# Analysis of imputed data

Yurii Aulchenko
Independent researcher & consultant
yurii [dot] aulchenko [at] gmail [dot] com

# Areas of expertise

- (infra)Structure and project evaluation and advise
- Methodological advise (study design, planning of analyses, methods, software)
- Methods, algorithms and software development
- Data analyses
- Teaching and training

# Analysis of imputed genotypic data in GWAS

- Short review of standard methods
- Methods for analysis of imputed data
- Open questions and problems

# Singe SNP analysis

- Analysis of each SNP in turn independent of others
- For each SNP, regression is performed, resulting in estimates of **regression coefficients**, their **standard errors** and the **p-value**

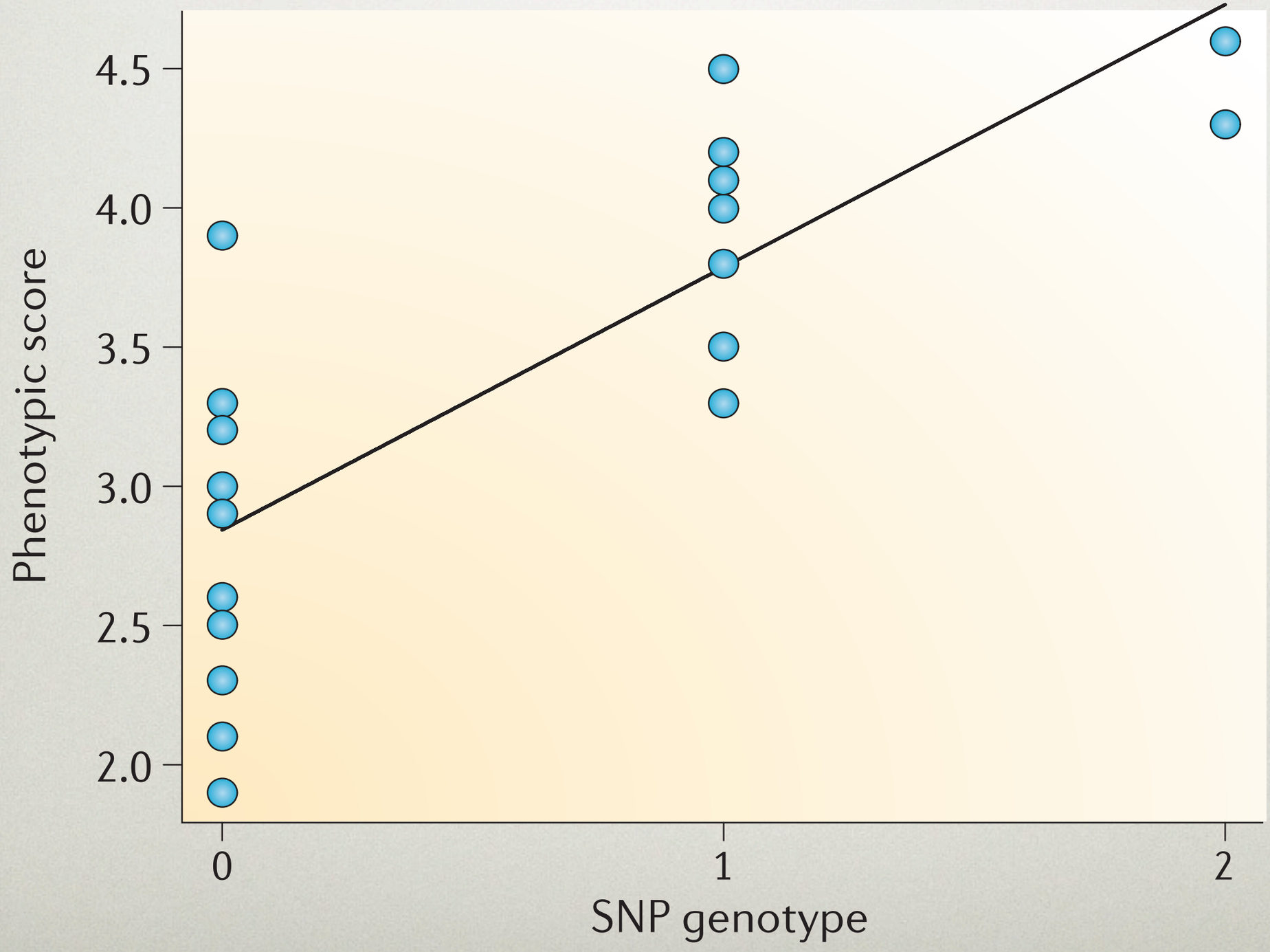
# Linear regression model

The value of the trait in *i*-th individual is assumed to follow linear model

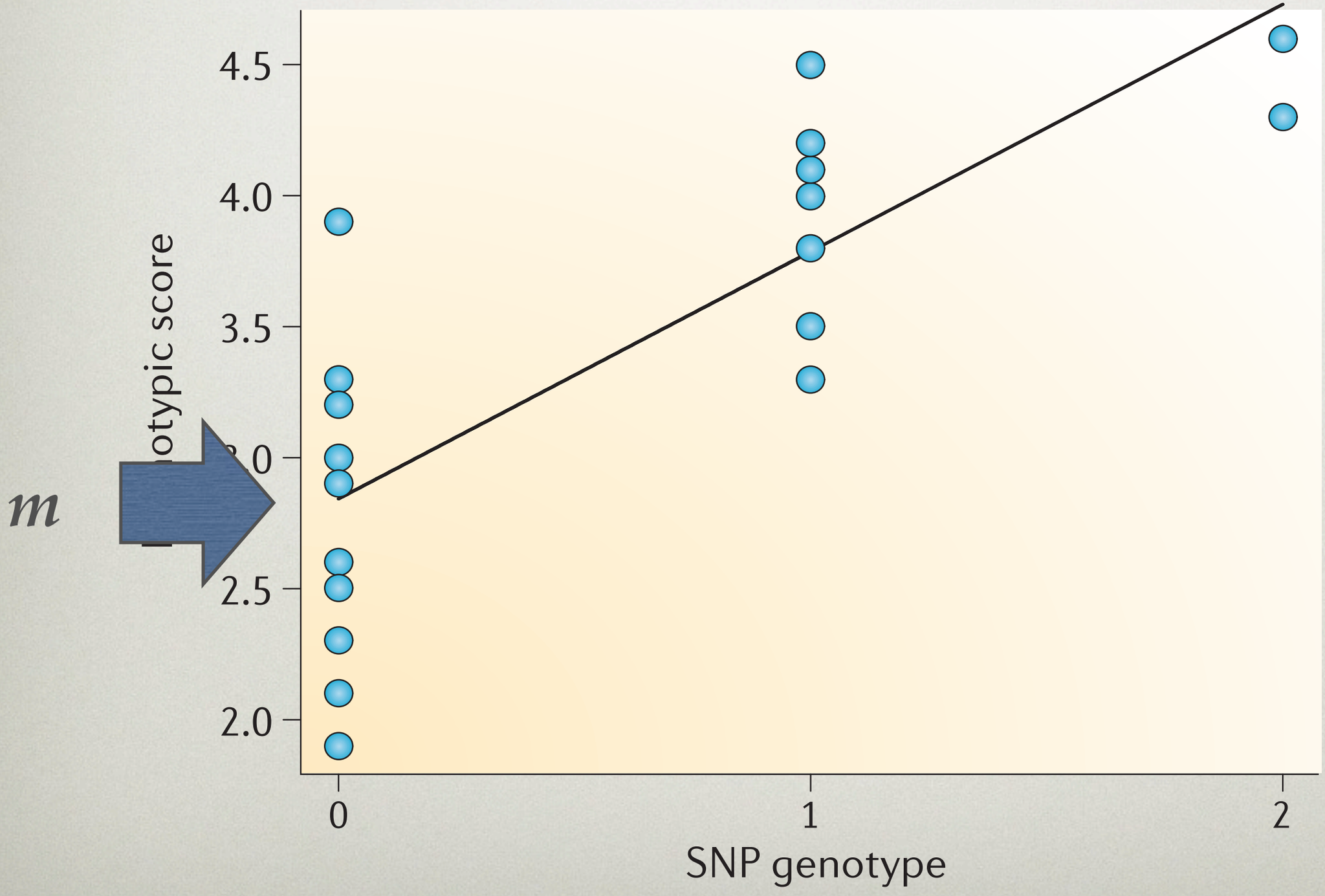
$$Y_i = m + b_g g_i + e_i$$

where $m$ is intercept, $g_i$ is the genotypic value (coded as 'B' allele dose - 0, 1 or 2), and $e_i$ is random residual error

