GWAS crash course

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@uliege.be

Genome-wide Association Studies

1 Setting the pace

Relevant questions and concepts

2 The rise of GWAs

3 Study Design Elements

3.a Marker level

3.b Subject level

3.c Gender level (not considered in this course)

4 Pre-analysis Steps

4.a Quality-Control

4.b Linkage disequilibrium

4.c Confounding by shared genetic ancestry

5 Analysis Steps

5.a Association / Regression

5.b Replication and Validation

5.c Causation &

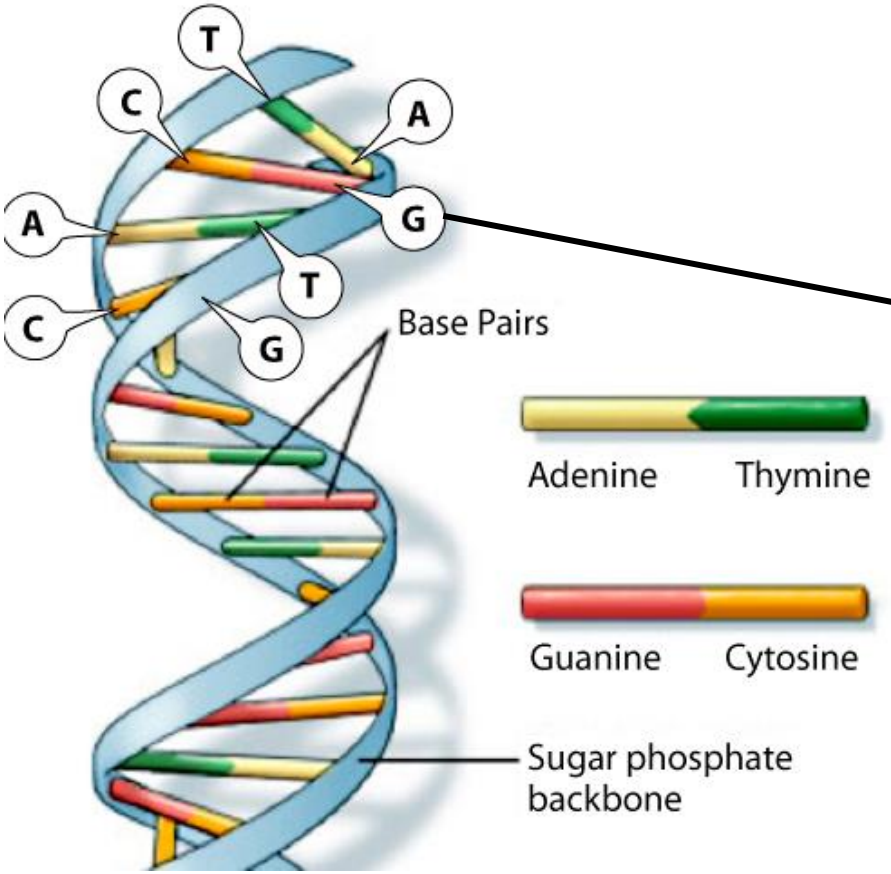
5.d Interpretation

1 Setting the pace

Play the following video on “molecular information”, and learn how information can be retrieved at different levels and scales

http://www.youtube.com/watch?v=00vBqYDBW5s

Types of genetic markers: single nucleotide polymorphisms



Single Nucleotide Polymorphisms (SNPs)	**Frequency in general population**
G	95%
A	5% > 1%

Types of genetic markers: single nucleotide polymorphisms or **SNPs**

- Variations in single base, i.e., one base substituted by another base
- In theory: four different nucleotides possible at base
- In practice: generally only two different nucleotides observed
- Definition strict and loose:
  - Strict: minor allele frequency ≥ 1%
  - Loose: ≥ 2 nucleotides observed in two individuals at position
- Nomenclature:
  - ss-number (submitted SNP number)
  - rs-number: searchable in dbSNP, mapped to external resources, unique
  - rs-numbers do not provide information about possible function of SNP
  - Alternative: nomenclature of Human Genome Variation Society

(Ziegler and Van Steen, Brazil 2010)

Types of genetic markers: single nucleotide polymorphisms

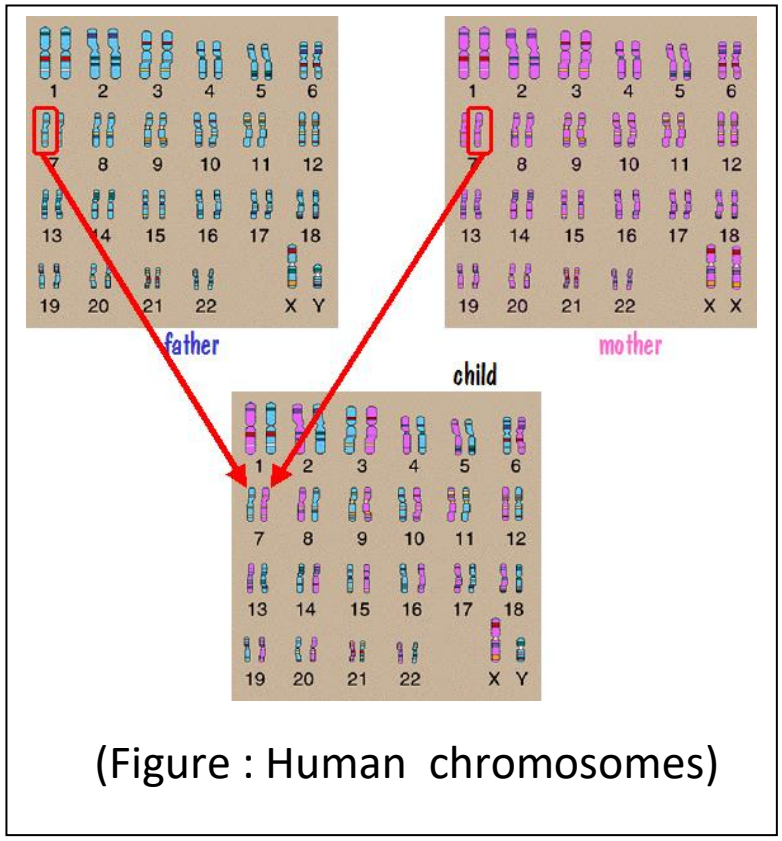
**Submissions received after reclustering of current build will appear as new rs# clusters in the next build.*

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency	Number of weight 1 SNPs	Number of weight 2+ SNPs
Homo sapiens	150	38.3	907,237,763	325,658,303 (135,967,291)	191,585,061	73,917,935	129,875,536		
Bos taurus	150	7.2	332,061,559	104,286,568 (12,102,319)	46,308,631	10,202	968		
Mus musculus	150	38.5	189,214,027	84,152,707 (6,466,270)	40,278,667	24,843,897	77	DIV:9911312 MNV:452 Named:6779 SNV:67883617	DIV:180165 MNV:2259 SNV:1647286
Sus scrofa	150	5.1	195,656,177	67,116,509 (8,107,358)	36,126,981	52	184		
Ovis aries	150	2.1	147,584,937	63,745,118 (3,570,277)	30,029,327	65	173		
Macaca mulatta	150	2.1	95,808,453	53,929,680 (2,760,325)	23,087,008	29	8,072	DIV:9 SNV:32798877	SNV:38416
Zea mays	150	1.1	86,608,237	58,915,360 (14,672,946)	13,436,128	90			
Gallus gallus	150	4.1	73,244,003	24,277,657 (15,305,602)	14,926,051	3,624,831	203		
Bos indicus	150	1.1	30,533,959	17,758,946 (621)	5,131,669		223		
Arabidopsis thaliana	150	9.2	15,307,574	13,412,809 (5,947)	9,174,636	299		DIV:4 MNV:5 SNV:1069121	MNV:1 SNV:338

Genes

- The **gene** is the basic physical unit of inheritance.
- Genes are passed from parents to offspring and contain the information needed to specify traits.
- They are arranged, one after another, on structures called chromosomes.
- A chromosome contains a single, long DNA molecule, only a portion of which corresponds to a single gene.



(Figure : Human chromosomes)

Gene Annotation

- An annotation (irrespective of the context) is a note added by way of explanation or commentary.
- **Genome annotation** is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do.
- Once a genome is sequenced, it needs to be annotated to make sense of it

→ links to giving an “interpretation”

Alleles

- Allele: one of several alternative forms of DNA sequence at specific chromosomal location
- Polymorphism: often used to indicate the existence of at least 2 alleles at a single “locus”
- **Homozygosity** (homozygous): both alleles identical at locus
- **Heterozygosity** (heterozygous): different alleles at locus

- Genetic marker (in this course): polymorphic DNA sequence at single locus

[Mutations ~polymorphisms (see later)]

How to generate a genetic map?

- To produce a genetic map, researchers collect blood or tissue samples from **family members** where a certain disease or trait is prevalent.
- Using various laboratory techniques, the scientists isolate DNA from these samples and examine it for the unique patterns of bases seen only in family members who have the disease or trait. These characteristic molecular patterns are referred to as polymorphisms, or markers.
- Before researchers identify the gene responsible for the disease or trait, DNA markers can tell them roughly where the gene is on the chromosome.

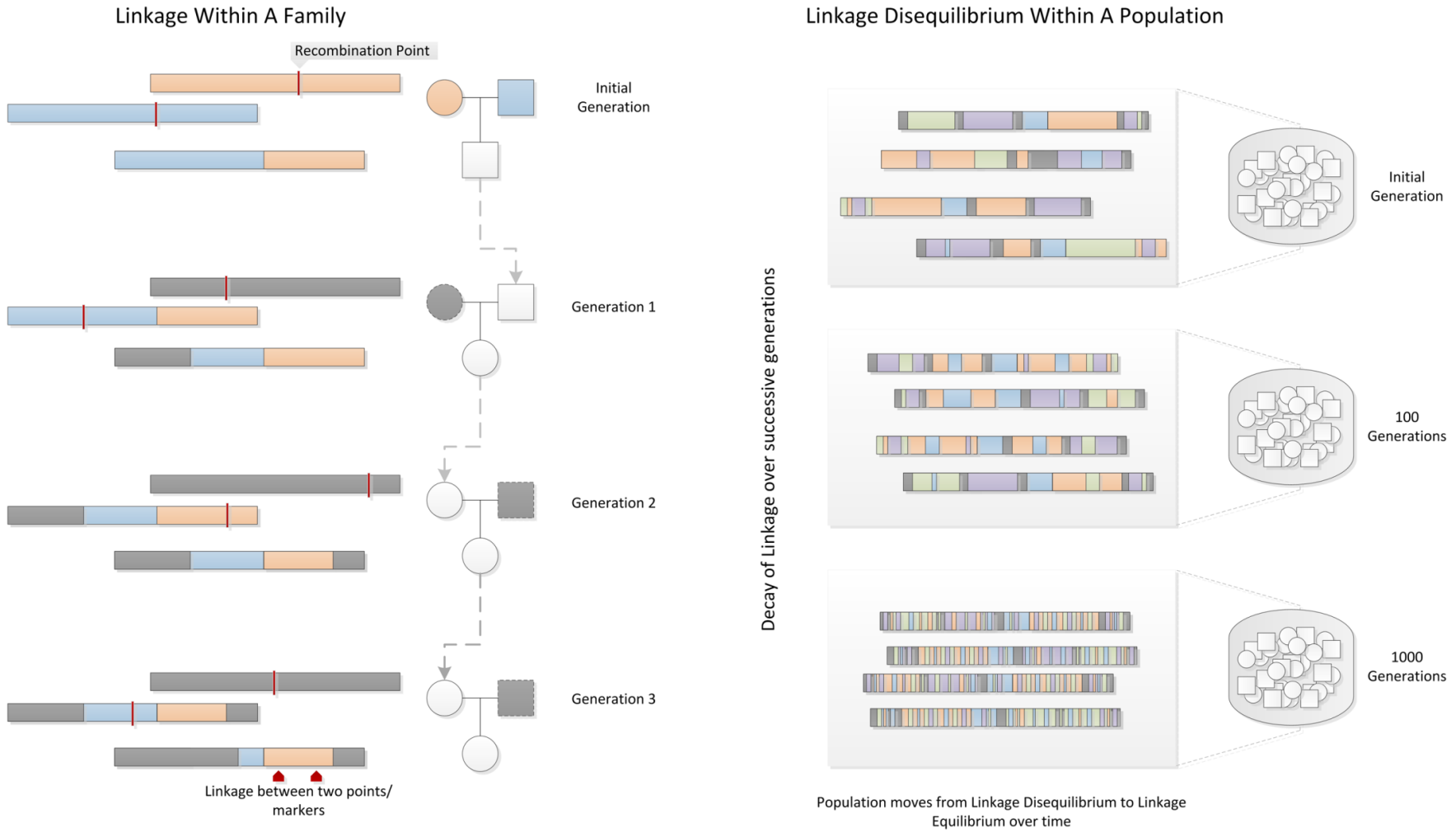
How is this possible?

How to generate a genetic map? (continued)

- This is possible because of a genetic process known as recombination.

> *As eggs or sperm develop within a person's body, the 23 pairs of chromosomes within those cells exchange - or recombine - genetic material. If a particular gene is close to a DNA marker, the gene and marker will likely stay together during the recombination process, and be passed on together from parent to child. So, if each family member with a particular disease or trait also inherits a particular DNA marker, chances are high that the gene responsible for the disease lies near that marker.*

How to generate a genetic map? (continued)



(Bush et al. 2012)

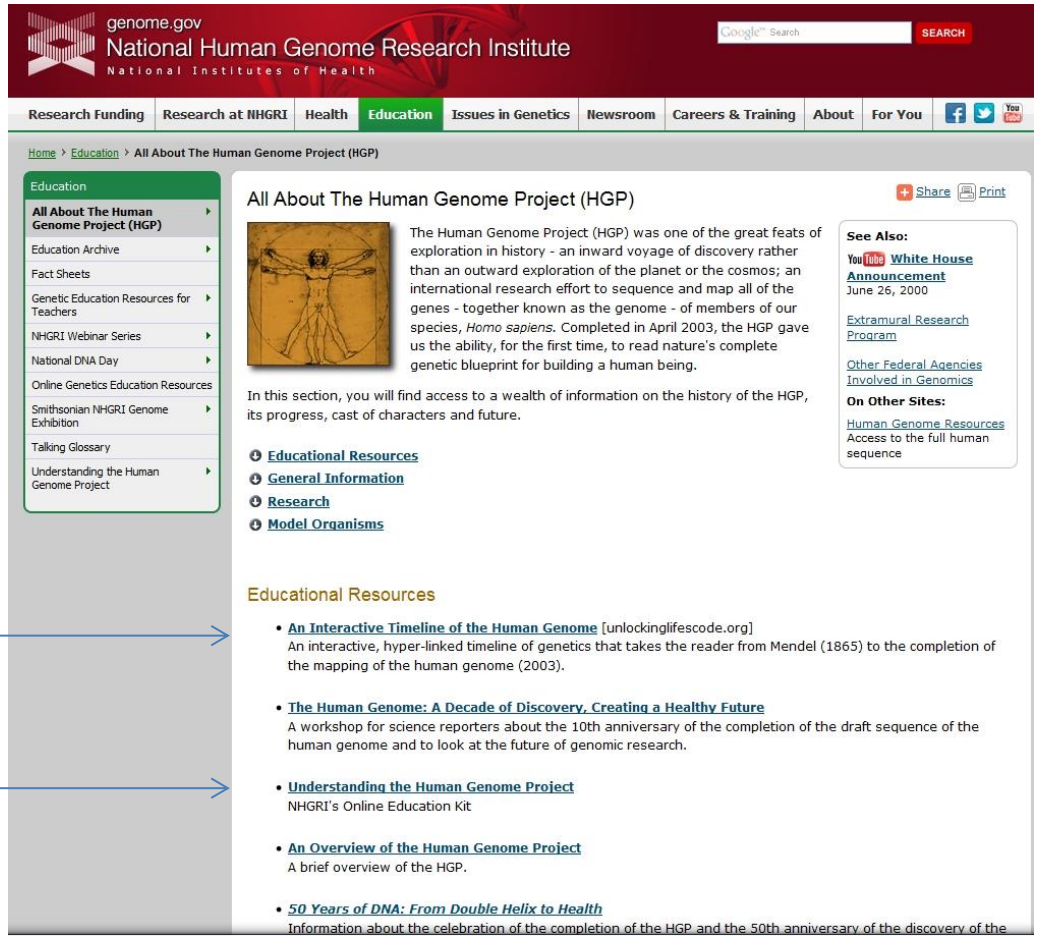
How to generate a genetic map? (continued)

- The more DNA markers there are on a genetic map, the more likely it is that one will be closely linked to a disease gene - and the easier it will be for researchers to zero-in on that gene.

