Association Studies

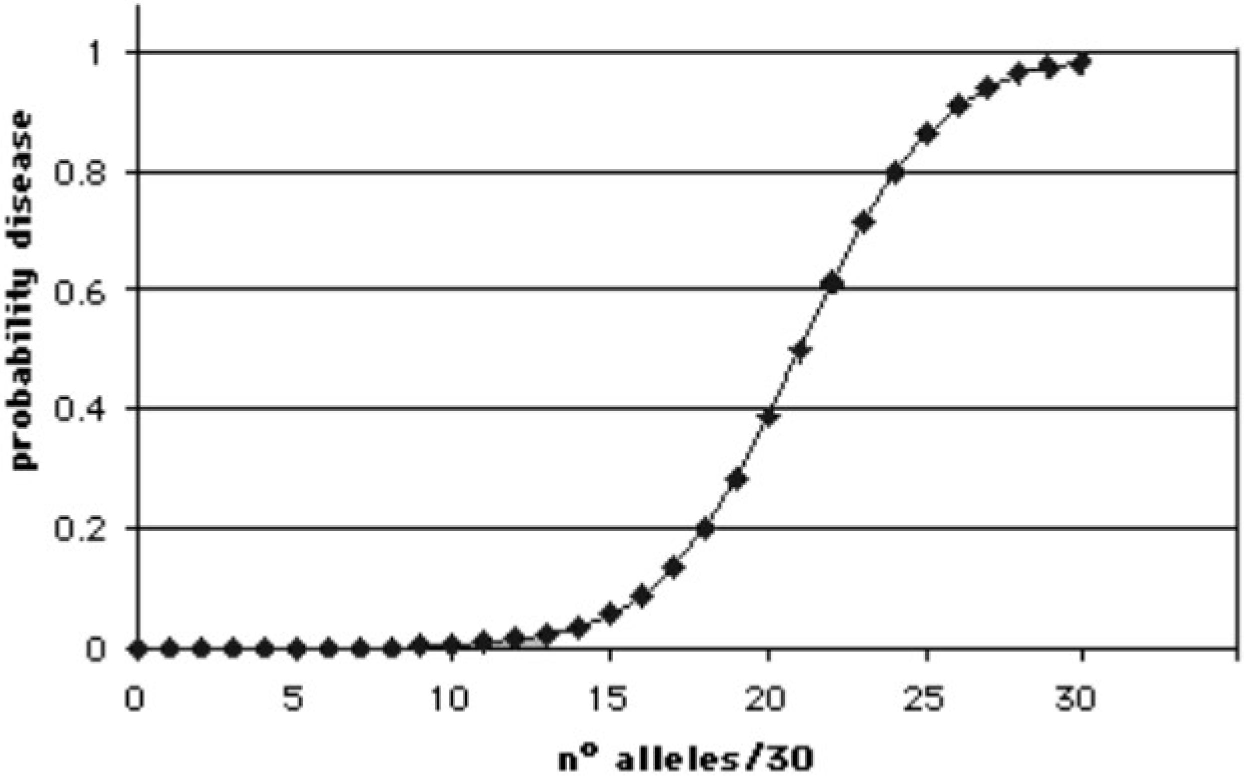
Prof. Pantelis Bagos

2021

Genetic Epidemiology

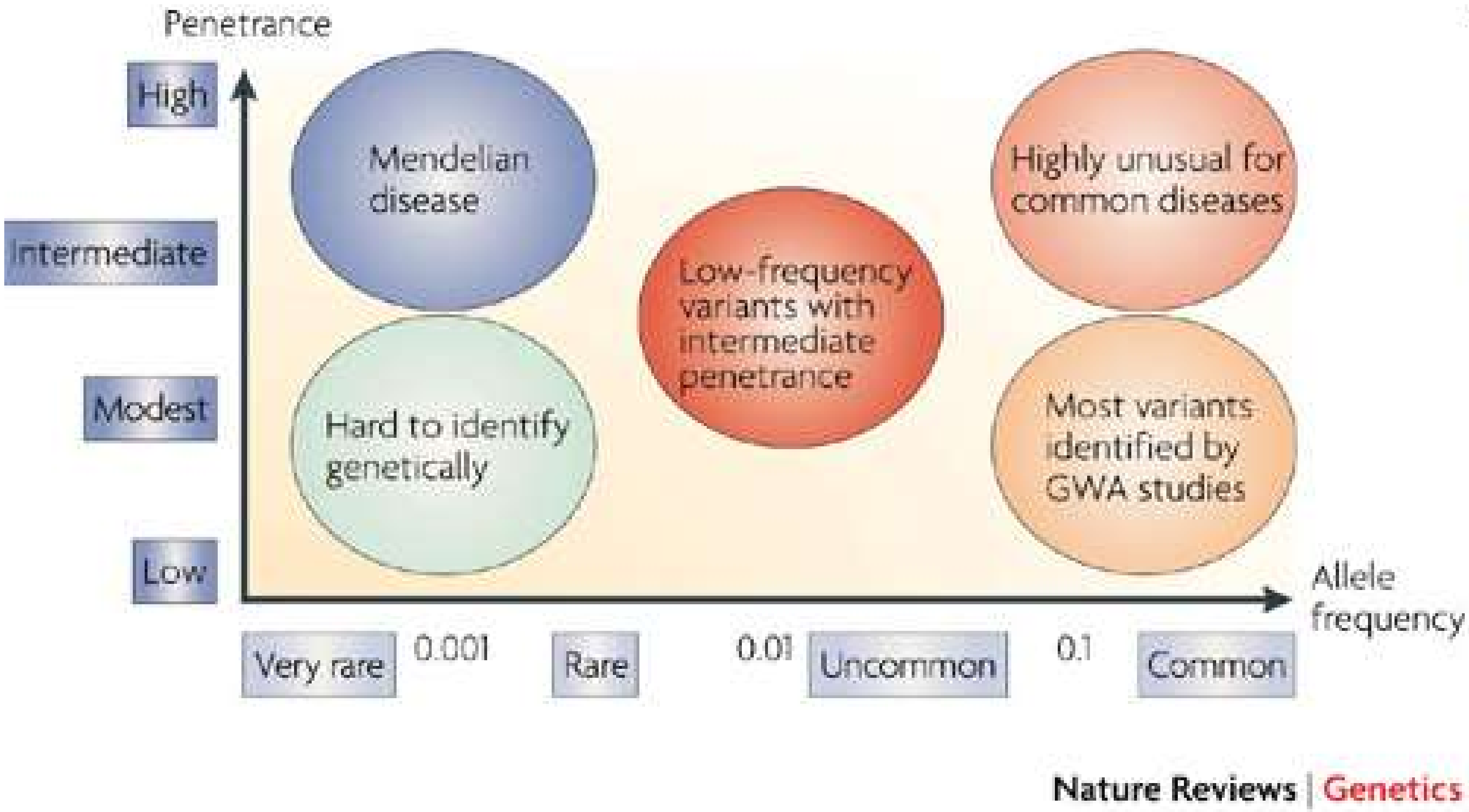
Genetic Epidemiology

- Studies the disease with the aim of deciphering
  - Whether it has a genetic background,
  - The heritability,
  - The mode of inheritance,
  - The genetic locus in which the responsible gene lies,
  - The gene and the allele that predisposes for the disease
  - The interactions with other genes or environmental factors

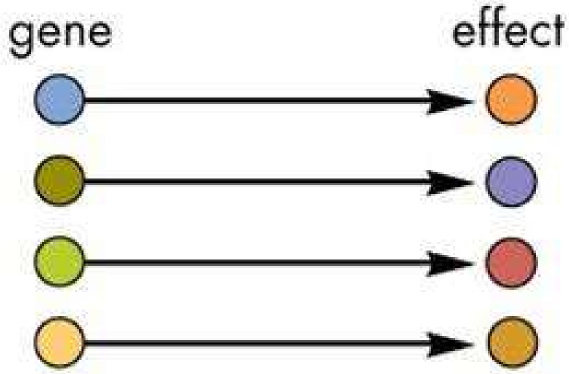


probability disease
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
nº alleles/30

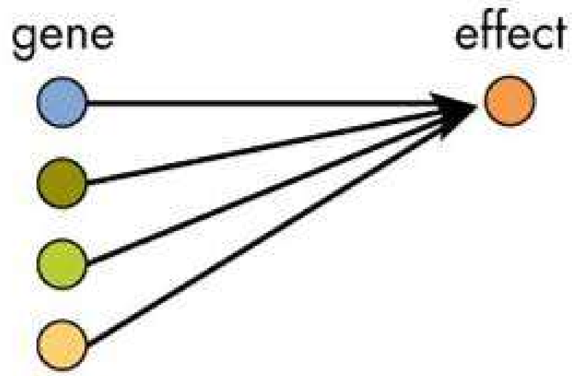
Penetrance



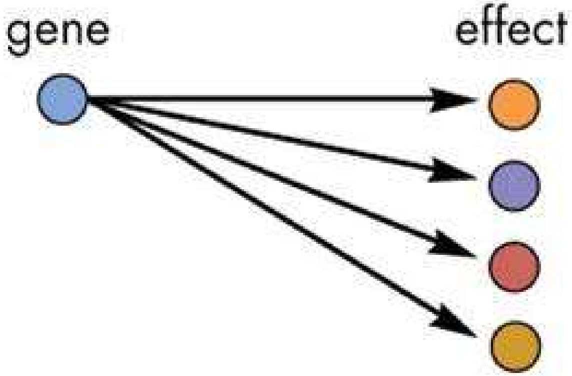
Penetrance=P(Disease | genotype)



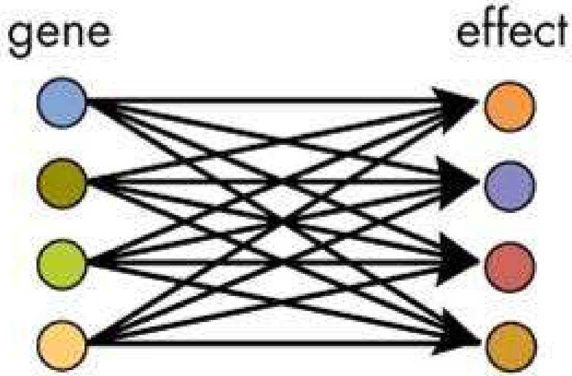
Each gene has a distinct biological effect.



