Norwegian Institute of Public Health

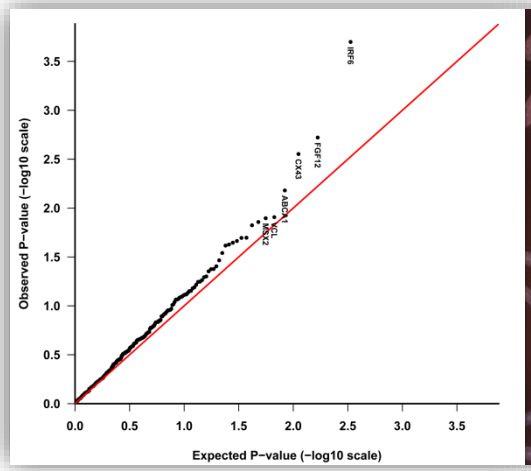
UNIVERSITY OF BERGEN

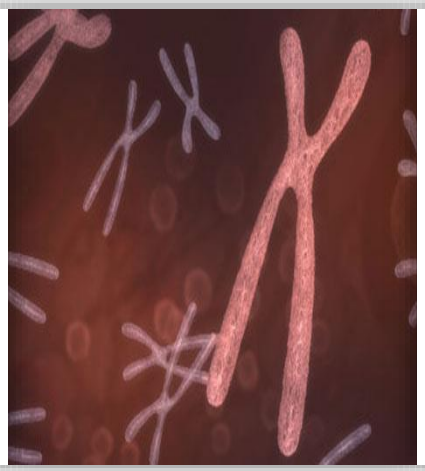
General Intro to Genetic Epidemiology

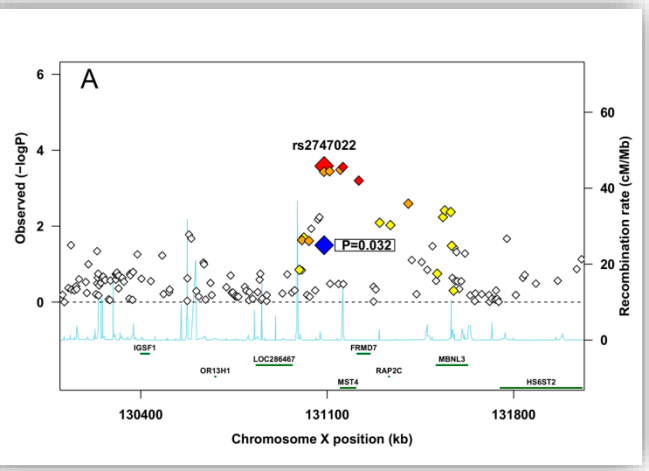
— Lecture 1, Parts I & 2 —

Anil Jugessur

Senior scientist, Norwegian Institute of Public Health, Oslo







NORBIS GENSTAT-B Course, 30th of May – 3rd of June, 2022

Lecture Outline

General introduction to genetic epidemiology (lecture I)

- Part I
  - What's a complex trait?
  - Genetic basis of complex traits
- Part II
  - Genetic approaches to studying complex traits
  - Candidate-gene analysis, GWAS, and GWAMA

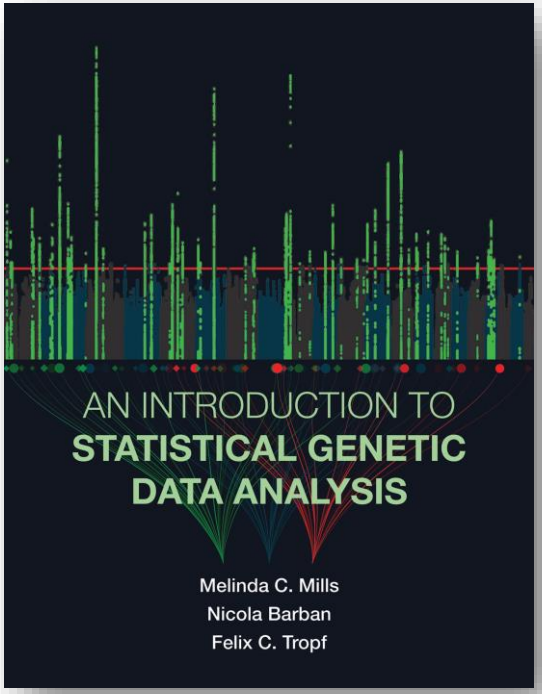
Lecture Outline

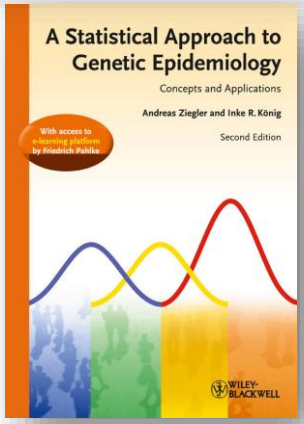
General introduction to genetic epidemiology (lecture I)