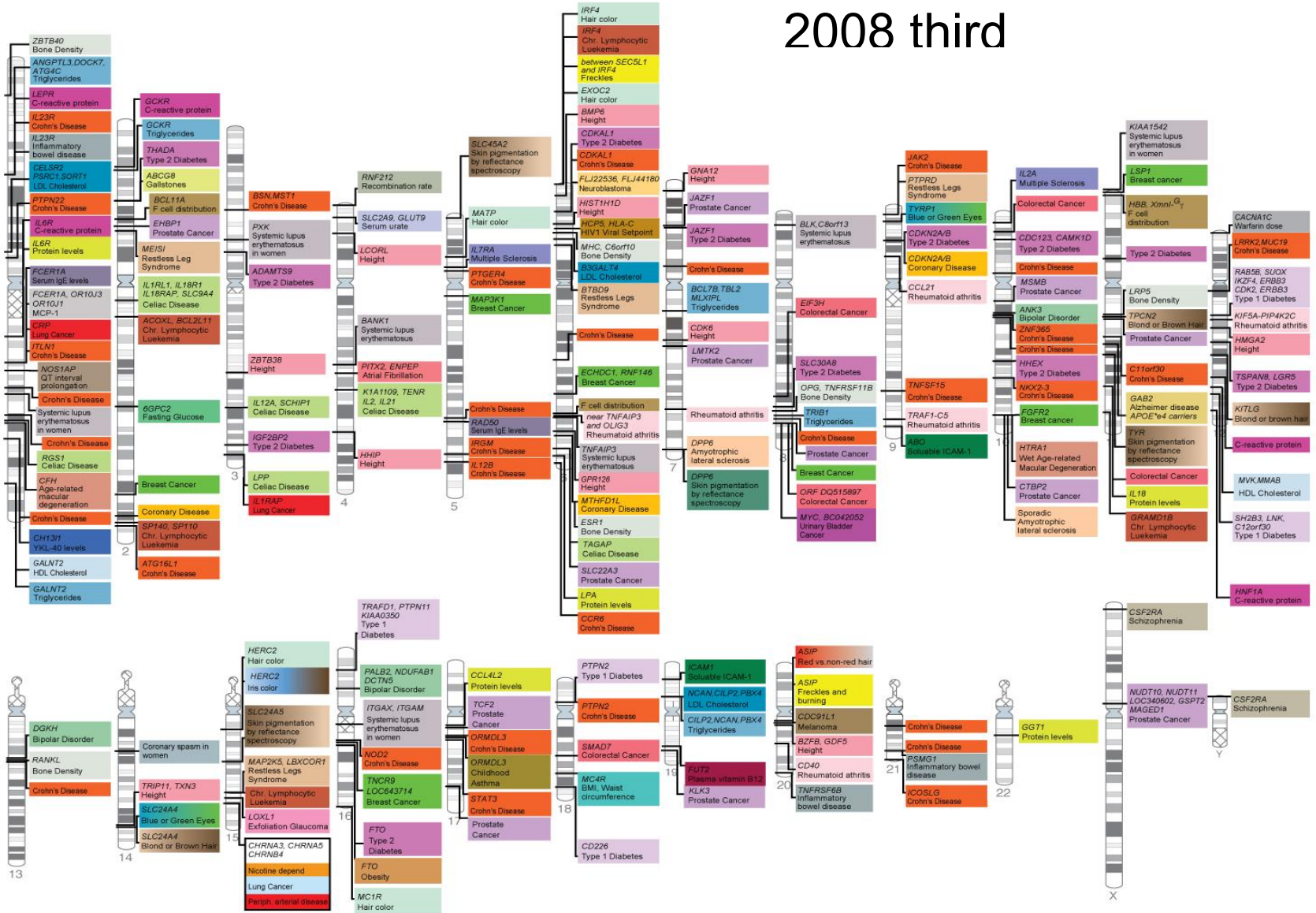
- One of the **first major achievements of the HGP was to develop dense maps of markers spaced evenly across the entire collection of human DNA**.

(http://www.genome.gov/10000715#al-3)

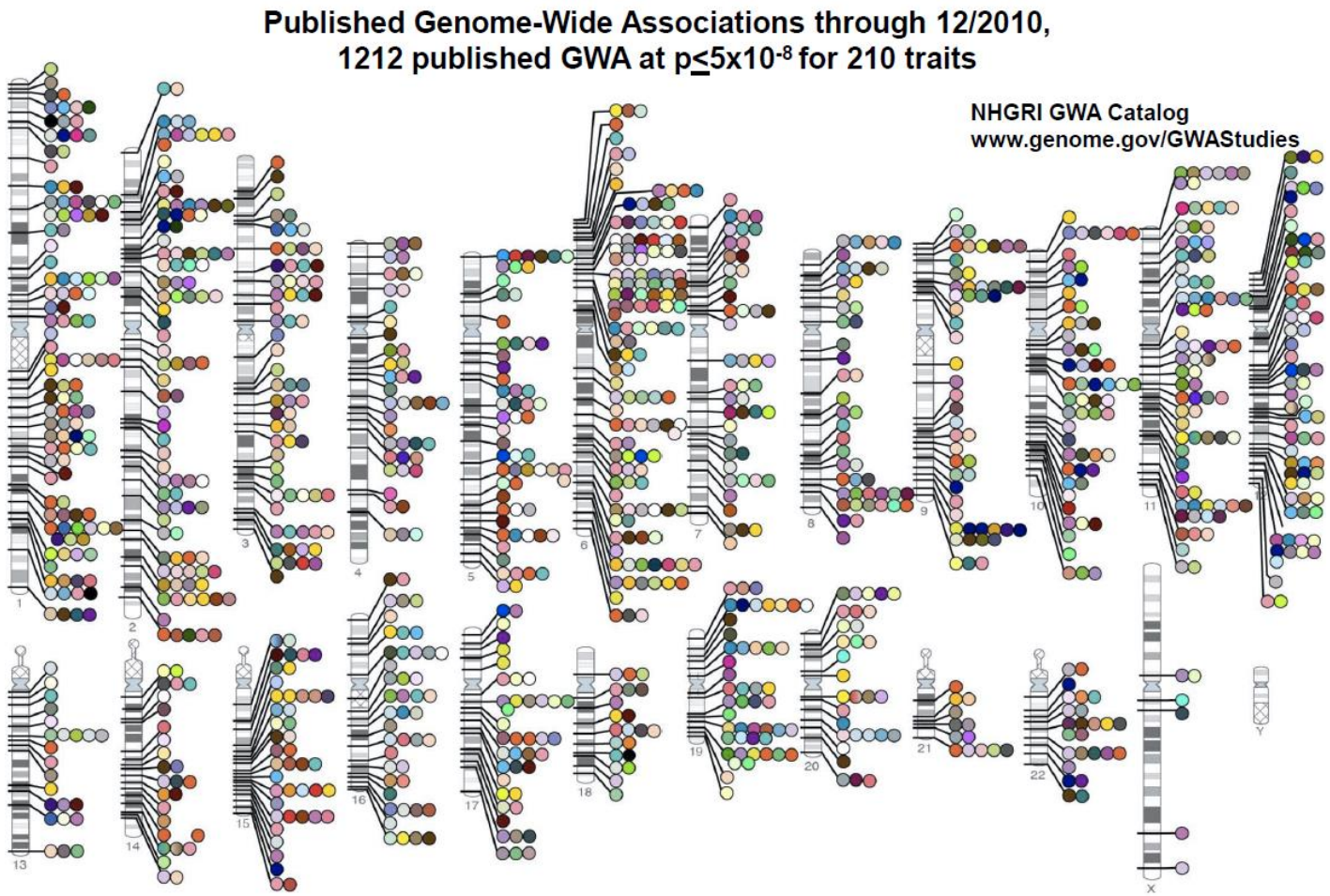
“The Human Genome Project”



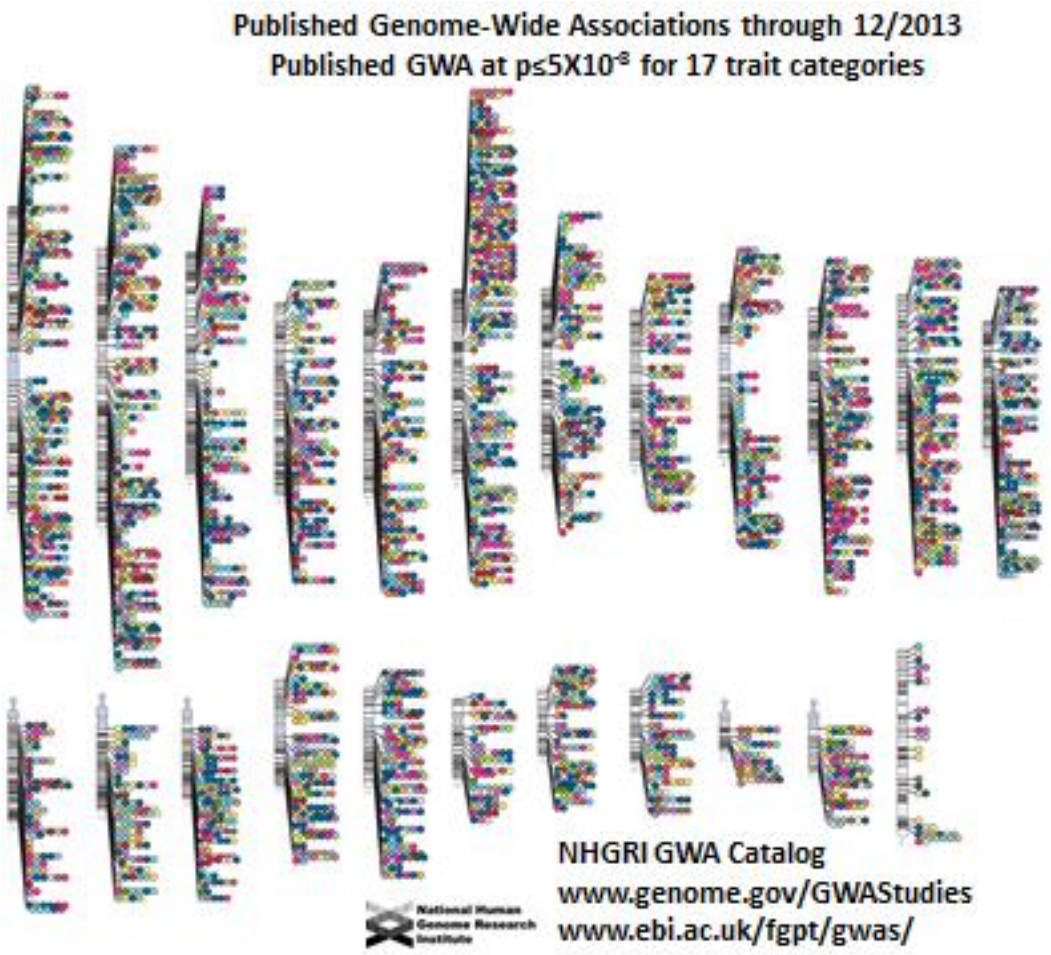
Historical overview: GWAs as a tool to “map” diseases



Historical overview: 210 traits – multiple loci (sites, locations)

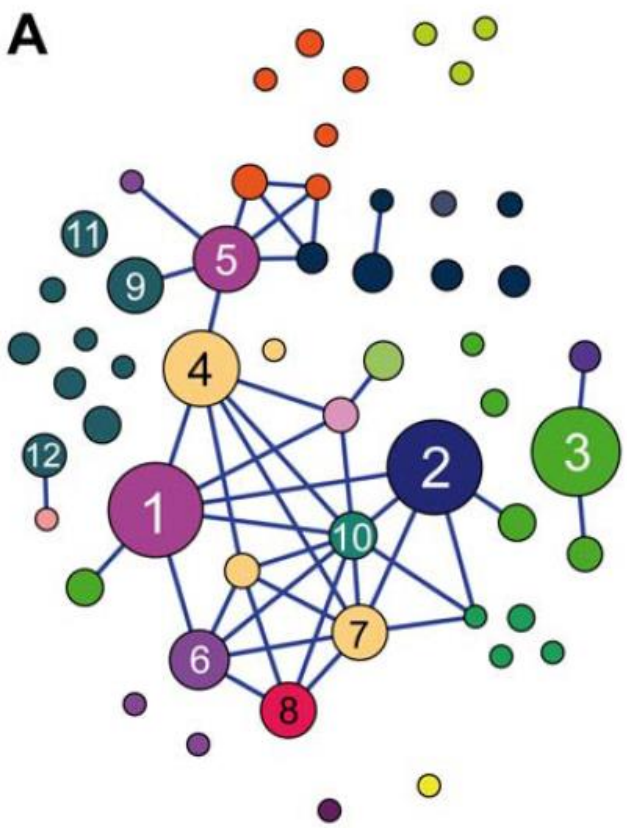


Historical overview: trait categories



- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurment
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

Historical overview: inter-relationships (networks)

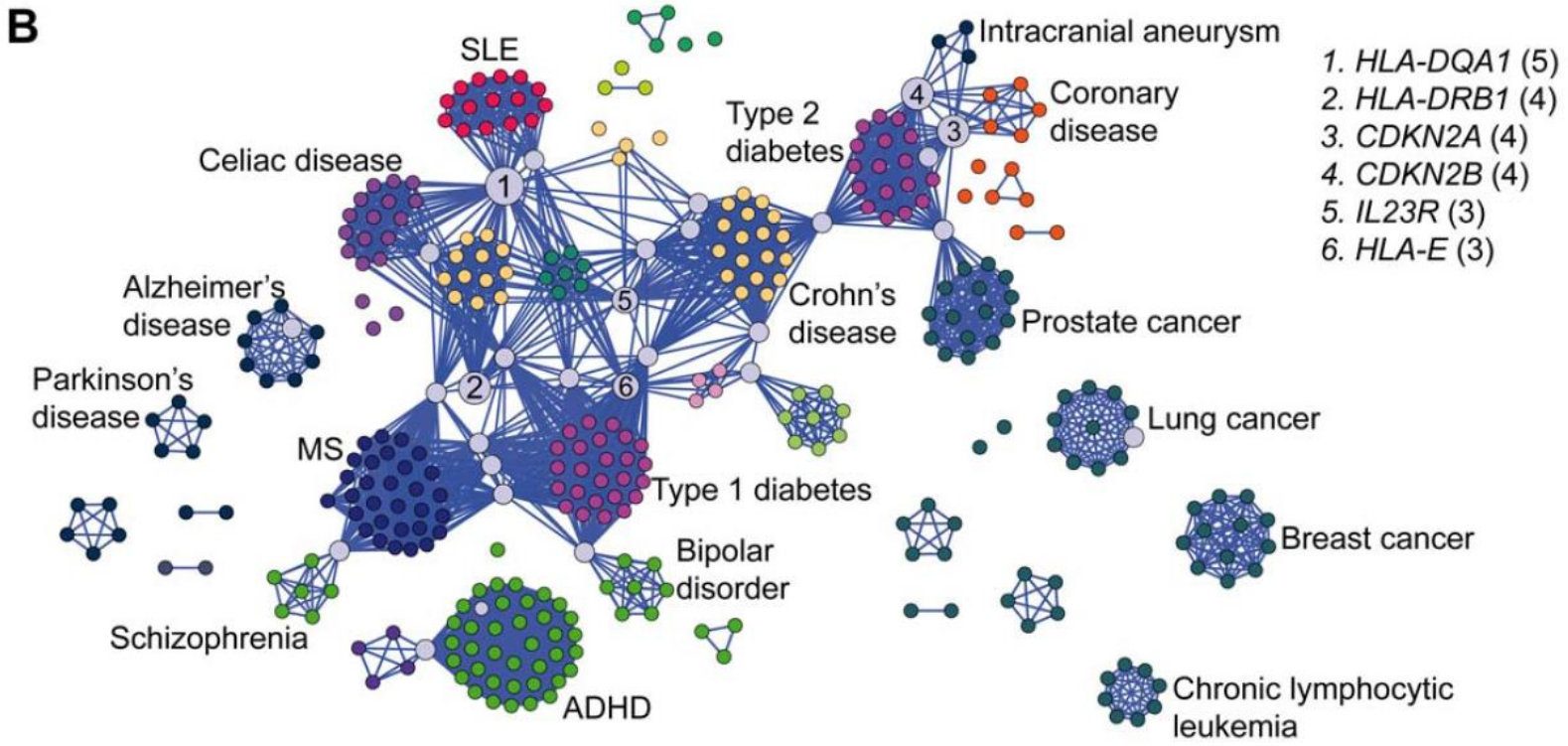


1. Type 1 diabetes (36)
2. Multiple sclerosis (36)
3. ADHD and conduct disorder (33)
4. Crohn's disease (27)
5. Type 2 diabetes (22)
6. Celiac disease (19)
7. Ulcerative colitis(17)
8. Systemic lupus erythematosus (17)
9. Prostate cancer (17)
10. Rheumatoid arthritis (13)
11. Breast cancer (12)
12. Lung cancer (11)

