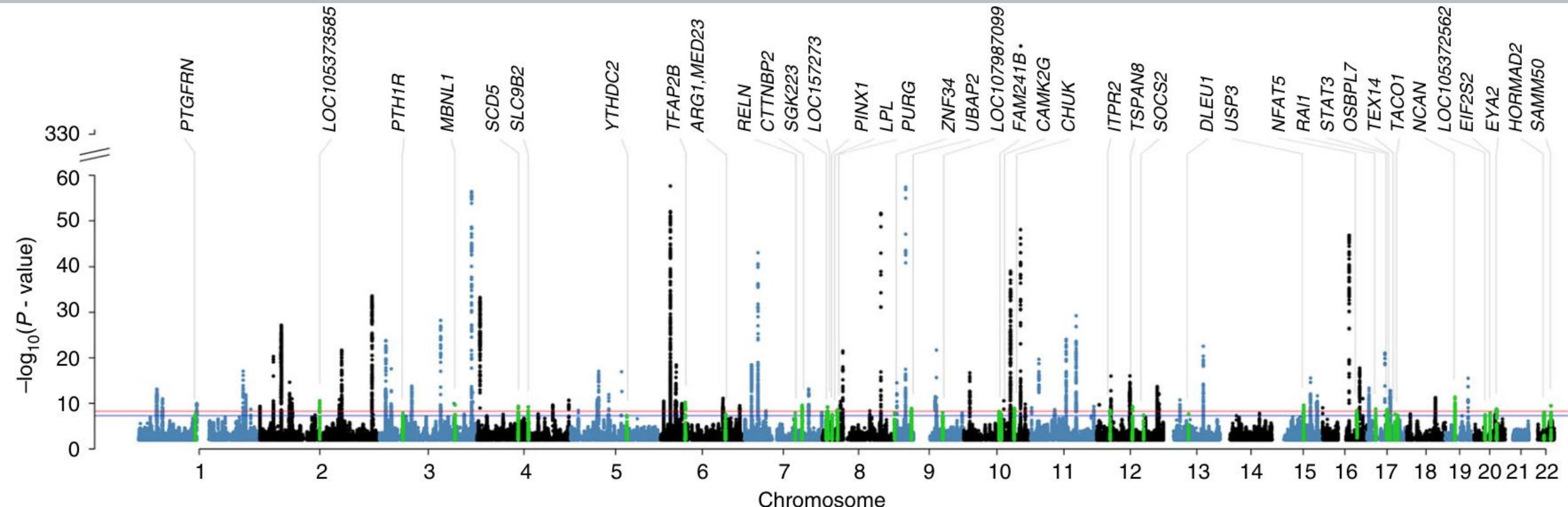


Genome-wide Association Studies (GWAS)

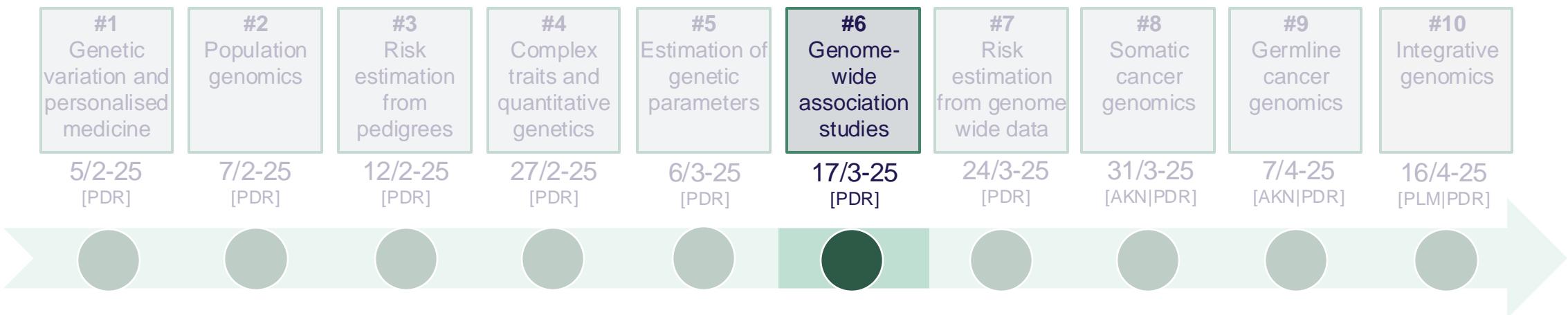
#6

PALLE DUUN ROHDE

palleldr@hst.aau.dk



LETS GET STARTED



AGENDA

- 08:15 – 08:30** Recap [*Complex traits and genetic parameters*]
- 08:30 – 09:00** Group presentations from last
- 09:00 – 09:15** Break
- 09:15 – 09:45** Lecture 1 [*Genetic associations + GWAS part 1*]
- 09:45 – 10:15** Exercise 1 + 2
- 10:15 – 10:30** Break
- 10:30 – 11:00** Lecture 2 [*GWAS part 2*]
- 11:00 – 11:55** Group work
- 11:55 – 12:00** Evaluation at Moodle

AGENDA

- | | |
|----------------------|---|
| 08:15 – 08:30 | Recap [<i>Complex traits and genetic parameters</i>] |
| 08:30 – 09:00 | Group presentations from last |
| 09:00 – 09:15 | Break |
| 09:15 – 09:45 | Lecture 1 [<i>Genetic associations + GWAS part 1</i>] |
| 09:45 – 10:15 | Exercise 1 + 2 |
| 10:15 – 10:30 | Break |
| 10:30 – 11:00 | Lecture 2 [<i>GWAS part 2</i>] |
| 11:00 – 11:55 | Group work |
| 11:55 – 12:00 | Evaluation at Moodle |

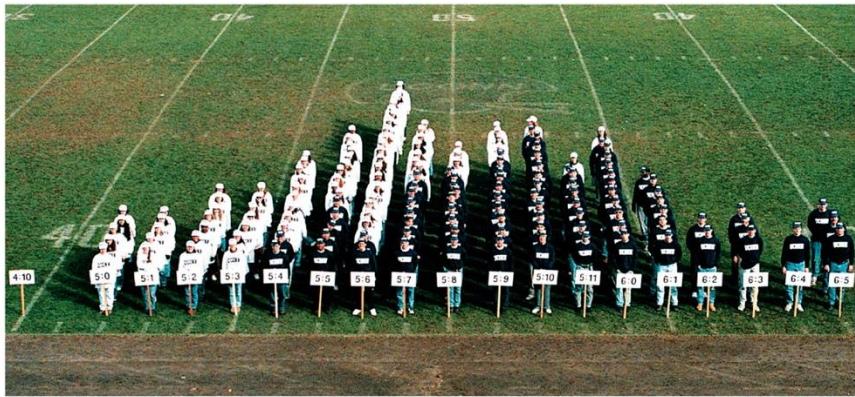
MULTIFACTORIAL INHERITANCE

- ❖ Monogenic (single gene/variant)
- ❖ Polygenic (multiple gene variants at the same time)
- ❖ Multifactorial (multiple gene variants + environment variation)



MULTIFACTORIAL TRAITS

Continuous variation

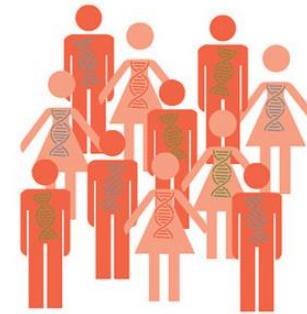


Categorical variation

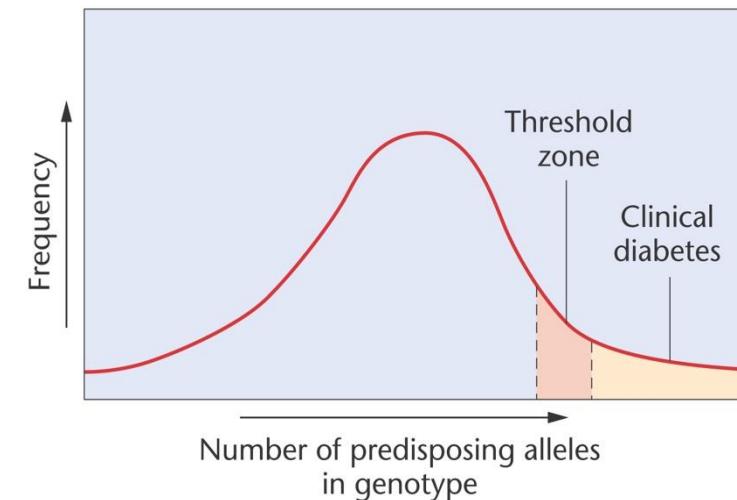


Dichotomous outcome

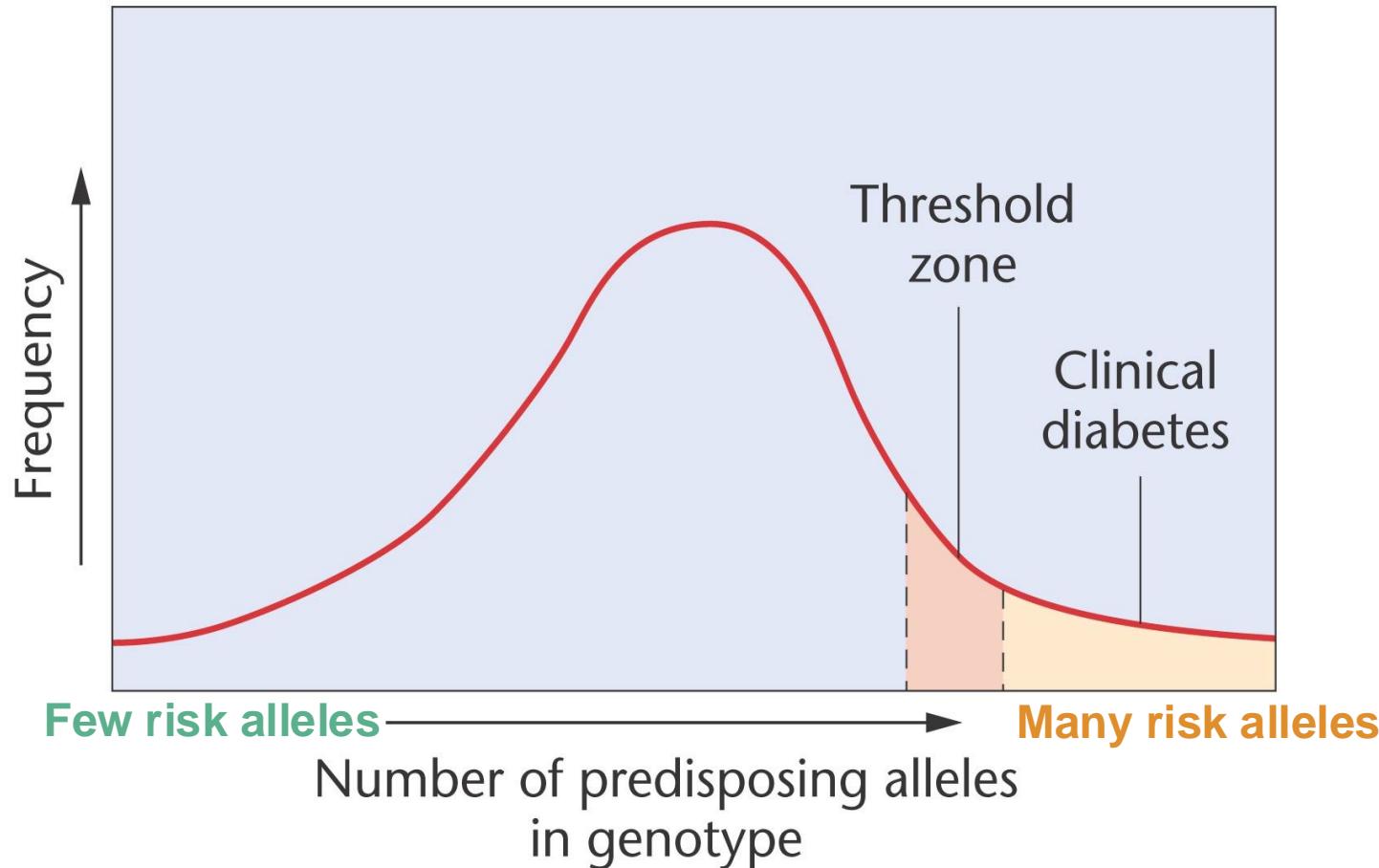
People without condition



People with condition



LIABILITY (THRESHOLD) MODEL



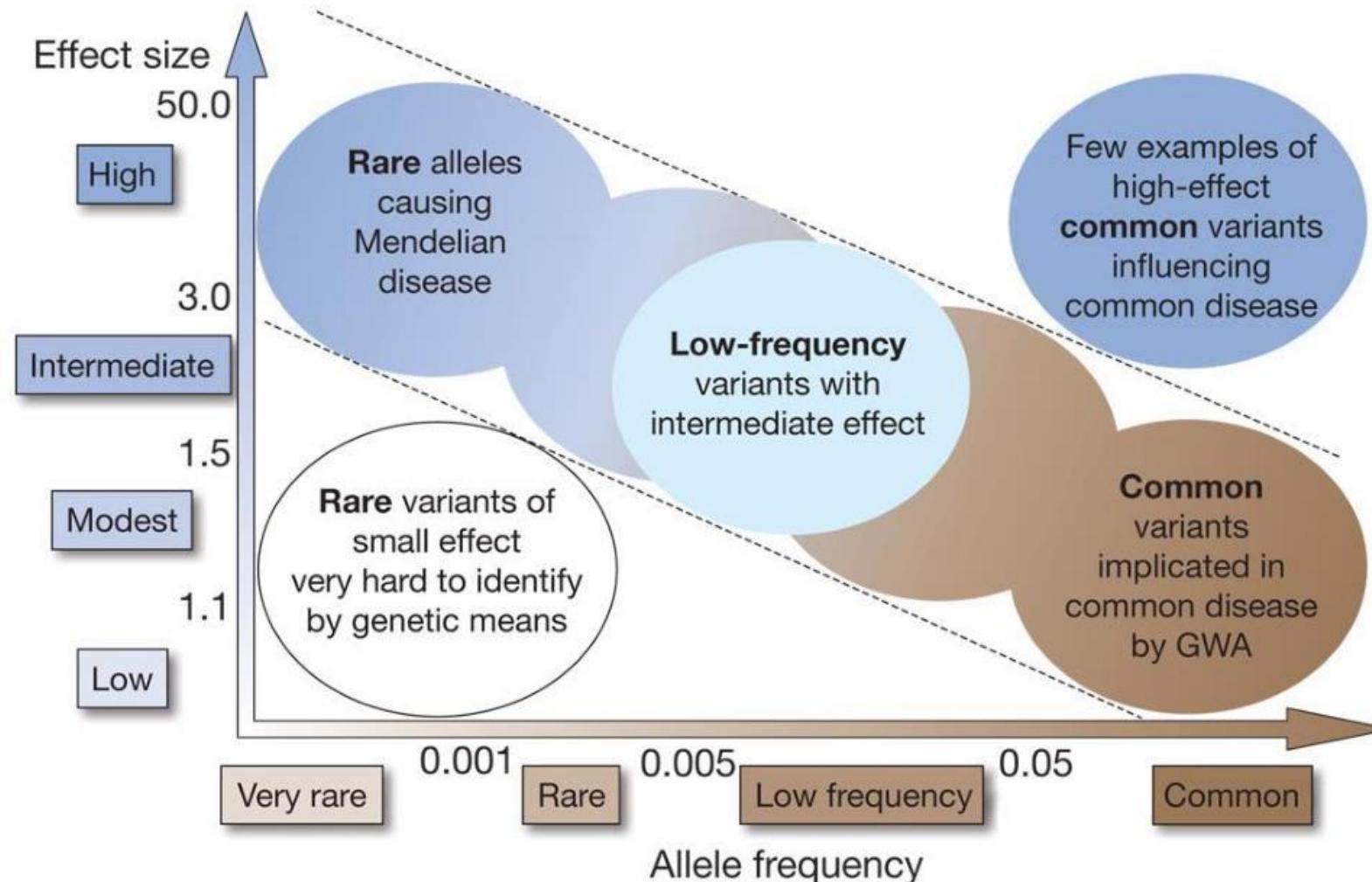
Liability model

Only individuals with a liability over a certain threshold will become affected

The **sum** of many genetic variants with **small effect/risk**.

Each locus follow Mendelian inheritance pattern, although the trait does not

GENETIC VARIATION



Multifactorial traits are caused
by the sum of MANY variants
exerting small effects

MULTIFACTORIAL TRAITS

Multifactorial traits = polygenic effect + environmental effect

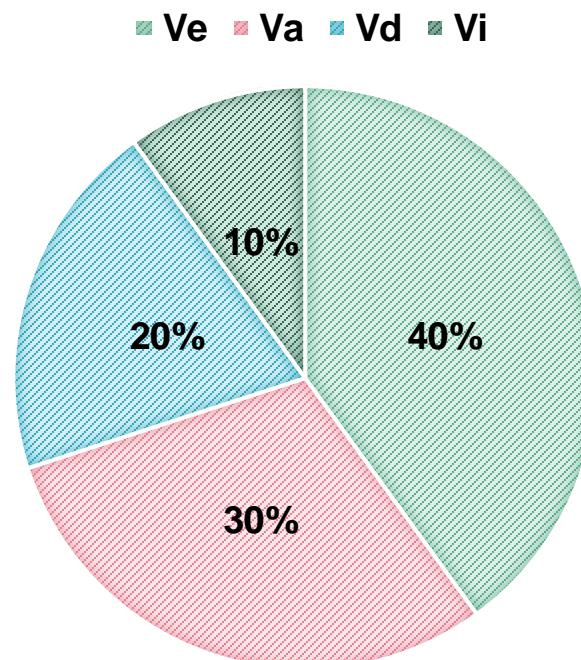
many genes/alleles

$$V_P = V_G + V_E$$

$$V_P = V_a + V_d + V_i + V_e$$

$$H^2 = V_G / (V_G + V_E)$$

$$h^2 = V_a / (V_a + V_E)$$



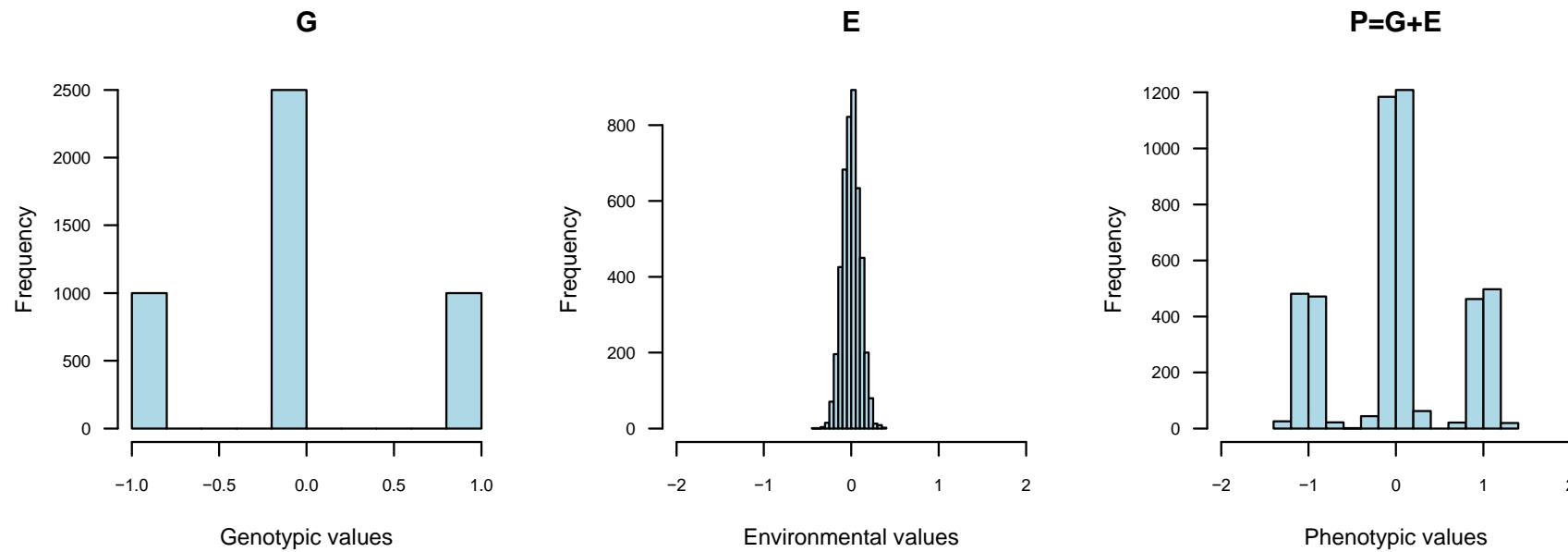
External effects that modulates the phenotypic value.

