Where are the bad guys?

Genome wide association studies

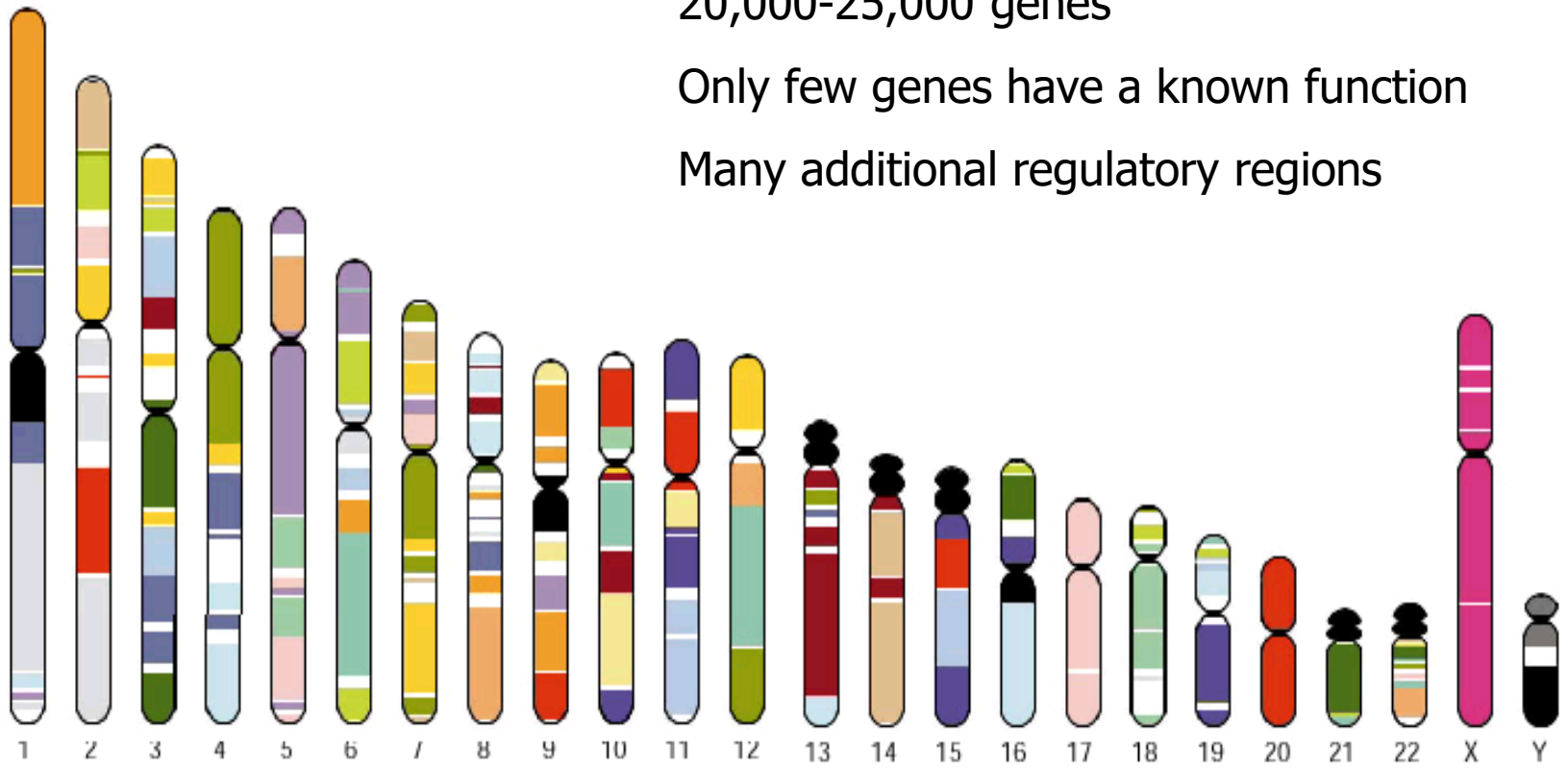
The human genome

~3,300,000,000 base pairs (~2 meter DNA in a single cell)

20,000-25,000 genes

Only few genes have a known function

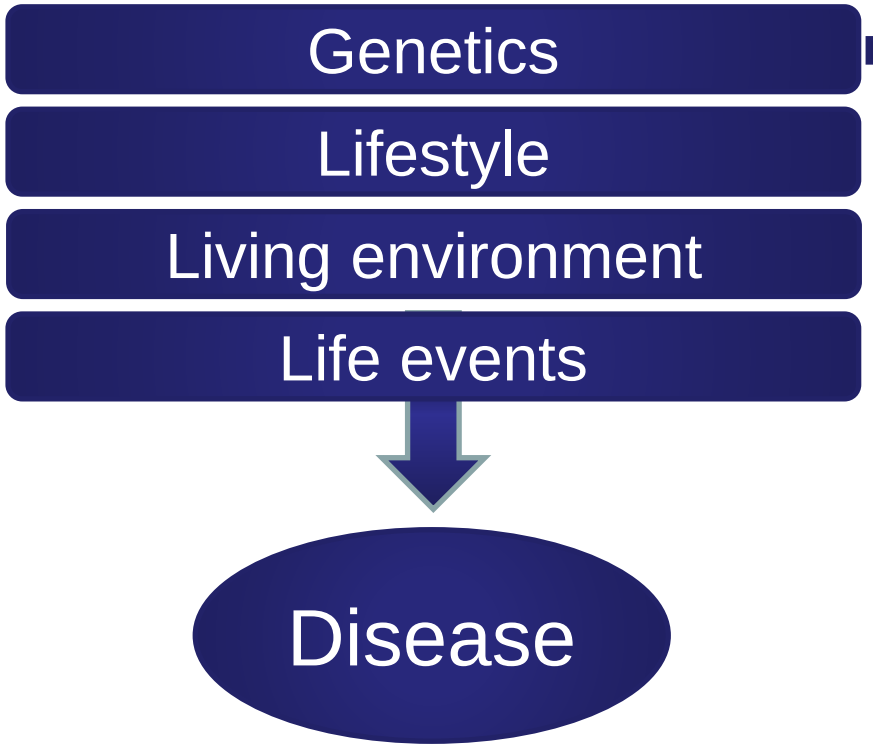
Many additional regulatory regions



Learning goals

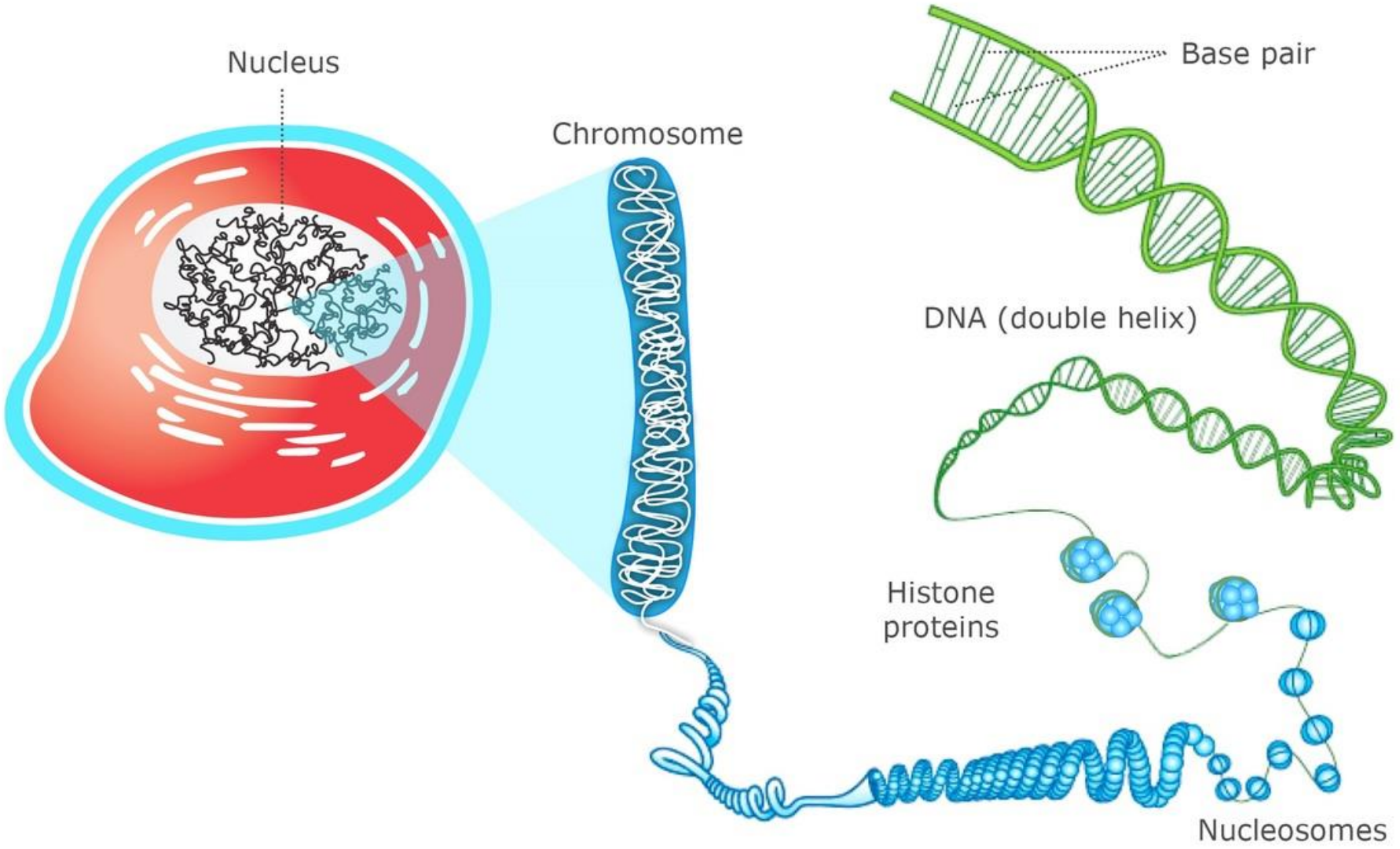
- After my introduction you are able to
  - Explain why genome wide association studies are being performed
  - List the prerequisits of a genome wide association study
  - Design a genome wide association study for an arbitrary trait

Disease mechanisms

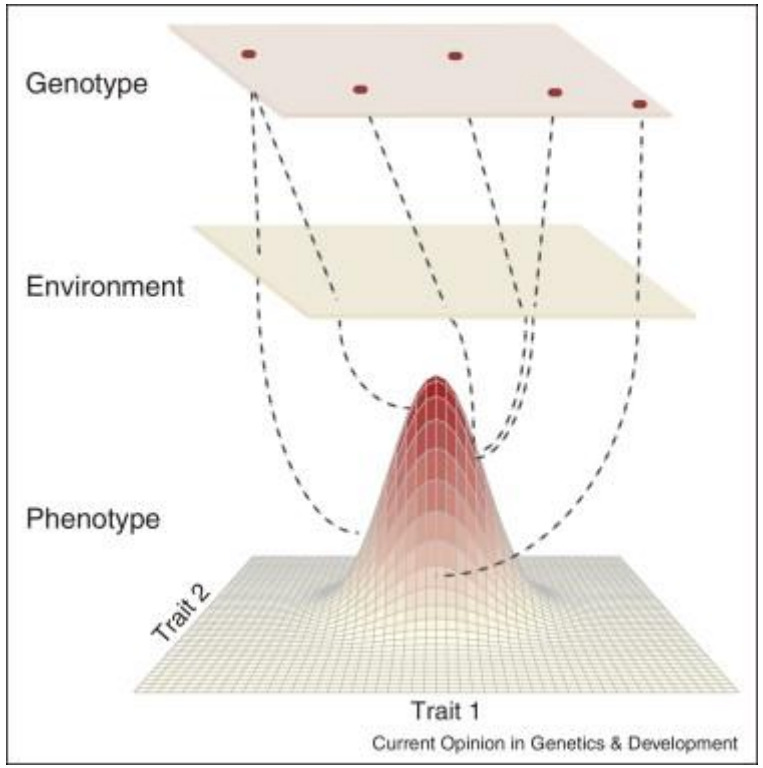


1. Invariable over time
2. Invariable over tissues
3. Almost all traits have a genetic component
4. Potential insight in
   a) Biology and mechanisms
   b) Potential drug targets
   c) Potential stratification of individuals according to lifetime risk
   d) Potential stratification of patients according to the most effective treatment

Desoxyribo Nucleic Acid

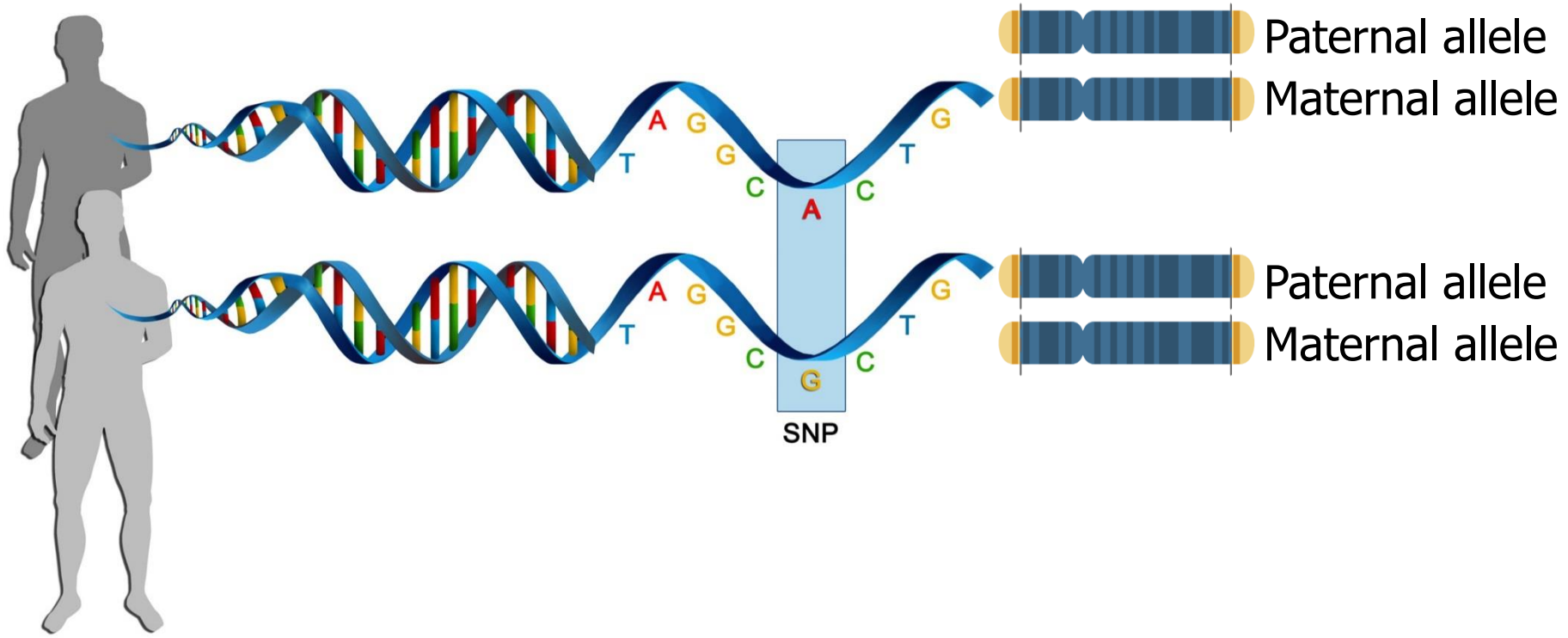


Heritability of a trait

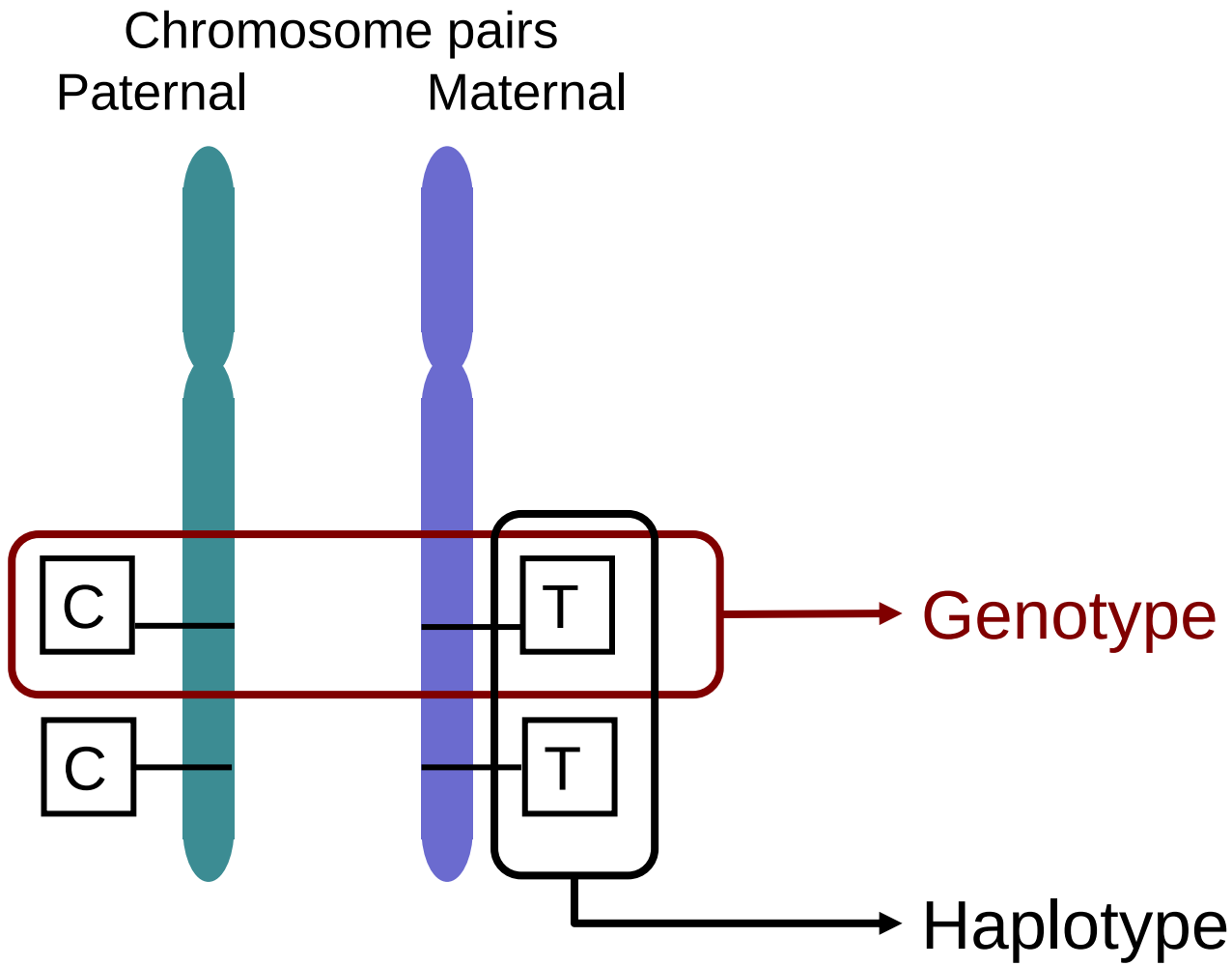


- How large is the contribution of the genotype in a trait?
- The size of the genetic component is not related with the amount of loci.

