Practical aspects in GWAS analysis

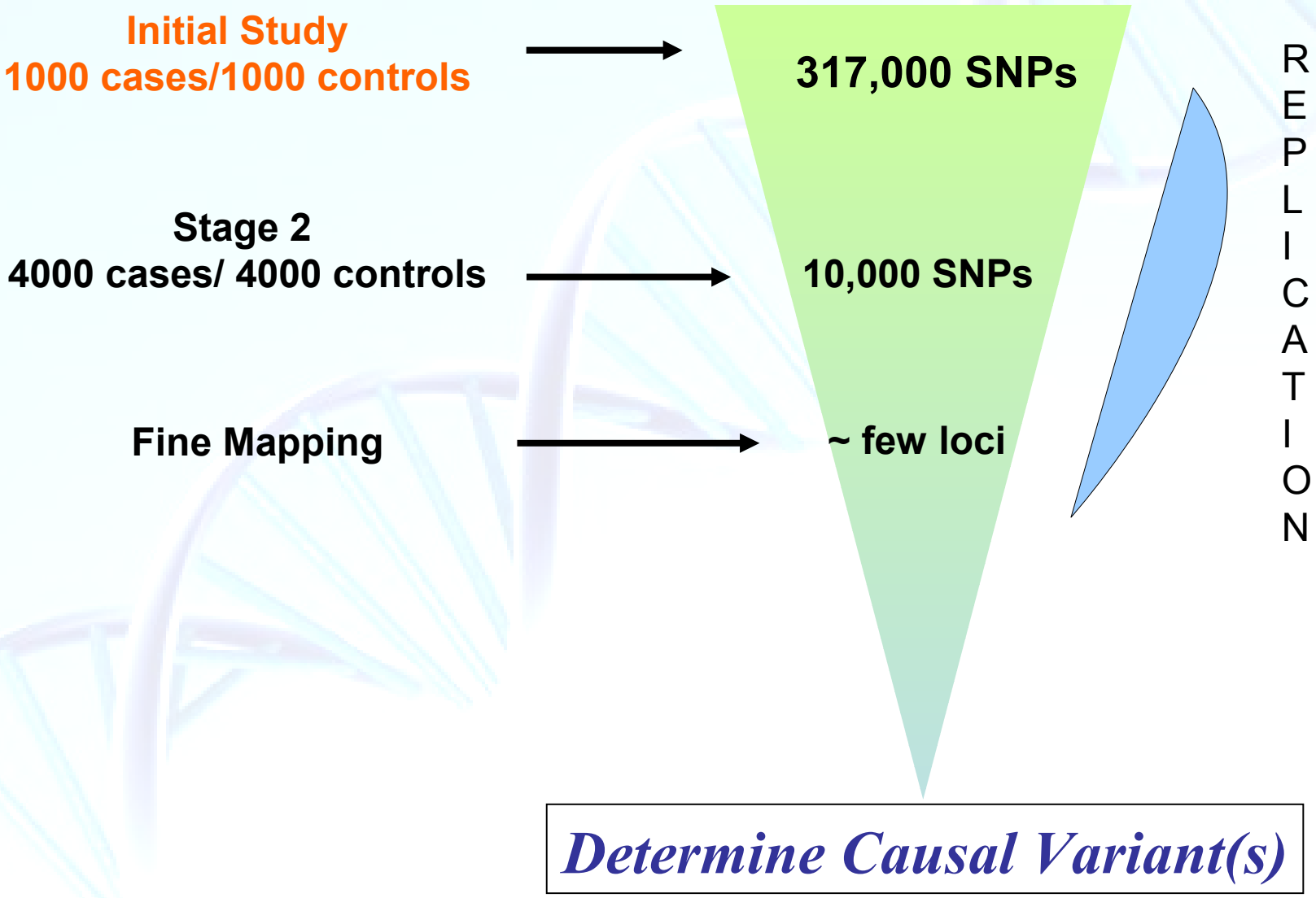
Genome Wide Association Studies

Cedric Coulonges

Chaire de bioinformatique

Conservatoire National des Arts et Metiers - Paris

Introduction



Contents

1) Genotyping chips

2) Data management

3) Stratification problem

4) Statistical analysis

5) Imputation

6) Meta analysis

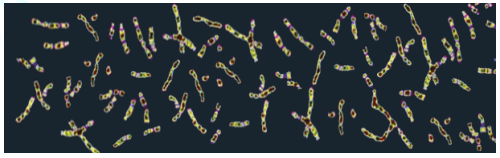
7) Example on an AIDS cohort (CNAM)

2002

Phase I: 1 million of SNPs 270 people (30 Trios Yoruba, 45 unrelated Japanese, 45 unrelated Chinese, 30 trios European ancestry)

Phase II: approximately 2,5 million of SNPs (same populations)

Phase III: 1115 people from 11 populations collecting from Illumina 1M and Affymetrix 6.0 merged



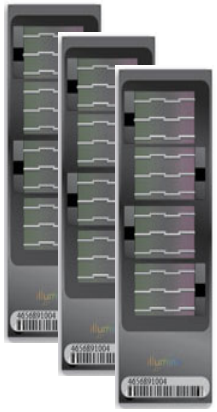
Future: 1000 Genomes Project

1. Genotyping chips

illumina

- Beadchips: 300 k, 370 k, 510 k, 660 k, and 1 mi assays. The DNA requirements are low, about 300 ng

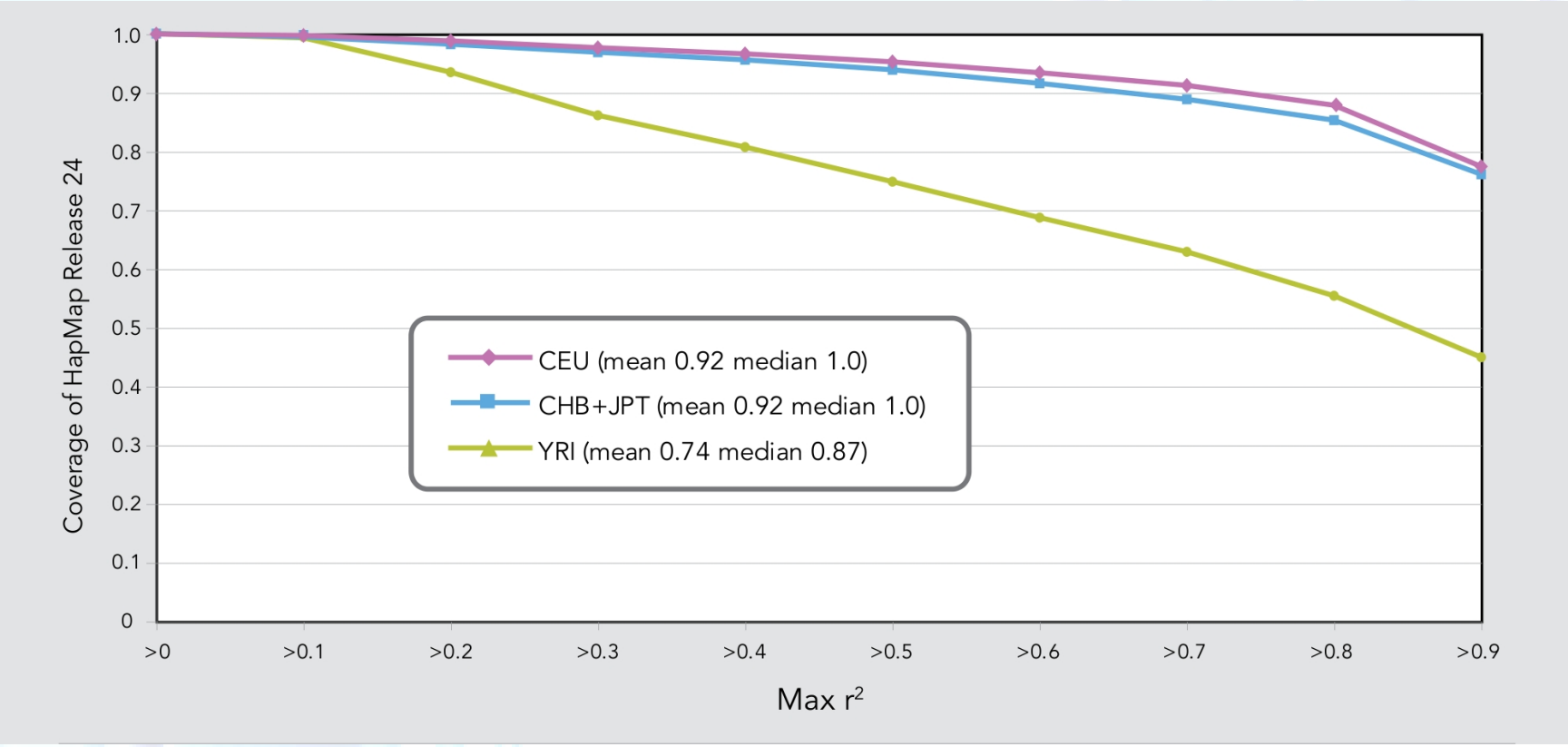
- The SNP selection strategy primarily on the HapMap project for a good coverage of SNP diversity