... or things that we cannot explain



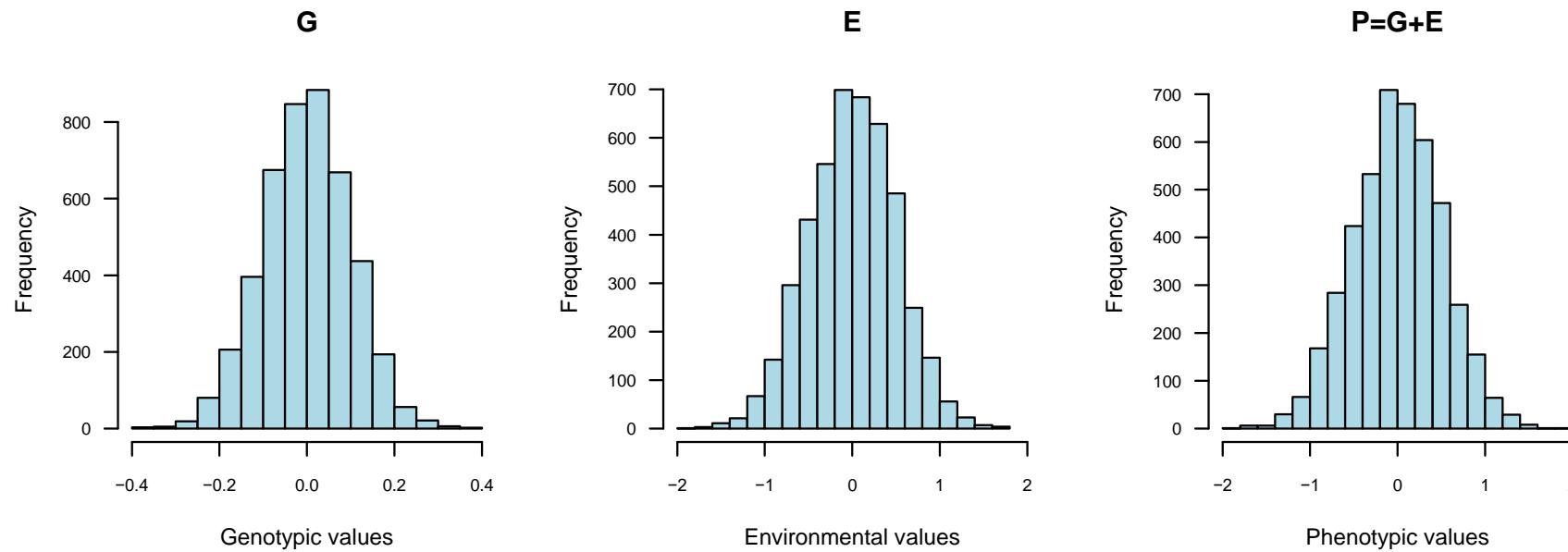
MULTIFACTORIAL TRAITS

Environmental variance (non-genetic factors) blurs phenotypic classes



MULTIFACTORIAL TRAITS

Genotypic and environmental variance creates infinite many phenotypic classes



BROAD-SENSE HERITABILITY

Broad-sense **heritability (H^2) describes the proportion of the phenotypic variance that is explained by genetic difference variation.**

$$V_P = V_G + V_E$$

$$H^2 = \frac{V_G}{V_P} = \frac{V_G}{V_G + V_E}$$

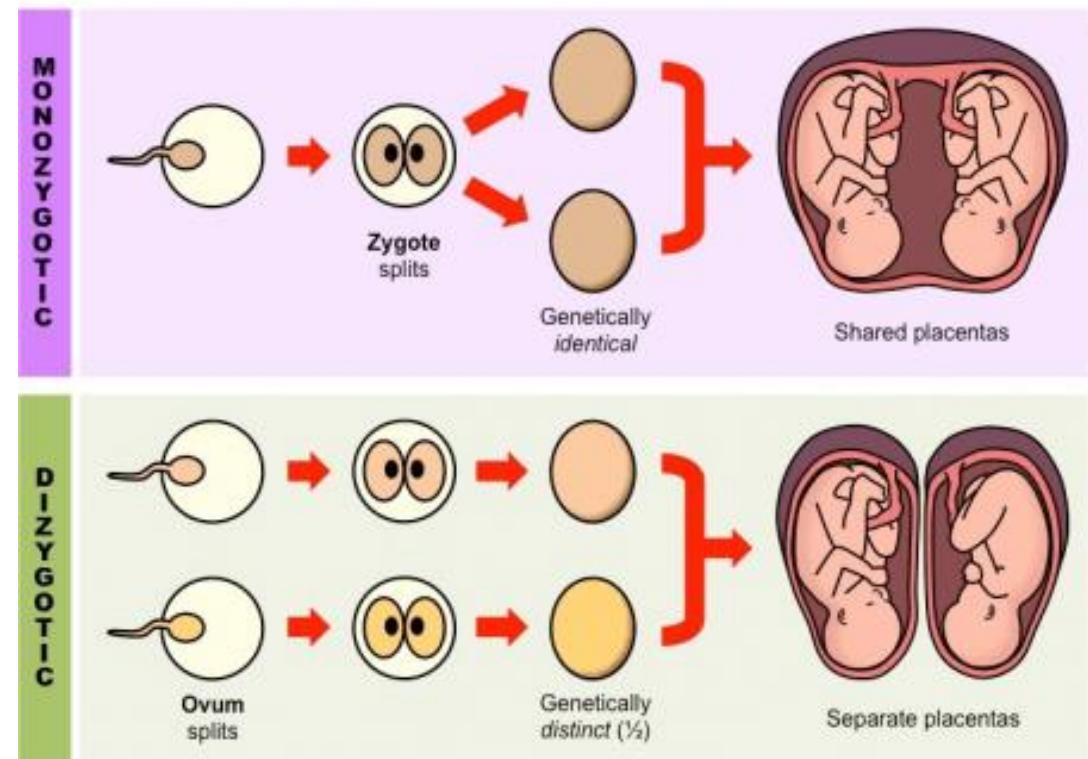
H^2 can take values between 0 and 1:

$H^2 = 0 \rightarrow$ all variation is due to environmental variation

$H^2 = 1 \rightarrow$ all variation is due to genetic variation

USING TWINS TO ESTIMATE H^2

Monozygotic twins (MZ): Difference in V_P is V_E



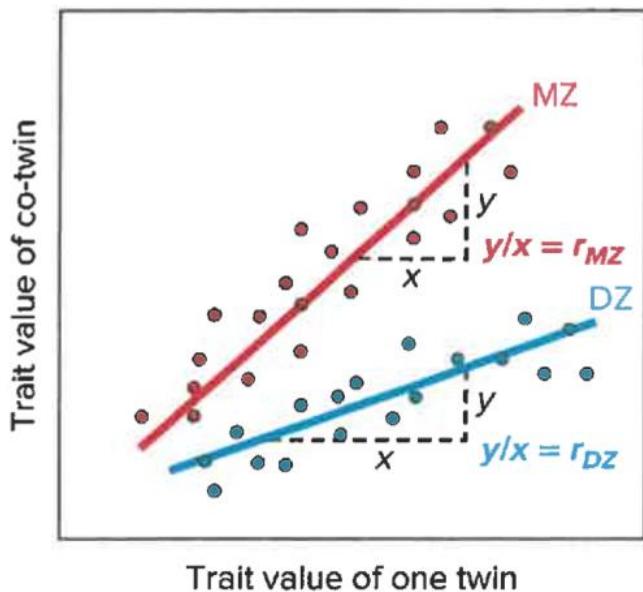
Dizygotic twins (DZ): Difference in V_P is $V_G + V_E$

Compare MZ [V_e] and DZ [$V_G + V_E$] twins – the difference is V_G

USING TWINS TO ESTIMATE H^2

QUANTITATIVE COMPLEX TRAITS

(b) MZ twins and DZ twins



Because DZ share 50% of the genetic material



$$H^2 = 2(r_{MZ} - r_{DZ})$$

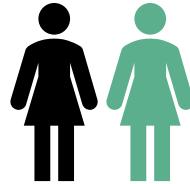
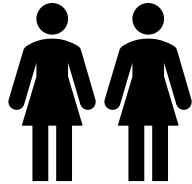
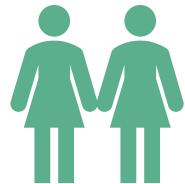
All phenotypic variation between MZ is environmental, whereas all phenotypic variation between DZ is both

Thus, the difference is genetic variation

USING TWINS TO ESTIMATE H^2

DICHOTOMOUS COMPLEX TRAITS

Concordance rate (C) = the frequency with which the other twin has the trait



Concordant pair: same phenotype

Discordant pair: only one in the pair has the trait

Important

$$H^2 = \frac{C_{MZ} - C_{DZ}}{1 - C_{DZ}}$$

There has to be a difference between C_{MZ} and C_{DZ} for a trait to be under genetic influence

The larger ratio $\frac{C_{MZ}}{C_{DZ}}$ the higher H^2

If $C_{MZ} < 1$ environmental exposures affect the trait

GROUP WORK

THE HERITABILITY OF HUMAN DISEASE

PART 1

- 1) Make 4 groups & prepare a 5-7 min presentation
 - Group 1 & 3 works with section 'Estimating heritability' p141-144
 - Group 2 & 4 works with section 'Biased heritability' p144-148

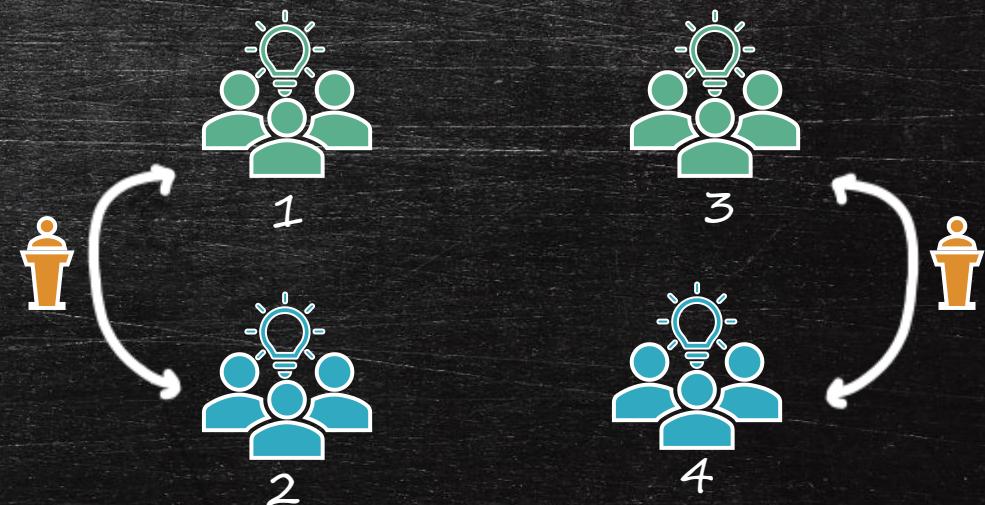
PART 2 – next time (17/3)

- Group 1 present to group 2 and *vise versa*
- Group 3 present to group 4 and *vise versa*

What did you find difficult?



Artiklen er på et meget højt og svært niveau, som kan være ret svær at forstå
- selvom man har læst den flere gange :(
:(

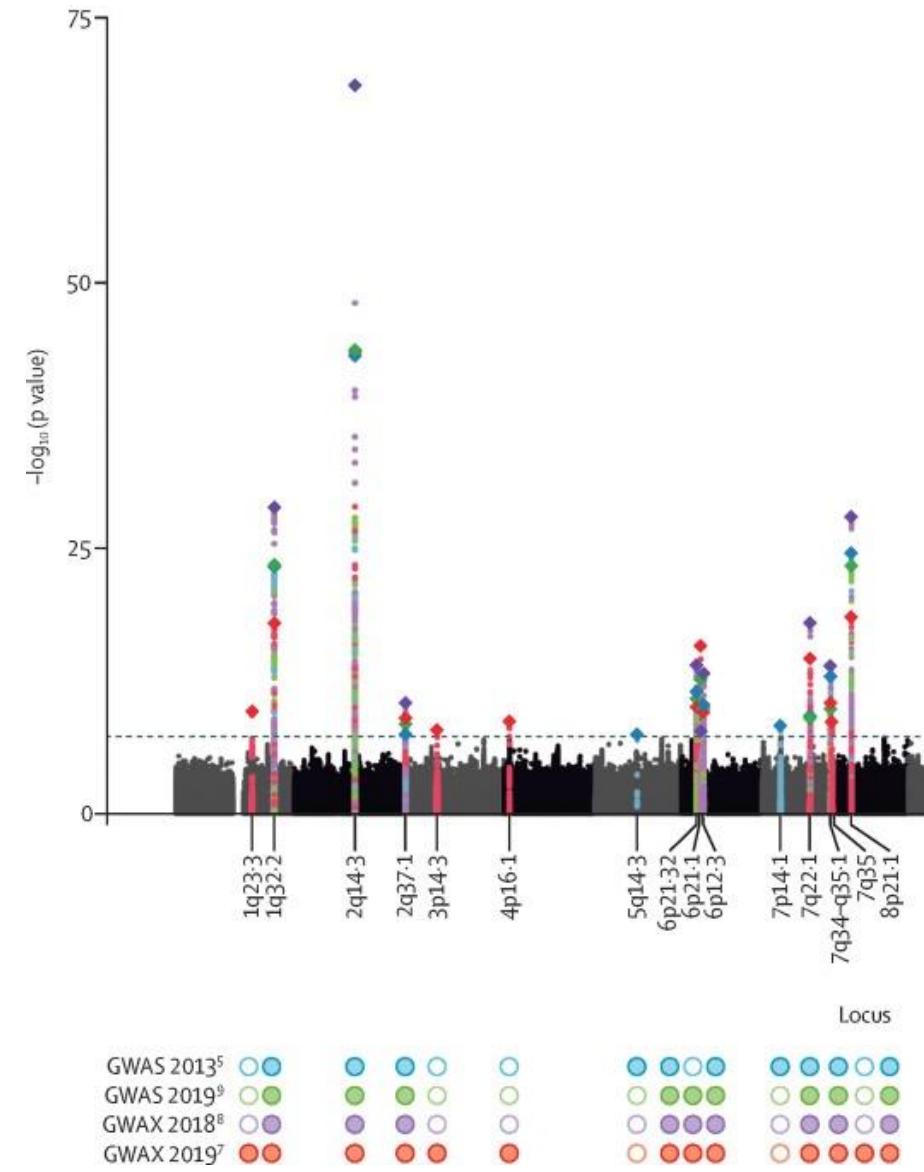


BREAK

AGENDA

- 08:15 – 08:30 Recap [*Complex traits and genetic parameters*]
- 08:30 – 09:00 Group presentations from last
- 09:00 – 09:15 Break
- 09:15 – 09:45** **Lecture 1** [*Genetic associations + GWAS part 1*]
- 09:45 – 10:15 Exercise 1 + 2
- 10:15 – 10:30 Break
- 10:30 – 11:00 Lecture 2 [*GWAS part 2*]
- 11:00 – 11:55 Group work
- 11:55 – 12:00 Evaluation at Moodle

GENETIC ASSOCIATION



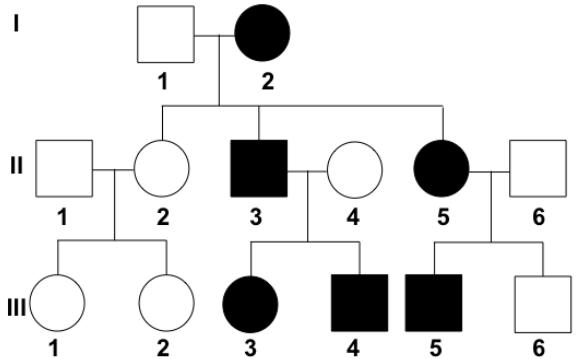
INHERITANCE PATTERN OF MULTIFACTORIAL TRAITS

The exact inheritance pattern depends on

- ❖ The number of risk genes/alleles involved
- ❖ The effect size distribution of the genetic risk variants
- ❖ The interaction among genetic variants
- ❖ The interaction with environmental exposures



Monogenic disorders



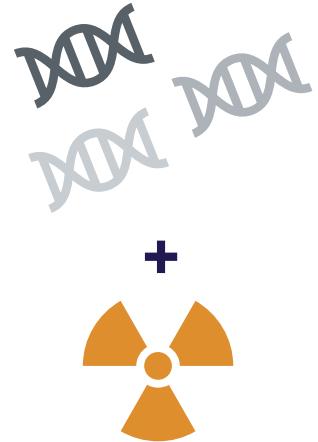
Mutation



Each genetic variant is **both** necessary and sufficient

Linkage analysis
(pedigree)

Common complex disorders



Each genetic variant is **neither** necessary nor sufficient

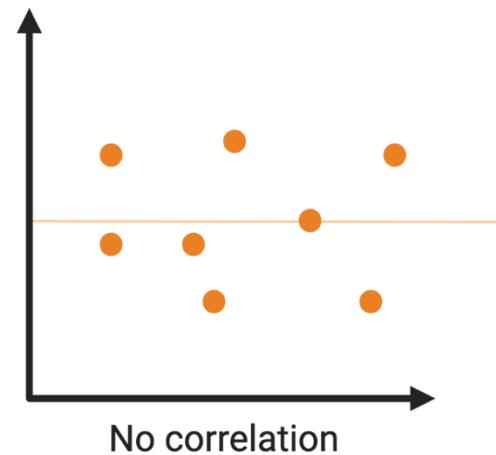
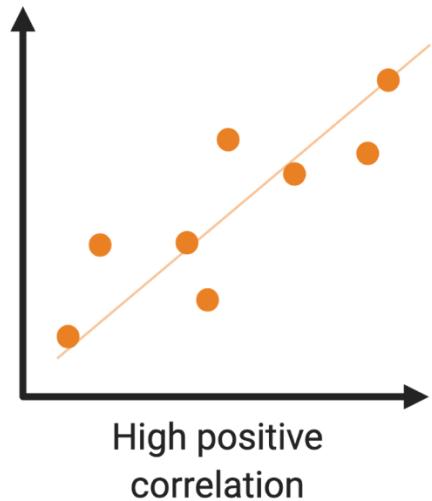
Association study
(unrelated population)



What does association mean

ASSOCIATION

An association defines a relationship between two entity objects based on common attributes.