Single Nucleotide Polymorphism



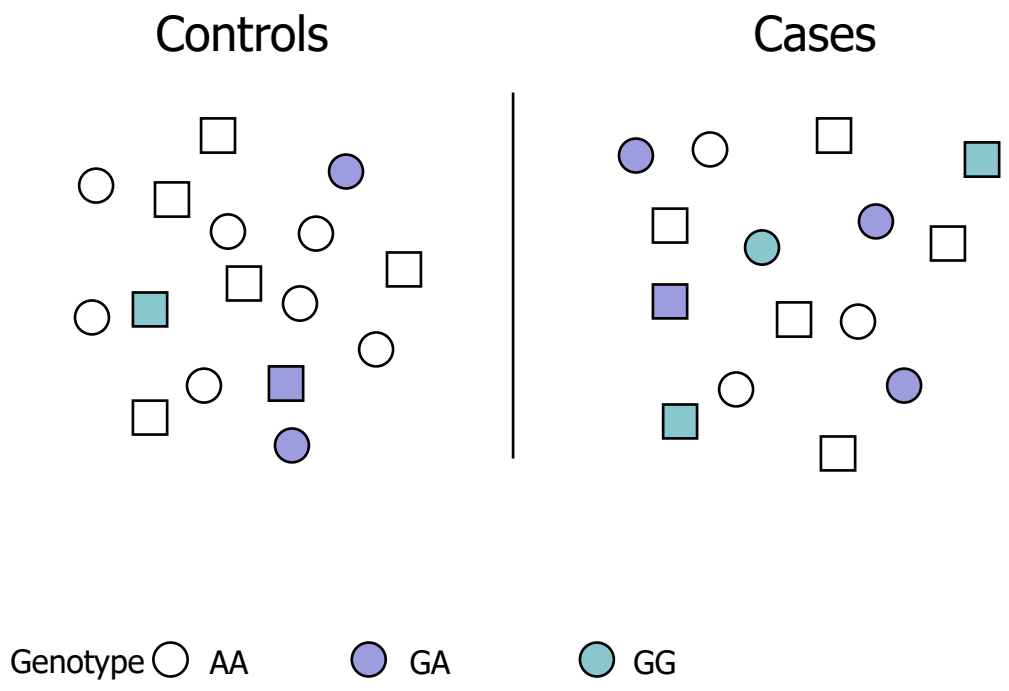
Genetic variation



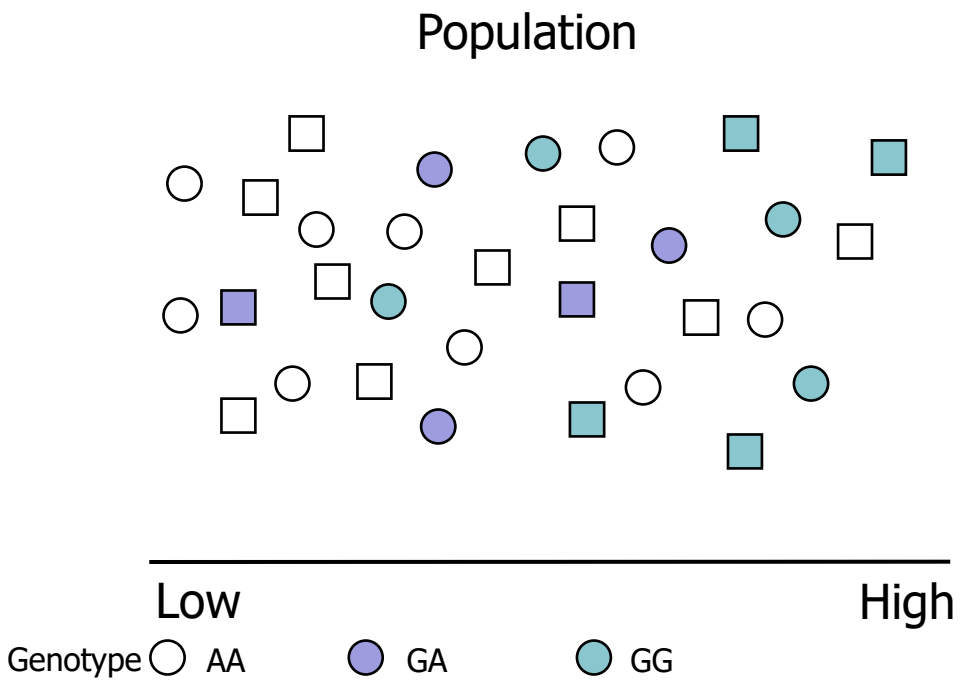
SNP rs429358 19:44,908,684

SNP rs7412 19:44,908,822

Case – Control study (qualitative trait)

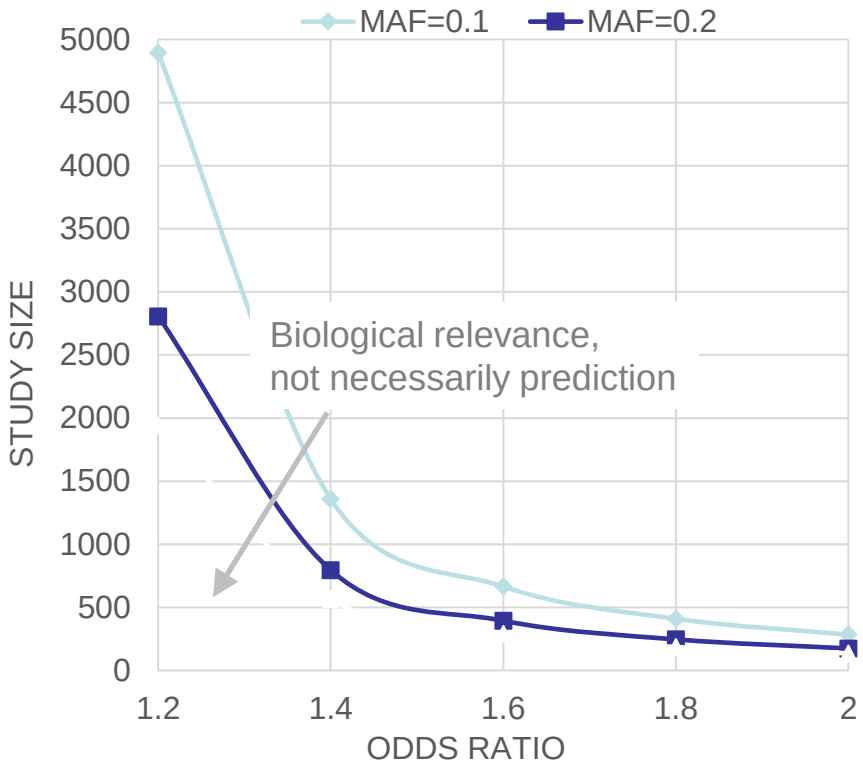


Biomarker study (quantitative trait)



Prerequisites for genetic association study

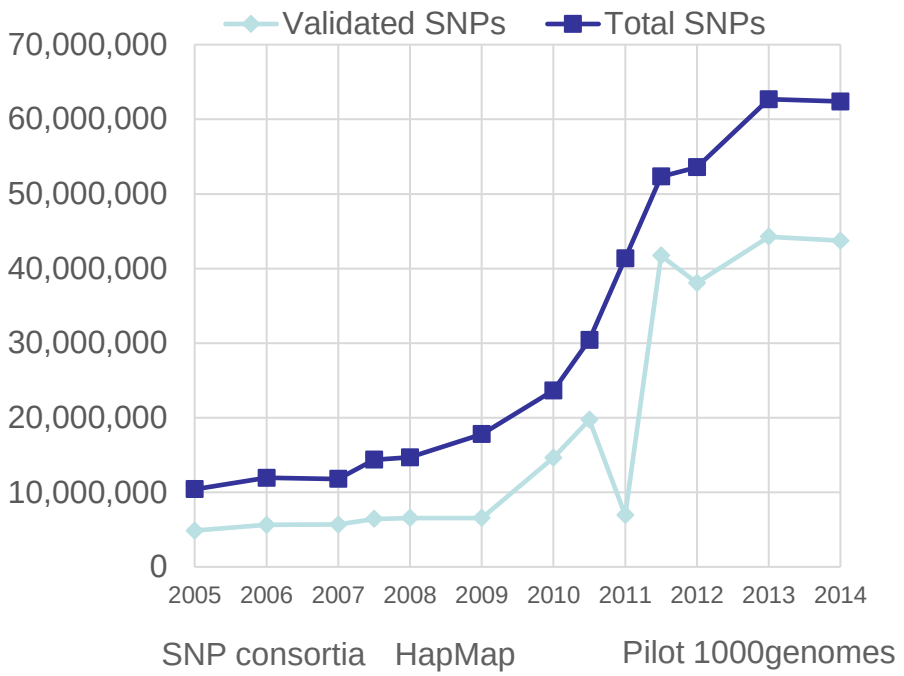
- Trait:
  - Genetic component
- Population
  - Sample size
- Genetic variation
  - Minor Allele Frequency



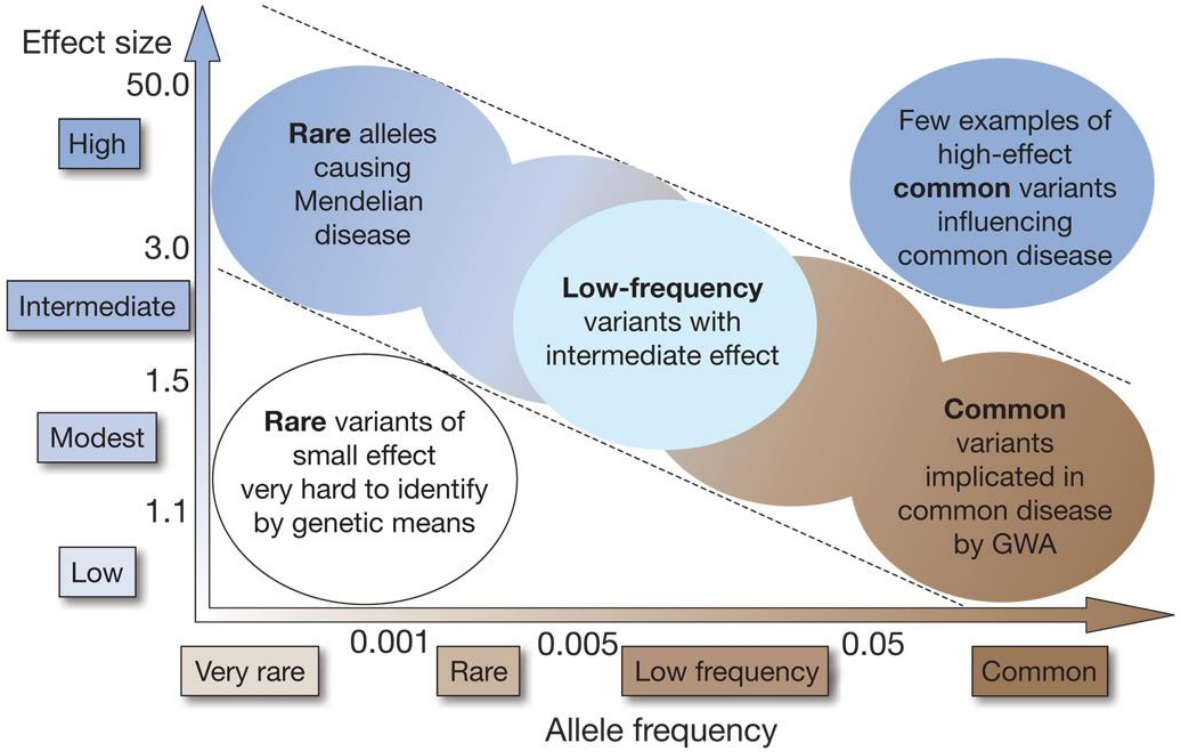
Study sizes assuming
Power=80%, P=0.05

Genome wide genetic association study

- Trait:
  - Genetic component
- Population
  - Sample size
- Genetic variation
  - Minor Allele Frequency
  - Genome wide genetic variation



Minor allele frequency and effect size

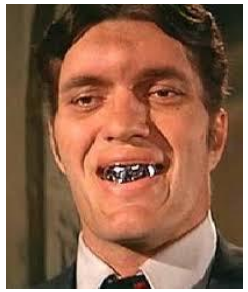


TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

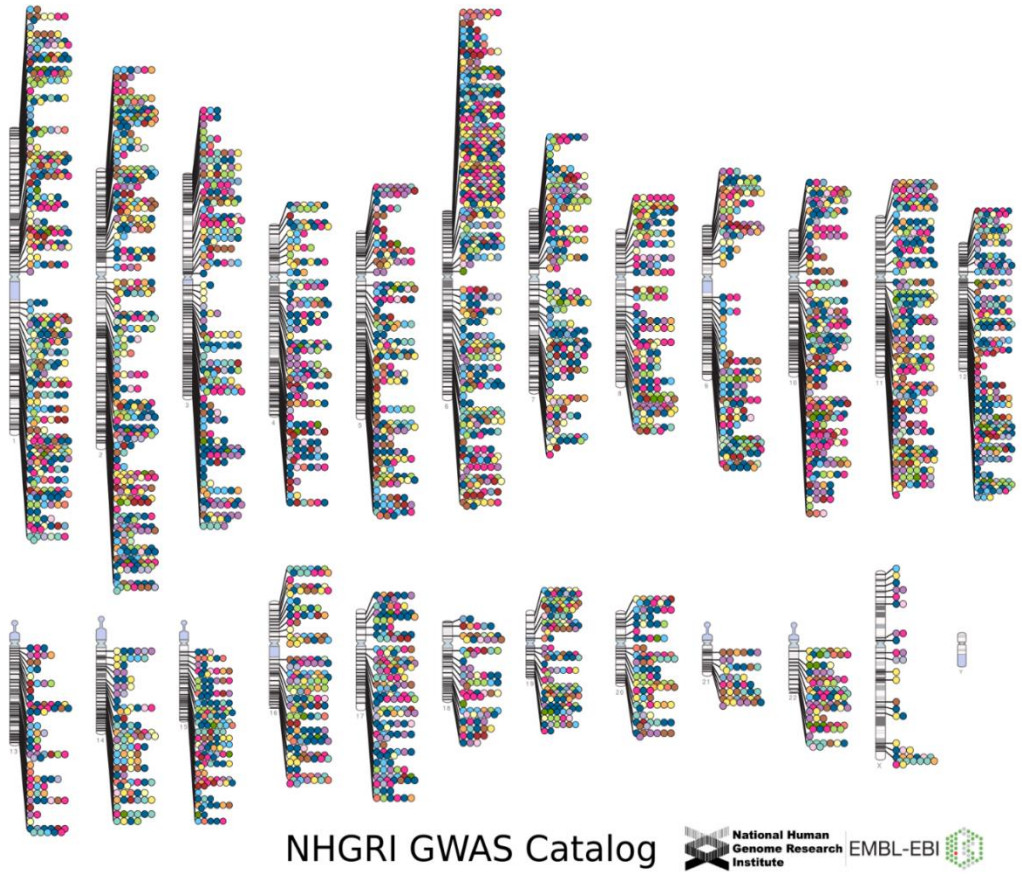
Prerequisites for GWAS

- Trait with an assumed/established genetic component
- Large population in which trait and genetic variation has been measured
  - Formation of large consortia
- Genetic variation
  - Common variants (MAF > 1%)
  - Localization of genetic variation
  - Technology
- Statistics and informatics

Some bad guys…

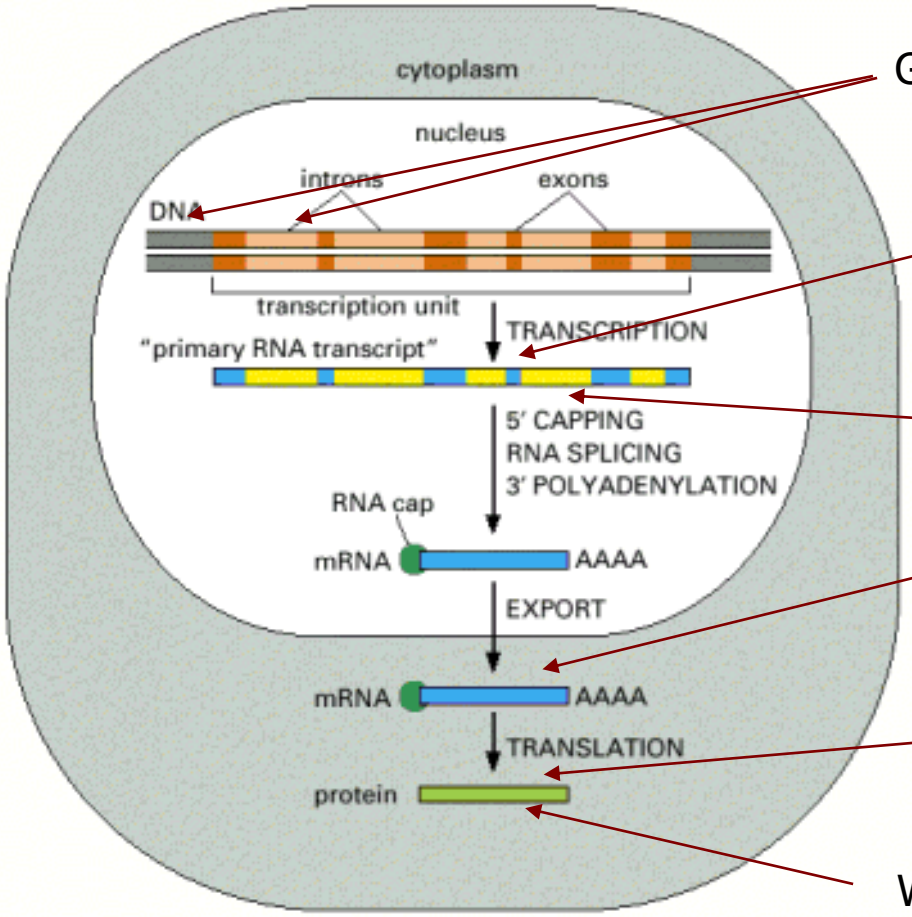


…have been identified



- 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

Biological effects of genetic variants



Gene expression

Stop codon

Splicing: e.g exon skipping

Binding miRNA: stability

Changes in silent 3rd base: stability mRNA

Wrong amino-acid

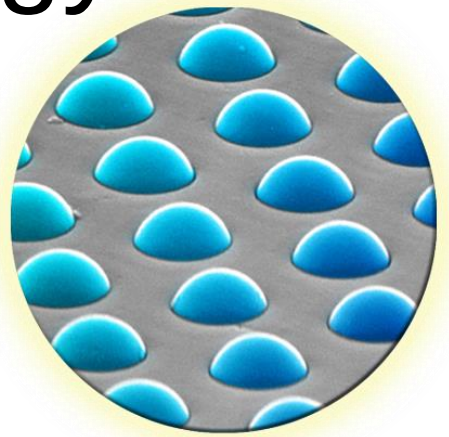
Prerequisites for GWAS

- Trait with an assumed/established genetic component
- Large population in which trait and genetic variation has been measured
  - Formation of large consortia
- Genetic variation
  - Common variants (MAF > 1%)
  - Localization of genetic variation
  - Technology
- Statistics and informatics

