Introduction

The term single nucleotide polymorphism (SNP) refers to a variation of a single base pair in a complementary DNA double strand. SNPs can be inherited and heritable genetic variants. SNPs represent the largest category of genetic variation in the human genome (Shen et al., 1999, Ramírez-Bello and Jiménez-Morales, 2017), occurring at a frequency of roughly one every 1000 base pairs (Shastry, 2009). SNPs are typically found in clusters which are unevenly distributed along the genome (Koboldt et al., 2006) and can be found in non-coding sequences (outside of genes, as well as in introns), or in exons that are transcribed into proteins. In the latter case, a distinction is made between so-called "silent" or "synonymous" SNPs, which change the nucleotide sequence but do not change the amino acid sequence derived from it due to the degeneracy of the genetic code, and "non-synonymous" SNPs that lead to an amino acid change at the codon in question. Even if a SNP is in a non-coding genomic region, it can influence gene transcription (regulatory SNPs) if, for example, it is in DNA segments to which transcription factors (enhancers, silencers) or RNA polymerases bind (promoters). SNPs in an intron can also result in a "cryptic" splice site, for example.

In the past, little attention was paid to SNPs, especially when they were located in non-coding gene segments. They were thought to be largely meaningless variants. Through large-scale genome-wide association studies, it is now known that some SNPs influence the risk of certain diseases in complex ways, many of which are not understood. This applies, for example, to individual non-coding SNPs in the human genes IKZF1, ARID5B and CEBPE and the associated risk of developing acute lymphoblastic leukemia (Papaemmanuil et al., 2009).

Explicitly, SNPs in the NOD2/CARD15 gene are associated with a higher incidence of Crohn's disease, an inflammatory bowel disease (Hugo et al., 2001, Ogura et al., 2001). Considerably, other SNPs can alter the effectiveness of medical treatment. Certain non-coding single nucleotide polymorphisms in the gene IL28B effect the efficacy of treatment of hepatitis C with pegylated interferon-alpha (Thomas et al., 2009).

Moreover, alleles characterized by single nucleotide substitutions need not have a negative effect on the individual. Some have no discernible consequences for the phenotype, while other variants are even beneficial to the organism. For example, Europeans owe lactose tolerance to a SNP in the intron of the gene MCM6, which is located 5' from LCT (lactase)(Bersaglieri and Sabeti 2004).

Furthermore, it has been hypothesized that variations in the human gene FOXO3 are responsible improving certain markers of longevity (Flachsbart et al., 2017). Two identified single nucleotide variants have been shown to be distinct enhancers of human longevity, associated with increased synthesis of FOXO3 mRNA in various tissues (Flachsbart et al., 2017).

In recent years, research on SNPs has advanced rapidly, requiring further applications in which they can be analysed more intensively. SNPip was created as a web application to support this progress and increase the availability of SNP summary statistics for human chromosome 22.

Methodology

Aim

The **SNPip** webserver exploits a large database containing SNP information for human chromosome 22. The main functions of SNPip involve searching a MySQL database by chromosomal location, gene name or SNP name for a subset of 5 subpopulations. The search query outputs basic statistics for allele frequency and genotype frequency for every subpopulation. Additionally, if multiple SNPs are selected it contains summary statistics for genetic diversity, haplotype diversity and Tajimas D. If multiple populations are selected it contains statistics of pairwise genetic distance and delta Tajimas-D.

Search by:

- Gene annotation
- SNP annotation
- Reference and Alternate allele

Output:

- Minor allele frequency
- Genotype frequency
- Shannon Diversity
- Haplotype Diversity
- Tajimas-D
- Hudson_FsT

Data

SNPip uses 1000 Genome Project Phase 3 variant data from ENSEMBLs ftp server (http://ftp.ensembl.org/pub/grch37/current/data_files/homo_sapiens/GRCh38/) (The 1000 Genomes Project Consortium 2015, Howe et al., 2021). This data originally comes from the 1000 Genomes Phase3 pipeline remapped from GRCh37 using dbSNP rsIDs remapped to GRCh38. Furthermore, it contains the super-populations of European, East Asian, South East Asian, American and African from 1000 Genomes browser. Super populations are defined by using the 1000G panel file. The database contains biallelic SNPs and only phased genotype data derived from preprocessing the phase 3 genotype data. SNPip contains only single nucleotide variants (SNPs). Indels and larger structural variants have been removed and only contains high confidence calls for ALT alleles and positions that have passed all filters.

SNPip contains SNPs located on chromosomes 22 (February 2022).

All of the 1000G variants have been preprocessed with *bcftools* before creating the database using the following criteria:

1. Remove SNPs with duplicate rsIDs.
2. Remove Indels, and multiallelic sites
3. Add gene annotation and rsIDs
4. Remove SNPs with duplicate chromosomal coordinates