- Cardiovascular diseases (Cv)
- Digestive system diseases
- Endocrine system diseases
- Eye diseases
- Immune system diseases (Is)
- Mental disorders
- Multiple diseases
- Musculoskeletal diseases (Ms)
- Ms, Sc, Is
- Neoplasms
- Nervous system diseases (Ns)
- Ns, Cv
- Ns, Is
- Ns, Ms
- Nutritional and metabolic diseases
- Nm, Es, Is
- Skin and connective tissue disease
- Sc, Is
- Urogenital diseases

(Barrenas et al 2009: complex disease network – nodes are diseases)

Historical overview: inter-relationships (networks)



(Barrenas et al 2009: complex disease GENE network – nodes are genes)

2 The rise of GWAs

(slide Doug Brutlag 2010)

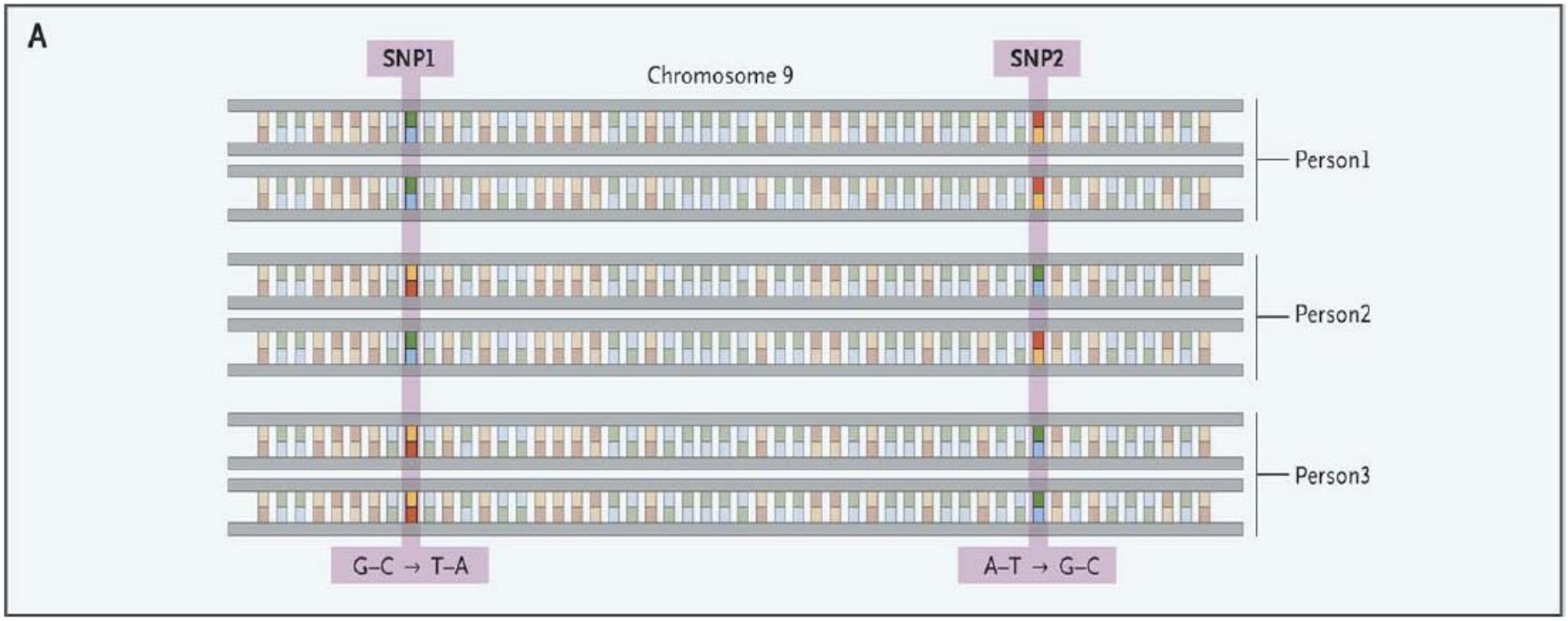
What are GWAs?

- A **genome-wide association study** is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular trait.

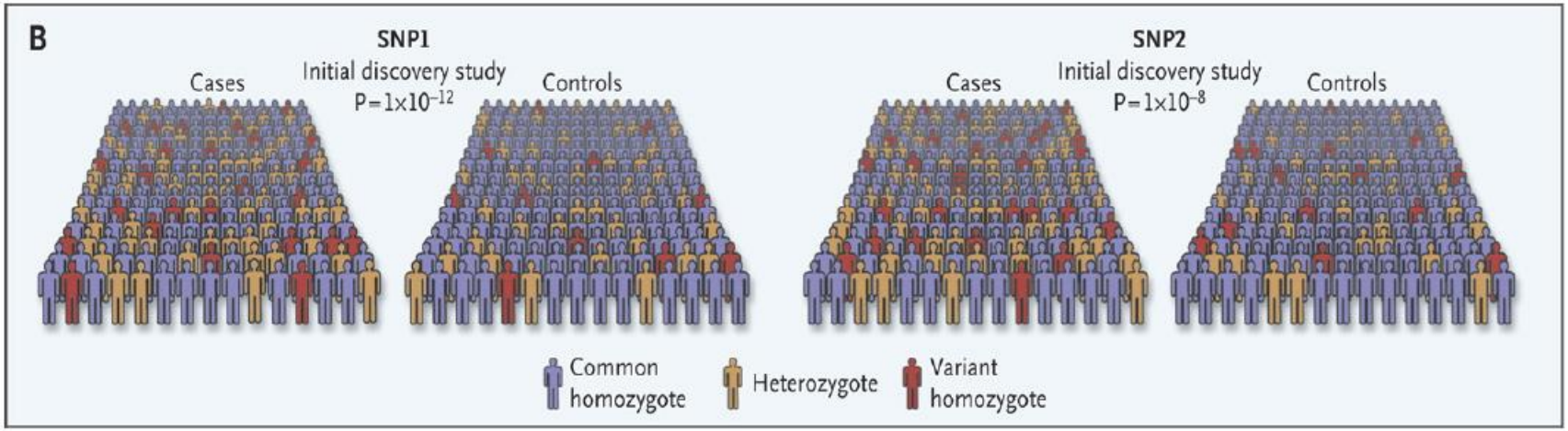
- **Recall**: a **trait** can be defined as a coded phenotype, a particular characteristic such as hair color, BMI, disease, gene expression intensity level, …

Genome-wide association studies: basic principles

The genome-wide association study is typically (but not solely!!!) based on a case‑control design in which single-nucleotide polymorphisms (SNPs) across the human genome are genotyped ... (Panel A: small fragment)



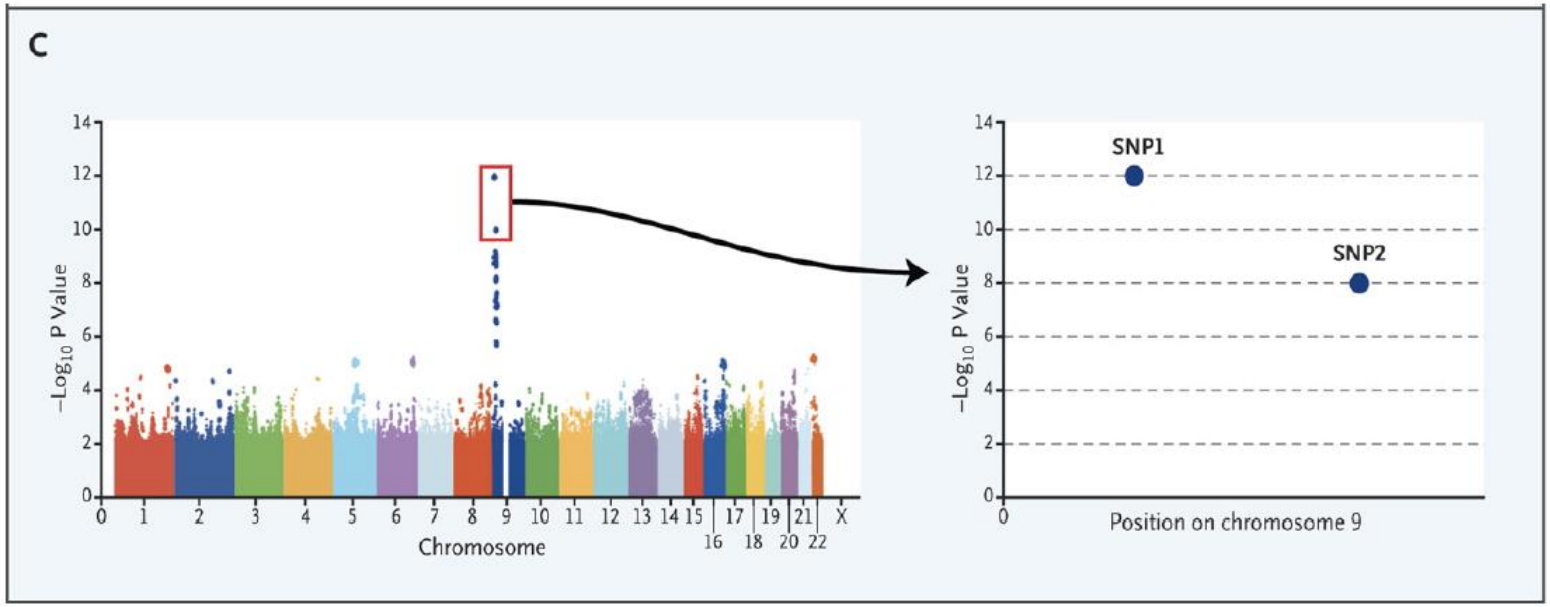
Genome-wide association studies: basic principles



- Panel B, the strength of association between each SNP and disease is calculated on the basis of the prevalence of each SNP in cases and controls. In this example, SNPs 1 and 2 on chromosome 9 are associated with disease, with P values of 10^{-12} and 10^{-8}, respectively

(Manolio 2010)

Genome-wide association studies: basic principles



- The plot in Panel C shows the P values for all genotyped SNPs that have survived a quality-control screen (each chromosome, a different color).
- The results implicate a locus on chromosome 9, marked by SNPs 1 and 2, which are adjacent to each other (graph at right), and other neighboring SNPs. (Manolio 2010)

View the GWAs catalogue (http://www.genome.gov/gwastudies/)

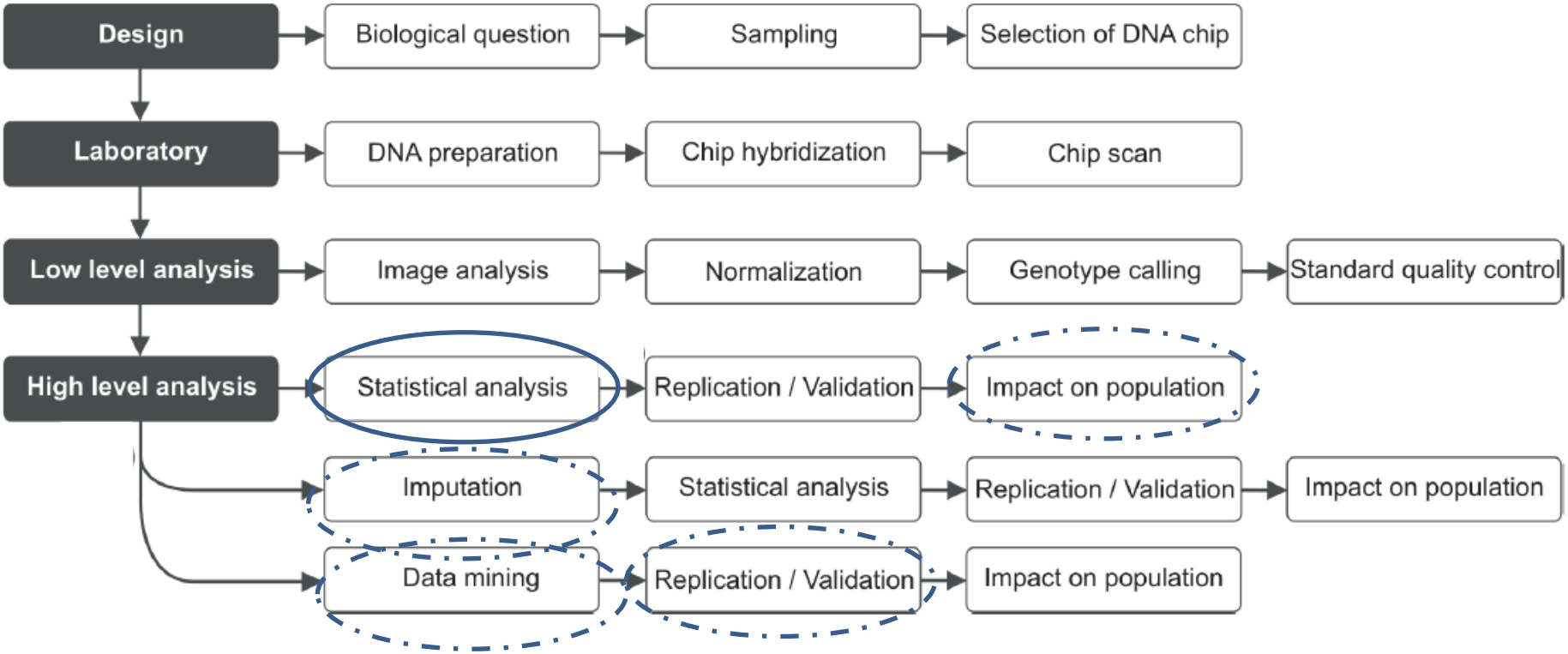
2317 studies (6/10/2014)

(Entries 1-50 of 2317)

Page 1 of 47 Next > Last >>

Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Description	Replication Sample Description	Region	Reported Gene(s)	Mapped Gene(s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
04/16/14	Chung CM March 03, 2014 *Diabetes Metab Res Rev* Common quantitative trait locus downstream of RETN gene identified by genome-wide association study is associated with risk of type 2 diabetes mellitus in Han Chinese: a Mendelian randomization effect.	Resistin levels	382 Han Chinese ancestry indiviudals	559 Han Chinese ancestry indiviudals	19p13.2	*RETN*	RETN - C19orf59	rs1423096-G		0.78	1×10^{-7}	.322 [0.25-0.40] ug/mL increase	Illumina [NR]	N
10/03/14	Zhang B January 21, 2014 *Int J Cancer* Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians.	Colorectal cancer	1,773 East Asian ancestry cases, 2,642 East Asian ancestry controls	6,902 East Asian ancestry cases, 7,862 East Asian ancestry controls	18q21.1	*SMAD7*	SMAD7	rs7229639-A	intron	0.145	3×10^{-11}	1.22 [1.15-1.29]	Affymetrix & Illumina [1,695,815] (imputed)	N
10/06/14	Xie T January 17, 2014 *Neurobiol Aging* A genome-wide association study combining pathway analysis for typical sporadic	Amyotrophic lateral sclerosis (sporadic)	250 Han Chinese ancestry cases, 250 Han Chinese ancestry controls	NA	View full set of 175 SNPs								Illumina [859,311] (pooled)	N
					NA	*RAB9P1*	NA	kgp22272527-?		NR	8×10^{-11}	NR		
					NA	*MYO18B*	NA	kgp8087771-?		0.2	2×10^{-10}	3.0327 [2.212039-4.157817]		
					12q24.33	*GPR133*	GPR133	rs11061269-?	intron	0.08	8×10^{-10}	3.7761 [2.49-5.74]		
					21q22.3	*TMPRSS2*	TMPRSS2 -	rs9977018-?		0.05	2×10^{-9}	NR		

Detailed flow of a genome-wide association study



(Ziegler 2009)

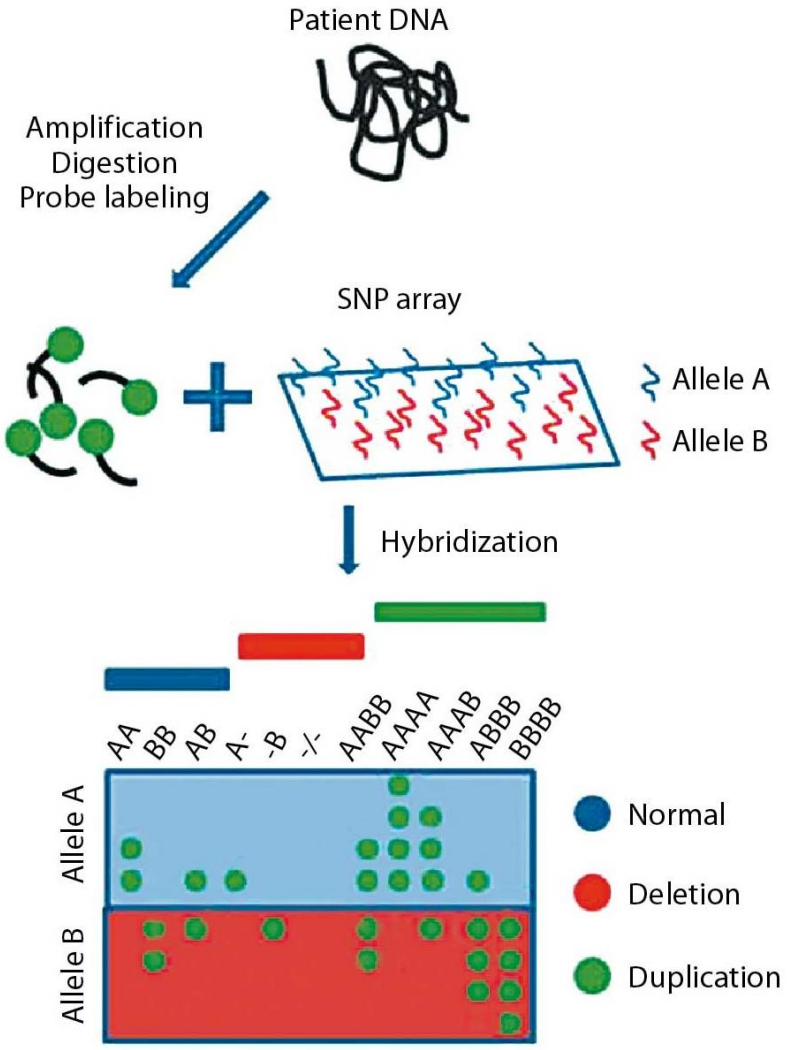
3 Study Design

Components of a study design for GWA studies

- The design of a genetic association study may refer to
  - study scale:
    - Genetic (e.g., hypothesis-drive, panel of candidate genes)
    - Genomic (e.g., hypothesis-free, genome-wide)
  - marker design:
    - Which markers are most informative in GWAs? Common variants-SNPs and/or Rare Variants (MAF<1%)
    - Which platform is the most promising? Least error-prone? Marker-distribution over the genome?
  - subject design

3.a Marker Level

- Costs may play a role, but a balance is needed between costs and chip/sequencing platform performance
- Coverage also plays a role (e.g., exomes only or a uniform spread).
- When choosing **Next Generation Sequencing platforms, also rare variants** can be included in the analysis, in contrast to the older **SNP-arrays** (see right panel).