Learning goals

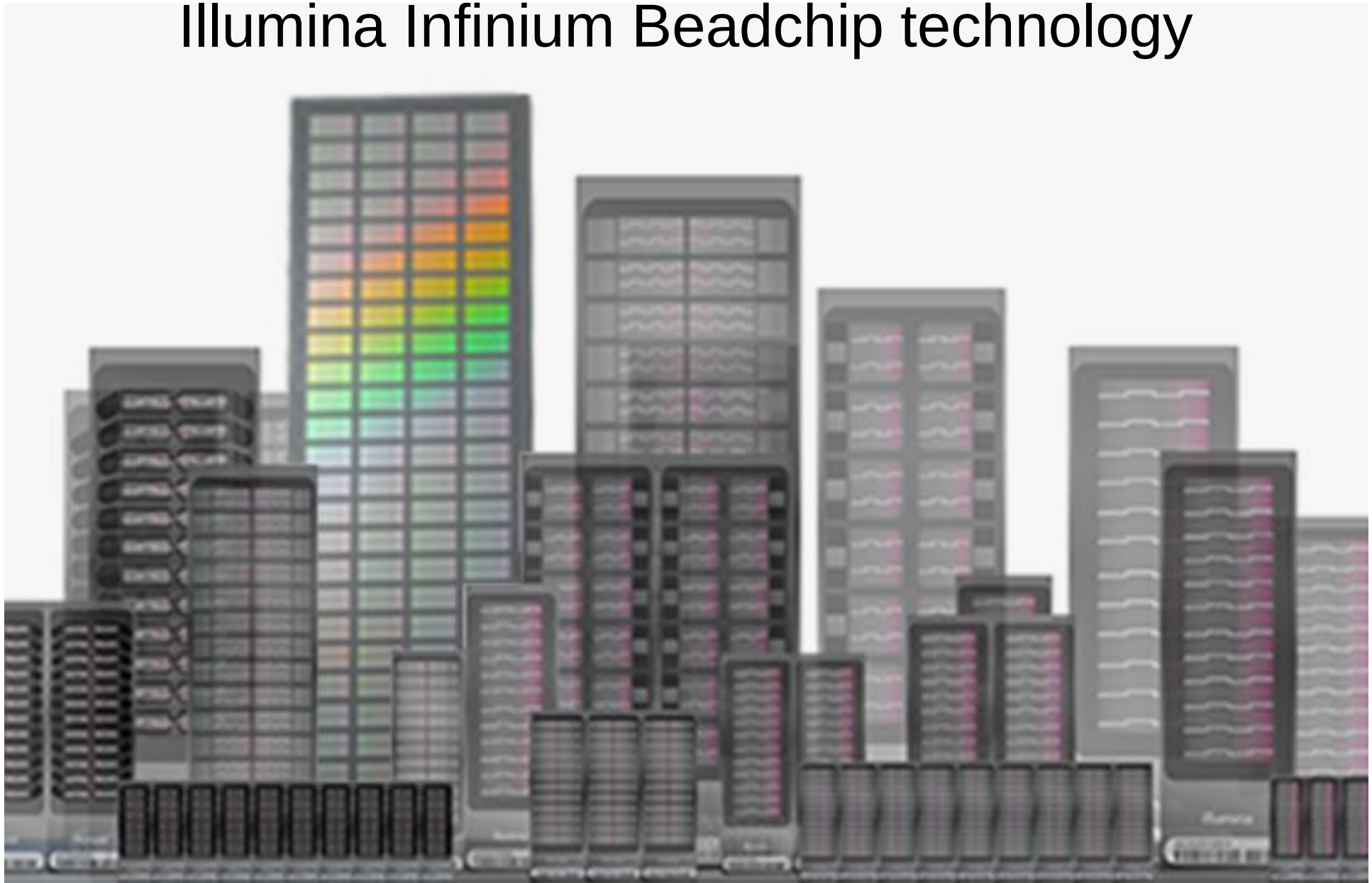
- After the second part you are able to
  - Interpret linkage disequilibrium
  - Explain the difference between D’ and R^2
  - Understand the principle of genetic imputation

Genotyping technology

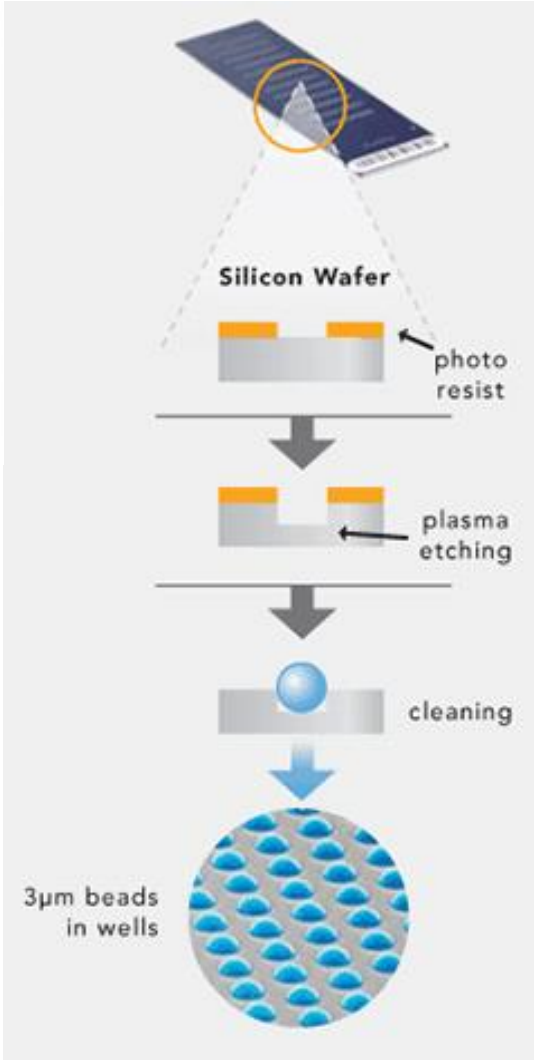


- Illumina Human OmniExpress
  - ~700k common SNPs
  - Copy Number Variations (CNV)
  - Very high data quality (call rate 99.84%)
  - Reasonable throughput (12 samples per chip)
  - Cost ~€ 200-300 per sample all in
- Other Illumina chips up to 2.5M SNPs

Genetic variation: Genotyping
Illumina Infinium Beadchip technology

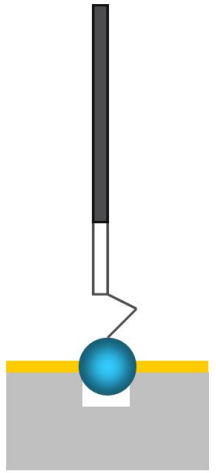


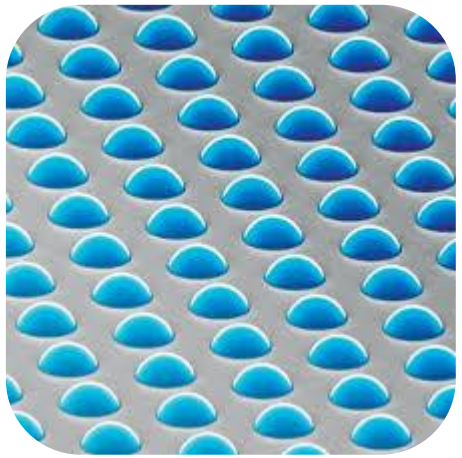
Genetic variation: Genotyping
Illumina Infinium Beadchip technology

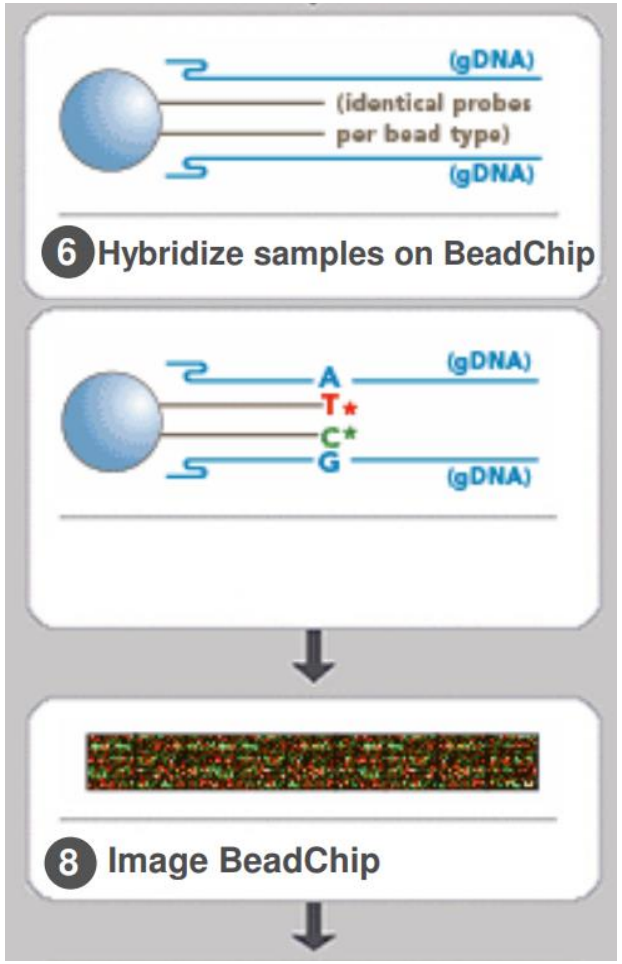


Specific Probe
(50 base nt)

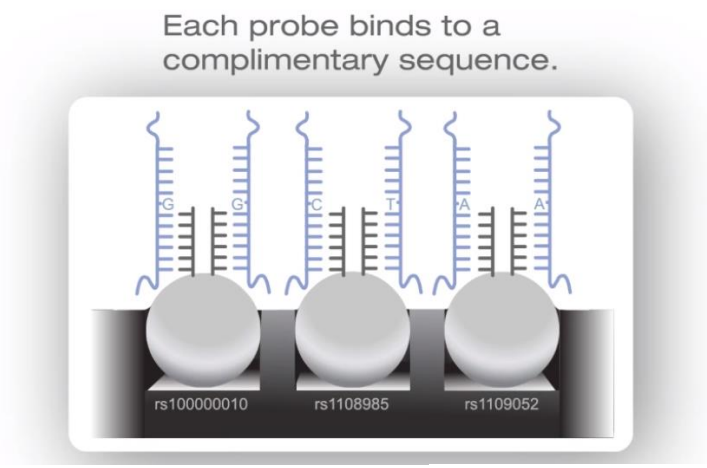
Bead Identifier
(30base nt)

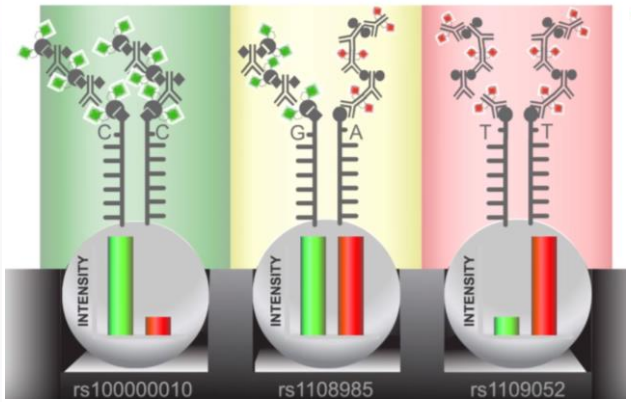


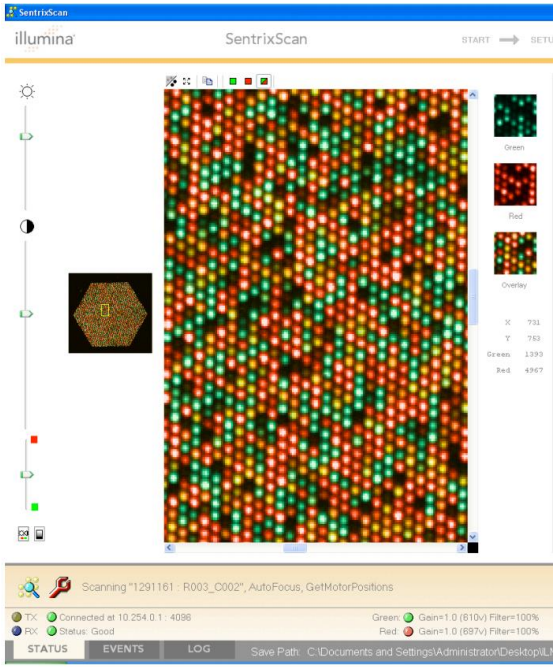


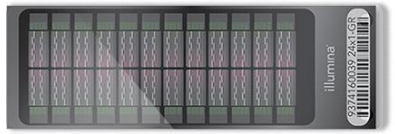


Genetic variation: Genotyping
Illumina Infinium Beadchip technology









Indirect approach of association

- Test all common genetic variation…
- …by genotyping a small subset of SNPs only (efficient!)
- Exploit ‘linkage disequilibrium’

Indirect approach of association

- Two SNPs

1. A/G with MAF=0.40
2. C/T with MAF=0.20

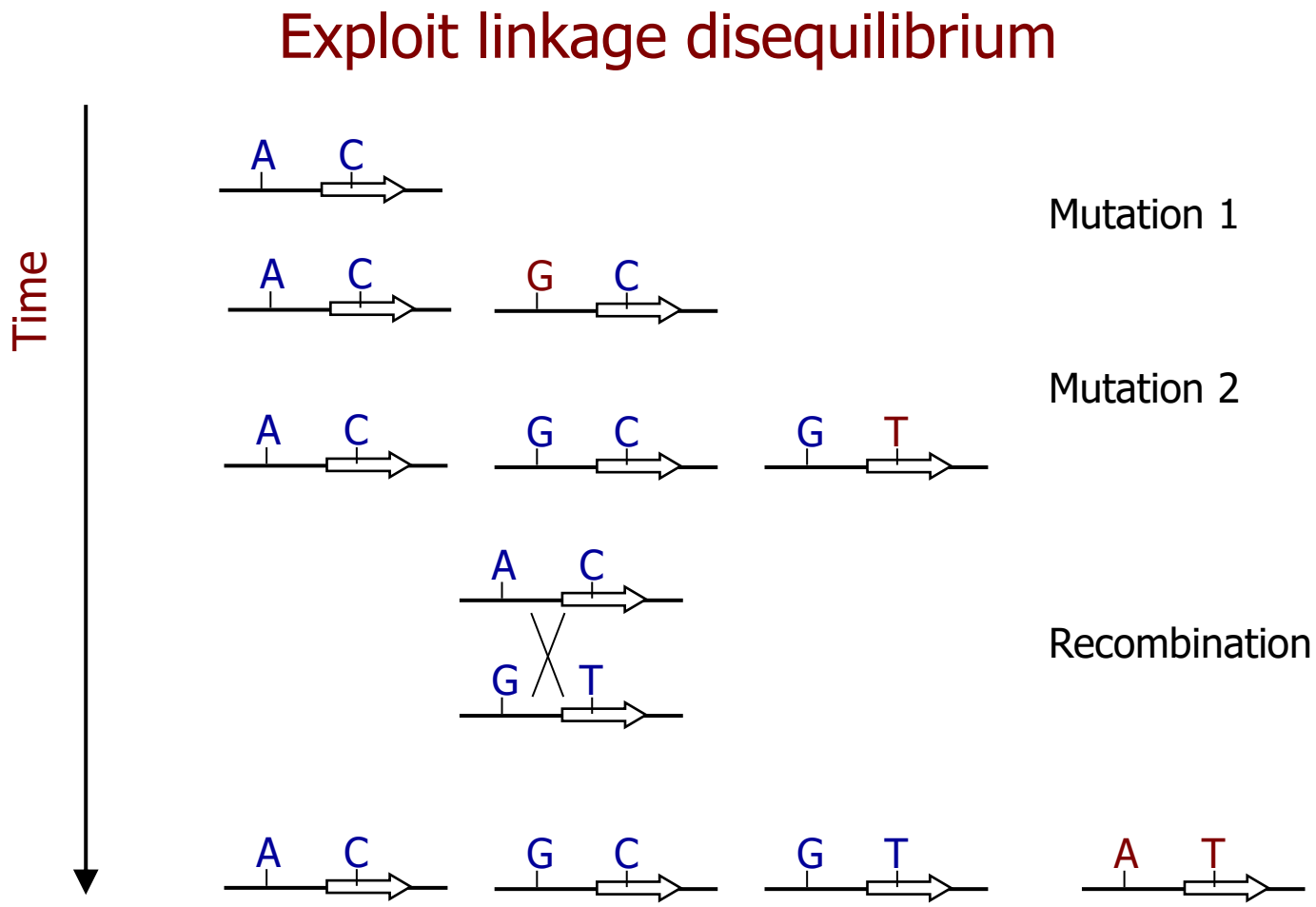
- Expectation combinations

1. A-C: 0.60x0.80=0.48
2. A-T: 0.60x0.20=0.12
3. G-C: 0.40x0.80=0.32
4. G-T: 0.40x0.20=0.08

- Frequently this does not hold for close by SNPs:

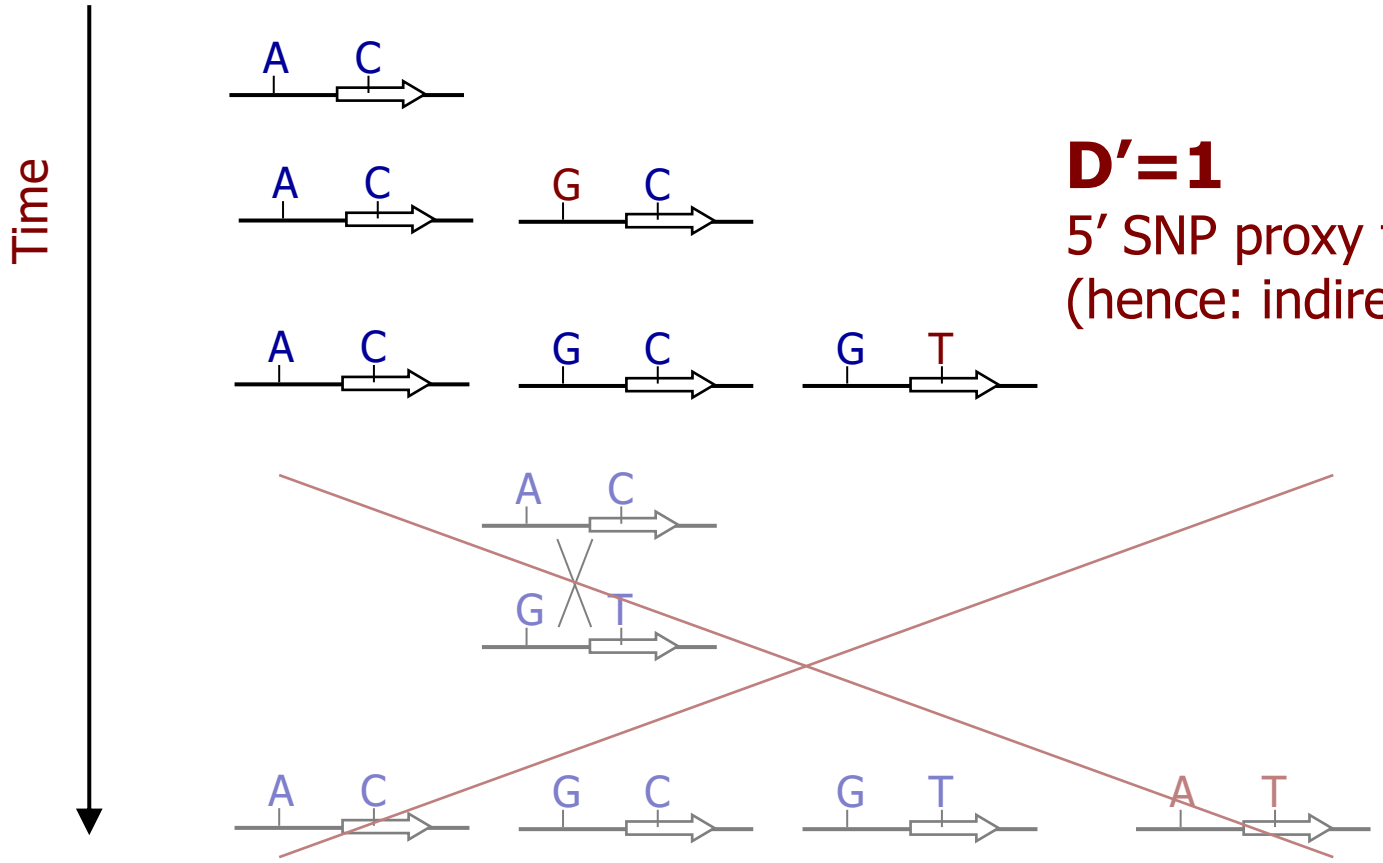
→ *DI*SEQUILIBRIUM

Indirect approach of association



Indirect approach of association

If no recombination:

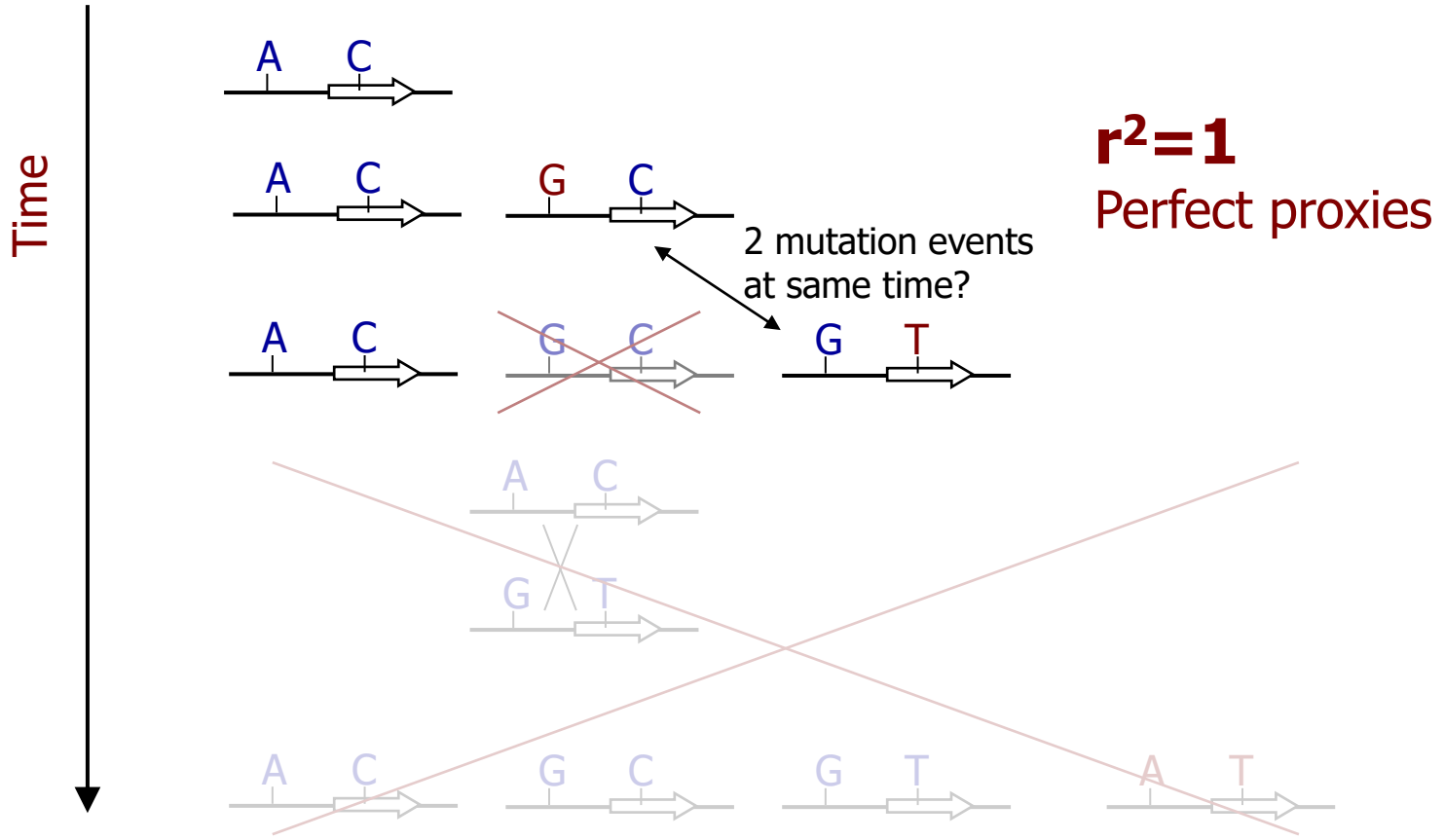


D'=1

5' SNP proxy for 3' SNP
(hence: indirect)

Indirect approach of association

If also equal allele frequencies:



$r^2=1$

Perfect proxies

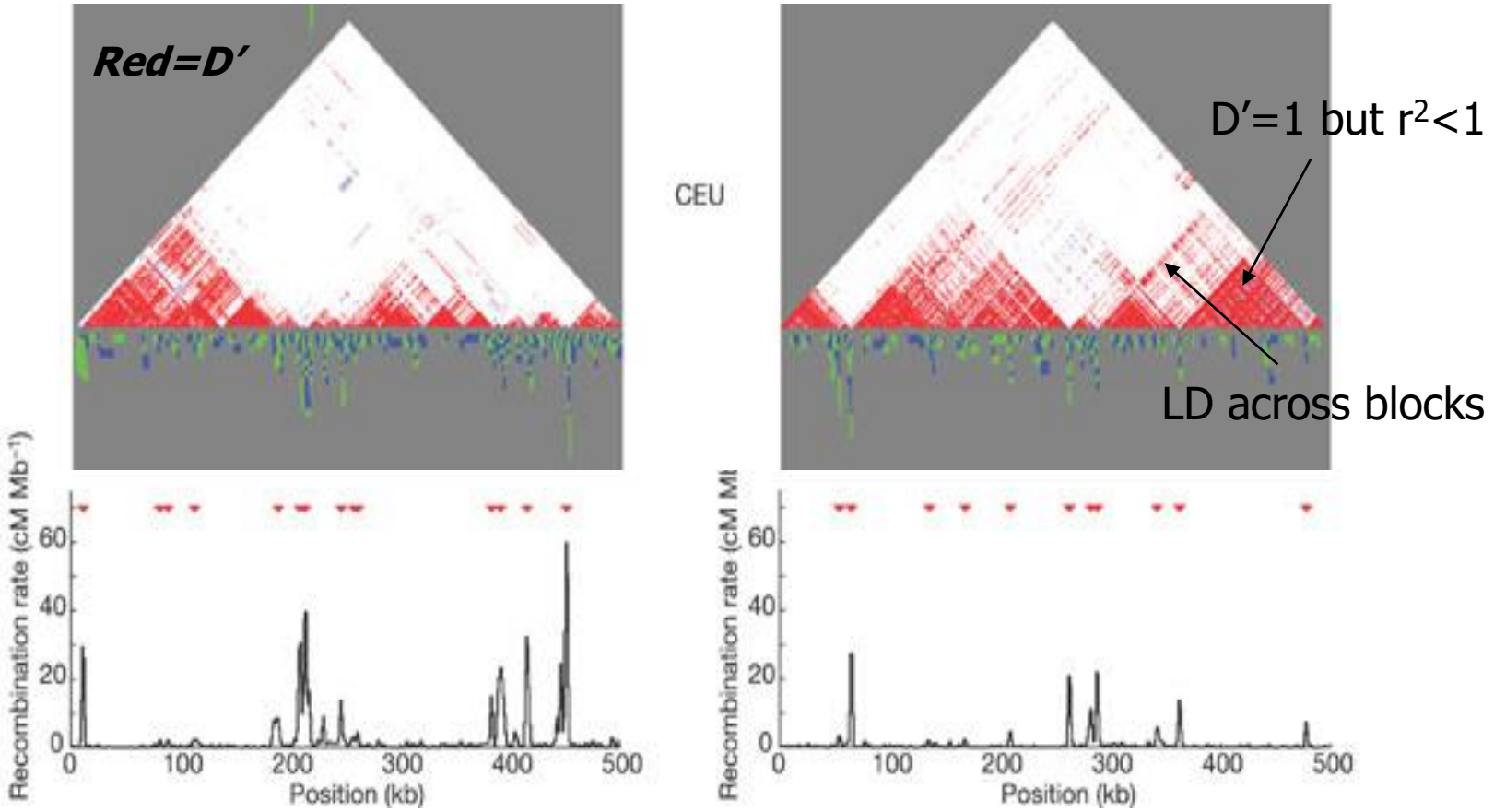
International HapMap project

- Samples HapMap phases 1+2
  - Yoruba, Nigeria (YRI):
    - n=90 (30 parent-offspring trios)
  - Ceph - Utah, USA (CEU):
    - n=90
  - Han Chinese, Beijing (CHB) + Japanese, Tokyo (JPT):
    - n=90

- Genotyping
  - 6,349,188 suspected SNPs assessed
  - 2,819,322 indeed polymorphic and MAF>0.05
  - No resequencing done but ENCODE regions (10 x 5Mb) sequenced as reference

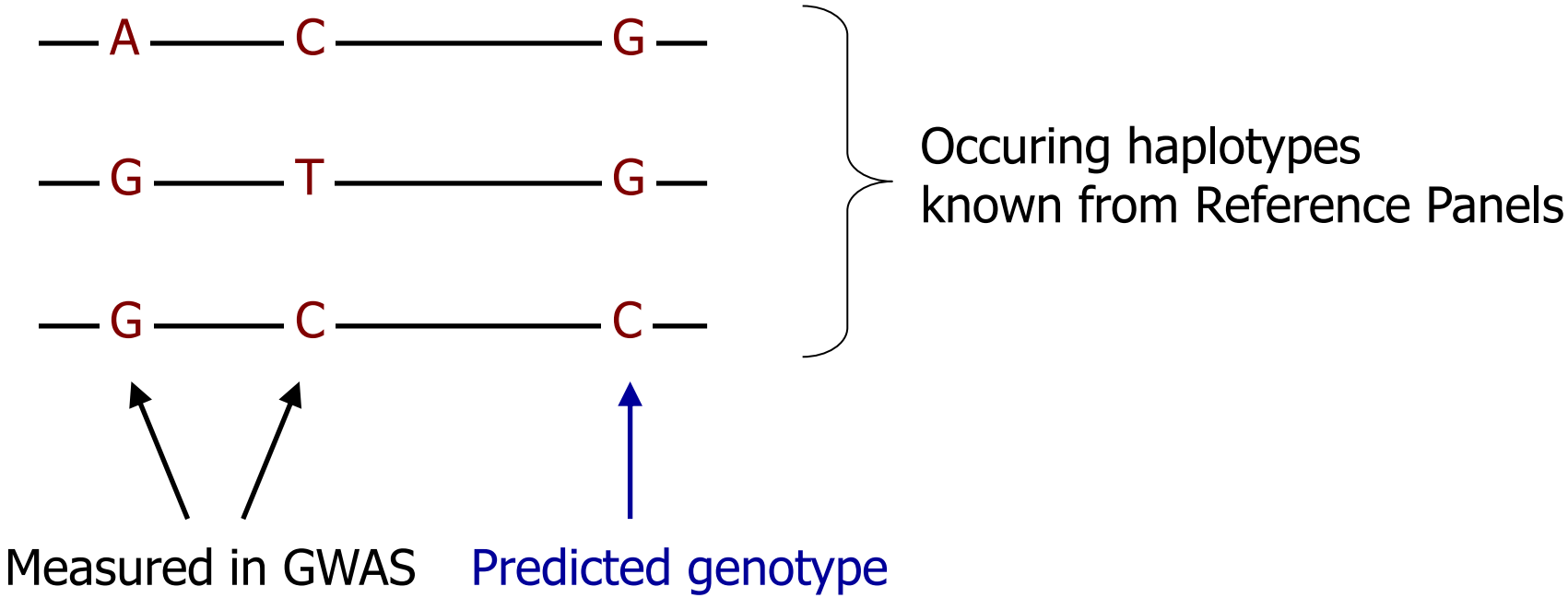
HapMap consortium, Nature 2005&2007

Genetic variation is limited



Genetic imputation

- Prediction of missing genotypes using LD



Imputation Reference Panels

- International HapMap project (2007)
  - CEU (European), YRB (African), JPT/CHB (Asian)
  - 270 samples
  - ~2.5M SNPs

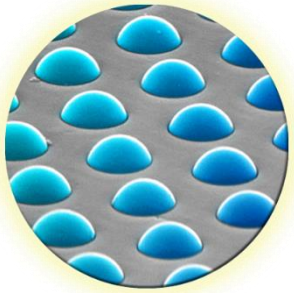
- 1000 Genomes Project (2010)
  - EUR (European), AFR (African), ASN/SAN (Asian), AMR (Americas)
  - 2,535 samples
  - ~30M SNPs + Indels

- The Haplotype Reference Consortium (2015)
  - Mainly EUR ancestry
  - 32,611 samples
  - 39.2 M SNPs

The Haplotype Reference Consortium

etic variation: Genotyping and imputation

- Illumina Global Screening Array (GSA) array
  - ~640k common SNPs
  - Very high data quality (call rate 99.84%)
  - High throughput (24 samples per chip)
  - Cost ~€ 50 per sample all in



- => Haplotype Reference Consortium imputation up to 40 Million SNPs!

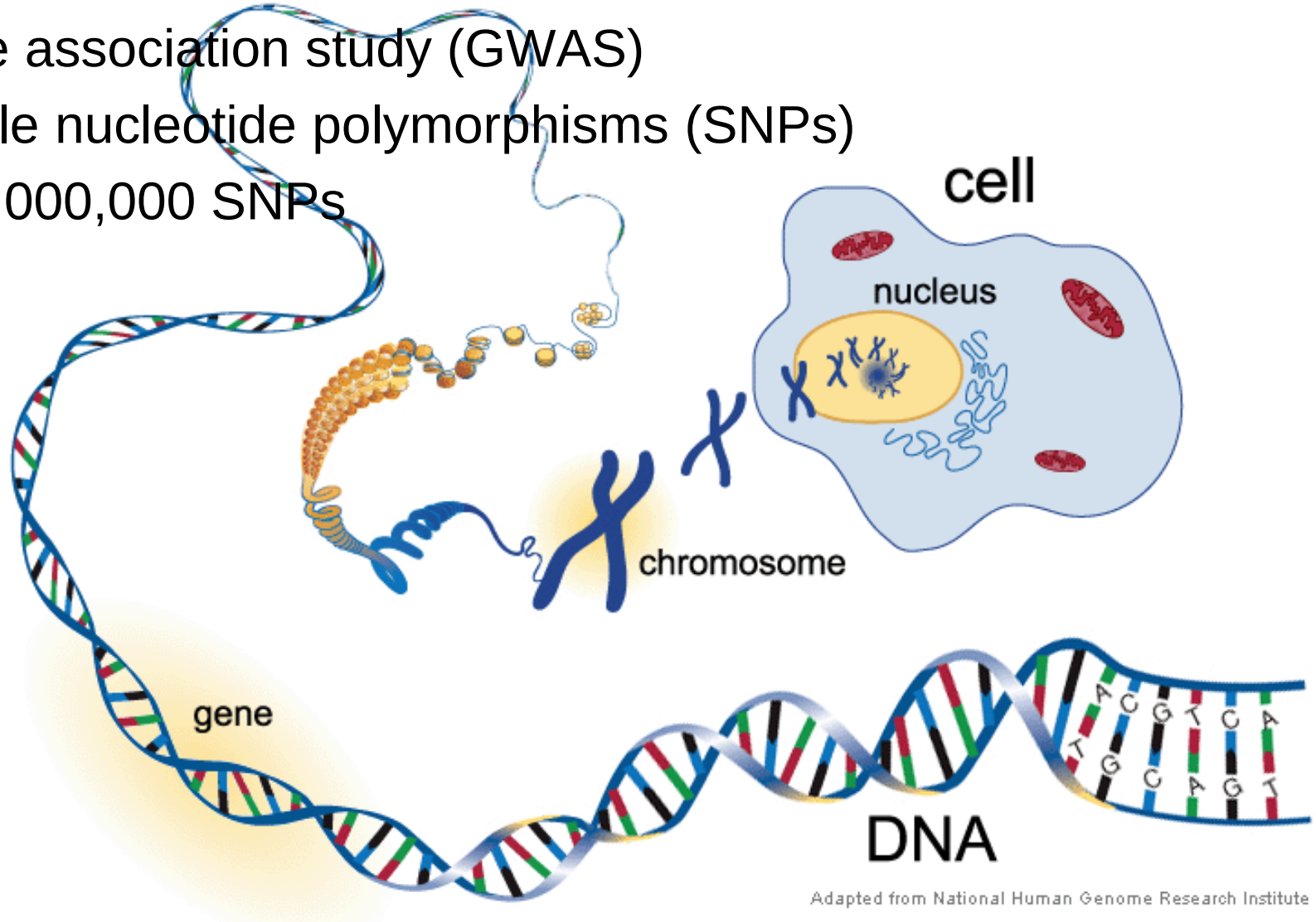
HELP!

I generated 400 billion data points

- 40,000,000 SNPs
- 10,000 individuals

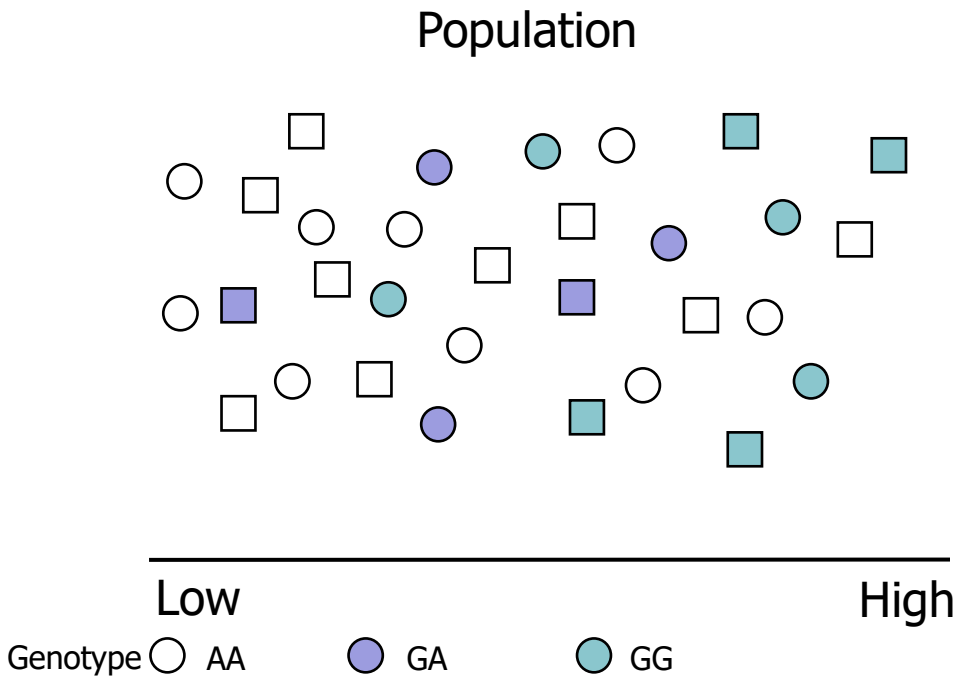
Practical

- Genome-wide association study (GWAS)
- Measure single nucleotide polymorphisms (SNPs)
- 300,000 – 40,000,000 SNPs



Practical

- Investigate whether different genotypes are associated with different levels of a biomarker
- Successful approach for many complex diseases/traits



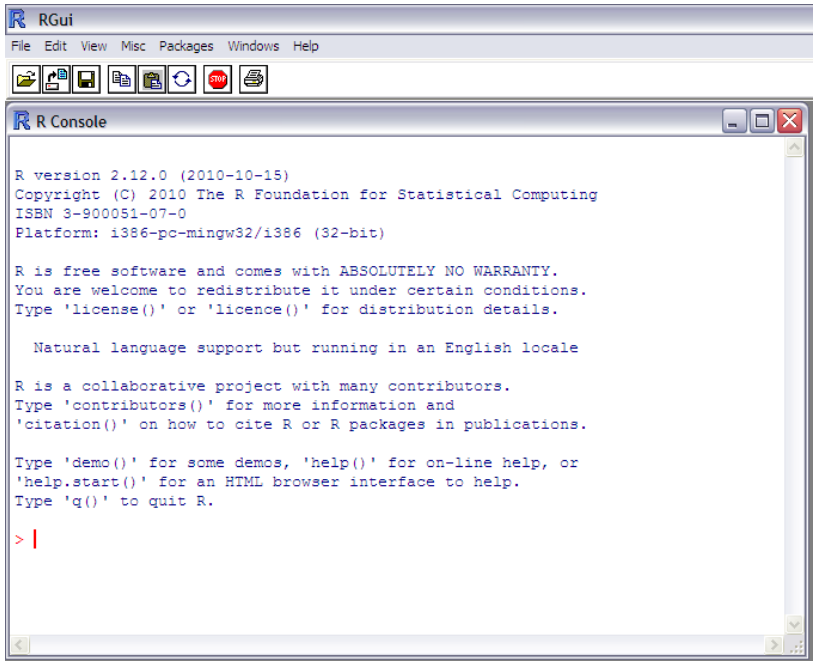
Conquer your fear of The Blinking Cursor

R

Rvtests

Rare variant test software for next generation sequencing data

1. Explore input files
2. Demo by Dina Vojinovic
3. Explore and visualize results




BREAK

Hardy Weinberg Equilibrium

- HWE is used to check the genotyping quality in GWAS analyses
- Allele frequencies should correspond with genotype frequencies
- Example SNP rsXX has 2 alleles:
  A with a frequency of 0.8
  G with a frequency of 0.2

Hardy Weinberg Equilibrium

- A= 0.8 is represented by p
  B = 0.2 is represented by q

=> p+q=1

- From allele frequencies the genotype frequencies can be calculated.