Example for 660k SNPs chip:
$r^2 > 0.8$ on genes, $r^2 > 0.7$ for others SNPs

- The best chip is actually the HumanOmni1-Quad:1,199,187 SNPs (median value of 1.5 kb for intermarker distance)

1. Genotyping chips (2)



Human660W-Quad Beadchip coverage
(hapmap >2.3 M SNPs)

1. Genotyping chips (3)

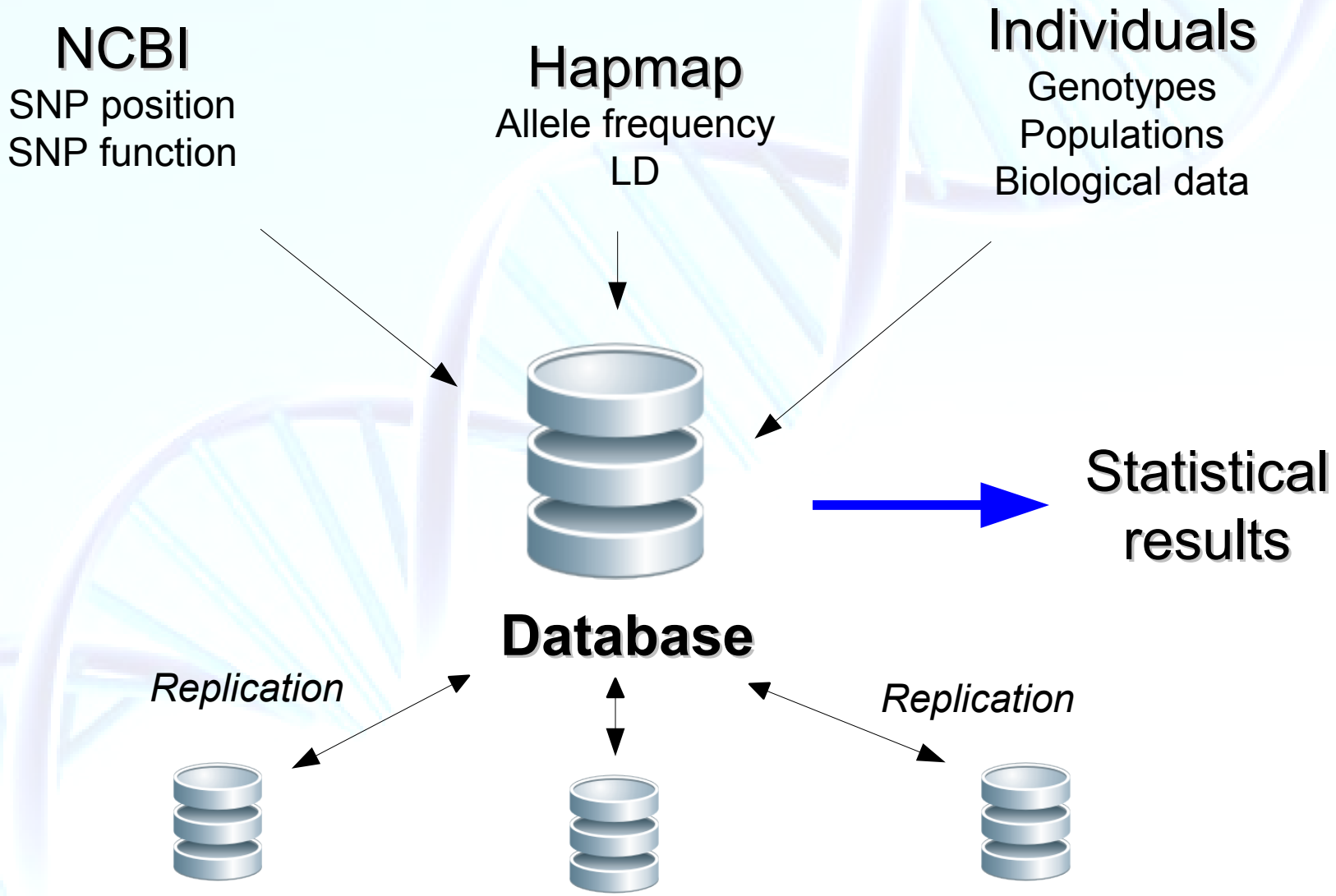
-The procedure allows the detection of 10,000–2,000,000 SNPs.

- The DNA requirements are low, about 300 ng

- SNPs were selected and tiled on arrays based on accuracy, and linkage disequilibrium analysis in three populations across the genome.

- The best chip is 6.0 chip: 1.8 million markers (median 1.3 kb for interSNP distance)

2. Data Management



2. Data Management (2)

Quality control on raw genotyping data

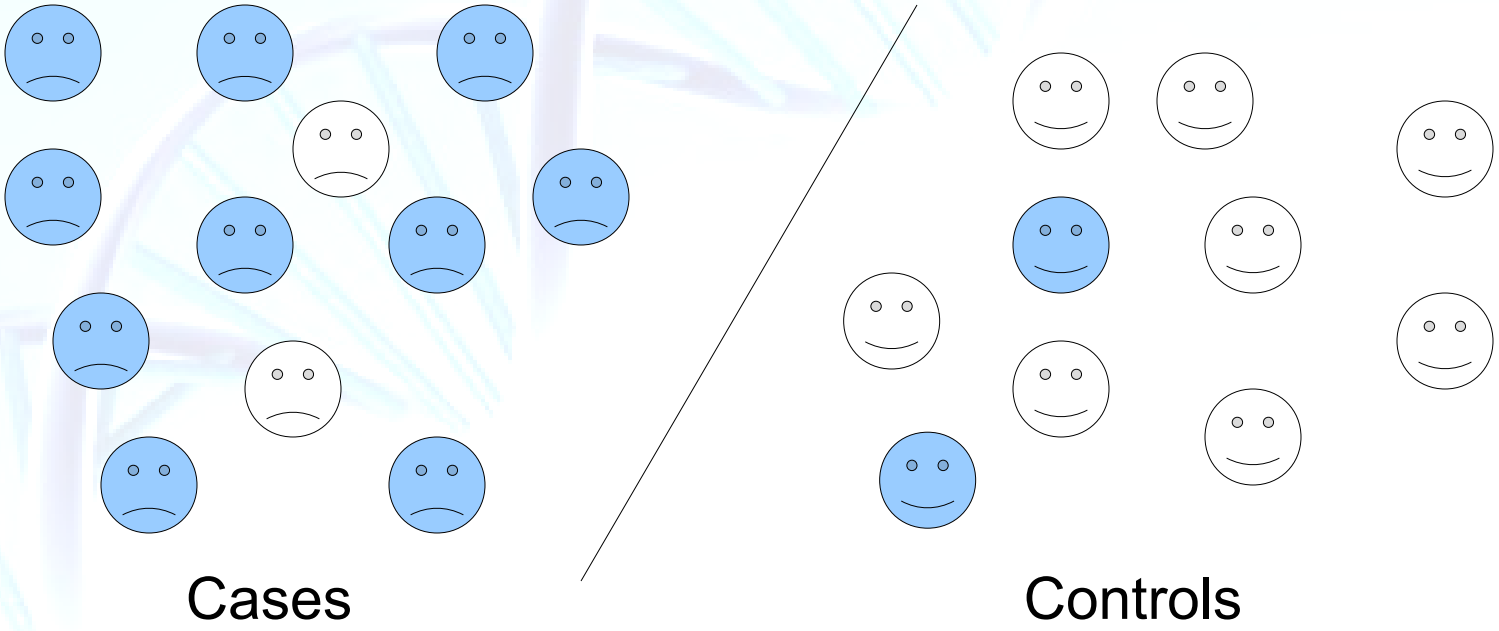
- Call rate (missing rate per individual)
  > 98% (2%) ?
- Call freq (missing rate per SNP)
  > 98% (2%) ?
- Minor allele frequency
  > 1% ?
- Hardy-Weinberg equilibrium (HW exact test)
  P < 0.001 ?