From common variants towards including rare variants

- Hypothesis 1 for GWAs: Common Disease – Common Variant (CDCV):

  - This hypothesis argues that **genetic variations with appreciable frequency** in the population at large, but **relatively low penetrance** (i.e. the probability that a carrier of the relevant variants will express the disease), are the major contributors to genetic susceptibility to common diseases (Lander, 1996; Chakravarti, 1999; Weiss & Clark, 2002; Becker, 2004).
  - The hypothesis speculates that the gene variation underlying susceptibility to common heritable diseases existed within the **founding population of contemporary humans** → explains the success of GWAs?

From common variants towards including rare variants

- Hypothesis 2 for GWAs: Common Disease – Rare Variant (CDRV):

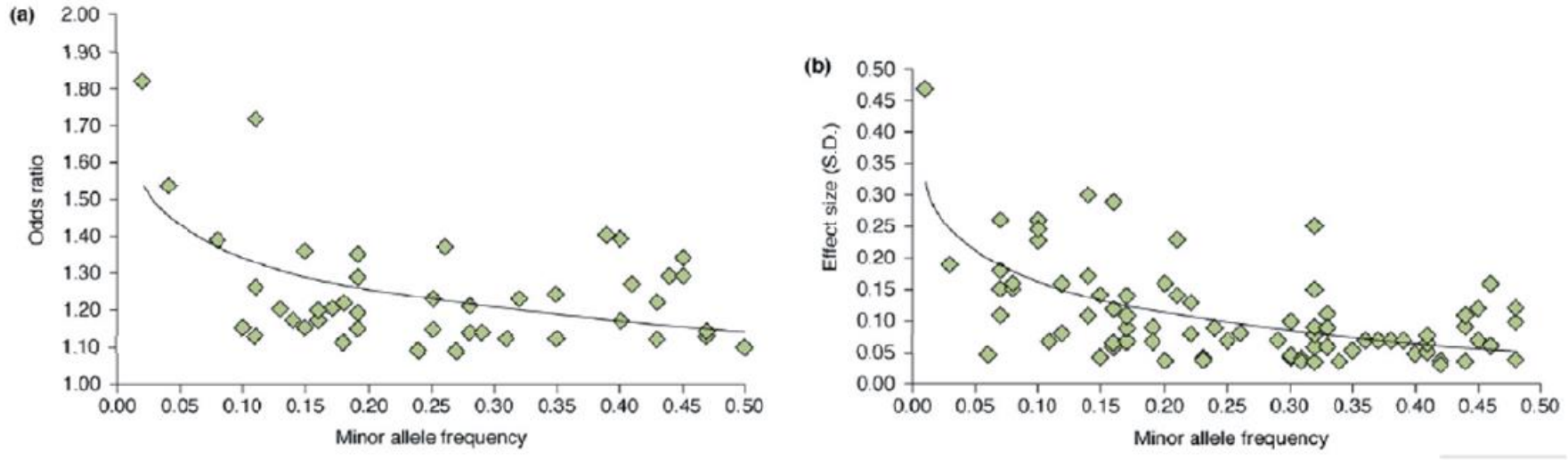
  - This hypothesis argues that **rare DNA sequence variations**, each with **relatively high penetrance**, are the major contributors to genetic susceptibility to common diseases.
  - Some argumentations behind this hypothesis include that by reaching an appreciable frequency for common variations, these variations are not as likely to have been subjected to negative selection. Rare variations, on the other hand, may be **rare because they are being selected against due to their deleterious nature**.

There is room for both hypothesis in current research !
(Schork et al. 2009)

Distribution of SNP "effects"

Dichotomous Traits

Quantitative Traits



Arking & Chakravarti 2009 Trends Genet

Food for thought:

- The higher the MAF, the lower the effect size
- Rare variants analysis is in its infancy in 2009 ….

3.b Subject Level

Aim	Selection scheme
Increased effect size	Extreme sampling: Severely affected cases vs. extremely normal controls
Genes causing early onset	Affected, early onset vs. normal, elderly
Genes with large / moderate effect size	Cases with positive family history vs. controls with negative family history
Specific GxE interaction	Affected vs. normal subjects with heavy environmental exposure
Longevity genes	Elderly survivors serve as cases vs. young serve as controls
Control for covariates with strong effect	Affected with favorable covariates vs. normal with unfavorable covariate

Morton & Collins 1998 Proc Natl Acad Sci USA 95:11389

Popular design 1: cases and controls

Avoiding bias – checking assumptions:

1. Cases and controls drawn from same population
2. Cases representative for all cases in the population
3. All data collected similarly in cases and controls

Advantages:

1. Simple
2. Cheap
3. Large number of cases and controls available
4. Optimal for studying rare diseases

Disadvantages:

1. Population stratification
2. Prone to batch effects and other biases
3. Case definition / severity
4. Overestimation of risk for common diseases

Popular design 2: family-based

Avoiding bias – checking assumptions:

1. Families representative for population of interest
2. Same genetic background in both parents

Advantages:

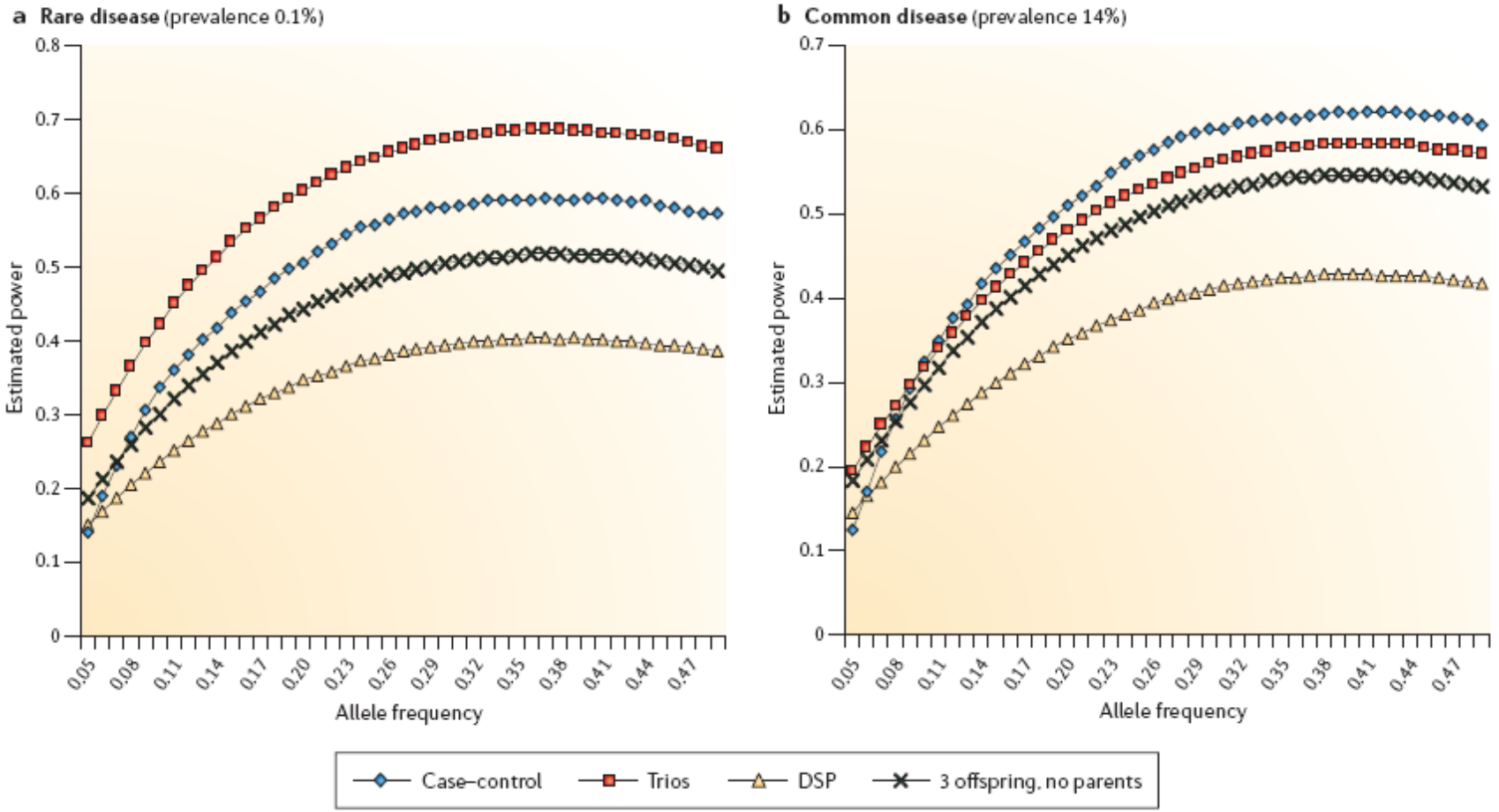
1. Controls immune to population stratification (no association without linkage, no "spurious" (false positive) association)
2. Checks for Mendelian inheritance possible (fewer genotyping errors)
3. Parental phenotyping not required (late onset diseases)
4. Simple logistics for diseases in children
5. Allows investigating imprinting ("bad allele" from father or mother?)

Disadvantages

1. Cost inefficient
2. Sensitive to genotyping errors
3. Lower power when compared with case-control studies

Some more power considerations

- Rare versus common diseases (Lange and Laird 2006)



4 Pre-analysis steps

4.a Quality control

Standard file format for GWA studies

Standard data format: tped = transposed ped format file

FamID	PID	FID	MID	SEX	AFF	$SNP1_1$	$SNP1_2$	$SNP2_1$	$SNP2_2$
1	1	0	0	1	1	A	A	G	T
2	1	0	0	1	1	A	C	T	G
3	1	0	0	1	1	C	C	G	G
4	1	0	0	1	2	A	C	T	T
5	1	0	0	1	2	C	C	G	T
6	1	0	0	1	2	C	C	T	T

ped file

Chr	SNP name	Genetic distance	Chromosomal position
1	SNP1	0	123456
1	SNP2	0	123654

map file

Standard file format for GWA studies (continued)

Chr	SNP	Gen. dist.	Pos	PID 1		PID 2		PID 3		PID 4		PID 5		PID 6	
1	SNP1	0	123456	A	A	A	C	C	C	A	C	C	C	C	C
1	SNP2	0	123654	G	T	G	T	G	G	T	T	G	T	T	T

tped file

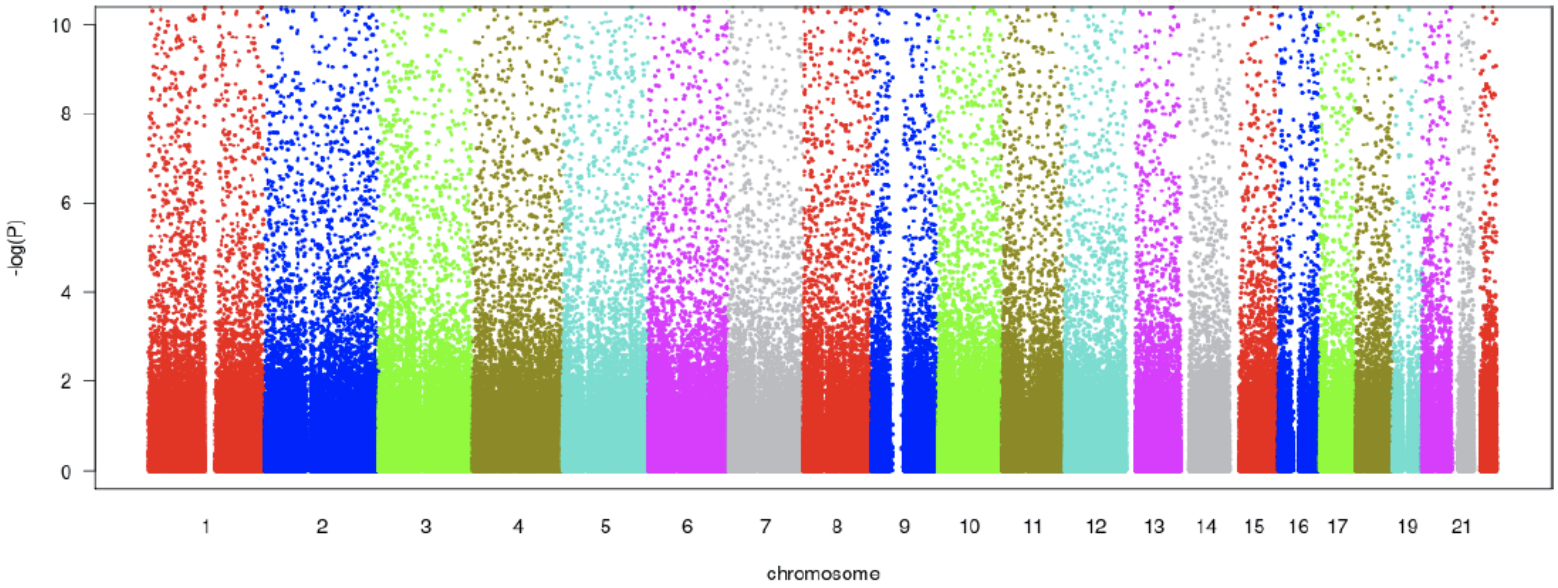
tfam file: First 6 columns of standard ped file

FamID	PID	FID	MID	SEX	AFF
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

tfam file

Why is quality control (QC) important?