Hardy Weinberg Equilibrium

- Formula of Hardy-Weinberg Equilibrium:
  $AA=p^2$
  $AG=2pq$
  $GG=q^2$
- In this case:
  $AA=0.8*0.8= 0.64$
  $AG=2*0.8*0.2=0.32$
  $GG=0.2*0.2=0.04$

NB: under the assumption of random mating

Hardy Weinberg Equilibrium

- HWE test:

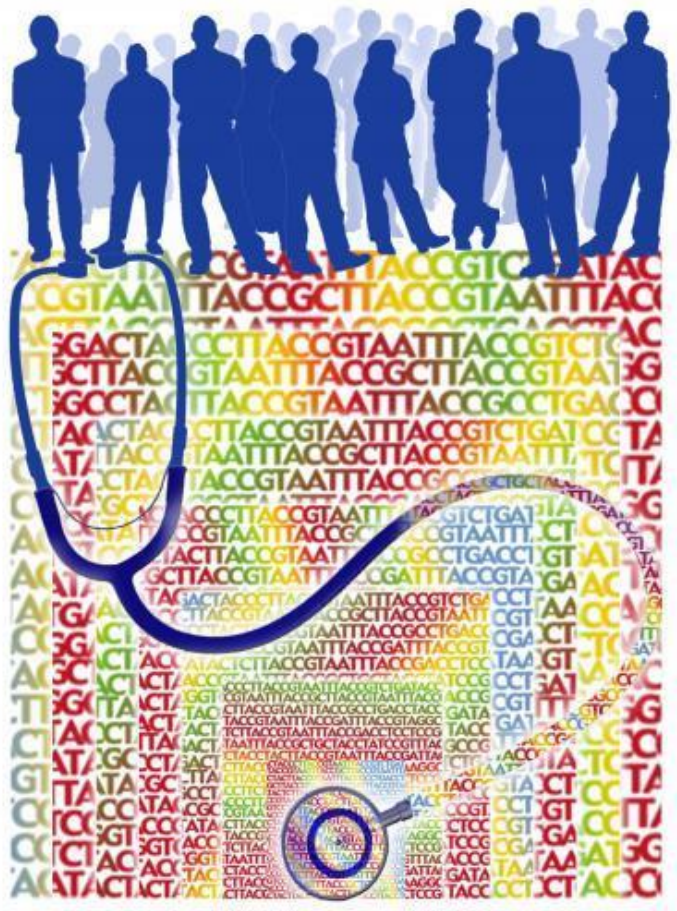
Genotypes	Observed N	Expected N	$(obs-exp)^2/exp$	
AA	60	64 (p^2 * 100)	0.25	
AG	20	32 (2pq * 100)	4.5	
GG	20	4 (q^2 * 100)	64	
Total	100		Chi^2=68.75	
A frequency	0.8		Df=1	P<0.05
G frequency	0.2			

Learning goals

- After this third part you are able to
  - Explain the principles of a genetic association analysis
  - Apply adjustment for multiple testing in genome wide association studies
  - Understand the importance of large sample size and replication of results

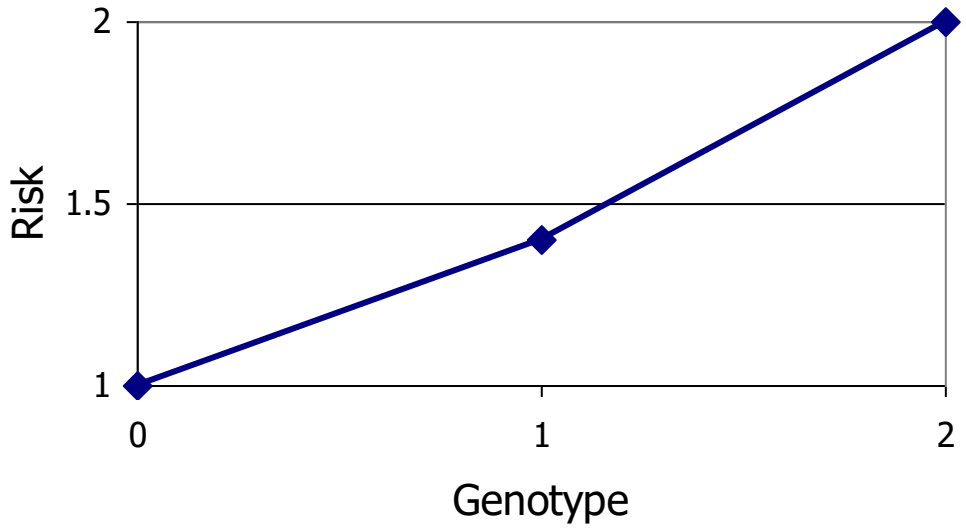
Statistical Analysis

- Keep it simple
  - Single SNP analysis
  - Easy to interpret
  - Minimal number of tests
  - All very basic statistical tests (X^2 or similar)



Statistical Analysis

- Cochrane-Armitage trend test
  - Same as linear-by-linear in SPSS
  - Additive effect: more risk alleles, more effect
  - Genotype coding: 0, 1, 2 (counting no. of rare alleles)
  - Plausible biological model
  - Robust against random fluctuations
  - Optimal power (df=1)



Regression analysis

$$\text{Trait} \sim \text{Constant} + \beta_1 \times \text{SNP}$$

Where the dependent variable “Trait” is

a) Quantitative trait => linear regression

b) Case-Control status => logistic regression

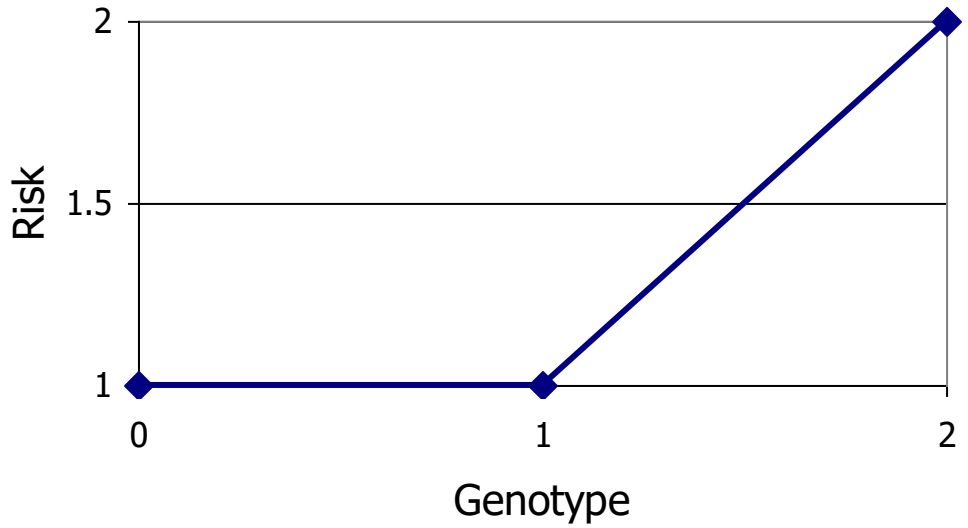
And where the independent variable “SNP” is coded as number of alternative alleles (0, 1 or 2)

Cochrane Armitage Trend test H0: $\beta_1 = 0$, No association

ASSOCIATION if β_1 is NOT EQUAL to 0

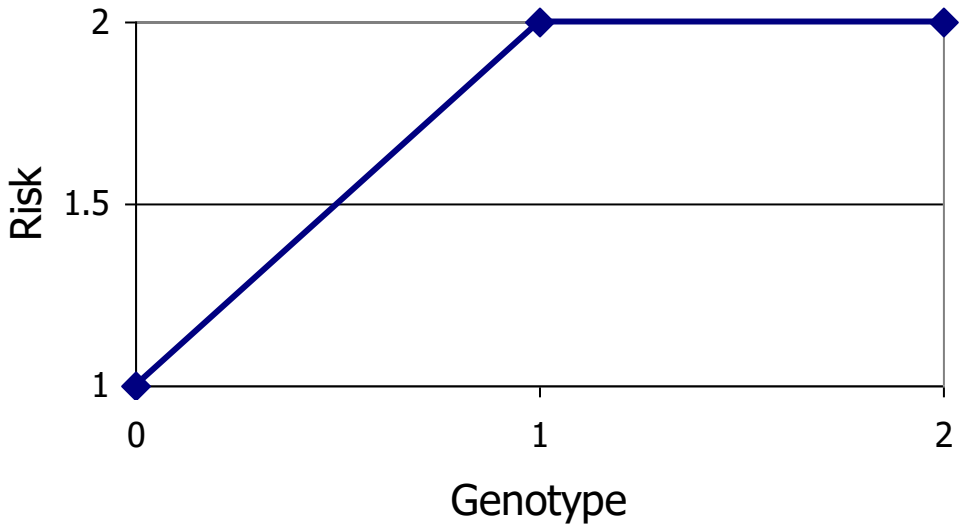
Alternatives

- Recessive test
  - Mendelian disease like Cystic Fibrosis
  - Assumes effect among rare homozygotes only

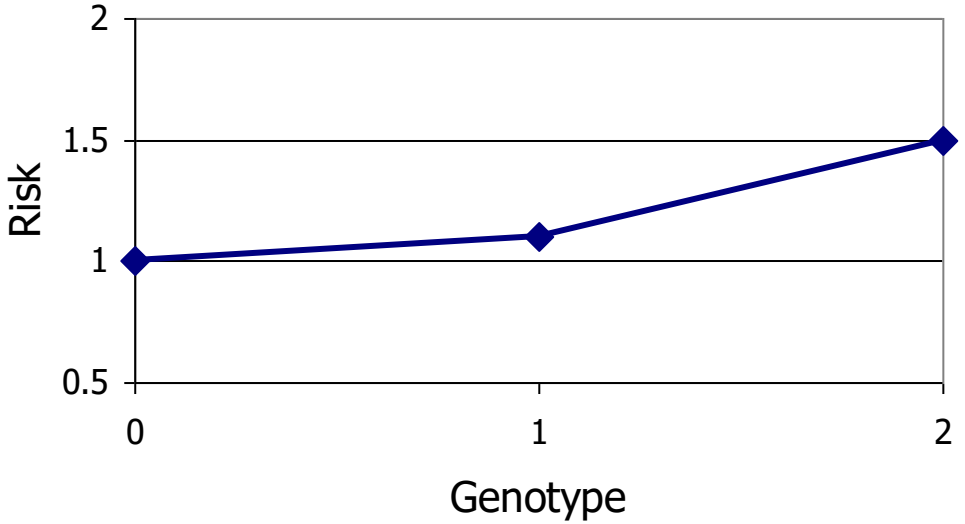


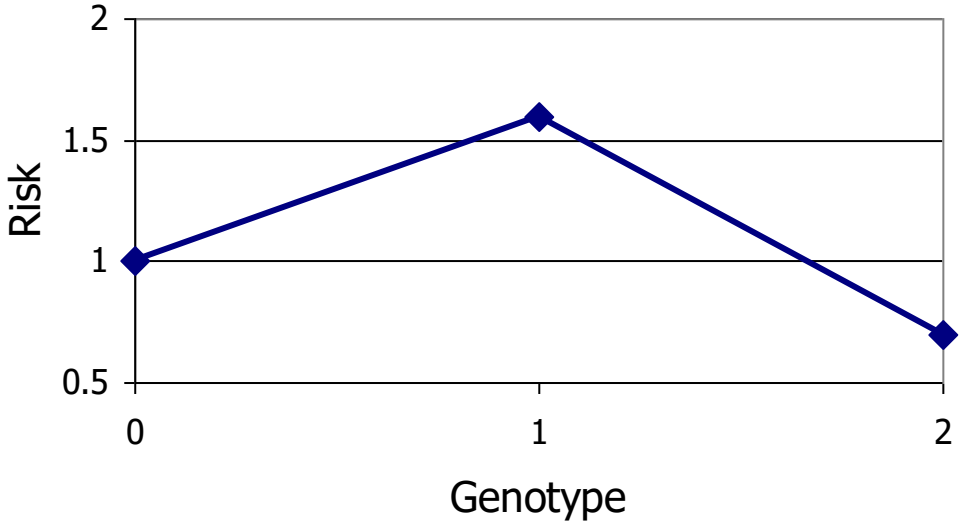
Alternatives

- Dominant test
  - Mendelian disease like Huntington's Disease
  - A single rare allele is sufficient for disease trait
  - Common homozygotes and heterozygotes same effect
  - Often trend test has sufficient power



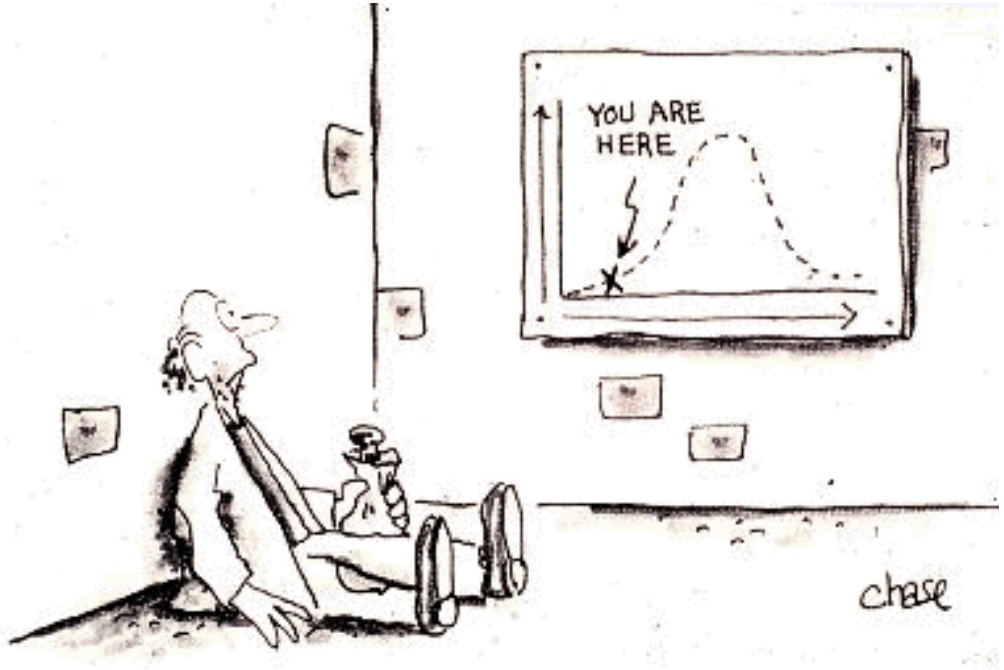
What if





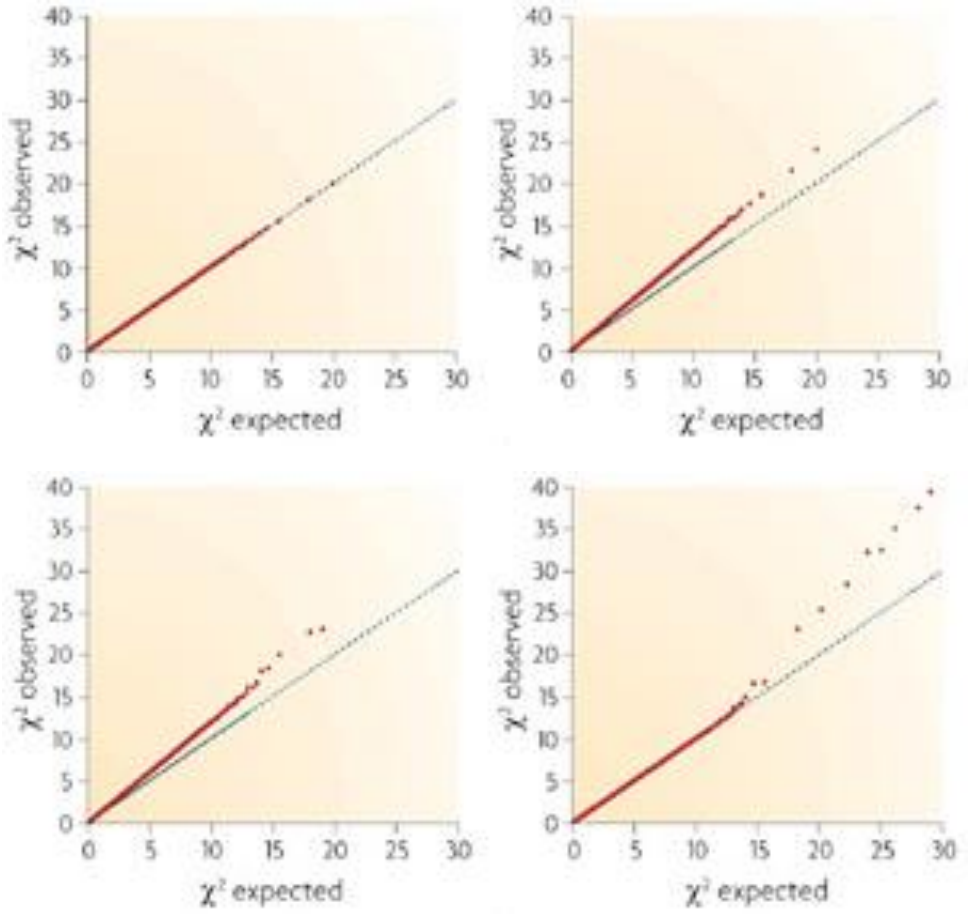
Significance in GWAs

- Consensus on genome-wide significance in GWAs
  - 1 million independent tests
  - $P < 5x10^{-8}$ ($0.05/10^6$)
  - But ignores ‘enrichment’ for low p-values