# Linear regression

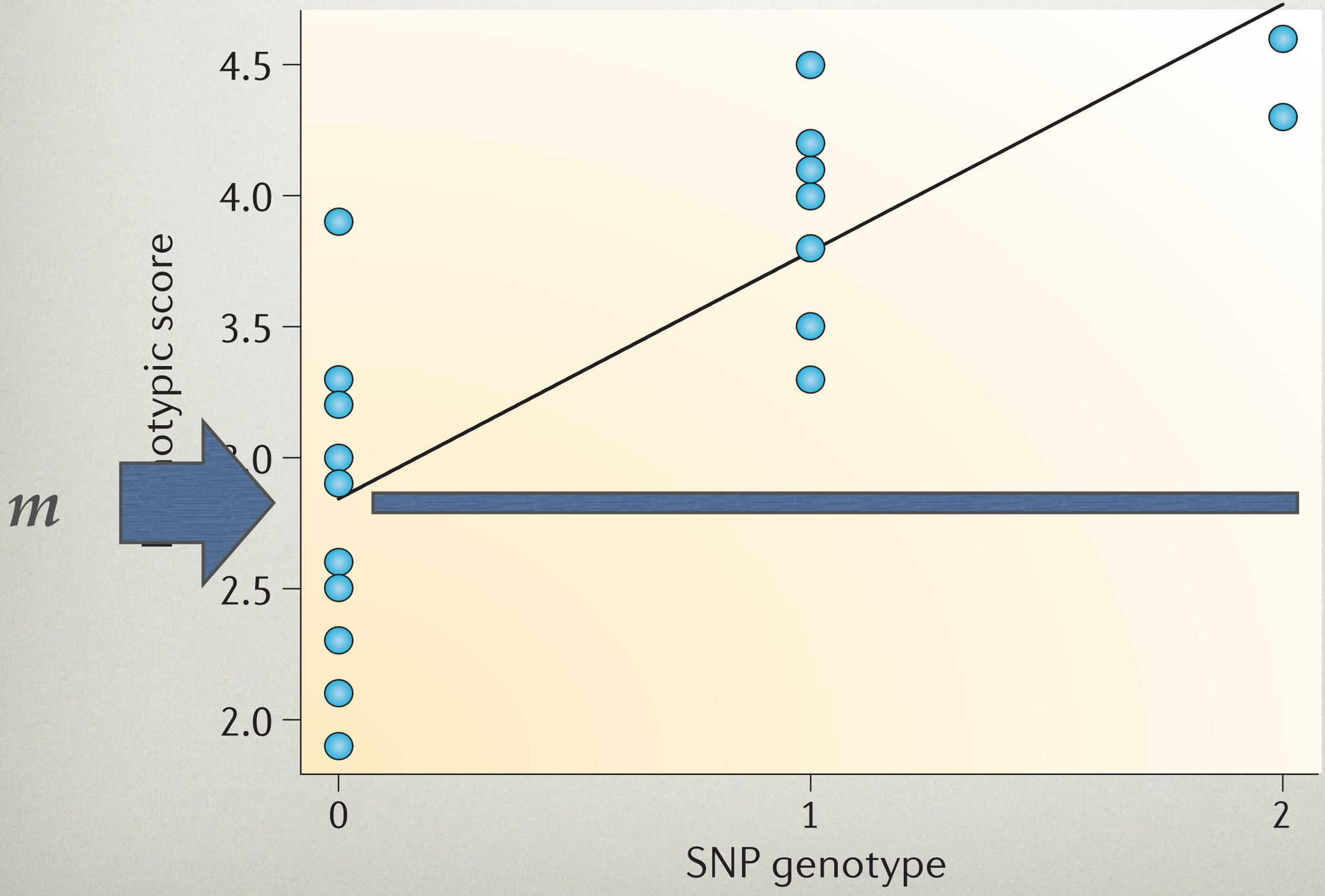


# Linear regression

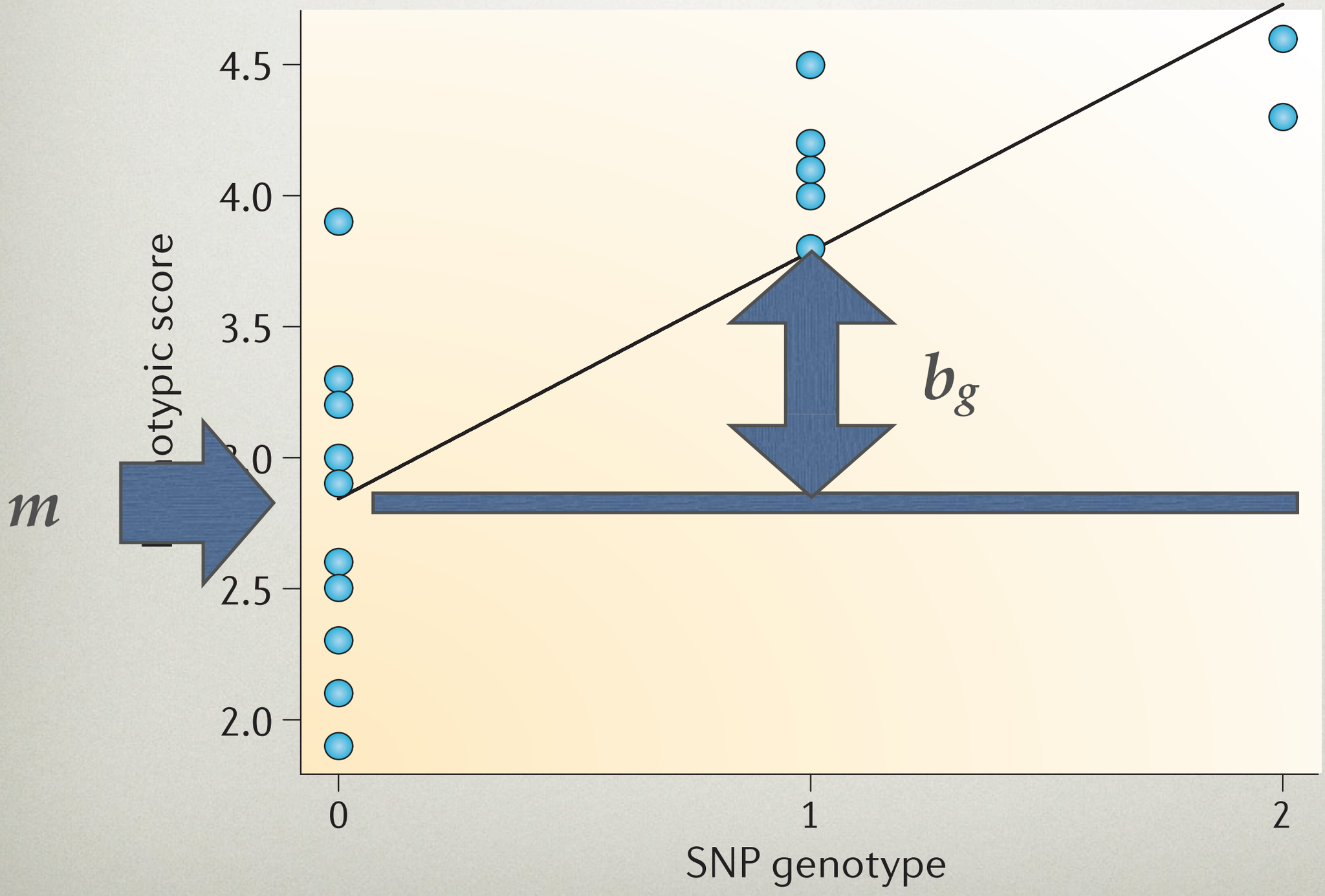



$m$

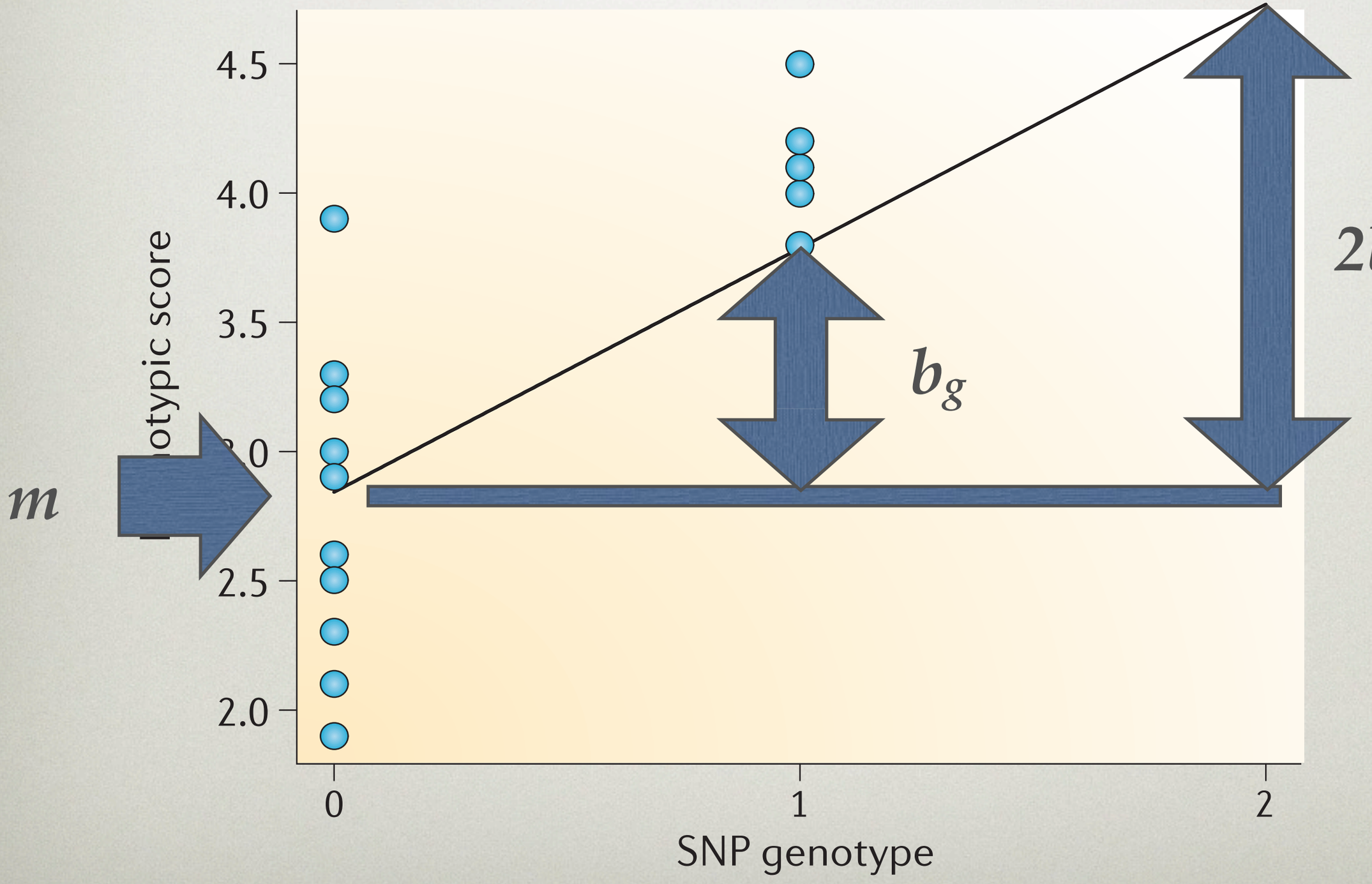
# Linear regression



# Linear regression



# Linear regression



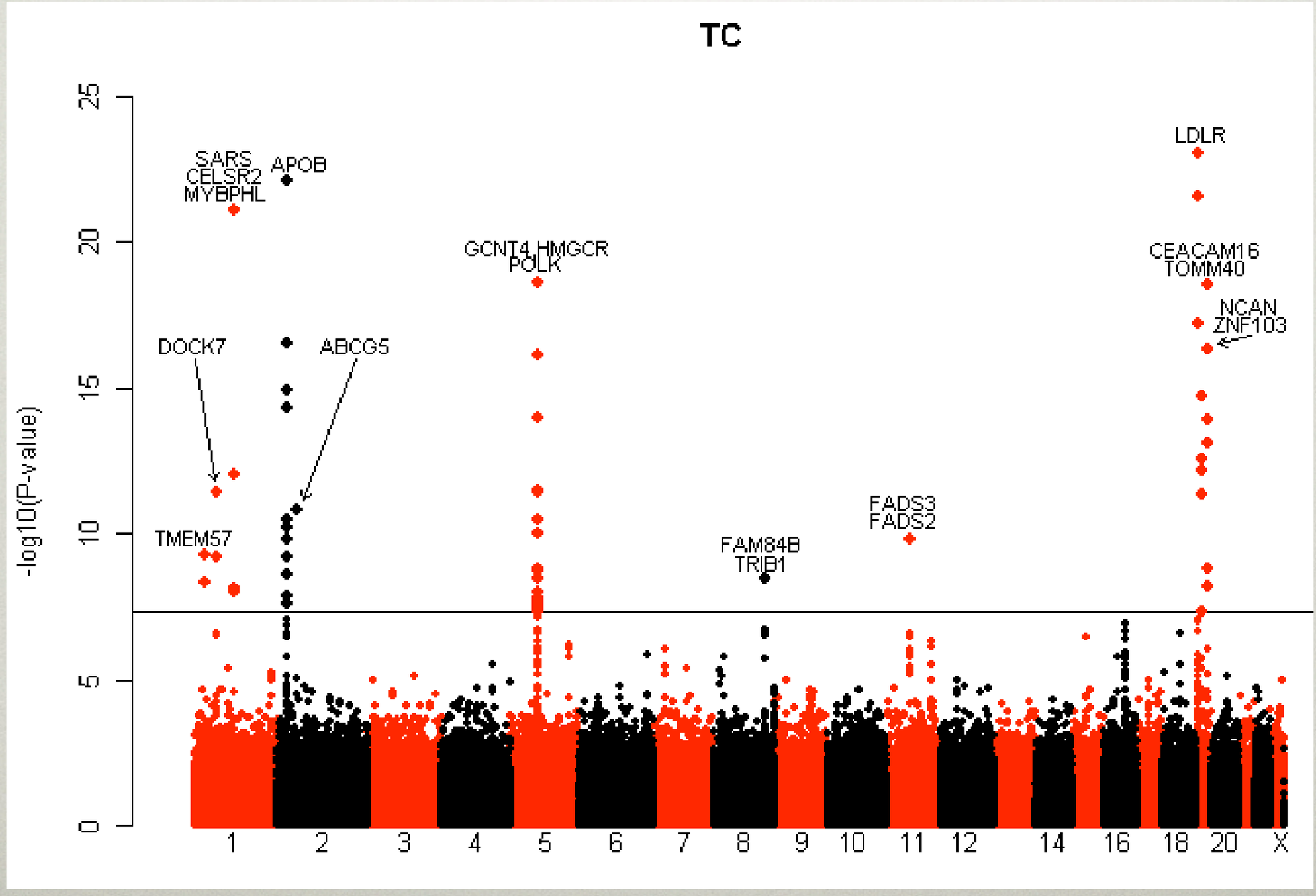
# Significance testing

The estimate of $b_g$ and its standard error $s_g$ are computed using standard methods

Under the null hypothesis, the test statistic $T^2=(b_g/s_g)^2$ is distributed as chi-squared with 1 degree of freedom (1df)

For specific $T^2$ we know the $p$-value: the probability that $b_g$ deviates from zero purely by chance

# As a result...


TC
-log10(P-value)
0
5
10
15
20
25
SARS
CELSR2
MYBPHL
APOB
DOCK7
ABCG5
TMEM57
GCNT4 HMGCR
POLK
FAM84B
TRIB1
FADS3
FADS2
LDLR
CEACAM16
TOMM40
NCAN
ZNF103
1
2
3
4
5
6
7
8
9
10
11
12
14
16
18
20
X