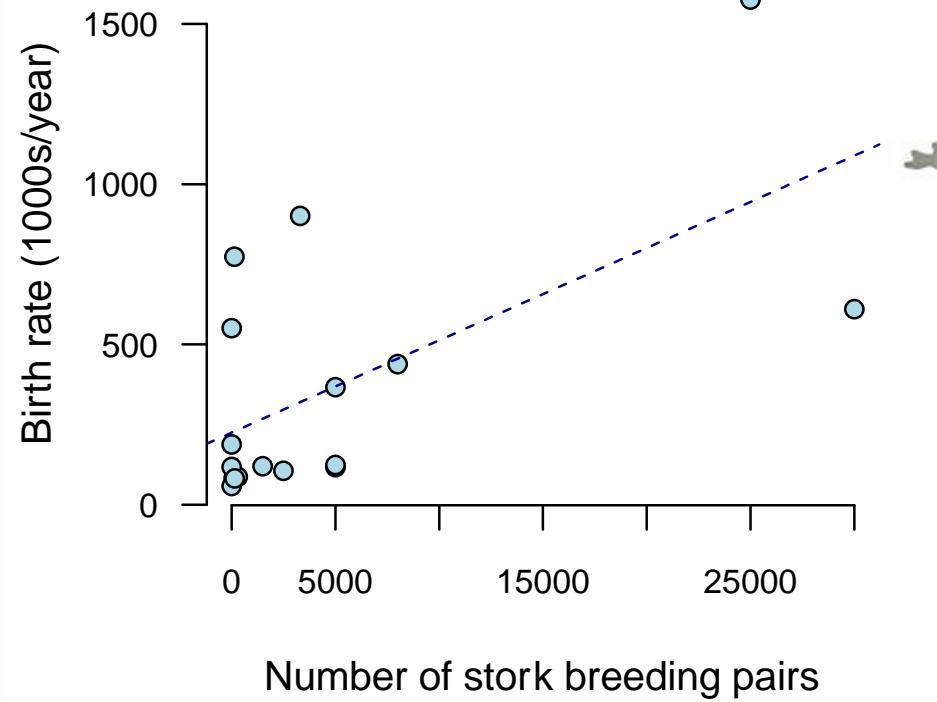
Is there an association between **exposure** and **outcome**?

ASSOCIATION NOT CAUSATION



Country	Area (km ²)	Storks (pairs)	Humans (10 ⁶)	Birth rate (10 ³ /yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries



Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.250e+02 9.356e+01 2.405 0.0295 *

Storks 2.879e-02 9.402e-03 3.063 0.0079 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



ASSOCIATION

Rare variants may have a **strong effect** on outcome (large OR)

Common variants have **small effect** on outcome (OR=1.1-1.8)

Is there an association between **genetic variant** and **disease status**?

OR>1

- allele is risk allele
- allele is seen more often among cases

OR<1

- allele is protective
- allele is seen more often among controls

		Outcome		Odds
		Case	Control	
Exposure	Carry variant (A -)	a	b	a/b
	Don't carry variant (aa)	c	d	c/d

Odds ratio (OR) is the ratio of the **odds of disease among the exposed to the odds of disease among the unexposed**

$$OR = \frac{a/b}{c/d} = \frac{a \times d}{c \times b}$$

ASSOCIATIONS

ALLELIC VS GENOTYPIC

We can count alleles OR genotypes. Example:

Association of rs6983267 on 8q24 with colorectal cancer [C/T, allele C is the risk allele]

	CC	CT	TT
Cases	250	375	150
Controls	460	940	500

GENOTYPIC ASSOCIATION

$$OR_{TT} = \frac{odds(disease|TT)}{odds(disease|TT)} = 1$$

$$OR_{CT} = \frac{odds(disease|CT)}{odds(disease|TT)} = \frac{375 \times 500}{940 \times 150} = 1.33$$

$$OR_{CC} = \frac{odds(disease|CC)}{odds(disease|TT)} = \frac{250 \times 500}{460 \times 150} = 1.81$$

Note, these ORs are relative to TT
(the lowest-risk genotype)

ASSOCIATIONS

ALLELIC VS GENOTYPIC

We can count alleles OR genotypes. Example:

Association of rs6983267 on 8q24 with colorectal cancer [C/T, allele C is the risk allele]

	C	T
Cases	875	675
Controls	1860	1940

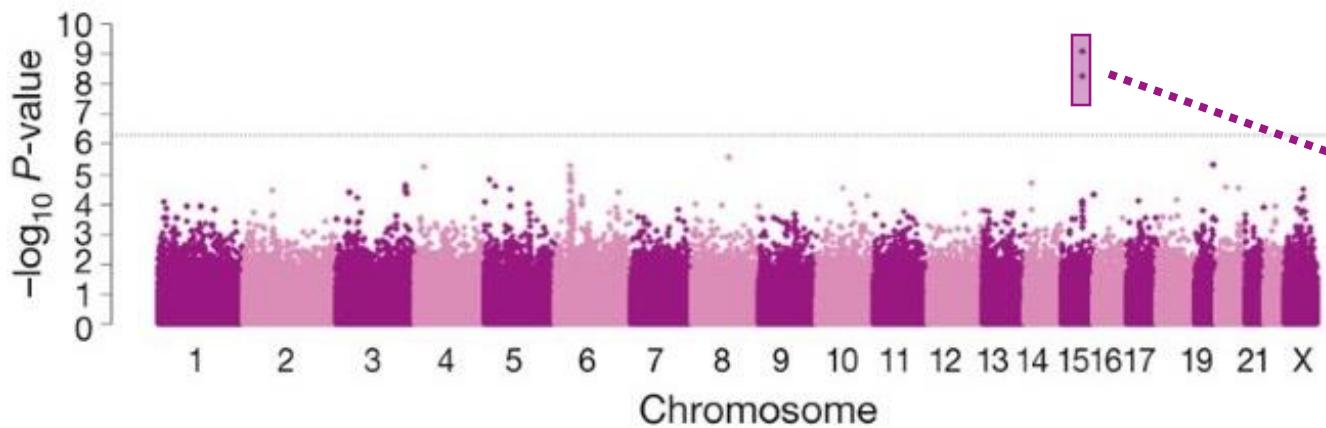
ALLELIC ASSOCIATION

Cases C alleles = $2 * 250 \text{ CC} + 375 \text{ CT} = 875$
 T alleles = $2 * 150 \text{ TT} + 375 \text{ CT} = 675$

Controls C alleles = $2 * 460 \text{ CC} + 940 \text{ CT} = 1860$
 T alleles = $2 * 500 \text{ TT} + 940 \text{ CT} = 1940$

$$OR_C = \frac{\text{odds(disease|C)}}{\text{odds(disease|T)}} = \frac{875 \times 1940}{1860 \times 675} = 1.35$$

GENETIC RISK FOR LUNG CANCER?



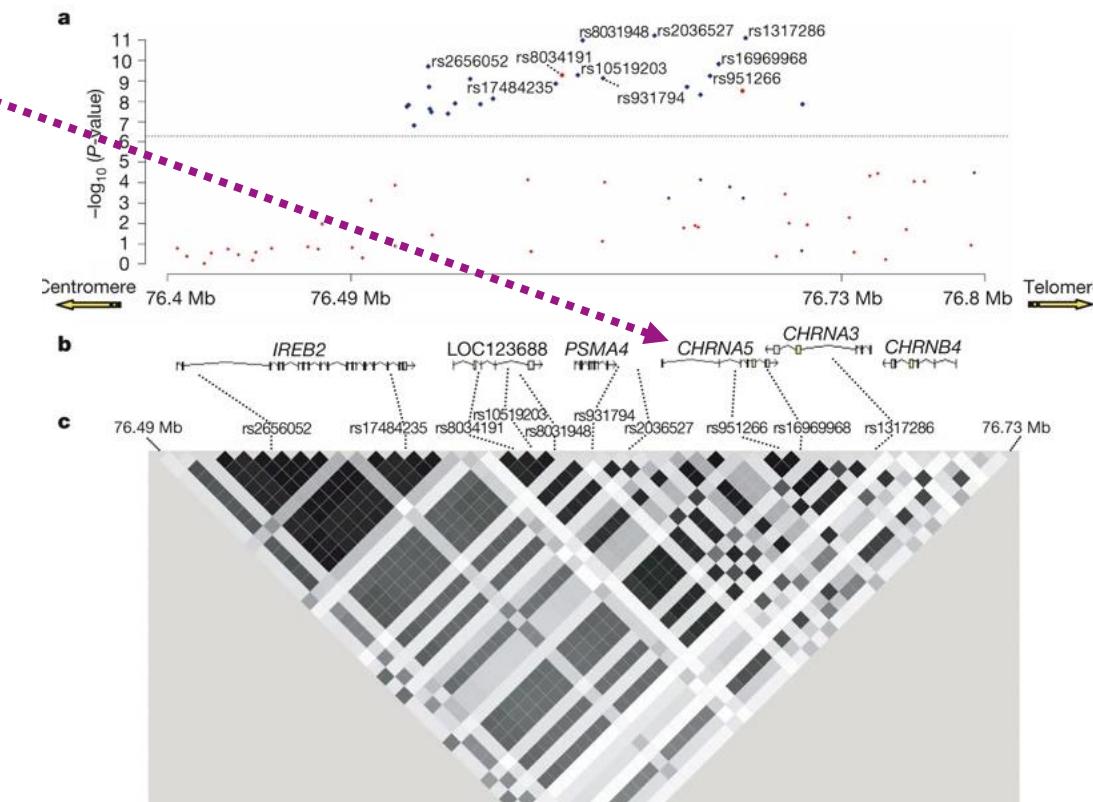
A genetic study of lung cancer (LC, 1989 cases and 2625 controls) found a nicotine receptor (*CHRNA3/5*) to be associated with the risk of developing lung cancer.

Does that then mean that *CHRNA3/5* is a risk loci for LC?

Risk loci for nicotine addition → addicted to smoking → increase LC risk

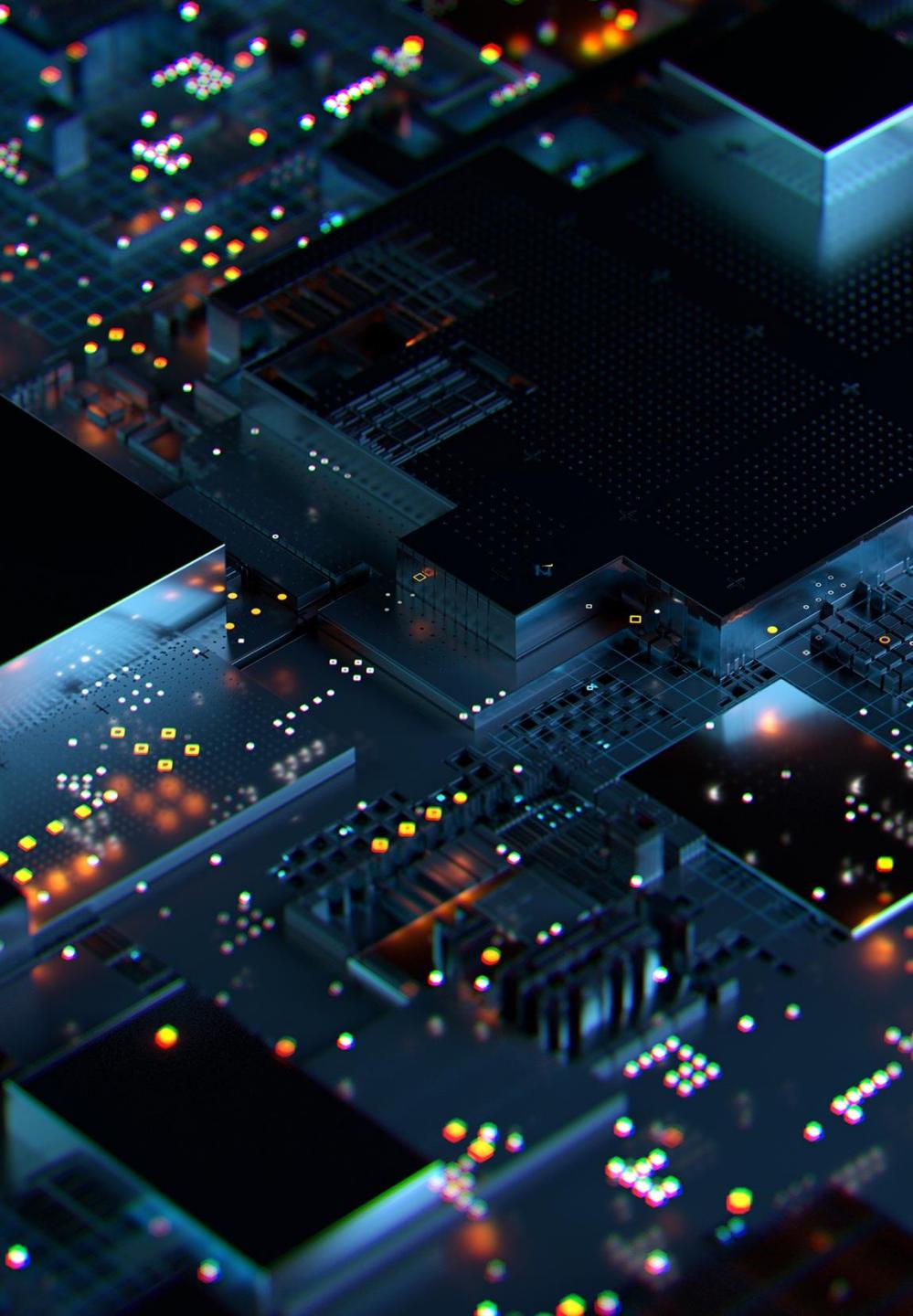
LETTERS

A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25



GENOME-WIDE ASSOCIATION ANALYSIS (GWAS) – PART 1

- What is a GWAS
- LD



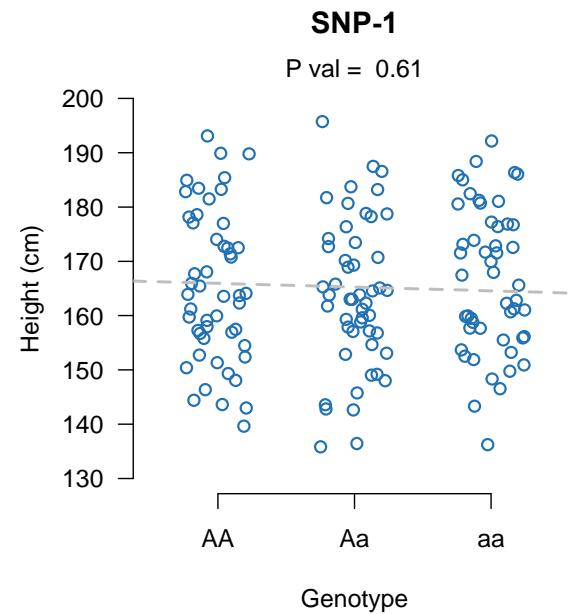
GENOME-WIDE ASSOCIATION STUDY (GWAS)

A systematic analysis of all common genetic variants without a priori hypothesis.

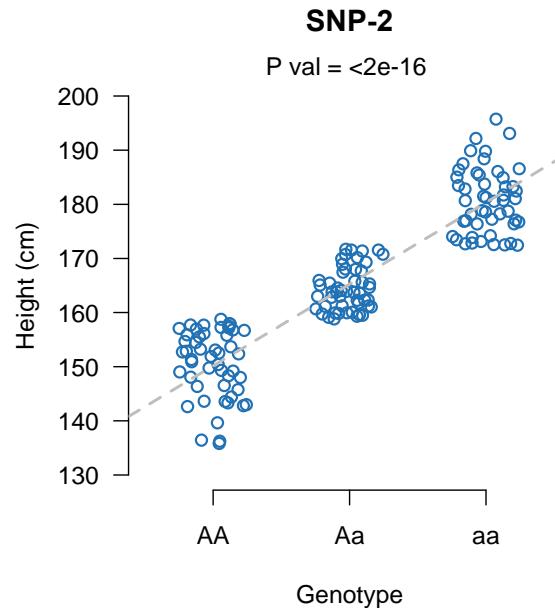
The aim is to identify risk variants for complex traits.

Testing one variant at a time

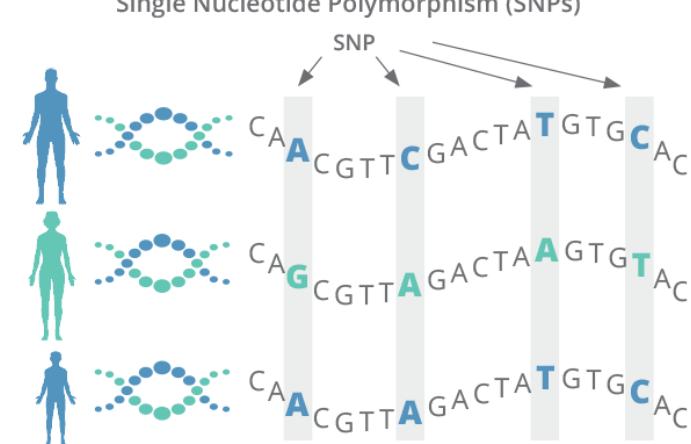
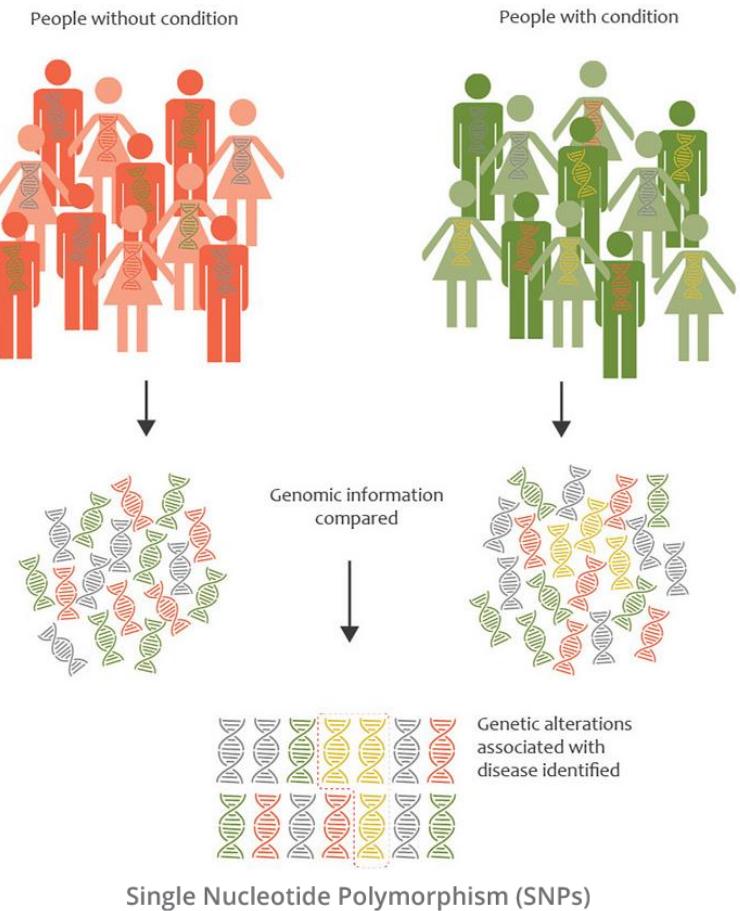
ASSOCIATION TESTING



No association

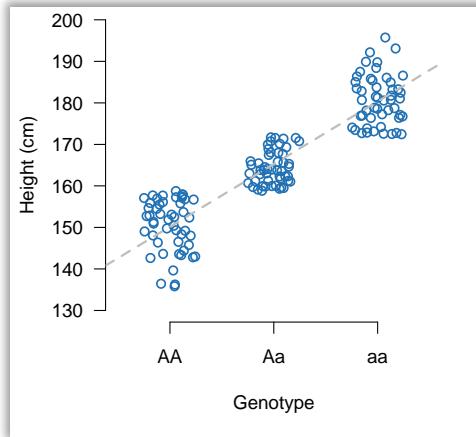


Association



THE BIOMETRIC MODEL

WHY THIS MODEL?