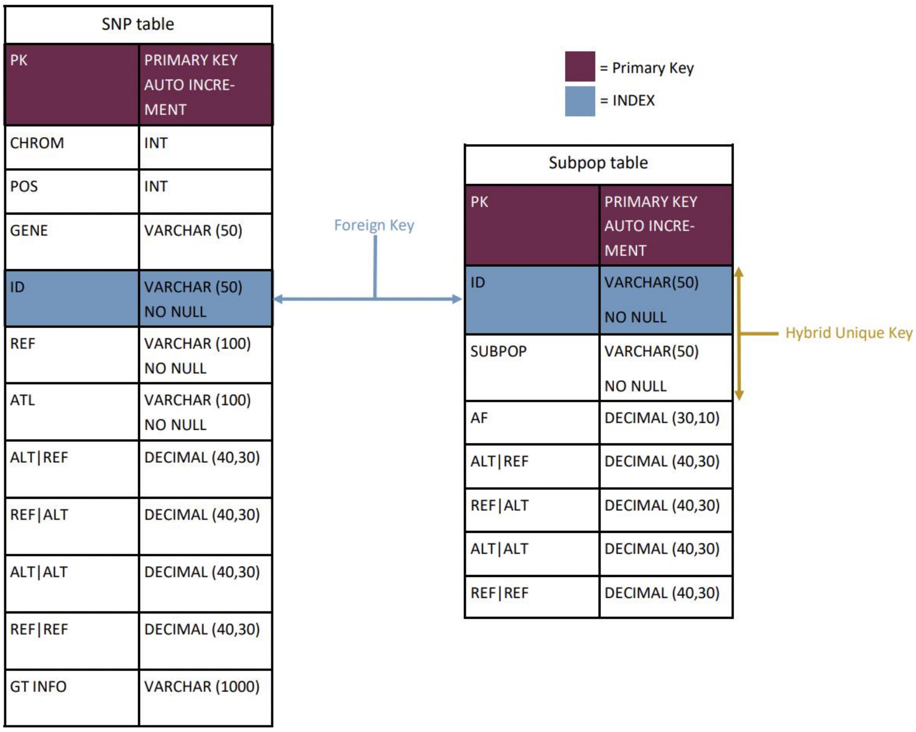
SNPip Super-Population Definitions

SNPip subpopulation definitions are derived from subsets of the following 1000G cohorts:

- **European (EUR)**
    - British in England and Scotland (GBR)
    - 91 Samples
- **East Asian (EAS)**
    - Han Chinese in Bejing, China (CHB)
    - 103 Samples
- **Africa (AFR)**
    - Esan in Nigeria (ESN)
    - 99 Samples
- **South East Asian (SAS)**
    - Bengali in Bangladesh (BEB)

  - 86 Samples
- **American Ancestry (AMR)**
  - Peruvian in Lima, Peru (PEL)
  - 85 Samples

SNPip Database Summary



SNP table	
PK	PRIMARY KEY AUTO INCRE-MENT
CHROM	INT
POS	INT
GENE	VARCHAR (50)
ID	VARCHAR (50) NO NULL
REF	VARCHAR (100) NO NULL
ATL	VARCHAR (100) NO NULL
ALT|REF	DECIMAL (40,30)
REF|ALT	DECIMAL (40,30)
ALT|ALT	DECIMAL (40,30)
REF|REF	DECIMAL (40,30)
GT INFO	VARCHAR (1000)

Subpop table	
PK	PRIMARY KEY AUTO INCRE-MENT
ID	VARCHAR(50) NO NULL
SUBPOP	VARCHAR(50) NO NULL
AF	DECIMAL (30,10)
ALT|REF	DECIMAL (40,30)
REF|ALT	DECIMAL (40,30)
ALT|ALT	DECIMAL (40,30)
REF|REF	DECIMAL (40,30)

The Database schema used in the SNPip application.

SNPip uses a relational database hosted on the google cloud sever, within a SQL instance, naturally allowing for remote access to the SNP data stored within it. The SNPip search app can then connect to the database through its public IP address.

The data for these subpopulations are stored in two tables. The first table is an SNP focused table containing information on the SNPs chromosome, position on said chromosome, the SNP ID, its alternate and reference ID as well as some aggregate genotype stats based on all five genotypes. In order to reduce the time it takes to retrieve information from the database, proper indexing of each table in the database had to be achieved. In the SNP table this is done through the use of an auto incrementing primary key, which assigned each row a numerical value which increased with each row added to the table. The RsID of each SNP was also used as an index inside this table, allowing for it to be used as a foreign key linking both tables. The ID column of the SNP table has a one (SNP table) to many (Subpopulation Table) relationship.

The other table contains information on each of the subpopulations. It consists of the SNP ID followed by its subpopulation and then its subpopulation specific allele frequency and genotype stats. Each SNP in this table has 5 entries, one for each subpopulation: British in England and Scotland (GBR), Han Chinese in Bejing, China (CHB), Esan in Nigeria (ESN), Bengali in Bangladesh (BEB), Peruvian in Lima, Peru (PEL). Due to each SNP having multiple entries in the subpopulation table, maintenance of the data in this table can become slow and cumbersome. In order to circumvent this, a hybrid primary key was assigned to each entry in the table. This key consisted of each rows' SNP ID and the subpopulation that its statistics are related to. In doing so, better indexing of the table is achieved as the duplicate SNP IDs have in essence been removed.

Through the use of a relational database and purposeful parsing of relevant information with regards to our aims, SNPip was able to reduce the size of said database by approximately 30% (the initial chromosome 22 vcf file was 10gb as opposed to the 7gb size of the current database).

The table below summarises the number of SNPs in each Subpopulation population within the SnPip database.

Subpopulation	Number of SNPs in the Database
Bengali in Bangladesh (BEB)	1,053,837
British in England and Scotland (GBR)	1,053,837
Han Chinese in Bejing, China (CHB)	1,053,837
Esan in Nigeria (ESN)	1,053,837
Peruvian in Lima, Peru (PEL)	1,053,837

Methods

Data Preprocessing

Phase 3 data from the 1000 Genomes project was selected as it contains recent sequencing and variant call information. For further processing the. *vcf.gz* and the *.tbi* data was loaded and stored in the same folder.

Subsetting

The processing of the 1000 genome data was done with *bcftools, vcftools* and *tabix* (Heng et al., 2009, Danecek et al., 2011, Danecek et al., 2021). To subset the data into subpopulations of interest the 1000G panel file was used and each subpopulation was copied into a *.txt* file containing only samples from this subpopulation. Subsequently, five *.txt* files containing only the samples of the used subpopulations were created.

```
$ bcftools view –force samples <all_populations.vcf.gz> -S peru.txt >
output_PEL.vcf
```

Annotation

Furthermore, after subsetting the data the gene annotation was added. A gene file for Chromosome 22 containing all start and end positions from the UCSC Table Browser (Kent et al., 2002) was downloaded and converted into a *bed* file and then indexed by *tabix.*

```
$ bgzip genes_for_annotation.bed
$ tabix -p bed genes_for_annotation.bed.gz
```

After doing that, the following command inserted the gene annotation inside INFO column of the *.vcf* file

```
$ bcftools annotate \
    -a genes_for_annotation.bed.gz \
    -c CHROM,FROM,TO,GENE \
    -h <(echo '##INFO=<ID=GENE,Number=1,Type=String,Description="Gene name">') \
    output_GBR.vcf.gz \
    -o annotated_GBR.vcf
```

Since the used *VCF* file was already annotated with rsIDs from the 1000G, no more annotation was needed.