# Multiple testing

- Hence nominal single test p-value corresponding to experiment-wise type 1 error rate of 5% is << 0.05
- Usual practice is to use a fixed threshold of $5 \times 10^{-8}$
- Note: this threshold is defined for GWAS of *common variants* in a population of *European ancestry*

# Non-additive models

... can be specified using linear model

$$Y_i = m + b_1 I(g_i=1) + b_2 I(g_i=2) + e_i$$

where $I(g_i=k)$ is an indicator variable taking the value of 1 if $g_i$ is equal to $k$ and zero otherwise

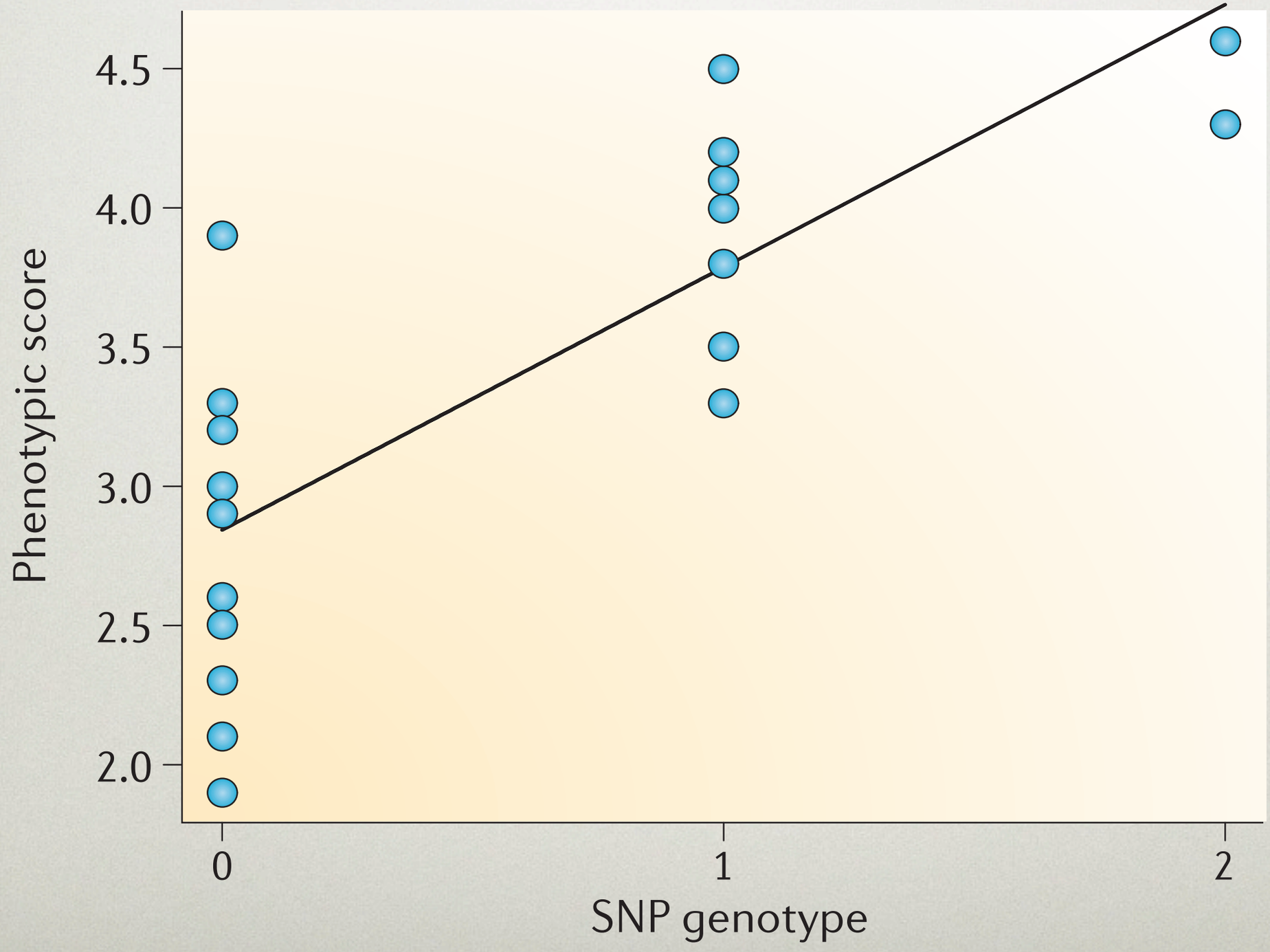
# Non-additive models

In other words, the expected value of the trait for the genotype

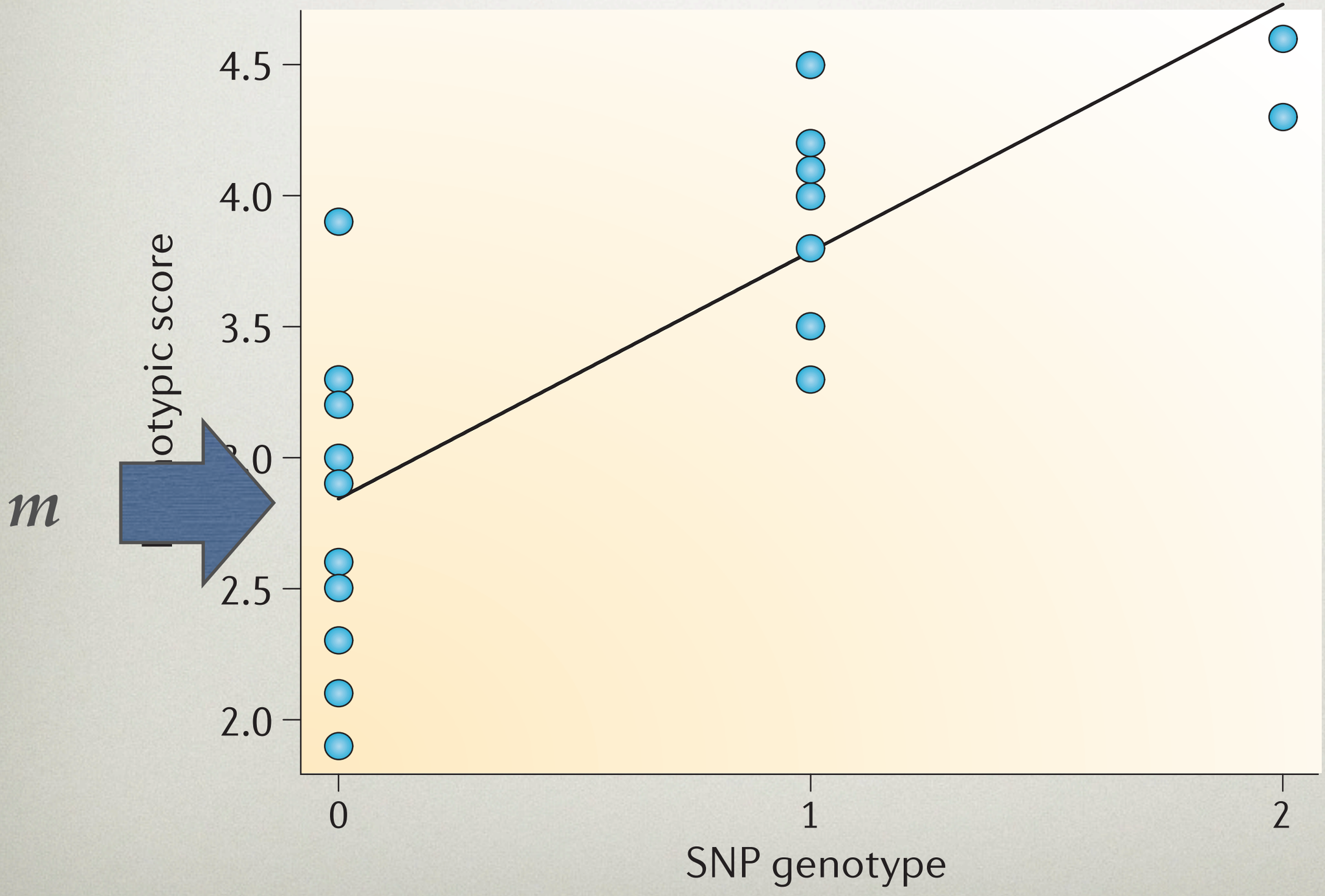
- AA ($g_i$=0) is $m$
- AB ($g_i$=1) is $m + b_1$
- BB ($g_i$=2) is $m + b_2$