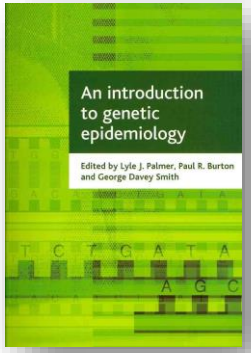
- Part I
  - What's a complex trait?
  - Genetic basis of complex traits
- Part II
  - Genetic approaches to studying complex traits
  - Candidate-gene analysis, GWAS, and GWAMA

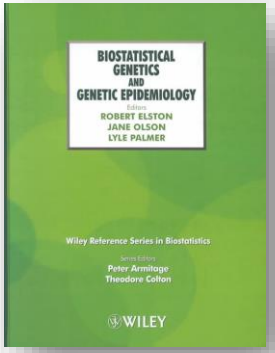
Course Book

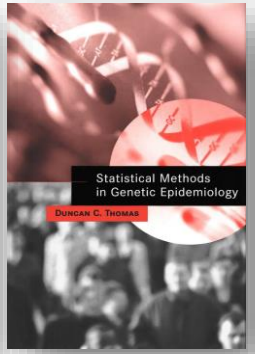
Course book

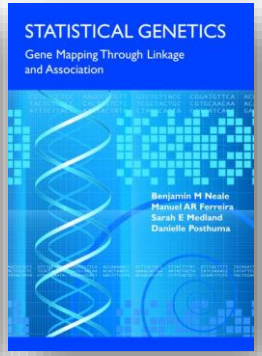


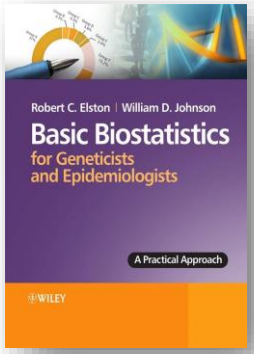


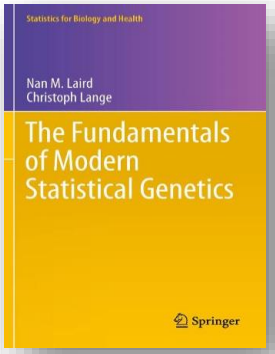


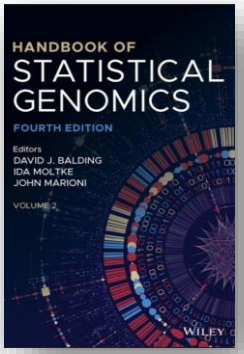


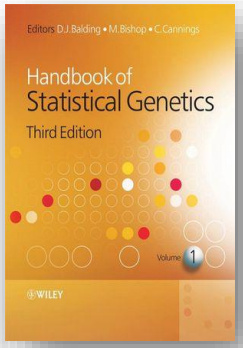








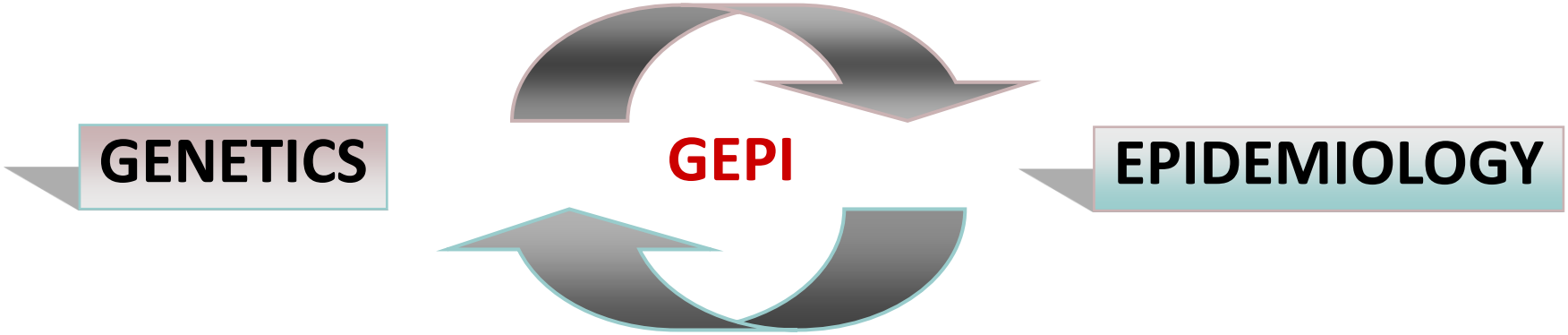




What Is Genetic Epidemiology?

In broad terms:

«The application of genetic principles and techniques to answering epidemiological questions»



Lots Of Definitions Out There…

Table 1–1. Some definitions of genetic epidemiology

N. E. Morton and C. S. Chung (1978): "A science that deals with the etiology, distribution, and control of disease in groups of relatives, and with inherited causes of disease in populations."

R. Ward (1979): "The primary objective of the genetic epidemiologist will be to identify the genetic contribution to the etiological pathway."

B. H. Cohen (1980): Genetic epidemiology is defined "as examining the role of genetic factors, along with the environmental contributors to disease, and at the same time, giving equal attention to the differential impact of environmental agents, nonfamilial as well as familial, on different genetic backgrounds."