Polygenic trait: Many genes contribute to a single effect.



Pleiotropy: A gene has multiple effects.

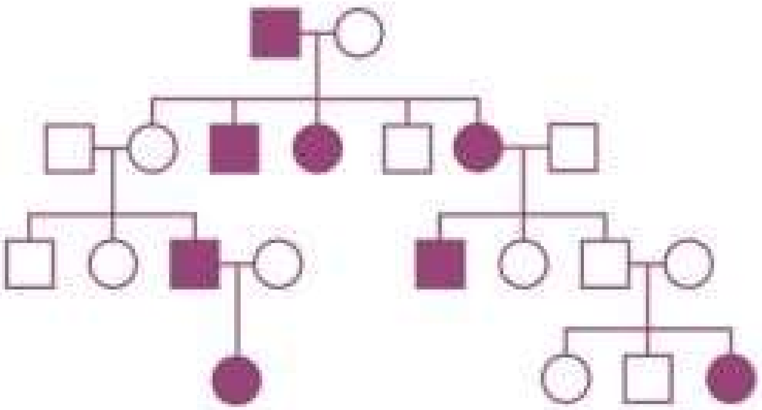


Polygenic traits and pleiotropy

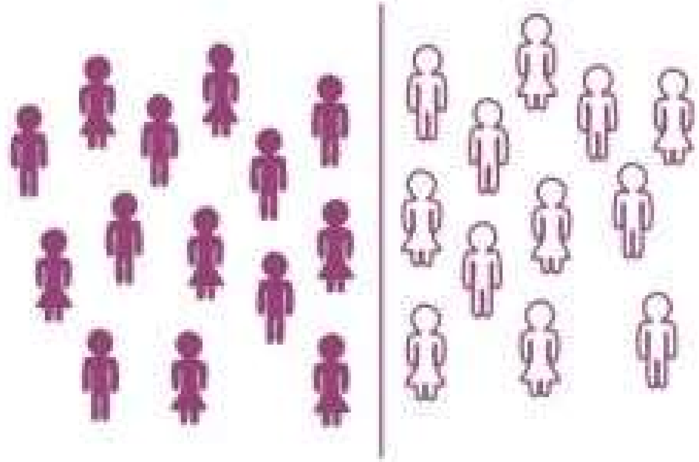
Study types

- Inheritance studies
  - Family history
  - Family studies (twins, adoptions etc)
  - Segregation studies
- Linkage studies
  - Aim to find the genetic locus in which the genes are
- Genetic association studies
  - Find the gene and quantify the risk
  - Family-based vs. population-based
  - GxG and GxE interactions
  - GAS vs. GWAS

Linkage analysis



Association analysis



Genetic Association Studies

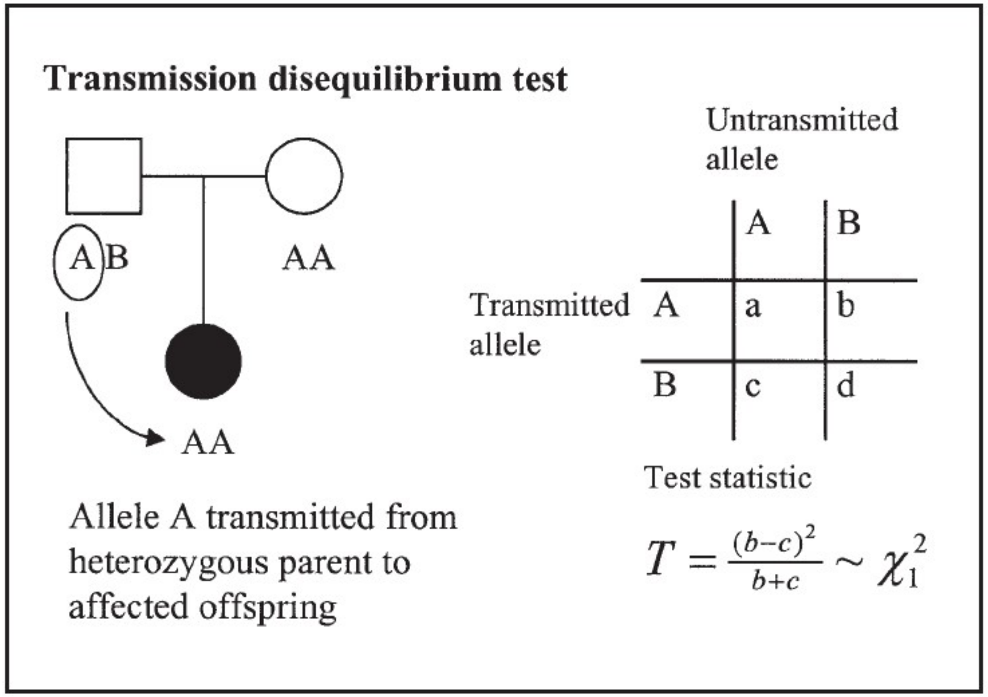
- Identify the allele that causes the disease
  - Use effect sizes like the OR
- In families vs. in population
  - case-control studies, TDT, family based studies
- Gene X Gene and Gene X environment interactions
  - Special designs (πχ case-only studies)
- GAS vs. GWAS
  - One candidate gene or million SNPs

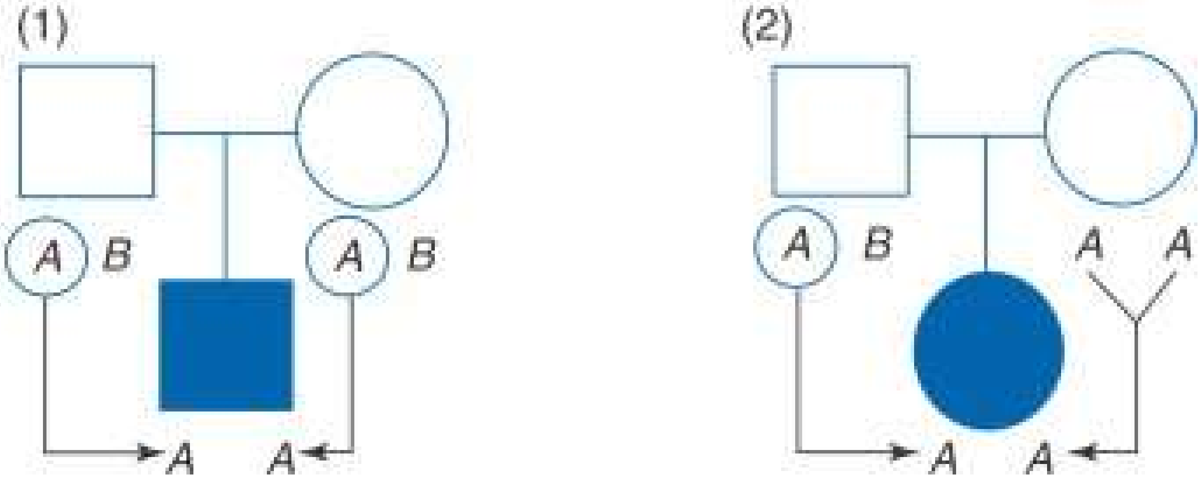
Family based genetic association studies

- Compares the allele in cases and in healthy parents
- Use theTransmission-Disequilibrium Test (TDT) which is equivalent to McNemar's χ^2
- TDT tests both linkage and association

- **Advantages:** controls for confounding (population stratification)

- **disadvantages:** low statistical power





Untransmitted allele

Transmitted allele		A	B
	A	a	b
	B	c	d

→ Test statistic

$$T = \frac{(b-c)^2}{b+c} \sim \chi_1^2$$

Extensions

- 1-TDT (one parent is available)
- Multi-allelic loci
- Sib-TDT (compares the marker genotypes in affected and unaffected offspring)
- Quantitative traits (cholesterol etc)
- X-linked genes
- And more…

1-TDT

TABLE 1. Case-parental control design when only one parental genotype is available

Case *genotype*	Parental genotype NN(0)	NM(1)	MM(2)
NN(0)	A_{00}	A_{01}	0
NM(1)	A_{10}	A_{11}	A_{12}
MM(2)	0	A_{21}	A_{22}

$$T_1 = \frac{A_{01} + A_{12} - A_{10} - A_{21}}{\sqrt{V}} = \frac{b_1 - c_1}{\sqrt{V}}.$$

$$V_1 = \Sigma(b_{1i} - c_{1i})^2,$$

- When diseases with onset in adulthood or in old age are studied, it may be impossible to obtain genotypes for markers in the parents of the affected offspring. This difficulty has limited the applicability of the TDT.
- Instead of using marker data from affected offspring and their parents, this method compares the marker genotypes in affected and unaffected offspring. The S-TDT does not reconstruct parental genotypes and does not depend on estimates of allele frequencies

Sibship and Sib Status	No. of Sibs with Genotype				M_1 Alleles in "Affected" Sibs, by Chance[a]	
	M_1M_1	M_1M_2	M_2M_2	M_1M_3	Mean	Variance
1 (7 sibs):						
Affected	2	1	...	...		
Unaffected	...	2	...	2	3.8571	.4082
2 (5 sibs):						
Affected	...	1	...	...		
Unaffected	...	1	2	1	.6000	.2400
3 (4 sibs):						
Affected	1	...	...	...		
Unaffected	...	1	2	...	.7500	.6875

Table 2. Total Number of Alleles in Affected and Unaffected Members of Sibships in Table 1

Sib Status	No. of Alleles			
	M_1	M_2	M_3	Total
Affected	8	2	0	10
Unaffected	7	12	3	22

HHRR

- The haplotype-based haplotype relative risk (HHRR), in an effort to increase power (i.e. to decrease the variance), uses the unmatched version of Table, since, under the null hypothesis, the two alleles of each parent are independent. The transition to the unmatched analysis is given in Table

Table 1. The 2x2 contingency table corresponding to a population-based case-control study in which allele *B* is considered the susceptibility allele. The total number of *B* and *A* alleles are compared between cases and controls. For brevity we denote n_{01}=2BB_0+AB_0, n_{00}=2AA_0+AB_0, n_{10}= 2AA_1+AB_1 and n_{11}=2BB_1+AB_1. The total number of cases' alleles is n_1 and controls' n_0 (i.e. the total number of cases is n_1/2 and that of controls n_0/2).