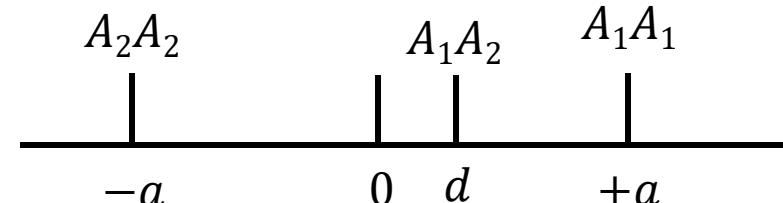
Consider a single locus with two alleles; A_1 and A_2 , with the frequencies p and q .

Under HWE, the genotype frequencies are p^2 , $2pq$, and q^2 .

Under the biometric model, the genotypic value of A_1A_1 is a , and $-a$ for A_2A_2 .

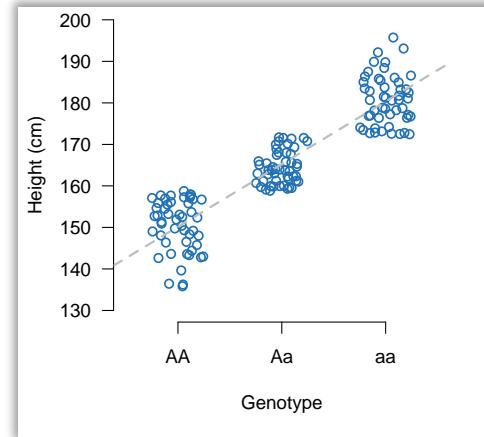
The genotype value for A_1A_2 is d .

In this model, every single genetic variant with non-zero effect on a phenotype contributes to the population mean ($P = G + E$).



THE BIOMETRIC MODEL

WHY THIS MODEL?

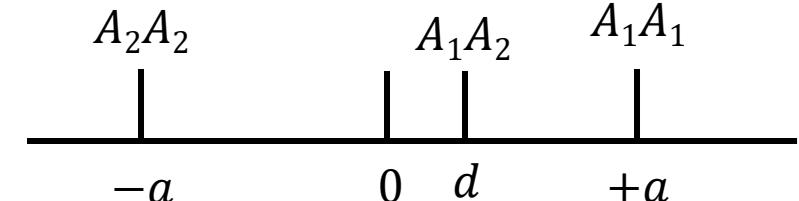
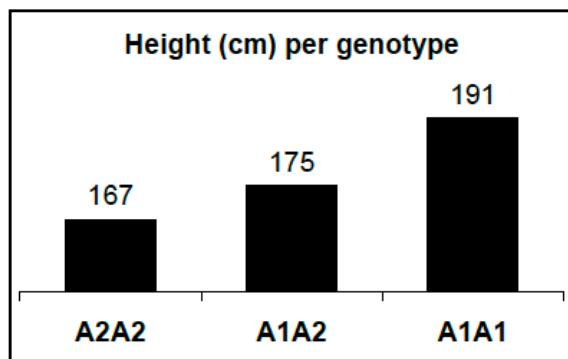


The mean effect (on the phenotype) is a function of the genotype effects ($-a$, a , and d) weighted by allele frequencies:

$$(ap^2) + (d2pq) + (-aq^2) = a(p - q) + d2pq$$

Polygenic traits are influenced by many genetic variants; thus, assuming additive and independent effects, the population mean is:

$$\mu = \sum_{i=1}^m a_i(p_i - q_i) + \sum_{i=1}^m d_i 2p_i q_i$$



HOW MANY SITES IN THE GENOME?

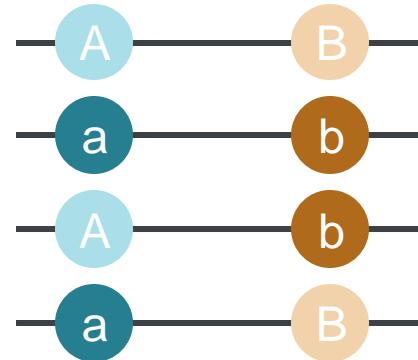
Should we test ALL 3,000,000,000 nucleotides within the genome?

No necessary, because of an old fried – Linkage Disequilibrium (LD)

WHAT IS LD?



Which gametes can
be produced?



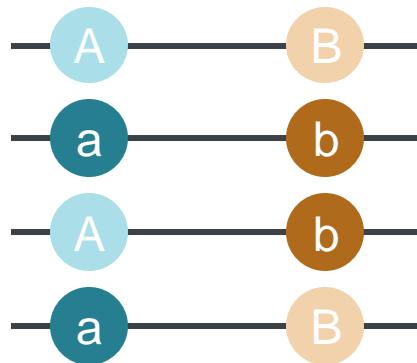
What are the
frequencies of the
alleles?

$$P(A), P(a)
P(B), P(b)$$

What are the
frequencies of the
haplotypes?

$$P(AB), P(Ab)
P(aB), P(ab)$$

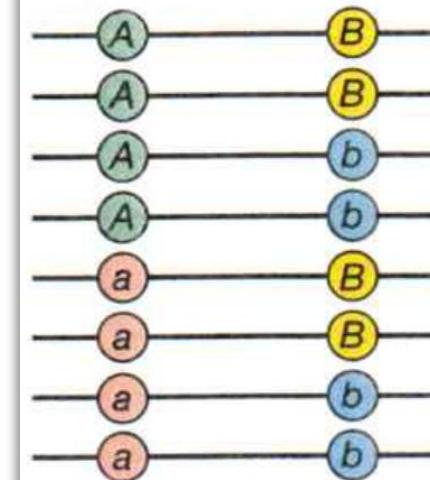
WHAT IS LD?



If there is random relationship among alleles at the two loci then the frequency of the haplotypes will be the product of the frequencies of the two alleles:

$$P(AB) = P(A) \times P(B)$$

(a) Linkage equilibrium



$$p_A = 0.5$$

$$p_a = 0.5$$

$$p_B = 0.5$$

$$p_b = 0.5$$

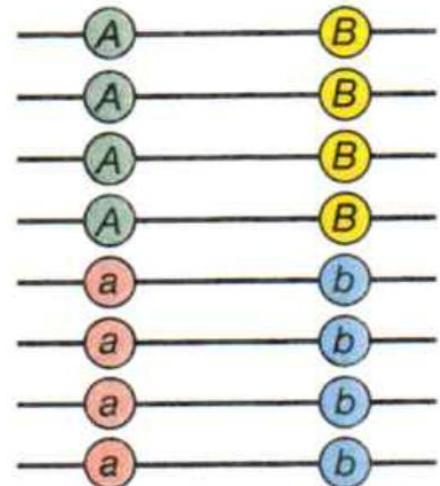
$$P_{AB} = 0.25$$

$$P_{Ab} = 0.25$$

$$P_{aB} = 0.25$$

$$P_{ab} = 0.25$$

(b) Linkage disequilibrium



$$p_A = 0.5$$

$$p_a = 0.5$$

$$p_B = 0.5$$

$$p_b = 0.5$$

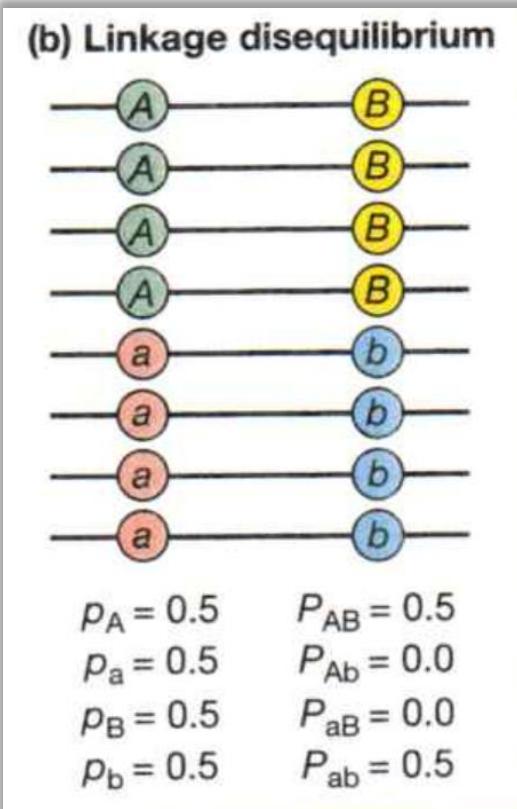
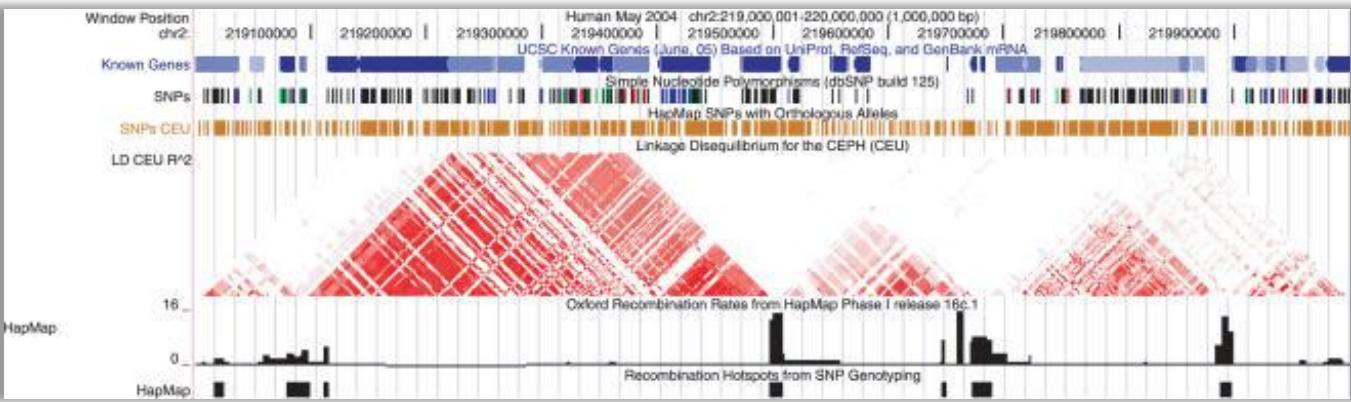
$$P_{AB} = 0.5$$

$$P_{Ab} = 0.0$$

$$P_{aB} = 0.0$$

$$P_{ab} = 0.5$$

WHAT IS LD?

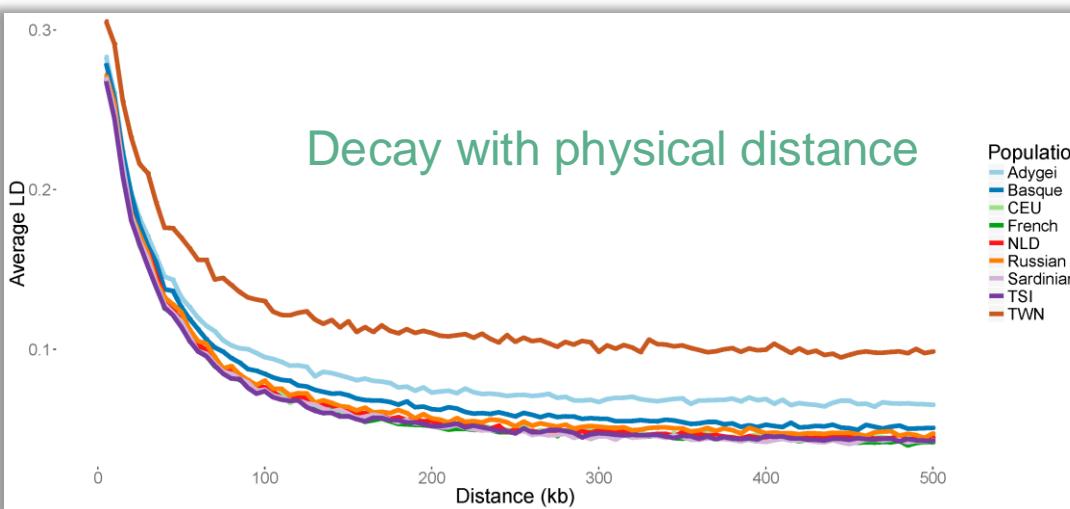


When the association between alleles at two loci is **non-random** they are said to be in **linkage disequilibrium**

The degree of LD can be measured in several ways – the simplest one is:

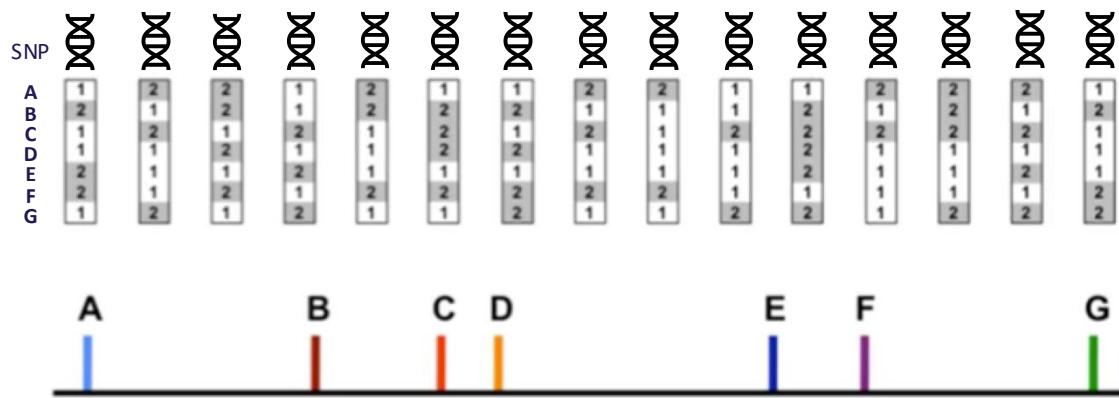
$$D = P_{AB} - P_A P_B$$

If $D=0$, no LD, if $D>0$ LD

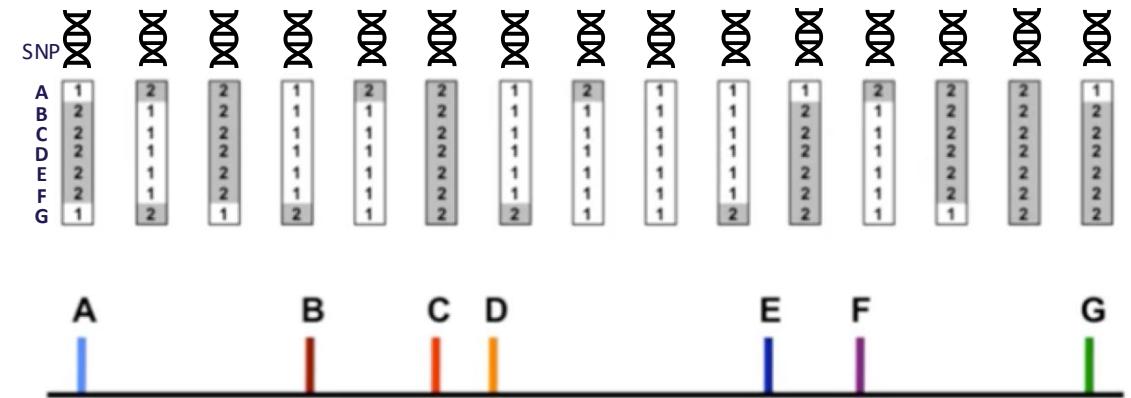


LD AND GENE MAPPING

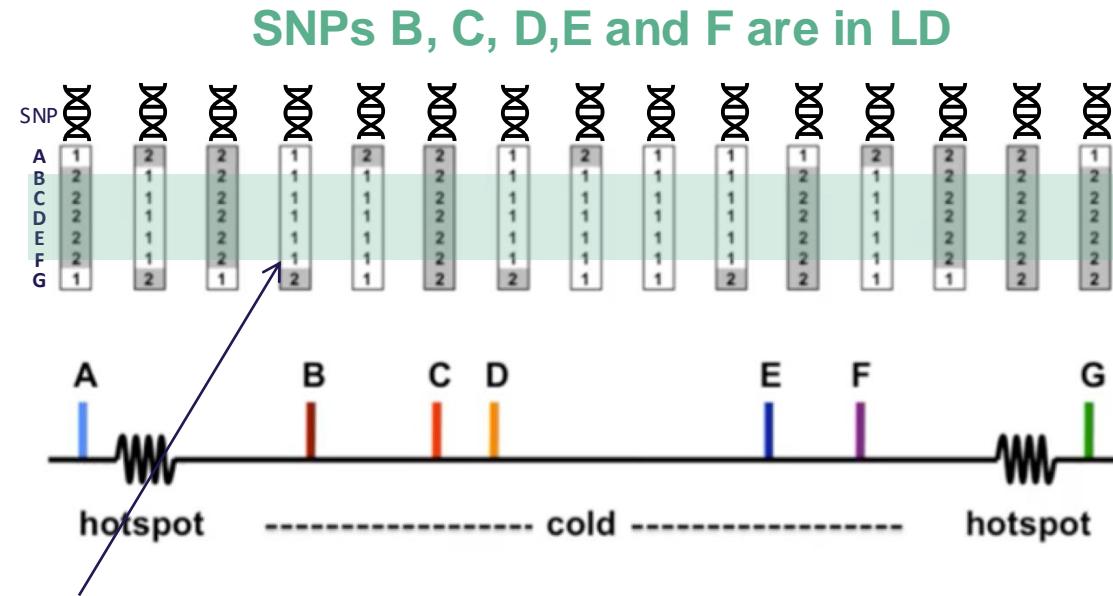
Linkage equilibrium – *random association*



Linkage disequilibrium – *non-random association*



LD AND GENE MAPPING



If you have allele 1 here, I know
what you are at the remaining sites
in this haploblok

HOW MANY SITES IN THE GENOME?

Should we test ALL 3,000,000,000 nucleotides within the genome?

No necessary, because of an old fried – Linkage Disequilibrium (LD)

Typically, we test 5,000,000 – 10,000,000 SNPs.