BEFORE QC → true signals are lost in false positive signals

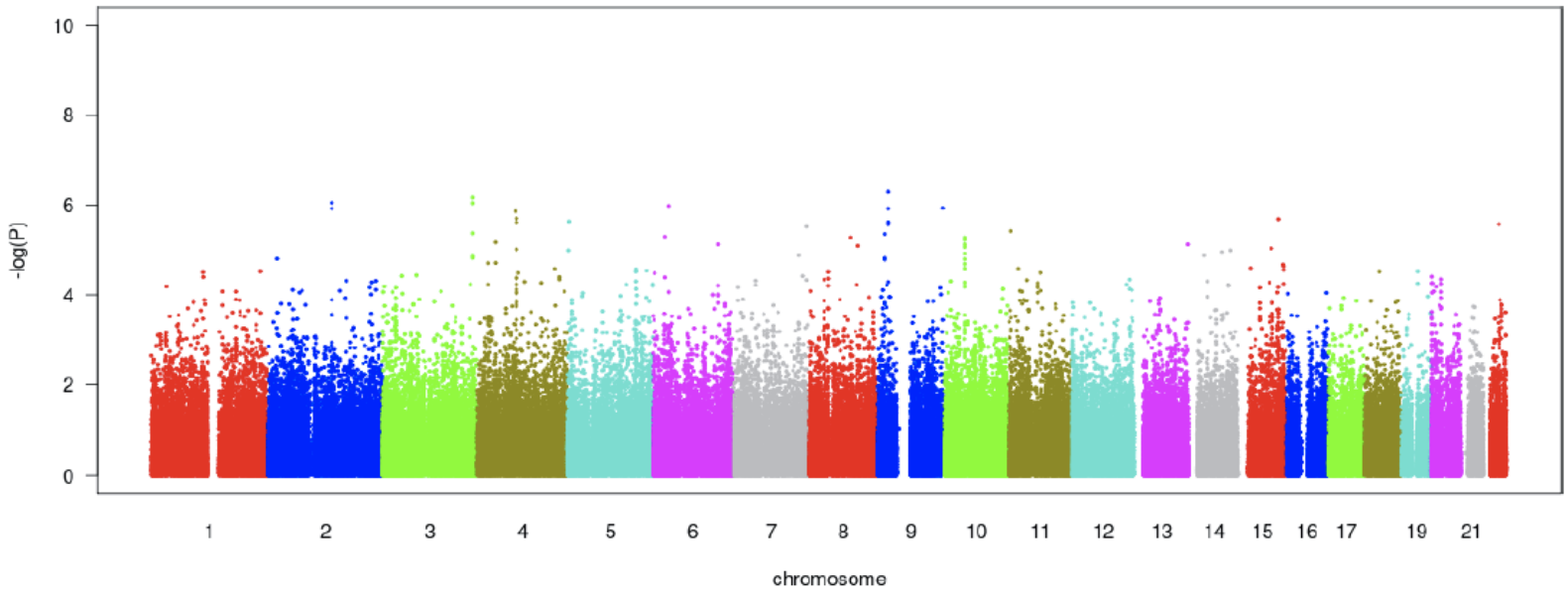


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

(Ziegler and Van Steen 2010)

Why is quality control important?

AFTER QC → skyline of Manhattan (→ name of plot: Manhattan plot):



Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

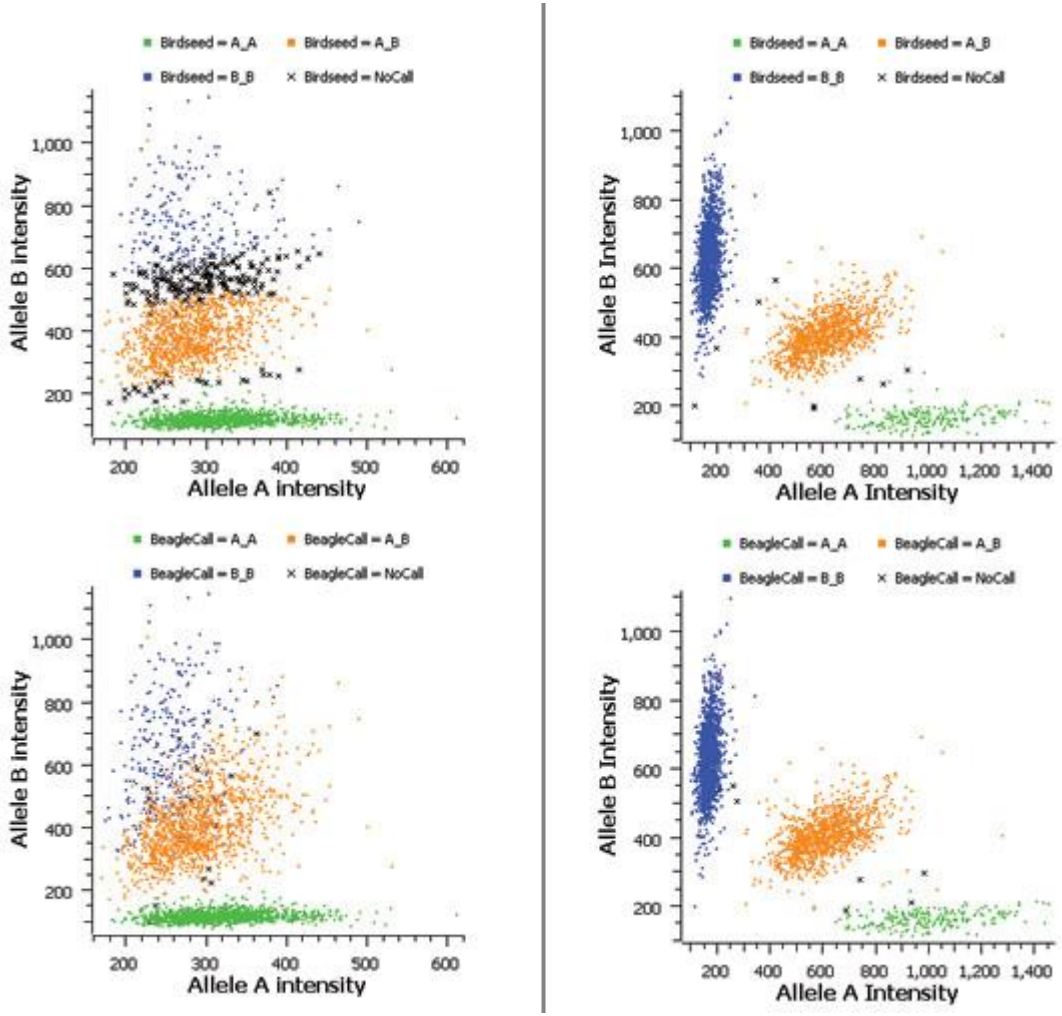
SNPs passing standard quality control: 270,701

(Ziegler and Van Steen 2010)

What is the standard quality control?

- Quality control can be performed on different levels:
  - Subject or sample level
  - Marker level (in this course: SNP level)
  - X-chromosomal SNP level (in this course not considered)
- Consensus on how to best QC data has led to the so-called “Travemünde criteria” (obtained in the town Travemünde) – see later

Marker level QC thresholds may be genotype calling algorithm dependent



Allele signal intensity genotype calling cluster plots for two different SNPs from the same study population.

Upper panels: Birdseed genotypes

Lower panels: BEAGLECALL genotypes.

The plots on the left show a SNP with poor resolution of A_B and B_B genotype clusters and the increased clarity of genotype calls that comes from using BEAGLECALL

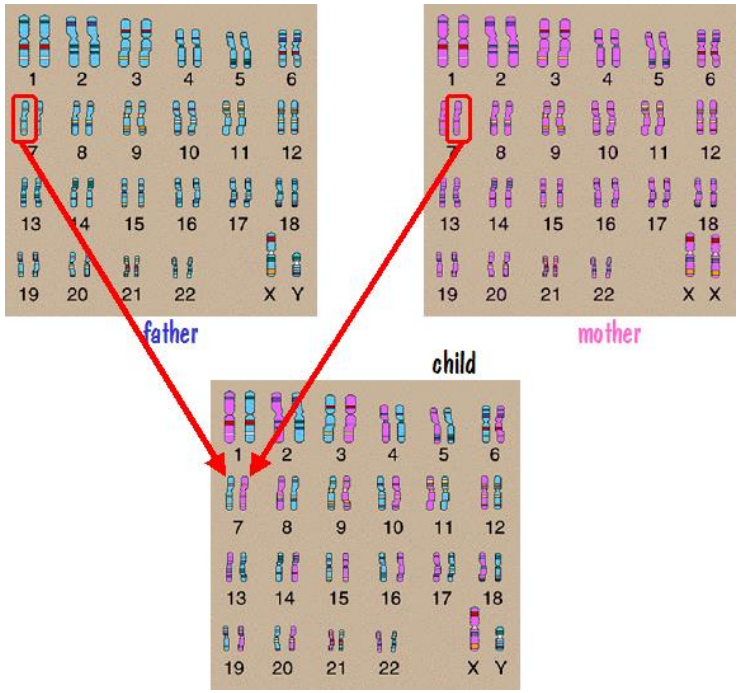
(Golden Helix Blog)

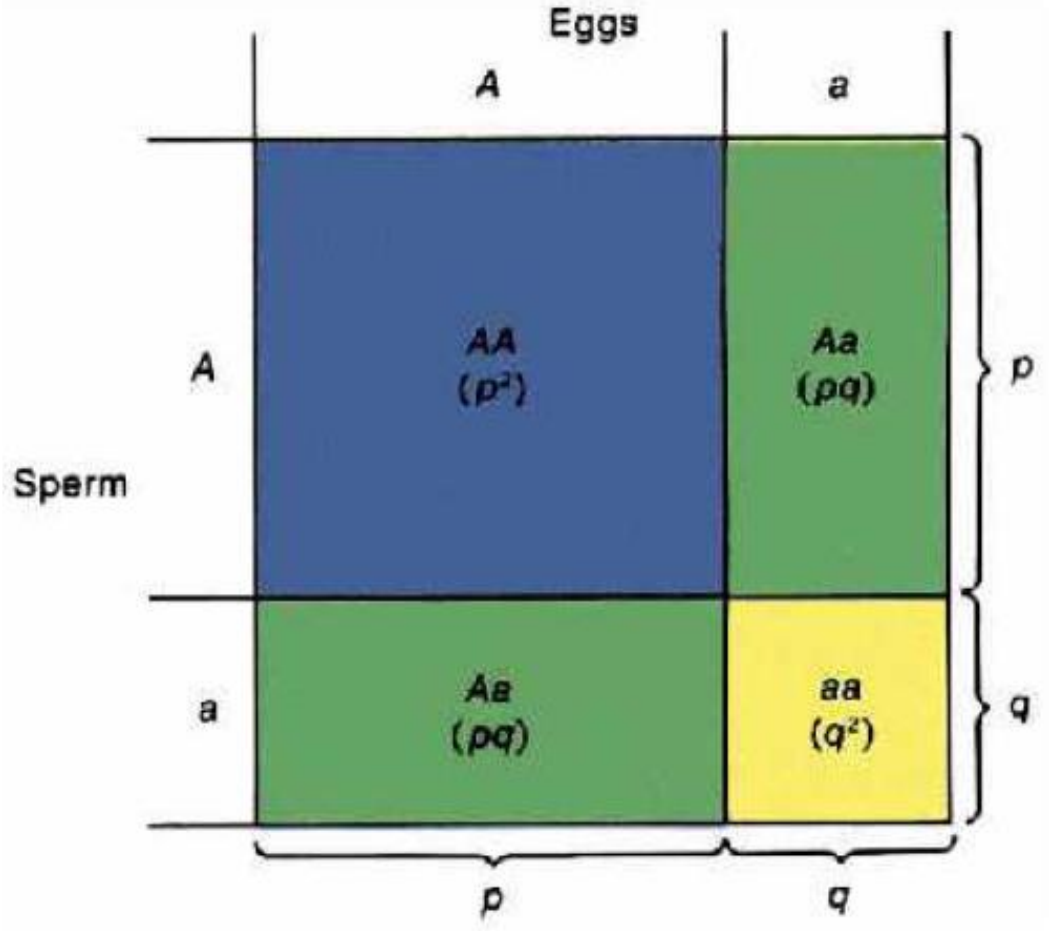
Quality control at the marker level

- **Minor allele frequency (MAF):**
  - Genotype calling algorithms perform poorly for SNPs with low MAF
  - Power is low for detecting associations to genetic markers with low MAF (with standard large-sample statistics)
- **Missing frequency (MiF)**
  - 1 minus call rate
  - MiF needs to be investigated separately in cases and controls because differential missingness may bias association results
- **Hardy-Weinberg equilibrium (HWE)**
  - SNPs excluded if substantially more or fewer subjects heterozygous at a SNP than expected (excess heterozygosity or heterozygote deficiency)

What is Hardy-Weinberg Equilibrium (HWE)?

Consider diallelic SNP with alleles A and a





What is Hardy-Weinberg Equilibrium (HWE)?

Consider diallelic SNP with alleles A_1 and A_2

- Genotype frequencies

  $P(A_1A_1) = p_{11}$, $P(A_1A_2) = p_{12}$, $P(A_2A_2) = p_{22}$
- Allele frequencies $P(A_1) = p = p_{11} + \frac{1}{2}p_{12}$, $P(A_2) = q = p_{22} + \frac{1}{2}p_{12}$

If

- $P(A_1A_1) = p_{11} = p^2$
- $P(A_1A_2) = p_{12} = 2pq$
- $P(A_2A_2) = p_{22} = q^2$

the population is said to be in HWE at the SNP

(Ziegler and Van Steen 2010)

The Travemünde criteria

Level	Filter criterion	Standard value for filter
Sample level	Call fraction	≥ 97%
	Cryptic relatedness	Study specific
	Ethnic origin	Study specific; visual inspection of principal components
	Heterozygosity	Mean ± 3 std.dev. over all samples
	Heterozygosity by gender	Mean ± 3 std.dev. within gender group
SNP level	MAF	≥ 1%
	MiF	≤ 2% in any study group, e.g., in both cases and controls
	MiF by gender	≤ 2% in any gender
	HWE	$p < 10^{-4}$

(Ziegler 2009)

The Travemünde criteria

Level	Filter criterion	Standard value for filter
SNP level	Difference between control groups	$p > 10^{-4}$ in trend test
	Gender differences among controls	$p > 10^{-4}$ in trend test
X-Chr SNPs	Missingness by gender	No standards available
	Proportion of male heterozygote calls	No standards available
	Absolute difference in call fractions for males and females	No standards available
	Gender-specific heterozygosity	No standard value available

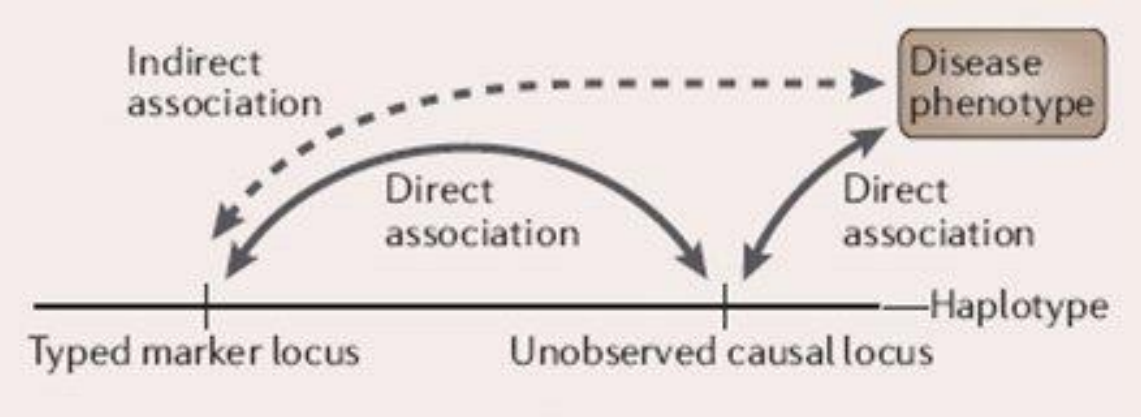
(Ziegler 2009)

4.b Linkage disequilibrium

- **Linkage Disequilibrium** (LD) is a measure of co-segregation of alleles in a population – linkage + allelic association

  Two alleles at different loci that occur together on the same chromosome (or gamete) more often than would be predicted by random chance.

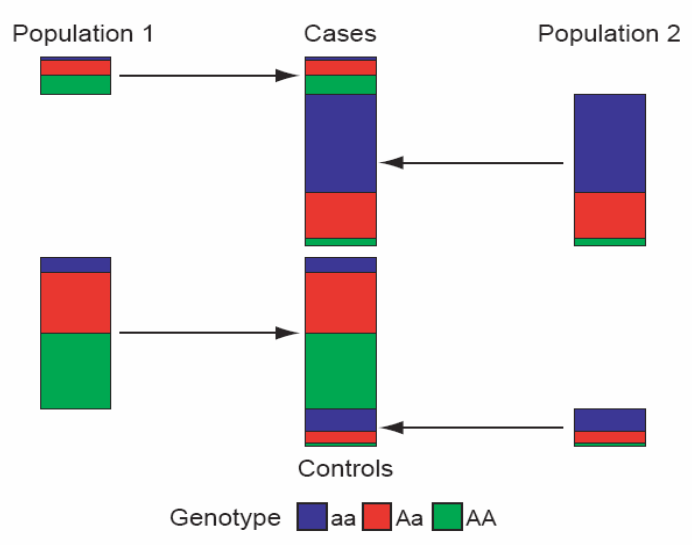
- It is a very important concept for GWAs, since it gives the rational for performing genetic association studies



4.c Confounding by shared genetic ancestry

What is spurious association?

- Typically, there are two characteristics present:
  - A difference in proportion of individual from two (or more) subpopulation in case and controls
  - Subpopulations have different allele frequencies at the locus.



