Tutorials - Linkage analyses

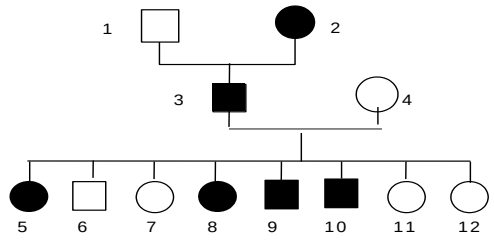
Exercise 1 :

The two pedigrees below represent families with affected subjects by neurofibromatosis (a disease characterized by the presence of skin, nerve or central nervous system tumors and pigment spots on the skin).

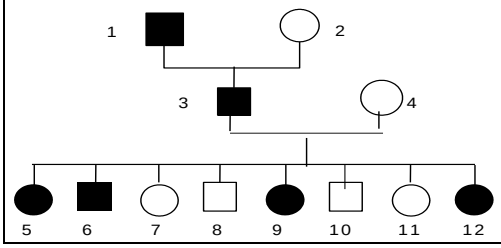
A probe of chromosome 17 recognizes a 3kb allele if the EcoRI site is present in the studied region and a 4.7 kb allele if this site is absent.

The results of the Southern Blot carried out for the individuals of each family are as follows:

Family 1



Family 2



Family 1	Band size	Family 2	Band size
ind 1	4,7 kb	ind 1	3kb 4,7kb
ind2	3 kb	ind2	3 kb
ind3	3kb 4,7kb	ind3	3kb 4,7kb
ind4	3kb	ind4	4,7kb
ind5	3kb	ind5	4,7kb
ind6	3kb 4,7kb	ind6	4,7kb
ind7	3kb 4,7kb	ind7	3kb 4,7kb
ind8	3kb	ind8	3kb 4,7kb
ind9	3kb	ind9	4,7kb
ind10	3kb	ind10	3kb 4,7kb
ind11	3kb 4,7kb	ind11	3kb 4,7kb
ind12	3kb 4,7kb	ind12	3kb 4,7kb

Q1 : How is the mode of inheritance? Write the genotype of the different individuals in the families.

Q2 : From family data, estimate the value of the recombination rate θ between these two loci; indicate whether there is a genetic linkage between the RFLP marker and the disease locus.

Q3 : Using the lod score method, we investigated whether there was a genetic linkage between the neurofibromatosis gene and the RFLP marker.

a) Determine the value of the statistic Z(θ) for θ=0 et 0,5 for each of the 2 families.
b) Calculate the lodscore in the total sample for these two values of θ.
c) The lodscore values calculated for different values of the recombination rate θ are reported below. Comment these results.

θ	**0**	**0.05**	**0,06**	**0,1**	**0,2**
fam1	2.41	2.23	2.19	2.04	1.63
fam2	- infinity	0.95	1	1.09	1.03
total	- infinity	3.18	3.19	3.13	2.66

d) The maximum lodscore for all of the two families is 3.19; the exact value of θ corresponding to this lodscore is 0.0625. What can you conclude from this result and is it consistent with your answer to Q2?

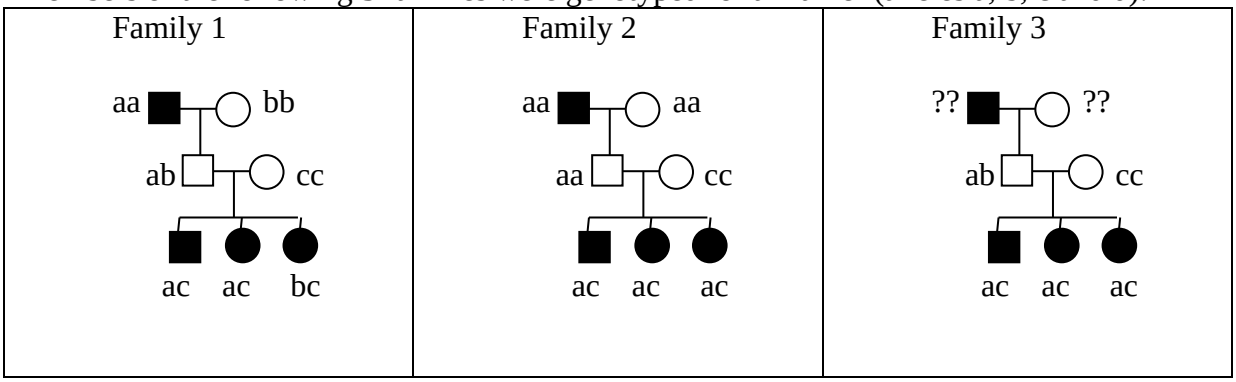
Q4 : Dans l'exemple présenté, les conclusions pouvaient être trouvées par analyse des enfants sans le calcul du lodscore. En quoi cette situation est-elle particulière? Quel est l'intérêt de la méthode du lodscore ?

