GENOM

IMPUTATION

Données GWAS et variations phénotypiques

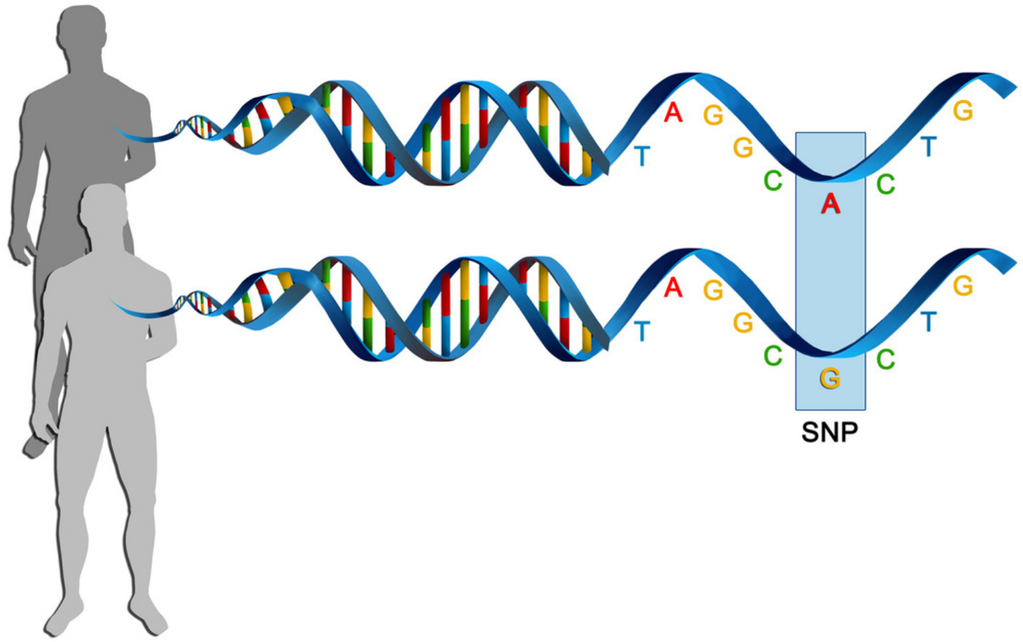
Alix BOUTHEROUE-DESMARAIS Kenza FLIOU Yuliya LIM
Sophie Garnier

INTRODUCTION - GWAS

GWAS - Genome Wide Association Studies

QTL - Quantitative Trait Loci mapping

BUT : Étudier l'association entre les variations génétiques (SNPs) et des traits phénotypiques particuliers



???

INTRODUCTION - GWAS

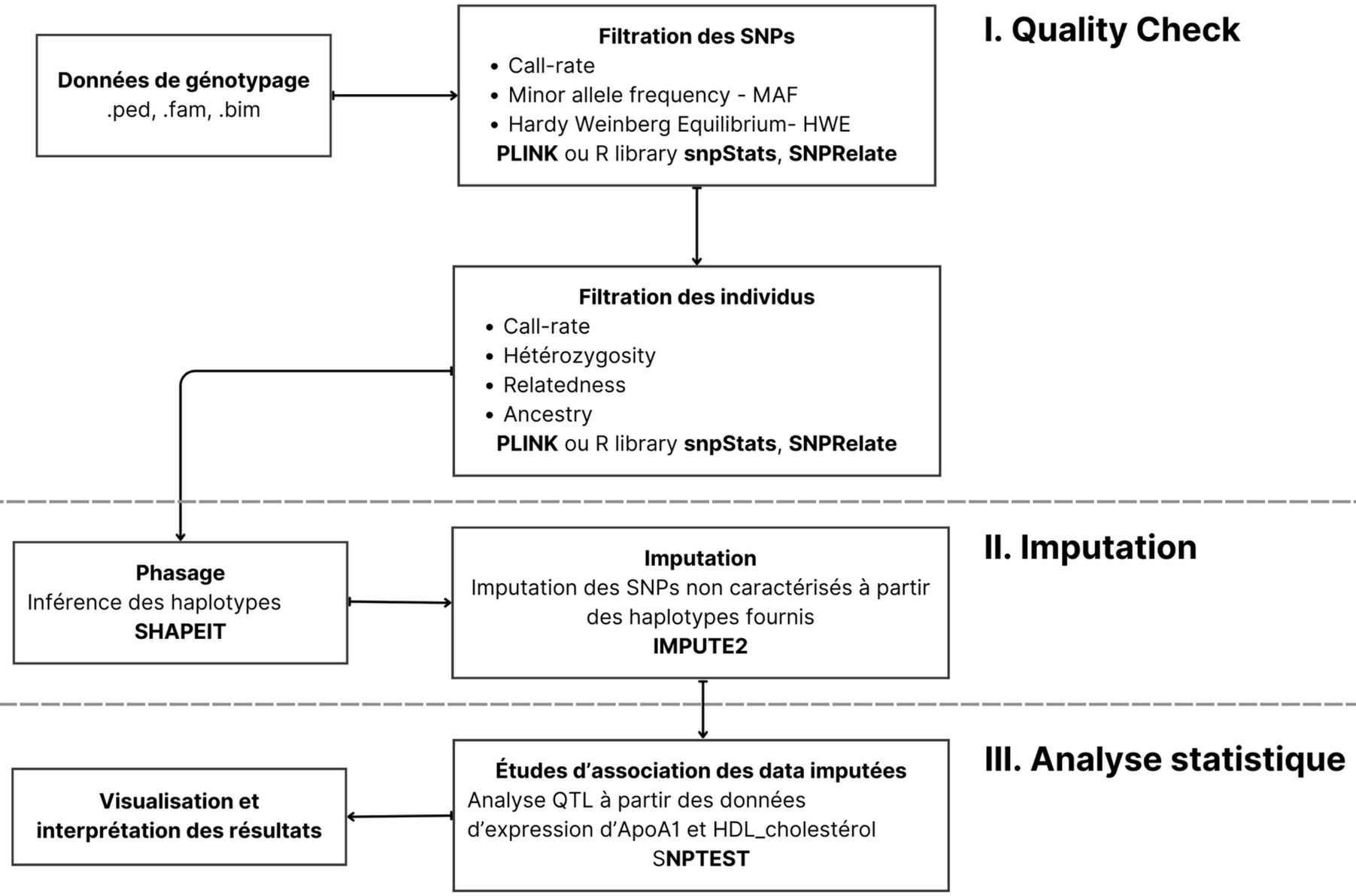
Description des données

- Cohort : 3300 individus sains
- 191 037 SNPs
- Données d’expression d’HDL_cholésterol et d’ApoA1:
  - HDL_cholésterol : lipoprotéines responsables du transport du cholestérol
  - ApoA1 : protéine présente dans 20 % de la population qui pourrait constituer un marqueur de risque cardio-vasculaire

Analyse QTL (quantitative trait locus)

But : identifier les SNPs associées à l’expression quantitative de ces deux protéines

PIPELINE


Données de génotypage
.ped, .fam, .bim
Filtration des SNPs
• Call-rate
• Minor allele frequency - MAF
• Hardy Weinberg Equilibrium- HWE
PLINK ou R library snpStats, SNPRelate
Filtration des individus
• Call-rate
• Hétérozygosity
• Relatedness
• Ancestry
PLINK ou R library snpStats, SNPRelate
I. Quality Check
Phasage
Inférence des haplotypes
SHAPEIT
Imputation
Imputation des SNPs non caractérisés à partir des haplotypes fournis
IMPUTE2
II. Imputation
Visualisation et interprétation des résultats
Études d'association des data imputées
Analyse QTL à partir des données d'expression d'ApoA1 et HDL_cholestérol
SNPTEST
III. Analyse statistique

QUALITY CHECK - plink ou R package

Filtrage des SNPs

- **SNPs call rate** (seuil à 98%) : conservation des SNPs présent chez 98% des individus

- **MAF** (seuil à 5%) : SNP peu représenté dans la population a peu de chances d'être significativement associé

- **HWE** (pvalue <1e-6) : violation de HWE
  - indication de sous-structure de la population
  - ou erreur de génotype.