By varying $m$, $b_1$ and $b_2$, any three genotypic means can be fit

# Linear regression

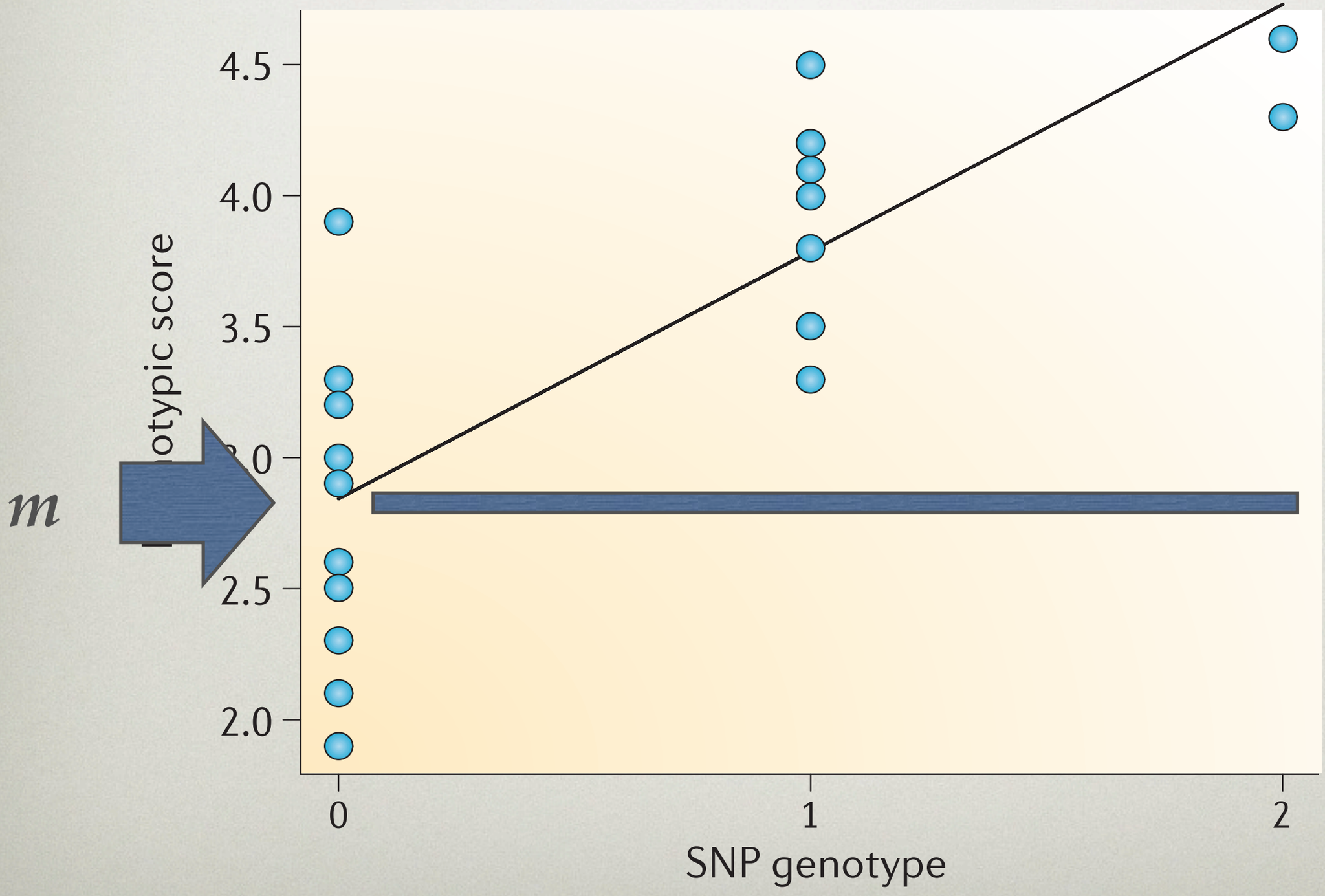


# Linear regression

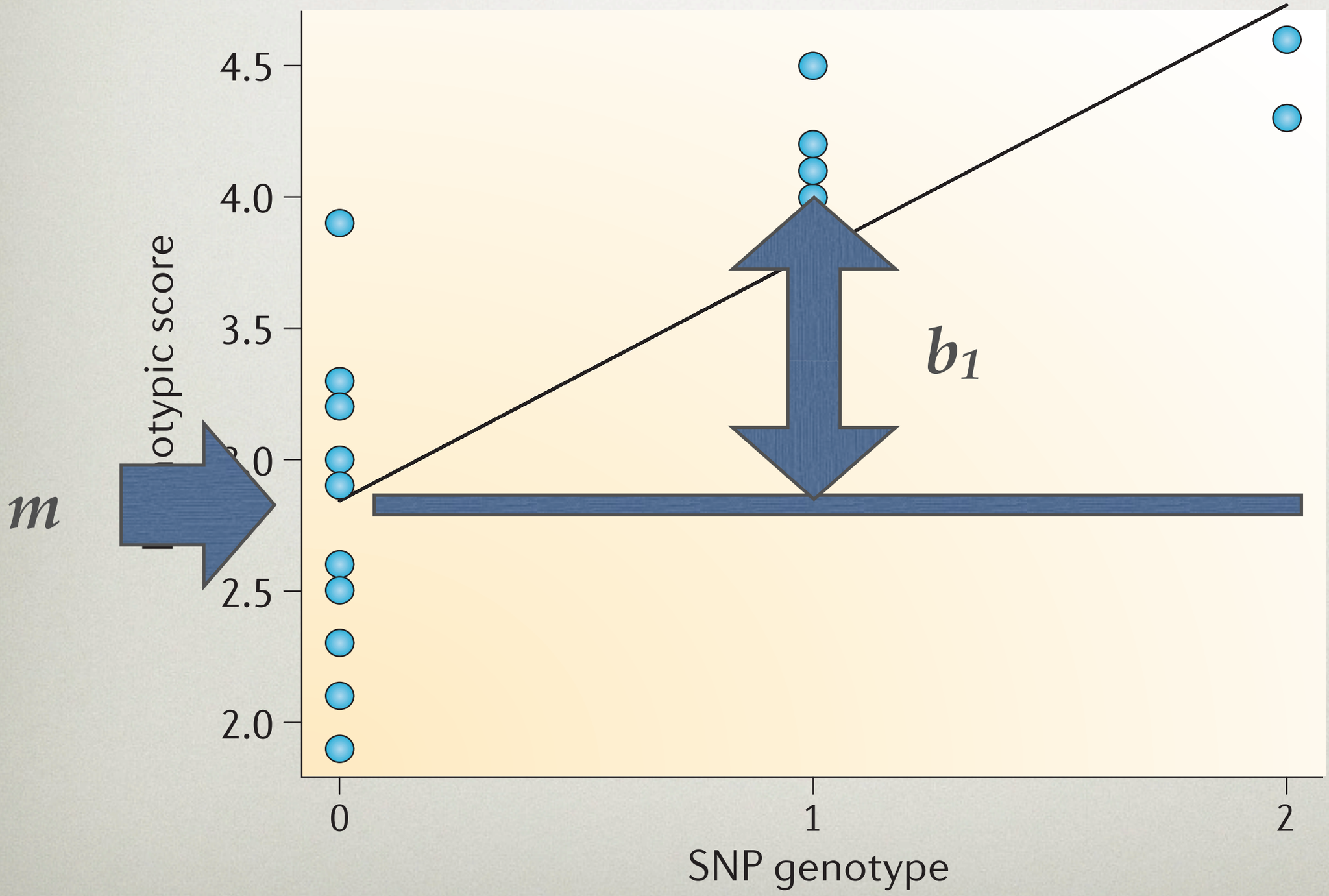



$m$

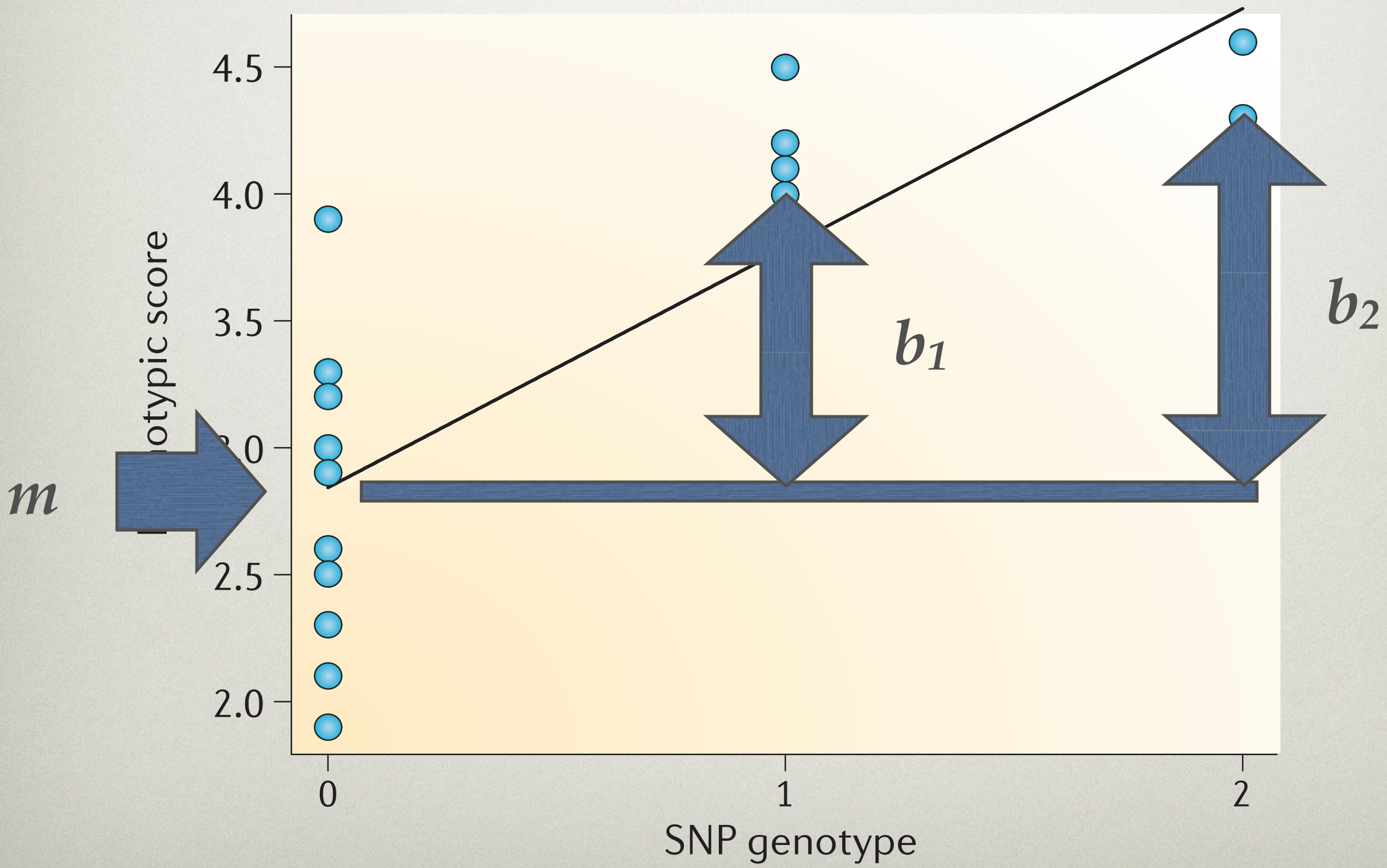
# Linear regression



# Linear regression



# Linear regression



# Other 1df models

Dominant B: $b_1 = b_2$

Recessive B: $b_1 = 0$

Over-dominant: $b_2 = 0$

... and additive: $b_2 = 2\ b_1$