Exercise 2 : Calculation of Lod-Score

We seek to locate the gene of a disease using the Lod-Score method. The genetic model underlying the disease has the following characteristics:

- autosomal recessive inheritance
- complete penetrance
- absence of phenocopy

Members of the following 3 families were genotyped for a marker (alleles a, b, c and d):



Q1 : Calculate the value of the Lod-score in the total sample for θ=0, knowing that there are in the studied sample 8 families of the same type as family 3 (and 1 family of the two other types).

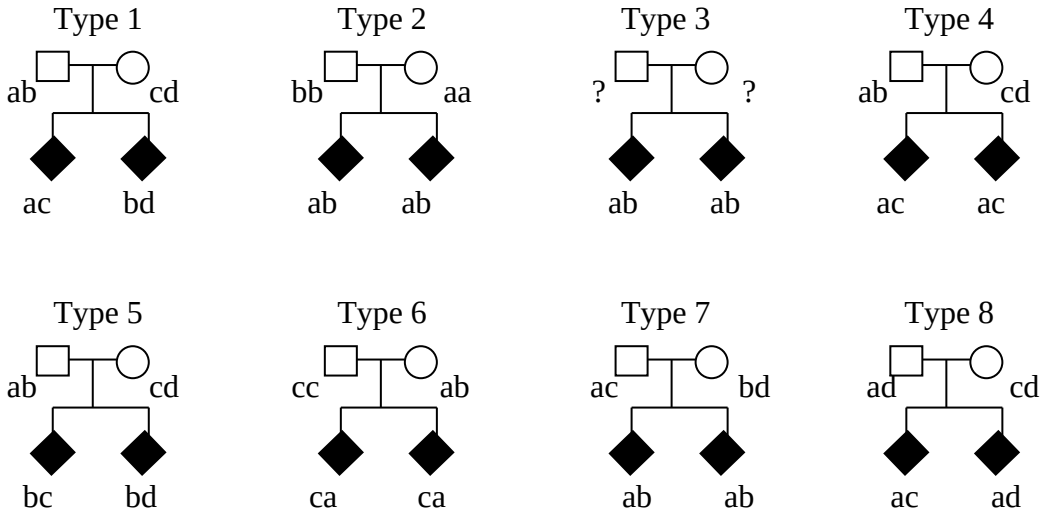
Q2 : The Lod-score results in the total sample for other values of θ are given in the table below. What can you conclude from this (include your results from Q1)?

θ	0.1	0.2	0.3	0.4	0.5
Lodscore	3.53463782	2.55480664	1.43250104	0.45519666	0

Exercice 3 : Linkage analyse : sib-pairs method

Subjects were genotyped for a marker with 4 alleles (named a, b, c and d and where f(a) = 0.1; f(b)=0.3; f(c)=0.4; f (d)=0.2) in a set of nuclear families (parents + 2 affected siblings). From the distribution of the different genotypes, we can classify families by "Family type".

Q1 : Among the following set of nuclear families, which ones cannot be used to directly perform the χ^2 test to test the genetic linkage between the disease gene and the marker?



Q2 : From a sample of families containing 41 type 4 families, 33 type 8 families and 26 type 1 families, indicate whether there is evidence for genetic linkage using a statistical test.

