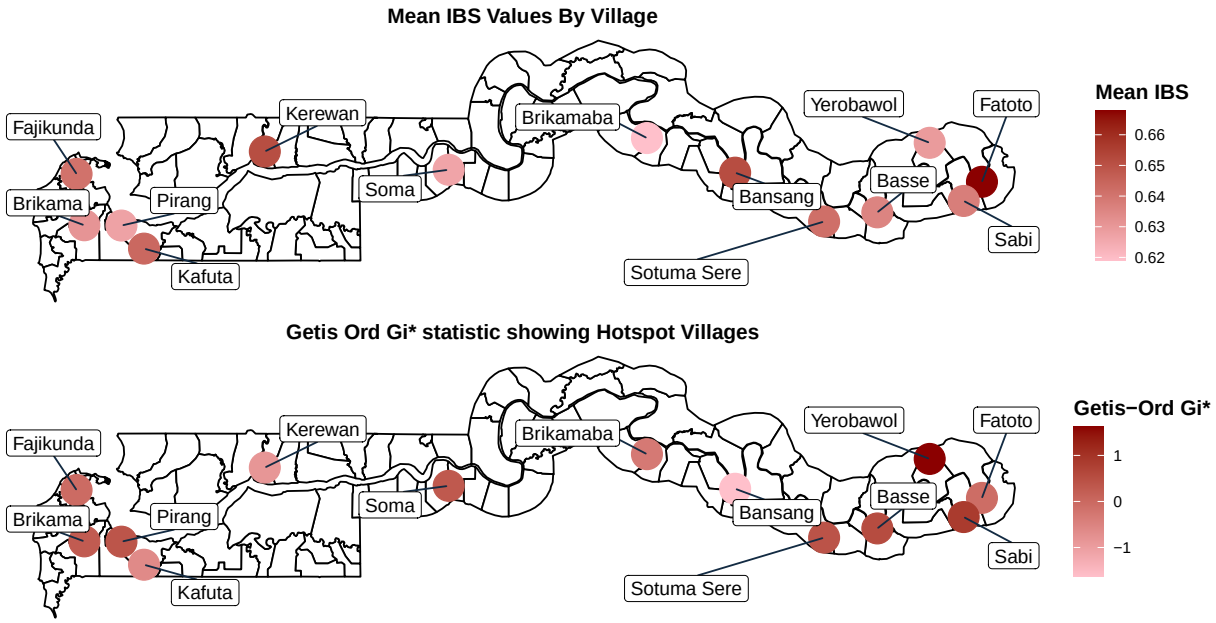
Nei's genetic distance (Nei 1972): A measure of the genetic distance between populations, taking into account the genetic identity and allelic frequencies. Used to infer evolutionary relationships and population structure.

Hotspot identification

Spatial analysis to identify hotspots of high transmission rates or clusters of genetically similar infections involves using statistical methods and indices that can detect areas with significantly higher incidences of disease or genetic similarity than would be expected by chance. This analysis can be crucial for targeting interventions and understanding the spatial dynamics of infectious diseases.