Filter for biallelic sites

In diploid systems multiallelic sites are often ignored because many modern methods struggle to include multi-allelic variants in an accurate manner and creating a strongly misleading association (Campbell et. al., 2016, Jiang et al., 2020). Therefore, we decided to not use them for further statistics. To delete multiallelic sites and exclude indels, the following command in *bcftools* was used.

```
$ bcftools view –max-alleles 2 –exclude-type indels annotated_GBR.vcf.gz
```

Calculate minor allele frequency

The allele frequency accounts for the relative frequency of an allele at a specified SNP. Since only biallelic data was used only the minor allele frequency (MAF) has been reported by using the following equation:

$$\frac{AC}{AN} = MAF$$

As described in the present formula (1) the total number of observed alleles (AC) between all samples was divided by the number of counts at s specific locus of the examined allele. Therefore, *bcftools* and *vcftools* has been used.

Option 1:

```
$ vcftools –vcf annotated_GBR.vcf.gz –freq --out freq_annotated_GBR
```

Option 2:

```
$ bcftools +fill-tags annotated_GBR.vcf.gz -o freq_annotated_GBR.vcf.gz
```

Decoding and Encoding Genotype Data

In order to get the genotype information inside the database, SNPip uses an encoder that converts the phased genotypes from 0 and 1 format into a string for each SNP. Explicitly, the genotype data was divided into the four possible genotypes of 0|0 (homozygous$_{ref}$), 0|1 (heterozygous$_{ref}$), 1|0 (heterozygous$_{alt}$) and 1|1 (homozygous$_{alt}$) and then encoded using the following algorithm:

1|0 = “a”

1|1 = “b”

0|1 = “c”

0|0 = “d”

Furthermore, to split the data into the different subpopulations that have been chosen SNPip created a nested_dictionary that was structure the following:

nested_dictionary= {'GBR' :{ 'a' : 'e' , 'b' : 'f' , 'c' : 'g' , 'd' :'h' } ,
'PEL' : { 'a' : 'i' , 'b' : 'j' , 'c' : 'k' , 'd' : 'l' } ,
'ESN' : { 'a' : 'm' , 'b' : 'n' , 'c' : 'o' , 'd' :'p' } ,
'BEB' : { 'a' : 'q' , 'b' : 'r' , 'c' : 's' , 'd' : 't' } ,
'CHB' : { 'a' : 'u' , 'b' : 'v' , 'c' : 'w' , 'd' : 'x' }}

In conclusion the resulting string of each SNPs genotype data was constructed out of letters from “e-z” and the index of the string informed the decoder of the specific sample the genotype referred to. In the next step when searching on the website the string gets decoded to the actual genotype information data by creating a 3D list from the nested_dictionary.

Statistics

For the calculation of Haplotype Diversity, Tajima-D and Hudson-FST the python package *scikit allel* has been used. Shannon Diversity as an index of genetic diversity has been computed separately. Genotype data from the database was decoded into a 3D-list of genotypes for every sample and every SNP to be used with *scikit allel.*

The list of genotype data is encoded in the following 3D list format from the database:

```
pre_arr [[[0,  1],[0,  0]],
          [[1,  0],[0,  0]],
          [[0,  0],[1,  1]]]
```

Haplotype Diversity

Haplotype Diversity represents the probability of two randomly sampled alleles being different. Here, a haplotype array for the region of interest gets created by using the genotype data stored in the database. Furthermore, the function *allel.haplotype_diversity* generates an estimate of haplotype diversity.

This is calculated by the following formula:

```
  # check inputs
g = GenotypeArray(pre_arr)

  # creates haplotype array
h = g.to_haplotype()

  # number of haplotypes
  n = h.n_haplotypes

  # compute haplotype frequencies
  f = h.distinct_frequencies()

  # estimate haplotype diversity
  Hd = (1 – np.sum(f**2) * n / (n-1)
```

→ $Hd = (\frac{1-\Sigma p_i^2 * n}{(n-1)})$

The estimate of haplotype diversity is calculated over the whole region selected by the user. Furthermore, for visualization SNPip computes the data in bins over the region to enable precise analysis. The values obtained from this function are in between 0 and 1. Where “0” accounts for a locus consisting of mainly one haplotype and “1” refers to a very diverse region with several haplotypes. The following graph describes a statistical output for haplotype diversity delivered by SNPip. By clicking the subpopulations on the right, it is possible to highlight only certain subpopulations. Furthermore, the haplotype diversity is calculated in small bins over the whole locus making it possible to analyze smaller partitions. Moreover, an info box opens by hovering over a specified bar showing additional information including subpopulation selected, region on chromosome and the value for haplotype diversity inside the selected range. To make comparison easier it is also possible to slide through the whole region and zoom into desired ranges. Additionally, a screenshot can be taken by clicking on the camera on the top right (Figure 1).