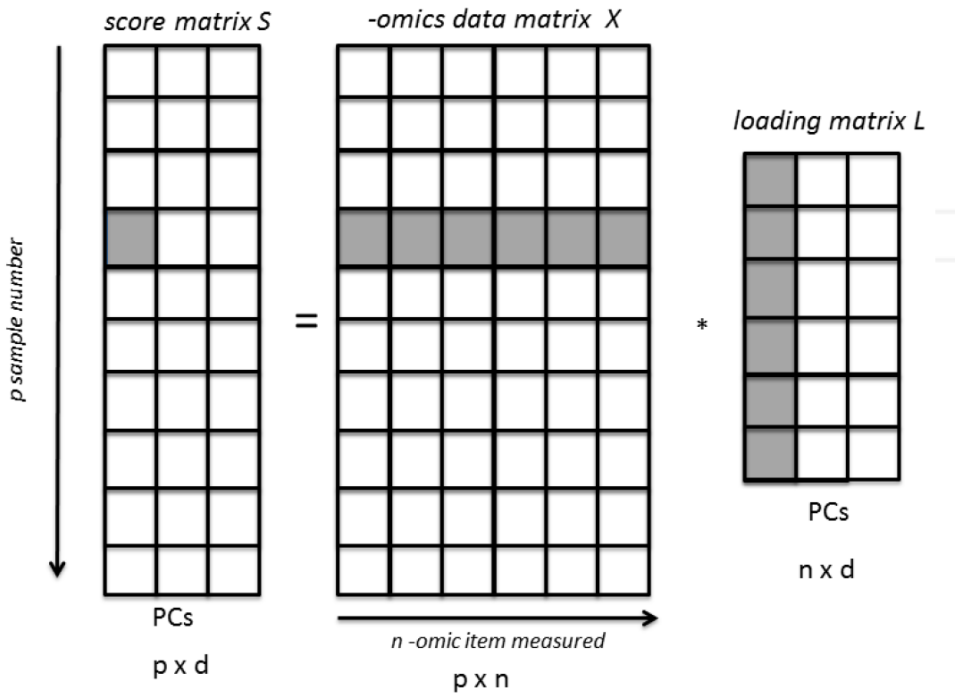
What are typical methods to deal with population stratification?

- Methods to deal with spurious associations generated by population structure generally require a number (at least >100) of widely spaced null SNPs that have been genotyped in cases and controls in addition to the candidate SNPs.

- These methods large group into:

  - **Principal components**: finding continuous axes of genetic variation
  - **Structured association methods**: "First look for structure (population clusters) and **second** perform an association **analysis** conditional on the cluster allocation"
  - **Genomic control methods**: "**First analyze** and second downplay association test results for over optimism"

Principal components



$$s_{4,1} = x_{4,1}.l_{1,1} + x_{4,2}.l_{2,1} + x_{4,3}.l_{3,1} + x_{4,4}.l_{4,1} + x_{4,5}.l_{5,1} + x_{4,6}.l_{6,1}$$

- Mathematical derivation: https://courses.cs.ut.ee/MTAT.03.227/2017_spring/uploads/Main/lecture-notes-9.pdf
- Applications in omics: http://cdn.intechopen.com/pdfs-wm/30002.pdf

Principal components

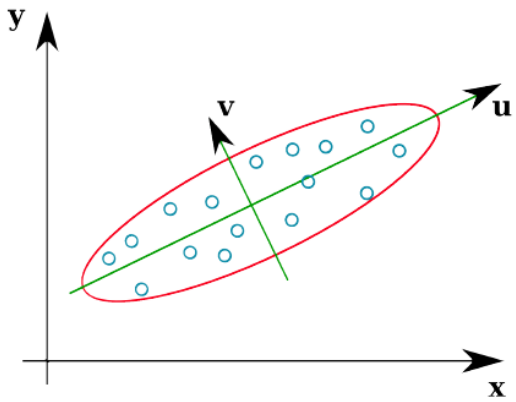


Figure 1: PCA for Data Representation

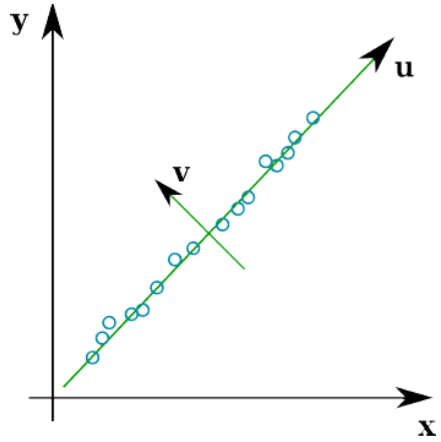


Figure 2: PCA for Dimension Reduction

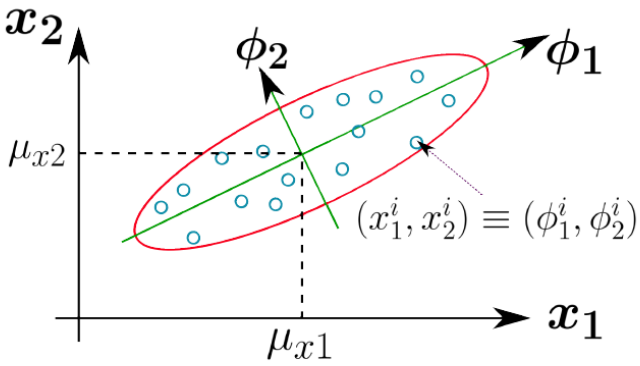
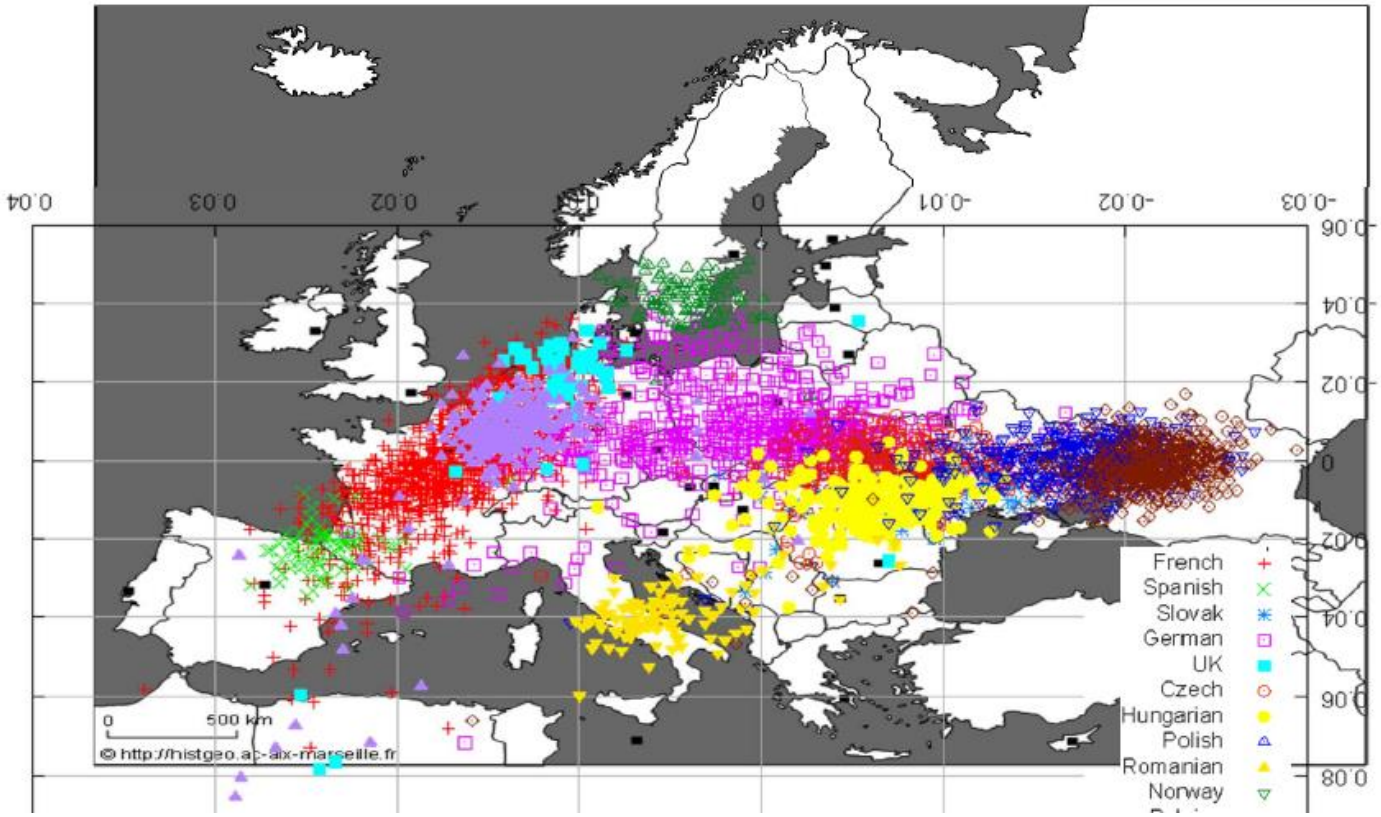


Figure 3: The PCA Transformation

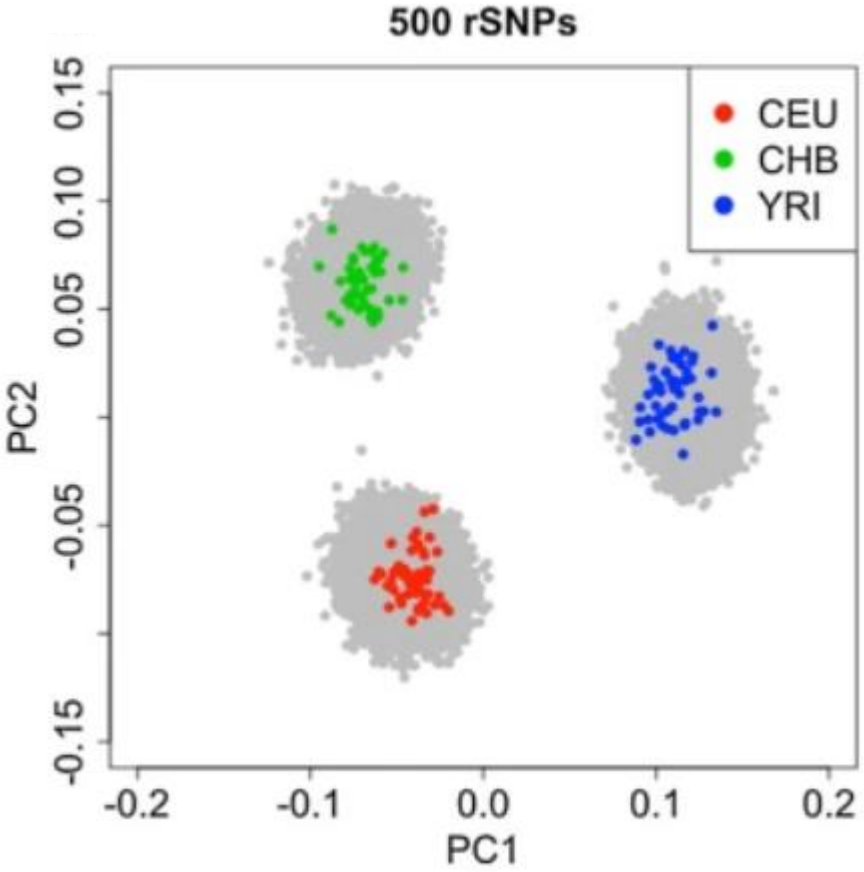
- Find eigenvectors of the covariance matrix for standardized (x1, x2, …) [→SNPs]
- These will give you the direction vectors indicated in Fig3 by phi_1 and phi_2
- These determine the axes of maximal variation

Principal components in statistical genetics

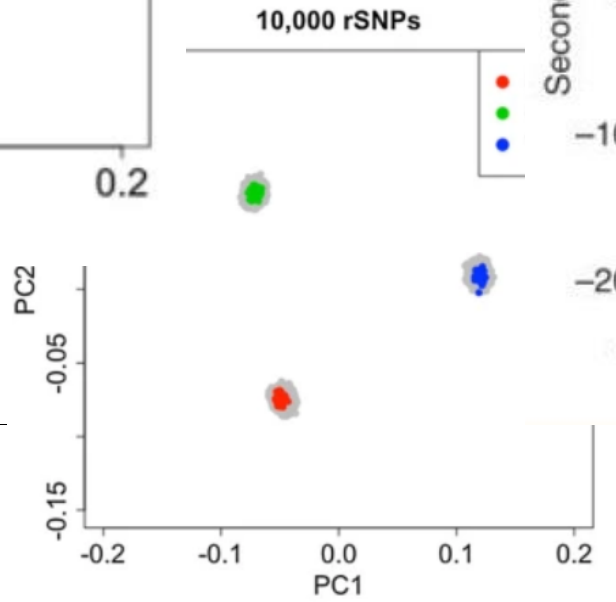
In European data, the first 2 principal components “nicely” reflect the N-S and E-W axes ! Y-axis: PC2 (6% of variance); X-axis: PC1 (26% of variance)

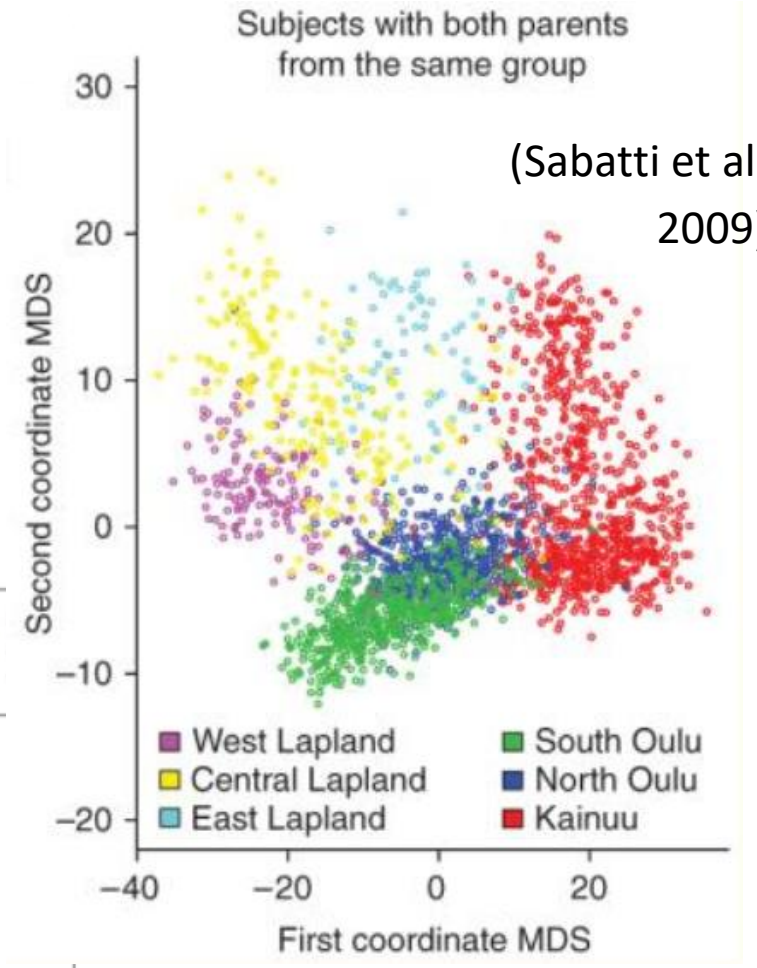


Principal components in statistical genetics: the more SNPs the better?



(Pardo-Seco et al. 2014)





(Sabatti et al. 2009)

Van Steen K

5 Analysis Steps

5.a Testing for Associations

The linear regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

- y: response variable.
- $x_1, \ldots, x_k$: regressor variables, independent variables.
- $\beta_0, \beta_1, \ldots, \beta_k$: regression coefficients.
- ϵ: model error.
  - Uncorrelated: $\text{cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$.
  - Mean zero, Same variance: $\text{var}(\epsilon_i) = \sigma^2$. (homoscedasticity)
  - Normally distributed.

Linear vs non-linear

Linear Models Examples:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$
$$y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \epsilon$$
$$\log y = \beta_0 + \beta_1 \left(\frac{1}{x_1}\right) + \beta_2 \left(\frac{1}{x_2}\right) + \epsilon$$

Nonlinear Models Examples:

$$y = \beta_0 + \beta_1 x_1^{\gamma 1} + \beta_2 x_2^{\gamma 2} + \epsilon$$
$$y = \frac{\beta_0}{1 + e^{\beta_1 x_1}} + \epsilon$$

Regression inference

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

- Least square estimation of the regression coefficients.
  $b = (X^T X)^{-1} X^T y.$
- Variance estimation for σ^2 (see later)
- Coefficient of Determination. R^2.
- Partial F test or t-test for $H_0 : \beta_j = 0$.