1. **Getis-Ord Gi* (Hotspot Analysis)**

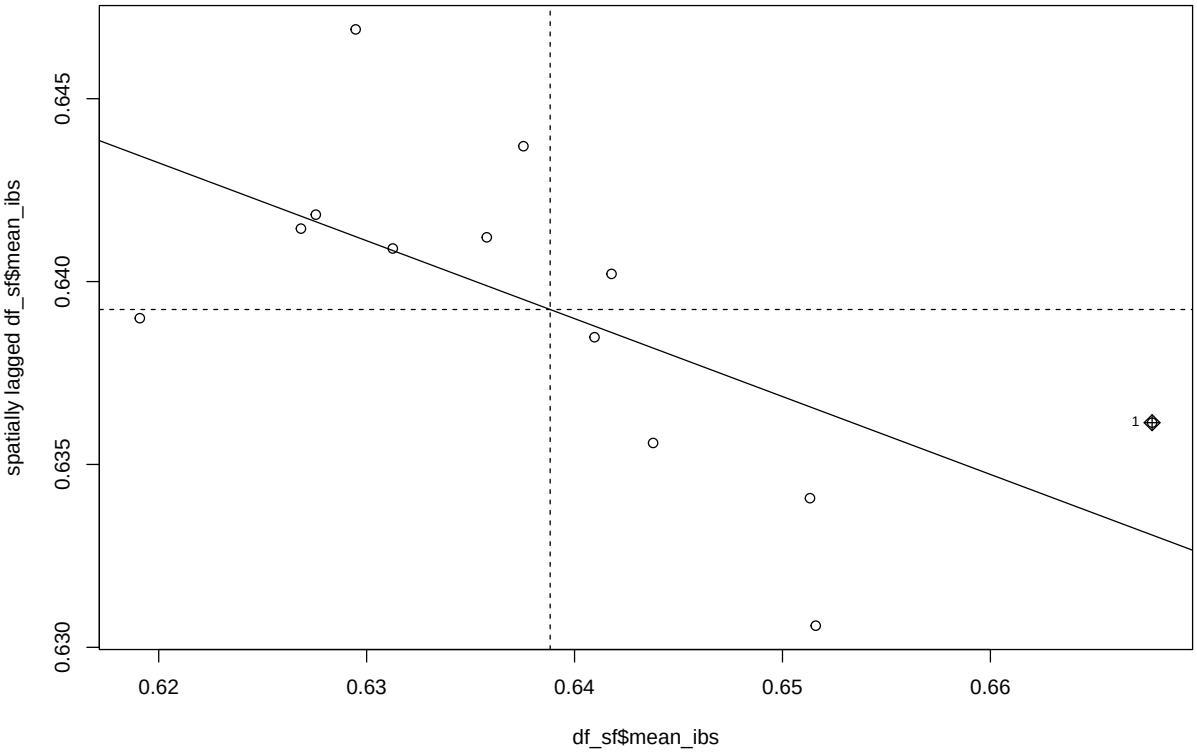
The Getis Ord Gi* statistic is what is used in "Hot spot analysis". It looks at how local spatial variance in a single variable, as mentioned by u/Drewddit. It was advanced by Artur Getis and Keith Ord. MGWR is built on Geographically weighted regression, which is built on OLS regression. GWR adds a spatial weighted matrix into a normal regression equation, as opposed to assuming that there is not local variance. Multiscale and Geographically and Temporally Weighted Regression both change that matrix to weigh variables differently in an analysis.

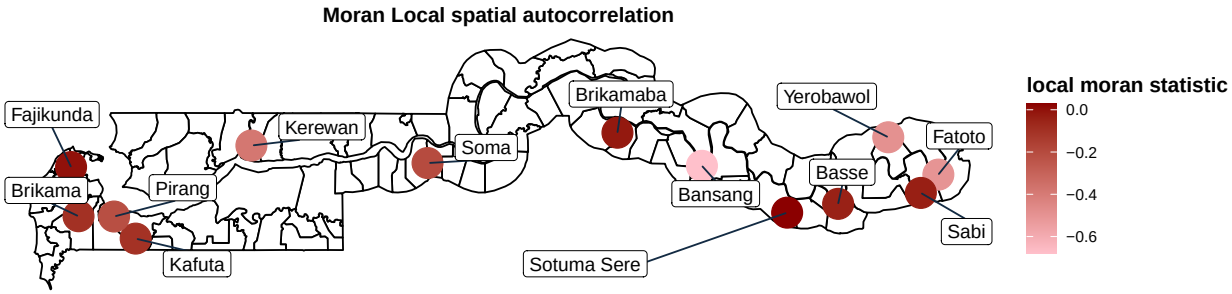


2. **Moran's I**

Moran's I is a measure of spatial autocorrelation that evaluates whether the pattern expressed is clustered, dispersed, or random across the study area. It compares the value of a variable at one location with the value of the same variable at nearby locations. High positive Moran's I values indicate clustering of similar values (e.g., high transmission rates or genetic similarity).

```
FALSE
FALSE   Moran I test under randomisation
FALSE
FALSE data:  df_sf$mean_ibs
FALSE weights: lw
FALSE
FALSE Moran I statistic standard deviate = -0.92682, p-value = 0.823
FALSE alternative hypothesis: greater
FALSE sample estimates:
FALSE Moran I statistic       Expectation          Variance
FALSE      -0.21298331       -0.08333333        0.01956844
```


spatially lagged df_sf$mean_ibs
0.630
0.635
0.640
0.645
0.62
0.63
0.64
0.65
0.66
df_sf$mean_ibs
1



3. **Geary's C**

Geary's C is another measure of spatial autocorrelation, similar to Moran's I, but it is more sensitive to differences between neighboring locations. It focuses on the dissimilarities between adjacent areas. Values significantly lower than 1 indicate spatial clustering of similar values.