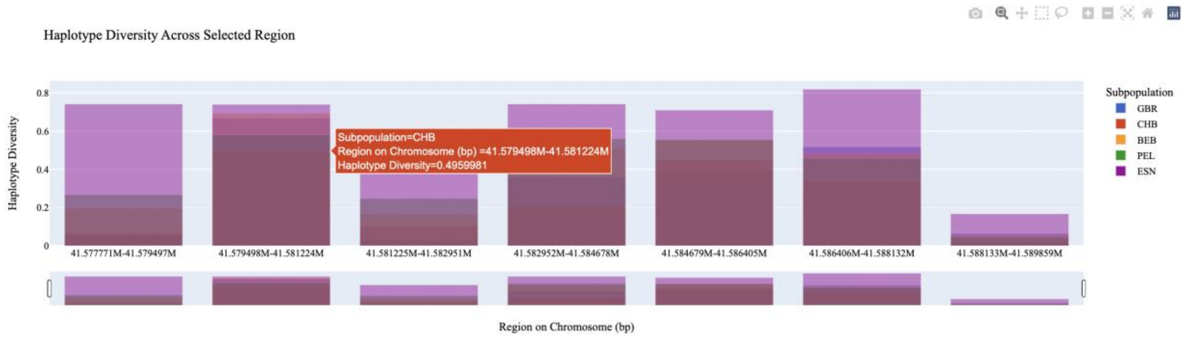
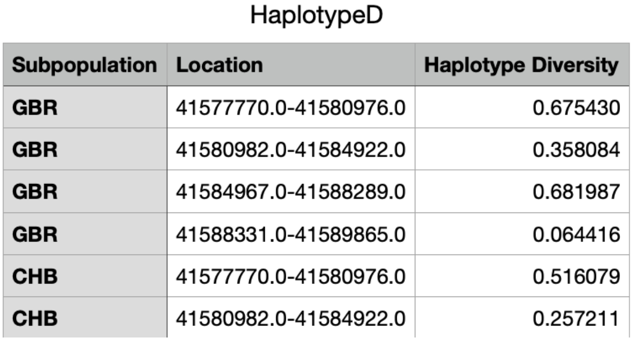


Figure 1 Visualization of Haplotype Diversity for PMM1. Showing distinct patterns of homozygosity between the selected populations within specified regions. Higher values near to 1 consider a higher probability of picking two different alleles by random choice. Values near 0 consider a very homozygous region between the samples picked.

For further analysis SNPip provides a downloadable result table next to the graph (Figure 2).



HaplotypeD

Subpopulation	Location	Haplotype Diversity
GBR	41577770.0-41580976.0	0.675430
GBR	41580982.0-41584922.0	0.358084
GBR	41584967.0-41588289.0	0.681987
GBR	41588331.0-41589865.0	0.064416
CHB	41577770.0-41580976.0	0.516079
CHB	41580982.0-41584922.0	0.257211

Figure 2 Downloadable result table for Haplotype Diversity within and between populations.

Tajima D

Natural populations can sustain a certain amount of variation which is dependent on the population size and the amount of variation occuring. Consequently, by generating a value to estimate how much variation a population should be able to sustain helps to understand and determine this problem. Being able to infer deviation from that expectation is useful to understand whether a particular kind of selection is going on. Consequently, Tajimas D is a test against neutrality to see if the SNPs are subject to selection within a population and computes an average number of pairwise differences in between base pairs of various population samples and compares it to the total number of variant sites in that sample. Subsequently, it compares two sequences on either evolving randomly versus one evolving in a non-random manner. Explicitly, mutations that occur randomly are unbiased (neutral) and the one's undergoing selection are biased. The test requires at least three homologous DNA sequences to identify these which do not go with the neutral theory and consequently have no effects on fitness and survival (Tajima, 1989).

The main input of *scikit allel* looks like this:

```
    # check inputs
g = GenotypeArray(pre_arr)
```

```
# count alleles
ac = g.count_alleles()

# calculate tajimas
allel.tajima_d(ac)
```

Having a closer look into the equation that follows *allel.tajima_d* gets us insights into the number of Segregating sites (S) and number of samples (n).

```
      # count segregating variants
 S = ac.count_segregating()
 If S < min_sites(3):
           return np.nan

# assume number of chromosomes samples is constant for all variants
 n = ac.sum(axis=1).max()
```

Additionaly a_1, pi (π) and theta (θ) are main factors to evaluate for Tajimas D. Here, π describes the mean pairwise difference within a given sequence meanwhile θ stands for the expectation of π. Theta in a purely neutral population with no special mutational events concludes a measurement for mutational drift.

```
      # (n-1)th harmonic number
 a1 = np.sum(1 / np.arrange(1, n)

# calculate Waterson theta
 theta_hat_w_abs = S / a1

      # calculate mean pairwise distance
 mpd = mean_pairwise_difference(ac, fill = 0)

# calculate theta_hat pi (sum difference over variants)
 theta_hat_pi_abs = np.sum(mpd)

    # calculate theta_hat pi (sum difference over variants)
 d= theta_hat_pi_abs - theta_hat_w_abs
```

Given certain circumstances π may not be right at that expectation (θ) and this measure is named Tajimas D. This variable is the normalized version of:

$$D = \pi - \theta$$

Where, π describes the observed mean pairwise difference and θ occurs as the expected mean pairwise difference within sequences inside a population. Tajimas D in a more accurate manner is calculated by the following:

$$a_1 = \sum_{i=1}^{n-1} 1/i$$

$$a_2 = \sum_{i=1}^{n-1} 1/i^2$$

$$b_1 = \frac{n+1}{3(n-1}$$

$$b_2 = \frac{2(n^2+n+3}{9n(n-1)}$$

$$c_1 = b_1 - 1/a_1$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

$$e_1 = \frac{c_1}{a_1}$$

$$e_2 = \frac{c_2}{a_1}$$

Inferring the sampling variance (how different these statistics are from sample to sample) computes Tajimas D in that equation:

$$D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}}$$

In neutral population we would expect:

$$E[\pi] = \theta$$

$$E[S] = a_1\theta$$

Given certain circumstances π may not be right at that expectation (θ) and this measure is named Tajimas D. Tajimas D lies with 95% probability inside -2 and 2. All values outside this range should be used carefully and may need further analysis.

$$If: D < 0$$

A negative Tajimas D concludes a selection removing variation resulting in the conclusion of recent expansion of population size. This could be due to bottleneck-effects or selective sweeps.

$$If: D > 0$$

A positive Tajimas D assumes a maintaining selection and a resulting contraction of population size caused by balancing selection.

$$If: D = 0$$

A neutral Tajimas D concludes that the population is evolving under the neutral theory and as it is expected from the per mutation-drift equilibrium. There is no evidence of selection events.

If SNPip generates NaN values this may be caused because the minimum segregating sites of three were not inside the calculation.