- Exclusion des **SNPs ambigus** (A/T et C/G) et tri-**alléliques**

Avant filtrage	**Après filtrage**	**Après filtrage**
191 037 SNPs	**129 769 SNPs**	**108 996 SNPs**
3 300 individus	3 300 individus	3 300 individus

QUALITY CHECK - plink ou R package

Filtrage des individus

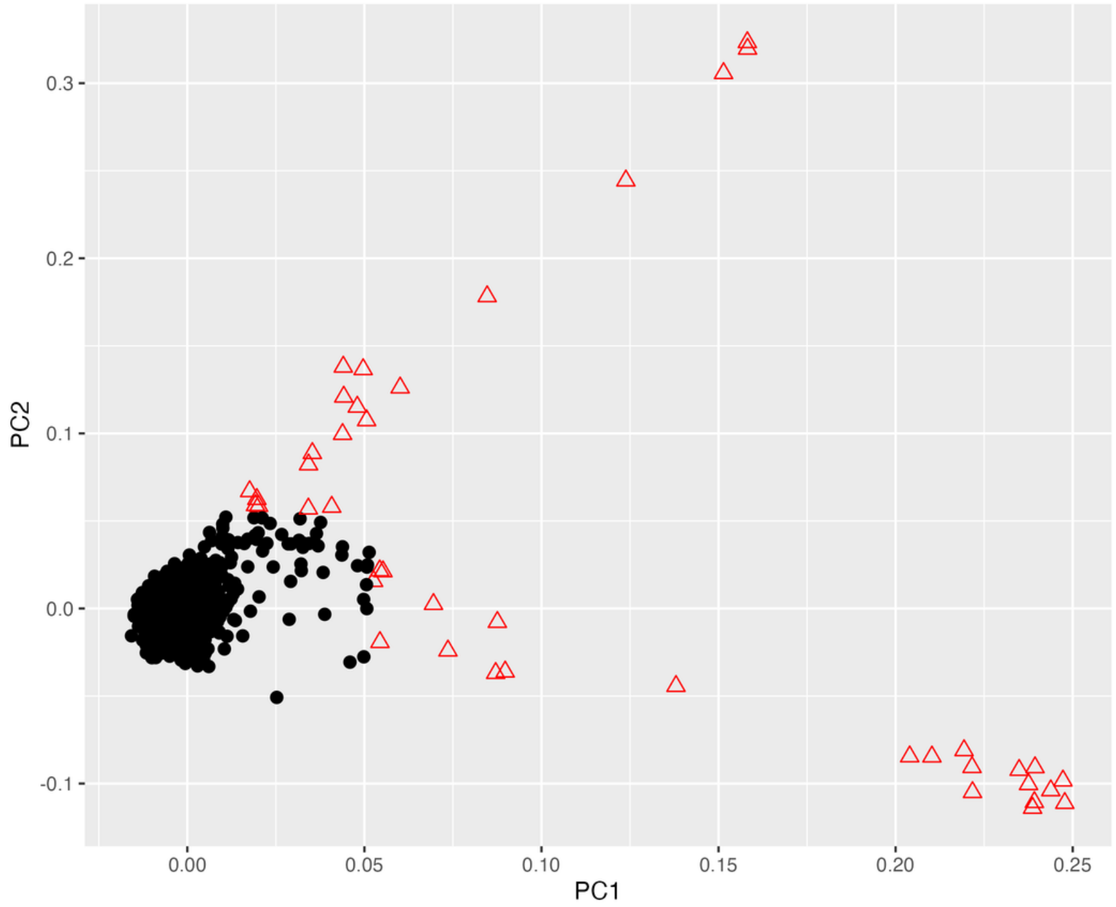
- **Sample call rate** (seuil à 95%) : conservation des individus qui présentent 95% des SNPs sélectionnés pour l'étude

- **Hétérozygotie** (± 3SD par rapport à la moyenne) :
  - un excès peut-être lié à une pauvre qualité du sample
  - une déficience une indication de consanguinité

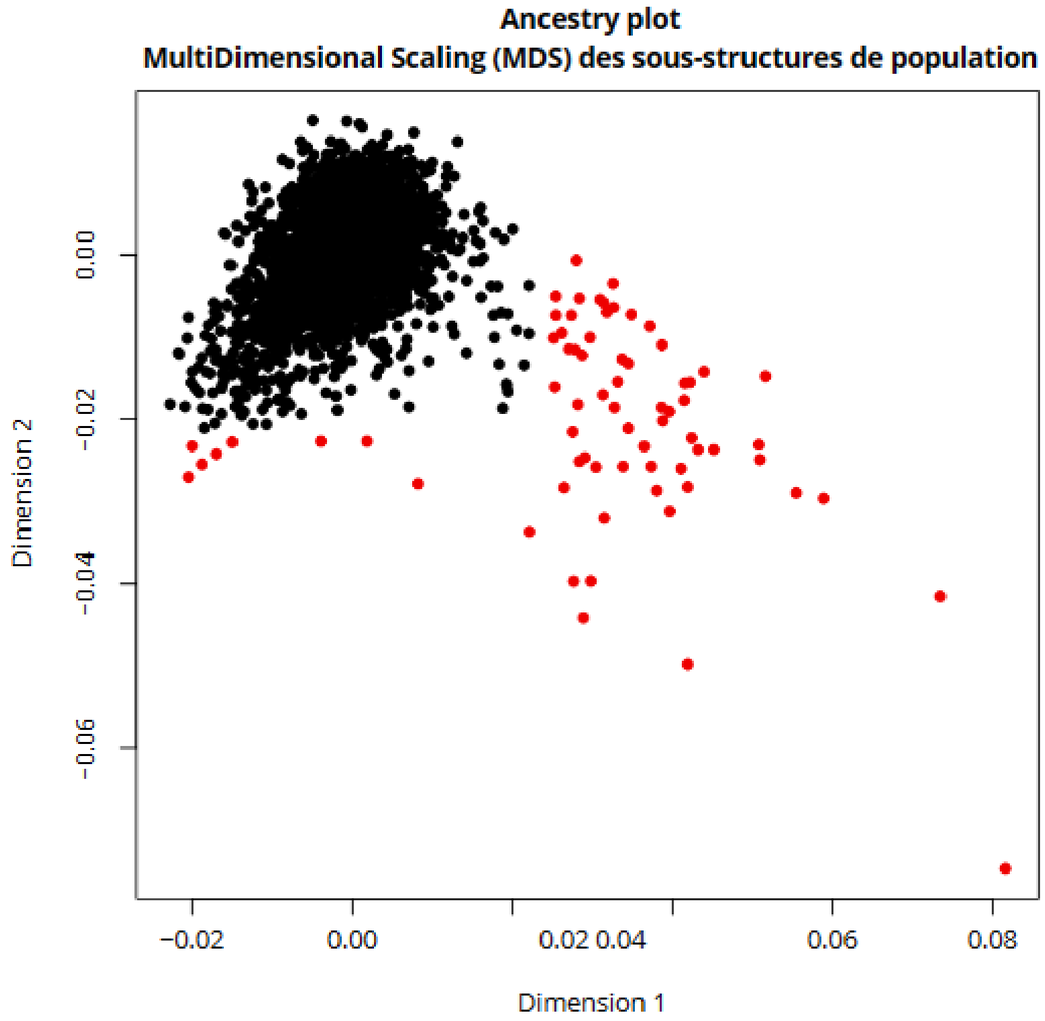
- **Relatedness ou Identity by descent (IBD)** (linkage disequilibrium pruning method)

- **Ancestry ou population substructure** (PCA - MDS) : les substructures de population dû à l'éthnicité peuvent rajouter un biais dans l'étude d'association. Un individu en dehors d'un groupe éthnique peut être dû à un sample de mauvaise qualité.