3. Population structure

The presence of subpopulations is possible because of admixture or different ancestry

→ Different allelic frequencies

The association found is not associated with disease but is due to population substructure



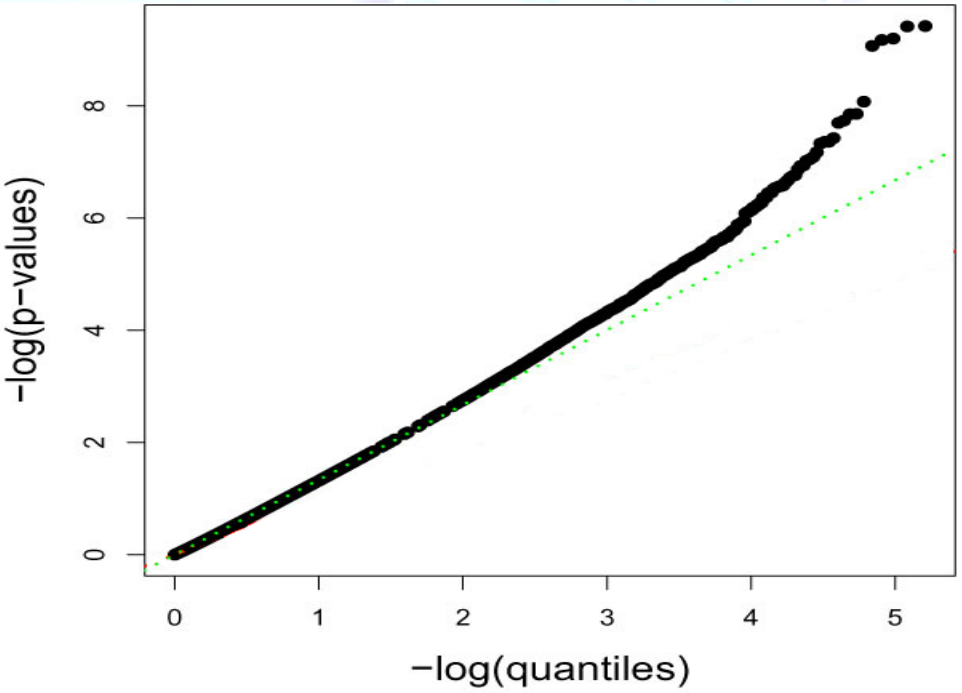
3. Population structure (2)

Genomic Control (inflation factor)

Devlin, B. and Roeder, K. (1999). Genomic control for association studies, Biometrics 55(4): 997–1004

For n independent SNPs, the statistics is inflated by a factor λ
It can be estimated by:
$\lambda = \text{median}(x^1, x^2 \ldots, x^n)/0.456$ where x is the 2x2 chi2 test

λ = 1.33

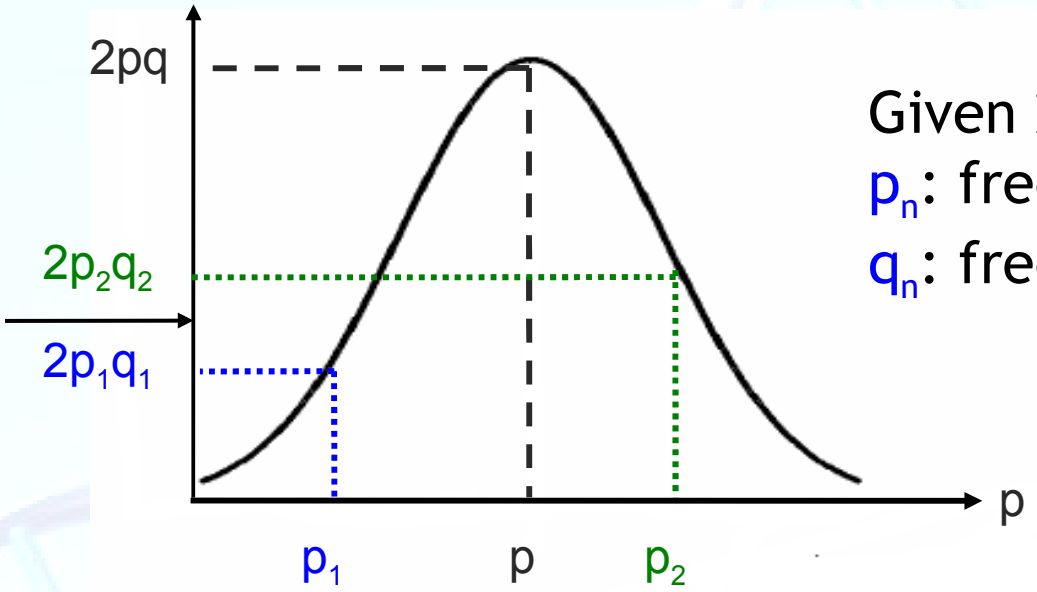


If λ>1.10, we consider that the stratification is very important

3. Population structure (3)

Structure software

- Stratification deviates Hardy Weinberg equilibrium



Given 2 populations:

p_n: frequency of allele A in pop n

q_n: frequency of allele a in pop n

- An “artificial” linkage disequilibrium appears between supposed independent loci

3. Population structure (4)

Structure software (2)

Selection of neutral and independent SNPs (>50)

Hypothesis: K subpopulations

↓

MCMC

↓

Likelihood score

3. Population structure (5)

Eigenstrat

Currently, the most-widely used software for stratification correction

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association. Nature Genetics 38:904-909

Principal Components Analysis (PCA):
Eigenvectors which distinguish subpopulations can be used as covariates

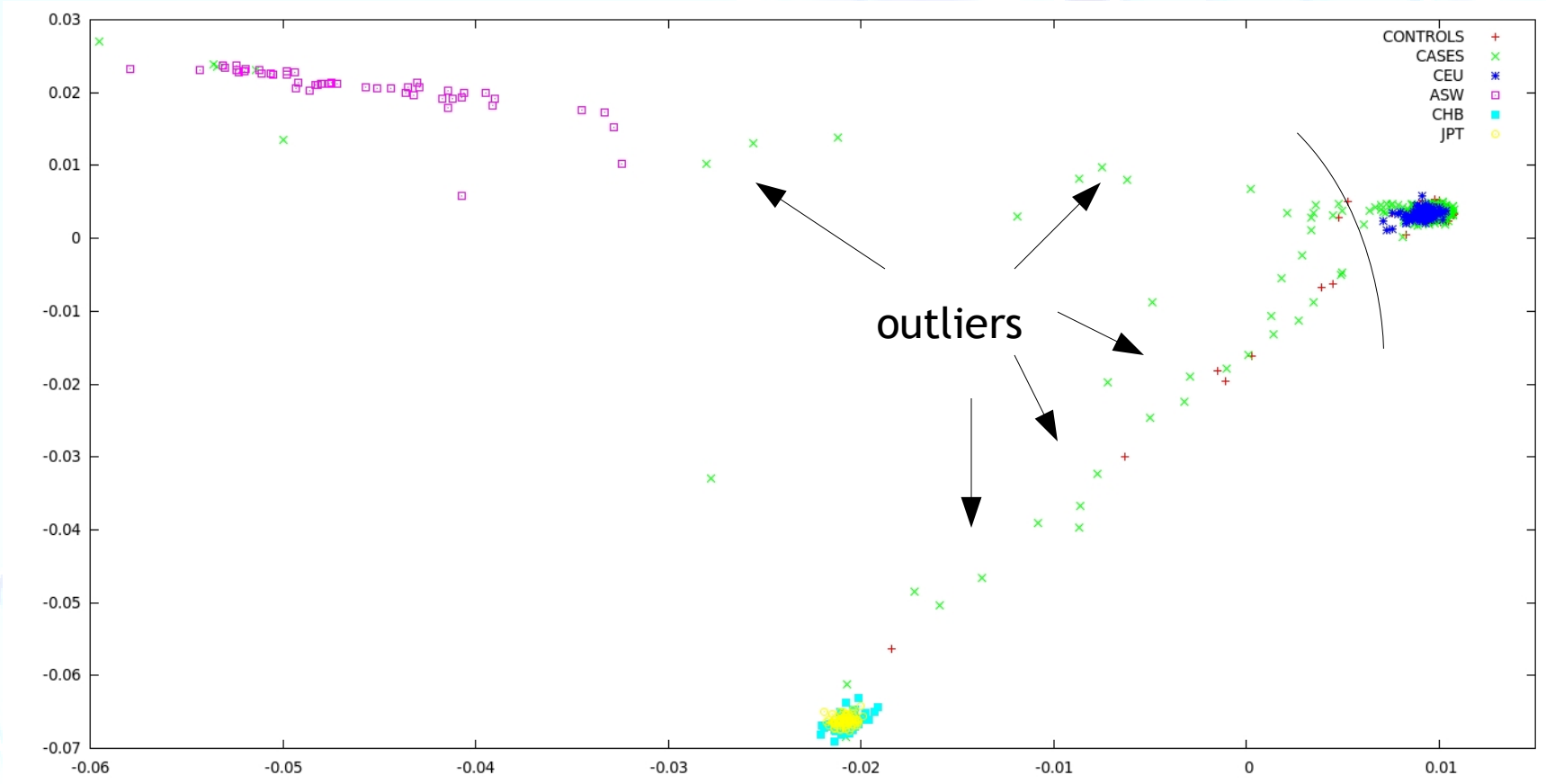
Outliers can be removed for further analysis