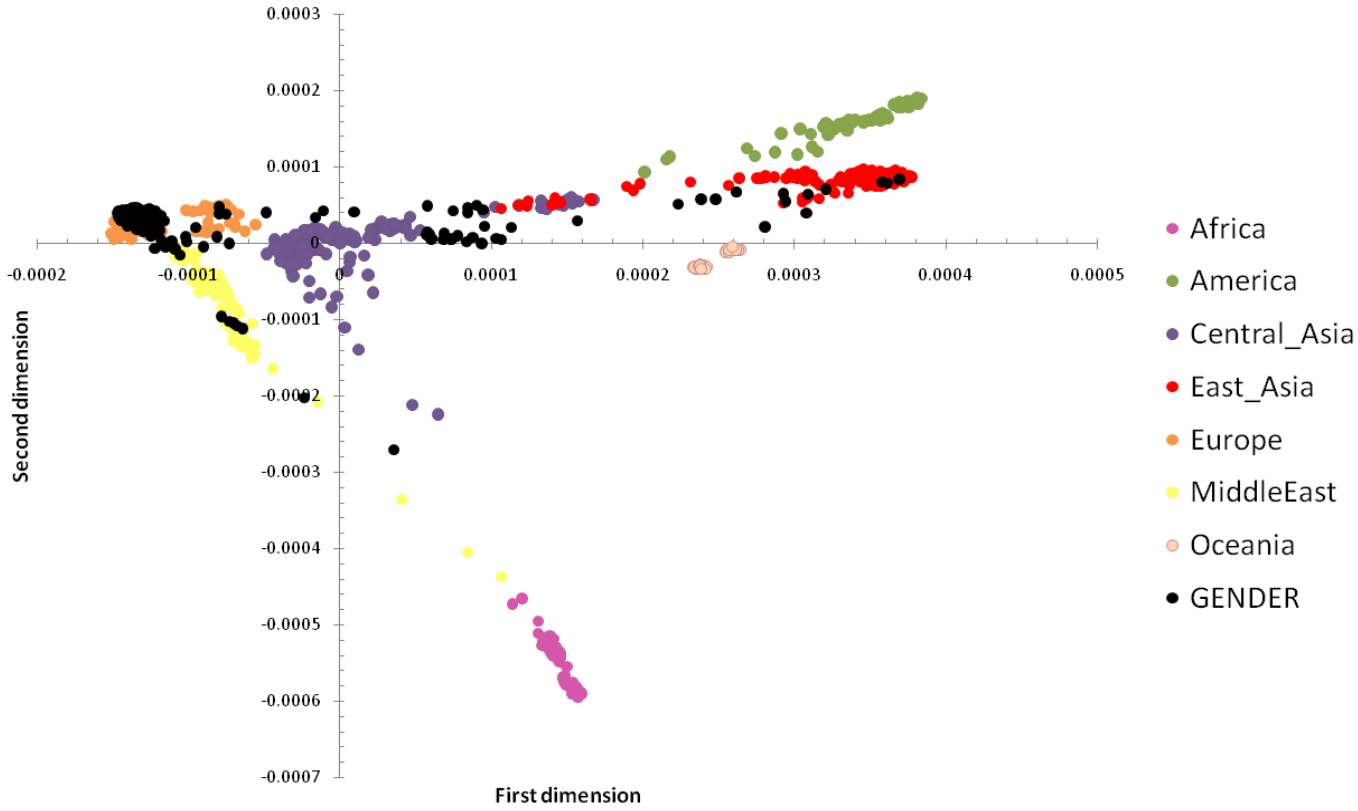
Bias and enrichment

- QQ plots allow to detect bias and enrichment for low p-values



Prevent population stratification

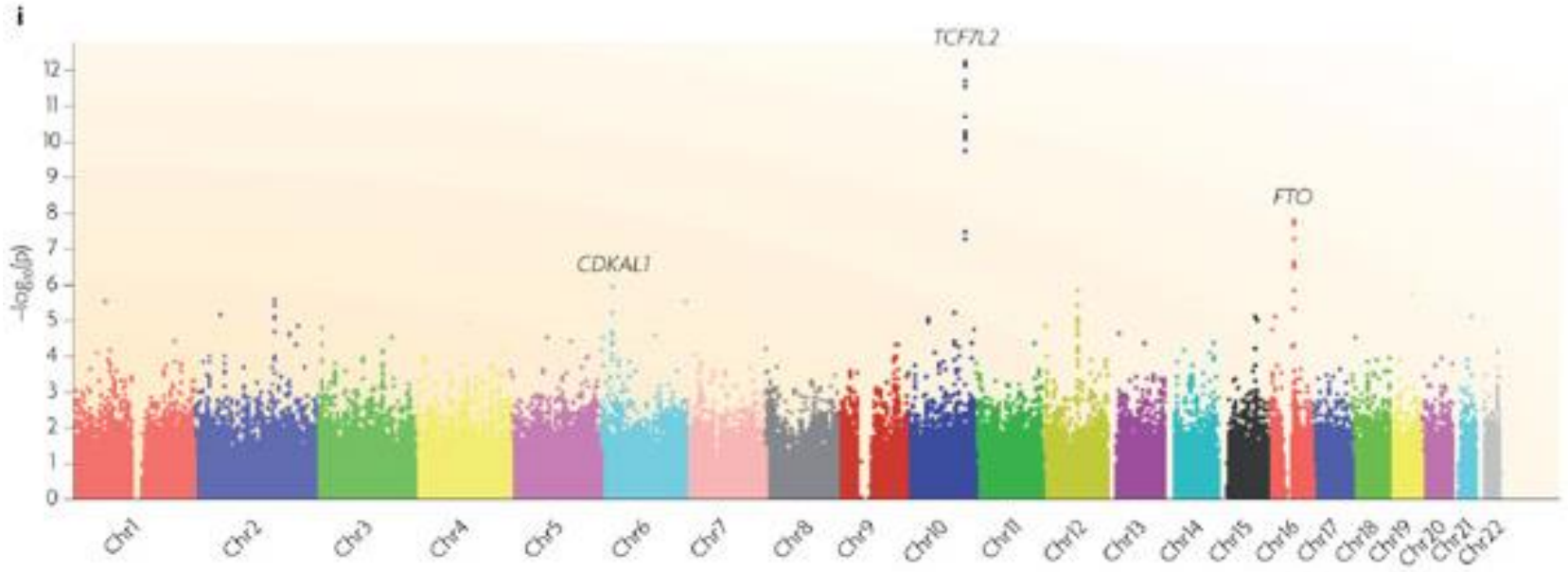
- Detect heterogeneity in origin of participants by comparing genotyping results to HapMap data



Sampietro *et al*. Hum Mol Genet. 2011 Dec 1;20(23):4748-57

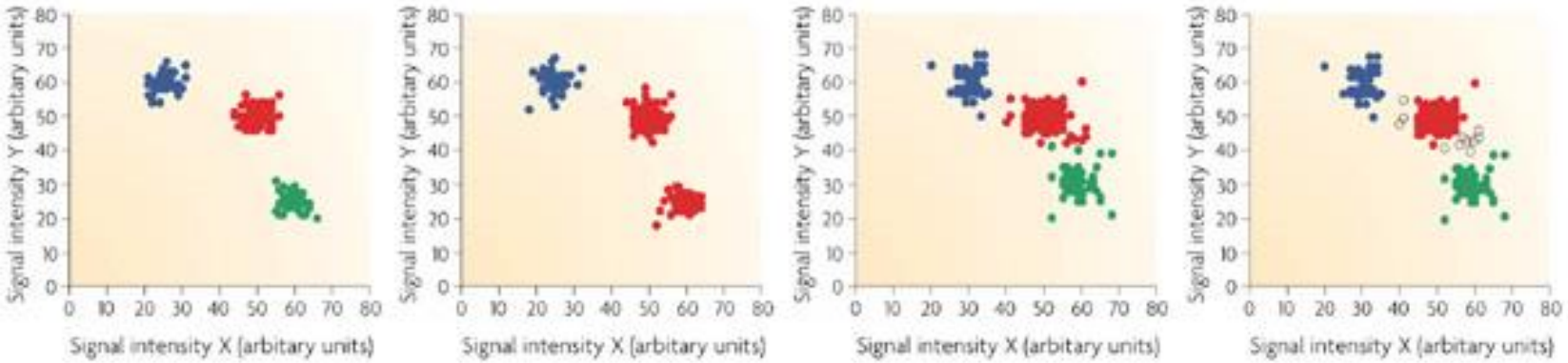
Visualization

- Manhattan plots are standard way to display GWAS results



Not all is fancy

- Check cluster plots of identified SNPs



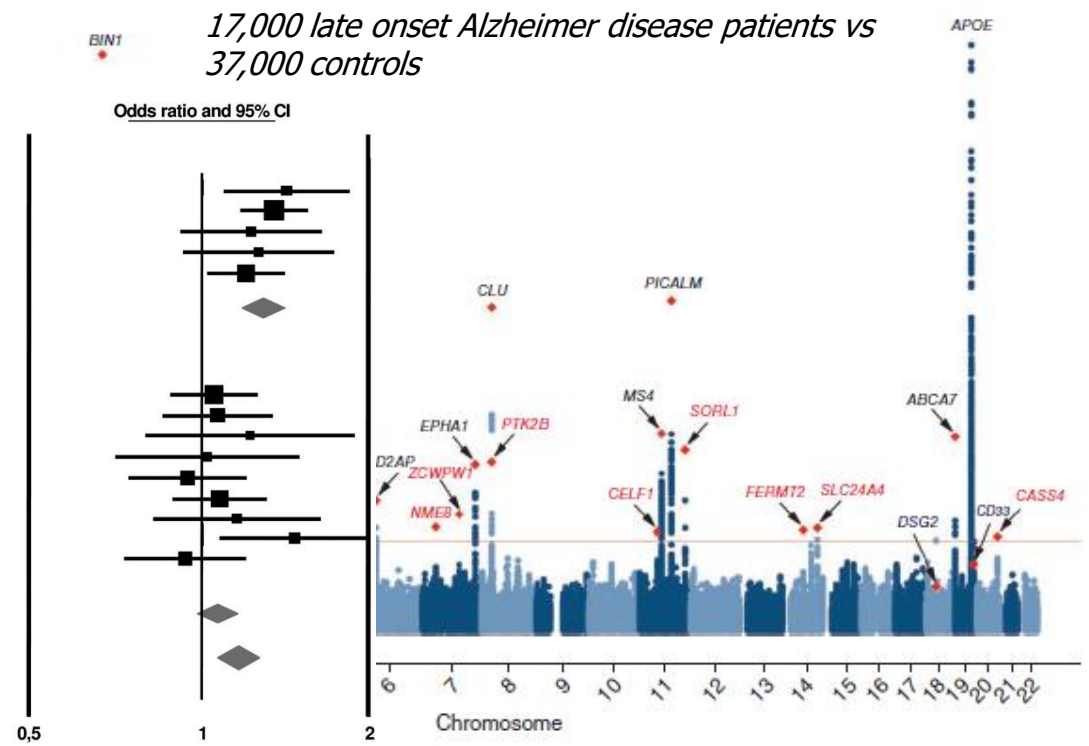
Association of multiple SNPs (in LD) may be considered technical validation

State of the art Meta-GWAS

- Combine multiple GWAS (n > 50,000)
- Follow-up in multiple cohorts (n > 50,000)
- Genome wide significant loci (P<5x10-8)

17,000 late onset Alzheimer disease patients vs 37,000 controls

	Nr cases	Nr controls	OR and 95% CI	p-Value
DISCOVERY				
ERF	149	665	1.42 (1.09-1.85)	8.30E-3
RS-I	563	1892	1.35 (1.18-1.56)	1.98E-5
RS-II	110	668	1.23 (0.92-1.65)	1.66E-1
RS-III	85	868	1.27 (0.93-1.74)	1.37E-1
TWINSUK	401	1698	1.20 (1.02-1.41)	2.48E-2
Overall Effect Discovery	**1308**	**5791**	**1.30 (1.19-1.42)**	**1.18E-8**
(Fixed-effects: I^2 = 0%)				
REPLICATION				
1958BC	315	2206	1.05 (0.88-1.26)	5.85E-1
AGES	173	1204	1.07 (0.85-1.34)	5.55E-1
CHINGFORD	48	337	1.23 (0.80-1.89)	3.57E-1
DSDBAC	81	219	1.03 (0.70-1.50)	8.98E-1
EPIFUND	139	503	0.94 (0.74-1.20)	6.34E-1
FOA	384	814	1.08 (0.89-1.31)	4.36E-1
GARP	67	925*	1.16 (0.82-1.64)	4.02E-1
HCS	90	2117	1.47 (1.08-2.00)	1.54E-2
SHIP	183	589	0.93 (0.73-1.20)	5.89E-1
Overall Effect Replication	**1480**	**7989**	**1.07 (0.98-1.16)**	**1.24E-1**
(Fixed-effects: I^2 = 0%)				
Overall Effect All Studies	**2788**	**13780**	**1.17 (1.10-1.24)**	**4.67E-7**
(Fixed-effects: I^2 = 28.4%)				



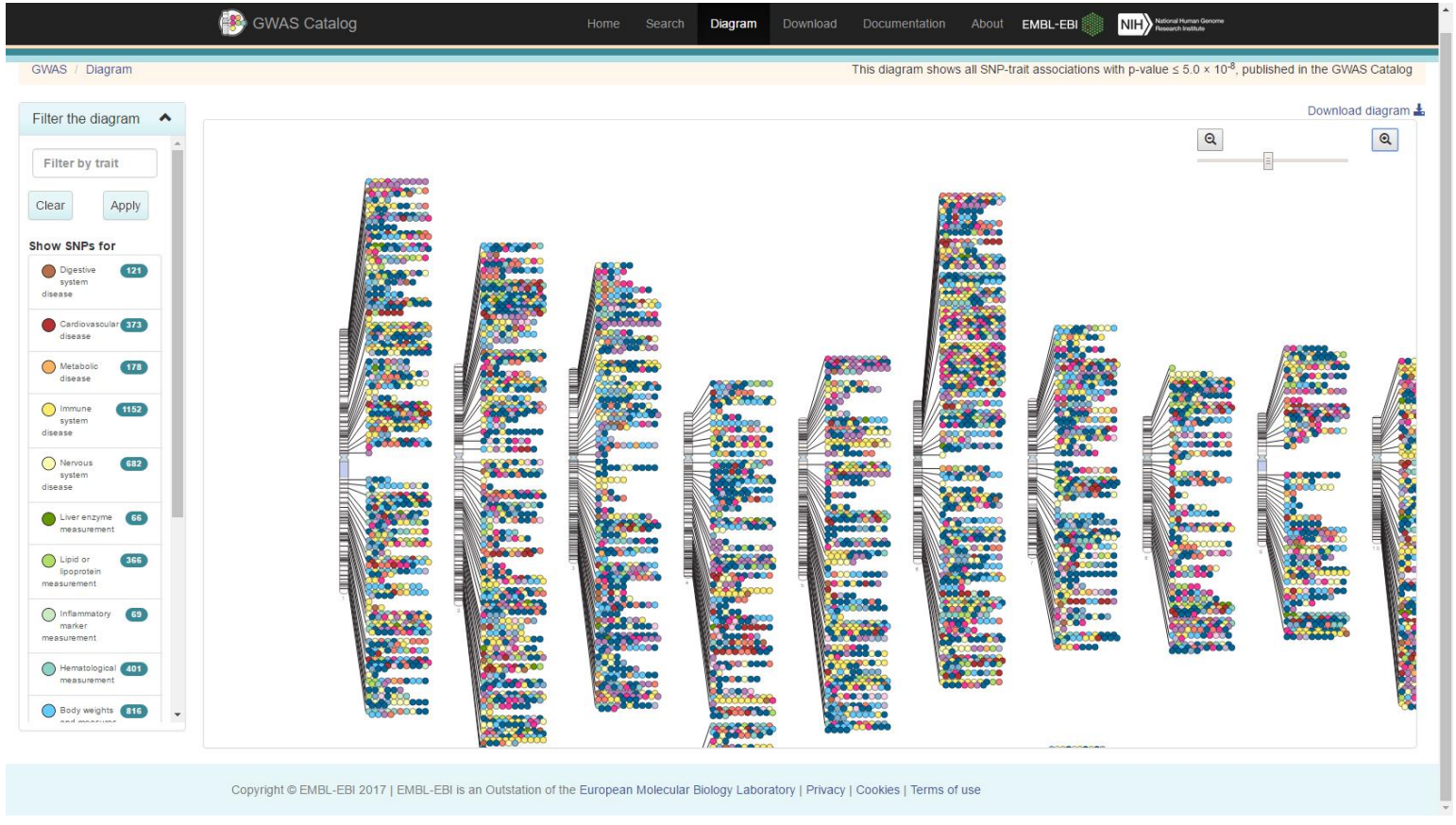
Lambert et al Nat Genet 2014

Replication

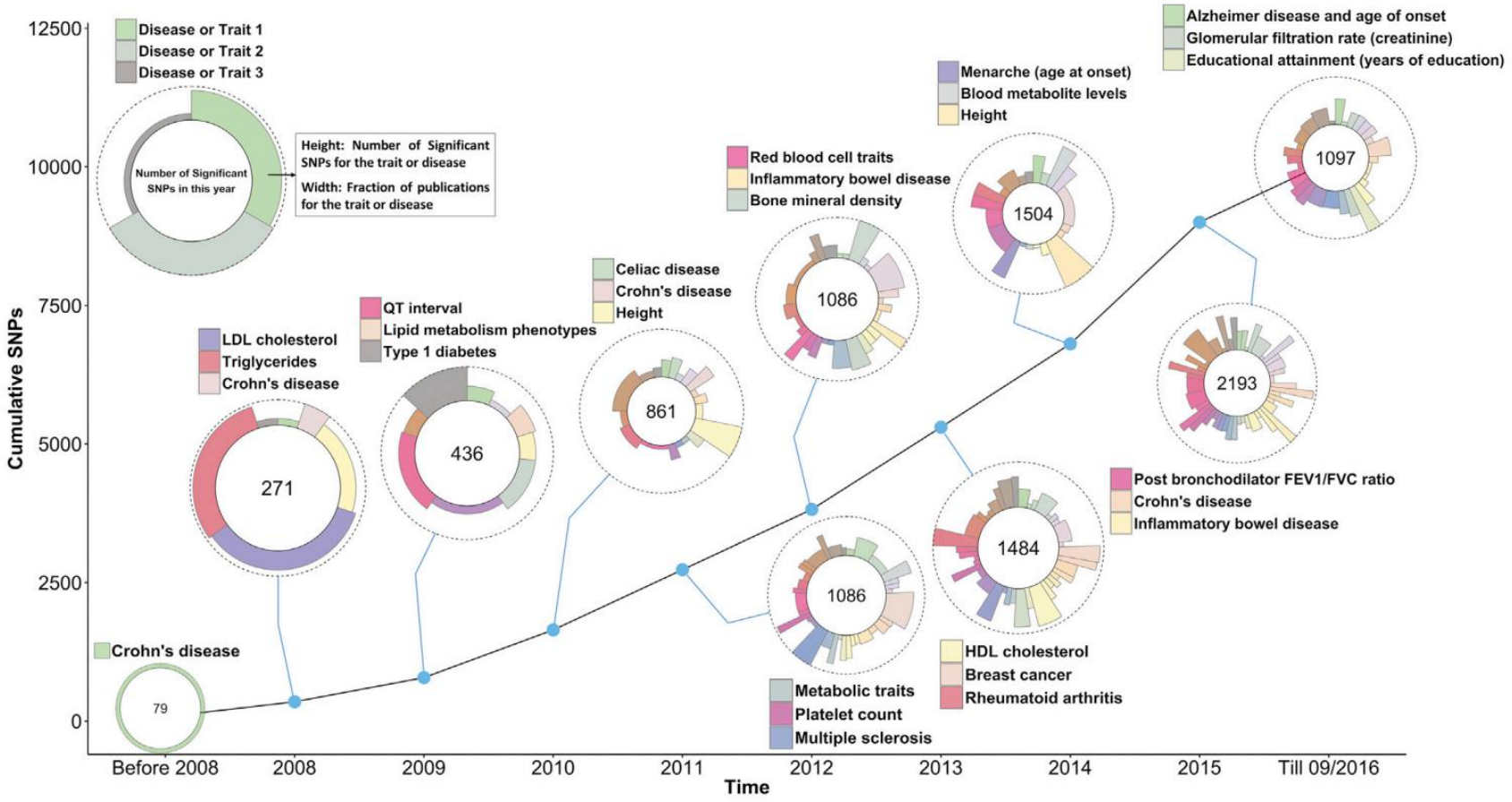
- Without replication no one will believe you
  - Same SNP
  - Same allele
  - Same phenotype
  - Same genetic model
- Considerations
  - Often ≥2 large replications required nowadays: collaboration is key
  - Replication in cohorts that do not have GWAS data available