Figure 3 shows visualization of Tajima D. The user will be able to analyze Tajima D in bins of 5% from the maximum length of the region specified. The range of bins can be seen on the x-axis. The legend on the right makes it possible to select and deselect a certain subpopulation to make differences more visible. By hovering over bars, it is possible to get an info box with additional information containing the selected subpopulation, region on the chromosome and the value of Tajimas D. In the following graph it is visible that the end region of *PMM1* shows no signs of selection events whereas the rest of the gene shows patterns of removing variation, especially for the subpopulations of Bangladesh and British citizens, with lower heterozygosity and an excess of rare alleles. This could be concluded to be part of a recent population expansion. Additionally, a sliding window is provided by SNPip to zoom into regions of interest and a results table for values calculated over the whole locus is downloadable next to the graph . In comparison to the Tajima D values generated for each bin it shows an overall negative Tajima D whereas the user can now in connection with the graph conclude where the strongest selective sweeps are happening for *PMM1*. (Figure 3, Figure 4)

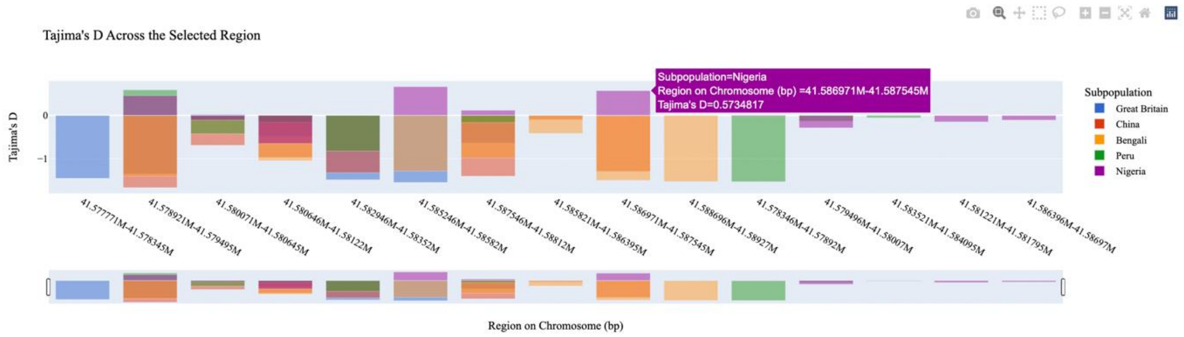
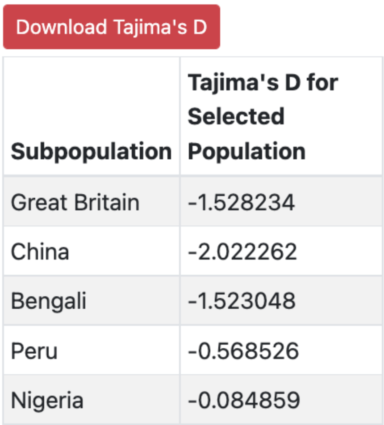


Figure 3 Visualization of Tajima D for PMM1. Showing distinct patterns of D between the selected populations within specified regions. An example explanation of the graph can be given by looking at Bengali in the regions of 41.578921M bp – 41.574945M bp and 41.586971M bp – 41.58972M bp where a recent selective sweep and lower average heterozygosity would be possible to conclude.



Subpopulation	Tajima's D for Selected Population
Great Britain	-1.528234
China	-2.022262
Bengali	-1.523048
Peru	-0.568526
Nigeria	-0.084859

Figure 4 Downloadable result table for values of Tajima D for every selected population. In comparison to the Tajima D values generated for each bin it shows an overall negative Tajima D whereas the user can now in connection with the graph conclude where the strongest selective sweeps are happening.

Hudson FST

The fixation index offered by Hudson et al., 2009 is a population genomics statistics used to show how much variation there is among different species or populations. It is looking at the proportion of genetic variation that is within population relative to that which is between populations:

$$FST = \frac{variation\ between\ populations - variation\ within\ populations}{variaton\ between\ populations}$$

$$FST = \frac{H_T - H_S}{H_T}$$

Inside the forwarded function H_T refers to the average heterozygosity between populations and H_S refers to the average heterozygosity within a population. Heterozygosity is a measure of genetic variation calculated by the sum of the squared allele frequencies:

$$H = 1 - \sum p_i^2$$

SNPip takes the estimator of Hudson which takes averages of FST according to the Weir & Cockerham (1984) definition since this is independent from sample sizes.

$$Hudson_{FST} = \frac{(FST_1 + FST_{2)}}{2}$$

Furthermore, SNPip computes an estimate of average FST within a region by just using the average of pairwise differences and does not make use of any topological structure of the gene as this should not change the results in a significant manner (Hudson et al., 1992, Bhatia, et al., 2013).

```
            # allele count and GenotypeArray creation
              ac1 = allel.GenotypeArray(pre_arr1).count_alleles()
              ac2 = allel.GenotypeArray(pre_arr2).count_alleles()

             # calculate tajimas
              allel.hudson_fst_d(ac1, ac2)

          # number of alleles
    an1 = np.sum(ac1, axis=1)
            an1 = np.sum(ac1, axis=1)

             # average siversity (heterozygosity) within each population
    within = (mean_pairwise_difference(ac1, an1) +
                               (mean_pairwise_difference(ac2, an2) / 2

   # divergence (heterozygosity) between each population
    Between = = (mean_pairwise_difference_between(ac1, ac2, an2,

             # define numerator and denominator for FST calculations
    num = between – within
    den = between

   # averaging fst over variants
    fst = np.sum(num)/np.sum(den)
```

The FST will most likely be a value between 0 and 1 where values near 0 will show less genetic distance and values near 1 show more genetic distance. At the border of 0.25 one can consider a significant deviation on the selected region between the populations selected is shown. FST values will be NaN if there are no heterozygous sites in the specified regions. Values that are negative can be considered 0 and no deviation of variance could be observed.