4. **Kulldorff's Spatial Scan Statistic**

Used in SaTScan software, this method scans the study area for clusters by moving a circular window across the area and evaluating the likelihood of observing the number of cases within the window by chance. It's widely used for detecting disease outbreaks and can be adapted for identifying clusters of genetically similar infections.

Temporal dynamics analysis

To analyze how genetic variation changes over time and understand the spread and dynamics of transmission, several indices and methods can be employed. These analyses help in tracking the evolution of pathogens, identifying the emergence of new variants, and understanding how these changes influence transmission dynamics.

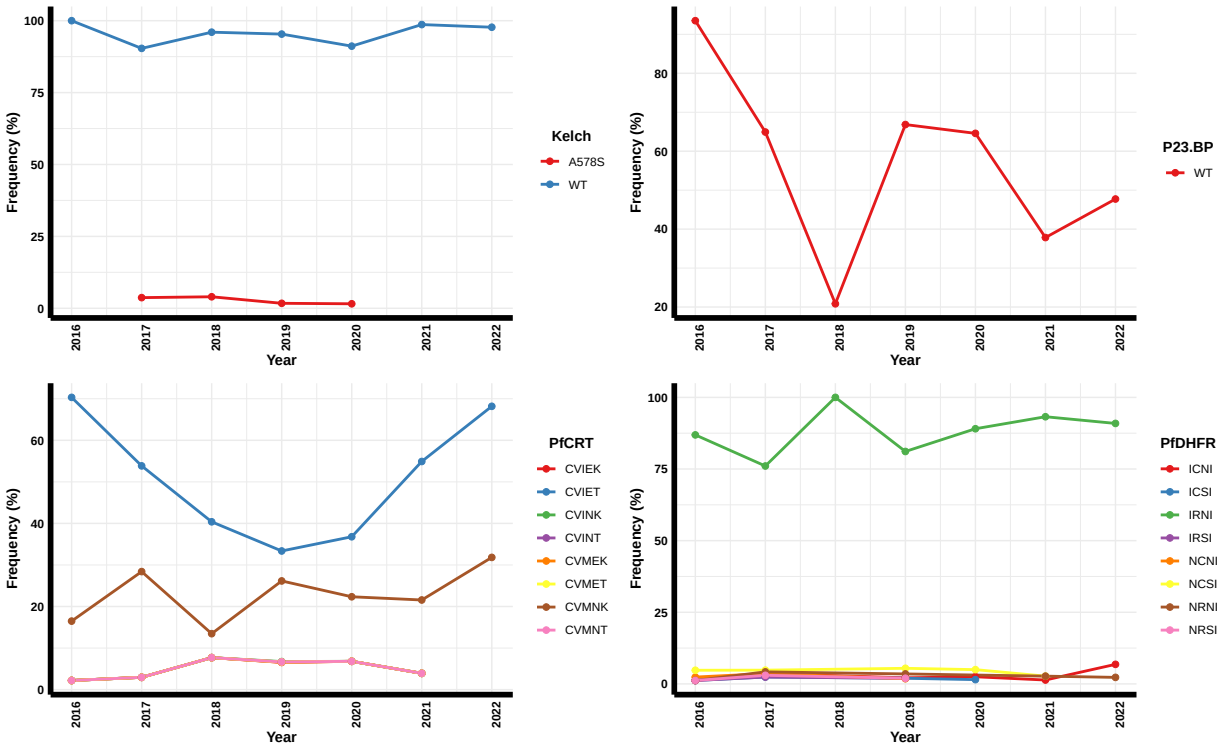
Temporal Fst (Fixation Index): Temporal Fst compares genetic differentiation between populations at different time points. It measures how genetic variance is distributed over time, providing insights into the population structure's temporal dynamics. A high Fst value indicates significant genetic differentiation, which could result from selection pressure, population bottlenecks, or founder effects.