The approach is powerful as well as fast

Thousands of markers

3. Population structure (6)

ASW:African ancestry, CEU: European ancestry,
CHB:Han Chinese in Beijing, China, JPT:Japanese



4. Statistical analysis

case/control association test

- Odds of allele 1 in disease

(a/(a+b))/(b/(a+b)) = a/b = e

- Similarly odds of allele 1 in healthy = c/d = f
- Odds ratio (OR) of allele 1 in disease vs healthy = e/f

	#Allele1	#Allele2
Disease	a	b
Healthy	c	d

Chi square statistics

Alternative: Fisher's exact test

4. Statistical analysis (2)

- Unbiased Fisher Test

A Fast, Unbiased and Exact Allelic Test for case-control association studies. Guedj, Wojcik et al. Human Heredity. 2006. 61: 210-221

- Cochran-Armitage trend test

does not assume Hardy-Weinberg equilibrium, as the individual is the unit of analysis (as permutation test)

- Genotypic (2 df) test

If A is the major allele and a is the minor:

(AA) vs (Aa) vs (aa)

- Dominant/recessive gene action (1df) test

(AA,Aa) vs (aa)

4. Statistical analysis (3)

Logistic regression

- Additive, dominant, recessive or genotypic model

Example: Additive model

$$Y = a + b.ADD + c_1.Cov_1 + c_2.Cov_2 + \ldots + c_n.Cov_n$$

With Cov_n is covariate,

Y is the phenotype (case or control)

Null hypothese is b=0

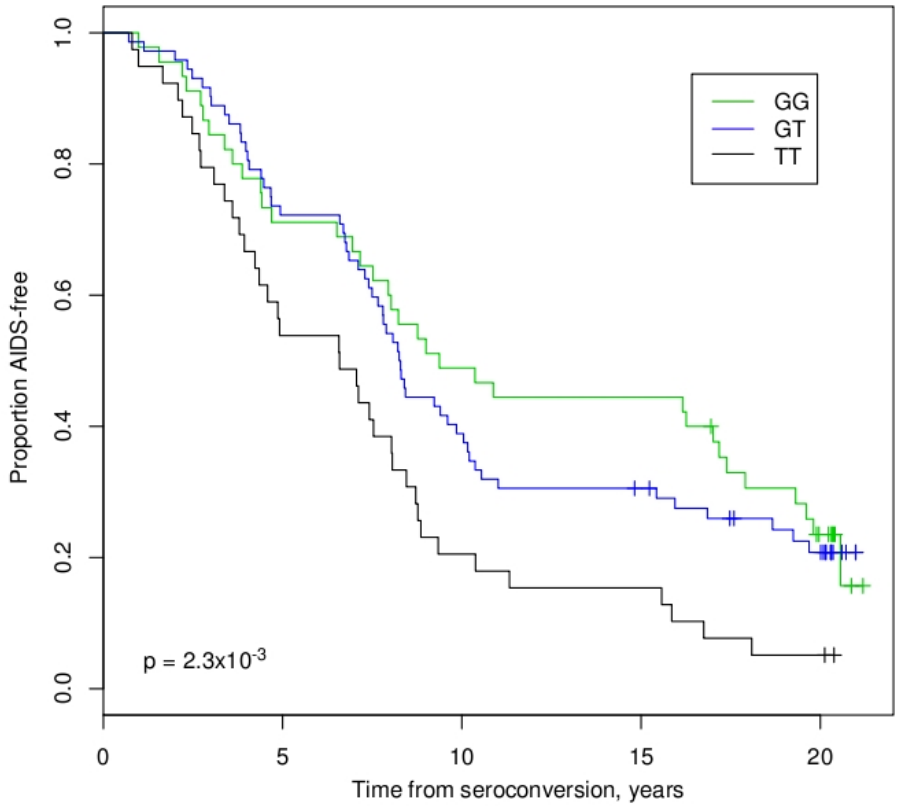
Linear regression

- The phenotype is a quantitative trait (case only)

→ Not accurate for low allele frequencies

4. Statistical analysis (4)

Cox Proportional-Hazards Regression for Survival Data



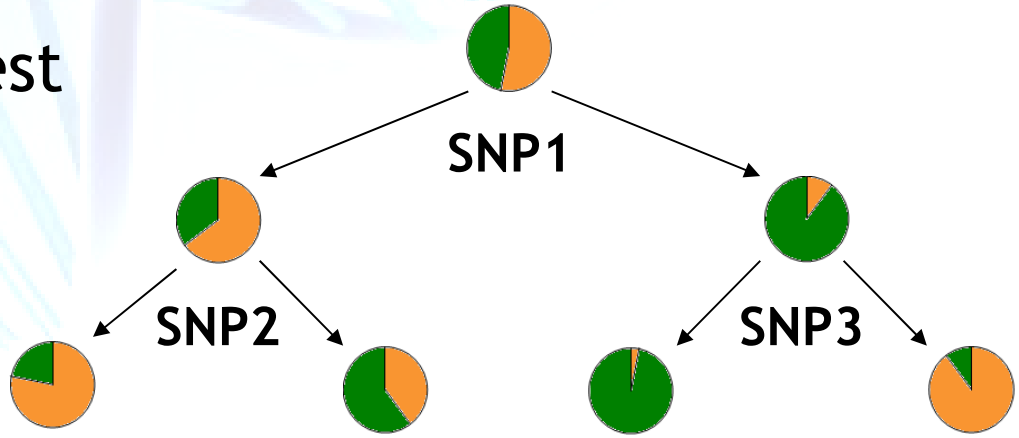
Kaplan–Meier plot

Survival time
Censoring
covariates

4. Statistical analysis (5)

Multimarkers tests

- Haplotypes approach
  → Neighboring SNPs
- Epistasis effect (combinations)
  → SNPs in different chromosomes
- Random forest



Combination allowing best population discrimination

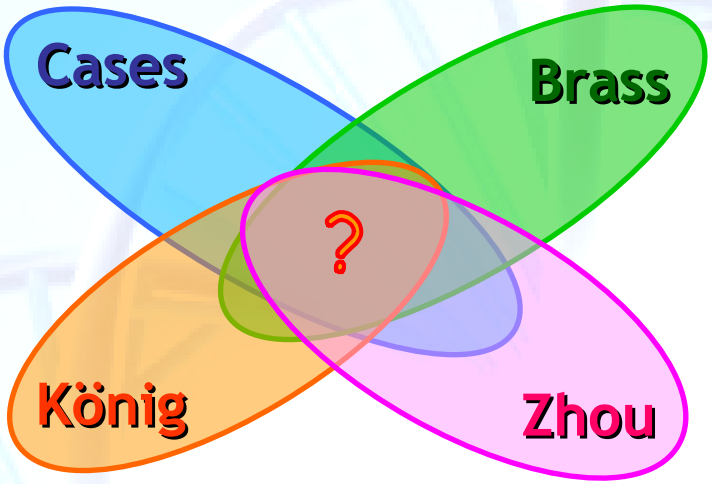
4. Statistical analysis (6)

Candidate gene approach

Attribute a weight for candidate genes of interest

3 siRNA genome-wide screenings in 2008 on AIDS: *Brass et al, Science, König et al, Cell, Zhou et al, Cell Host Microbe*

Approximately 250 genes / study (25000 genes in GWAS)



Calculate in each intersection if the pvalues in each gene are better than expected

4. Statistical analysis (7)

Multi testing problem

- Bonferroni correction very stringent

$Pvalue_c$ = pvalue x N (where N is the number of independent tests)

- False Discovery Rate (FDR) more powerful

Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". Journal of the Royal Statistical Society, Series B (Methodological) 57 (1): 125–133

Group of SNPs → a false positive rate

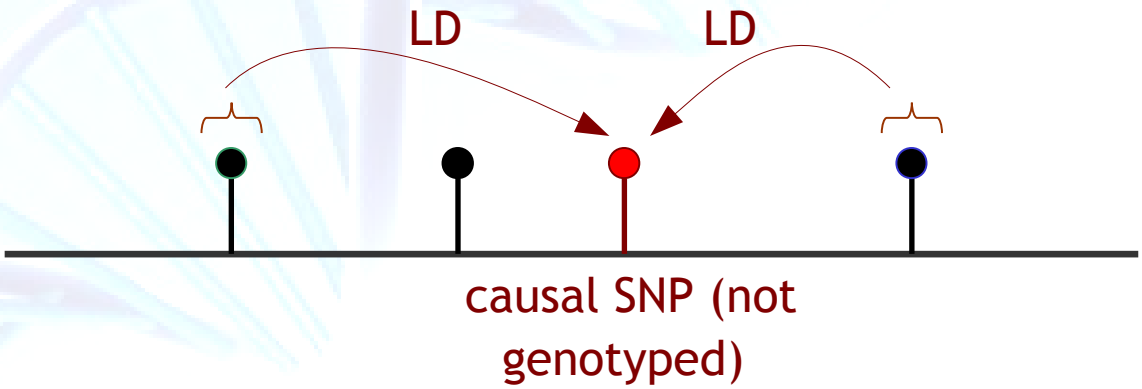
- Local FDR

kerfdr: A semi-parametric kernel-based approach to local FDR estimations. Guedj, Célisse, Robin and Nuel. BMC Bioinformatics (2009)

