

# Assignment 2

|  |  |  |
| --- | --- | --- |
| Date:  Course:  Name:  Student number: | 20/04/2025  MMG1344H  Maria Eleni Fafouti  1010799094 | |
|  | |  |
|  | |  |
|  | |  |

# Part 2

Task 1:

To indicate whether the two SNPs (rs683 and rs910) on the TYRP1 gene are fixed or variable in specific subpopulations (ASW, CEU, CHB, LWK, MKK, YRI), I first unzipped the main file and the files of all subpopulations. Then, I proceeded to define functions to help me extract the information for each of the two specific SNPs in a population file and then compute the allele frequencies by taking into account the genotypes of all individuals. I ran this by for-looping over the subpopulation .hmap files and got the following output:

A screen shot of a computer

AI-generated content may be incorrect.

Looking at the output above, we can observe that none of the SNPs are fixed: both alleles are present in all individuals, across all populations studied. However, those alleles are present at different frequencies. Below is a table that indicates which allele is more fixed for each population, for both of the two the two SNPs:

|  |  |  |  |
| --- | --- | --- | --- |
| **Population** | **SNP** | **More Fixed Allele** | **Frequency** |
| ASW | rs683 | C | 0.8169 |
| ASW | rs910 | A | 0.8310 |
| CEU | rs683 | A | 0.6512 |
| CEU | rs910 | C | 0.6709 |
| CHB | rs683 | C | 0.9878 |
| CHB | rs910 | A | 0.9878 |
| LWK | rs683 | C | 0.8537 |
| LWK | rs910 | A | 0.8494 |
| MKK | rs683 | C | 0.6871 |
| MKK | rs910 | A | 0.6959 |
| YRI | rs683 | C | 0.9018 |
| YRI | rs910 | A | 0.9049 |

Here’s a barplot visualizing these numbers:

A graph of different colored bars

AI-generated content may be incorrect.

We can observe that in the CHB population (Han Chinese in Beijing, China) one of the alleles is present at the same frequency (0.9878) for both SNPs, while the other allele is present in roughly only ~1% of the population. This means that for the CHB population, both SNPs are very close to being fixed. This case is also an example where allele frequency between the two SNPs does not differ. All other populations also show great similarity in this aspect: the frequency of the prevalent allele is very similar for both SNPs (eg. YRI: 0.9049 and 0.9018). However the identity of the prevalent allele is different (C for rs683 and A for rs910). It seems as if the allele frequencies for each of those SNPs are opposite from each other: All frequencies for C in rs683 match to all frequencies for A in rs910 and vice versa: all frequencies for A in rs683 match to all frequencies for C in rs910.

Task 2:

To calculate the fixation index for the indicated subpopulations (ASW, CHD, GIH, LWK, MXL, MKK, TSI) and compare it to the Li et al. 2012 data, we will first need to obtain the allele frequencies from all the populations. After computing this with the functions defined in the previous task, we get the following outcome (one table per SNP):

**SNP rs683:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Population** | **Fixed?** | **Allele A Frequency** | **Allele C Frequency** | **Fixed Allele** |
| ASW | False | 0.1831 | 0.8169 | None |
| CHD | False | 0.0143 | 0.9857 | None |
| GIH | False | 0.3012 | 0.6988 | None |
| LWK | False | 0.1463 | 0.8537 | None |
| MEX | False | 0.4366 | 0.5634 | None |
| MKK | False | 0.3129 | 0.6871 | None |
| TSI | False | 0.5779 | 0.4221 | None |

**SNP rs910:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Population** | **Fixed?** | **Allele A Frequency** | **Allele C Frequency** | **Fixed Allele** |
| ASW | False | 0.831 | 0.169 | None |
| CHD | False | 0.9857 | 0.0143 | None |
| GIH | False | 0.689 | 0.311 | None |
| LWK | False | 0.8494 | 0.1506 | None |
| MEX | False | 0.5652 | 0.4348 | None |
| MKK | False | 0.6959 | 0.3041 | None |
| TSI | False | 0.3506 | 0.6494 | None |

To compute Fst values, we will use the formula provided in the assignment instructions. From this, we observe that we will need to calculate two different kinds of Heterozygosity: total Heterozygosity which refers to all populations and subpopulation-specific Heterozygosity. Below, we have the general formula for heterozygosity, as per the assignment instructions.