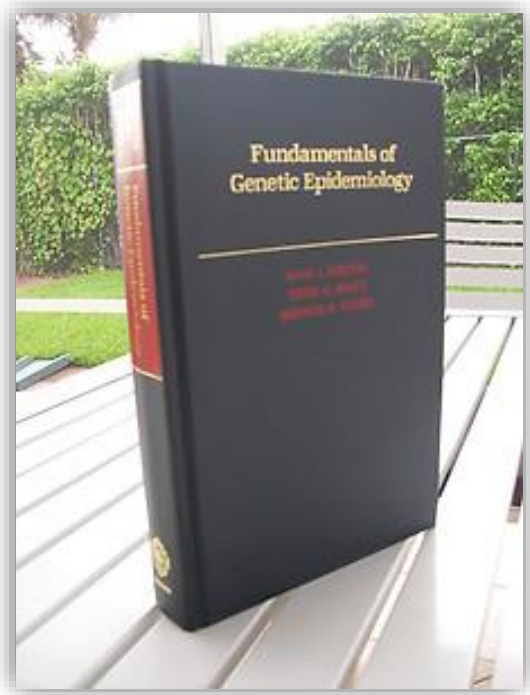
P. Phillippe (1982): "Genetic epidemiology studies the interaction between genetic and environmental factors at the origin of disease."

M.C. King et al. (1984): "Genetic epidemiology is the study of how and why diseases cluster in families and ethnic groups."

D.C. Rao (1984): "Genetic epidemiology is an emerging field with diverse interests, one that represents an important interaction between the two parent disciplines: genetics and epidemiology. Genetic epidemiology differs from epidemiology by its explicit consideration of genetic factors and family resemblance; it differs from population genetics by its focus on disease; it also differs from medical genetics by its emphasis on population aspects."

D.F. Roberts (1985): argues the distinction of genetic epidemiology from epidemiology in general. Genetic epidemiology "is not merely the application of the central concept of epidemiology, the study of the distribution of disease in space and time, to genetic disease. Instead, in genetic epidemiology, the concept is extended to include the additional variables of the genetic structure of the population, with the object of elucidating the etiology of disease in which there may be a genetic component."