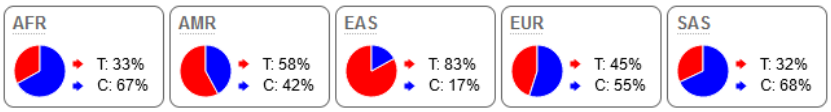
Tutorials - Association analyses

Part I : SNP selection and analysis of study populations

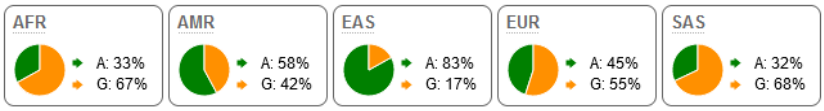
Exercice 1

Allele frequencies of 12 SNVs on a region of chromosome 1 have been reported in figure 1 below.

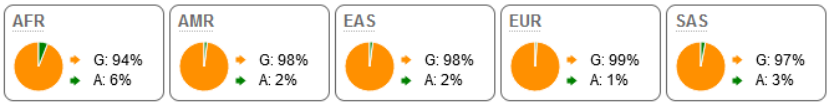
1 rs1217404 – chromosome 1 : 113849937



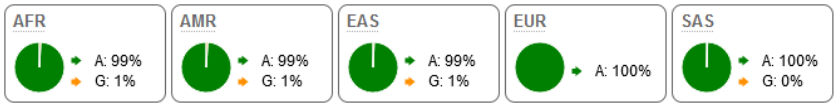
2 rs1217405 – chromosome 1 : 113850010



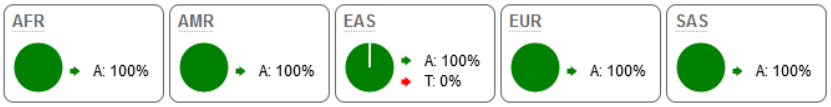
3 rs532580695 – chromosome 1 : 113850272



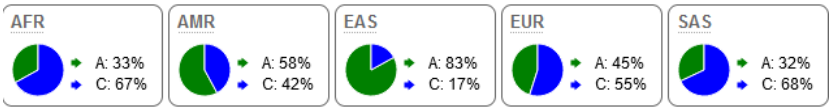
4 rs141976030 – chromosome 1 : 113850278



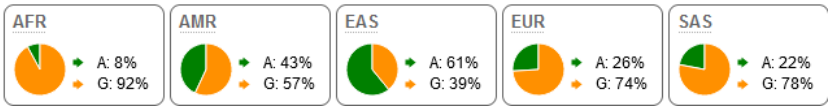
5 rs61496190 - chromosome 1 : 113850301



6 rs1217406 – chromosome 1 : 113850531

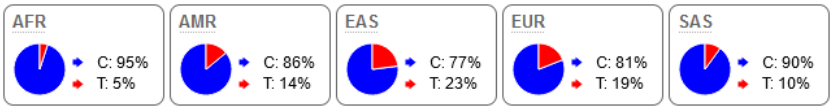


7 rs1217407 – chromosome 1 : 113851126

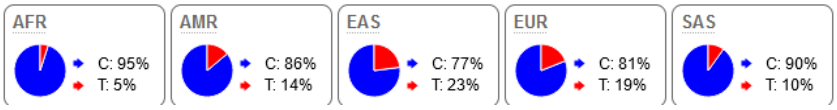


8 rs5777169 – chromosome 1 : 113851157-113851170

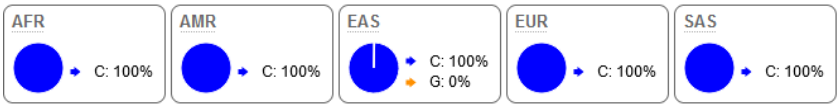
9 rs3765598 – chromosome 1 : 113851841



10 rs4839346 – chromosome 1 : 113853007



11 rs180914933 – chromosome 1 : 113853169



12 rs1217408 – chromosome 1 : 113853658

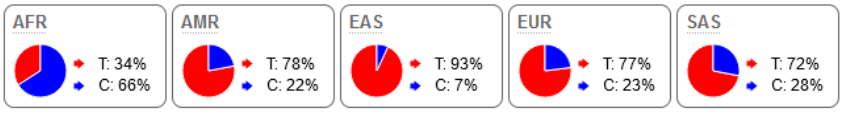


Figure1 :
Allele frequencies of SNVs from 1000 Genomes Project in five populations
AFR : African
AMR : American
EAS : East Asian
EUR : European
SAS : South Asian

https://www.ensembl.org/

Q1/ Describe profiles of allele frequencies observed for these 12 SNVs.