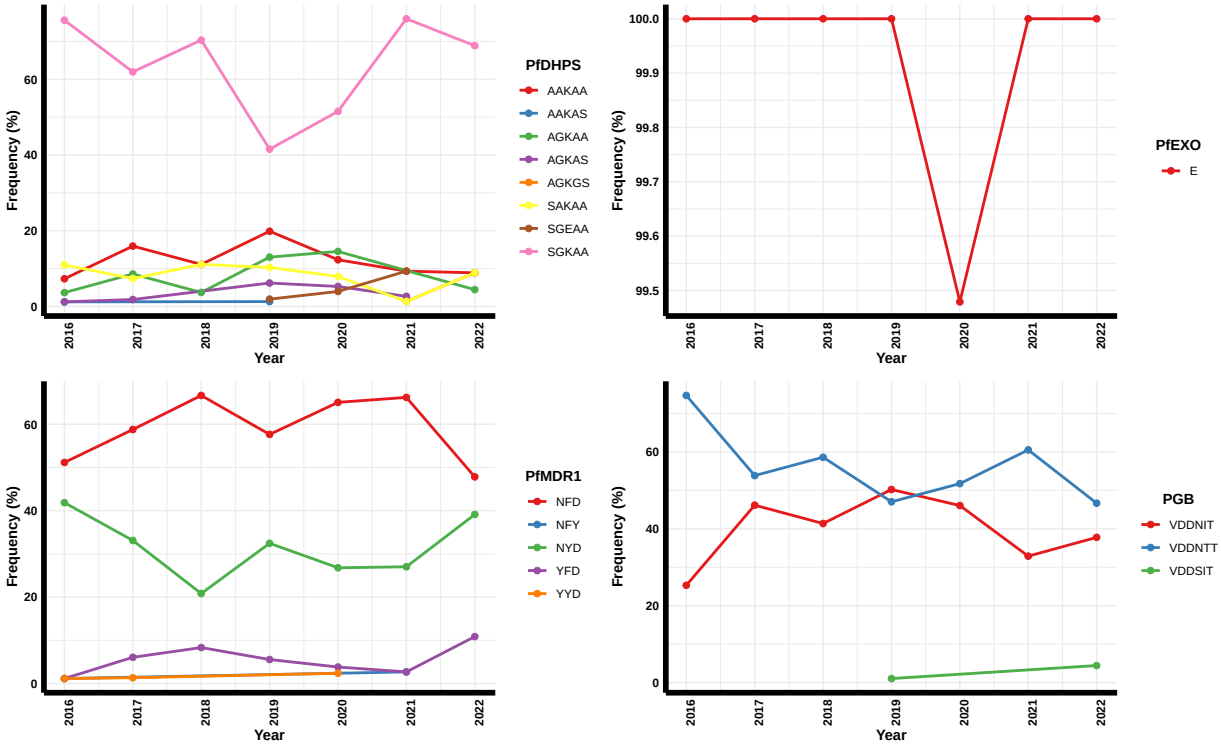
Haplotype diversity over time: Haplotype diversity measures the uniqueness of the genetic variants present in a population. Tracking changes in haplotype diversity over time can indicate the emergence or disappearance of strains, reflecting how genetic diversity is influenced by transmission dynamics, selection pressures, and population size changes.