		Allele		
		B	*A*	Total
Status	*Cases*	n_{11}	n_{10}	n_1
	Controls	n_{01}	n_{00}	n_0
	Total			n_0+ n_1

Table 2. Presentation of the data in a family-based study using the Transmission Disequilibrium Test (TDT). The transmitted alleles are contrasted against the non-transmitted ones and the OR is given by the ratio of the discordant pairs (b/c). For comparison with Table 1 we denote $a+b=w=n_{11}$ and $c+d=x=n_{10}$.

		Non-transmitted allele		
Transmitted Allele		*B*	*A*	Total
	B	a	b	w
	A	c	d	x
	Total	y	z	n_1

Table 3. Presentation of the data of a family-based study under the Haplotype-based Haplotype Relative Risk (HHRR). The transmitted alleles are contrasted against the non-transmitted alleles of parents that form a "pseudo-control" population. The OR is given by wz /xy. To make the connection with the data in Table 1 we have to notice that the first rows of the tables are identical ($n_{11}=w$ and $n_{10}=x$)

	Allele		
	B	*A*	Total
Transmitted	w	x	n_1
Non-transmitted	y	z	n_1
Total	$w+y$	$x+z$	$2n_1$

Remarks

- From a historical point of view, it is worth-noting that Falk and Rubinstein were the first to propose the use of untransmitted alleles to form a single pseudocontrol genotype (Falk & Rubinstein, 1987).
- Later, Terwilliger and Ott extended this idea and they discussed, for the first time, the use of McNemar’s test, although they concluded that it was less powerful than the unmatched analysis that corresponds to the HHRR test (Terwilliger & Ott, 1992).
- Few years later, the McNemar’s statistic was reformulated and presented as the TDT test that is now widely used (Spielman & Ewens, 1996; Spielman et al, 1993).

Population based genetic association studies

- Compares the allele in cases and unrelated controls
- Typical epidemiological design

▪ **advantages** statistical power, large sample

▪ **disadvantages:** requires testing to control for confounding due to ethnicity (population stratification)

	Exposed	Unexposed
Cases	α	β
Controls	γ	δ

Odds Ratio

$$OR = \frac{\alpha\delta}{\beta\gamma}, \quad se_{\log OR} = \sqrt{\frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma} + \frac{1}{\delta}}$$

X^2 as criterion of association

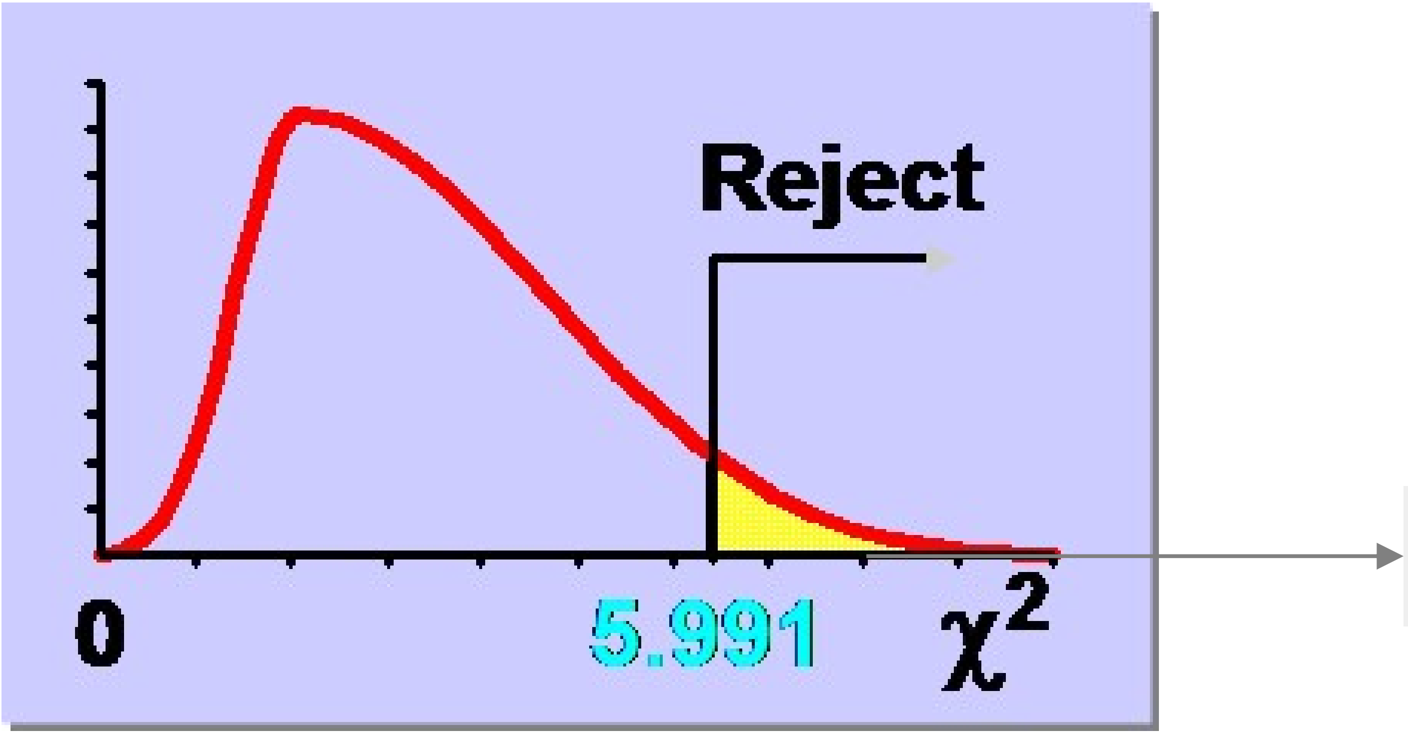
	Genotype			
	AA	*AB*	*BB*	*Total*
Disease				
yes	140 (320*380)/890=136,6	125 (320*390)/890=140,2	55 (320*120)/890=43,2	320
no	240 (570*380)/890=243,4	265 (570*390)/890=249,8	65 (570*120)/890=76,8	570
Σύνολο	380	390	120	890

$$X^2 = \sum\frac{(O-E)^2}{E} = \frac{(140-136{,}6)^2}{136{,}6} + \frac{(125-140{,}2)^2}{140{,}2} + \frac{(55-43{,}2)^2}{43{,}2} + \frac{(240-243{,}4)^2}{243{,}4} + \frac{(265-249{,}8)^2}{249{,}8} + \frac{(65-76{,}8)^2}{76{,}8} =$$

$$0{,}08 + 1{,}65 + 3{,}22 + 0{,}05 + 0{,}92 + 1{,}81 = 7{,}73$$

X^2 as criterion of association

- H0: disease and exposure are unrelated
- H1: there is a relation
- ❖ α = 0.05
- ❖ df. = (*r*-1)*(*c*-1)=(2-1)*(3-1)=2



$\chi 2 = 7{,}73$

Collapsing the table

	Genotypes		
	AA	**AB**	**BB**
Cases	α	β	γ
Controls	δ	ε	ζ

- The Pearson χ^2, performs a model-free approach
- In order to assume a particular model we need to have a 2x2 table, i.e. merging AA+AB or AB+BB

Example

	Genotypes		
	AA	**AB**	**BB**
Cases	105	225	119
Controls	132	206	87

```
. tabi  132 206 87\ 105 225 119, all

           |                  col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |       132        206         87 |       425 
         2 |       105        225        119 |       449 
-----------+---------------------------------+----------
     Total |       237        431        206 |       874 

          Pearson chi2(2) =   8.2316   Pr = 0.016
 likelihood-ratio chi2(2) =   8.2524   Pr = 0.016
               Cramér's V =   0.0970
                    gamma =   0.1628  ASE = 0.056
          Kendall's tau-b =   0.0915  ASE = 0.032
```

```
. tabi  132 293\ 105 344, all
```

row	col 1	2	Total
1	132	293	425
2	105	344	449
Total	237	637	874

```
          Pearson chi2(1) =   6.5050   Pr = 0.011
 likelihood-ratio chi2(1) =   6.5111   Pr = 0.011
               Cramér's V =   0.0863
                    gamma =   0.1922  ASE = 0.074
          Kendall's tau-b =   0.0863  ASE = 0.034
```

```
. tabi  338 87\ 330 119, all
```

row	col 1	2	Total
1	338	87	425
2	330	119	449
Total	668	206	874

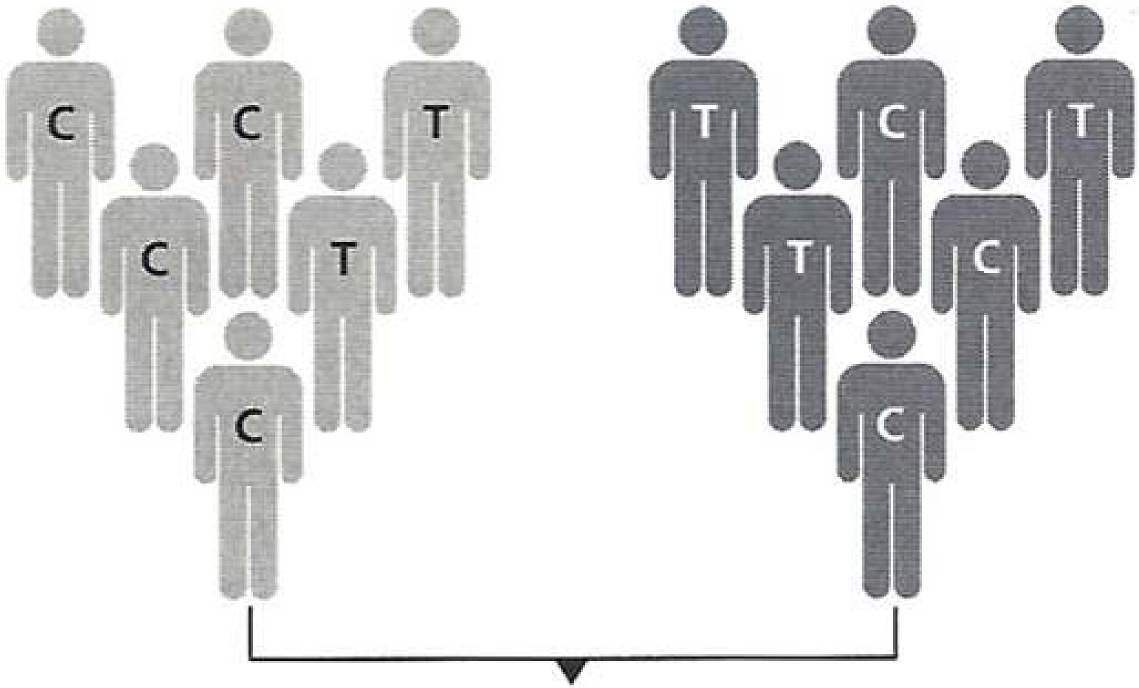
```
          Pearson chi2(1) =   4.4110   Pr = 0.036
 likelihood-ratio chi2(1) =   4.4277   Pr = 0.035
               Cramér's V =   0.0710
                    gamma =   0.1670  ASE = 0.078
          Kendall's tau-b =   0.0710  ASE = 0.034
```

Case-control study for genetic association

Cases (n=1,000)
(express the trait)

vs.