The visualization of genetic distances shows two distinct lines. The red line contributes to a significant difference between populations at 0.25% meanwhile the grey line refers to an upper threshold of FST between continental populations. Moreover, it is visible which populations differ the most from each other which for the region of 41.57777M bp – 41.578869M bp on *PMM1* (Chromosome 22) would be ESN vs. GBR (0.155), ESN vs. CHB (0.15) and ESN vs. BEB (0.13) with FSTs between 0.13 – 0.155. Subsequently ESN shows the most distinct patterns of SNPs inside this region in comparison to other populations. Consequently, SNPip makes it visible that the distances in between GBR vs CHB (~ 0), GBR vs. BEB (~ 0) and CHB vs BEB (~ 0) are near 0 showing no distinct patterns of variation inside this region (Figure 5). Additionally, the slider on the bottom allows to slide through different regions across *PMM1* and compare the results. The step size is calculated by 10% of the gene length. It is furthermore, possible to download these graphs at certain positions, when hovering over it by taking a automatically downloaded screenshot on the top right of the graph.

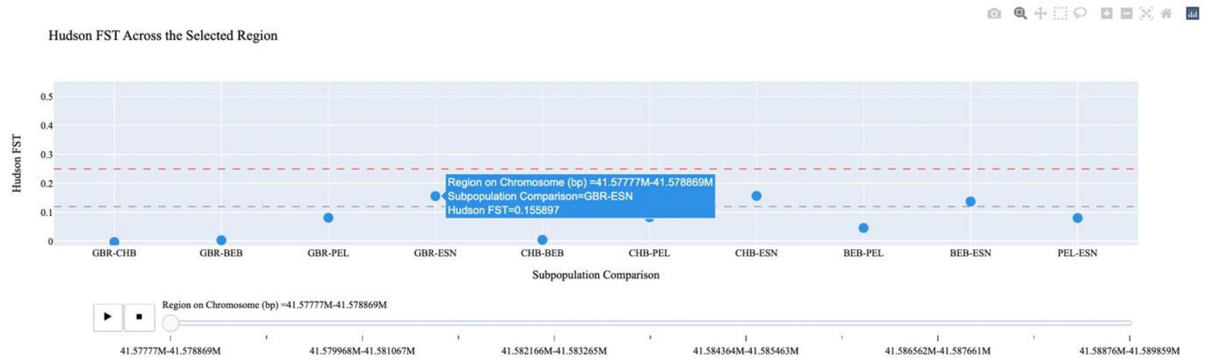
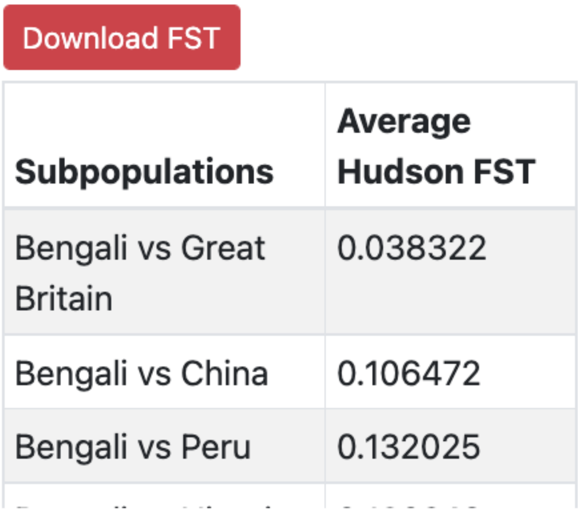


Figure 5: Hudson_FST of the gene PMM1 in the region between 41.57777M - 41578869M bp on Chromosome 22. The red line contributes to a significant difference between populations at (25%). The grey line refers to an upper threshold of FST between continental populations (12%). For the specified region the populations of GBR vs. ESN and CHB vs. ESN as well as BEB vs. ESN show to most genetic distance between each other.

Moreover, next to each graph a table of the average FST over the whole region is downloadable (Figure 6)



Subpopulations	Average Hudson FST
Bengali vs Great Britain	0.038322
Bengali vs China	0.106472
Bengali vs Peru	0.132025

Figure 6 Downloadable table next to the Hudson_FST visualization, where average FSTs over the whole selected regions are calculated.

Shannon’s Diversity Index

Shannon diversity index (SDI, H) is widely used in ecology and population genetic studies due to its adaptive properties to describe variation at multiple levels (M.K Konopiński, 2020). As opposed to measures based on heterozygosity or allele number, SDI weighs alleles in proportion to their population. Using allele frequency in SDI, SNpip calculates genetic diversity. The allele frequency for a SNP variant in each population is used to describe the genetic diversity between each population (A. Chao, 2015).

$H= -\sum[(p_i)*\ln((p_i)]$ $H= -\sum[pi*ln(pi]$

- p_i = allele frequency (AF) where the proportion of alternate alleles (AC, Allele Count) found divided by the total number of alleles (AN, Allele Number) found.
  - $p_i = \frac{AC}{AN}$
  - p_i= for each population (i = population 1, 2, 3)
- Minimum diversity (H_{min}) value is at 0 where there is no diversity, only the reference allele.
- Maximum diversity (H_{max}) is the log () of the total number of alleles

SDI calculated using a python algorithm with the SDI formula:

1. Select allele frequency for the SNP variant for each population.
2. For each population, multiply the af by the natural log of af.
3. Sum all the numbers from step 2.
4. Multiply the sum by -1 to get a positive value.

The visualization of Shannon diversity shows a line graph with peaks on certain positions where the selected populations differ the most on each other. By hovering over these peaks an info box containing the rsID, SDI values and positional information will open. The graph is also downloadable by clicking on the camera in the top right corner of the graph (Figure 7).