QUALITY CHECK - plink ou R package



Ancestry plot - ACP des sous-structures de population



 Exclusion des samples qui s’éloignent de ±3SD de la moyenne d’un des deux axes du MDS (● △)

QUALITY CHECK - plink ou R package

Filtrage des individus

Avant filtrage	Après filtrage	Après filtrage
191 037 SNPs	129 769 SNPs	108 996 SNPs
3 300 individus	**3 175 individus**	**3 004 individus**

IMPUTATION - ShapeIT + Impute2

BUT : imputation des SNPs manquants à partir d'haplotypes de références (1000 Genomes) avec l'appui des haplotypes parentaux

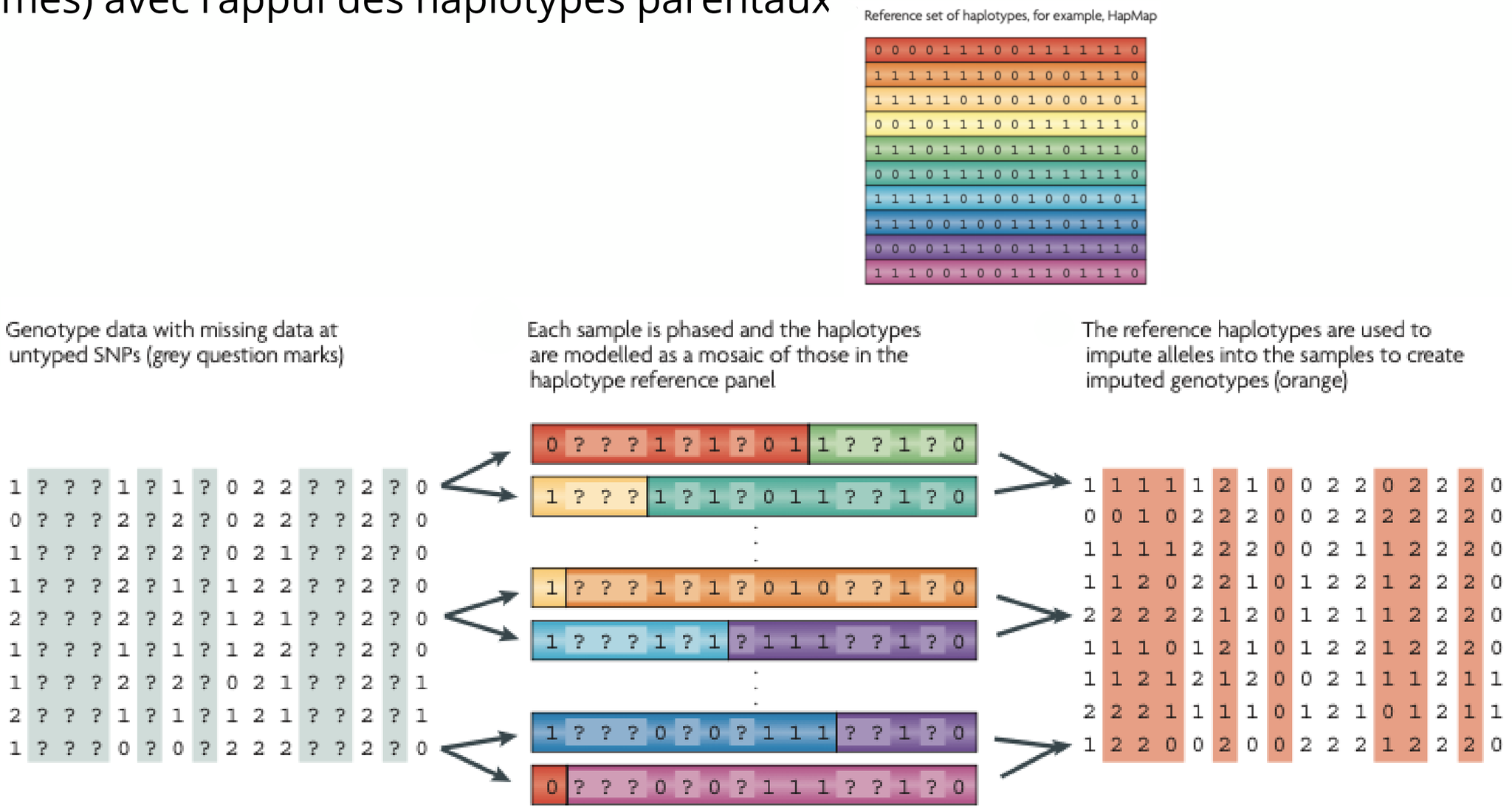
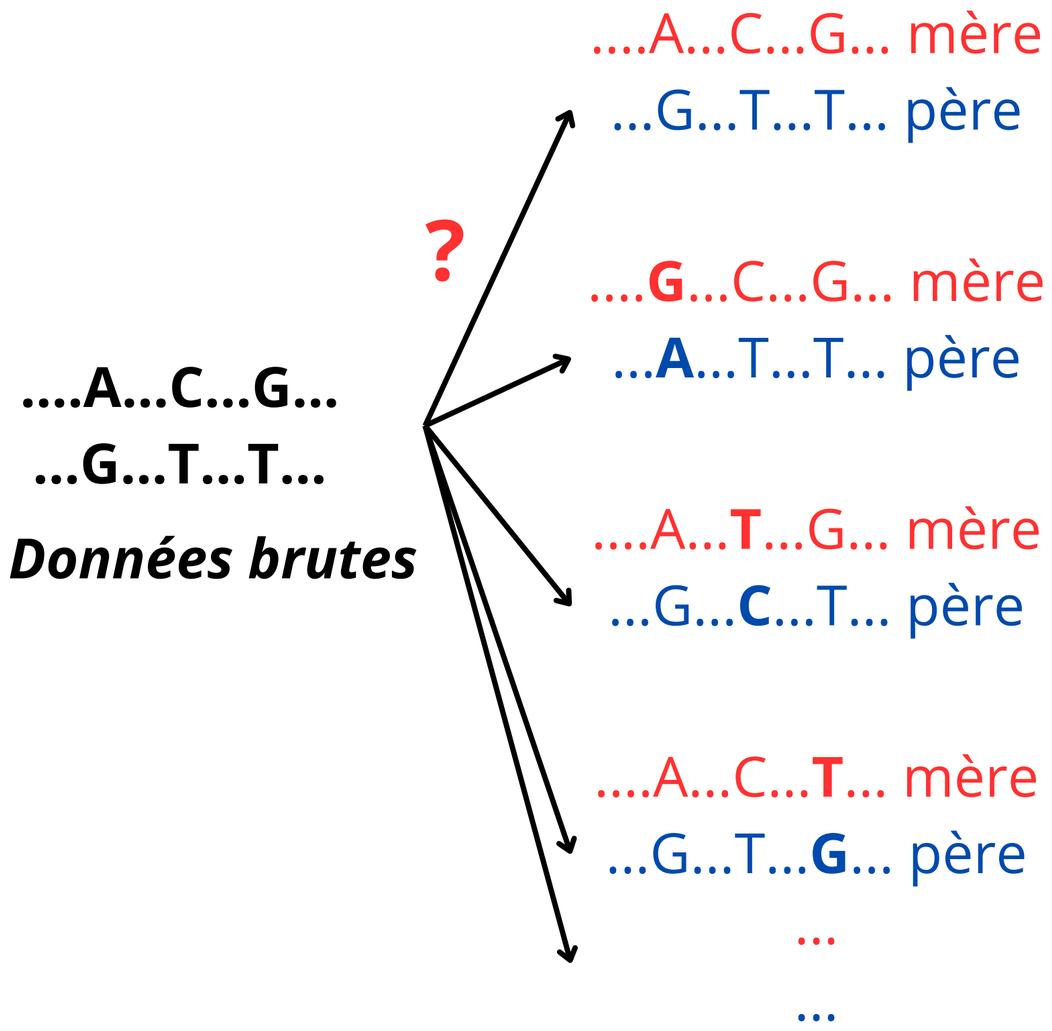


figure issue de l'article : "Genotype imputation for genome-wide association studies" - Jonathan Marchini* and Bryan Howie‡