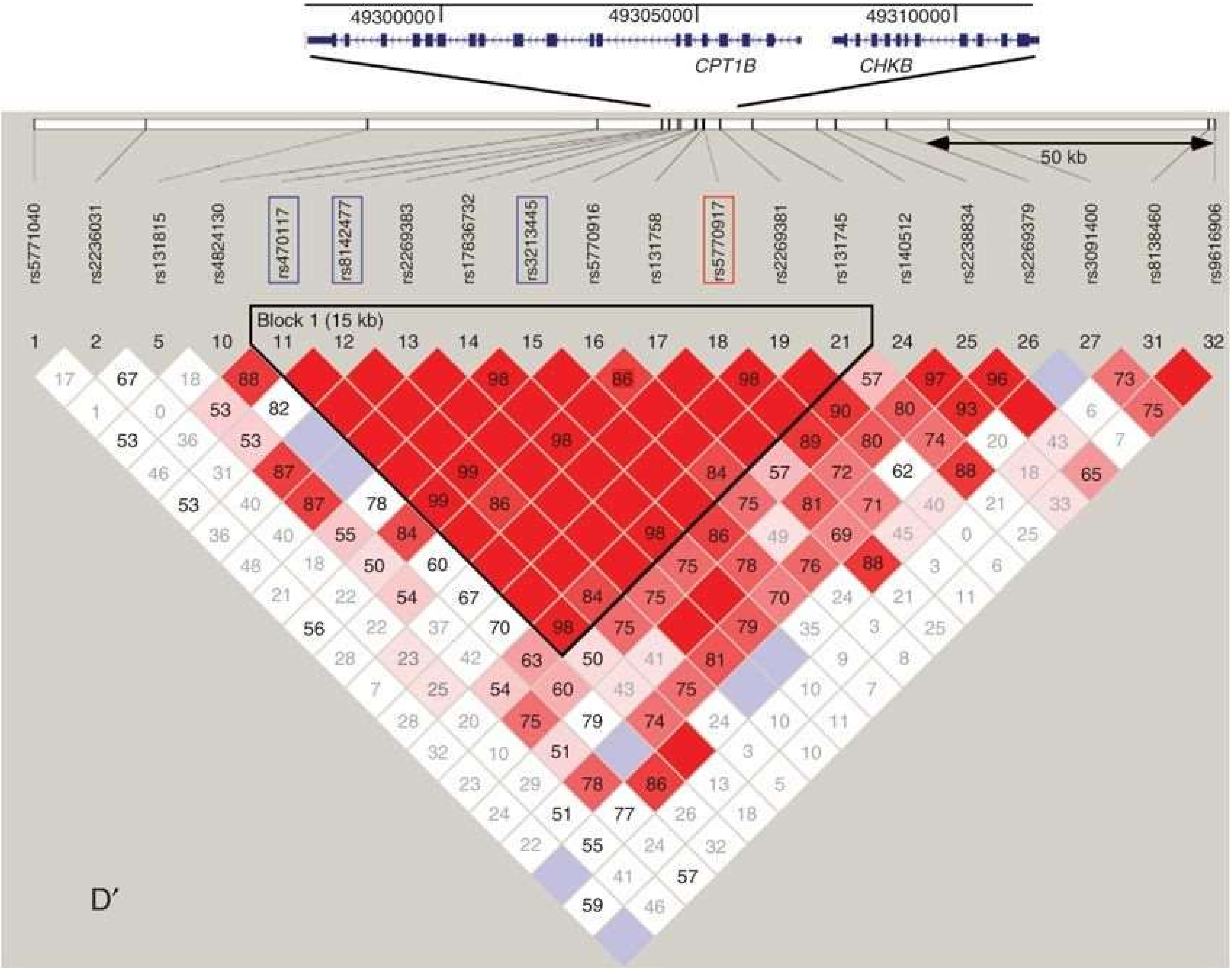
Controls (n=1,000)
(do not express the trait)



	C	T
Cases	62%	38%
Controls	49%	51%

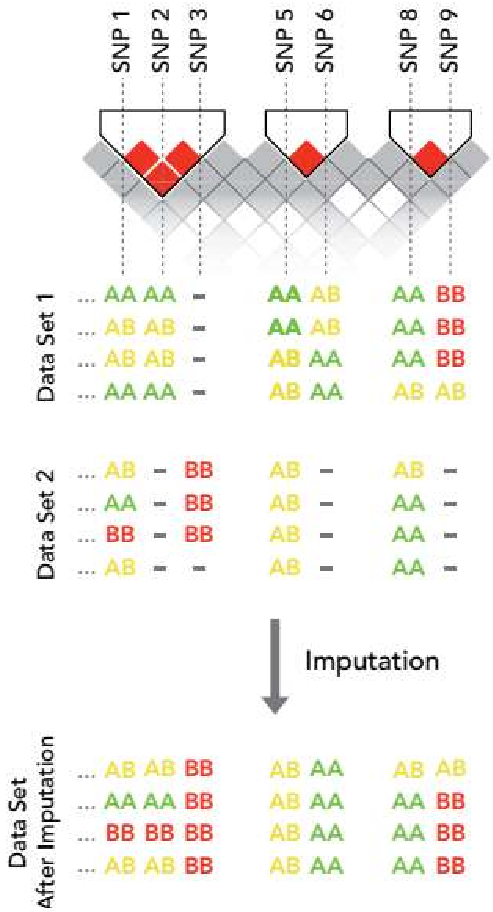
GWAS

- Million of SNPs
- Different platforms (need for imputation)
- The statistical analysis is simple(i.e. OR, CATT, SMD) but there are complications
- Basic issues: multiple comparisons, quality control, data sharing, population stratification
  - Teo, Y.Y. (2008) Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure, Curr Opin Lipidol, 19, 133-143
  - Zeggini, E. and Ioannidis, J.P. (2009) Meta-analysis in genome-wide association studies, Pharmacogenomics, 10, 191-201
  - Ziegler, A., Konig, I.R. and Thompson, J.R. (2008) Biostatistical aspects of genome-wide association studies, Biom J, 50, 8-28



49300000
49305000
49310000
CPT1B
CHKB
50 kb
rs5771040
rs2236031
rs131815
rs4824130
rs470117
rs8142477
rs2269383
rs17836732
rs3213445
rs5770916
rs131758
rs5770917
rs2269381
rs131745
rs140512
rs2238834
rs2269379
rs3091400
rs8138460
rs9616906
Block 1 (15 kb)
1 2 5 10 11 12 13 14 15 16 17 18 19 21 24 25 26 27 31 32
D′

FIGURE 1: IMPUTATION OVERVIEW



SNPs 1–9 form three blocks of high LD, indicated by the red diamonds between the SNPs. Data Sets 1 and 2 represent a total of eight individuals genotyped using two different arrays at SNPs 1–9. The imputed data set contains genotypes for all SNP loci, with estimated genotypes filling in the missing data from Data Set 2. For example, SNP 2 is genotyped in Data Set 1 but not Data Set 2. Due to strong LD between SNPs 1–3, the individual genotypes for SNP 2 can be inferrred in Data Set 2 based on those present in Data Set 1.

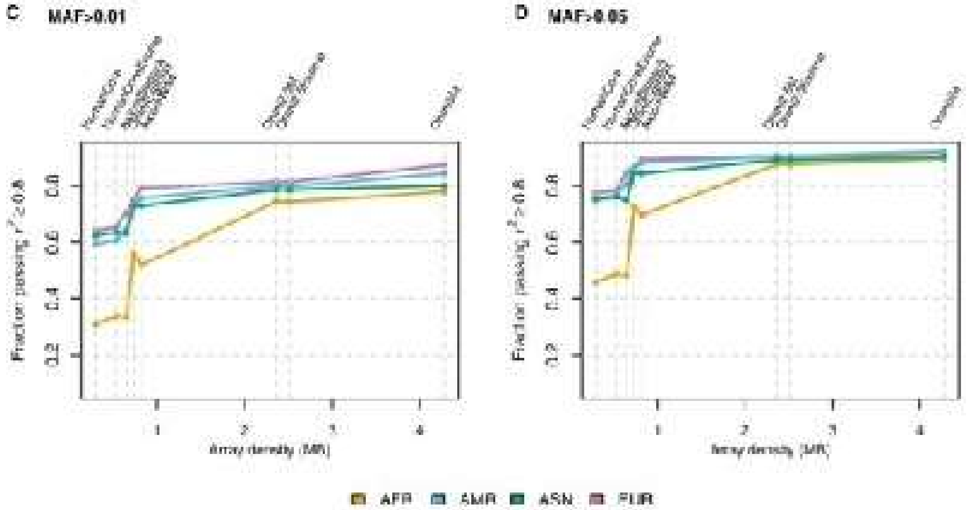
TABLE 1: COMMONLY USED IMPUTATION SOFTWARE PACKAGES

SOFTWARE NAME	INSTITUTION	URL
MACH	University of Michigan[1,2]	http://www.sph.umich.edu/csg/abecasis/MaCH/tour/imputation.html
BEAGLE	University of Auckland[3]	http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html
IMPUTE	Oxford University[4,5]	http://mathgen.stats.ox.ac.uk/impute/impute.html
PLINK	Massachusetts General Hospital / Broad Institute[6]	http://pngu.mgh.harvard.edu/~purcell/plink/

Genomic Coverage of GWAS Chips

- estimated by the percent of common SNPs having an r^2 of 0.8 or greater with at least 1 SNP on the platform.