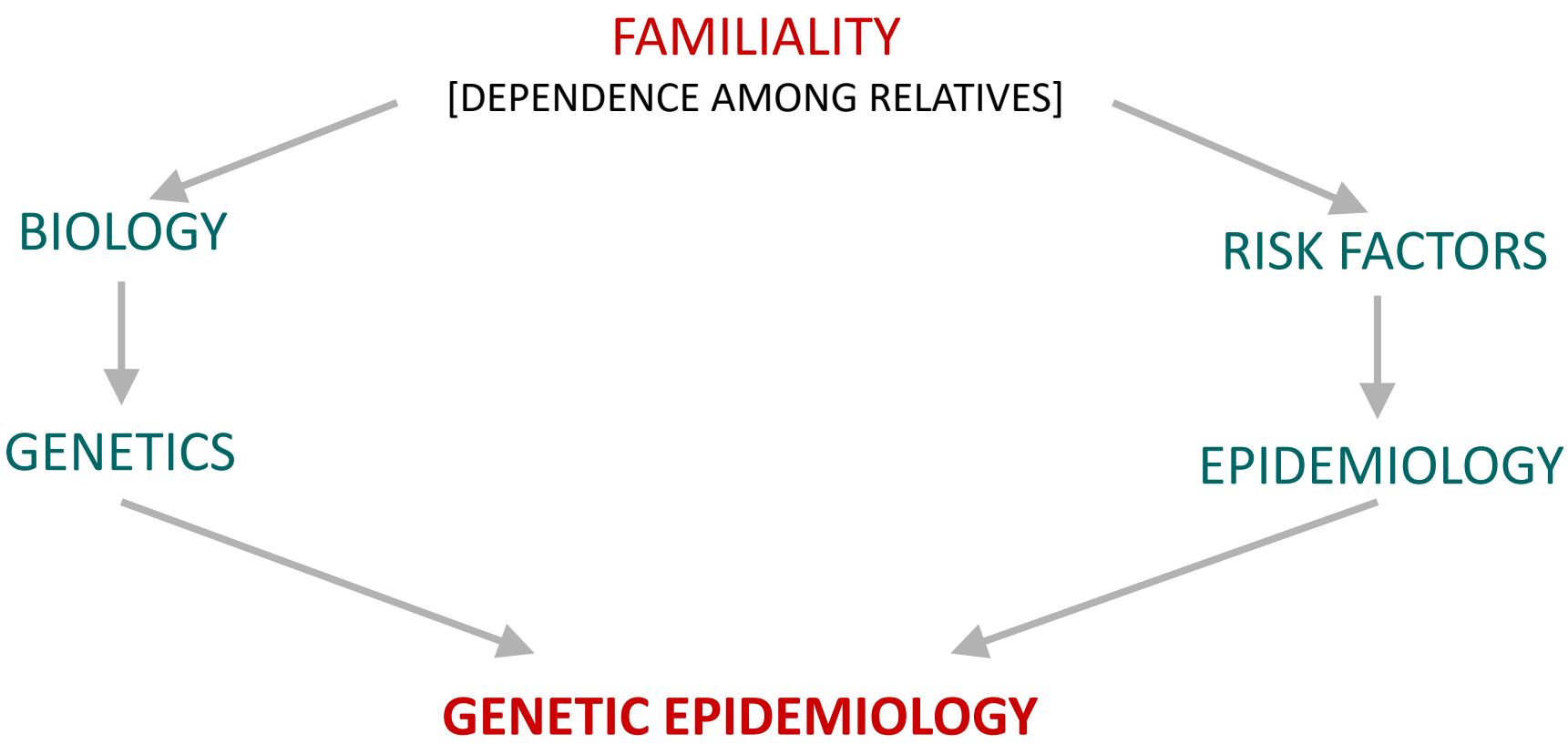
E.A. Thompson (1986a): "Genetic epidemiology is the analysis of the familial distributions of traits, with a view to understanding any possible genetic basis."



Prof of biostats @ WASH-U.

Source (book): «Fundamentals of Genetic Epidemiology», Eds. Khoury, Beaty and Cohen. Oxford Press, 1993.

Growing Convergence Of Different Fields



"Less divergence in terminology and methodology, but an increased conversation, collaboration and convergence across the fields."

Building A Team

– E.G. For a Study of Birth Defects –

TEAM

CLINICIANS
- MEDICAL GENETICIST
- DYSMORPHOLOGIST
- PLASTIC SURGEON
- SPECIALIST NURSE
- SPEECH PATHOLOGIST
- NUTRITIONIST
- GENETIC COUNSELOR
- PEDIATRIC DENTIST