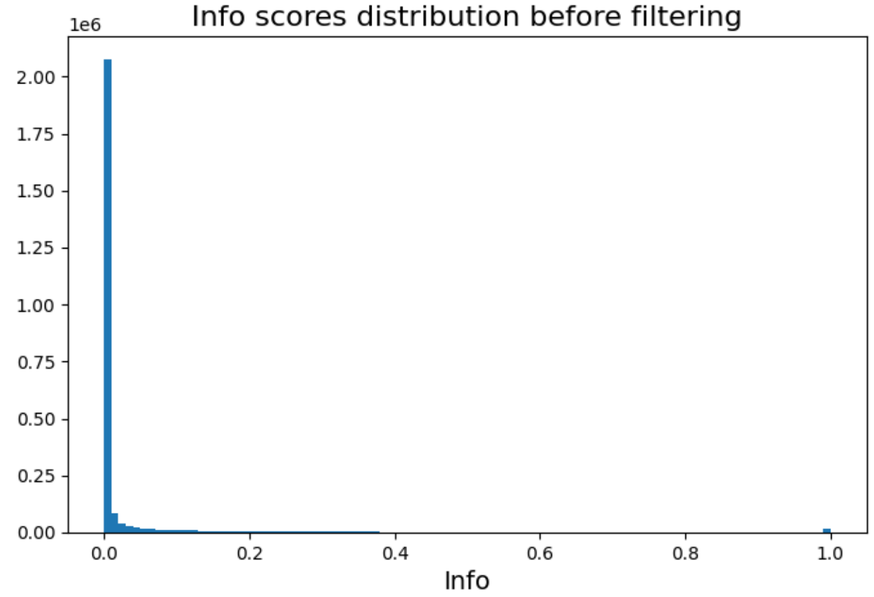
IMPUTATION - Phasage avec SHAPEIT

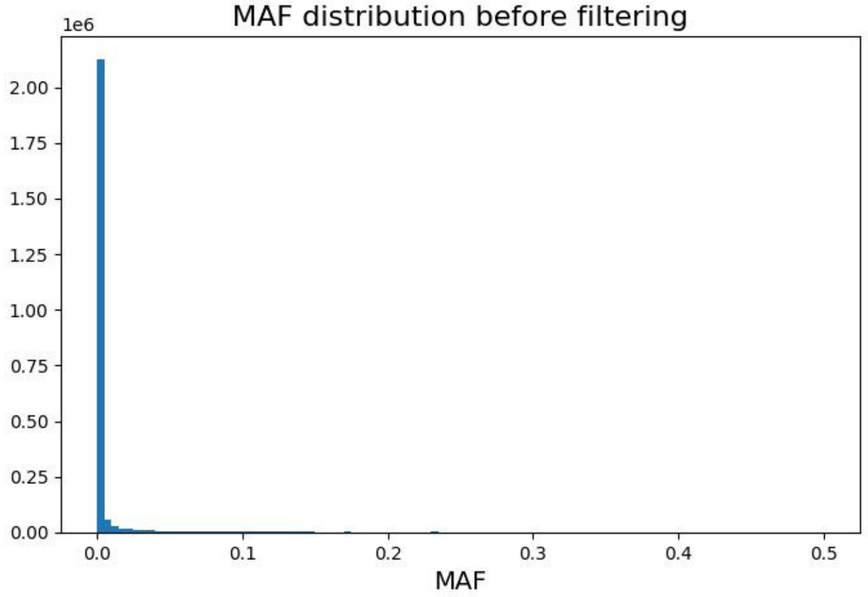
BUT : reconstruction des haplotypes parentaux les plus probables à partir des SNPs présents et du taux de recombinaison