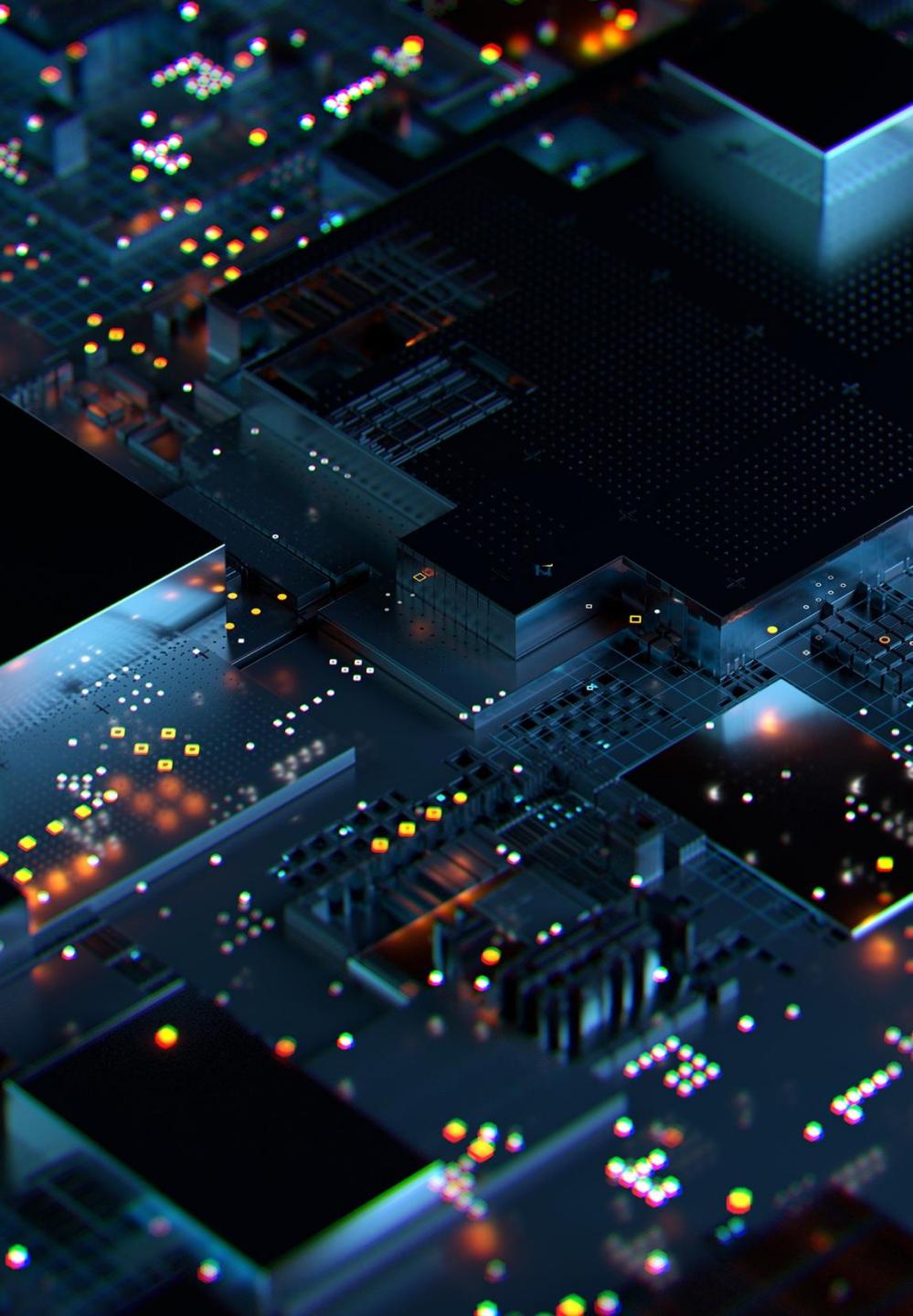
AGENDA

- 
- 08:15 – 08:30 Recap [*Complex traits and genetic parameters*]
- 08:30 – 09:00 Group presentations from last
- 09:00 – 09:15 Break
- 09:15 – 09:45 Lecture 1 [*Genetic associations + GWAS part 1*]
- 09:45 – 10:15 Exercise 1 + 2
- 10:15 – 10:30 Break
- 10:30 – 11:00 Lecture 2 [*GWAS part 2*]
- 11:00 – 11:55 Group work
- 11:55 – 12:00 Evaluation at Moodle

BREAK

GENOME-WIDE ASSOCIATION ANALYSIS (GWAS) – PART 2

- What is a GWAS
- LD
- GWAS by steps



GWAS BY STEPS

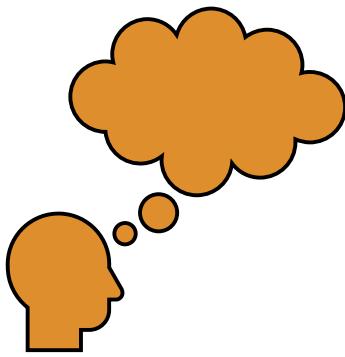


GWAS BY STEPS

1. Select trait/disease



SELECT PHENOTYPE



Do we know anything about the phenotype already?

Is it heritable?

Do we know whether it is monogenic or polygenic?

Is it a common or rare disorder?

GWAS BY STEPS

1. Select trait/disease
2. Extract genetic variants



GENOTYPING



Because of LD you do not have to analyse all 3,000,000,000 variants in the genome.

Typically, we genotype $\frac{1}{2}$ - 1 million variants

Because of LD we can impute (“guess” what variants are next to the genotyped variant) up to 10 million common genetic variants.

IMPUTATION USING HAPLOTYPES

The true haplotypes

A	T	C
G	C	A

This individual has inherited a chromosome with alleles A-T-C from one parent, and G-C-A from the other parent

We observe only the genotypes

A/G T/C C/A

Genotype data does not carry information about the haplotypes.

We do not know whether A at SNP1 is coming from the same parent as T or C at SNP2

Different haplotypes

A	C	A
A	C	C
A	T	A
A	C	A
G	C	A
G	C	C
G	T	A
G	T	C

Phasing = estimate the most likely haplotypes

IMPUTATION

Genotypes

Study sample

.....	A.....	A.....	A.....
.....	G.....	C.....	A.....

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTCTTCTGTGC
CGAAGCTCTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTCTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTCTTCTGTGC

From a sequencing study

Study sample

.....	A.....	A.....	A.....
.....	G.....	C.....	A.....

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTCTTCTGTGC
CGAAGCTCTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTCTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTCTTCTGTGC

Study sample

cgagAtctcccgAcctcAtgg
cgaAGctcttttCtttcAtgg

Reference haplotypes

CGGCCCGGGCAATTTTTTTT
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTCTTCTGTGC
CGAAGCTCTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTCTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTCTTCTGTGC

GWAS BY STEPS

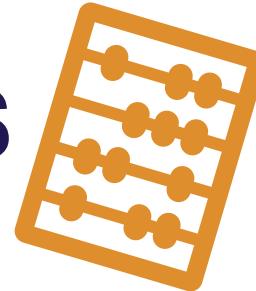


3. GWAS

2. Extract genetic variants

1. Select trait/disease

STATISTICAL ANALYSIS



For each genetic variant [genotyping chip or after imputation] measure the degree of association between SNP and disease/trait

Fischers exact test

		Outcome	
		Case	Control
Exposure	Carry variant (A-)	a	b
	Don't carry variant (aa)	c	d

$$OR = \frac{a \times d}{c \times b}$$

Linear regression

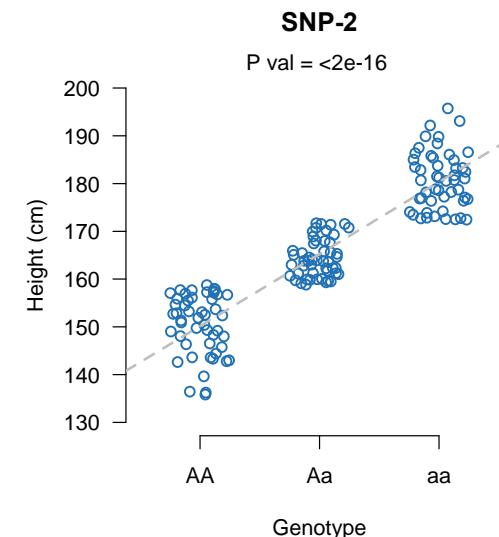
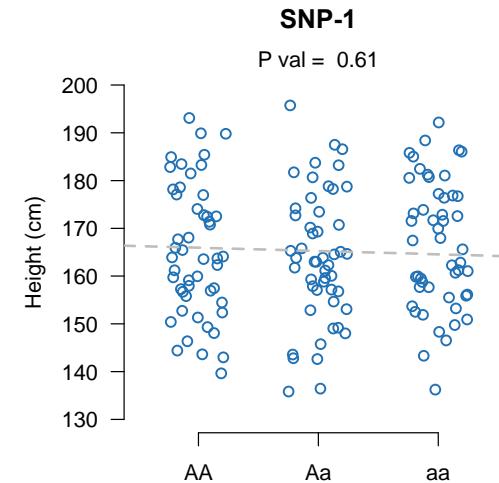
Allow us to account for confounding effects like, sex, age and ancestry.

If trait follows Gaussian distribution:

- Linear regression
- Linear Mixed Model

If trait is dichotomous:

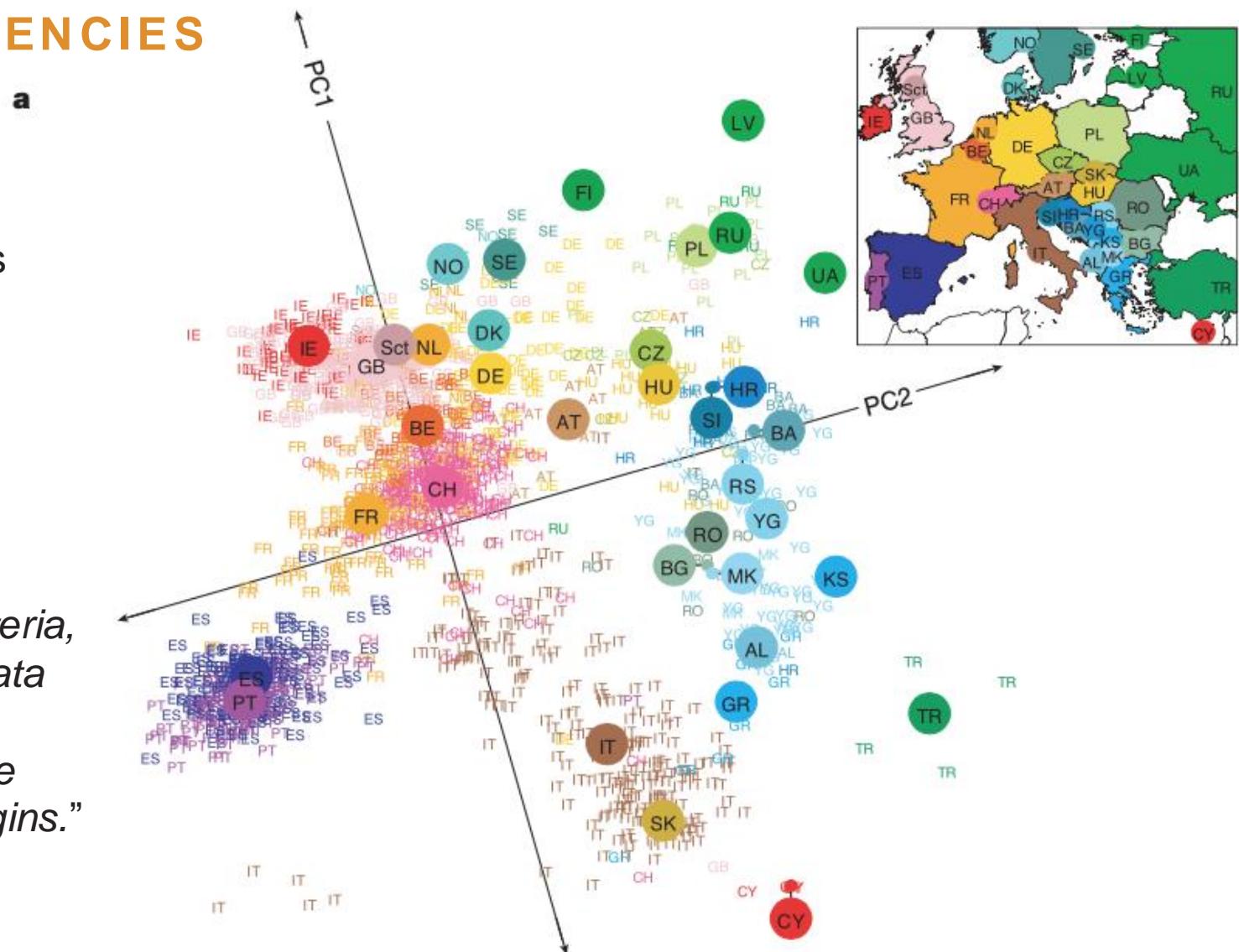
- Linear model with non-linear transformation
[logistic regression with $\text{logit}(x)$]



ANCESTRY AFFECTS ALLELE FREQUENCIES

They surveyed genetic variation in a sample of **3,192 European** individuals which were genotyped at 500,568 loci using the Affymetrix 500K single nucleotide polymorphism (SNP) chip.

“Our main result holds even when we relax nearly all of these stringency criteria, we focus our analyses on genotype data from 197,146 loci in 1,387 individuals (Supplementary Table 2), for whom we have high confidence of individual origins.”