Tests in GWAS using the regression framework

- **Example 1:**

$$Y = \beta_0 + \beta_1 SNP + \varepsilon$$

  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 2$ (this links to df in variance estimation)
  - $\mathrm{df_R} = \mathrm{n} - 1$ (this links to df in variance estimation)

It can be shown that for testing $\beta_1 = 0$ versus $\beta_1 \neq 0$

- $F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} : \frac{SSE(F)}{df_F} = \frac{b_1^2}{s^2(b_1)} = (t^*)^2$

Why is the t-test more flexible?

Tests in GWAS using the regression framework

- **Example 2:**

$$Y = \beta_0 + \beta_1 SNP + \beta_2 PC_1 + \beta_3 PC_2 + \varepsilon$$

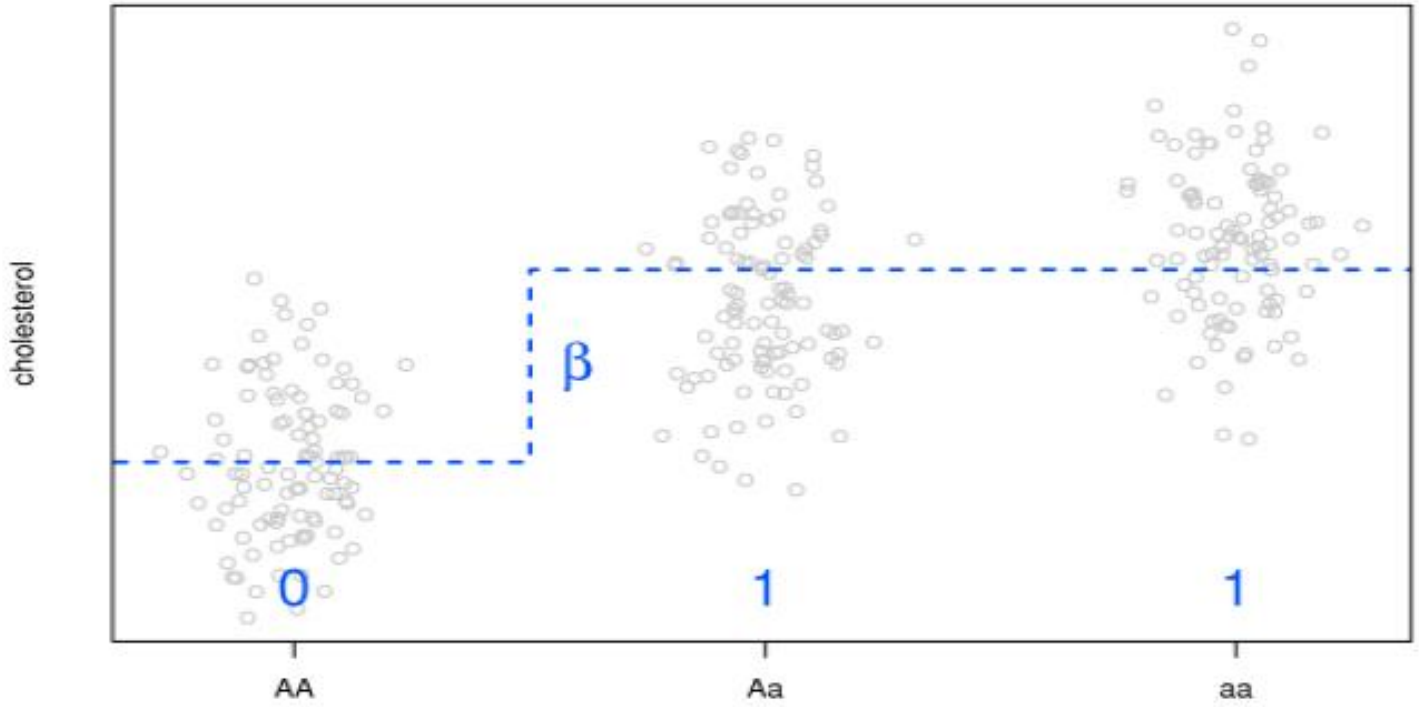
  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - $\mathrm{df_F} = \mathrm{n} - 4$
  - $\mathrm{df_R} = \mathrm{n} - 3$

How many dfs would the corresponding F-test have?

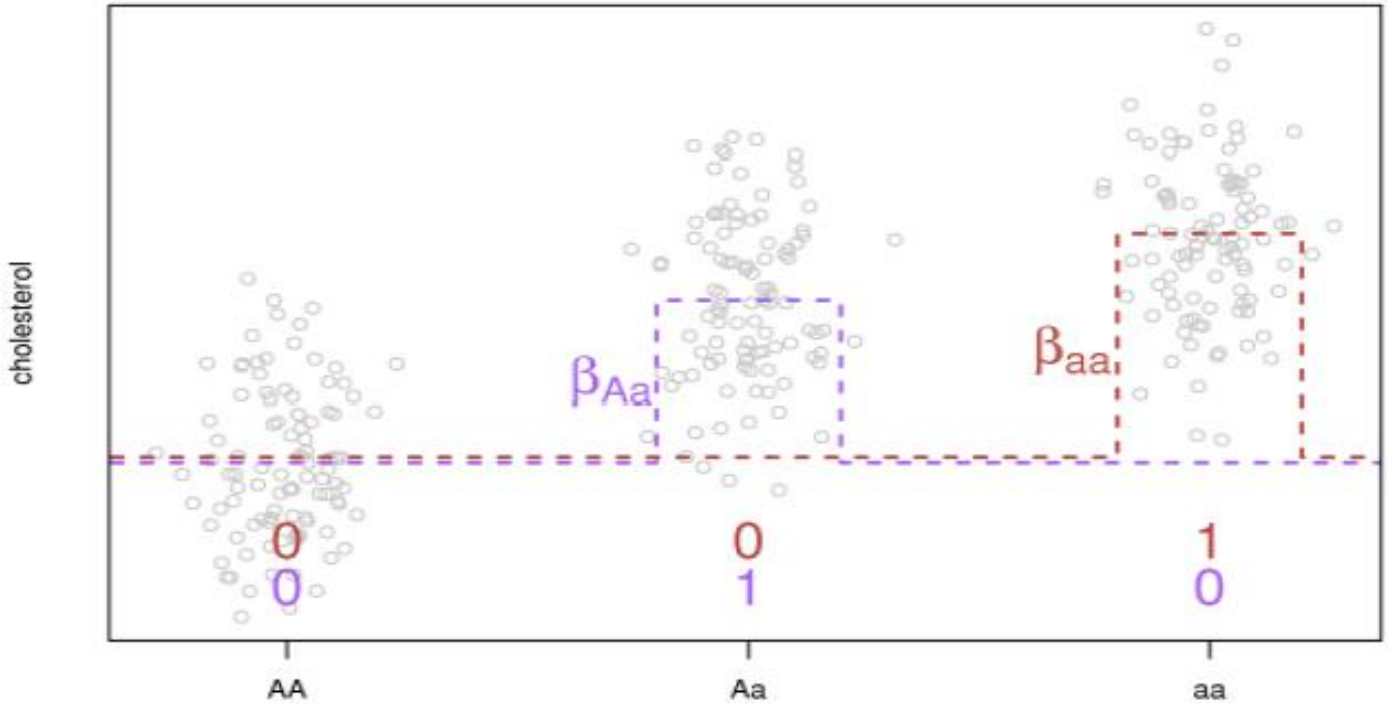
The impact of different encoding schemes for SNPs

	Coding scheme for statistical modeling/testing					
Indiv. genotype	**X1**	**X1**	**X2**	**X1**	**X1**	**X1**
	Additive coding	**Genotype coding (general mode of inheritance)**		**Dominant coding (for a)**	**Recessive coding (for a)**	**Advantage Heterozygous**
AA	**0**	**0**	**0**	**0**	**0**	**0**
Aa	**1**	**1**	**0**	**1**	**0**	**1**
aa	**2**	**0**	**1**	**1**	**1**	**0**

Which encoding scheme provides a good fit to the data?

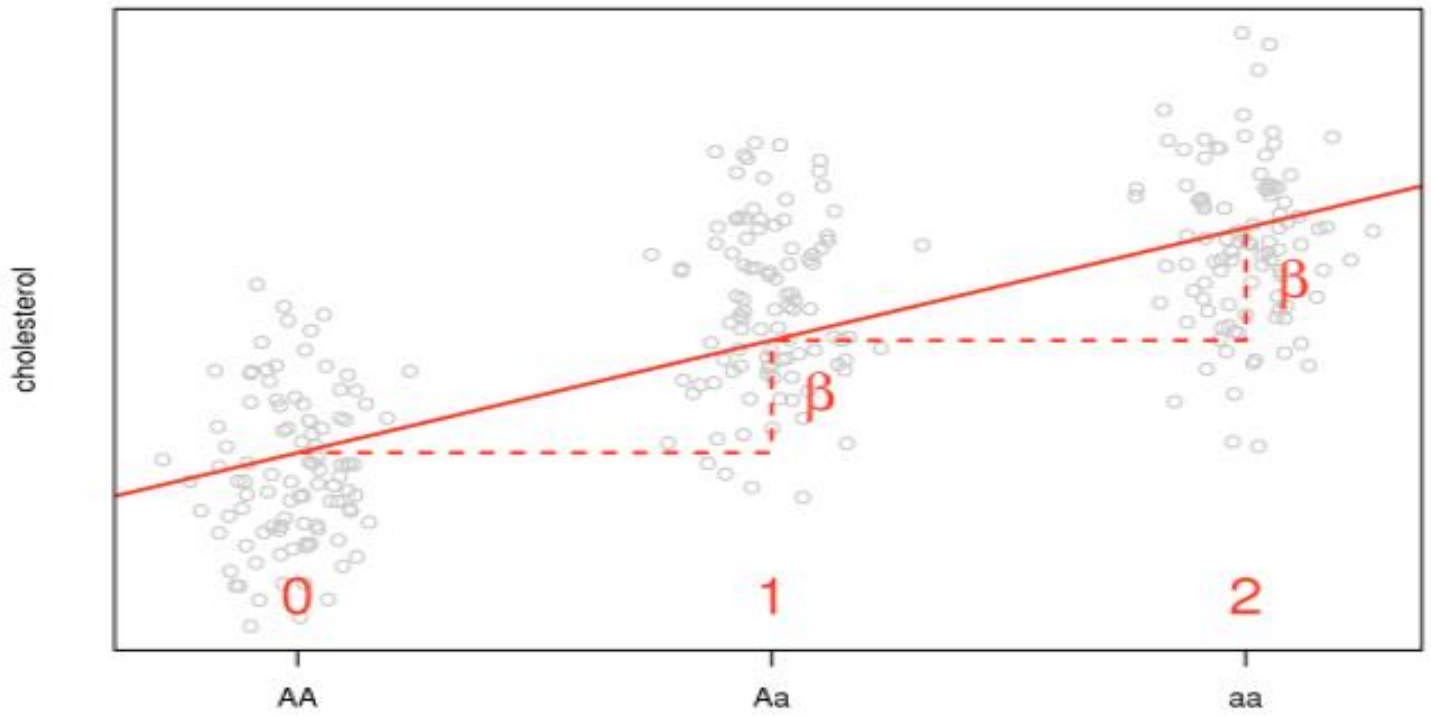


Which encoding scheme provides a good fit to the data?



Robust vs overkill ?

Which encoding scheme provides a good fit to the data?



Most commonly used

Regression analysis in R

Syntax	Model	Comments
Y ~ A	$Y = \beta_o + \beta_1 A$	Straight-line with an implicit y-intercept
Y ~ -1 + A	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
Y ~ A + I(A^2)	$Y = \beta_o + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function **I( )** allows terms in the model to include normal mathematical symbols.
Y ~ A + B	$Y = \beta_o + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
Y ~ A:B	$Y = \beta_o + \beta_1 AB$	A model containing only first-order interactions between A and B.
Y ~ A*B	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is Y ~ A + B + A:B.
Y ~ (A + B + C)^2	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 AC$	A model including all first-order effects and interactions up to the n^{th} order, where n is given by **( )^n**. An equivalent code in this case is Y ~ A*B*C – A:B:C.

Model diagnostics are model-dependent …

- There are 4 principal assumptions which justify the use of **linear regression** models for purposes of prediction:
  - **linearity** of the relationship between dependent and independent variables
  - independence of the errors (no serial correlation)
  - homoscedasticity (constant variance) of the errors
    - versus time (when time matters)
    - versus the predictions (or versus any independent variable)
  - normality of the error distribution. (http://www.duke.edu/~rnau/testing.htm)

- To check **model assumptions**: go to **quick-R** and regression diagnostics (http://www.statmethods.net/stats/rdiagnostics.html)

QQ plots for model diagnostics – Q for Quantile

- Quantiles are points in your data below which a certain proportion of your data fall.

  What is the 0.5 quantile for normally distributed data?

- Here we generate a random sample of size 200 from a normal distribution and find the quantiles for 0.01 to 0.99 using the quantile function:

```
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))
```

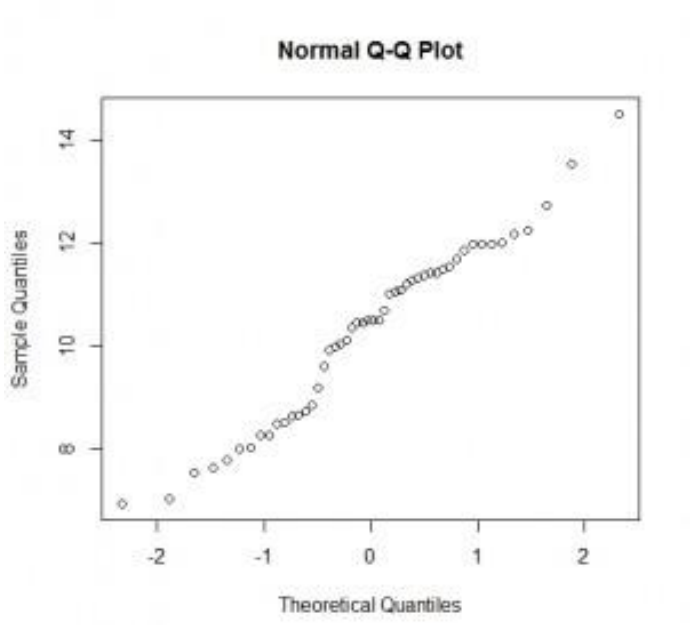
- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
  The number of quantiles is selected to match the size of your sample data.

  The quantile function in R offers 9 different quantile algorithms!

  See help(quantile)

QQ plots for model diagnostics – Q for Quantile

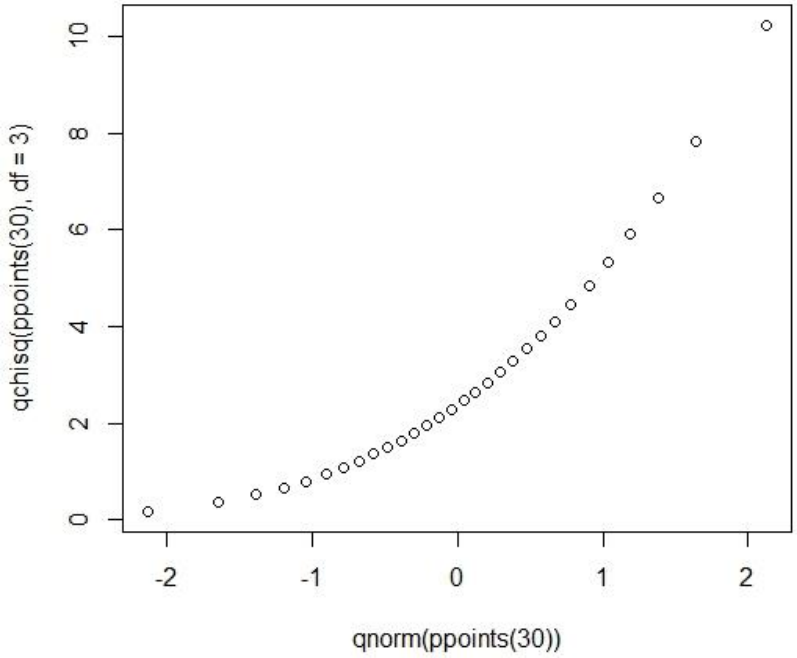
- A Q-Q plot is a scatterplot created by plotting **two sets of quantiles** against one another.
- If both sets of quantiles come from the same distribution, we should see the points forming a line that's roughly straight.
- Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Examples of QQ plots: no straight line

- QQ plot of a distribution that's skewed right; a Chi-square distribution with 3 degrees of freedom against a Normal distribution

qqplot(qnorm(ppoints(30)), qchisq(ppoints(30),df=3))



Testing for association between case/control status and a SNP

- Fill in the table below and perform a chi-squared test for independence between rows and columns → **genotype test → 2 df**