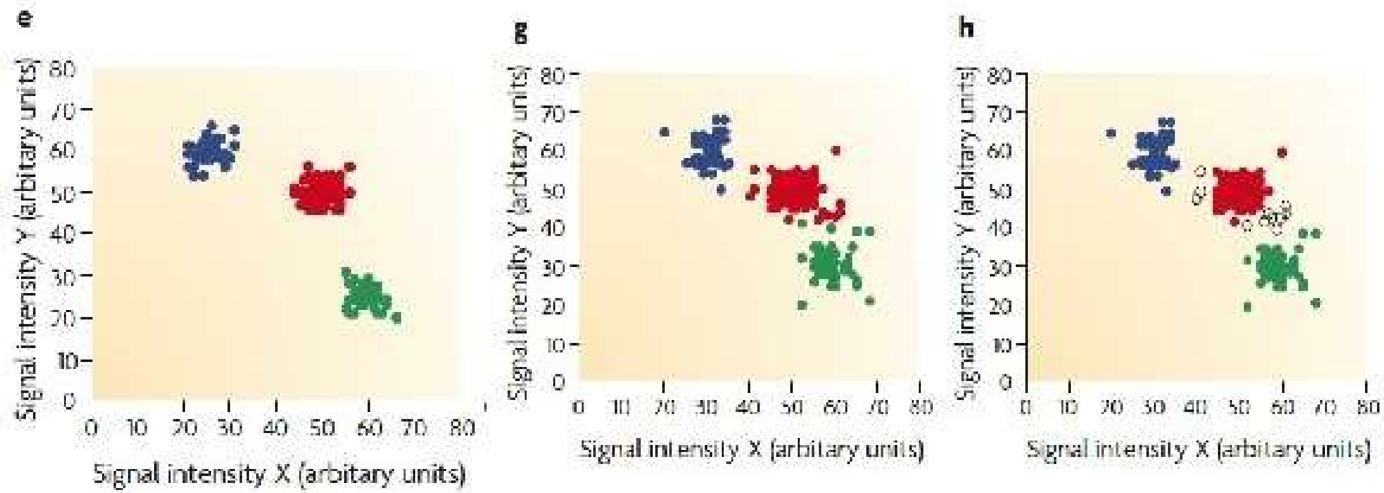
- Platforms comprising 500,000 to 1,000,000 SNPs capture ~67 -89% of common SNPs in populations of European and Asian ancestry and 46-66% in populations of African ancestry.



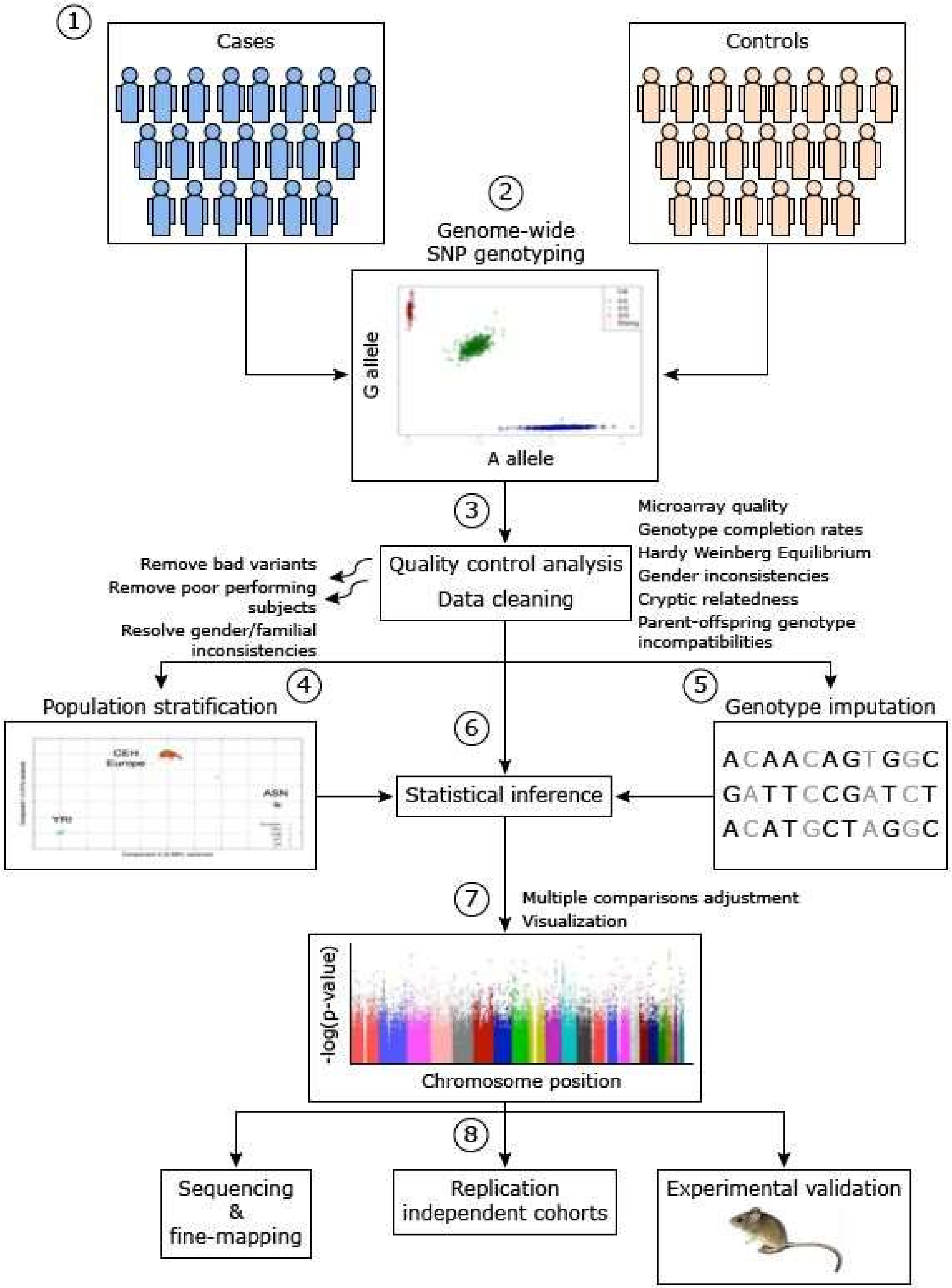
Nelson et al. G3 (Bethesda) 2013; 3: 1795–1807.

Genotyping and Quality Control in GWAS

- Genotype "calling" is based on intensities for the two alleles at each genetic marker
- Genotyping errors, must be diligently sought and corrected.
- Established quality control features should be applied both on a per-sample and a per-SNP basis.

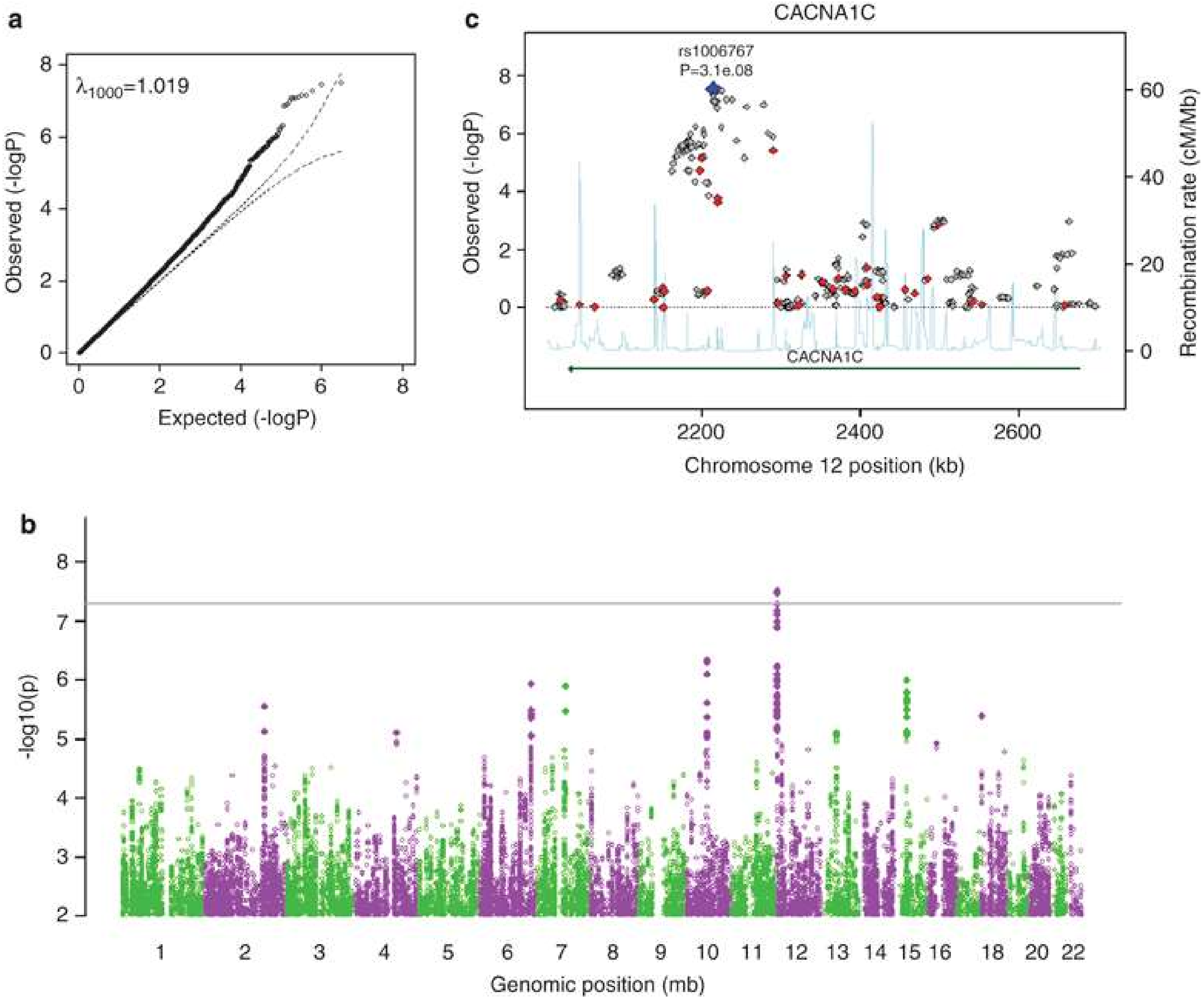


McCarthy et al. Nat Rev Genet 2008; 9:356-369



1
Cases
Controls
2
Genome-wide
SNP genotyping
G allele
A allele
3
Quality control analysis
Data cleaning
Remove bad variants
Remove poor performing subjects
Resolve gender/familial inconsistencies
Microarray quality
Genotype completion rates
Hardy Weinberg Equilibrium
Gender inconsistencies
Cryptic relatedness
Parent-offspring genotype incompatibilities
4
Population stratification
CEH
Europe
ASN
YRI
5
Genotype imputation
ACAACAGTGGC
GATTCCGATCT
ACATGCTAGGC
6
Statistical inference
7
Multiple comparisons adjustment
Visualization
-log(p-value)
Chromosome position
8
Sequencing
&
fine-mapping
Replication
independent cohorts
Experimental validation

GWAS



a
λ_{1000}=1.019
Observed (-logP)
Expected (-logP)
c
CACNA1C
rs1006767
P=3.1e.08
Observed (-logP)
Recombination rate (cM/Mb)
CACNA1C
Chromosome 12 position (kb)
b
-log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22
Genomic position (mb)

Statistical methods

Table 1 The 2×3 contingency table with the distribution of cases and controls in a traditional GAS or GWAS concerning a single biallelc locus.