Linkage disequilibrium r^2 between each pair of SNV was estimated in a particular sub-population (JPT) and reported in the table below.

Table 1 : r^2 values between SNVs in JPT population for a region of 12 SNVs in chromosome 1

SNP												
rs1217404	rs1217404											
rs1217405	1	rs1217405										
rs532580695	-	-	rs532580695									
rs141976030	-	-	1	rs141976030								
rs61496190	-	-	-	-	rs61496190							
rs1217406	1	1	-	-	-	rs1217406						
rs1217407	0.335	0.335	-	-	-	0.335	rs1217407					
rs5777169	0.229	0.229	-	-	-	0.229	0.784	rs5777169				
rs3765598	0.085	0.085	-	-	-	0.085	0.374	0.382	rs3765598			
rs4839346	0.085	0.085	-	-	-	0.085	0.374	0.382	1	rs4839346		
rs180914933	-	-	-	-	-	-	-	-	-	-	rs180914933	
rs1217408	0.54	0.54	-	-	-	0.54	0.181	0.108	-	-	-	rs1217408

Q2/ How this r^2 is calculated and how interpret its value ?
Q3/ Explain why some data are missing
Q4/ How this table will help to determine tagSNPs ?

Exercice 2

tagSNPs are genotyped in cases and controls for an association study. Before testing SNP association, a Principal Component Analysis (PCA) is conducted. Figure 2 shows PCA for SNPs genotyped in samples from 4 reference populations (CEU, YRI, CHB & JPT) and in cases of the study.

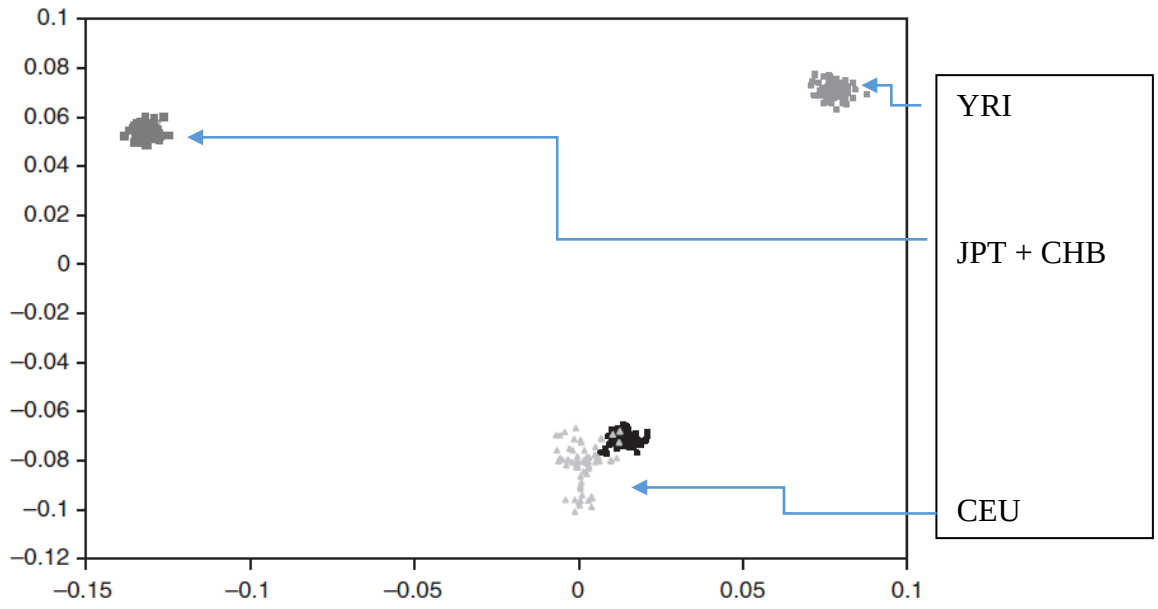


Figure 2 : PCA on SNPs data for 4 reference populations CEU, CHB, JPT et YRI (in grey), and for cases of the study (in black). From Hosking et al., 2011.