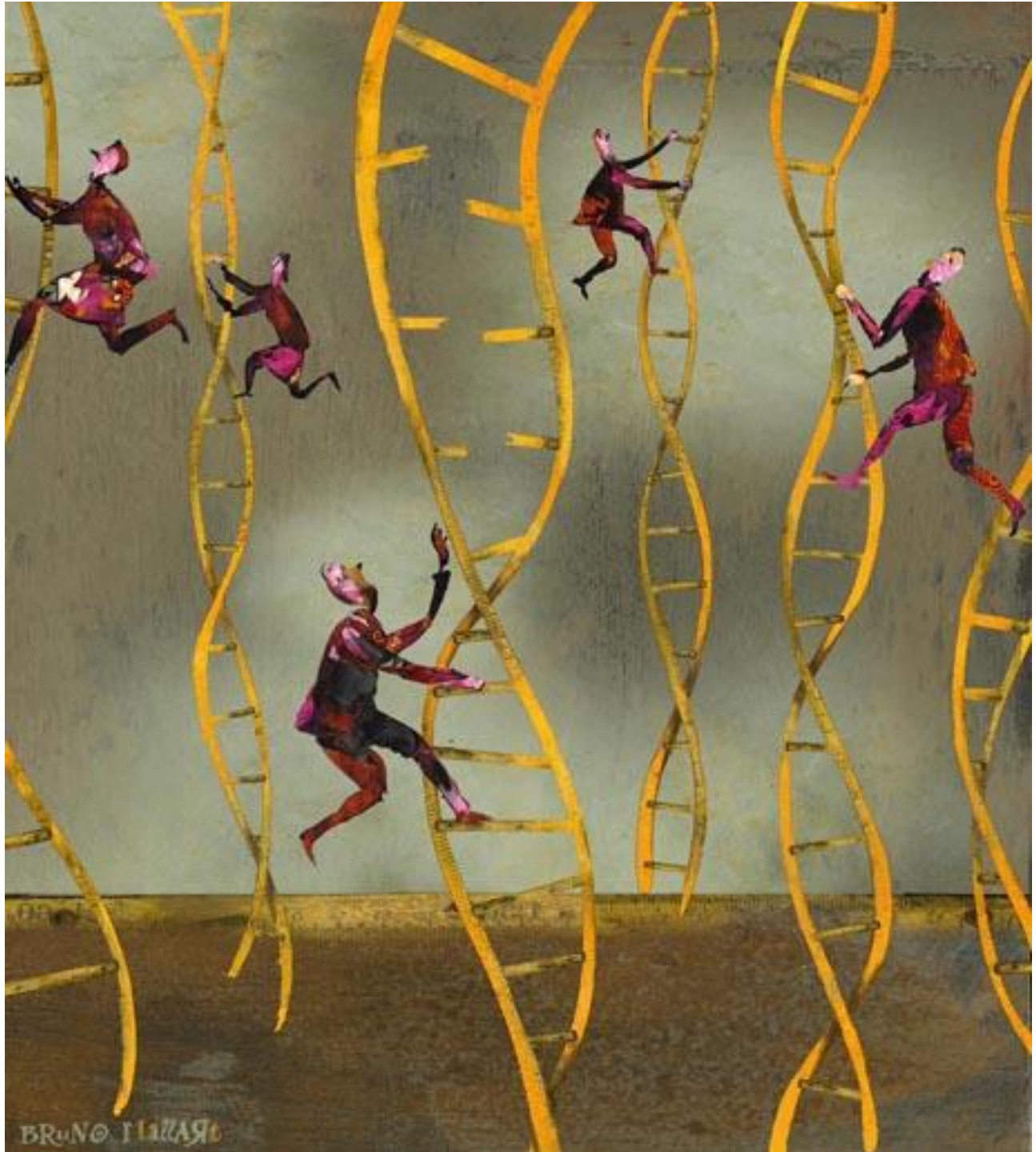
A Catalog of GWAS

- Database of all GWAS results https://www.ebi.ac.uk/gwas/
- 19/06/2017: 2982 publications and 36,948 unique SNP-Trait associations $P<10^{-8}$



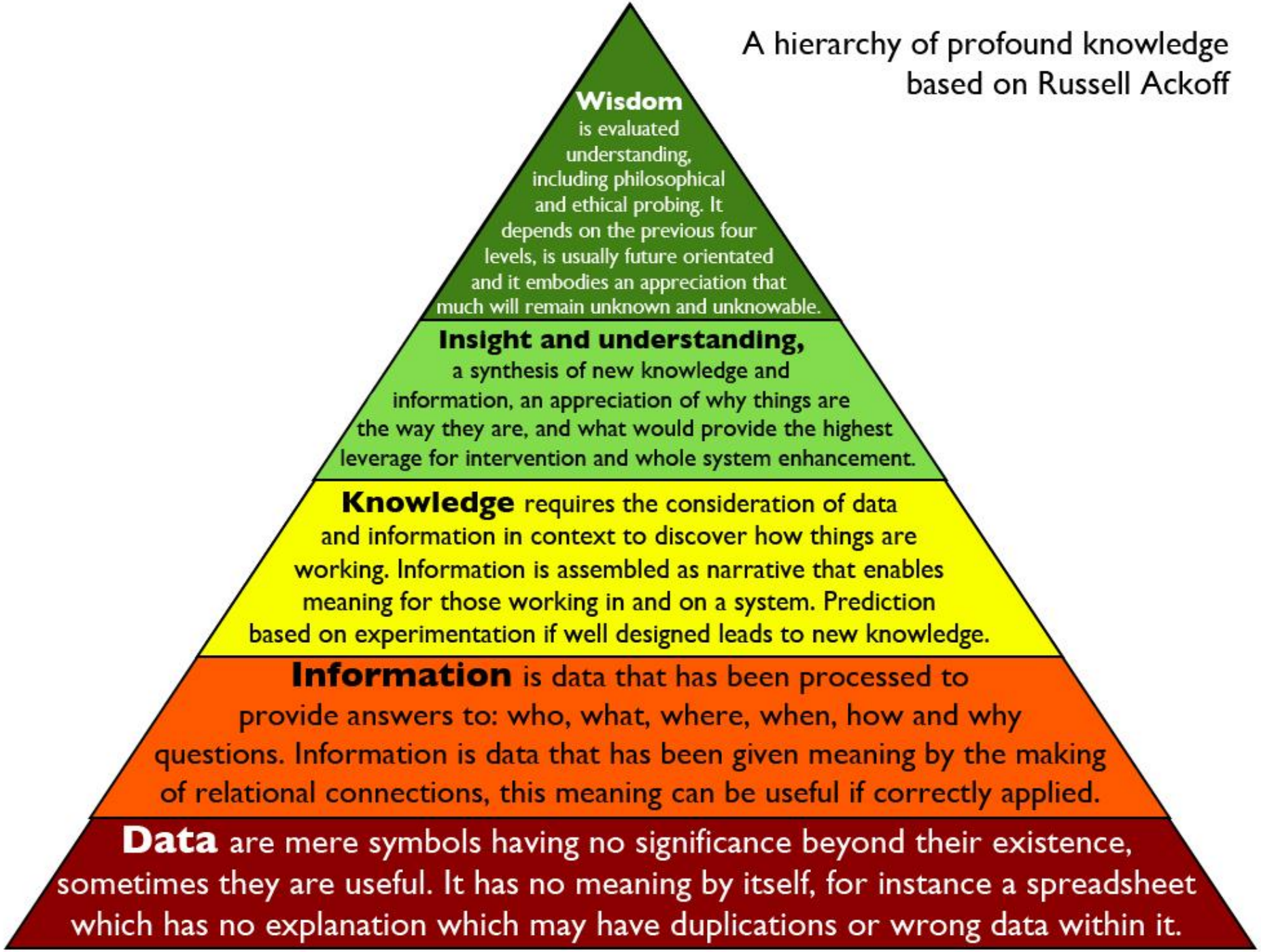
GWAS SNP-Trait discovery Timeline





BREAK BREAK BREAK BREAK BREAK

A hierarchy of profound knowledge
based on Russell Ackoff



Learning goals

- After this fourth part you are able to
  - Explain potential factors contributing to the missing heritability of traits
  - Understand that genetic loci associated with disease are no good predictors for disease
  - Provide an outlook how mechnisms underlying GWAS results may be investigated to establish causality

What we do not know after al those GWASes

- Missing heritability

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration[72]	5	50%
Crohn's disease[21]	32	20%
Systemic lupus erythematosus[73]	6	15%
Type 2 diabetes[74]	18	6%
HDL cholesterol[75]	7	5.2%
Height[15]	40	5%
Early onset myocardial infarction[76]	9	2.8%
Fasting glucose[77]	4	1.5%

TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

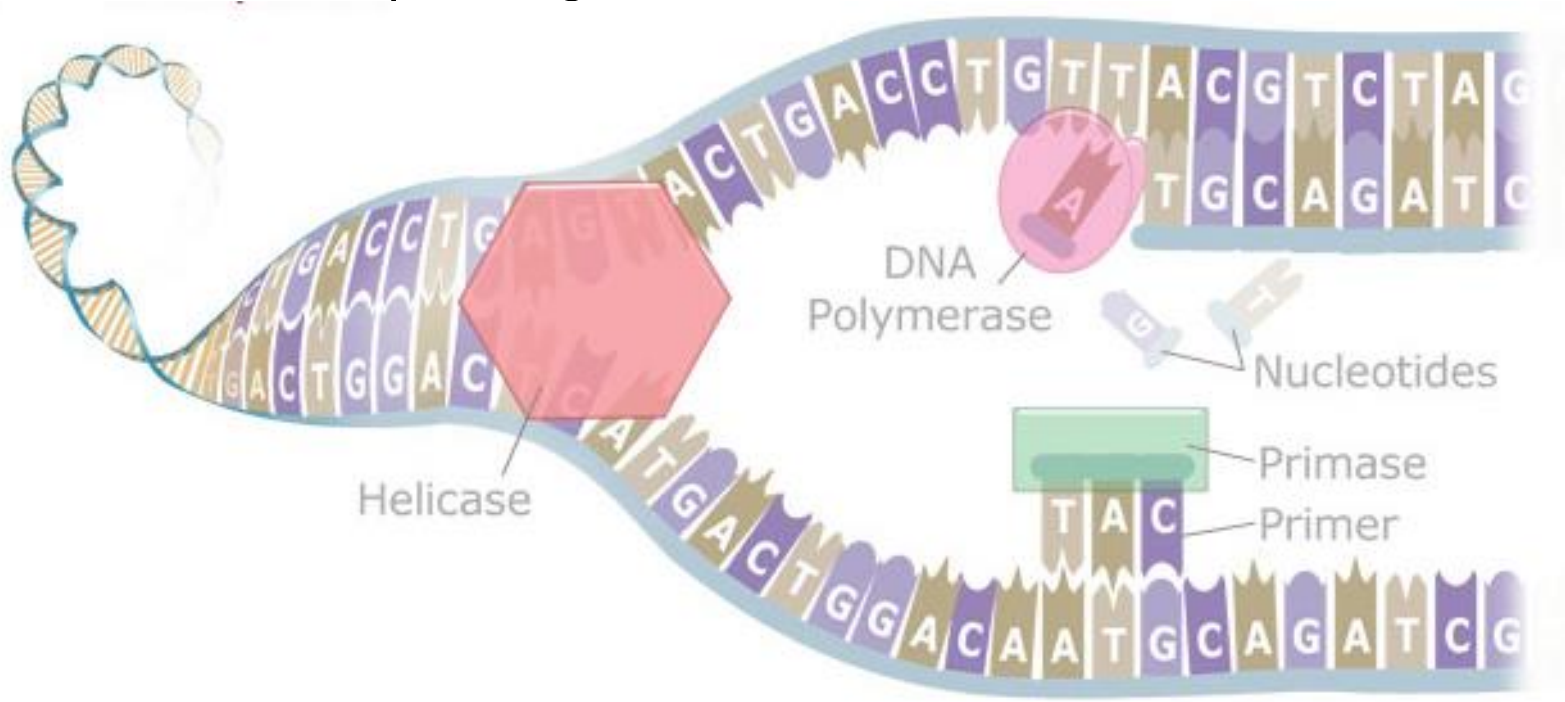
Incomplete detection genetic variation

- Factors
  - Common SNPs missed with (previous) arrays
  - Structural variation, copy number variants, in/dels
    - Much fewer known than SNPs

LV Wain *et al. The Lancet* **374** (9686):340-350 (2009)

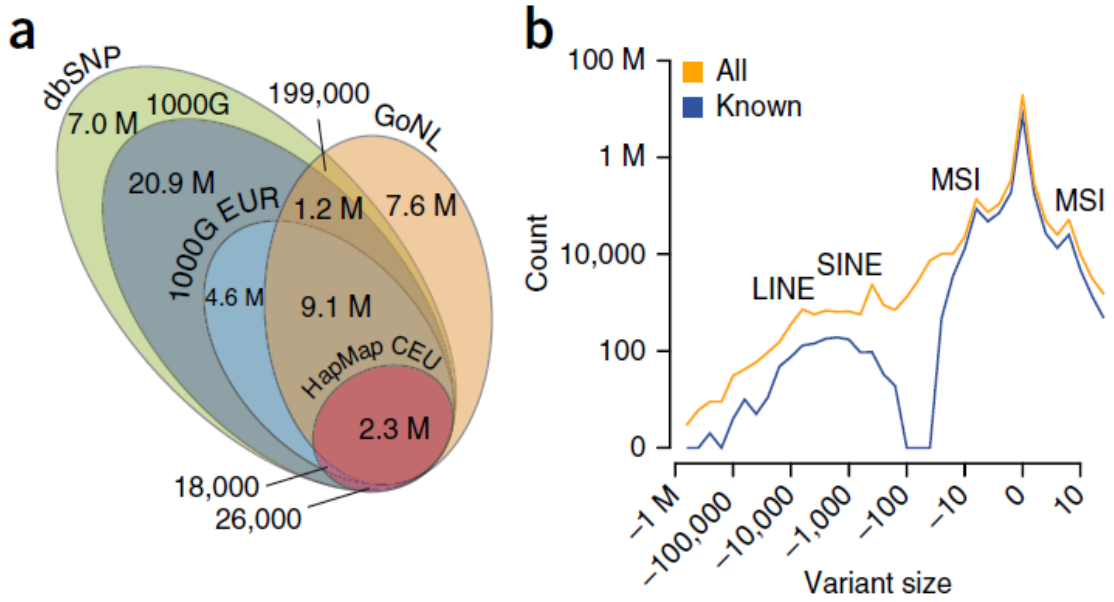
Incomplete detection genetic variation

- Factors
  - Common SNPs missed with (previous) arrays
  - Structural variation, copy number variants, in/dels
    - Much fewer known than SNPs
  - Rare variants: Sequencing



Current state-of-the-art

- Genome of The Netherlands (Go.NL)
  - 769 samples
  - ~20M SNPs
  - Insertion and deletions (Indels)



L Francioli *et al. Nature Genetics* (2014)

Prediction by single SNP

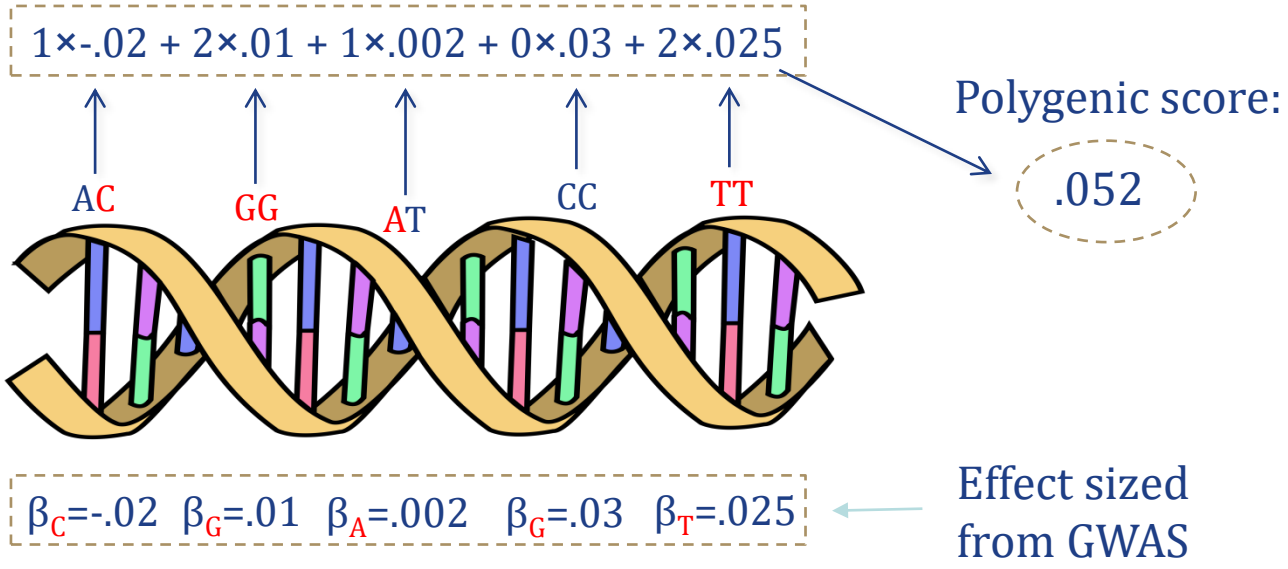
- Limited or no prediction despite biological insight

	Region	Candidate gene(s)	Weight†	Reference†	Risk allele	Risk allele frequency	Other allele	Coronary heart disease (total n=19790)	
								Pooled HR (95% CI)‡	p value
rs17465637	1q41	*MIA3*	1·14	15	C	0·75	A	0·99 (0·87–1·12)	0·854
rs11206510	1p32	*PCSK9*	1·15	15	T	0·84	C	0·94 (0·81–1·09)	0·431
rs646776	1p13	*CELSR2–PSRC1–SORT1*	1·19	15	T	0·79	C	0·96 (0·84–1·09)	0·512
rs6725887	2q33	*WDR12*	1·17	15	C	0·11	T	1·14 (0·96–1·35)	0·126
rs9818870	3q22	*MRAS*	1·15	16	T	0·10	C	0·88 (0·73–1·06)	0·174
rs3798220	6q26	*LPA*	1·68	18	C	0·01	T	2·07 (1·39–3·09)	$3·8\times10^{-3}$
rs9349379	6p24	*PHACTR1*	1·12	15	C	0·44	T	1·16 (1·04–1·29)	0·008
rs4977574	9p21	*CDKN2A–CDKN2B*	1·29	15	G	0·43	A	1·21 (1·08–1·34)	0·001
rs1746048	10q11	*CXCL12*	1·17	15	C	0·84	T	1·13 (0·97–1·33)	0·113
rs2259816	12q24	*HNF1A*	1·08	16	T	0·36	G	1·02 (0·91–1·14)	0·774
rs3184504	12q24	*SH2B3*	1·13	17	T	0·40	C	1·03 (0·92–1·15)	0·568
rs1122608	19p13	*LDLR*	1·15	15	G	0·79	T	1·00 (0·87–1·14)	0·988
rs9982601	21q22	*SLC5A3–MRPS6–KCNE2*	1·20	15	T	0·14	C	1·29 (1·07–1·57)	0·009

Ripatti *et al. Lancet* (2010)

Prediction by Polygenic Risk Score

- Polygenic Scores capture (part of) someone's genetic "risk" by summing all risk alleles weighted by the effect sizes estimated in a Genome-Wide Association Study (GWAS)



Prediction by Polygenic Risk Score

- By summing the collective effect sizes of many SNPs you can quantify part of the genetic “risk” in an **independent** dataset
- Polygenic Scores generally improve when adding SNPs that individually didn’t reach genome-wide significance