# Interaction models

The value of the trait in $i$-th individual is assumed to follow linear model

$$Y_i = m + b_f F_i + b_g g_i + b_{fg} F_i g_i + e_i$$

where $m$ is intercept, $F_i$ is the value of some "factor", $g_i$ is the genotypic value, and $e_i$ is random residual error

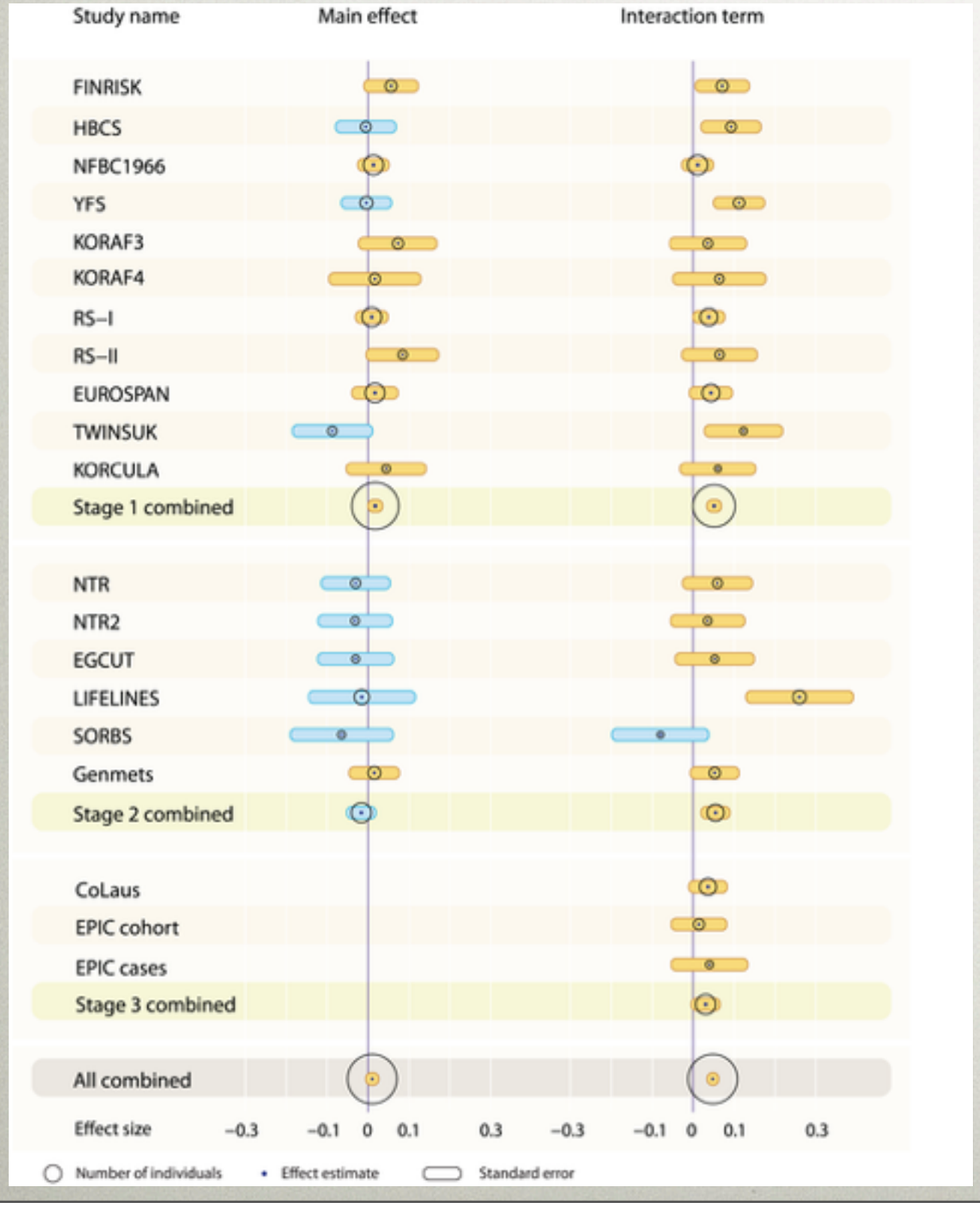
# What could "F" be?

- An environment (gene-environment interaction)
- Indicator of transmitting parent (imprinting models)
- Other genotype (gene-gene)
- ... etc.