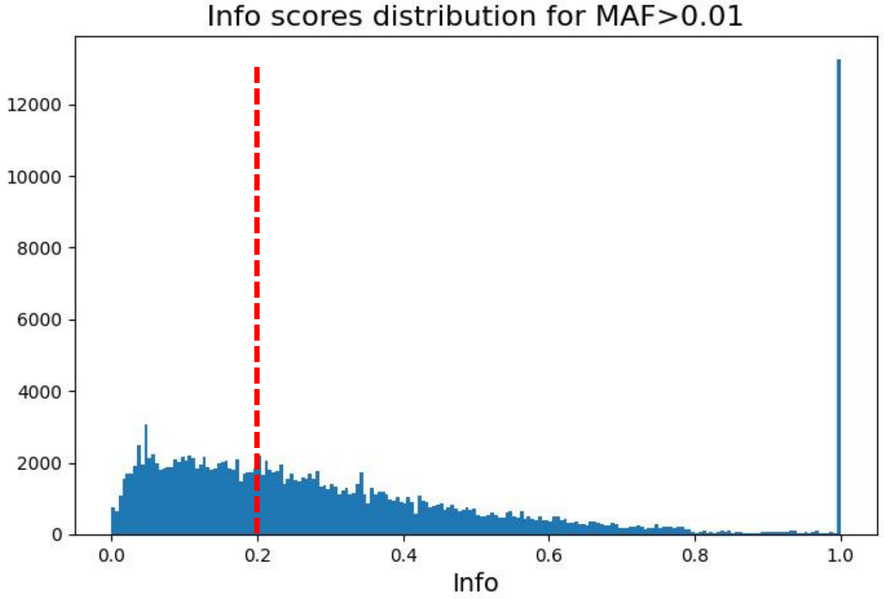
RESULTATS - Imputation QC

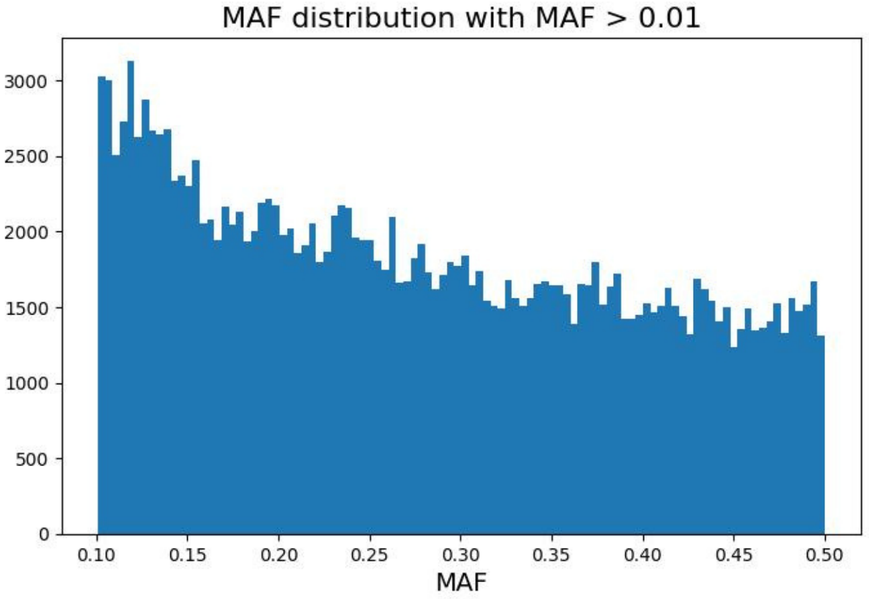
Chromosome 16





2 529 729 SNPs





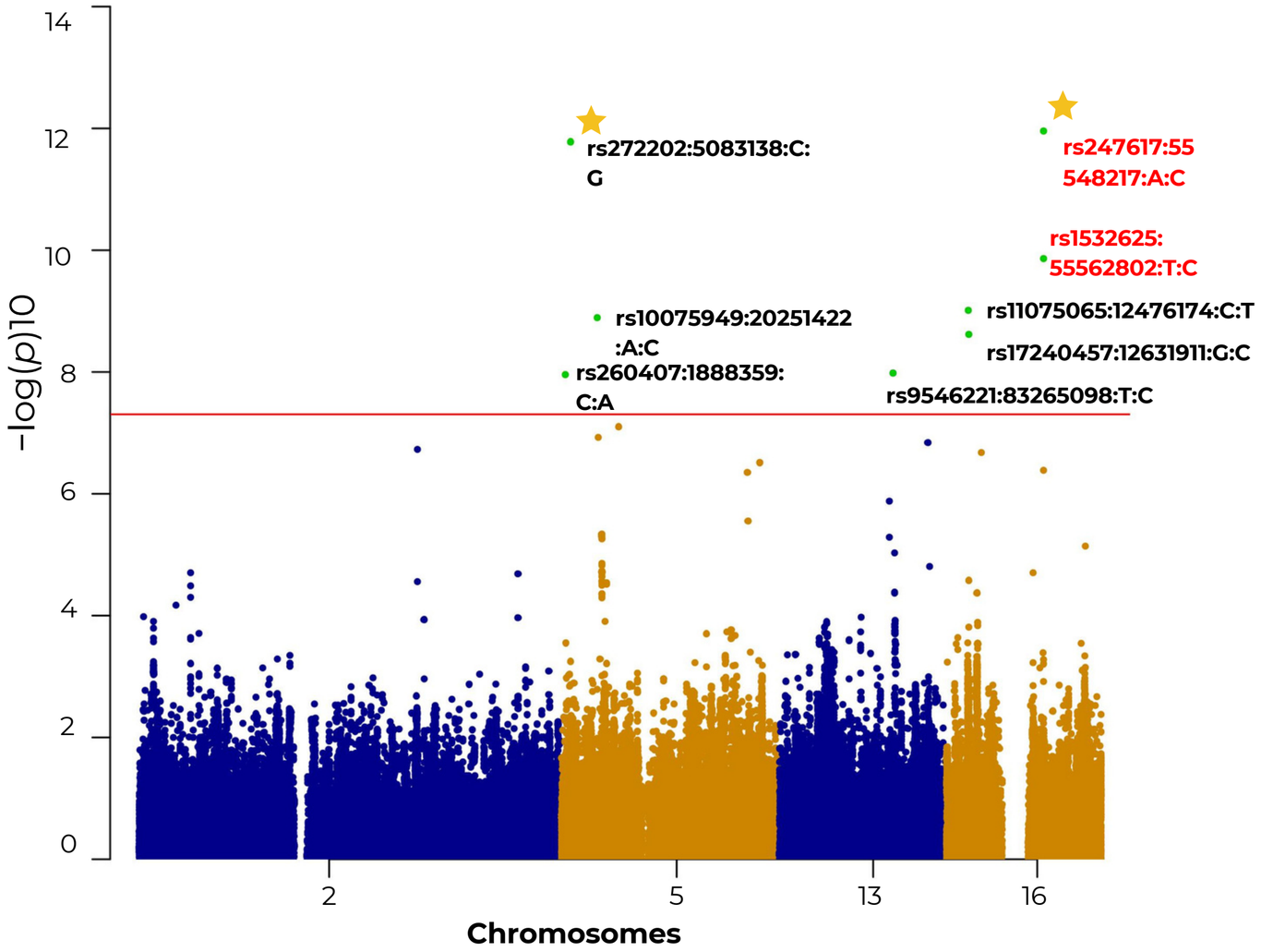
Number of imputations with MAF > 0.01 and info > 0.2: 108 725 SNP

RESULTATS - Imputation QC

		Imputation		QC control
Chr 2	41 805 SNPs	→	6 479 005 SNPs	→ 244 476 SNPs
Chr 5	32 968 SNPs	→	5 322 569 SNPs	→ 269 502 SNPs
Chr 13	19 373 SNPs	→	2 862 754 SNPs	→ 158 831 SNPs
Chr 16	15 120 SNPs	→	2 529 729 SNPs	→ 108 725 SNPs

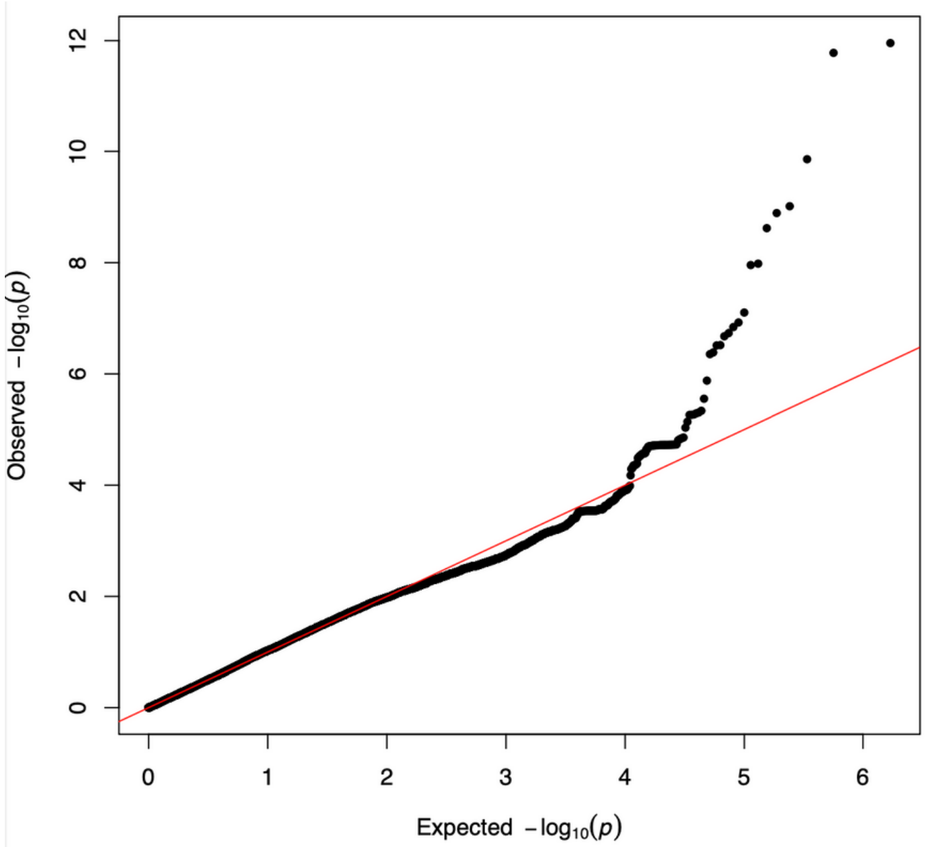
RESULTATS - Analyse QTL Apolipoproteine (ApoA1)