Distribution of drug resistant haplotypes

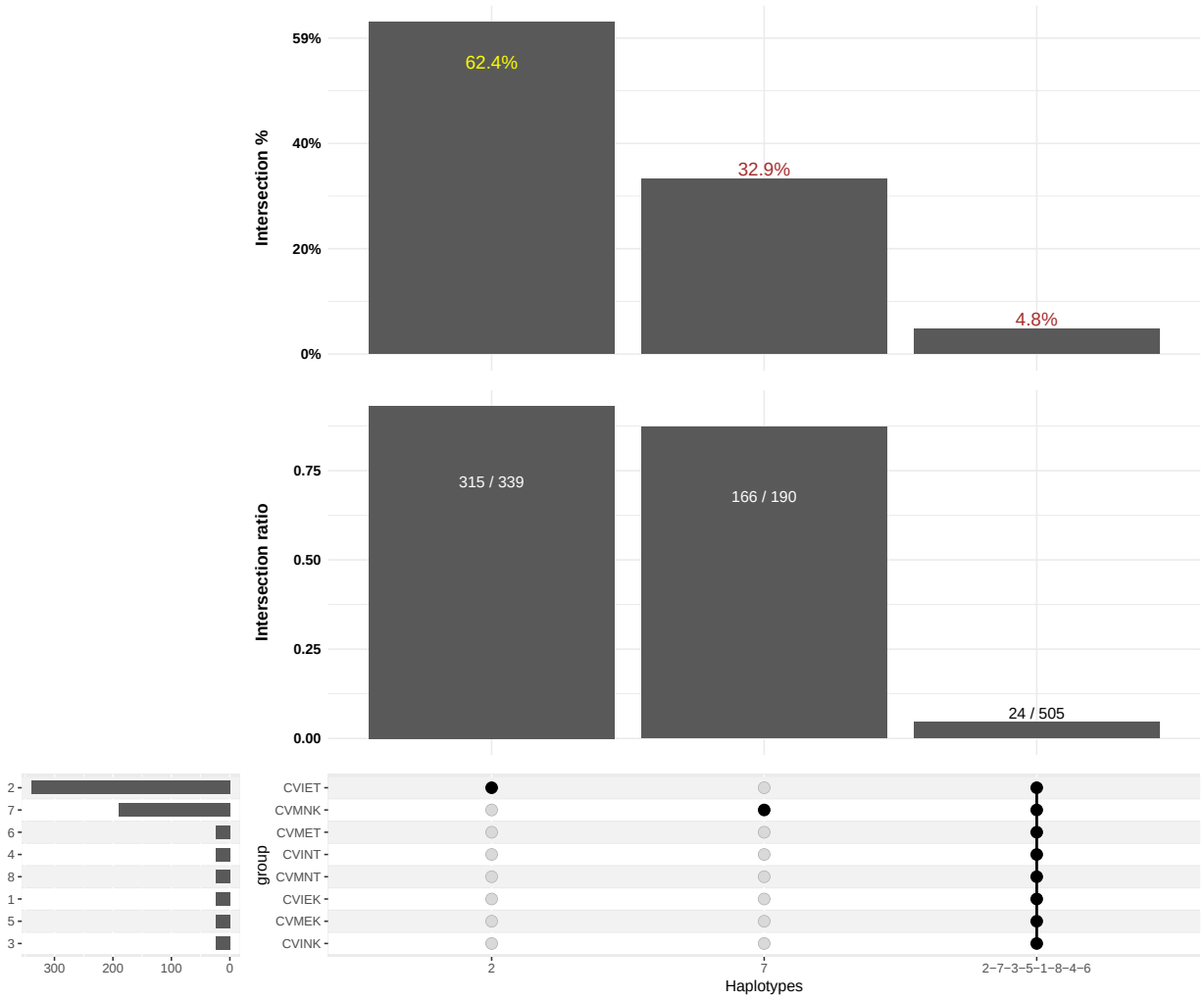
Since the SNPs are linked to drug resistance, we analyzed the prevalence of these resistance-associated SNPs in different locations and over time.