OPEN ACCESS Freely available online

PLoS GENETICS

# A Genome-Wide Screen for Interactions Reveals a New Locus on 4p15 Modifying the Effect of Waist-to-Hip Ratio on Total Cholesterol




- A meta-analysis of genome-wide association (GWA) data from 18 population-based cohorts with European ancestry (maximum *N* = 32,225).
- Eight further cohorts (*N* = 17,102) for replication
- SNP *rs6448771* demonstrated genome-wide significant interaction with waist-to-hip-ratio (WHR) on total cholesterol (TC) with a combined *P*-value of $4.79 \times 10^{-9}$

# Analysis of imputed genotypic data in GWAS

- Short review of standard methods
- **Methods for analysis of imputed data**
- Open questions and problems

# Imputed data

We can not tell the exact genotype, but can estimate posterior probability distribution: $P(g)=\{p_{AA},p_{AB},p_{BB}\}$

Directly typed SNPs: either AA, AB or BB. The probability distribution is *degenerate* (e.g. {*0,1,0*} that is to say AB)

For imputed SNPs, the distribution is not degenerate

# Imputing: guess the "?"

Reference Haplotypes

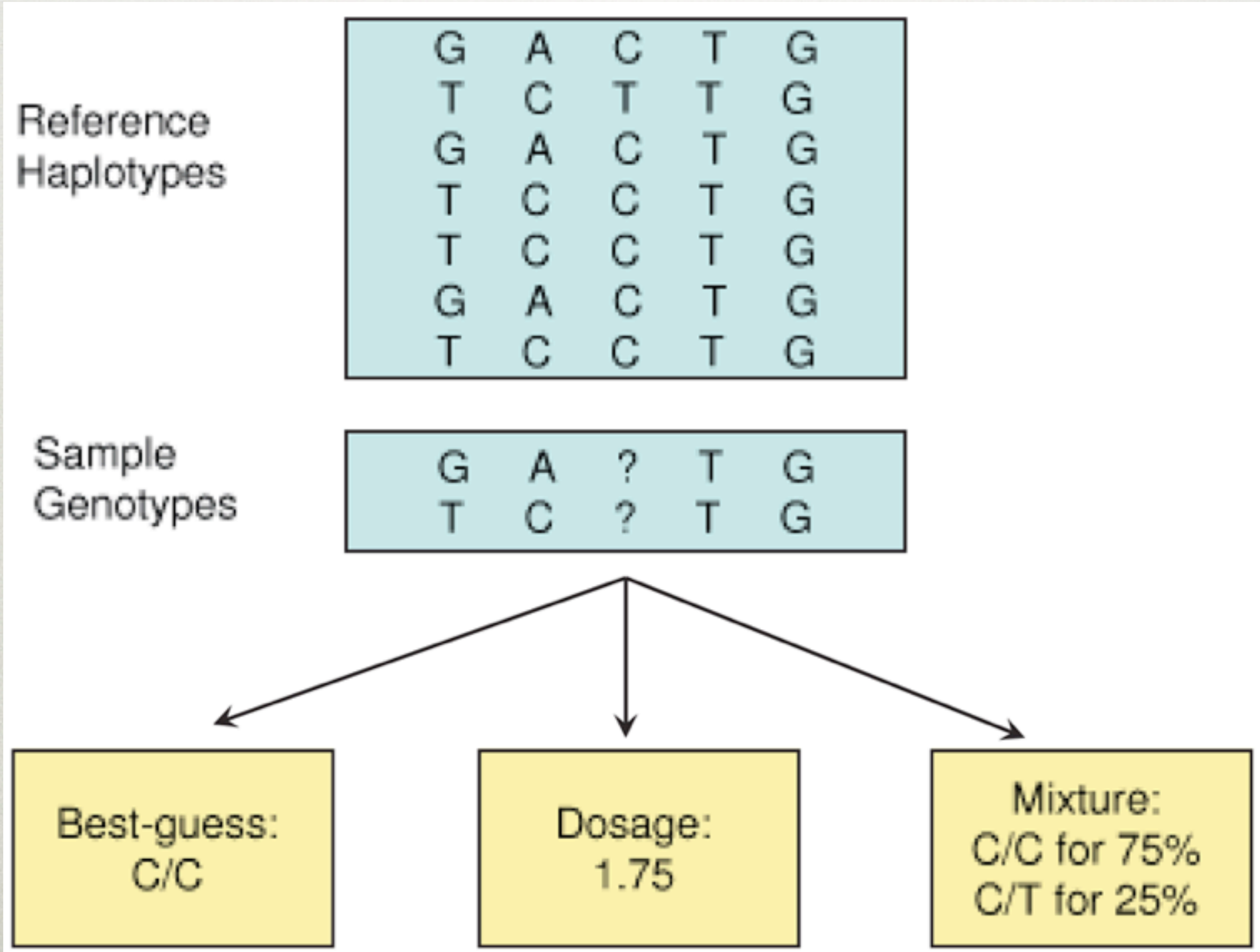
| | | | | |
|---|---|---|---|---|
| G | A | C | T | G |
| T | C | T | T | G |
| G | A | C | T | G |
| T | C | C | T | G |
| T | C | C | T | G |
| G | A | C | T | G |
| T | C | C | T | G |

Sample Genotypes

| | | | | |
|---|---|---|---|---|
| G | A | ? | T | G |
| T | C | ? | T | G |

Zheng et al., 2011

# Imputing: guess the "?"




Zheng et al., 2011

# How can we analyze imputed data?

- Instead of genotypes, we have probabilities that certain person at certain locus has certain genotype
- ???

# Best guess

- Best guess: take the genotype with maximal posterior probability and treat it as if it was true, directly typed
- Drawback: biased estimates, reduced power

# Regression onto probabilities

Use the model

$$E[Y_i] = m + b_1 P(g_i=1) + b_2 P(g_i=2)$$

Note this is very similar to model for directly typed SNPs, with probabilities used instead of indicator variables. Different genetic models can be formulated in the same way.

# Maximum likelihood based on probabilities

Define individual likelihood as

$$L_i = SUM_{gi=\{0,1,2\}}\ P(g_i)\ P(Y_i \mid g_i)$$

Where

$$P(Y_i \mid g_i) = Normal(E[Y_i \mid g_i], s2)$$

and

$$E[Y_i \mid g_i] = m + b_1\ I(g_i=1) + b_2\ I(g_i=2)$$

# Maximum likelihood based on probabilities

Define joint likelihood as the product of individual likelihoods

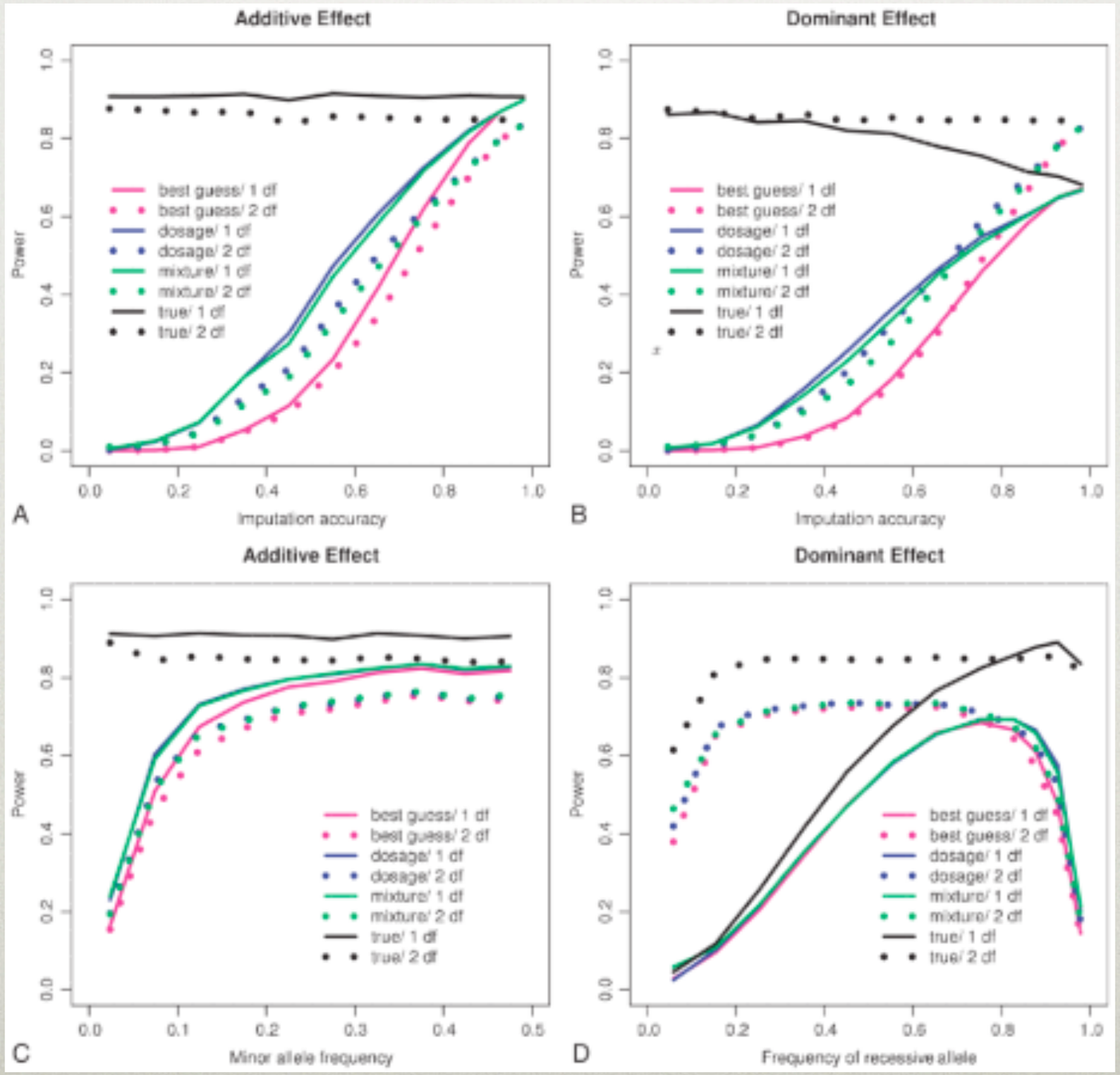
Maximize the likelihood over the parameters involved