Q5/ What is your conclusion about this PCA concerning cases of the study ?

Q6/ How to complete this preliminary analysis ?

Part II: Association Tests

Exercice 3 : familial (TDT) and population-based (cases-controls) association analyses

Six SNPs in a candidate gene were genotyped in 600 cases and 600 controls (alleles are noted 1 and 2 for each SNP). Cases have a multifactorial disease called M and information on an intermediate quantitative phenotype[1] (called IQP) is also available for all individuals (cases and controls).

The genotypes of each SNP were coded as follows:

- 0 corresponds to genotype 11
- 1 corresponds to genotype 12
- 2 corresponds to genotype 22

Q1: What type of coding (general, additive, dominant, recessive) was used to indicate the genotype of individuals for the different SNPs?

Q2: Is the Hardy-Weinberg model verified in the controls for SNP n°2, knowing that the observed counts for each genotype are as follows:

Genotype :	0	1	2
Observed counts	35	250	315

NB: We will assume that the Hardy-Weinberg model is verified for the other SNPs, except for SNP n°6.

Q3: Assuming that the allele frequency in the controls is representative of the allele frequency in the population from which they come, and knowing that:

- the allele frequency of allele 1 in controls for SNP5 is equal to 0.275
- the frequency of haplotype 11 for SNPs 2 and 5 is equal to 0.266 in the studied population (haplotype 11 corresponds to the combination of allele 1 for SNP2 and allele 1 for SNP5)

Calculate the value of the linkage disequilibrium r^2 between SNPs 2 and 5. What can you conclude? Is it necessary to study the 2 SNPs?

NB: SNPs 1, 3, 4 and 6 are not in linkage disequilibrium with SNP2, nor with each other.

[1] The analysis of an intermediate quantitative phenotype can sometimes make it easier to identify genetic factors involved in a disease. Example of a PQI: IgE corresponds to a PQI of asthmatic disease since most asthmatic subjects have a high IgE level. Thus, finding genetic factors involved in this PQI could provide a better understanding of genetic susceptibility to asthma. However, since the correlation between M and PQI is not 100% (some individuals with high IgE levels do not have asthma), certain genetic factors may be specific for high IgE levels.

To find likelihood ratio test(LRT): (SNP1 as example)
(-2)x(lnH0) - (-2)x(lnH1) <=> 1663.6 - 1661 = 2.6 ~3.84 (1:df)

Q4: The results of a logistic regression analysis are shown in the table below for each of the SNPs

	Estimate	Std. Error	z value	Pr(>|z|)	Likelihood	LRT
SNP1	0.13005	0.08075	1.611	0.107	1661.0	**2.6 (H0 n.r.)**
SNP2	-0.55926	0.09424	-5.935	2.94e-09 ***	1627.1	**36.6 (H0 r.)**
SNP3	-0.01928	0.08016	-0.240	0.810	1663.5	**0.1 (H0 n.r)**
SNP4	0.08102	0.08221	0.985	0.324	1662.6	**2 (H0 n.r)**
SNP5	-0.43184	0.09378	-4.605	4.13e-06 ***	1641.9	**21.7 (H0 r.)**

D

ORs (find beta hat -> exp(beta))
12vs11->exp(-0.56) = 0.57
22vs11->exp(2x-0.56) = 0.33

CI = exp[beta ± 1.96*root_var_beta]
CI(lower) = exp[beta - ….
CI(upper) = exp[beta + ….