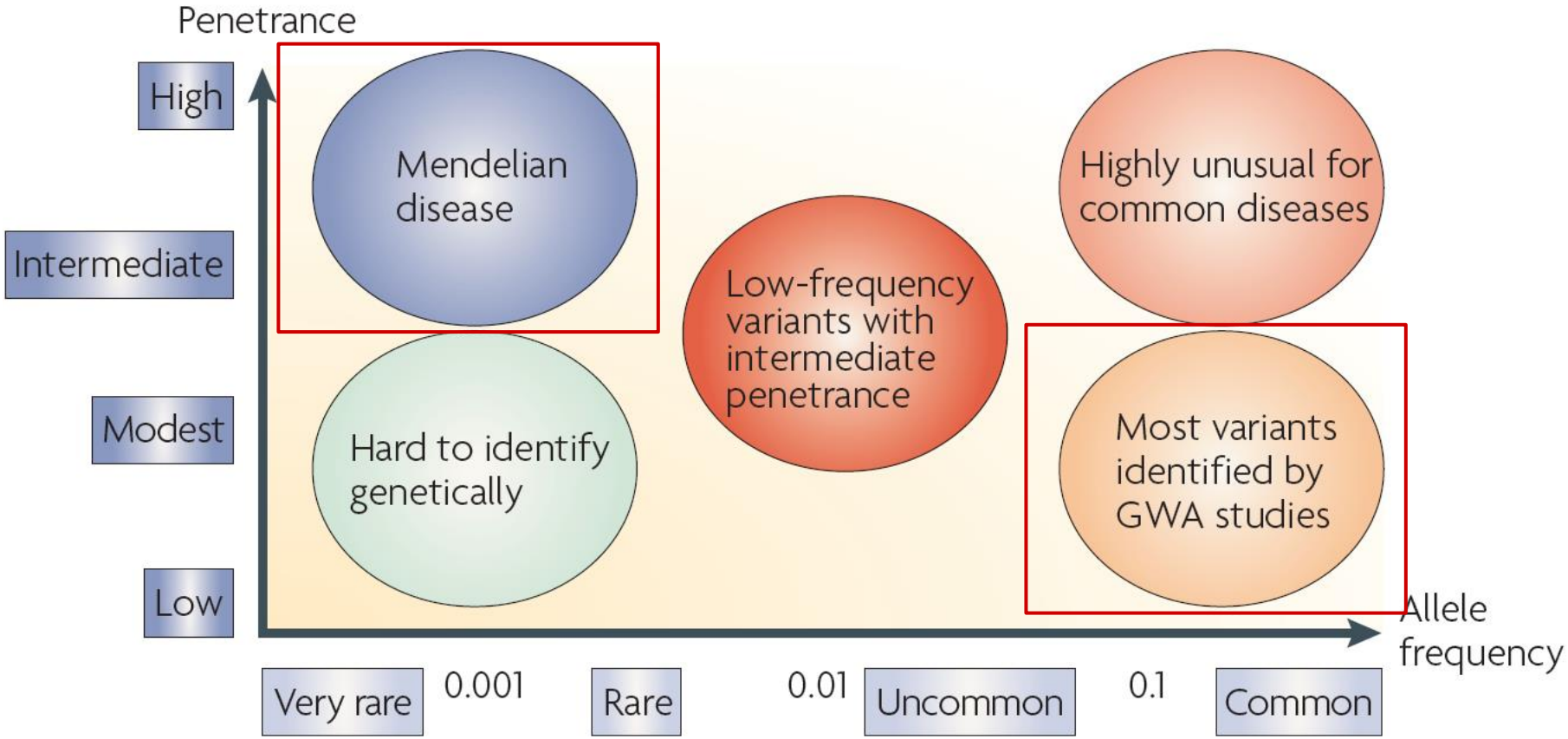
DATA ANALYSTS
- EPIDEMIOLOGIST
- STATISTICAL GENETICIST
- BIOSTATISTICIAN
- BIOINFORMATICIAN