Prediction by Polygenic Risk Score

NATURE GENETICS LETTERS

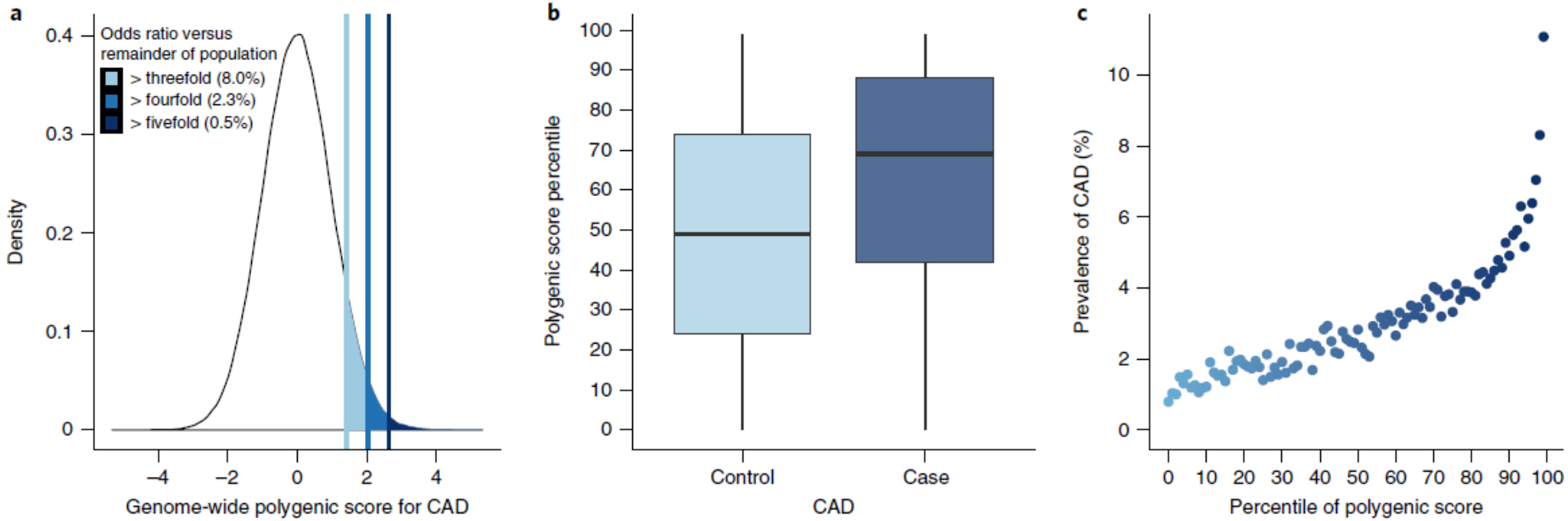
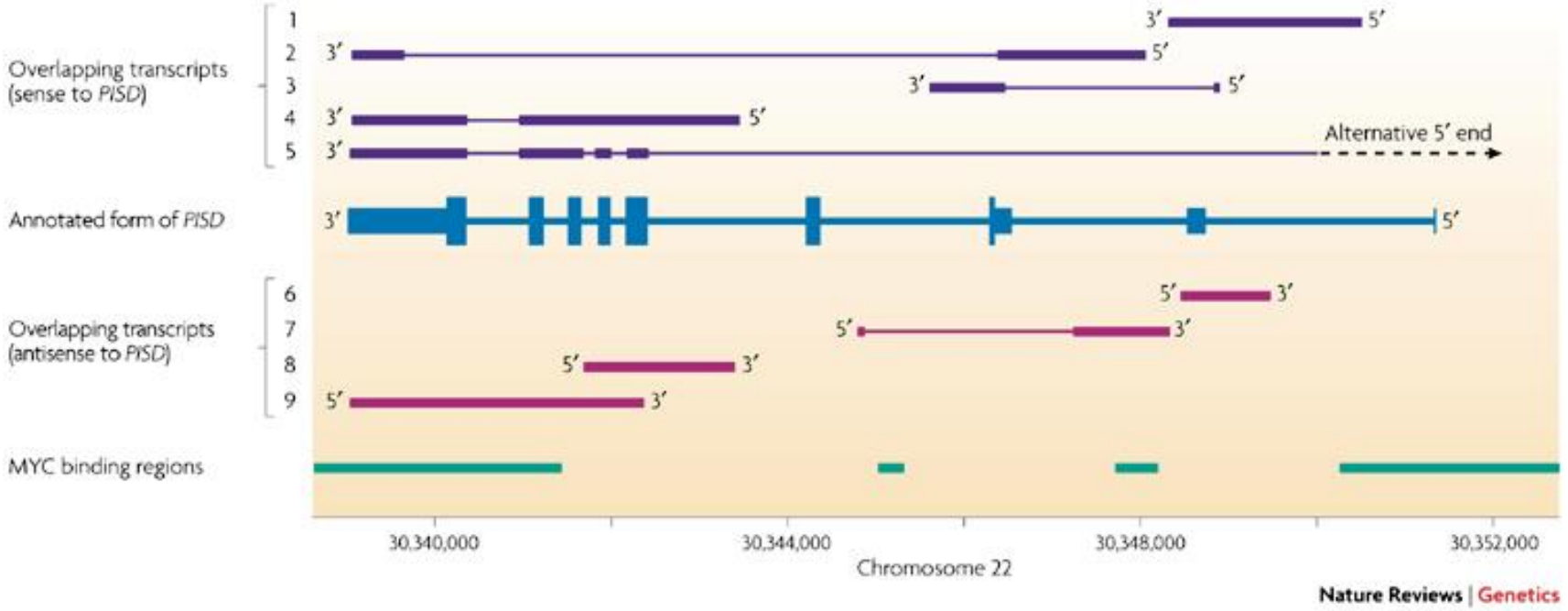


Fig. 2 | Risk for CAD according to GPS. a, Distribution of GPS_{CAD} in the UK Biobank testing dataset ($n = 288,978$). The *x* axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, GPS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping. **c**, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the GPS_{CAD}.

A gene is not what it used to be

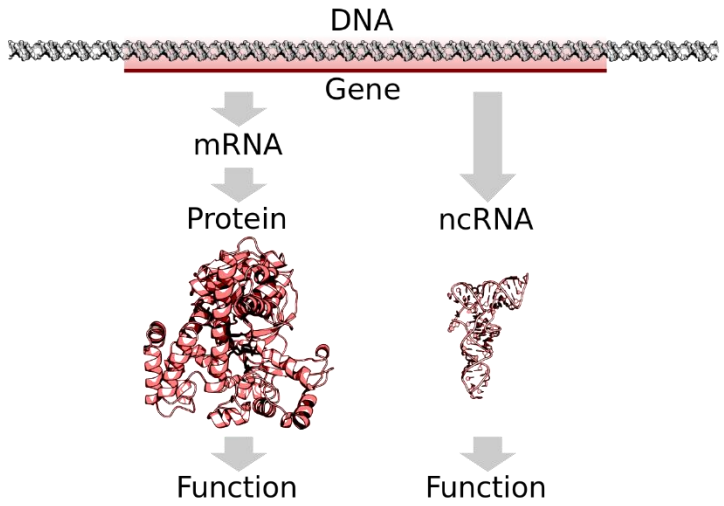
- Multiple use of same sequence
  - Phosphatidylserine decarboxylase (PSID)



Kapranov et al. Nat Rev Genet 2007 & ENCODE publications

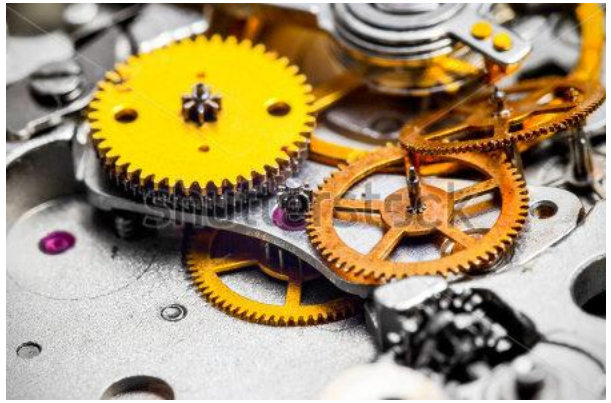
A gene is not what it used to be

- Up to 90% of genomic DNA is transcribed
  - 1-2% encodes exons, ~15% exons+introns
- Alternative initiation of trancription: ~60%
  - Alternative TSS 10s-100s kb away
  - Encode: 90% of genes have unannotated exon/TSS
- Alternative splicing: 60%
- Transcripts with anti-sense counterpart: 60%
- Alternative polyadenylation
- Gene fusions
- Trans splicing
- RNA synthesis at enhancers (eRNA)
- Relevant transcription factor binding occurs anywhere



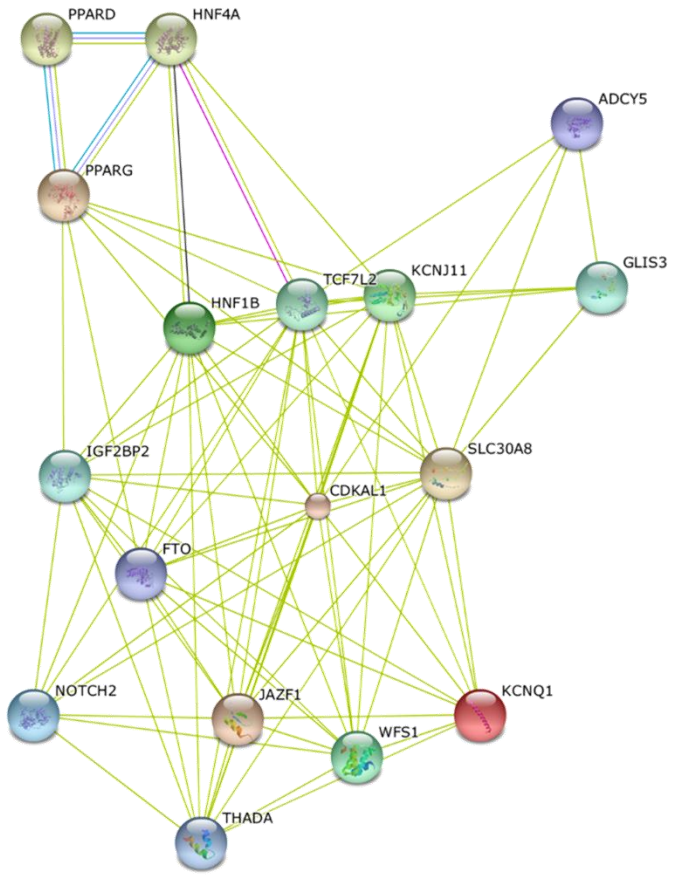
Underlying Mechanism

- Pathway analyses:
- DAVID
  - **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery
  - Gene Functional Classification Tool Based on Gene Ontology
- STRING
  - **S**earch **T**ool for the **R**etrieval of **I**nteracti**ng** **G**enes/Proteins
  - Known and predicted protein-protein interactions
- DAPPLE
  - **D**isease **A**ssociation **P**rotein-**P**rotein **L**ink **E**valuator
  - Physical connectivity among proteins encoded for by genes according to protein-protein interactions reported in the literature.



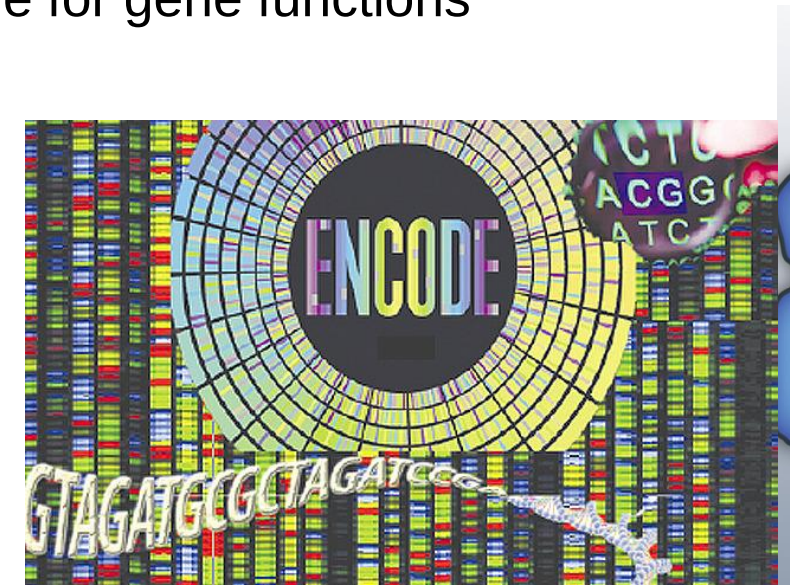
Underlying Mechanism

STRING



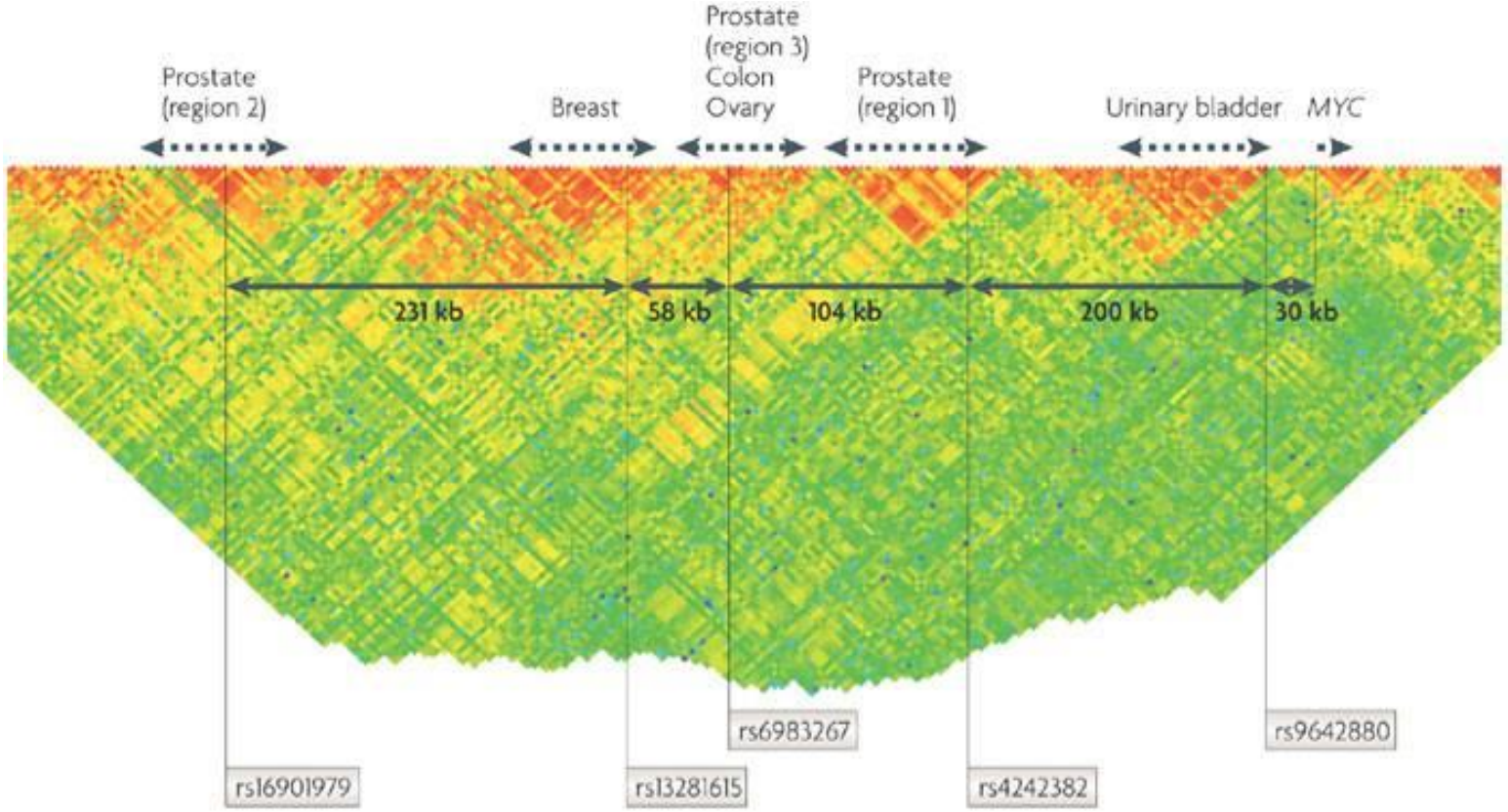
Causality

- Combine knowledge from public databases using USCS genome browser
  - Databases with biological knowledge
    - ENCODE
    - Haploreg
  - Databases with functional knowledge
    - GTEX for eQTLs
    - Entrez Gene for gene functions



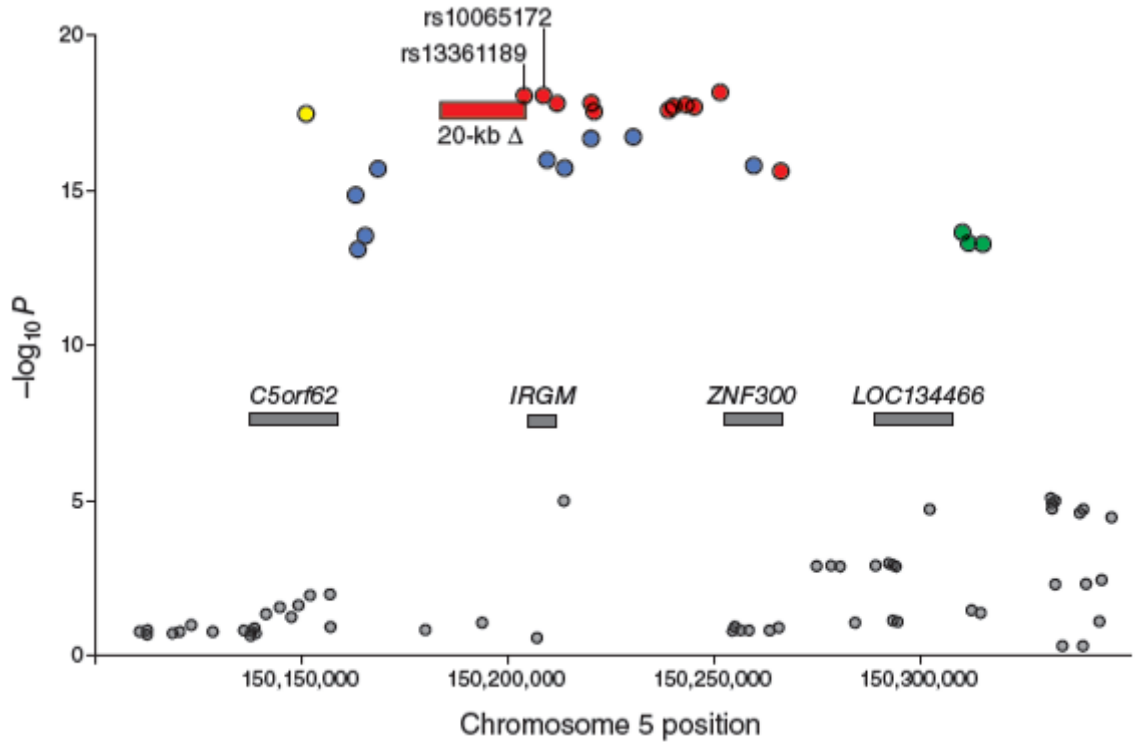
Truly novel findings

- ‘Gene deserts’
  - 8q24 region confers susceptibility to various cancers
  - Functional studies are very much required



Proven causality

- Rarely causality is proven or even plausible



Brest et al Nat Genet 2011. Locus associated with Crohn's Disease: a synonymous variant in the *IRGM* coding region alters a binding site for miR-196 and modulates IRGM-dependent autophagy.

Challenges in GWAS studies

- Weaknesses of designs: cross-sectional, broad phenotypes
- Complex effects: interactions, allele specific effects
- Incomplete detection of genetic variation

Complex effects

- Interaction
  - Interactions within biomolecular networks: rare combinations of common variants (epistasis – as routinely seen in yeast)
- Allele specific effects
  - Estimate: ~15% of type 2 diabetes heritability due to known variants involves parent-of-origin effects.

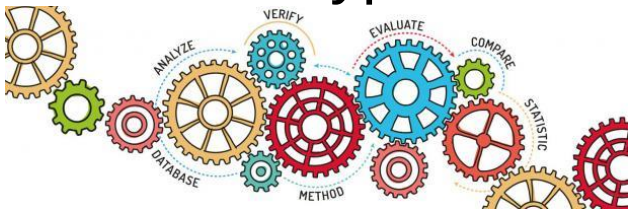
Table 1 | Parental-origin-specific analyses of disease-susceptibility variants

Disease, SNP [alleles]*		Standard case-control test		Tests of association with parental origins						
NCBI build 36 position, *N*	*M*, F_{con}	OR	*P*‡	Paternal allele§		Maternal allele§		2-d.f. test‖	Paternal vs maternal (case only)	
				OR	*P*	OR	*P*	*P*	n12:n21¶	*P*
T2D, rs2334499 [T/C]										
C11 1,653,425,	34,706,									
1,468 (discovery)	0.412	1.11	0.017	1.41	4.3×10^{-9}	0.87	0.020	3.5×10^{-9}	437:276	7.0×10^{-9}
783 (replication)		1.02	0.71	1.23	0.0055	0.84	0.023	0.0018	222:157	8.0×10^{-4}
2,251 (combined)		1.08	0.034	1.35	4.7×10^{-10}	0.86	0.0020	5.7×10^{-11}	659:433	4.1×10^{-11}

Kong et al. Nature 2009

Life after genome-wide studies

- Go big: meta-analysis
- Increase genetic detail (rare variants)
- Smarter clinical end-points
- Detailed intermediate phenotypes (biomarkers) and system approaches (vertical genomics)
- Pathway analyses
- Acquire biological knowledge from public databases
- Re-analysis publicly available data (e.g. interactions)
- Functional studies to prove causality (!)
- General: more hypothesis-driven, more depth



Upload answers to Brightspace, TurnItIn

Filename:*Yourname*_*YourStudentnr*_GWAS_2021