The root_var_beta equates to the std.error in the table

=> Estimate corresponds to the regression coefficient b associated with the tested variable, knowing that logit P = a + b SNP1 (here, a = intercept and b = regression coefficient associated with the variable SNP1)
=> Std. Error corresponds to the standard deviation of b
=> Z-value corresponds to the Wald test = Estimate / Std. Error (note: H0: b=0 ; H1: b≠0)
=> Pr(>|z|) corresponds to the p-value associated with the Wald test
*=> Likelihood corresponds to -2*ln(L(H1) of the model considered. Note that for a model not including any SNP, -2*ln(L(H0))= 1663.6*

CI(12vs11) = [(0.57 - 1.96*0.094 =) _0.47_ , (0.57 + 1.96*0.094=) _0.69_]

CI(22vs11) = [(2*0.33 - 1.96*0.094 =) _0.47_ , (2*0.33 + 1.96*0.094=) _0.69_]

a/ Does the phenotype studied correspond to M or PQI? **Remember to mult by 2, bcs you did that for 22vs11**

b/ Why are there no results for SNP6? **SNP6 was removed during quality control, possibly due to heterogeneity between SNP6 cases and control.**

c/ For each of the SNPs, perform the likelihood ratio test and indicate the SNP(s) associated with the studied phenotype. **SNP2 and SNP5 highly significant contributors; LRT confirms that SNP1, 3, 4 not rejected, but validates the results for SNP2 and SNP5**

d/ For SNP2, calculate the odds ratio and the associated confidence interval with genotype 12 compared to genotype 11. Deduce the odds ratio and the associated confidence interval with genotype 22 compared to genotype 11. Conclude.

Q5 : For one additional SNP (A and B alleles), the distribution of genotypes in cases and controls is shown in the table below.

	Cases	**Controls**	**ROW**
BB	50	150	**200**
AB	75	80	**155**
AA	95	60	**155**
COLUMN	**220**	**290**	**520**

a) Calculate the Odds ratio (and confidence intervals) associated with the AB genotypes on the one hand and AA on the other hand given the reference genotype BB.

b) Calculate the Odds ratio (and the confidence interval) associated with the B allele compared to the reference allele A.

c) What can you conclude?

Q6 : SNP2 was genotyped in a sample of trios (2 parents + one affected child). The distribution of different trios according to the genotype of the parents and children is given in the table below. Is one of the marker alleles associated with the disease? Compare to Q4 results and comment.

Father's genotype	**Mother's genotype**	**Child's genotype**	**Number of trios**
11	11	11	1
11	12	11 12	4 8
11	22	12	4
12	11	11 12	4 8
12	12	11 12 22	7 8 9
22	11	12	4
22	12	12 22	25 3
22	22	22	16

Part III: Generalizations of association tests

Exercise 4 : haplotypic analysis

Two SNPs were genotyped in cases and controls (SNP A of A1 and A2 alleles and SNP B of B1 and B2 alleles). We know that the A1 allele is never found with the B1 allele in the population studied. The numbers of genotypes for the 2 SNPs in cases and controls are shown in the table below:

SNP A	SNP B	Cases	Controls
A1A1	B2B2	24	49
A2A2	B1B2	35	39
A2A2	B2B2	68	26
A1A2	B1B2	27	29

Q1/ Determine the observed haplotypes in cases and controls.

Q2/ Perform a haplotype association test and conclude.

Q3/ Calculate the ORs (and CIs) associated respectively with the A2B1 and A2B2 haplotypes compared to the reference haplotype A1B2.

Exercise 5 : Interaction GxE

A genetic marker SNP (bi-allelic with alleles 1 and 2) was genotyped in cases and controls. Information on an environmental factor was also collected and is summarized in the tables below:

E+			
	allele 1	allele 2	total
Controls	192	288	480
Cases	120	480	600

E-			
	allele 1	allele 2	total
Controls	432	288	720
Cases	480	120	600

Q1/ Is there an interaction between the environmental factor and the SNP (for each stratum, calculate the odds ratio associated with allele 1 compared to allele 2) ?

Q2/ How would you write the logit of P to test the SNP x environment interaction by logistic regression considering additive coding for the SNP? How would you calculate the odds ratio associated with genotype 22 compared to genotype 11 for individuals exposed to the environmental factor?