	$AA\ (g_0)$	$AB\ (g_1)$	$BB\ (g_2)$	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Pearson Chi-square

$$T^2_{\chi^2_2}=\sum_{j=0}^{2}\frac{\left(r_j-n_j r/n\right)^2}{n_j r/n}+\sum_{j=0}^{2}\frac{\left(s_j-n_j s/n\right)^2}{n_j s/n}$$

Logistic Regression (Odds Ratio)

$$\text{logit}\ [P(\text{case}|g_j)]=\alpha+\beta_1 x_1+\beta_2 x_2$$

Cochran-Armitage Trend Test (CATT)

$$Z_{CATT(x)}=\frac{U_x}{\sqrt{\text{var}_{H_0}(U_x)}}=\frac{\sqrt{n}\sum_{i=0}^{2}x_i\left(sr_i-rs_i\right)}{\sqrt{rsn\left[n\sum_{i=0}^{2}x_i^2 n_i-\left(\sum_{i=0}^{2}x_i n_i\right)^2\right]}}\sim N(0,1)$$

Bagos PG. **Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis.** Statistical Applications in Genetic and Molecular Biology, 2013

Robust methods

- The methods are designed to have the maximum statistical power irrespective of the mode of inheritance

MERT

$$Z_{MERT} = \frac{Z_{CATT(0)} + Z_{CATT(1)}}{\sqrt{2\left(1+\rho_{CATT(0,1)}\right)}} \sim N(0,1)$$

MAX

$$Z_{MAX} = \max(|Z_{CATT(0)}|, |Z_{CATT(1/2)}|, |Z_{CATT(1)}|)$$

MIN2

$$MIN2 = \min\left(P_{T^2_{\chi^2_2}}, P_{Z_{CATT(1/2)}}\right)$$

Bagos PG. **Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis.** Statistical Applications in Genetic and Molecular Biology, 2013

et al. (2008). Recently, Zang *et al.* found that $Z_{CATT(0)}$, $Z_{CATT(1/2)}$ and $Z_{CATT(1)}$ are linearly dependent, a result that allowed them to develop faster algorithms for calculating the statistical significance of MAX (Zang *et al.*, 2010). Thus, the *P*-value for the MAX statistic is given by:

$$P(Z_{MAX} < t) = 2\int_0^{t(1-\omega_1)/\omega_0} \Phi\left(\frac{t - \rho_{CATT(0,1)} z_0}{\sqrt{1 - \rho_{CATT^2_{(0,1)}}}}\right)\phi(z_0)dz_0$$
$$+2\int_{t(1-\omega_1)/\omega_0}^{t} \Phi\left(\frac{(t - \omega_0 z_0)/\omega_1 - \rho_{CATT(0,1)} z_0}{\sqrt{1 - \rho_{CATT^2_{(0,1)}}}}\right)\phi(z_0)dz_0$$
$$-2\int_0^{t} \Phi\left(\frac{-t - \rho_{CATT(0,1)} z_0}{\sqrt{1 - \rho_{CATT^2_{(0,1)}}}}\right)\phi(z_0)dz_0 \quad (4)$$

where:

$$\omega_0 = \frac{\rho_{CATT(0,1/2)} - \rho_{CATT(0,1)}\rho_{CATT(1/2,1)}}{1 - \rho_{CATT^2_{(0,1)}}} \quad (5)$$

$$\omega_1 = \frac{\rho_{CATT(1/2,1)} - \rho_{CATT(0,1)}\rho_{CATT(0,1/2)}}{1 - \rho_{CATT^2_{(0,1)}}} \quad (6)$$

Zang Y, Fung WK, Zheng G. Simple algorithms to calculate the asymptotic null distributions of robust tests in case-control genetic association studies in R. Journal of Statistical software. 2010 Feb 17;33(8).

MIN2 is an interesting robust approach that was adopted by investigators of the Wellcome Trust Case-Control Consortium (WTCCC, 2007). They applied the $\chi_2{}^2$ along with the CATT(½) and, subsequently, chose the minimum of the *P*-values:

$$MIN2 = \min(P_{T^2_{\chi^2_2}}, P_{Z_{CATT(1/2)}}) \quad (7)$$

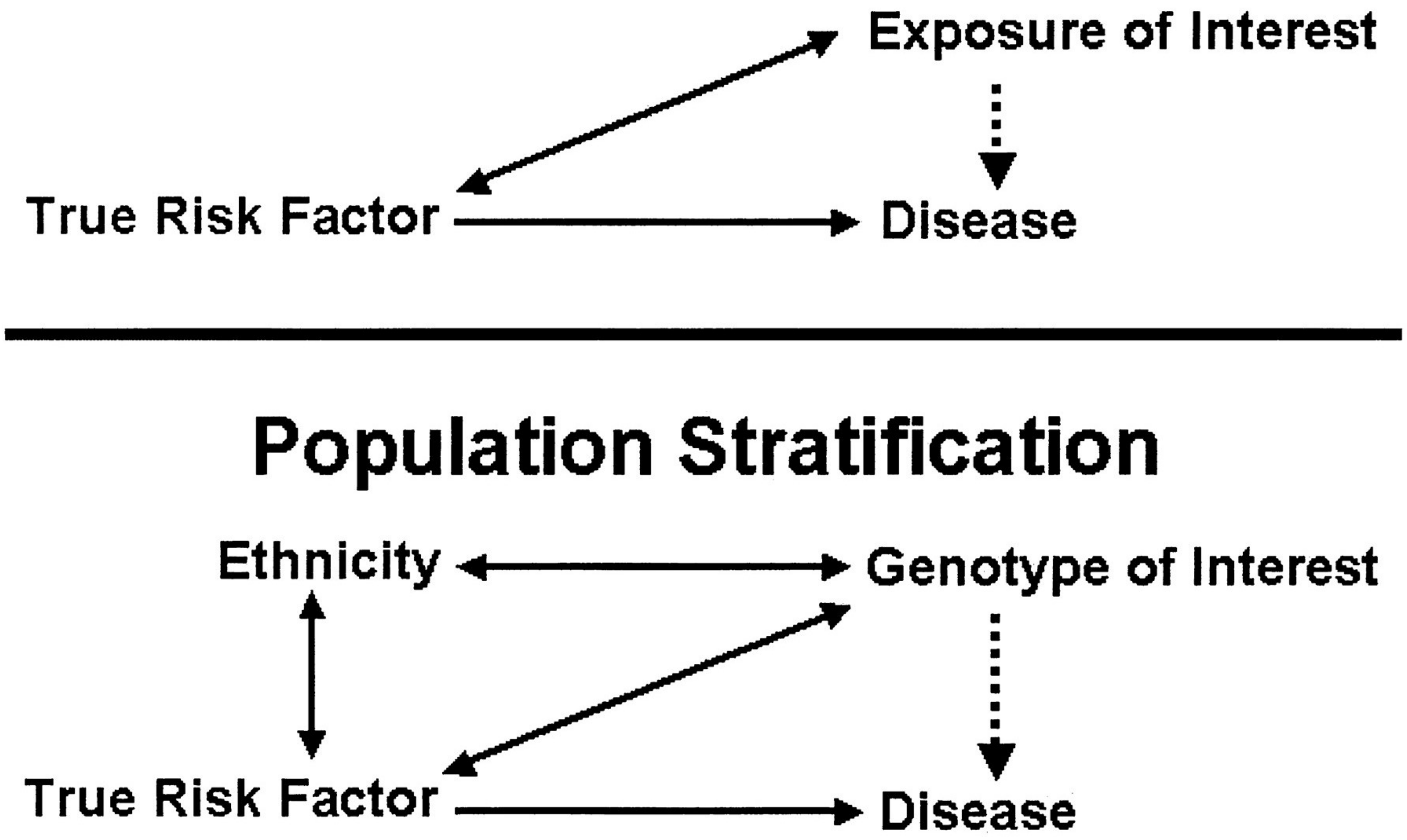
The use of MIN2 is justified by simulations showing that $\chi_2{}^2$ has 5% less power compared with MAX and outperforms MERT, except when the additive model holds (Zheng *et al.*, 2006). However, MIN2 is not a proper *P*-value since the statistics are correlated and multiple tests are performed. Later, Joo *et al.* (2009) derived the joint distribution needed in order to calculate a proper *P*-value:

$$P\left(Z^2_{CATT(1/2)} < t_1, T^2_{\chi^2_2} < t_2\right) = \begin{cases} 1 - \frac{1}{2}e^{-\frac{t_1}{2}} - \frac{1}{2}e^{-\frac{t_2}{2}} + \frac{1}{2\pi}\int_{t_1}^{t_2} e^{-\frac{u}{2}} \arcsin\left(\frac{2t_1}{u} - 1\right) du, t_1 < t_2 \\ 1 - e^{-\frac{t_2}{2}}, t_1 > t_2 \end{cases} \quad (8)$$