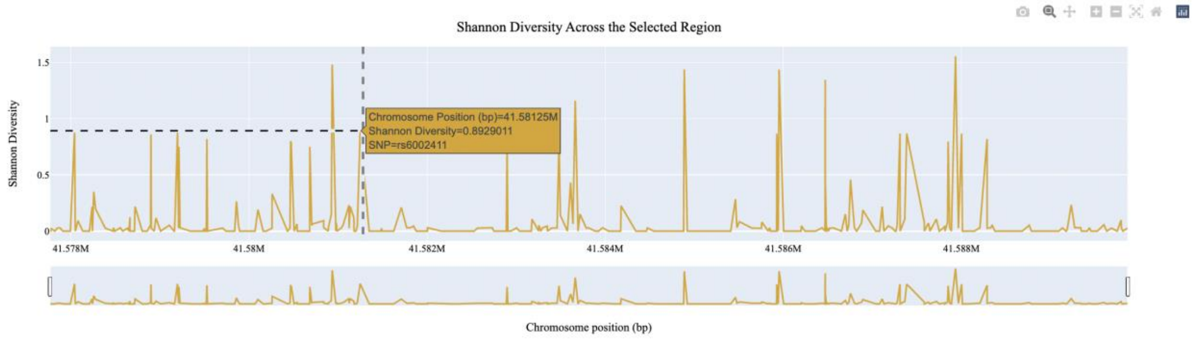
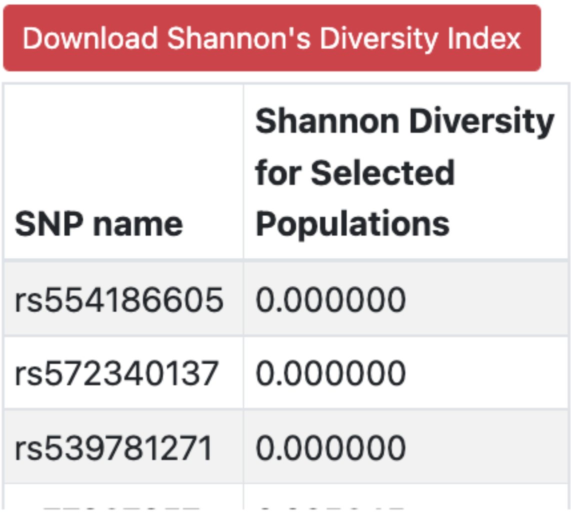


Figure 7 Vizualisation of Shannon's Diversity Index on all SNPs in between PMM1. The graph is calculated based on the populations selected (here: ESN, PEL, CHB, BEB and GBR) and shows differences in specific positions. In the bottom there is a sliding window making it possible to specify certain regions and zoom into these and evaluate the distribution of Shannon 's Diversity Index of PMM1.

Considerably, also for Shannon's Diversity Index SNPip provides a table to download the Shannon Diversity results calculated for every single SNP inside the selected region (Figure 8).



SNP name	Shannon Diversity for Selected Populations
rs554186605	0.000000
rs572340137	0.000000
rs539781271	0.000000

Figure 8 Downloadable table next to the visualization of Shannon's Diversity index to get a table of results containing all results of Shannon Diversity for every SNP and the selected populations.

Web Application

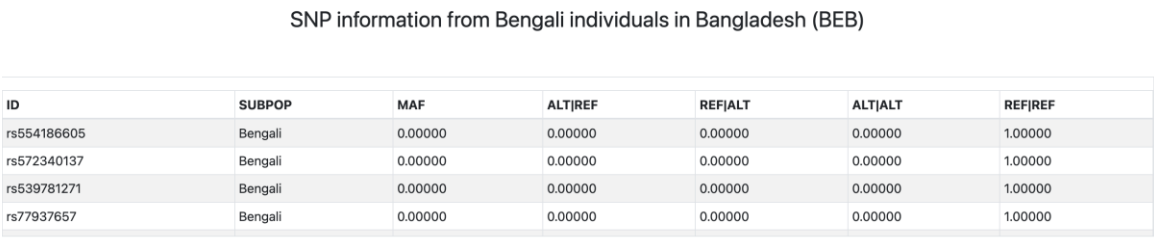
Input

SNPip search form can take 3 different types of inputs, the SNP name (RSID), the gene name and the genomic position. The input form for SNP name can only take a single RSID of interest, which will then returning information regarding that specific SNP. Likewise, only one gene name can be searched; although, depending on the gene multiple SNPs with complementary information will be return and relevant statistical data. Lastly, the genomic position takes a start and end number which can also return multiple SNP but also genes if multiple genes within the range.

The user can also select the population of interest, there are 5 different sub population the user can choose from: GBR, CHB, ESN, BEB, PEL (please look at Population and Sub-population subsection in Data).

Output

When searching with snp name, the output given is the RSID, subpopulation, minor frequency and genotype.



SNP information from Bengali individuals in Bangladesh (BEB)

ID	SUBPOP	MAF	ALT|REF	REF|ALT	ALT|ALT	REF|REF
rs554186605	Bengali	0.00000	0.00000	0.00000	0.00000	1.00000
rs572340137	Bengali	0.00000	0.00000	0.00000	0.00000	1.00000
rs539781271	Bengali	0.00000	0.00000	0.00000	0.00000	1.00000
rs77937657	Bengali	0.00000	0.00000	0.00000	0.00000	1.00000

Figure 9 Shows the SNP information for all the SNPs in the PMM1 gene.

For gene name and genomic position, the output will be similar to SNP name search, but for multiple SNP's if gene or position contains multiple SNP’s.
Additionally, when multiple SNPs are returned from searching using gene name and genomic position statistics such as Tajima's D is calculated given the populations selected, check out the statistics section for more detailed explanation on the statistical output.

Output files

SNPip download function allows users to download any table which has been generated by user search, this includes the SNP information table for every population and the statistical summary data. The plotly function also allows user to take a screenshot and download statistical data generated on graph.

Output file names

- SNP_Inforamtion.tsv
- Hudson_FST.tsv
- Tajima_D.csv
- Shannon_Diversity.tsv
- Haplotype_Diversity.tsv

Limitations and Scalability

Statistics

The bin sizes for every statistic cannot be individually set, this may be part of future updates. Explicitly this would make analysis and comparison easier and would provide support for smaller regions in bigger genes as well as for bigger regions in smaller genes.
Furthermore, sample size itself can affect the result of any statistic but the scalability of increasing the number of samples available is not a problem.