Maximum Likelihood Ratio test can be used to test nested models and draw statistical inferences

# Power in large samples (small effects)

vs.
accuracy

vs. MAF

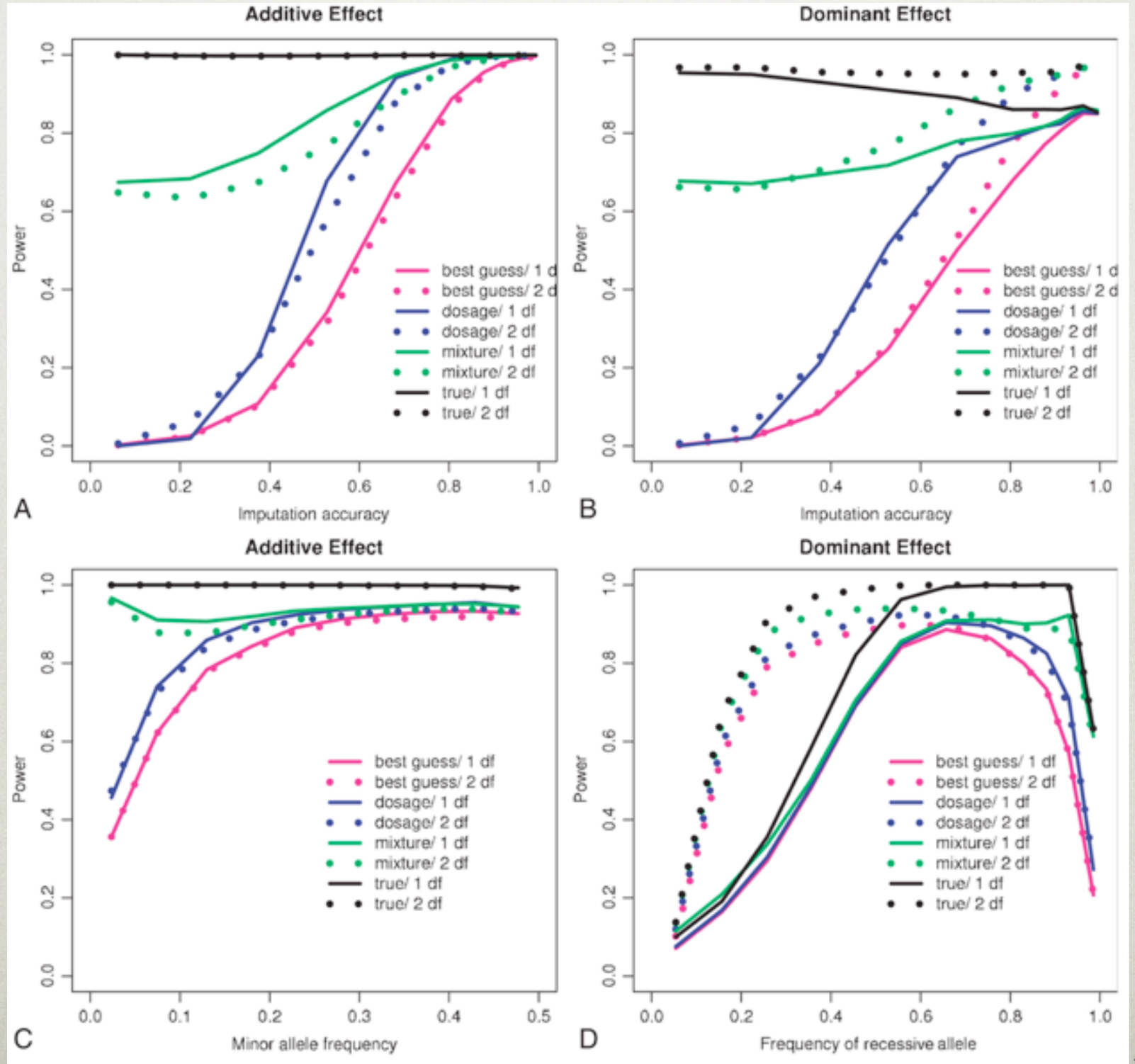



Zheng et al., 2011

# Power in small samples (large effects)

vs. accuracy

vs. MAF




Zheng et al., 2011

# Analysis of imputed genotypic data in GWAS

- Short review of standard methods
- Methods for analysis of imputed data
- **Open questions and problems**

# Additional problems

- Multi-locus analysis: is problematic as no information about joint distribution of genotypes is retained after standard imputation procedures
- GxE analysis (and expected for other interaction analyses) represents a methodological challenge

# λ's going all the way around 1

- Rotterdam study: **population-based** cohort used for genetic research for over 15 years
- In GWAS performed over many traits, always λ < 1.05
- GxE results for some traits:

| | Environmental factor | | | |
|---|---|---|---|---|
| | cov 1 | cov 2 | cov 3 | cov 4 |
| trait 1 | 1.13 | 1.13 | 1 | 1.14 |
| trait 2 | 0.98 | 1.04 | 1.02 | 1.04 |
| trait 3 | 1.12 | **1.22** | 1 | 1.09 |
| trait 4 | 1.05 | 1.01 | 1.03 | 0.97 |
| trait 5 | 1.1 | 1.09 | 1.07 | 1.01 |
| trait 6 | 1.02 | 1.01 | 0.92 | 1.03 |
| trait 7 | 0.94 | 0.95 | **0.89** | 1 |

# Solution: use robust (co)variances

- Solution coming from Thomas Lumley
- Implemented in ProbABELv0.1-1 (Aulchenko *et al.*, BMC Bioinformatics, 2010)

| | Environmental factor | | | |
|---|---|---|---|---|
| | cov 1 | cov 2 | cov 3 | cov 4 |
| trait 1 | 1.03 | 1.04 | 1.03 | 1.02 |
| trait 2 | 1.03 | 1.01 | 1.03 | 1.02 |
| trait 3 | 1.02 | 1.04 | 1.03 | 1.02 |
| trait 4 | 1.04 | 1.03 | 1.03 | 1.01 |
| trait 5 | 1 | 1.02 | 1.03 | 1.01 |
| trait 6 | 1.03 | 1.01 | 1.02 | 1.01 |
| trait 7 | 1.02 | 1.03 | 1.03 | 1.01 |

- Still not 100% satisfactory!
- What about Mixed Models?

# Conclusions

- Using regression onto genotype probabilities is a valid and powerful method for standard scenarios
- Use of ML/mixture method can give extra power in case of small samples and large effects
- Caution should be exercised in interaction analyses with imputed data