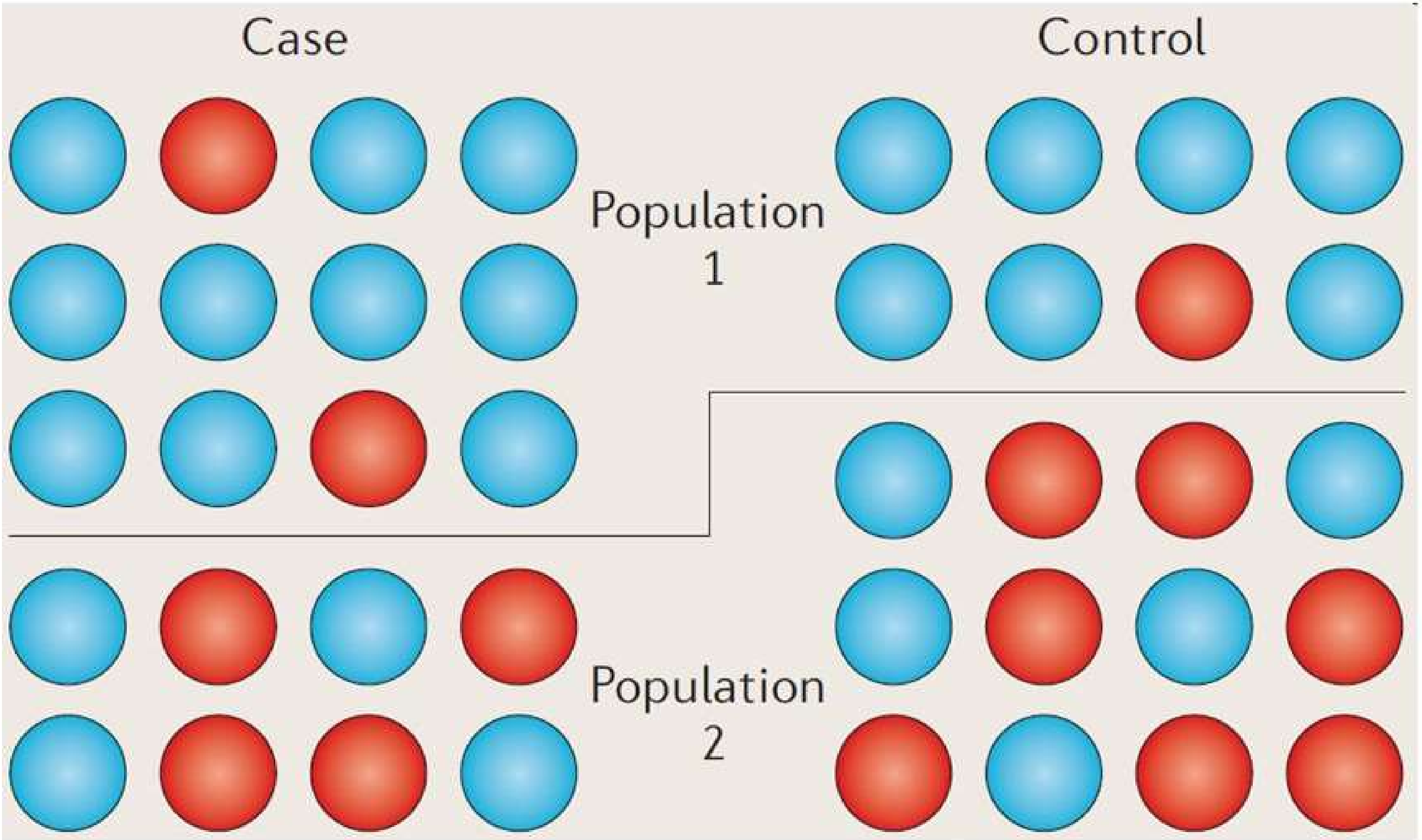
Unlike MAX, MIN2 is independent of the allele frequency.

Joo J, Kwak M, Ahn K, Zheng G. A Robust Genome-Wide Scan Statistic of the Wellcome Trust Case–Control Consortium. Biometrics. 2009 Dec;65(4):1115-22.

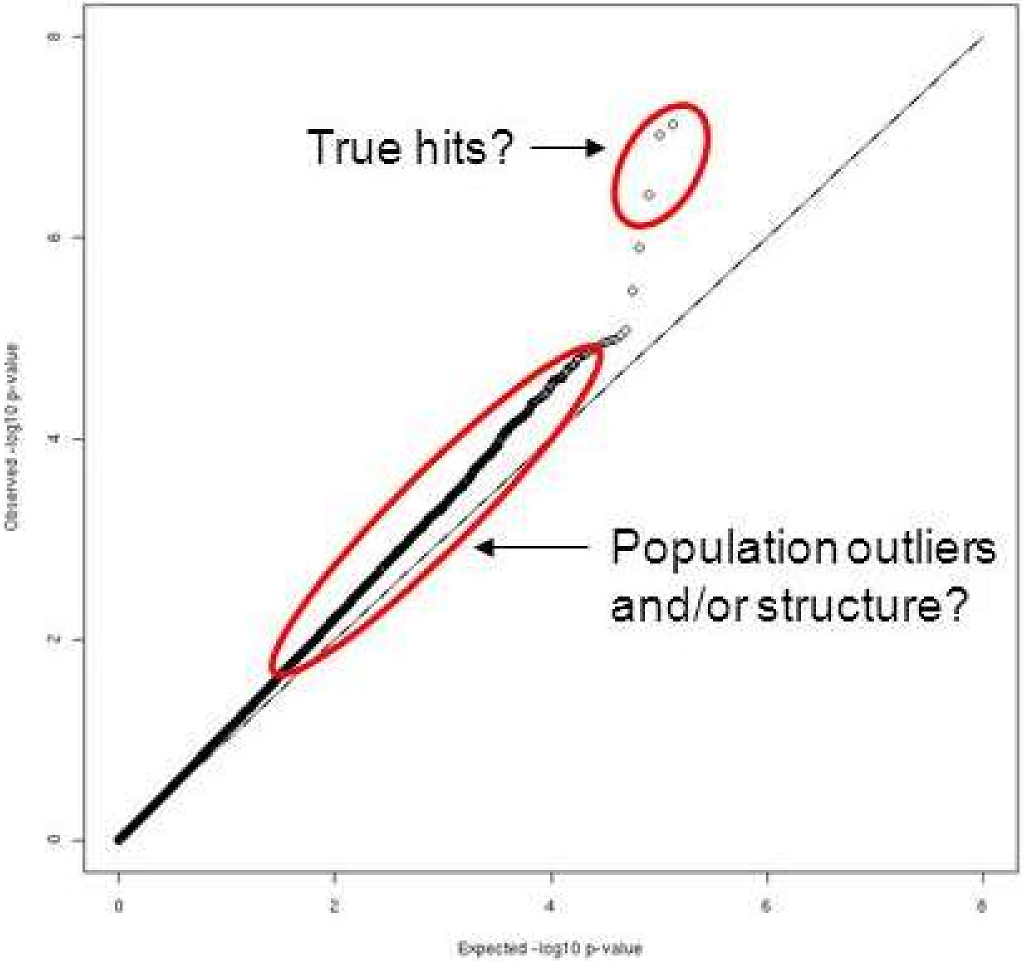
Confounding



Population Stratification



Balding, Nature Reviews Genetics 2010



Inflation factor λ = 1.11

Genomic Control

Let $\chi_1^2, \ldots, \chi_L^2$ be the χ^2-statistics at the null markers. The same type of test statistic is selected and applied to all null loci and the marker loci are tested formally for association. The inflation factor λ for the variance can then be estimated by

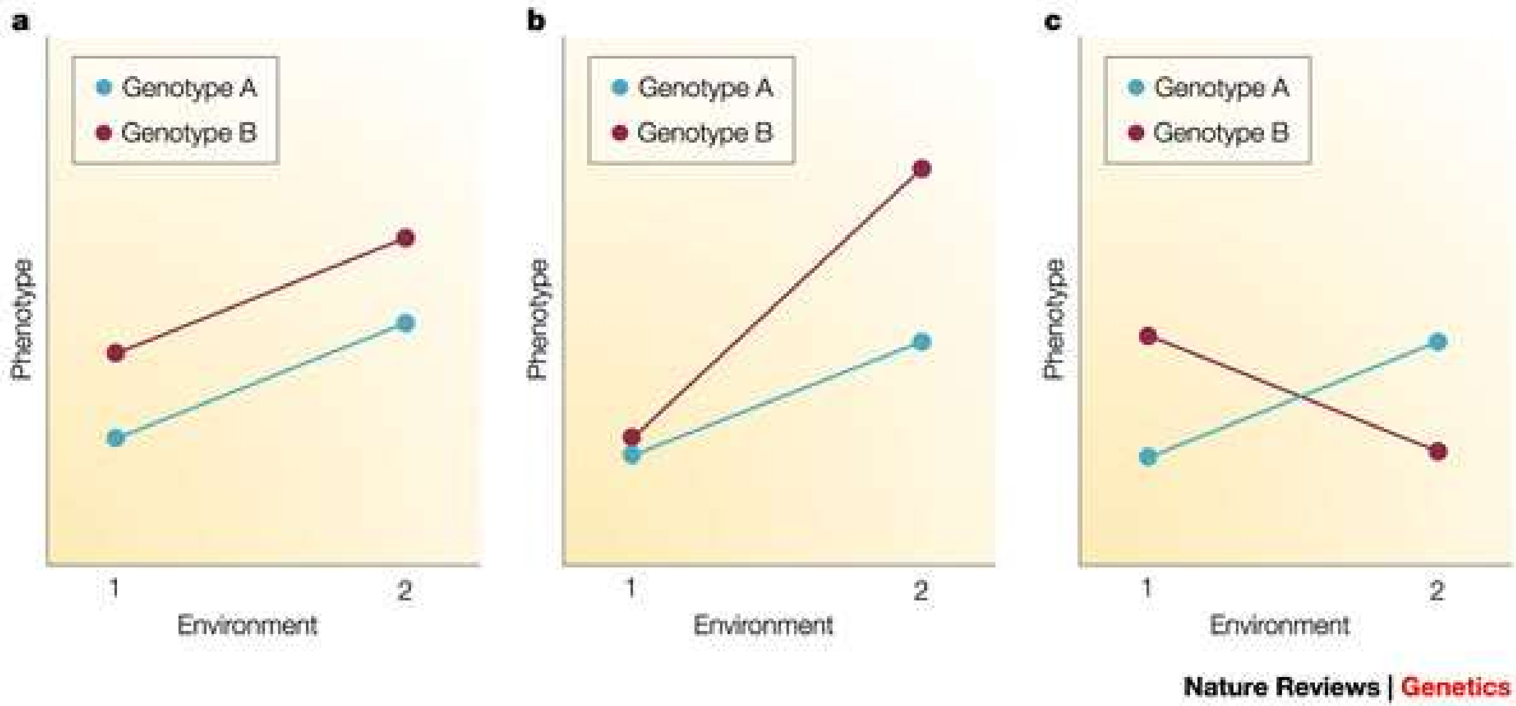
$$\hat{\lambda} = \frac{0.4549}{\text{median}(\chi_1^2, \ldots, \chi_L^2)}.$$

The value of 0.4549 corresponds to the median for the χ^2-distribution with 1 df. The test statistic, e.g., χ_T^2 or χ_L^2, for the marker locus of interest is then adjusted by

$$\chi_{GC}^2 = \hat{\lambda}\, \chi_L^2 \sim \chi_1^2$$

for the alleles test, and similarly for the trend test χ_T^2. For a codominant test we use the median value of a χ_2^2 distribution in the numerator of $\hat{\lambda}$.