To calculate the Heteozygosity for a biallelic SNP in one population, the formula will be the following. This is done for each of the SNPs:

**SNP rs683**

|  |  |  |
| --- | --- | --- |
| Population | Allele Frequencies | Heterozygosity (H\_S) |
| ASW | A: 0.1831, C: 0.8169 | 0.2991 |
| CHD | C: 0.9857, A: 0.0143 | 0.0282 |
| GIH | A: 0.3012, C: 0.6988 | 0.4210 |
| LWK | C: 0.8537, A: 0.1463 | 0.2498 |
| MEX | C: 0.5634, A: 0.4366 | 0.4920 |
| MKK | C: 0.6871, A: 0.3129 | 0.4300 |
| TSI | C: 0.4221, A: 0.5779 | 0.4879 |

**SNP rs910**

|  |  |  |
| --- | --- | --- |
| Population | Allele Frequencies | Heterozygosity (H\_S) |
| ASW | A: 0.8310, C: 0.1690 | 0.2809 |
| CHD | A: 0.9857, C: 0.0143 | 0.0282 |
| GIH | A: 0.6890, C: 0.3110 | 0.4286 |
| LWK | A: 0.8494, C: 0.1506 | 0.2558 |
| MEX | A: 0.5652, C: 0.4348 | 0.4915 |
| MKK | A: 0.6959, C: 0.3041 | 0.4232 |
| TSI | A: 0.3506, C: 0.6494 | 0.4554 |

To calculate the total Heterozygosity, we will need to compute the mean allele frequency per allele across all populations. Below is an example calculation for the Adenine (the mean frequency across all populations), the same will be computed for cysteine and will be shown in a table below:

**SNP rs683**

|  |  |
| --- | --- |
| Allele | Mean Frequency ( |
| A | 0.2818 |
| C | 0.7182 |

**SNP rs910**

|  |  |
| --- | --- |
| Allele | Mean Frequency ( |
| A | 0.7095 |
| C | 0.2905 |

These values can be plugged into the following formula to calculate the total Heterozygosity (across all populations):

After obtaining the values for both and , we can calculate the Fst (Fixation Index), we will use the total Heterozygosity (calculated above), along with the average heterozygosity for all populations. This is the formula for the latter:

After obtaining all those values we can calculate the FST per SNP:

Here are the results:

|  |  |  |
| --- | --- | --- |
|  | **rs683** | **rs910** |
| **H\_T** | 0.4047 | 0.4122 |
| **H\_S(mean)** | 0.3440 | 0.3377 |
| **FST** | 0.1501 | 0.1808 |

The Li et al. paper reports the following table of 30 SNPs and their FST values. All values lay above 0.5, while some of them are as high as 0.82. This is quite different from what we found analyzing the populations that were not part of the analysis in the Li et al paper. The trend for higher FST values persists when looking at the specific SNPs that were also analyzed in this assignment (FST values highlighted in pink in the upcoming table). The fixation index provides insight into the genetic differentiation between the populations analyzed. In our case, for both SNPs there is very low genetic differentiation for the populations analyzed, while in Li et al. this is not the case. Values above 0.50 indicate that the populations analyzed have strong genetic differences. For reference, a value of 1 for the FST index means complete genetic separation. The contrast in FST values underscores the importance of sampling strategy in population genetics, as different population sets can reveal vastly different levels of genetic differentiation.

A shorter version of table 1 from the paper, containing the FST information for each unique SNP can be found in the following page. able in the next page.

*Table 1. Li et al. 2012*

|  |  |  |  |
| --- | --- | --- | --- |
| SYMBOL | SNP | mALLELE | FST |
| ADAM17 | rs6705408 | C | 0.507805 |
| ADD3 | rs1127522 | G | 0.618752 |
| **APPBP2** | **rs9893667** | **C** | **0.822717** |
| APPBP2 | rs6503971 | G | 0.508963 |
| C4orf46 | rs3210749 | C | 0.598659 |
| CYTIP | rs267992 | T | 0.529737 |
| DCK | rs4643786 | C | 0.653711 |
| DPYD | rs291592 | C | 0.502731 |
| EXOC5 | rs10150771 | C | 0.660272 |
| EXOC5 | rs3742577 | C | 0.655322 |
| HEXA | rs3087652 | T | 0.55013 |
| HNF1B | rs1058166 | C | 0.533675 |
| KIF3C | rs6731972 | C | 0.578667 |
| LARS2 | rs267266 | A | 0.517315 |
| MAP3K3 | rs15637 | G | 0.512258 |
| NUBPL | rs7148746 | G | 0.815249 |
| OXTR | rs237884 | G | 0.502445 |
| PDE7B | rs7753890 | C | 0.674353 |
| RSPO4 | rs516714 | G | 0.513515 |
| SLC25A19 | rs7198 | G | 0.638151 |
| SMNDC1 | rs1050755 | C | 0.531575 |
| TCF7L2 | rs1056877 | T | 0.692805 |
| TLN2 | rs16945912 | T | 0.504656 |
| TMX3 | rs309202 | C | 0.615157 |
| TTC14 | rs7645817 | G | 0.554868 |
| TYRP1 | rs683 | C | 0.556808 |
| TYRP1 | rs910 | A | 0.56451 |
| XYLT1 | rs8044934 | A | 0.592521 |
| XYLT1 | rs8046024 | T | 0.592521 |
| ZNF586 | rs10401255 | C | 0.570075 |