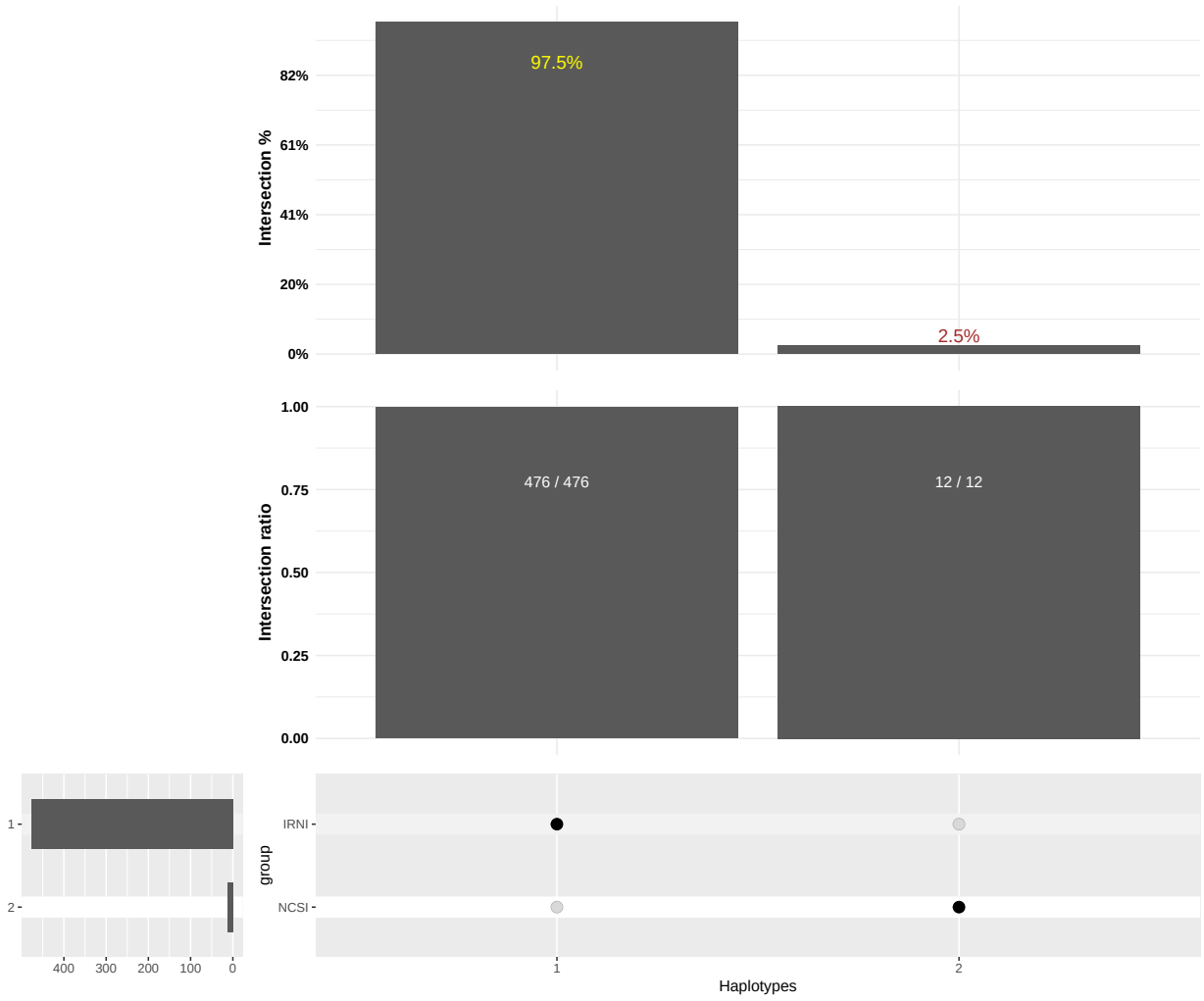
Estimating the prevalence of molecular markers associated with drug resistance involves identifying specific genetic mutations or patterns within a pathogen's genome that confer resistance to antimalarial drugs. This process is crucial in understanding the spread of resistance and informing treatment guidelines.