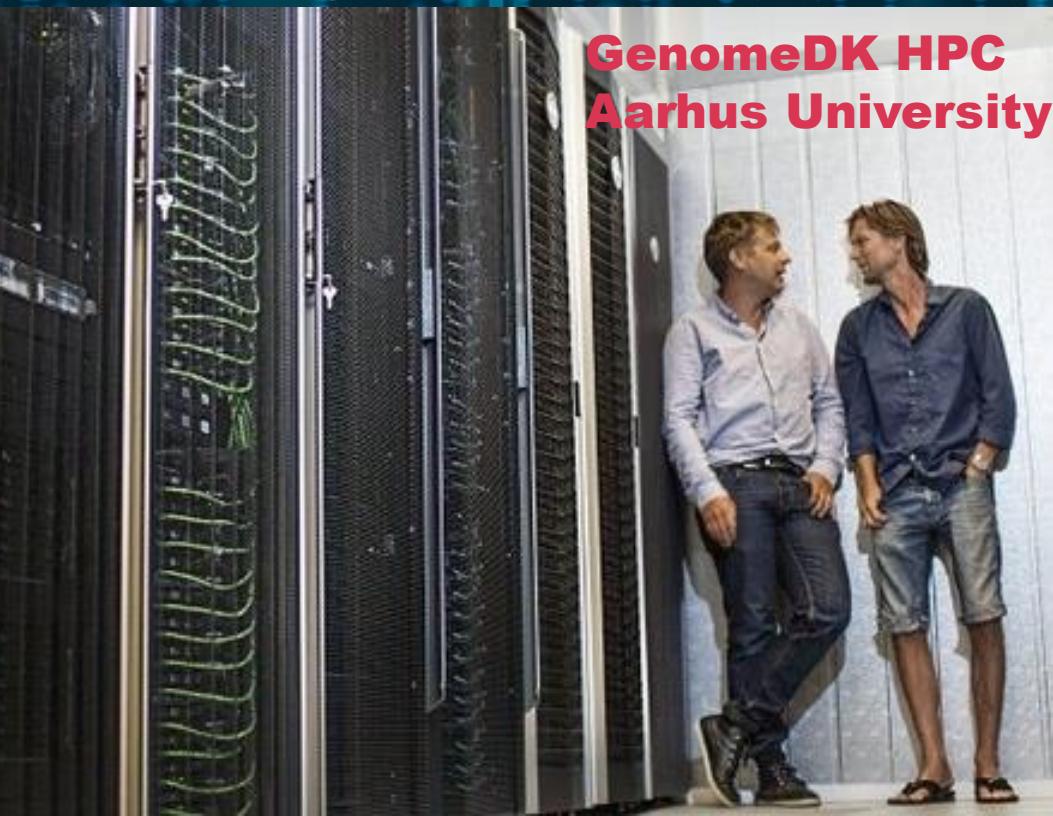




... and specialized software designed to handle large genetic data sets



GenomeDK HPC
Aarhus University



GWAS BY STEPS

1. Select trait/disease
2. Extract genetic variants
3. GWAS
4. Summaries 10M linear regressions



chromosome	base_pair_location	ID	REF	ALT	A1	TEST	OBS_CT	BETA	SE	T_STAT	p_value
1	752721	1:752721:G	A	G	A	ADD	1173	-0.0123326	0.0574282	-0.371466	0.710358
1	794332	1:794332:G	G	A	A	ADD	1173	0.0167258	0.0751683	0.222512	0.823954
1	840753	1:840753:T	T	C	C	ADD	1173	0.034768	0.0424801	0.818454	0.413265
1	845635	1:845635:C	C	T	T	ADD	1173	0.010575	0.0584211	0.209734	0.833912
1	845938	1:845938:G	G	A	A	ADD	1173	0.0138978	0.0501171	0.277307	0.781593
1	846078	1:846078:C	C	T	T	ADD	1173	0.0154452	0.052242	0.295648	0.767551
1	846398	1:846398:G	G	A	A	ADD	1173	0.0097726	0.0507151	0.192696	0.84723
1	846888	1:846888:C	C	T	T	ADD	1173	0.0130275	0.0521993	0.249572	0.802962
1	846864	1:846864:G	G	C	C	ADD	1173	0.0154452	0.052242	0.295648	0.767551
1	847228	1:847228:C	C	T	T	ADD	1173	-0.0016347	0.0394749	0.07474657	
1	847491	1:847491:G	G	A	A	ADD	1173	0.00426234	0.0584764	0.0844423	0.932719
1	848090	1:848090:G	G	A	A	ADD	1173	0.00819704	0.050492	0.102928	0.918038
1	848445	1:848445:G	G	A	A	ADD	1173	-0.0174173	0.0505568	-0.34451	0.730525
1	848456	1:848456:G	G	A	G	ADD	1173	-0.0174173	0.0505568	-0.34451	0.730525
1	848738	1:848738:C	C	T	T	ADD	1173	0.0051974	0.050492	0.102928	0.918038
1	850062	1:850062:A	A	T	T	ADD	1173	0.00426234	0.0584764	0.0844423	0.932719
1	850123	1:850123:C	C	T	T	ADD	1173	0.00819704	0.050492	0.102928	0.918038
1	851190	1:851190:G	G	A	A	ADD	1173	0.00880137	0.0504533	0.15859	0.87402
1	851204	1:851204:G	G	C	C	ADD	1173	0.00799277	0.0503183	0.158844	0.873819
1	852664	1:852664:C	C	T	T	ADD	1173	-0.002825601	0.0508564	-0.0561584	0.955225
1	852758	1:852758:G	G	C	C	ADD	1173	-0.002825601	0.0508564	-0.0561584	0.955225
1	853239	1:853239:A	A	G	G	ADD	1173	-0.002825601	0.0508564	-0.0561584	0.955225
1	854250	1:854250:G	G	A	G	ADD	1173	-0.00834348	0.0508285	-0.0675762	0.941635
1	858040	1:858040:C	C	A	A	ADD	1173	0.00473544	0.05084763	0.093818	0.925272
1	858051	1:858051:C	C	T	T	ADD	1173	0.00473544	0.0504763	0.093815	0.925272
1	864002	1:864002:G	G	C	C	ADD	1173	0.0184693	0.0500163	0.369265	0.711997
1	865219	1:865219:G	G	A	A	ADD	1173	0.00750316	0.05084318	0.148787	0.881754
1	866893	1:866893:T	T	C	T	ADD	1173	0.006032305	0.0422389	0.149697	0.881029
1	866938	1:866938:G	G	A	A	ADD	1173	0.00705206	0.0485498	0.1452584	0.884535
1	867635	1:867635:C	C	T	T	ADD	1173	0.00750316	0.0504318	0.148778	0.881754
1	872352	1:872352:G	G	C	C	ADD	1173	0.009660345	0.04686866	0.204822	0.837747
1	877147	1:877147:G	G	A	A	ADD	1173	-0.0115154	0.0473281	-0.243309	0.887089
1	881627	1:881627:G	G	A	G	ADD	1173	0.0520509	0.0422311	1.1899	0.234236
1	882033	1:882033:G	G	A	A	ADD	1173	-0.0068909	0.0470492	-0.146444	0.883596
1	886659	1:886659:T	T	C	T	ADD	1173	0.206878	0.0865538	0.39702	0.0169728
1	889238	1:889238:G	G	A	A	ADD	1173	0.21628	0.0890664	0.424994	0.052507
1	890164	1:890164:G	G	A	A	ADD	1173	0.00331508	0.0468322	0.0707863	0.94358
1	894573	1:894573:G	G	A	G	ADD	1173	0.141979	0.0688431	0.26235	0.839345
1	897564	1:897564:T	T	C	T	ADD	1173	0.179375	0.0798978	0.24505	0.024958
1	897738	1:897738:C	C	T	T	ADD	1173	0.179541	0.0802621	0.23693	0.0254787
1	898467	1:898467:C	C	T	T	ADD	1173	0.21628	0.0890664	0.424994	0.052507
1	900730	1:900730:G	G	A	G	ADD	1173	0.118829	0.057338	0.154345	0.873819
1	903321	1:903321:G	G	A	G	ADD	1173	0.00298207	0.0468322	0.0707863	0.94358
1	903426	1:903426:C	C	T	T	ADD	1173	-0.00879923	0.0476482	-0.167881	0.866767
1	908823	1:908823:G	G	A	A	ADD	1173	0.00223841	0.0556691	0.0402093	0.967933
1	909309	1:909309:T	T	C	C	ADD	1173	0.0218233	0.0549488	0.397322	0.691203
1	910473	1:910473:G	G	A	A	ADD	1173	-0.0200586	0.0576243	-0.348092	0.727834
1	911916	1:911916:C	C	T	T	ADD	1173	-0.0308612	0.0547515	-0.563659	0.573094
1	912049	1:912049:T	T	C	T	ADD	1173	0.0461071	0.0412247	-1.11843	0.263611
1	913610	1:913610:G	G	A	A	ADD	1173	-0.0380612	0.0547515	-0.563659	0.573094
1	913889	1:913889:G	G	A	G	ADD	1173	-0.0569182	0.0413668	-1.37594	0.169104
1	914333	1:914333:C	C	G	C	ADD	1173	0.0631388	0.0415457	0.262426	0.12772
1	914852	1:914852:C	C	G	C	ADD	1173	0.0654599	0.0414123	-1.36336	0.173031
1	914940	1:914940:T	T	C	T	ADD	1173	0.0612227	0.0411078	-1.48932	0.136672
1	916598	1:916598:G	G	A	A	ADD	1173	0.0395153	0.0508336	-0.777347	0.437111
1	916662	1:916662:A	A	C	C	ADD	1173	-0.0288419	0.0546679	-0.527584	0.597888
1	916834	1:916834:G	G	A	G	ADD	1173	-0.0672691	0.0411436	-1.63499	0.102321
1	917315	1:917315:G	G	A	A	ADD	1173	-0.0189492	0.0521825	-0.363134	0.71657
1	917492	1:917492:C	C	T	T	ADD	1173	-0.0288419	0.0546679	-0.527584	0.597888
1	917640	1:917640:G	G	A	A	ADD	1173	-0.0686789	0.0471771	-1.45577	0.145725
1	918238	1:918238:C	C	G	G	ADD	1173	-0.0686789	0.0471771	-1.45577	0.145725
1	918270	1:918270:C	C	T	T	ADD	1173	-0.0283636	0.0546403	-0.519097	0.608379
1	918384	1:918384:T	T	G	G	ADD	1173	-0.0861227	0.0410178	-1.48932	0.136672
1	918573	1:918573:A	A	G	A	ADD	1173	-0.0115154	0.0473281	-0.243309	0.887089
1	918617	1:918617:G	G	A	A	ADD	1173	-0.0283636	0.0546403	-0.519097	0.608379
1	919127	1:919127:T	T	C	C	ADD	1173	-0.0760187	0.0473807	-1.60442	0.108891
1	919419	1:919419:T	T	C	T	ADD	1173	-0.0131869	0.0458342	-0.287709	0.773621
1	919501	1:919501:G	T	G	G	ADD	1173	-0.021286	0.0410945	-0.519796	0.604573
1	919855	1:919855:G	G	A	A	ADD	1173	0.0272325	0.0574697	0.473859	0.636589
1	920503	1:920503:G	G	A	A	ADD	1173	-0.0424973	0.0559222	-0.791225	0.424848
1	920640	1:920640:C	C	T	T	ADD	1173	-0.0394462	0.0533635	-0.739197	0.459935
1	920977	1:920977:T	T	C	C	ADD	1173	-0.0289775	0.0516726	-0.672046	0.501687
1	921660	1:921660:C	C	T	T	ADD	1173	-0.0976773	0.0474452	-0.52784	0.597888
1	922483	1:922483:T	T	C	C	ADD	1173	-0.0341491	0.0546375	-0.624462	0.532446
1	924111	1:924111:T	T	A	A	ADD	1173	-0.0672691	0.0411436	-1.63499	0.102321
1	924161	1:924161:G	G	A	A	ADD	1173	-0.0186178	0.0546375	-0.624462	0.532446
1	924274	1:924274:G	G	A	A	ADD	1173	-0.0727421	0.0410178	-1.48932	0.136672
1	924452	1:924452:T	T	C	C	ADD	1173	-0.036718	0.0546403	-0.672046	0.501687
1	925057	1:925057:C	C	A	A	ADD	1173	-0.0976773	0.0474452	-0.52784	0.597888
1	925390	1:925390:C	C	A	A	ADD	1173	-0.023847	0.0516726	-0.672046	0.501687
1	925745	1:925745:T	T	C	C	ADD	1173	-0.0425746	0.0517009	-0.972044	0.33323
1	925785	1:925785:T	T	C	C	ADD	1173	-0.0365947	0.0451504	-0.851437	0.387184
1	926155	1:926155:G	G	A	A	ADD	1173	-0.031984	0.0565031	-0.672046	0.501687
1	926206	1:926206:C	C	A	A	ADD	1173	-0.0365947	0.0451504	-0.851437	0.387184
1	926212	1:926212:G	G	A	A	ADD	1173	-0.0305162	0.0484997	-0.672046	0.501687
1	926242	1:926242:G	G	A	A	ADD	1173	-0.0305162	0.0484997	-0.672046	0.501687
1	926250	1:926250:T	T	C	C	ADD	1173	-0.0383894	0.0454643	-0.79398	0.42765
1	926251	1:926251:G	G	A	A	ADD	1173	-0.0316708	0.0565252	-0.561864	0.574316
1	926256	1:926256:C	C	A	A	ADD	1173	-0.0365947	0.0451504	-0.851437	0.387184
1	926260	1:926260:T	T	C	C	ADD	1173	-0.0363704	0.0565252	-0.561864	

MANHATTAN PLOT

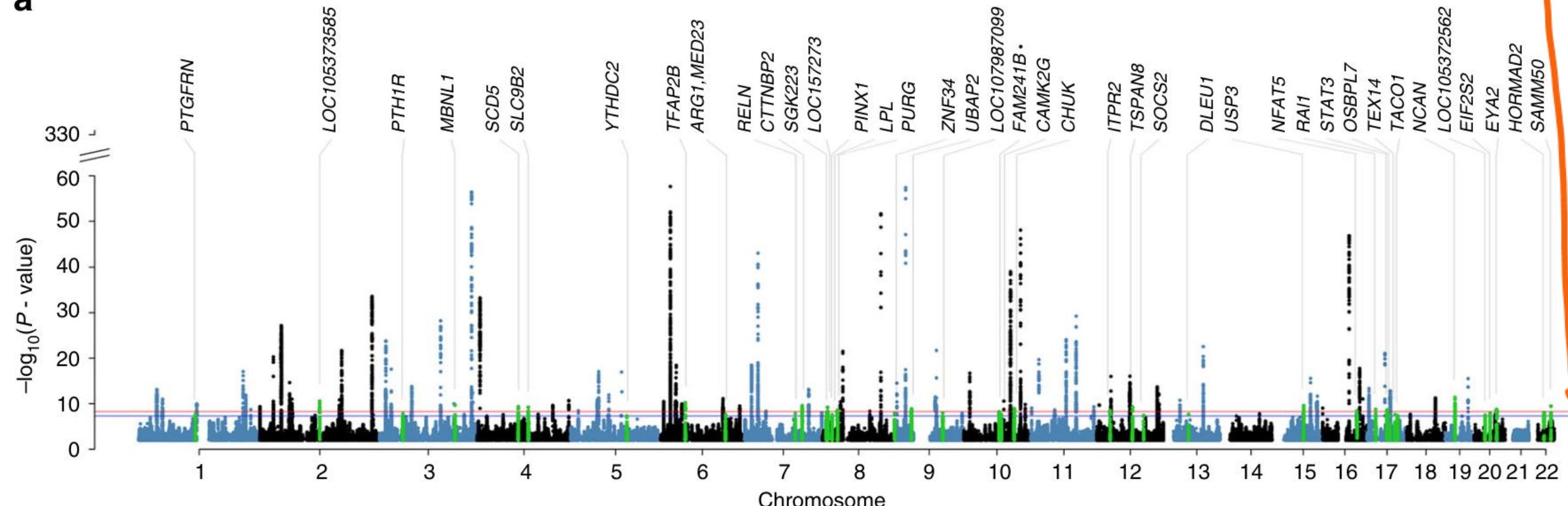
For each variant, plot the $-\log_{10}(P\text{-value})$ as function of chromosomal position.

$$P=0.05 \rightarrow -\log_{10}(0.05) = 1.3$$

$$P=0.001 \rightarrow -\log_{10}(0.001) = 3$$

$$P=0.000000005 \rightarrow -\log_{10}(0.000000005) = 8.3$$

a



HYPOTHESIS TESTING

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis (H_0)	Type I error α False positive	Correct outcome True positive
Accept null hypothesis (H_0)	Correct outcome True negative	Type II error β False negative

The probability (P) of making a type I error is denoted by α ; we reject the null hypothesis if the inferred P value is less than the significance level ($\alpha=0.05$). i.e., the probability of rejecting the null hypothesis when it should be accepted.

Why multiple testing correction? If we test 500,000 SNPs, then by chance we expect 25.000 SNPs to be significant (if $\alpha=0.05$) → i.e., **25.000 false-positive associations**.

One solution is to correct for number of tests performed; Bonferroni correction;

$$\text{Corrected } P\text{-value} = P \times n_{tests} \leq 0.05 \text{ OR } \frac{\alpha}{n_{tests}} = \text{new significance threshold}$$

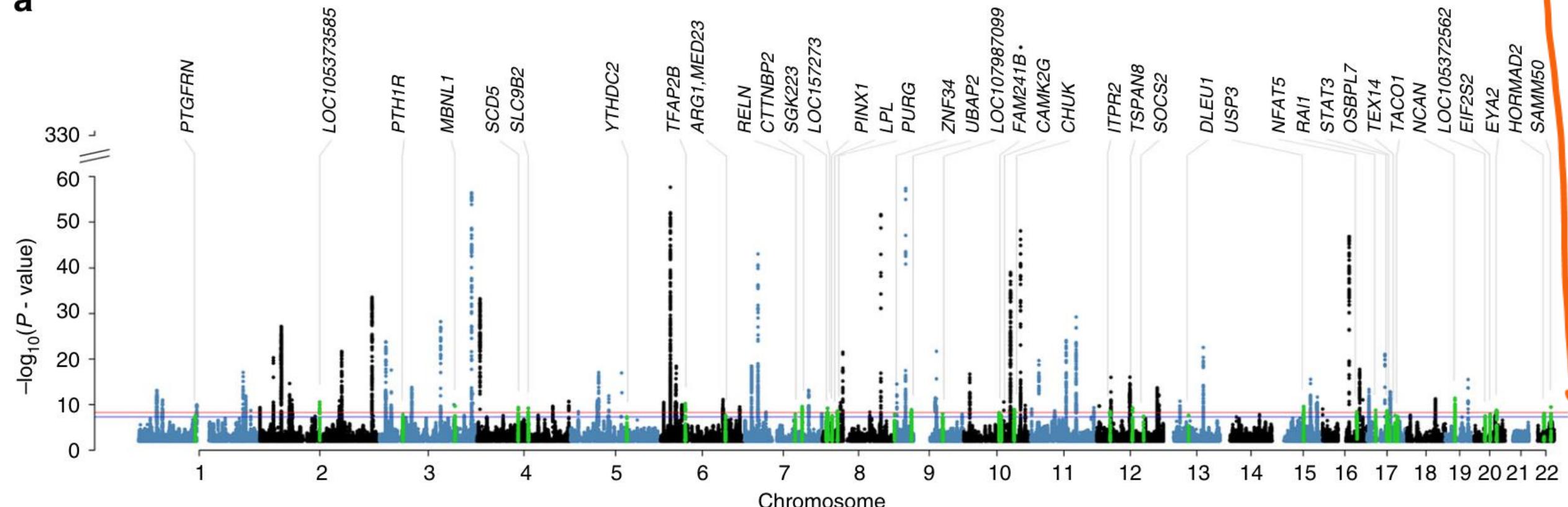
MANHATTAN PLOT

For each variant, plot the $-\log_{10}(P\text{-value})$ as function of chromosomal position.

Genome-wide significance level?

-adjust for no. of independent statistical tests
(1,000,000 independent genomic regions [LD])

a



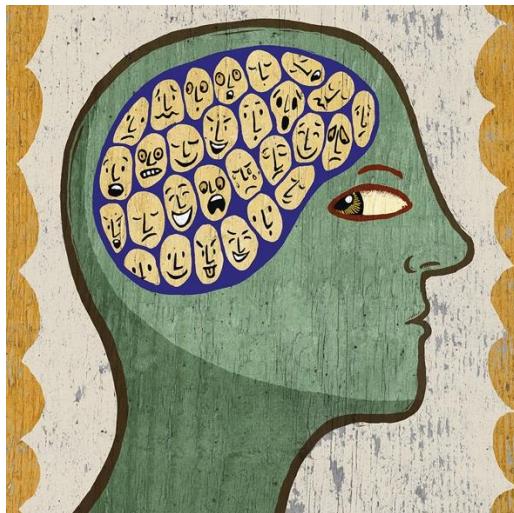
$$P=0.05 \rightarrow -\log_{10}(0.05) = 1.3$$

$$P=0.001 \rightarrow -\log_{10}(0.001) = 3$$

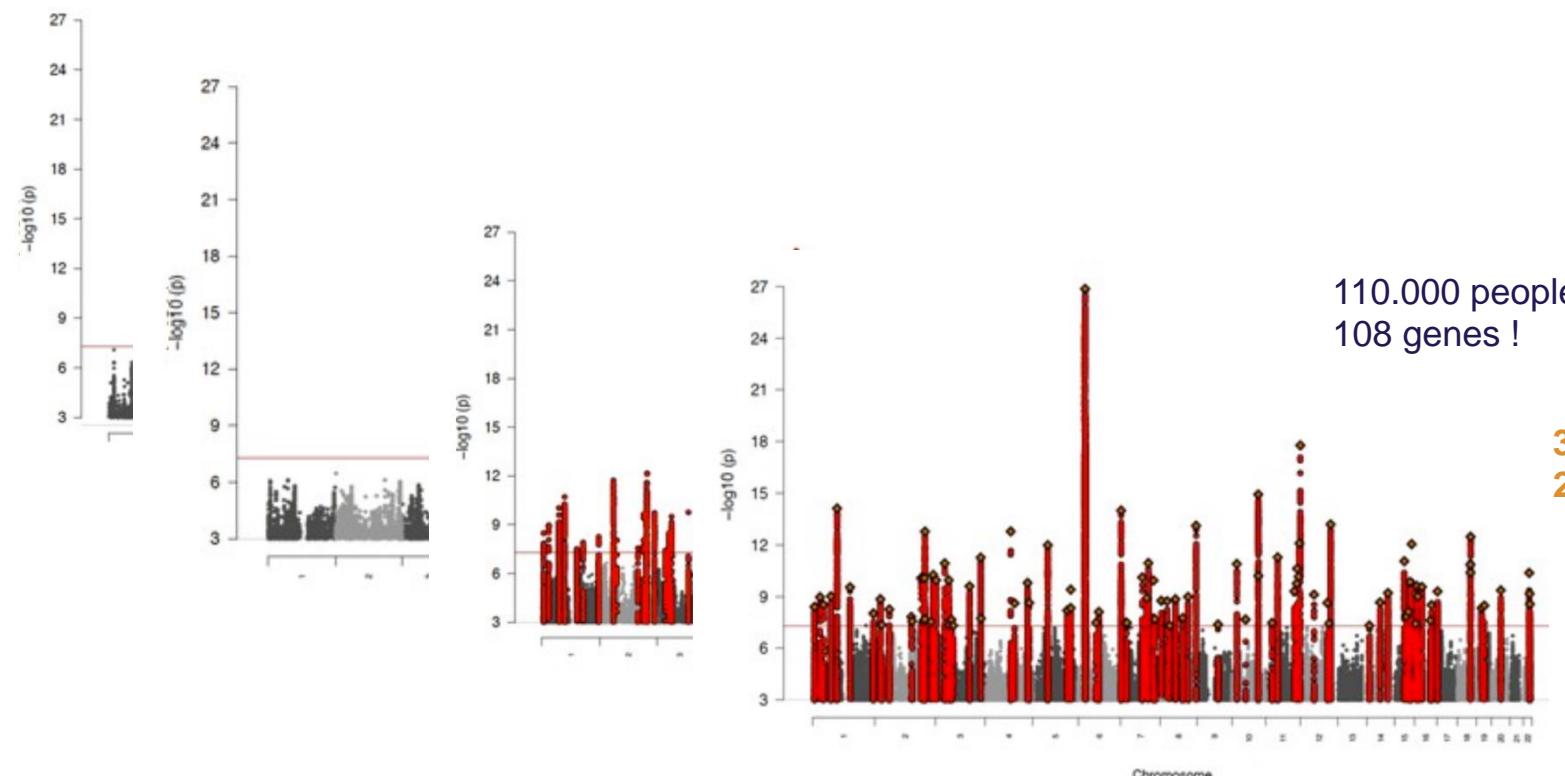
$$P=0.000000005 \rightarrow -\log_{10}(0.000000005) = 8.3$$

POWER IS EVERYTHING IN GWAS

Schizophrenia



Has high heritability ($h^2 = 0.85$)
The population prevalence is 1%
Emerge in late teens
Molecular aetiology is unknown