	AA	**Aa**	**aa**
Cases			
Controls			

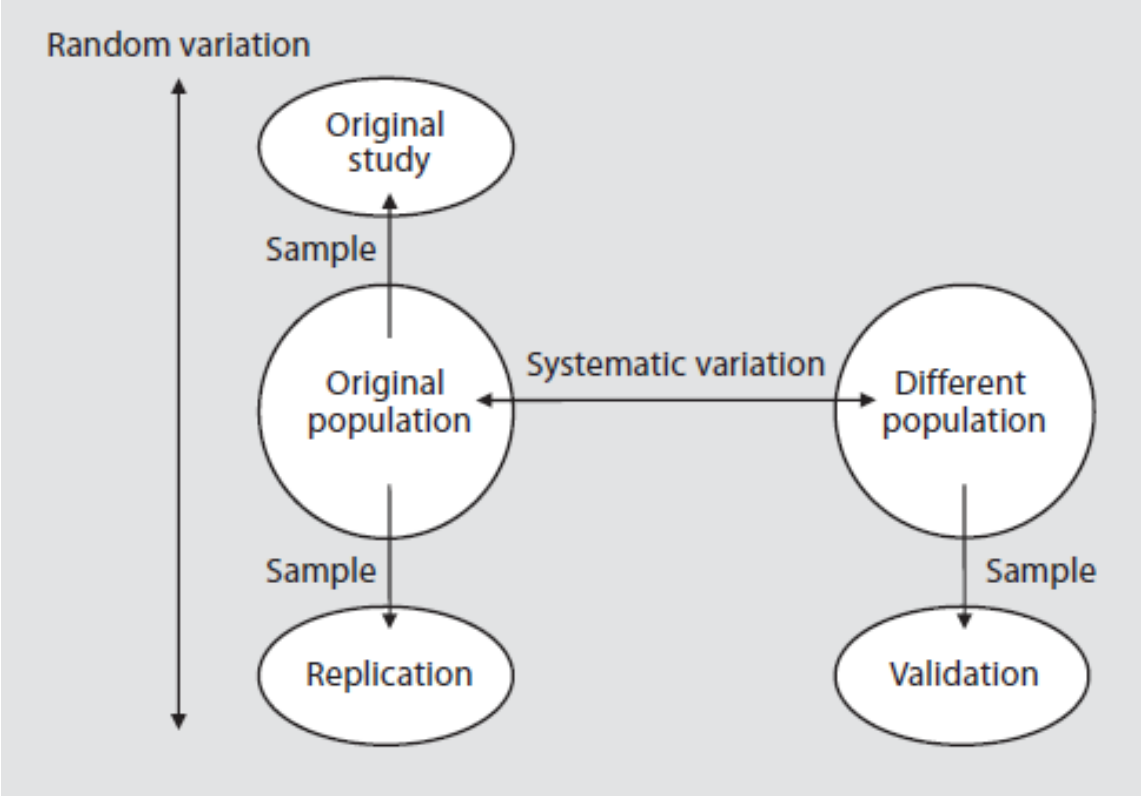
Sum of entries = cases+controls

- Fill in the table below and perform a chi-squared test for independence between rows and columns → **allelic test (ONLY valid under HWE) → 1df**

	A	**a**
Cases		
Controls		

Sum of entries is
2 x (cases + controls)

5.b Replication and validation



(Igl et al. 2009)

Guidelines for replication studies

- Replication studies should be of sufficient size to demonstrate the effect
- Replication studies should conducted in independent datasets
- Replication should involve the same phenotype
- Replication should be conducted in a similar population
- The same SNP should be tested
- The replicated signal should be in the same direction
- Joint analysis should lead to a lower *p*-value than the original report
- Well-designed negative studies are valuable

Note that SNPs are most likely to replicate when they
- show modest to strong statistical significance,
- have common minor allele frequency,
- exhibit modest to strong **genetic effect size** (~strength of association)

5.c Causation

"Association does not imply causation"

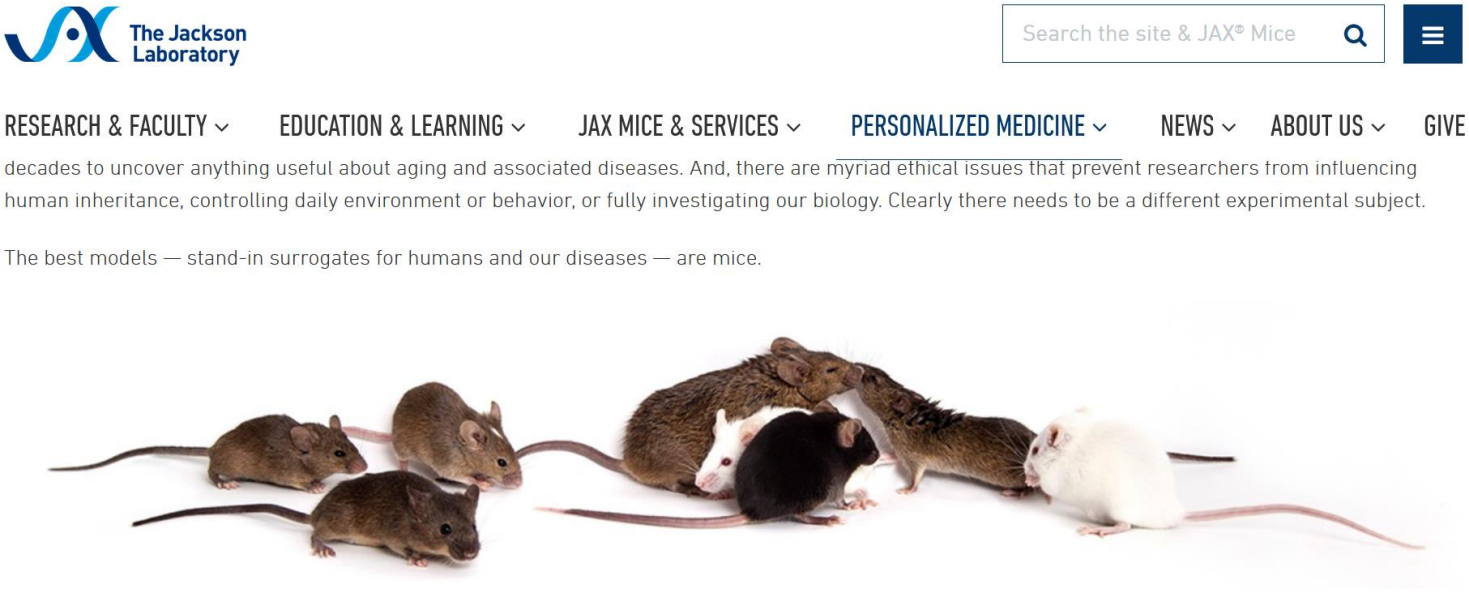
- Meaning:

  Just because two things correlate does not necessarily mean that one causes the other.

- As a seasonal example, just because people in Belgium tend to spend more in the shops when it's cold and less when it's hot doesn't mean cold weather causes high street spending.

Establishing causation: wet lab experiments in model organisms

- Gene knock-out experiments

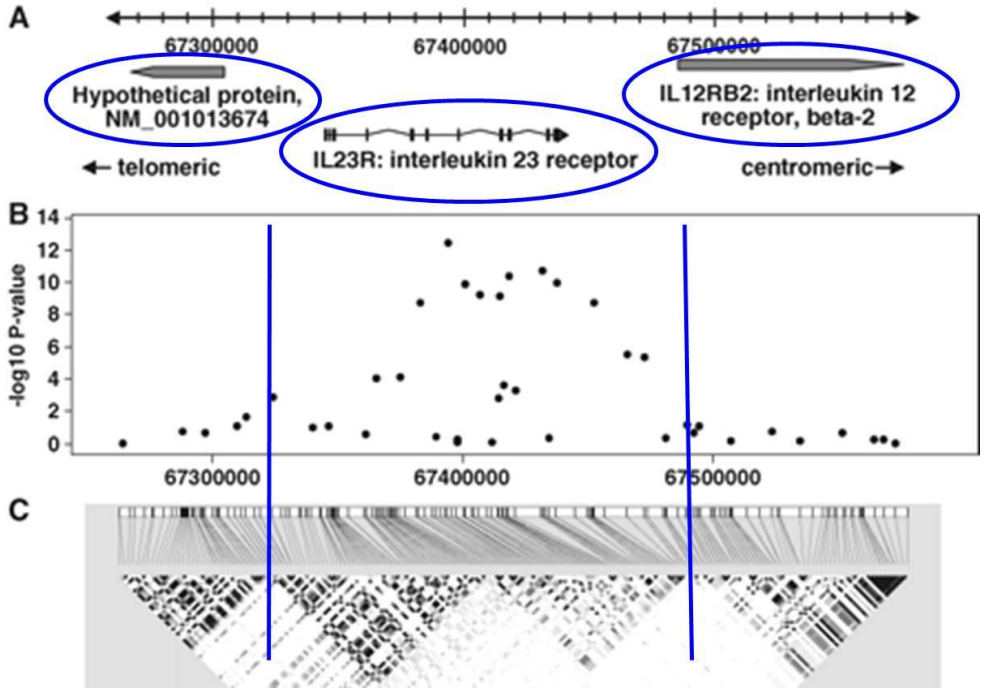


(https://www.jax.org/about-us/why-mice)

- The findings of animal experiments may not always be directly applicable to the human situation because of genetic, anatomic, and physiologic differences or the entity of exposures a human being has experienced

Establishing causation: dry lab

- As opposed to association studies that benefit from LD, the main challenge in identifying causal variants at associated loci analytically (**finemapping**) lies in distinguishing among the many closely correlated variants due to LD



(Duerr et al 2006)

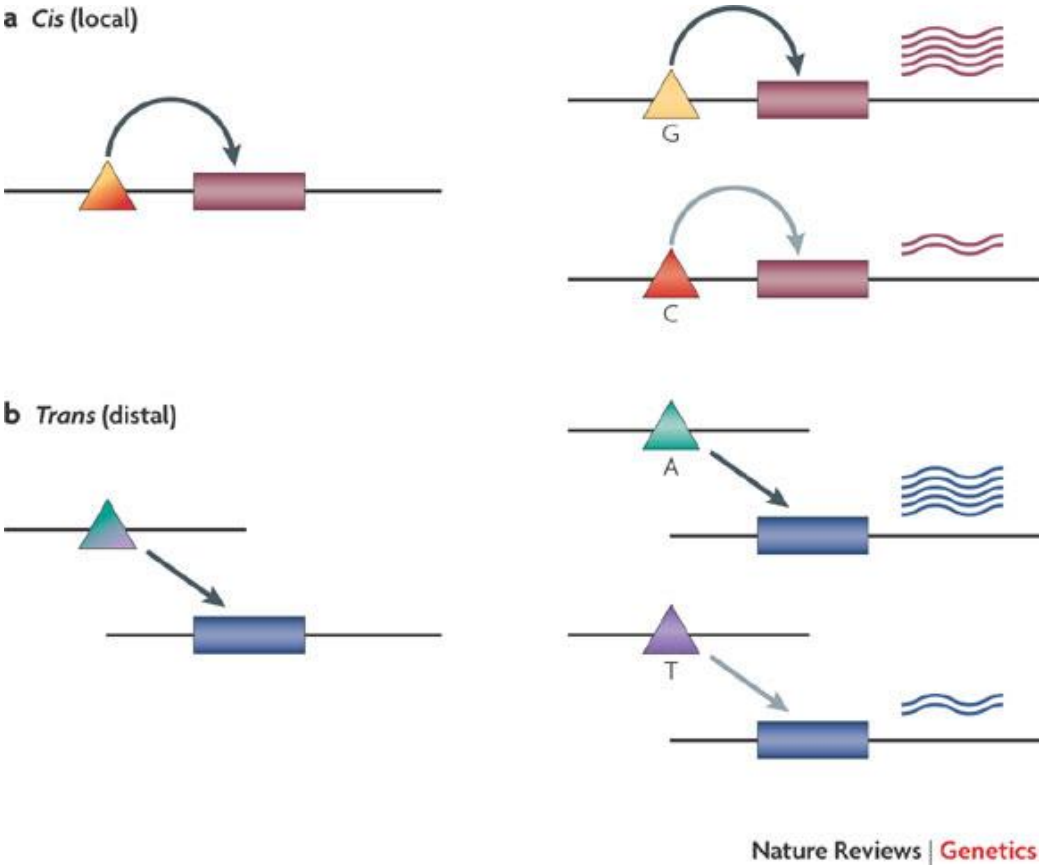
5. d Interpretation

Functional genomics analyses: incl transcriptomics

- One of the fundamental needs for the interpretation of the effects of genome variants is the understanding of the specific biological effect such variants have in the cell, which provides a handle to the biology of the disease or organismal phenotype.
- GWAS have demonstrated that the majority of such variants are found in non-coding regions of the genome and are therefore likely to be involved in gene regulation. Hence, there should be interpretational advantages in analyzing these variants in the context of gene expression (in cells/tissues)
- **An eQTL** is a locus that explains a fraction of the genetic variance of a gene expression phenotype.

(Nika and Dermitzakis 2013)

Functional genomics analyses: incl transcriptomics



(Cheung and Spielman 2009)

- Cis-acting variants are found close to the target genes and trans-acting variants are located far from the target genes, often on another chromosome.
- Different allelic forms of the cis- and trans-acting variants have different influence on gene expression.

Functional genomics analyses: incl transcriptomics

DOI: 10.1038/s41467-017-01261-5 OPEN

Functional mapping and annotation of genetic associations with FUMA

Kyoko Watanabe[1], Erdogan Taskesen[1,2], Arjen van Bochoven[3] & Danielle Posthuma[1,4]

"Colocalization analysis" (not to be confused with protein colocalization)

- Estimates the posterior probability that the same variant is causal in both a GWAS and eQTL study while accounting for the uncertainty of LD
- Example statistical methods following a Bayesian statistical framework: eCAVIAR (Hormozdiari et al. 2016), COLOC (Giambartolomei et al. 2014)
- Posterior support for the following hypotheses:

  H0: no causal variants for either trait;

  H1: a causal variant for disease association (GWAS) only;

  H2: a causal variant for gene expression association (eQTL) only;

  H3: two distinct causal variants, one for each trait;

  H4: a single causal variant common to both traits (co-localization).

Changing units of analysis: from SNPs to genes

European Journal of Human Genetics (2019) 27:811–823
https://doi.org/10.1038/s41431-018-0327-8

ESHG

ARTICLE

Comparison of methods for multivariate gene-based association tests for complex diseases using common variants

Jaeyoon Chung[1,2] · Gyungah R. Jun[2,3,4] · Josée Dupuis[4] · Lindsay A. Farrer[1,2,4,5,6,7]

Received: 13 December 2017 / Revised: 30 October 2018 / Accepted: 4 December 2018 / Published online: 25 January 2019

Abstract

Complex diseases are usually associated with multiple correlated phenotypes, and the analysis of composite scores or disease status may not fully capture the complexity (or multidimensionality). Joint analysis of multiple disease-related phenotypes in genetic tests could potentially increase power to detect association of a disease with common SNPs (or genes). Gene-based tests are designed to identify genes containing multiple risk variants that individually are weakly associated with a univariate trait. We combined three multivariate association tests (O'Brien method, TATES, and MultiPhen) with two gene-based association tests (GATES and VEGAS) and compared performance (type I error and power) of six multivariate gene-based methods using simulated data. Data ($n = 2000$) for genetic sequence and correlated phenotypes were simulated by varying causal variant proportions and phenotype correlations for various scenarios. These simulations showed that two multivariate association tests (TATES and MultiPhen, but not O'Brien) paired with VEGAS have inflated type I error in all scenarios, while the three multivariate association tests paired with GATES have correct type I error. MultiPhen paired with GATES has higher power than competing methods if the correlations among phenotypes are low ($r < 0.57$). We applied these gene-based association methods to a GWAS dataset from the Alzheimer's Disease Genetics Consortium containing three neuropathological traits related to Alzheimer disease (neuritic plaque, neurofibrillary tangles, and cerebral amyloid angiopathy) measured in 3500 autopsied brains. Gene-level significant evidence ($P < 2.7 \times 10^{-6}$) was identified in a region containing three contiguous genes (*TRAPPC12*, *TRAPPC12-AS1*, *ADI1*) using O'Brien and VEGAS. Gene-wide significant associations were not observed in univariate gene-based tests.

Changing units of analysis: from SNPs to (genes to) pathways

- **A biological pathway** is an example of a biosystem, that can consist of *interacting* genes, proteins, and small molecules.
- A biosystem, or biological system, is a group of molecules that interact in a biological system.
- Another type of biosystem is a disease, which can involve components such as genes, biomarkers, and drugs.
- The NCBI BioSystems Database currently contains records from several source databases: KEGG, BioCyc (including its Tier 1 EcoCyc and MetaCyc databases, and its Tier 2 databases), Reactome, the National Cancer Institute's Pathway Interaction Database, WikiPathways, and Gene Ontology (GO).

(https://www.ncbi.nlm.nih.gov/Structure/biosystems/docs/biosystems_about.html)

Questions?