Each SNP → probability to be a false positive

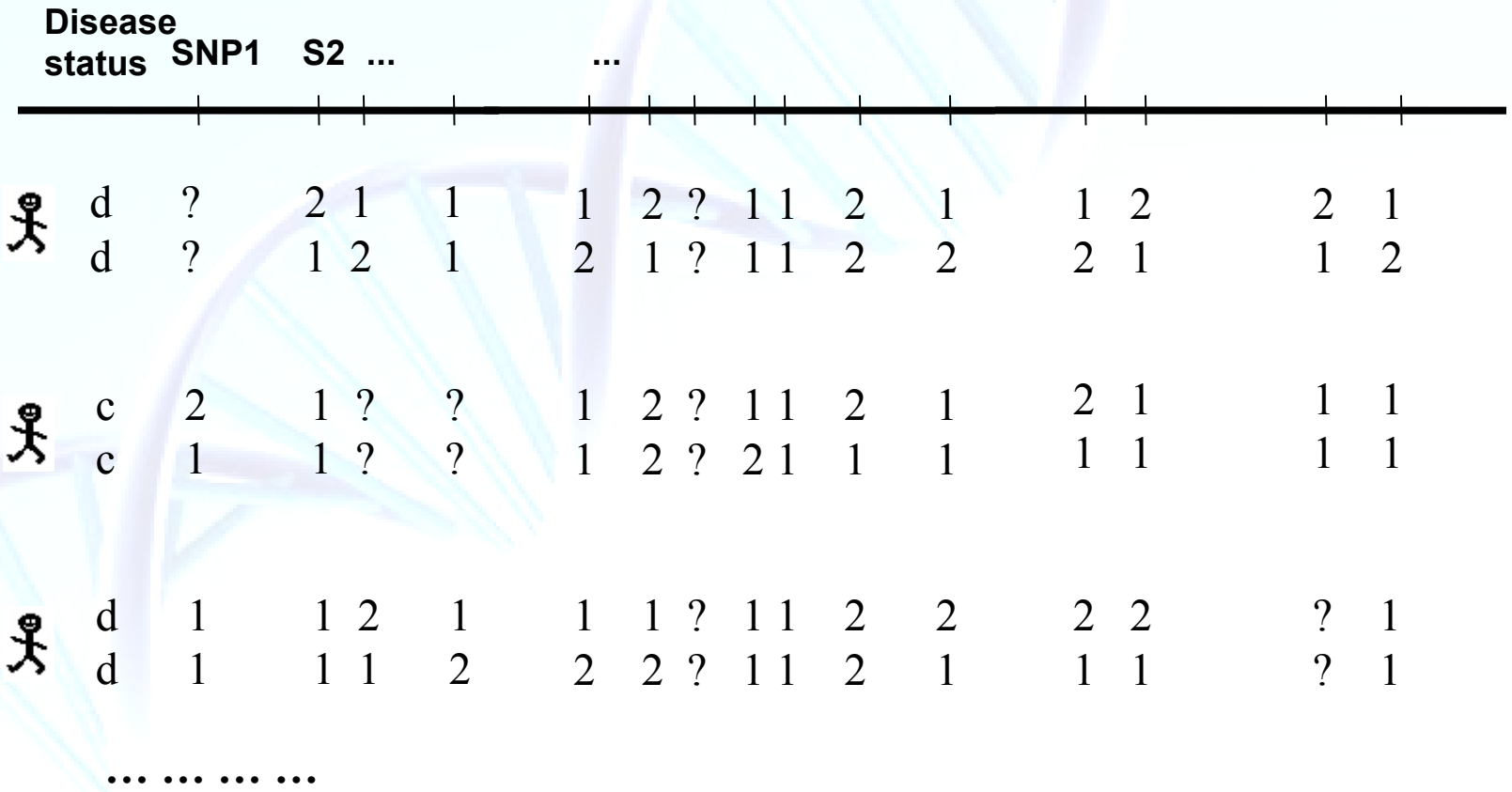
5. Imputation

Neighboring SNPs are often correlated (linkage disequilibrium)

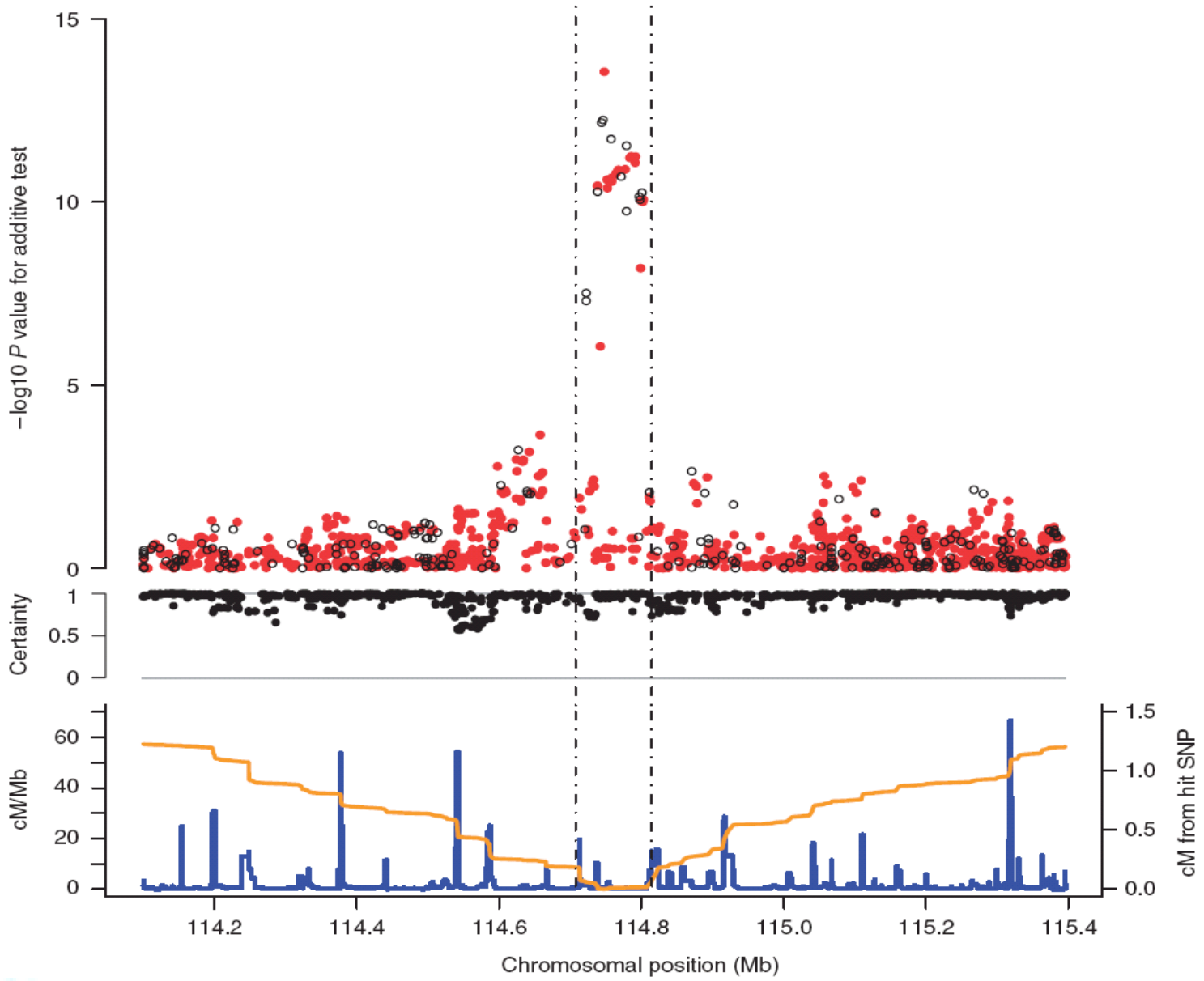


5. Imputation (2)

Haplotyping methods (EM, HMM) allow to impute missing data from huge resource panels such as the Hapmap project



5. Imputation (3)



5. Imputation (4)

- Increases power and may help for identification of the real causal locus

- Allows to compare multiple studies even if the genotyping platforms are different

- The accuracy of imputation can be evaluated in a similar fashion as haplotyping methods

One of the best software is actually Impute 2.1 (Marchini et al, 2007)

Other software available: Beagle, MACH...