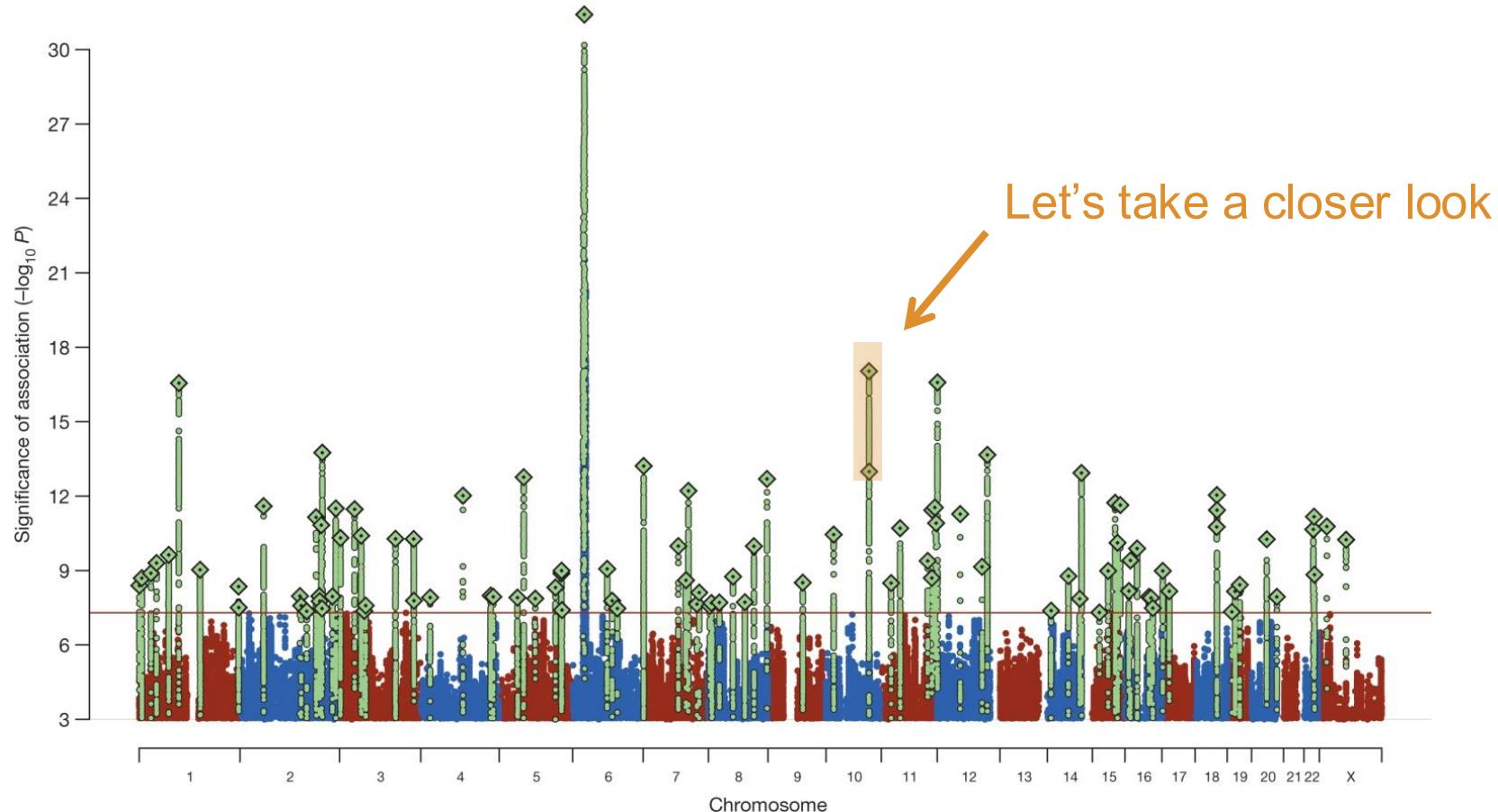


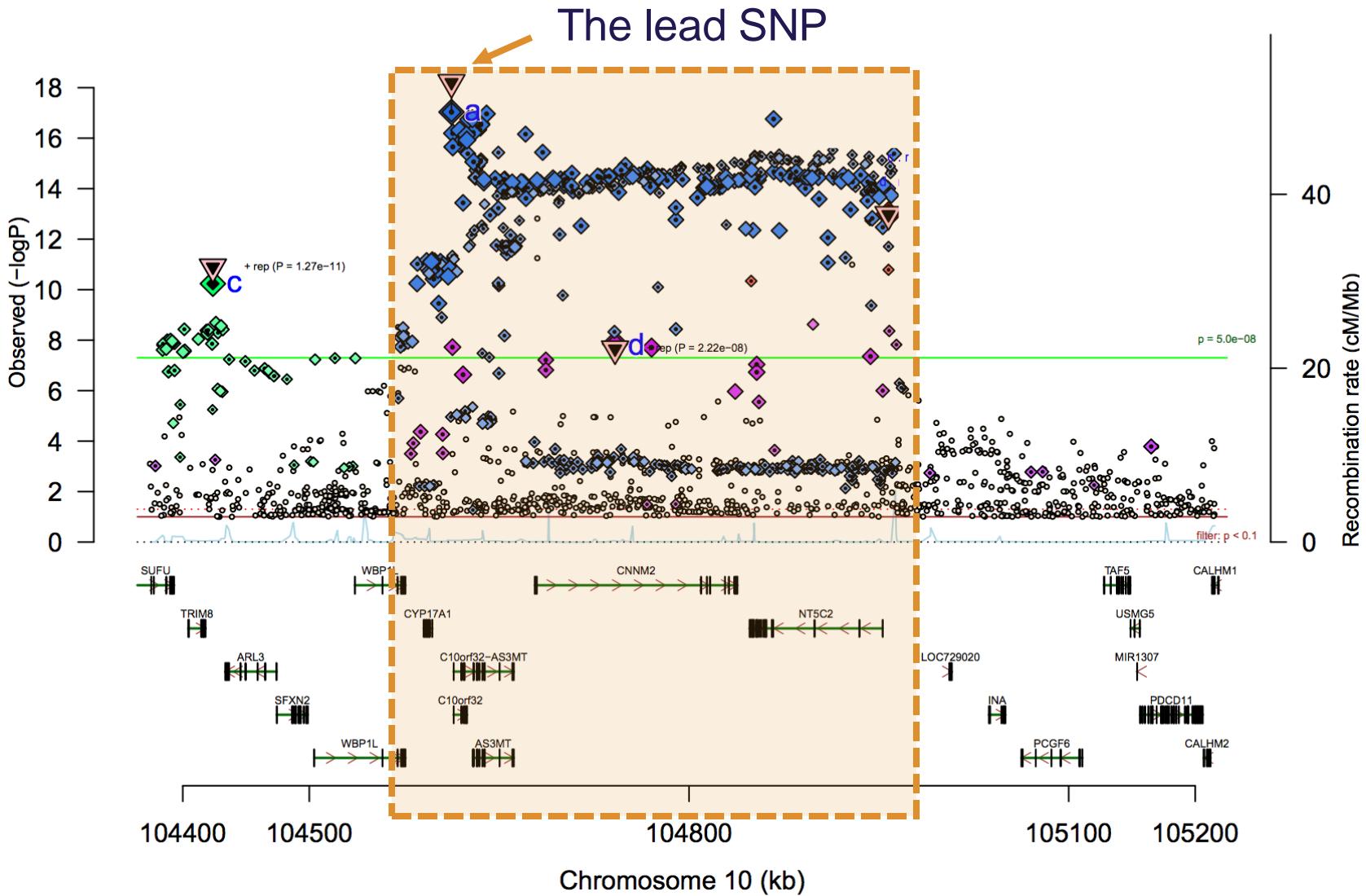
GWAS BY STEPS



- 5. Find the causal variant**
- 4. Summaries 10M linear regressions**
- 3. GWAS**
- 2. Extract genetic variants**
- 1. Select trait/disease**

Biological insights from 108 schizophrenia-associated genetic loci





All the “blue” variants are in LD – which gene is associated with SCZ?

Statistical “**Fine mapping**” fitting multiple variants at a time [adjust of LD]

FINE MAPPING

Statistical “**Fine mapping**” fitting multiple variants at a time [adjust of LD]

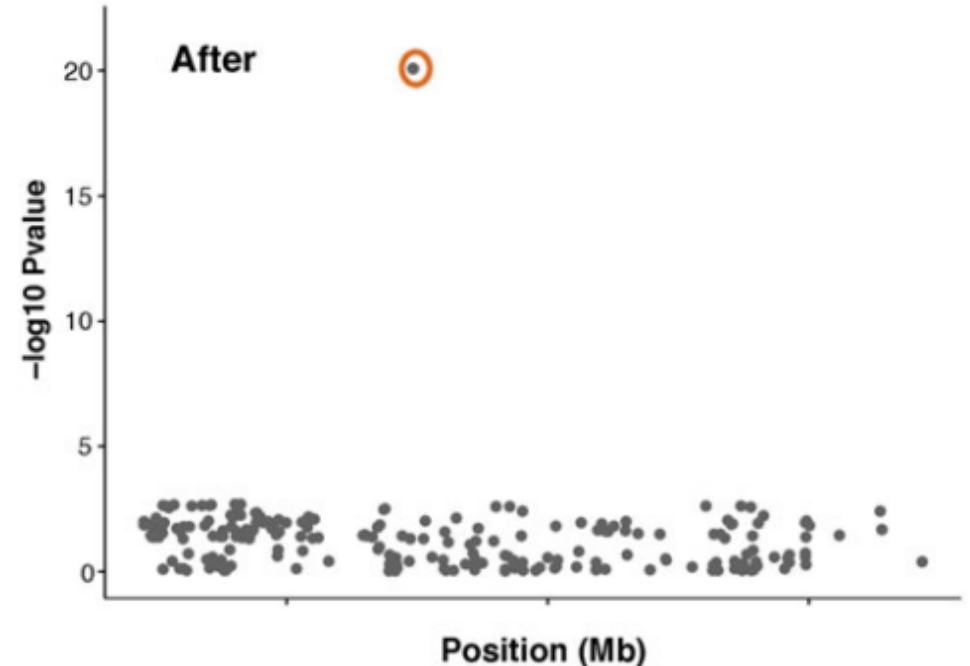
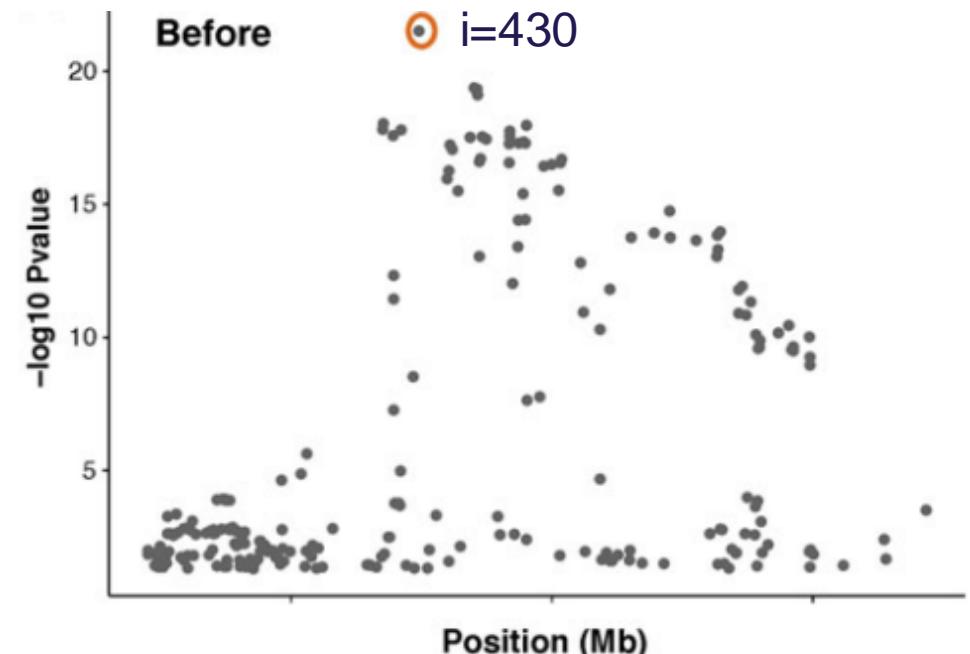
$$y = X_i \beta_i + g + e$$

$$y = X_{i=430} \beta_{i=430} + g + e$$

Condition on index variant

$$y = X_{i=430} \beta_{i=430} + X_{i=431} \beta_{i=431} + g + e$$

⋮
⋮



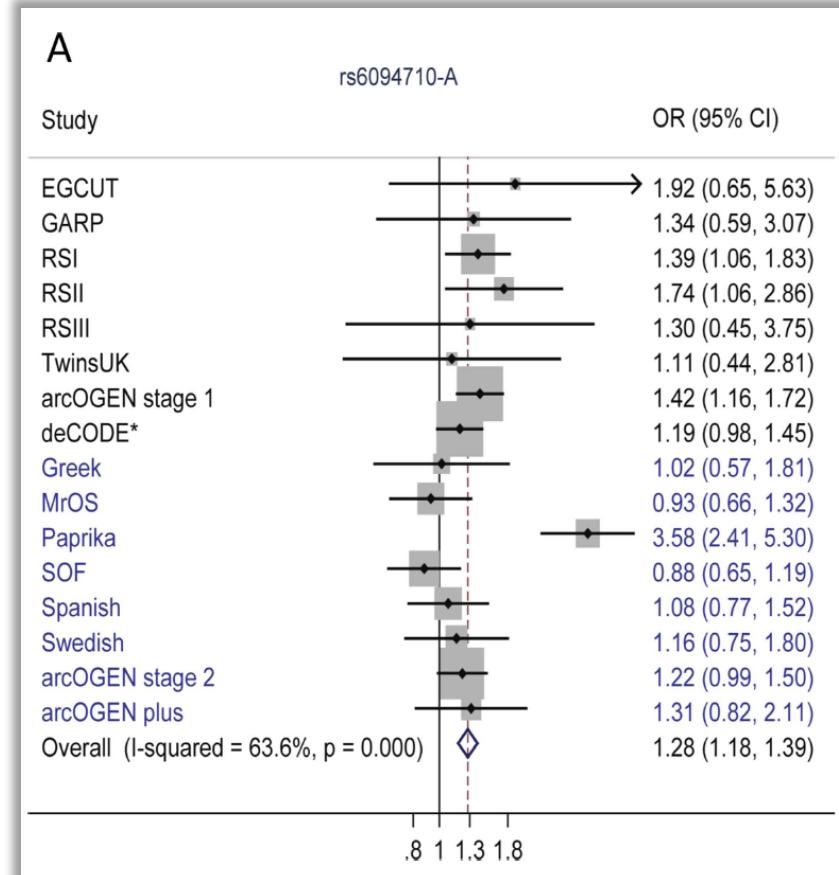
GWAS BY STEPS



- 1. Select trait/disease**
- 2. Extract genetic variants**
- 3. GWAS**
- 4. Summaries 10M linear regressions**
- 5. Find the causal variant**
- 6. Meta analysis**

META ANALYSIS

- Meta-analysis is a set of methods that allows the quantitative combination of data from multiple studies
- Meta-analysis of GWA datasets can increase the power to detect association signals by increasing sample size and by examining more variants throughout the genome than each dataset alone
- Assume two independent estimates $\hat{x}_1 = 1.0$ and $\hat{x}_2 = 2.0$, and that the precision of the first estimate is twice that of the second (precision = $1/SE^2$)
- $\hat{x}_{meta} = \frac{(2\hat{x}_1 + \hat{x}_2)}{2+1} = \frac{2+2}{3} = 1.33$



doi:10.1136/annrheumdis-2012-203114

$$\hat{\beta}_{l,F} = \frac{w_{1l}\hat{\beta}_{1l} + \dots + w_{Kl}\hat{\beta}_{Kl}}{w_{1l} + \dots + w_{Kl}} \quad \text{studies } 1, \dots, K$$

$$SE_{l,F} = (w_{1l} + \dots + w_{Kl})^{-\frac{1}{2}}, \quad \text{where the weight}$$

$$w_{kl} = \frac{1}{SE_{kl}^2} \text{ is the inverse-variance of study } k.$$

GWAS BY STEPS

7. Spurious associations

6. Meta analysis

5. Find the causal variant

4. Summaries 10M linear regressions

3. GWAS

2. Extract genetic variants

1. Select trait/disease



SPURIOUS ASSOCIATIONS

NEWS & VIEWS

Beware the chopsticks gene

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not. His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes.



Molecular Psychiatry (2000) 5, 11-13



Cases



Controls

SPURIOUS ASSOCIATIONS

NEWS & VIEWS

Beware the chopsticks gene

The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use, enough to account for nearly half of the observed variance. When the experiment was repeated with students from a different university, precisely the same marker lit up. Eureka! The delighted scientist popped a bottle of champagne and quickly submitted an article to *Molecular Psychiatry* heralding the discovery of the 'successful-use-of-selected-hand-instruments gene' (SUSHI).



Cases
 $p=0.50$



Controls
 $p=0.01$

SUSHI gene

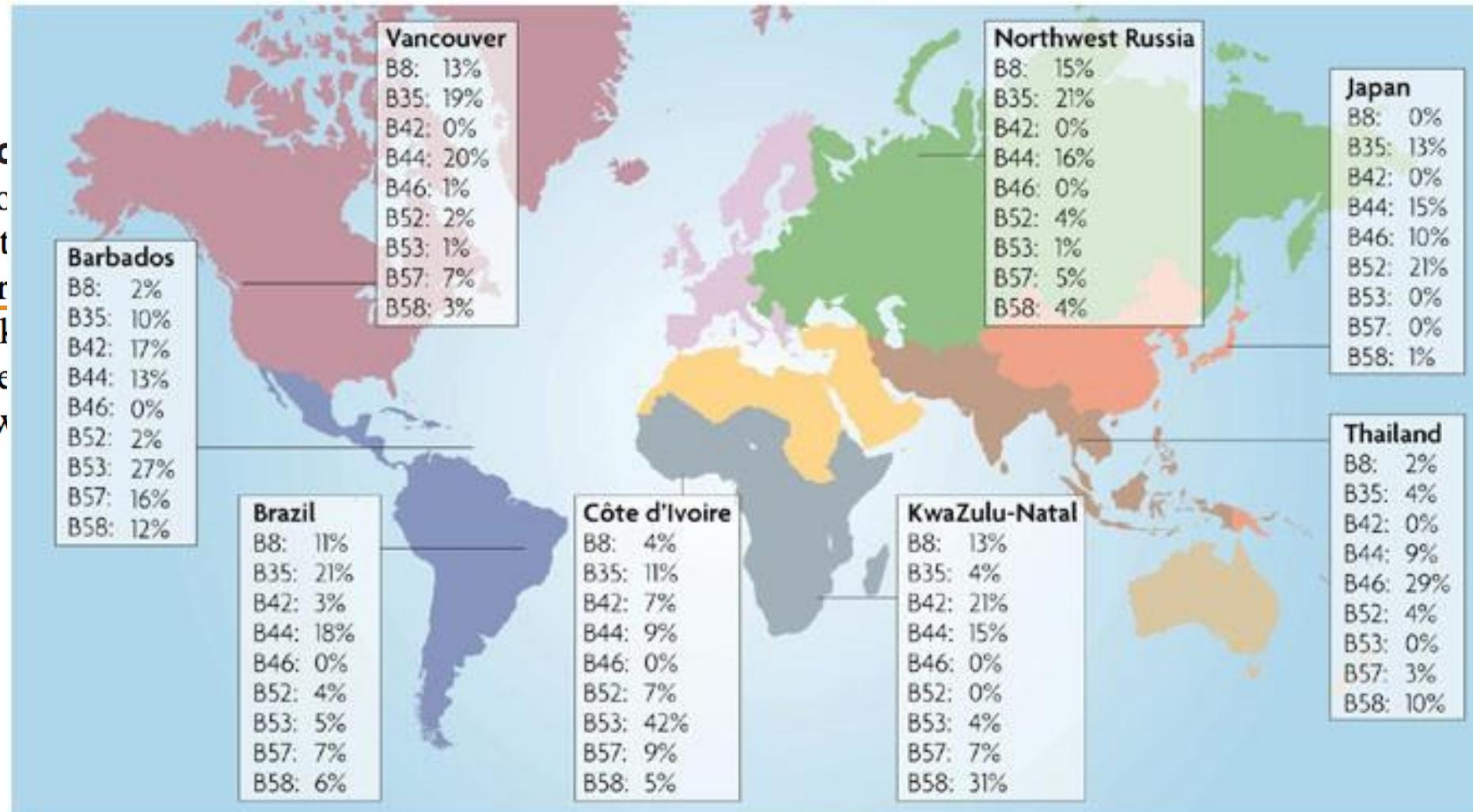
Are there any problems?

SPURIOUS ASSOCIATIONS

NEWS & VIEWS

Beware the chopsticks

It took another 2 years to find the first HLA class I antigen gene that has not been associated with disease. It turned out to have different allele frequencies in different populations. Of course differ in chopstick use is not a plausible explanation for such reasons. Even though the associations were statistically significant, they were readily replicated, they were not biologically plausible.