Manhattan plot for ApoA1 associations



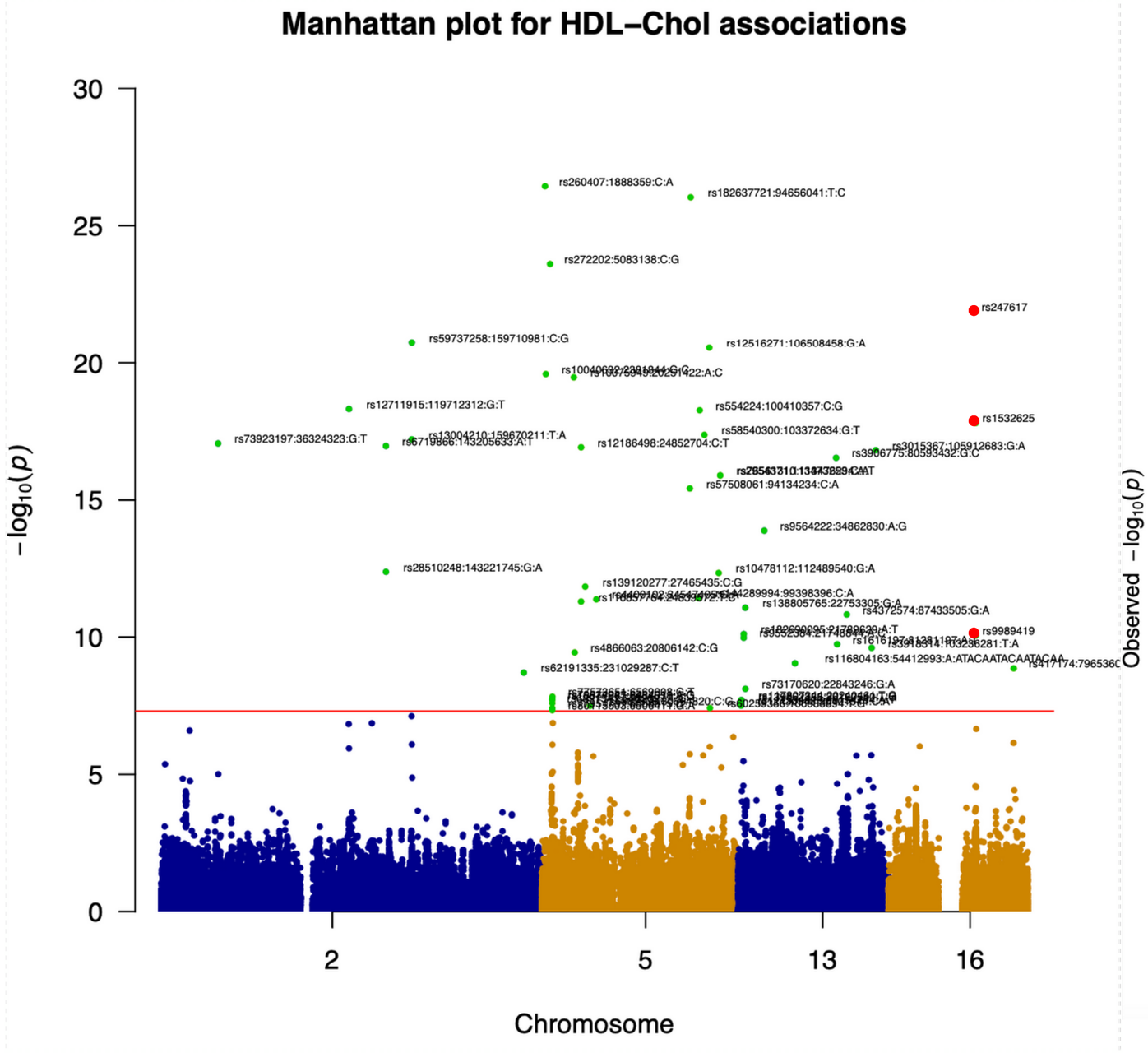
8 SNPs significatifs identifiés (en rouge SNPs génotypés) :