6. Meta analysis

- Replications are difficult
  - diversity of studies (population, set-point, experiments, analysis…)
  - disease complexity
  - errors (experimentation, statistics…)
- P Fisher's combined $\chi^2 = -2\sum_{i=1}^{k} \ln(p_i)$ Df= 2k
- Z-score
  - Can introduce weight
  - More accurate when compared pvalue are asymetric
- Meta analysis is essential to publish

Ressources

PLINK *http://pngu.mgh.harvard.edu/~purcell/plink/*
Whole genome association analysis toolset

Genabel *http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/*
An R library for Genome-wide association analysis

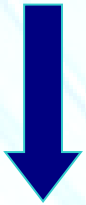
Probabel *http://mga.bionet.nsc.ru/~yurii/ABEL/*
Package for genome-wide association analysis of imputed data

Impute, SNPtest
http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html
Imputation software and analyse

Haploview *http://www.broadinstitute.org/mpg/haploview*
User-friendly interface for various analysis

7. Example of an AIDS cohort

Less than 10% of genetic factors on AIDS have been identified (O'Brien SJ et al - Nat Genet 2004)

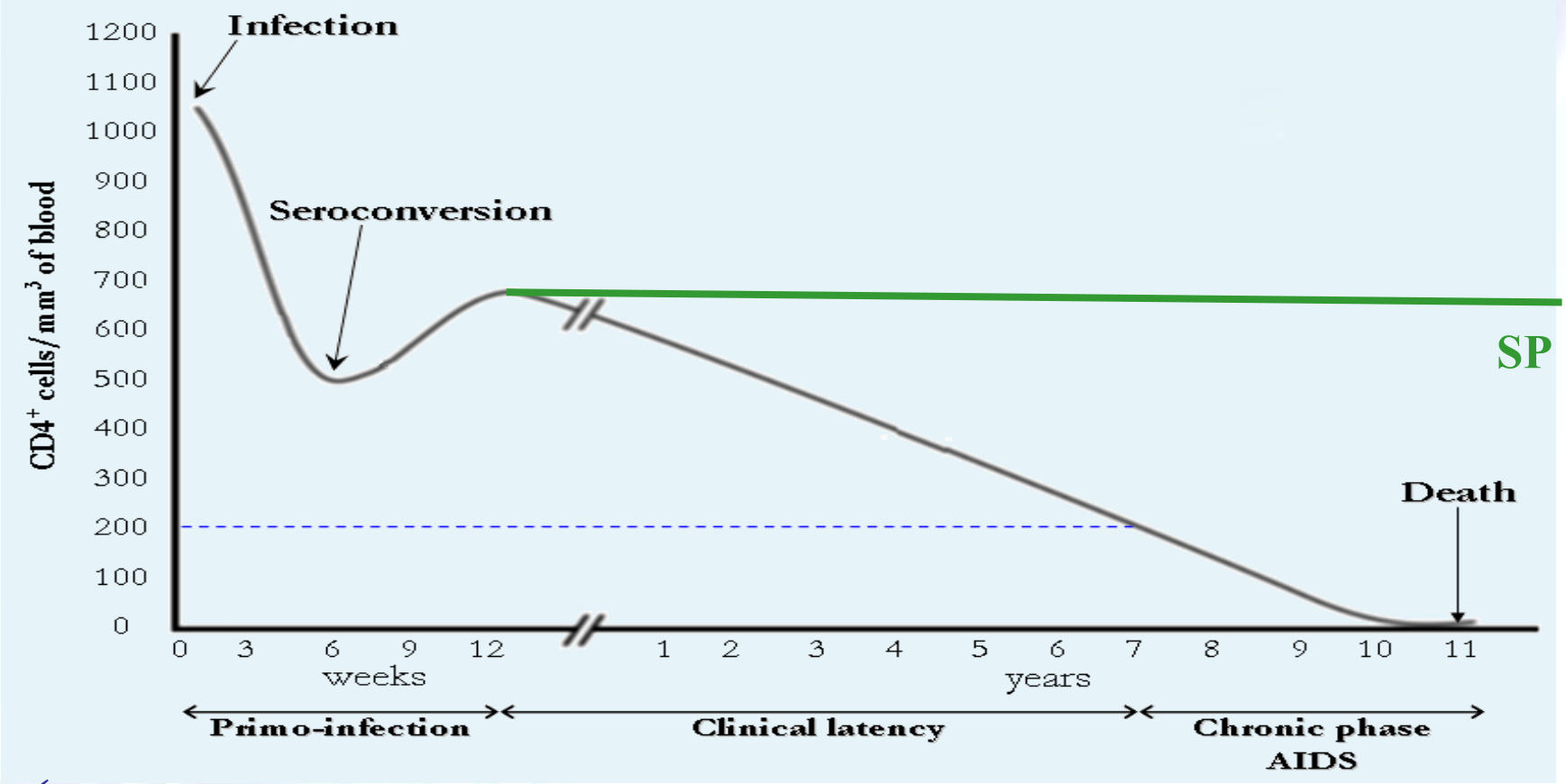


Genome Wide Association Study

Illumina 300K

2007

7. Example of an AIDS cohort (2)



✓ asymptomatic HIV-1 infection for more than 8 years with no treatment and with a CD4 T-cell count above 500 CD_4^+/mm^3 (Slow Progressor SP)

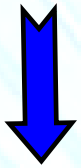
✓ ~5% of seropositive population ➡ **300 NP**

7. Example of an AIDS cohort (3)

The GRIV cohort : 300 slow progressors (SP)

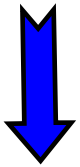
Enrichment in genetic factors involved in slow progression