HLA class I diversity is illustrated by the prevalence of nine HLA-B molecules

GWAS BY STEPS

8. Post hoc analyses

7. Spurious associations

6. Meta analysis

5. Find the causal variant

4. Summaries 10M linear regressions

3. GWAS

2. Extract genetic variants

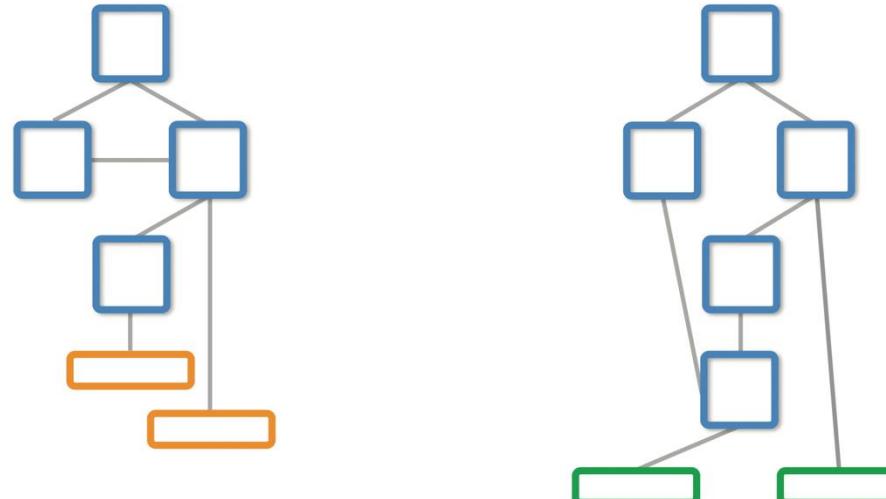
1. Select trait/disease

GWAS POST HOC ANALYSES

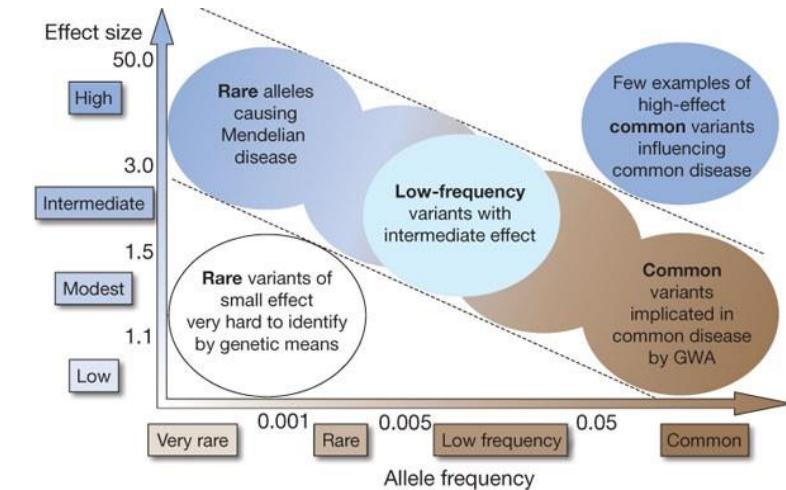
GENE SET ENRICHMENT



Genes (+/- regulatory sequences)
connected in **biological pathways** and **networks**



or other networks such as
protein interactions or **metabolites**



Looking for variants with small effects

In a **GWAS** we go through all SNPs one by one.

Gene enrichment analyses we examine whether a group of SNPs (within a biological entity) display a more extreme association signal than by chance.

HOW COMMON ARE GWASes?



$$g(x+h) - g(x)$$

$$= \lim_{h \rightarrow 0} \frac{x^2 + 2xh + h^2 - x^2}{h}$$

$$= \lim_{h \rightarrow 0} \frac{2xh + h^2}{h}$$

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

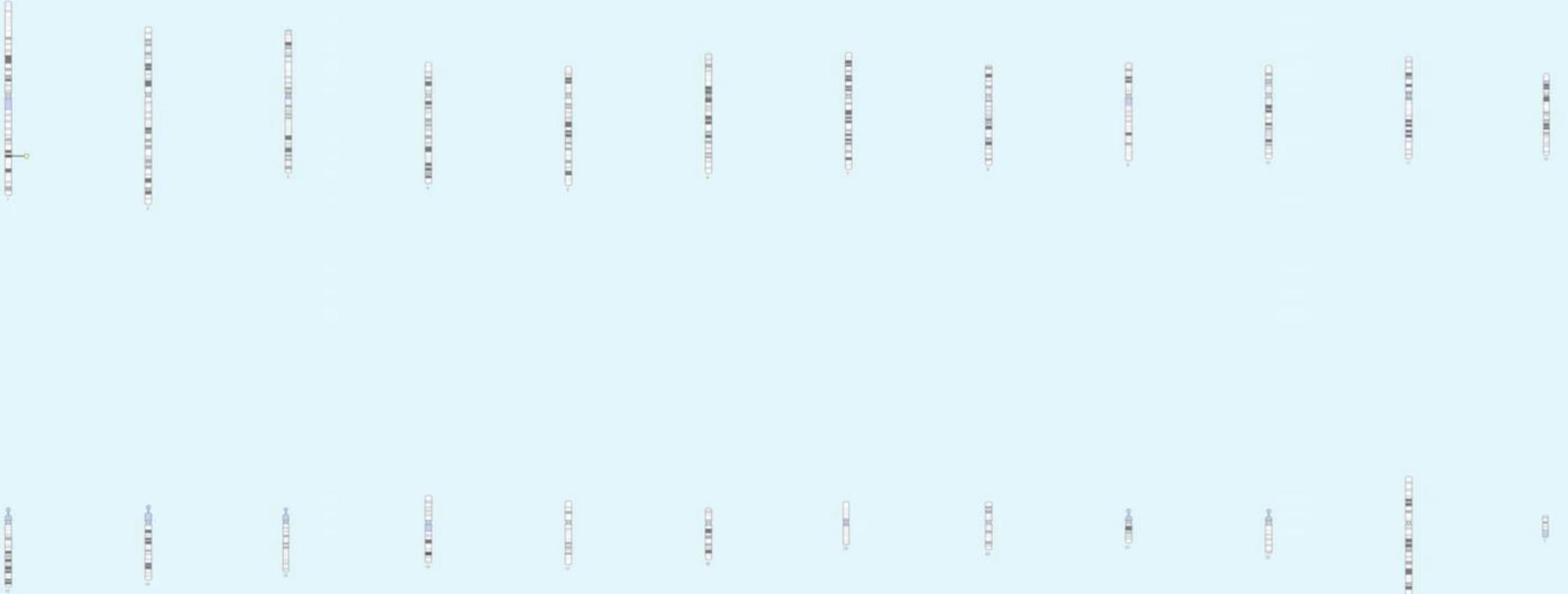
$$\begin{aligned} &= \lim_{h \rightarrow 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} \\ &= \lim_{h \rightarrow 0} \frac{(\sqrt{x+h} - \sqrt{x})(\sqrt{x+h} + \sqrt{x})}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \lim_{h \rightarrow 0} \frac{x+h-x}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \lim_{h \rightarrow 0} \frac{1}{\sqrt{x+h} + \sqrt{x}} \end{aligned}$$

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

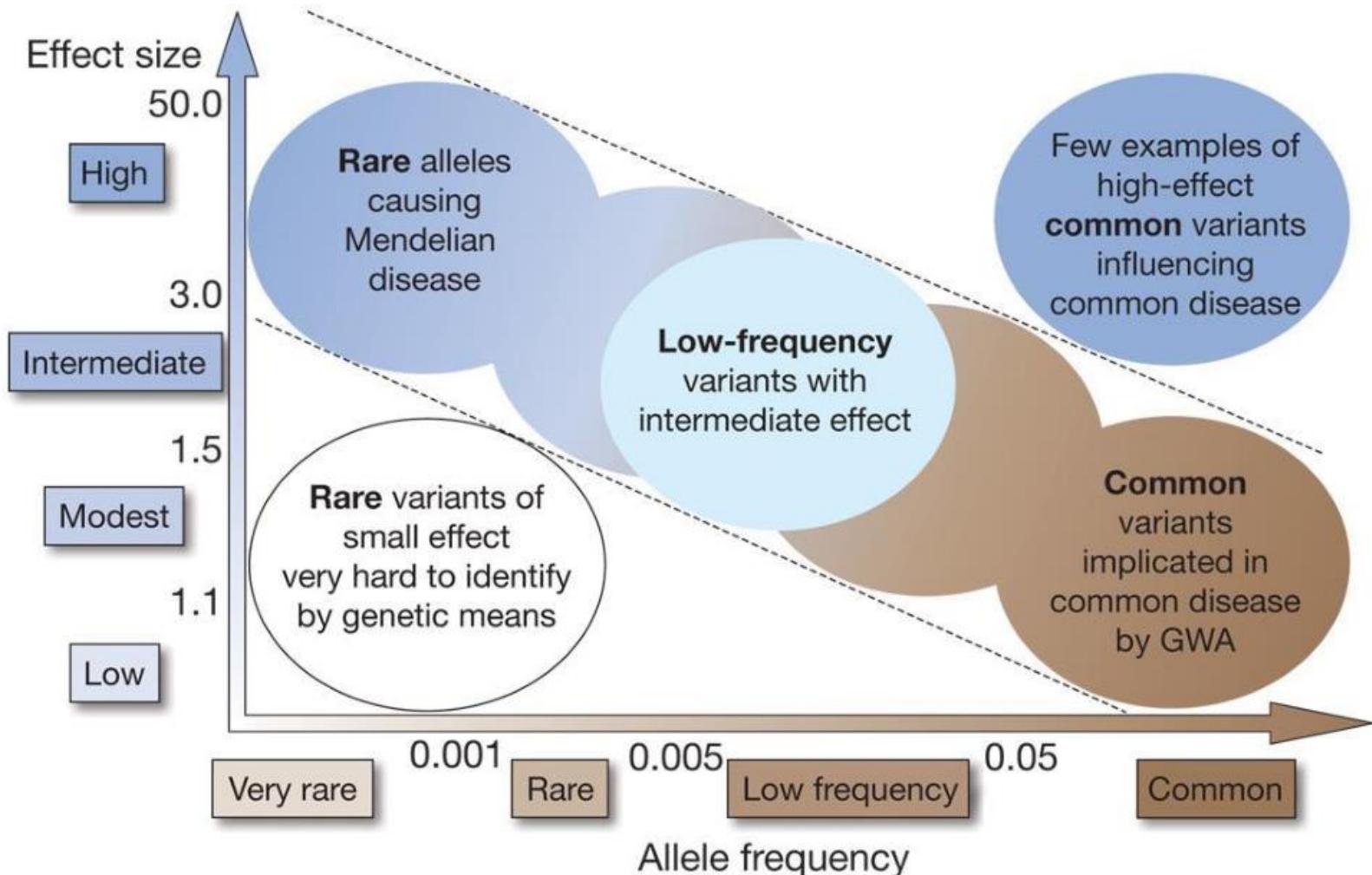
$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

2006 Jan

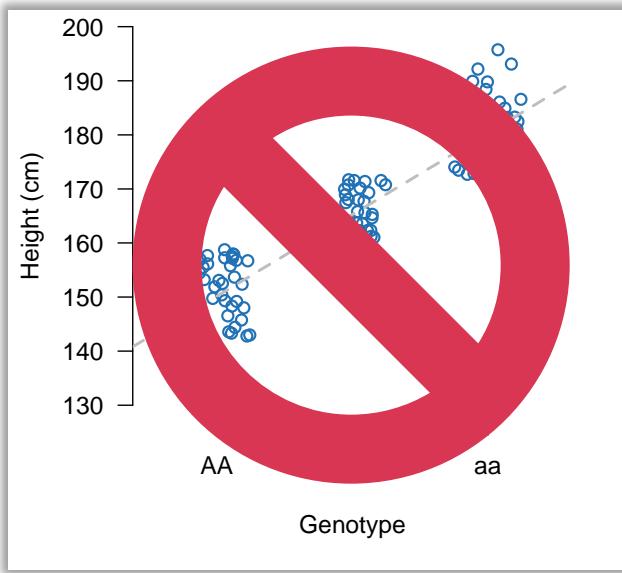


www.ebi.ac.uk/gwas

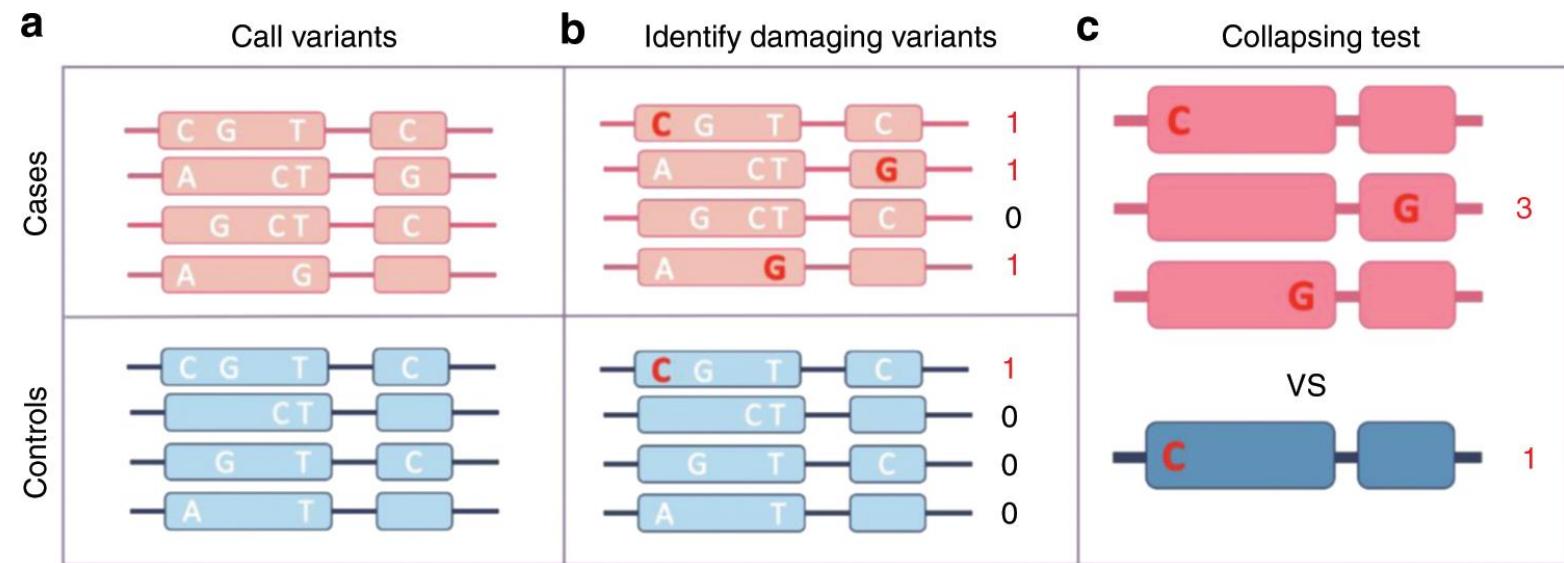


What about rare genetic variants for common diseases?

RARE VARIANT TESTING



If variation at the locus is rare
association testing is not possible
(no aa or few Aa individuals exists)

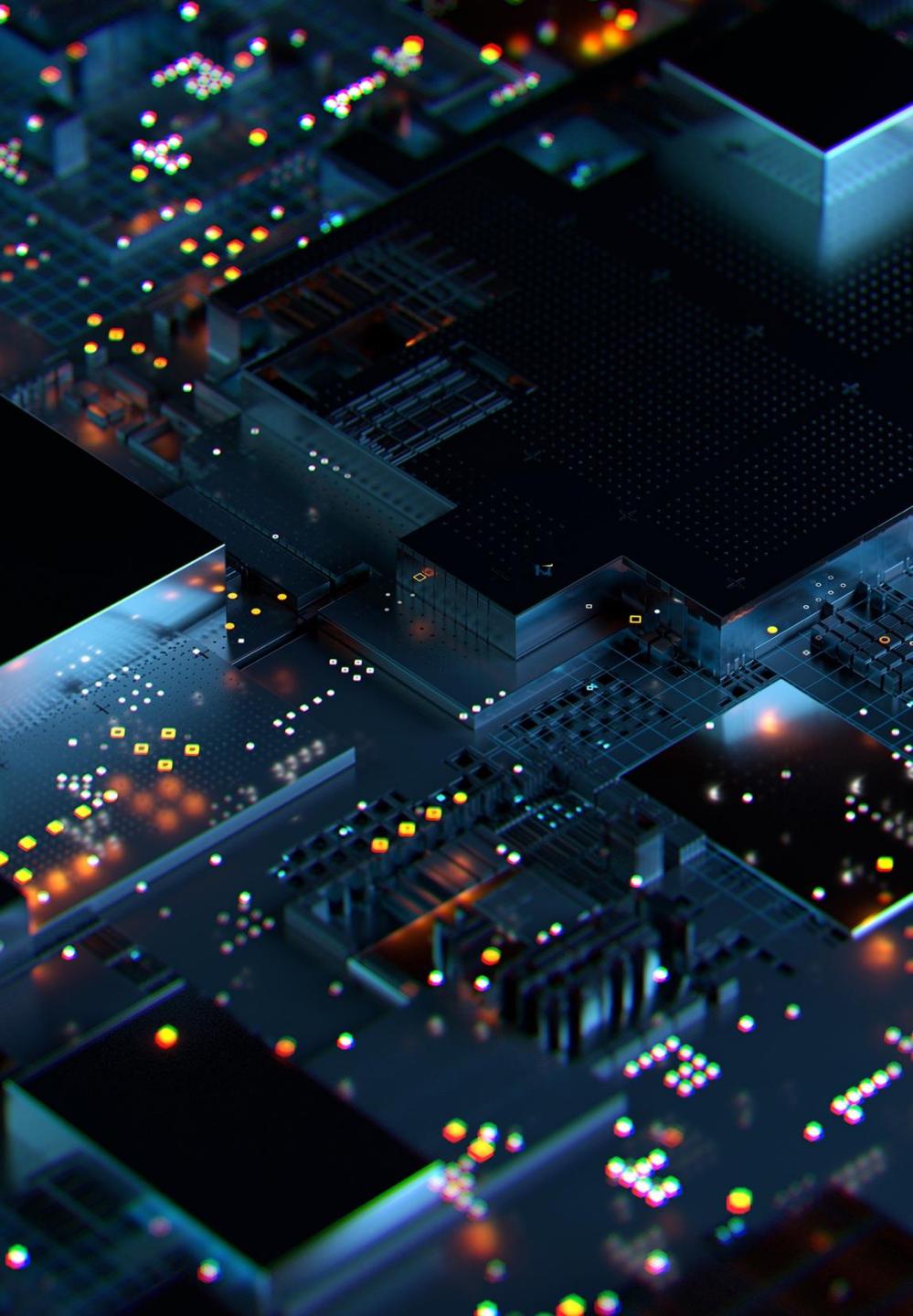


Test the “burden” of all rare variants within a gene.

Often variants are grouped by functional consequence
(missense, pLOF etc)

GENOME-WIDE ASSOCIATION ANALYSIS (GWAS)

- What is a GWAS
- LD
- GWAS by steps



GROUP WORK

GWAS AND FUTURE DIRECTION

- 1) Go into your 'complex traits' group [45 min]
 - Discuss what did you learn from the GWAS you selected
 - Read pages 3-10 in the ICDA white-paper
 - Discuss how the reccomdations could be important for your trait
- 2) Plenum discuss [10 min]



YOUR OPPINION MATTERS

MOODLE EVALUATION



List the two most important things you learned today	What did you find difficult?	What did you find easy?	Improvements for next session?
+ <input type="text"/>	+ <input type="text"/>	+ <input type="text"/>	+ <input type="text"/>