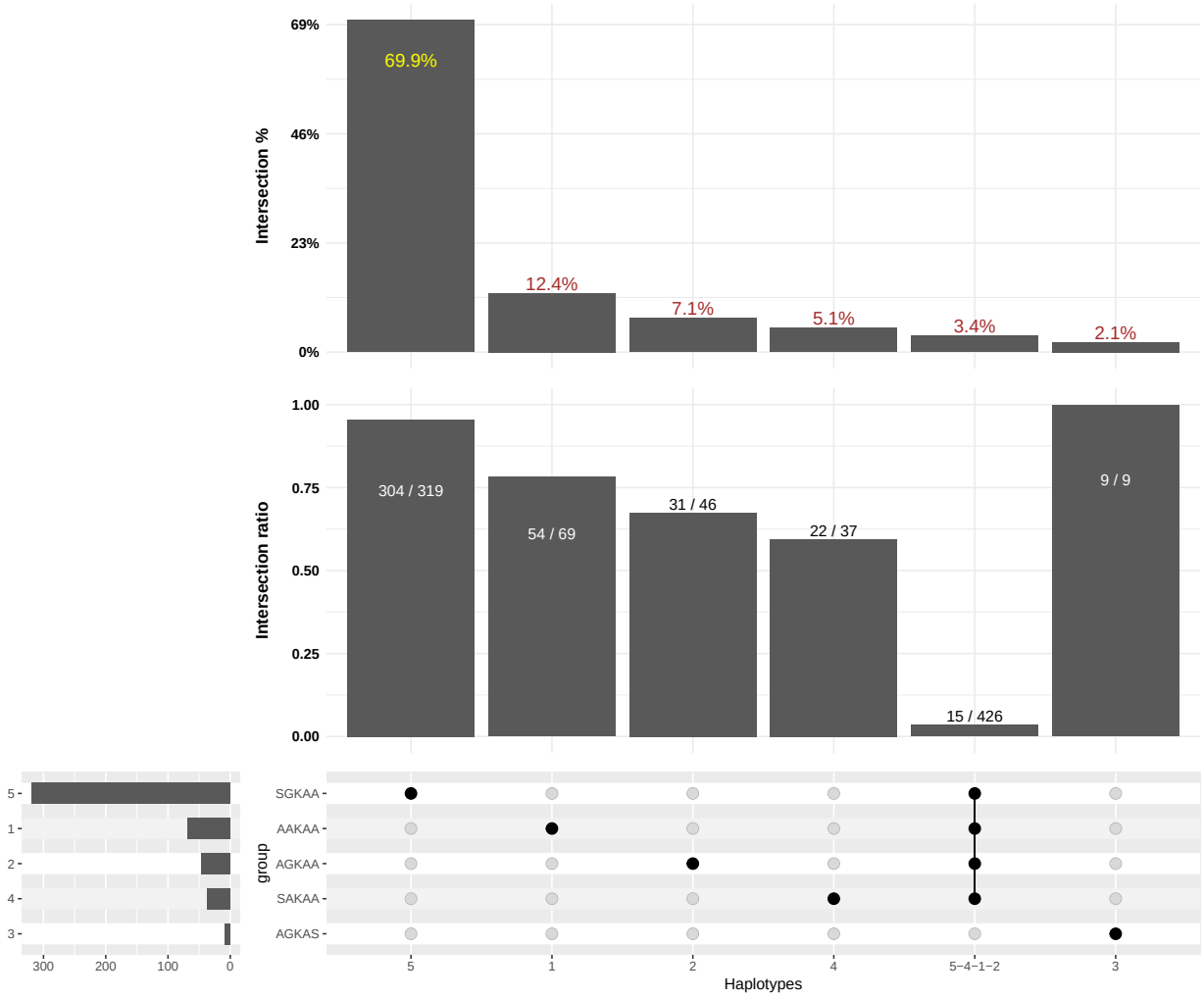


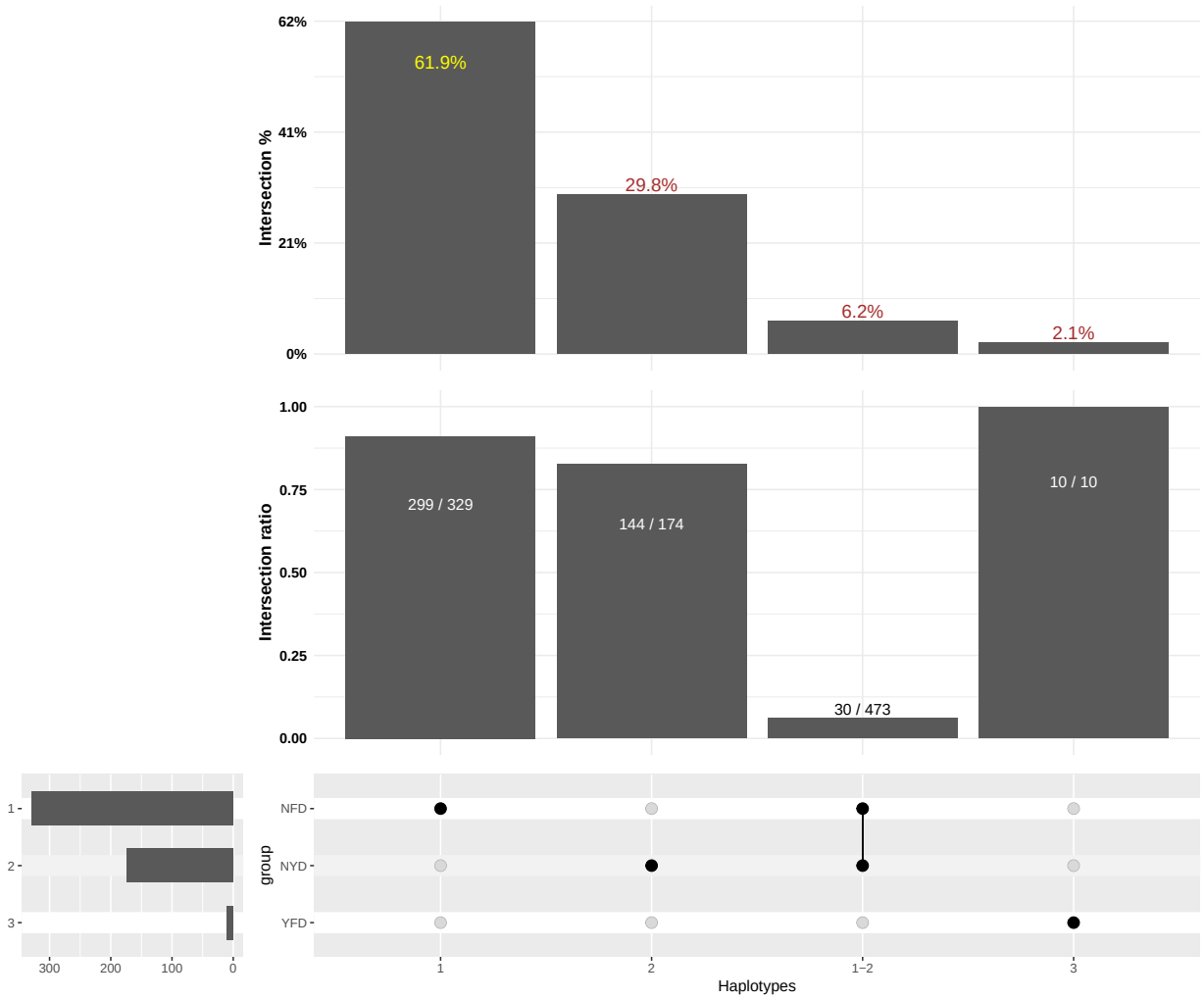


We also explored the joint distribution of all combinations of haplotypes in all antimalarial resistance genes and particularly in *PfCRT*, *PfDHFR*, *PfDHPS* and *PfMDR1* respectively. Raw combinations of mutations were visualized using the **UpSetR** package in R.


Intersection %
59%
40%
20%
0%
62.4%
32.9%
4.8%
Intersection ratio
0.75
0.50
0.25
0.00
315 / 339
166 / 190
24 / 505
2
7
6
4
8
1
5
3
300
200
100
0
group
CVIET
CVMNK
CVMET
CVINT
CVMNT
CVIEK
CVMEK
CVINK
2
7
2−7−3−5−1−8−4−6
Haplotypes


97.5%
2.5%
82%
61%
41%
20%
0%
Intersection %
1.00
0.75
0.50
0.25
0.00
Intersection ratio
476 / 476
12 / 12
1
2
400
300
200
100
0
IRNI
NCSI
group
1
2
Haplotypes


Intersection %
69%
46%
23%
0%
69.9%
12.4%
7.1%
5.1%
3.4%
2.1%
Intersection ratio
1.00
0.75
0.50
0.25
0.00
304 / 319
54 / 69
31 / 46
22 / 37
15 / 426
9 / 9
group
SGKAA
AAKAA
AGKAA
SAKAA
AGKAS
5
1
2
4
3
300
200
100
0
5
1
2
4
5−4−1−2
3
Haplotypes


62%
41%
21%
0%
Intersection %
61.9%
29.8%
6.2%
2.1%
1.00
0.75
0.50
0.25
0.00
Intersection ratio
299 / 329
144 / 174
30 / 473
10 / 10
1
2
3
300
200
100
0
group
NFD
NYD
YFD
1
2
1−2
3
Haplotypes