VS 697 Controls



More powerful than usual seroconverter cohorts

Better Odds ratios



Sample adapted for genome wide

7. Example of an AIDS cohort (4)

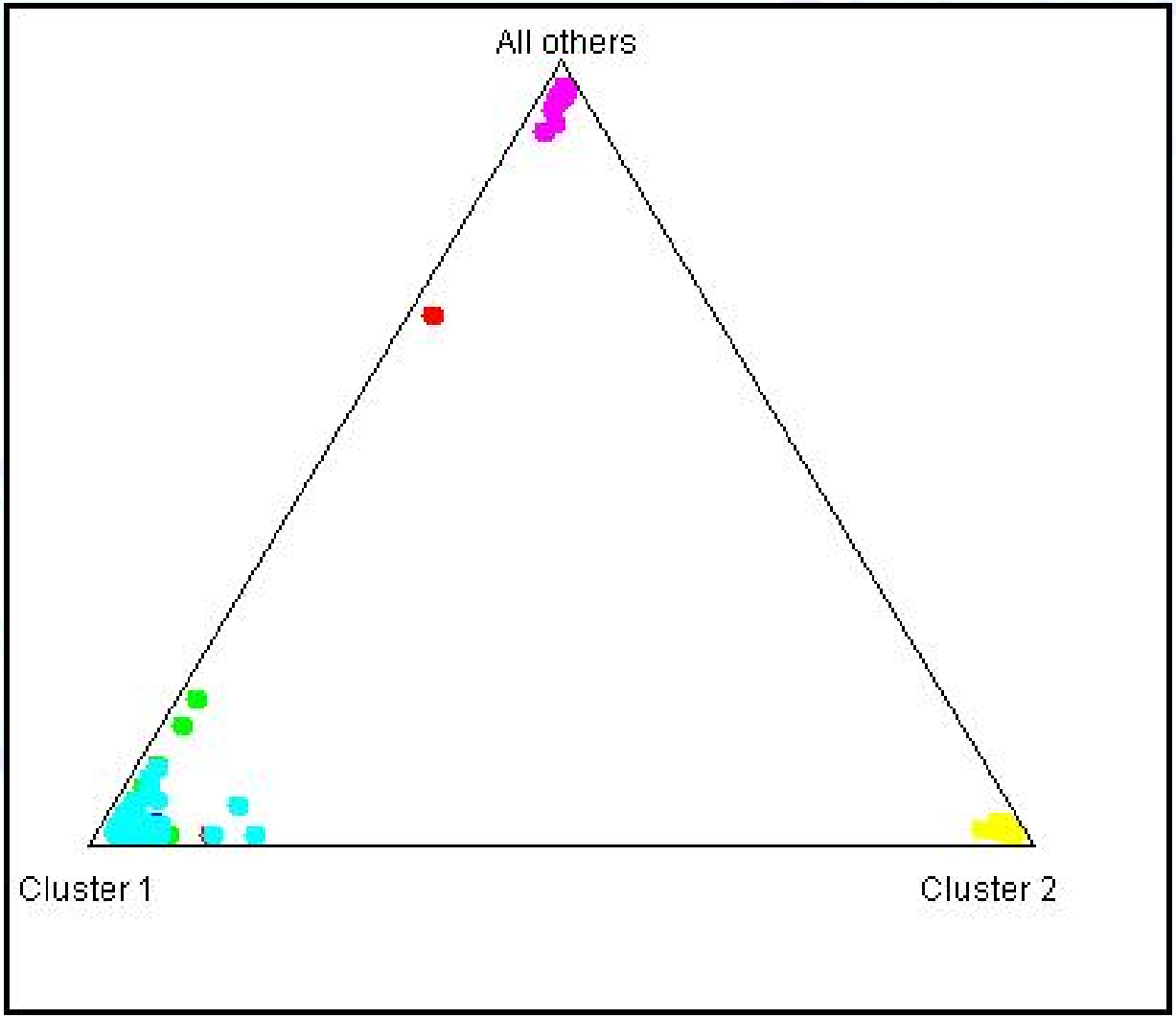
Structure / Stratification

Legend:
Red/green: Non progressors

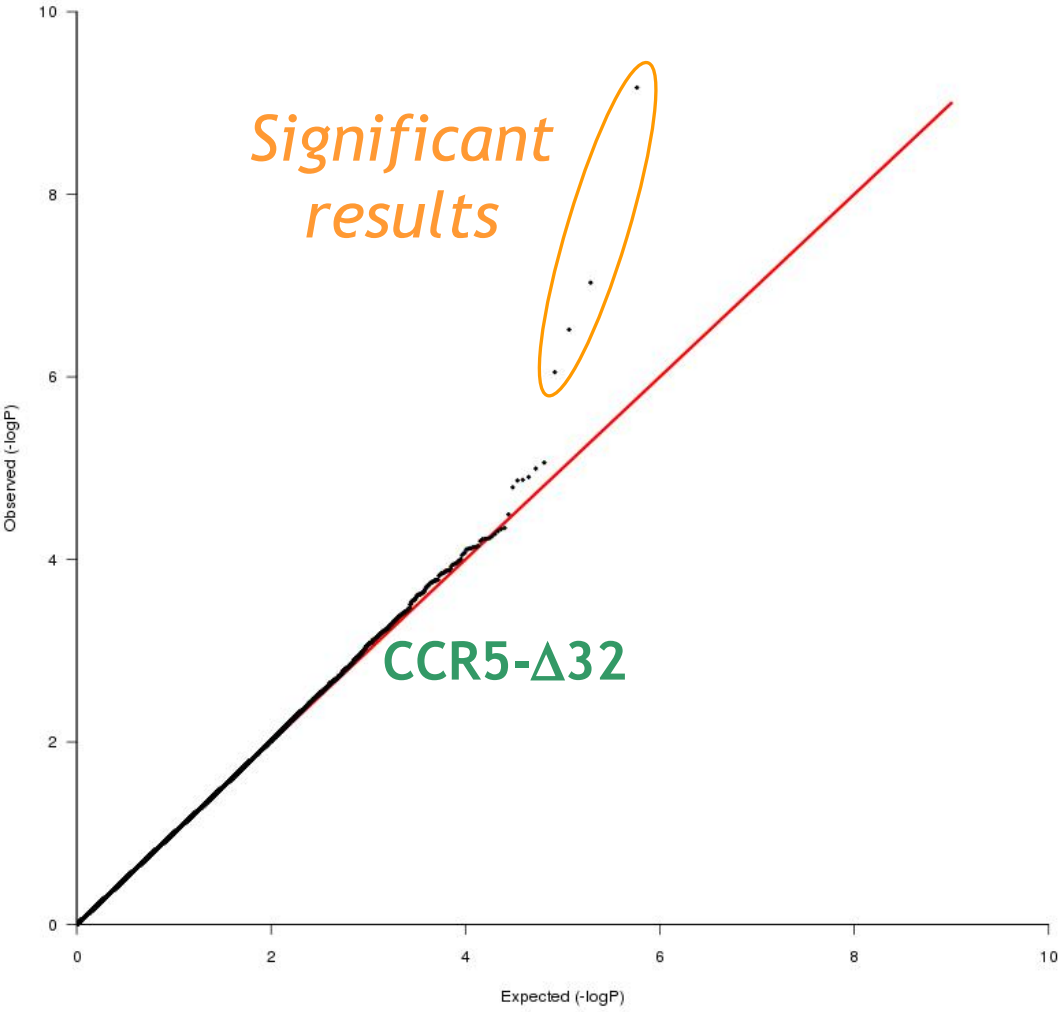
Blue: CEU

Yellow: Asian

Purple: African



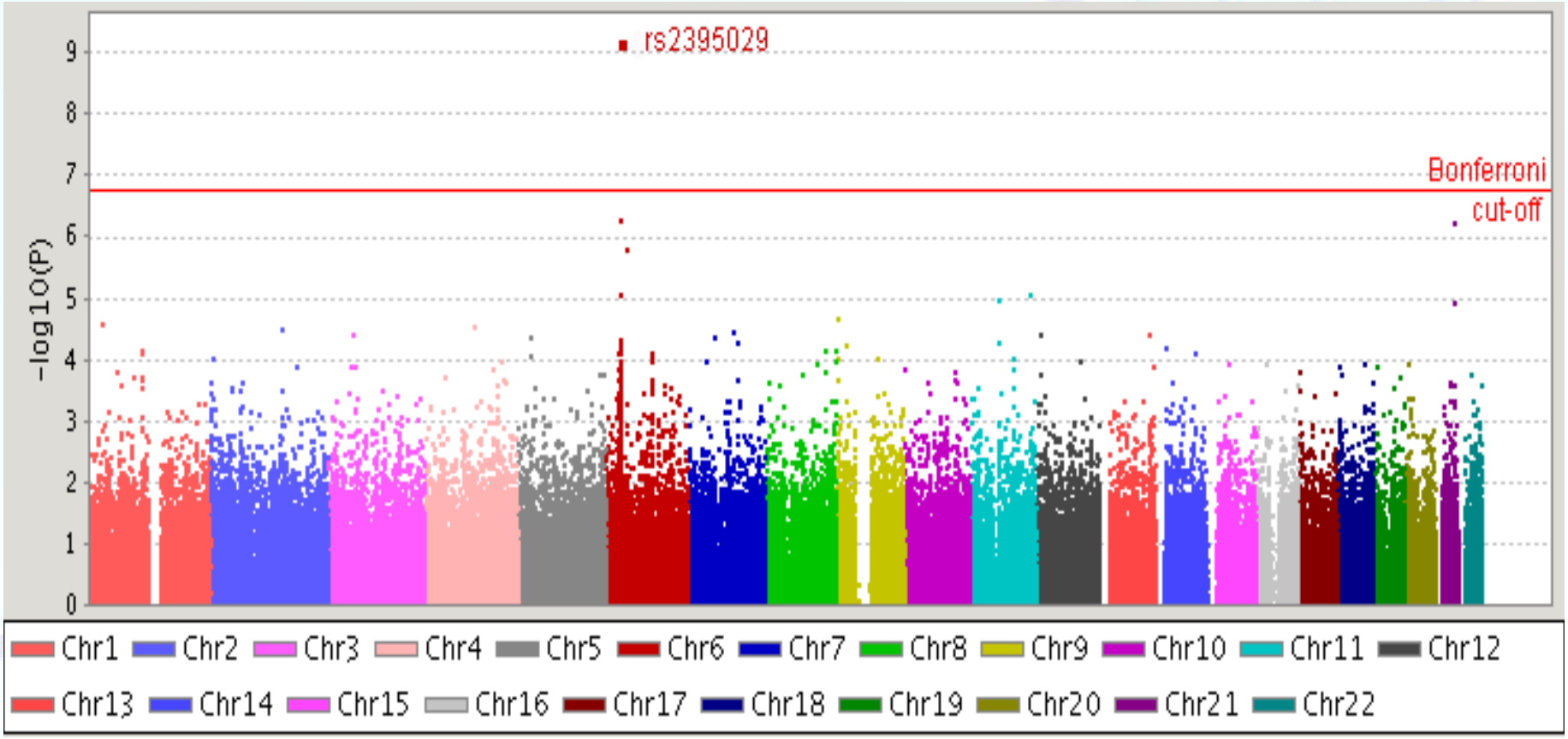
7. Example of an AIDS cohort (5)



Limou et al, Journal of Infectious Diseases (2008)