- chr5: 3 SNP
- chr13: 1 SNP
- chr16: 4 SNP
- Les plus significatifs ⭐



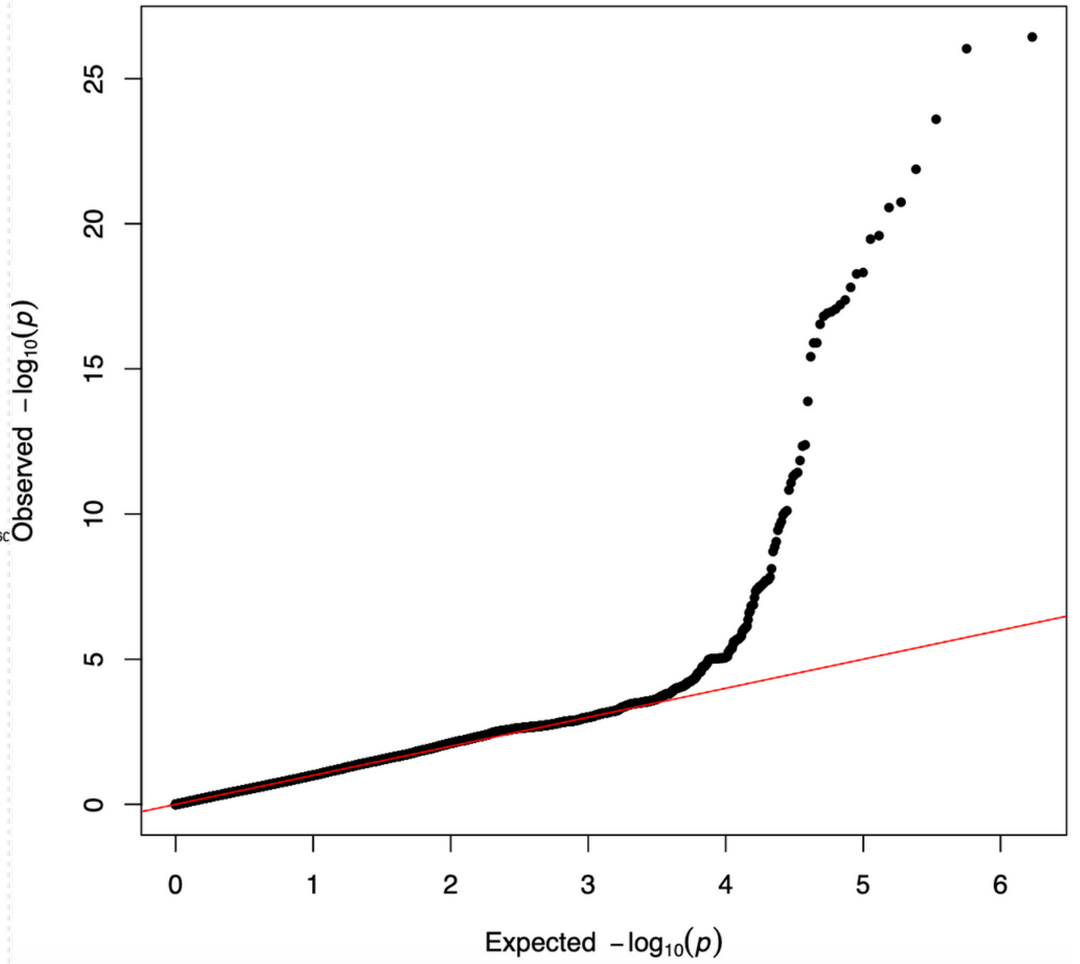
QQ plot for ApoA1 associations

RESULTATS - Analyse QTL HDL-Cholestérol



52 SNPs significatifs identifiés :

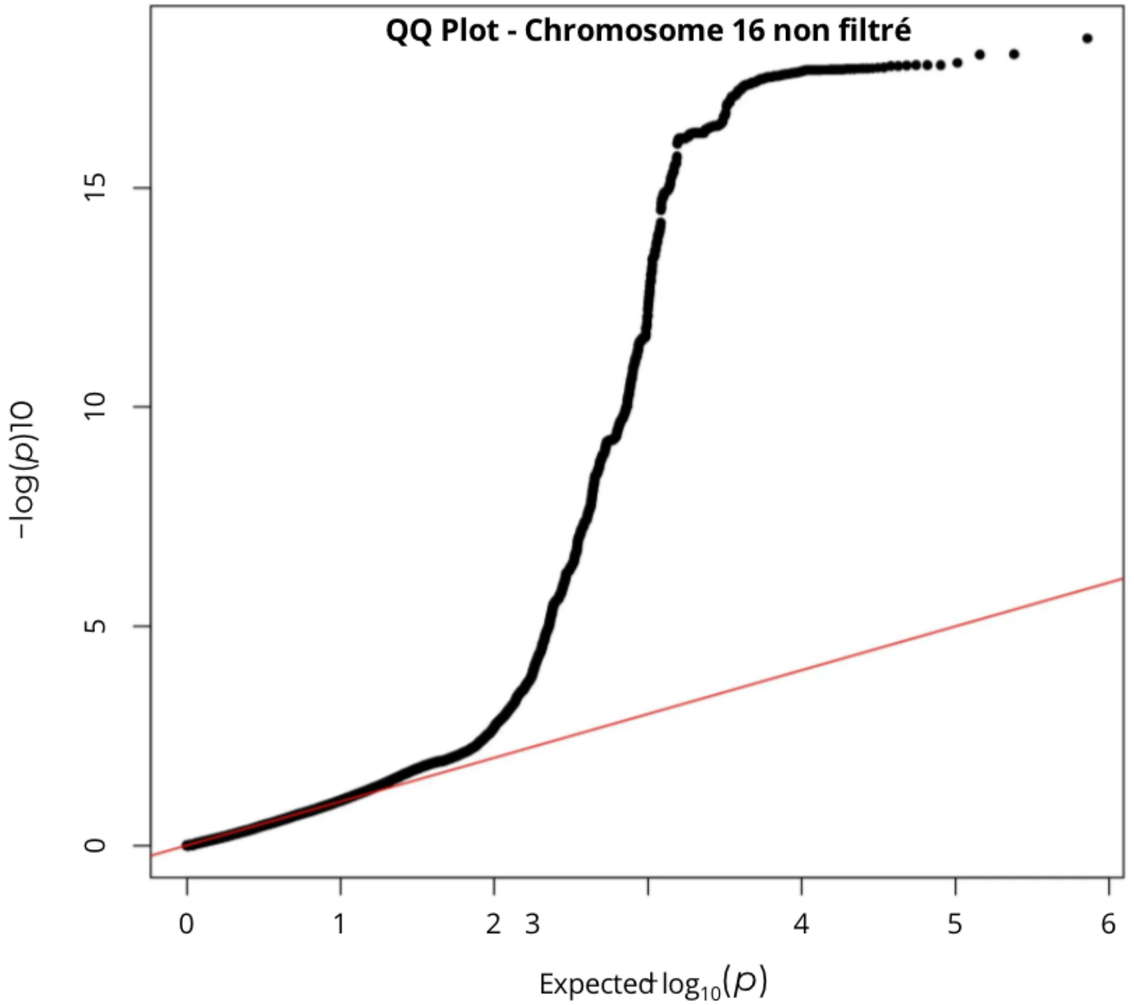
- 3 SNPs génotypés
- 49 imputés

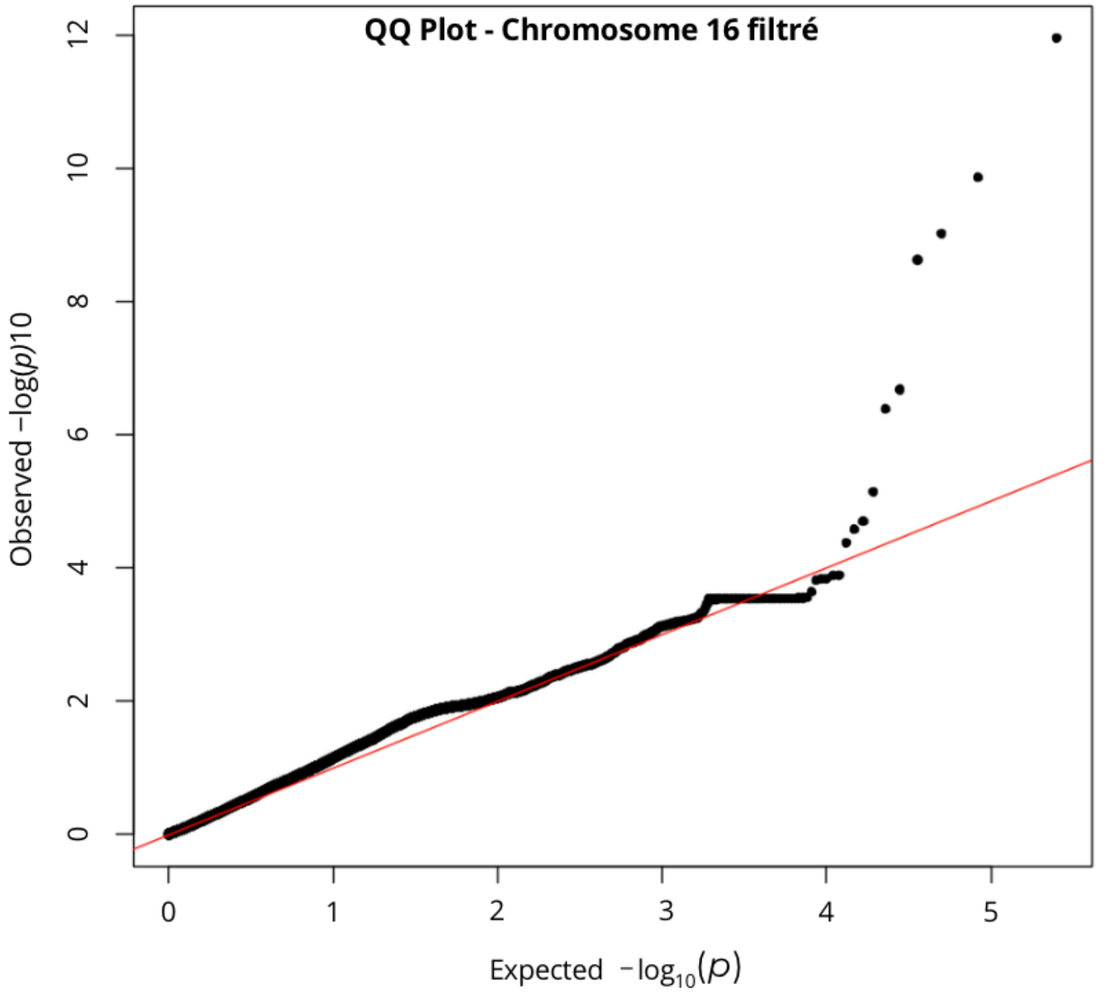


QQ plot for HDL associations

L’importance de filtrage post-imputation

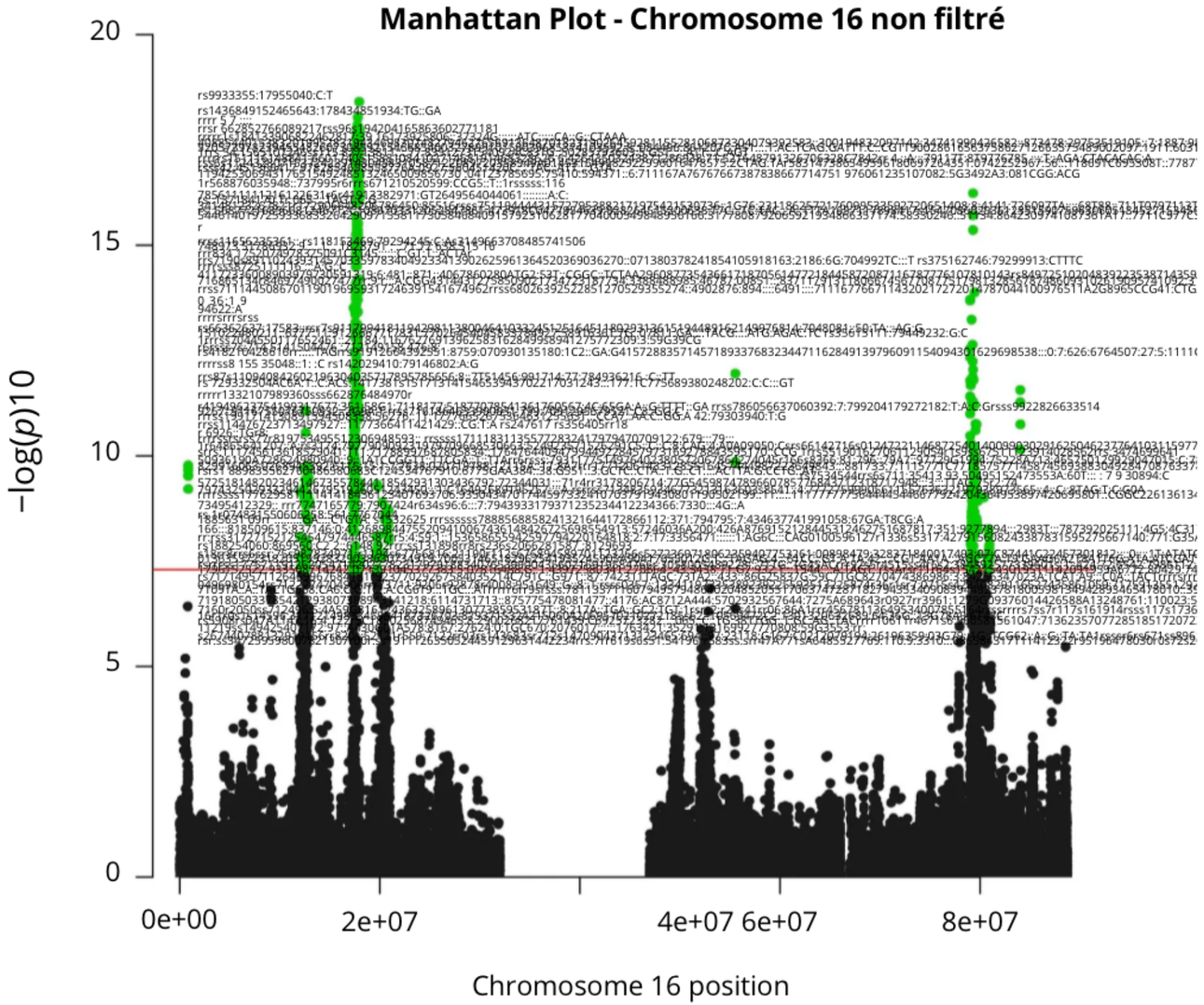
ApoA1

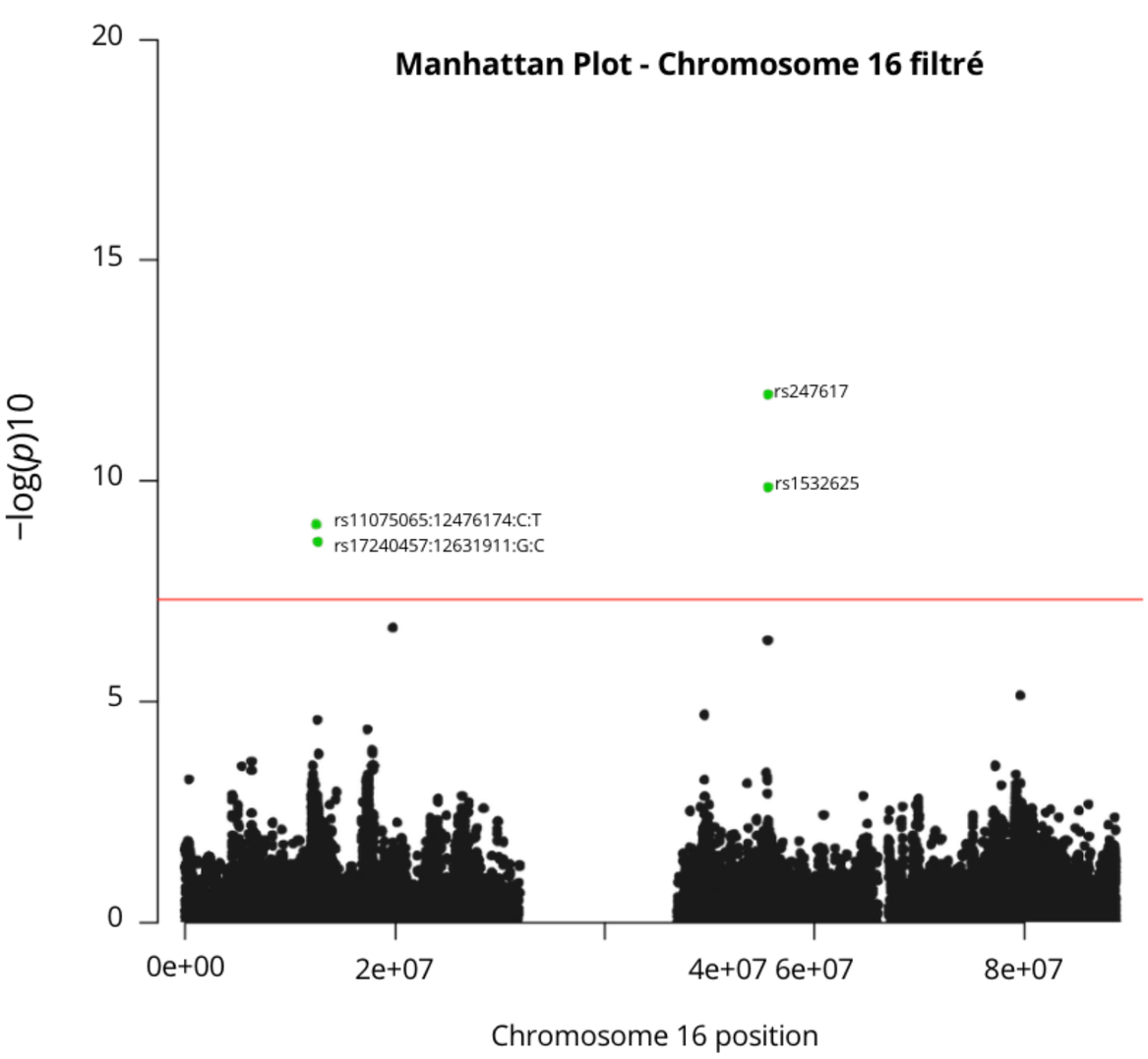




L’importance de filtrage post-imputation

ApoA1





CONCLUSION & DISCUSSION

- ApoA1 : Positions d'intérêt localisées sur les chromosomes 5, 13 et 16
  - Regional plot
  - HaploView

- HDL_cholésterol : Filtrage

- Différents paramètres/Benchmark : Temps computationnel/mémoire
  - Autres programmes de phasage/imputation?

- Comparer nos résultats à ceux sans imputation

- Elargir analyse au génome entier

- Nombreux SNPs imputés (x4), dont plusieurs significativement associés

Références

- “Genotype imputation for genome-wide association studies”, Jonathan Marchini* and Bryan Howie‡ (2010)
- “A guide to genome-wide association analysis and post-analytic interrogation”, Eric Reed,Andrea S. Foulkesa*†(2015)
- “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis”, Andries T. Marees, et al. (2017)
- “Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray”, Jonathan R. I. Coleman, et al. (2016)
- “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies”, Bryan N. Howie, et al. (2009)
- “Shape-IT: new rapid and accurate algorithm for haplotype inference”, Olivier Delaneau, Cédric Coulonges and Jean-François Zagury (2008)

ANNEXE

INTRODUCTION - GWAS

PRINCIPE

Population d'individus

âge, sexe, phénotype (malade sain ou données d'expression d'un gène)...

⬇

Génotypage des SNPs

⬇

QUALITY CHECK

⬇

+ IMPUTATION : inférer des variants manquants à partir des variants observés

⬇

Étude d'association

Étude statistique d'associations entre les variants génétiques et phénotype associé