GxE interactions



$$g(E(Y)) = \beta_0 + \beta_1 \times X + \beta_2 \times E + \beta_3 \times X \times E,$$

TABLE I. Data for a unmatched case-control study with a binary genetic factor and a binary environmental exposure

	$G=0$		$G=1$		
	$E=0$	$E=1$	$E=0$	$E=1$	Total
$D=0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D=1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

$$\psi = \underbrace{\frac{\text{Odds-ratio between } G \text{ and } E \text{ among cases}}{\text{Odds-ratio between } G \text{ and } E \text{ among controls}}}_{=1 \text{ under } G\text{–}E \text{ independence and rare disease}}.$$

$$\hat{\beta}_{CC} = \log\left(\frac{r_{001}r_{010}r_{100}r_{111}}{r_{000}r_{011}r_{101}r_{110}}\right).$$

$$\hat{\beta}_{CO} = \log\left(\frac{r_{100}r_{111}}{r_{101}r_{110}}\right).$$

TABLE 2. Gene-environment interaction analysis in the context of a case-control study

Exposure*	Susceptibility genotype	Cases	Controls	Odds ratio†
–	–	*a*	*b*	1.0
–	+	*c*	*d*	$OR_g = bc/ad$
+	–	*e*	*f*	$OR_e = be/af$
+	+	*g*	*h*	$OR_{ge} = bg/ah$

* –, absent; +, present.

† Under an additive model: $OR_{ge} = OR_g + OR_e - 1$.
Under a multiplicative model: $OR_{ge} = OR_g \times OR_e$.

$$\mathrm{SIM} = \mathrm{OR}_{ge}/\mathrm{OR}_g \times OR_e$$

TABLE 4. Gene-environment interaction analysis in the context of a case-only study*

Exposure	Susceptibility genotype	
	−	+
−	*a*	*b*
+	*c*	*d*

* COR, case-only odds ratio = *ad*/*bc*. Under assumption of independence between exposure and genotype among controls: COR = $OR_{ge}/OR_e \times OR_g$ = SIM, where SIM is the synergy index.

$$COR = OR_{ge}/(OR_e \times OR_g) \times Z,$$

TABLE 3. Case-control analysis of the interaction between maternal cigarette smoking, transforming growth factor alpha (*Taq* I) polymorphism, and the risk of cleft palate. Adapted from Hwang et al. (11)

Smoking	*Taq* I polymorphism	No. of cases	No. of controls	Odds ratio*,†	95% confidence interval
−	−	36	167	1.0	Referent
−	+	7	34	1.0	0.3–2.4
+	−	13	69	0.9	0.4–1.8
+	+	13	11	5.5	2.1–14.6

* Crude odds ratios are presented.
† Odds ratio based on a case-only study is 5.1 (95% confidence interval 1.5–18.5) ((13 × 36)/(13 × 7)).

marked departure from multiplicative effects of the genotype and the exposure. The COR obtained from this analysis is 5.1, comparable with the SIM of 6.1 obtained from the regular case-control analysis. Also, the assumption of independence between exposure and genotype among controls is reasonable.

Reproducibility

Table 1. Examples of Some Reported Reproducibility Concerns in Preclinical Studies

Author	Field	Reported Concerns
Ioannidis et al (2009)[22]	Microarray data	16/18 studies unable to be reproduced in principle from raw data
Baggerly et al (2009)[23]	Microarray data	Multiple; insufficient data/poor documentation
Sena et al (2010)[24]	Stroke animal studies	Overt publication bias: only 2% of the studies were negative
Prinz (2011)[1]	General biology	75% to 80% of 67 studies were not reproduced
Begley & Ellis (2012)[2]	Oncology	90% of 53 studies were not reproduced
Nekrutenko & Taylor(2012)[25]	NGS data access	26/50 no access to primary data sets/software
Perrin (2014)[26]	Mouse, in-vivo	0/100 reported treatments repeated positive in studies of ALS
Tsilidis et al (2013)[27]	Neurological studies	Too many significant results, overt selective reporting bias
Lazic & Essioux (2013)[28]	Mouse VPA model	Only 3/34 used correct experimental measure
Haibe-Kains et al (2013)[29]	Genomics/cell line analysis	Direct comparison of 15 drugs and 471 cell lines from 2 groups revealed little/no concordant data
Witwer (2013)[30]	Microarray data	93/127 articles were not MIAME compliant
Elliott et al (2006)[31]	Commercial antibodies	Commercial antibodies detect wrong antigens
Prassas et al (2013)[32]	Commercial ELISA	ELISA Kit identified wrong antigen
Stodden et al (2013)[33]	Journals	Computational biology: 105/170 journals noncompliant with National Academies recommendations
Baker et al (2014)[34]	Journals	Top tier fail to comply with agreed standards for animal studies
Vaux (2012)[35]	Journals	Failure to comply with their own statistical guidelines

ALS indicates amyotrophic lateral sclerosis; MIAME, minimum information about a microarray experiment; NGS, next generation sequencing; and VPA, valproic acid (model of autism).

Begley, C.G. and J.P. Ioannidis, **Reproducibility in science: improving the standard for basic and preclinical research.** Circ Res, 2015. **116**(1): p. 116-26.

Grading the credibility of molecular evidence for complex diseases (1)

Table 1 Effect sizes in the pre-molecular era and in the molecular era[a]

Effect sizes	Putative frequency	Typical examples of postulated risk factors	
		Pre-molecular era	Molecular era
Large (RR > 5)	Rare	Smoking and lung cancer	APOE and Alzheimer's disease[31]
			BRCA1 and breast cancer[32]
Moderate (RR 2–5)	Uncommon	Moderate obesity and cholesterol gallstones	NOD2 and Crohn's disease[33]
			HLA shared epitopes and rheumatoid arthritis[34]
Small (RR 1.2–2)	Common	Racial descent and hypertension	FcγRIIa and SLE[35]
			GSTM1 and bladder cancer[36]
Very small (RR 1–1.2)	Unclear frequency[a]	Passive smoking and lung cancer	GSTM1 and lung cancer[37]
			MTHFR and ischaemic stroke[38]

RR: relative risk.

[a] Presented examples reflect current state of knowledge and are subject to possible refutation in the future; for small and very small effect sizes, it is uncertain whether these risk factors are true, even when evidence is based on large sample sizes from several studies.

Ioannidis, J.P., **Commentary: grading the credibility of molecular evidence for complex diseases.** Int J Epidemiol, 2006. **35**(3): p. 572-8; discussion 593-6.

Grading the credibility of molecular evidence for complex diseases (2)

Table 2 Typical credibility of research findings according to effect size and extent of replication

Effect size (relative risk)	Replication	Typical credibility (%)
Large (>5)	None	10–60
	Limited	30–80
	Extensive	70–95
Moderate (2–5)	None	5–20
	Limited	10–40
	Extensive	50–90
Small (1.2–2)	None	<5
	Limited	2–20
	Extensive	10–70
Very small (1–1.2)	None	<1
	Limited	1–5
	Extensive	2–30

Ioannidis, J.P., **Commentary: grading the credibility of molecular evidence for complex diseases.** Int J Epidemiol, 2006. **35**(3): p. 572-8; discussion 593-6.

Table 3 Proposed grading of credibility in molecular evidence

First axis: Effect size		
	1.1	Very small or small effect size (relative risk < 2)
	1.2	Moderate effect size (relative risk 2–5)
	1.3	Large effect size (relative risk > 5)
Second axis: Amount and replication of evidence		
	2.1	Single or few scattered studies
	2.2	Meta-analyses of group data
	2.3	Large-scale evidence from inclusive networks
Third axis: Protection from bias		
	3.1	Clear presence of strong bias in the evidence
	3.2	Uncertain about the presence of bias
	3.3	Clear strong protection from bias
Fourth axis: Biological credibility		
	4.1	No functional/biological data or negative data
	4.2	Limited or controversial functional/biological data
	4.3	Convincing functional/biological data
Fifth axis: Relevance		
	5.1	No clinical or public health applicability
	5.2	Limited clinical or public health applicability
	5.3	Considerable clinical/public health applicability

Ioannidis, J.P., **Commentary: grading the credibility of molecular evidence for complex diseases.** *Int J Epidemiol,* 2006. **35**(3): p. 572-8; discussion 593-6.

References

- Lewis CM. Genetic association studies: design, analysis and interpretation. Brief Bioinform. 2002 Jun;3(2):146-53.
- Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7(10):781-91.
- Cordell HJ, Clayton DG. Genetic association studies. Lancet. 2005;366(9491):1121-31.
- Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001, 358: 1356-1360
- Bagos PG. Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis. Statistical Applications in Genetic and Molecular Biology, 2013
- Ziegler, A., Konig, I.R. and Thompson, J.R. (2008) Biostatistical aspects of genome-wide association studies, Biom J, 50, 8-28
- Ioannidis, J.P., Commentary: grading the credibility of molecular evidence for complex diseases. Int J Epidemiol, 2006. 35(3): p. 572-8; discussion 593-6.