7. Example of an AIDS cohort (6)

Global results

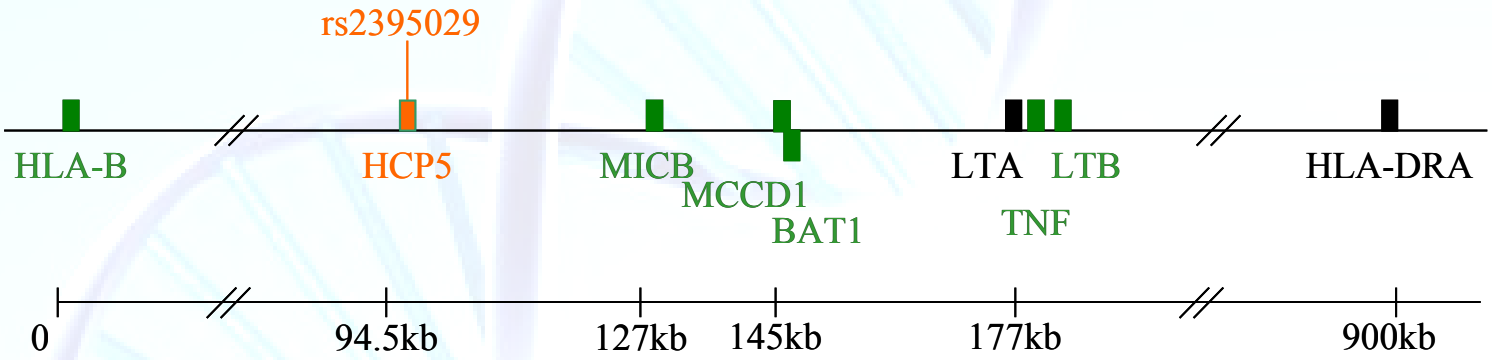


Manhattan plot

7. Example of an AIDS cohort (7)

- ***HCP5*** SNP (rs2395029 *P* = 6.79x10^{-10} odds ratio= 3.47)

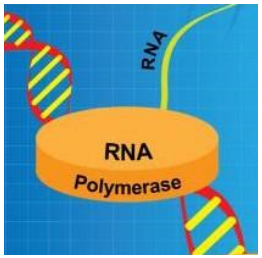
LD with major immunity genes (SNPs and haplotypes):
*HLA-B*57*, *MICB*, *TNFa*, *BAT1*, *LTB* et *MCCD1*



LD complexity → **Not easy to discriminate causal variants**

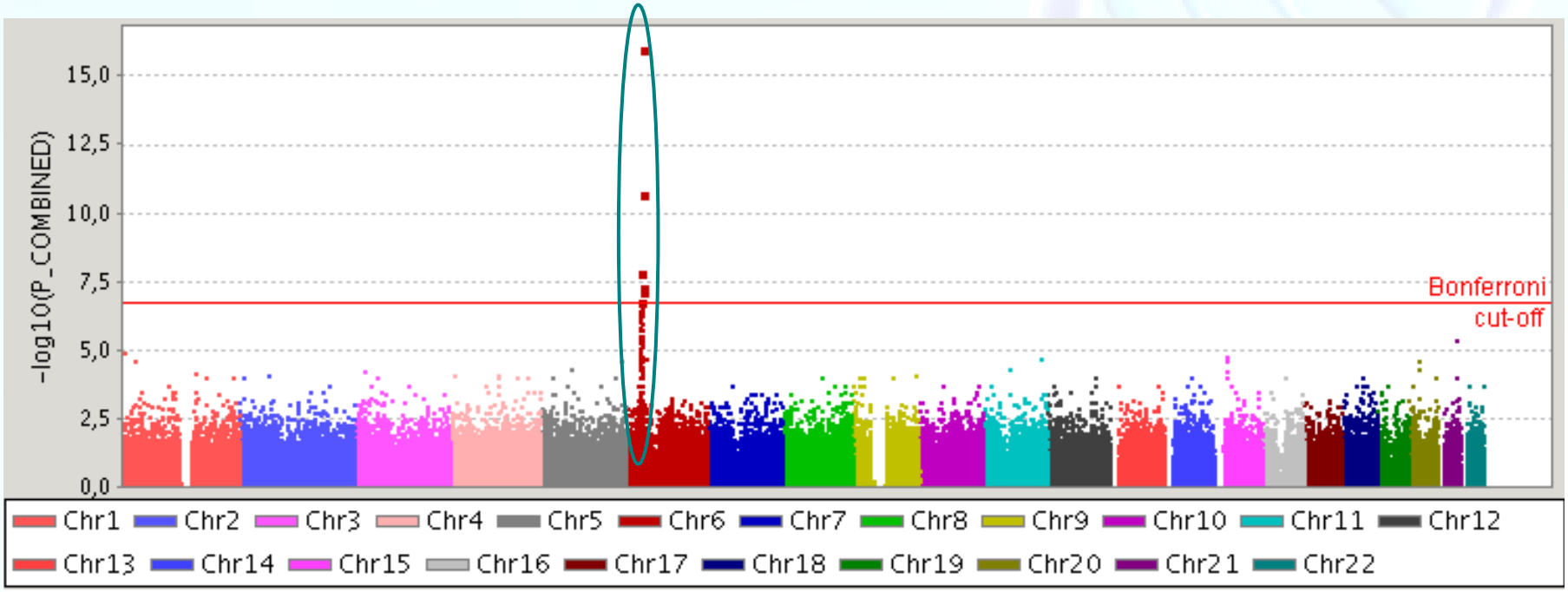
7. Example of an AIDS cohort (8)

- ***RNF39/ZNRD1*** locus (P = 9.2x10^{-7} ~ *HCP5* independent)
  - HLA region (chr 6)
  - Protective effect
  - Subunit of RNA polymerase
    - Interacts with HIV-1 during transcription process
    - Viral replication control



Major role of HLA genes in HIV progression

7. Example of an AIDS cohort (9)

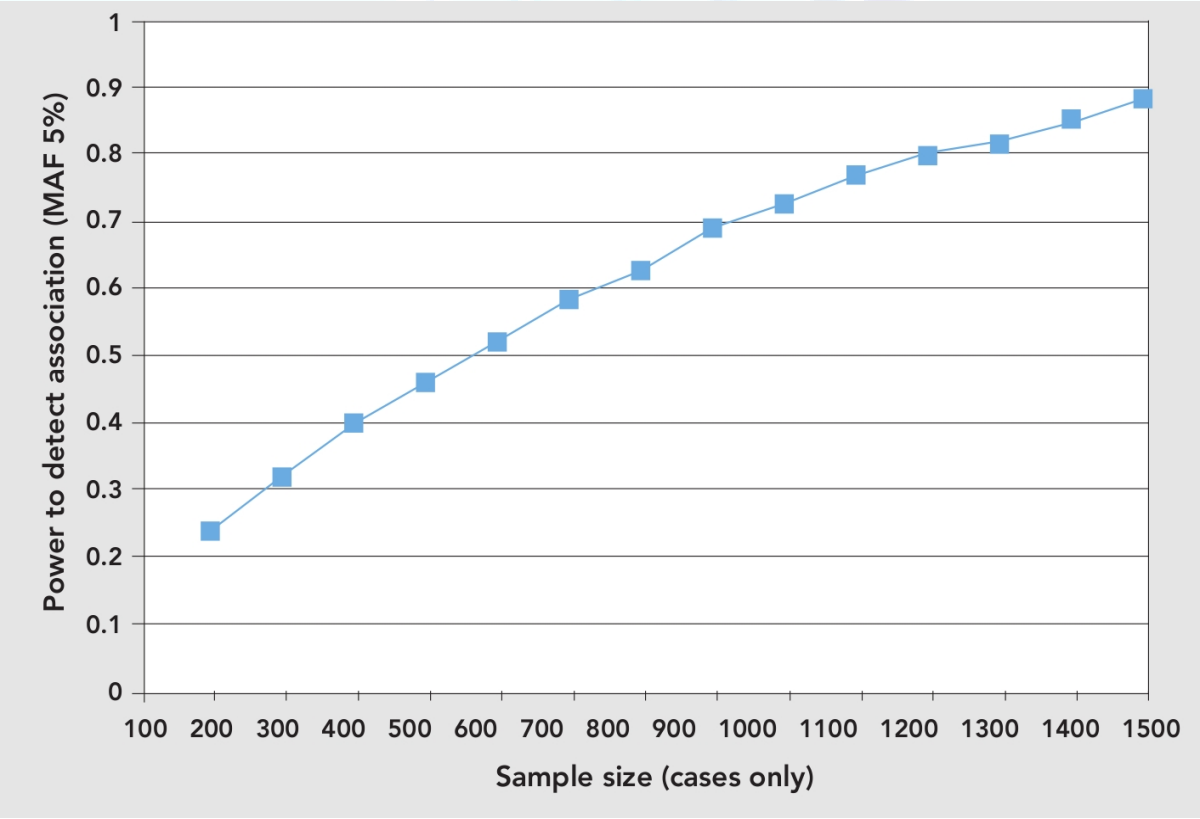


Meta analysis, with first GWAS, **the Euro-CHAVI GWAS** (500K): 486 HIV-1 seroconverters at all stages of disease (viral setpoint study)

Conclusion

GWAS are powerful but:
Need to increase number of cases for low OR
Need for meta analysis but difficulties for replication

OR=1.3
MAF=5%



Acknowledgments

CNRS 6241
Christine Sinoquet

Chaire de bioinformatique CNAM Paris
Pr Jean-Francois Zagury
Sophie Limou
Sigrid Le clerc
Olivier Delaneau
Lieng Taing

Paris 6 INSERM
Pr Amu Therwath