Smaller/ bigger bin sizes for haplotype diversity can affect the results. It is likely that smaller regions will give a lower value of haplotype diversity whereas bigger regions will increase the value of haplotype diversity since the probability of finding different alleles by random manner will rise with more and more samples. Considerably, it is more likely to get more heterozygous sites in bigger regions. Additionally, our bin size calculations can affect the last bin to be smaller than the rest of the bins which could affect the results, users should be aware of this problem and always look at the x-axis and the given ranges for each bin. To solve this problem a further search in which the specified bin could be searched as the central bin would allow the user to determine the true statistic for the full-size bin.

Database

Generally spoken our MySQL provides very good scalability on offering search for further Chromosomes and Populations in the future. The Decoder/ Encoder can be adjusted for more subpopulations and the storage of MySQL shows no limitation on this subject by now.

Search Time and Range

By now SNPip offers a chromosome wide search in ranges of in maximum 1 million bp to ensure user-friendly search times but these problems can be solved with future updates since we are aiming to enable chromosome wide search without facing memory problems.

Additiolly SNPip aims to get derived allele frequencies into the database and also aims to provide selection for several statistics.

In general, even more statistics can be added to ensure more population genomics research analysis in the future.

References

1. Danecek P, Bonfield JK, et al. Twelve years of SAMtools and BCFtools. Gigascience (2021) 10(2):giab008

2. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, 15 August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352

3. The 1000 Genomes Project Consortium. A global reference for human genetic variation.*Nature* 526**,** 68–74 (2015). https://doi.org/10.1038/nature15393

4. Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011

5. UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

6. Campbell IM, Gambin T, Jhangiani S, et al. Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. *Hum Mutat*. 2016;37(3):231-234. doi:10.1002/humu.22944

7. Jiang Y, Chen S, Wang X, et al. Association Analysis and Meta-Analysis of Multi-Allelic Variants for Large-Scale Sequence Data. *Genes (Basel)*. 2020;11(5):586. Published 2020 May 25. doi:10.3390/genes11050586

8. F Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism., *Genetics*, Volume 123, Issue 3, 1 November 1989, Pages 585–595, https://doi.org/10.1093/genetics/123.3.585

9. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992 Oct;132(2):583-9. doi: 10.1093/genetics/132.2.583. PMID: 1427045; PMCID: PMC1205159

10. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Res. 2013 Sep;23(9):1514-21. doi: 10.1101/gr.154831.113. Epub 2013 Jul 16. PMID: 23861382; PMCID: PMC3759727.

11. Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, José Carlos Marugán, Thomas Maurel, Aoife C McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P Sakthivel, Ahamed I Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, Ameya Chaubal, Nishadi De Silva, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth R IIsley, Nick Langridge, Jane E Loveland, Fergal J Martin, Jonathan M Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Fiona Cunningham, Andrew D Yates, Daniel R Zerbino, Paul Flicek Ensembl 2021**.** *Nucleic Acids Res.* 2021, vol. 49(1):884–891 PubMed PMID: 33137190. doi:10.1093/nar/gkaa942

12. E Papaemmanuil, FJ Hosking, J Vijayakrishnan, A Price, B Olver, E Sheridan, SE Kinsey, T Lightfoot, E Roman, JA Irving, JM Allan, IP Tomlinson, M Taylor, M Greaves, RS Houlston: *Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia*. In: *Nat Genet.*, 2009, 41(9), S. 1006–1010, doi:10.1038/ng.430, PMID 19684604

13. Hugot et al.: *Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's Disease*. In: *Nature*, Band 411, Mai 2001, S. 599–603, PMID 11385576

14. Ogura et al.: *A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease*. In: *Nature*, Band 411, Mai 2001, S. 603–606, PMID 11385577

15. DL Thomas, CL Thio, MP Martin, Y Qi, D Ge, C O'Huigin, J Kidd, K Kidd, SI Khakoo, G Alexander, JJ Goedert, GD Kirk, SM Donfield, HR Rosen, LH Tobler, MP Busch, JG McHutchison, DB Goldstein, M Carrington: *Genetic variation in IL28B and spontaneous clearance of hepatitis C virus.* In: *Nature*, 2009, 461(7265), S. 798–801. doi:10.1038/nature08463, PMID 19759533

16. T. Bersaglieri, P. C. Sabeti u. a.: *Genetic signatures of strong recent positive selection at the lactase gene.* In: *American Journal of Human Genetics.* Band 74, Nummer 6, Juni 2004, S. 1111–1120, doi:10.1086/421051, PMID 15114531, PMC 1182075

17. Friederike Flachsbart, Janina Dose, Liljana Gentschew, Claudia Geismann, *weitere 29 Autoren sowie* Almut Nebel: *Identification and characterization of two functional variants in the human longevity gene FOXO3.* In: *Nature Communication* 8/2017: 2063, 1–12. doi:10.1038/s41467-017-02183-y

18. Konopiński, M. Shannon diversity index: a call to replace the original Shannon's formula with unbiased estimator in the population genetics studies (2020). Peerj, 8, e9391. doi: 10.7717/peerj.9391

19. A. Chao, L. Jost, T. Hsieh, K. Ma, W. Sherwin and L. Rollins, "Expected Shannon Entropy and Shannon Differentiation between Subpopulations for Neutral Genes under the Finite Island Model", PLOS ONE, vol. 10, no. 6, p. e0125471, 2015. Available: 10.1371/journal.pone.0125471.

20. Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

21. Koboldt, D.C., Miller, R.D., Kwok, P.-Y., 2006. Distribution of Human SNPs and Its Effect on High-Throughput Genotyping. Hum Mutat 27, 249–254. https://doi.org/10.1002/humu.20286

22. Ramírez-Bello, J., Jiménez-Morales, M., 2017. [Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases]. Gac Med Mex 153, 238–250.

23. Shastry, B.S., 2009. SNPs: impact on gene function and phenotype. Methods Mol Biol 578, 3–22. https://doi.org/10.1007/978-1-60327-411-1_1

24. Shen, L.X., Basilion, J.P., Stanton, V.P., 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci U S A 96, 7871–7876.