COMPUTATIONAL SKILLS
- IT-EXPERT
- DATABASE MANAGER
- DATAMINER

CORE BIOLOGISTS
- HUMAN MOLECULAR GENETICIST
- DEVELOPMENTAL BIOLOGIST
- MOUSE GENETICIST

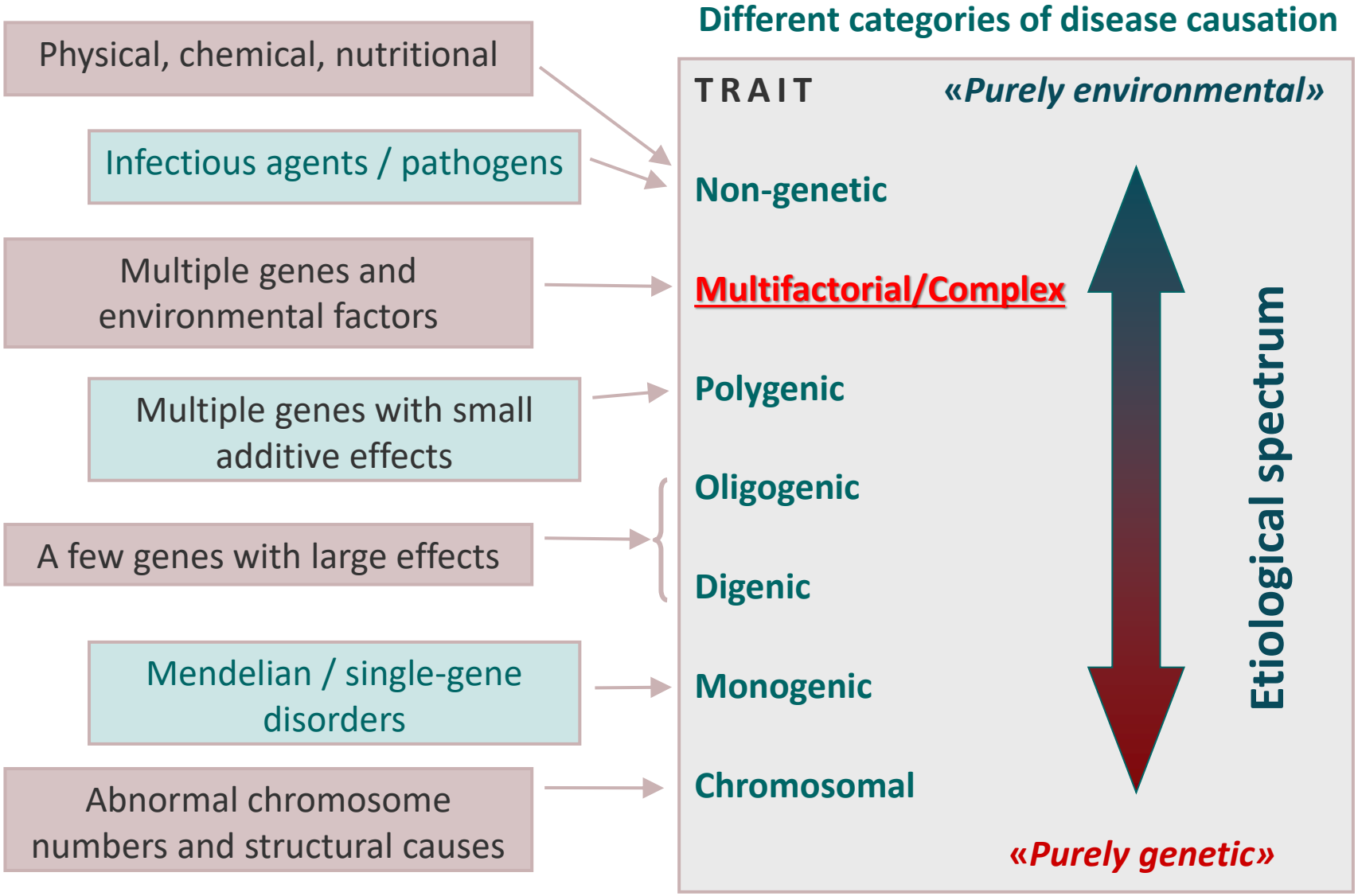
MANY MORE!!
- FIELD WORKERS
- VOLUNTEERS
- NGOs

Mendelian disorder vs. Complex trait



Reference: Mark McCarthy *et al.* Nat Genet Rev | 356 | May 2008 | volume 9

What's A Complex Trait?



Source: Haines & Pericak-Vance (1998). In the book: ***Approaches to genetic mapping in complex human diseases.*** Wiley-Liss, Inc.

Common Features Of Complex Traits

- **Unlike Mendelian diseases, complex traits are relatively common**
- **Heterogeneity at several levels:**
  - **Genetic heterogeneity:**
    - **«*locus*» and «*allelic*» heterogeneity**
- **Incomplete penetrance** ⇒ *not all individuals with the mutant genotype express the phenotype*
- **Effect of a gene can be masked by:**
  - **Phenocopies** ⇒ *an environmentally-caused phenotype mirrors a genetically-caused trait*
  - **Pleiotropy** ⇒ *the mutant genotype affects different traits or organs*
- **Complex interactions:**
  - **«gene-gene» and «gene-environment» interactions**
- **Stochastic effects** ⇒ *random or chance events; biological processes are error-prone!*

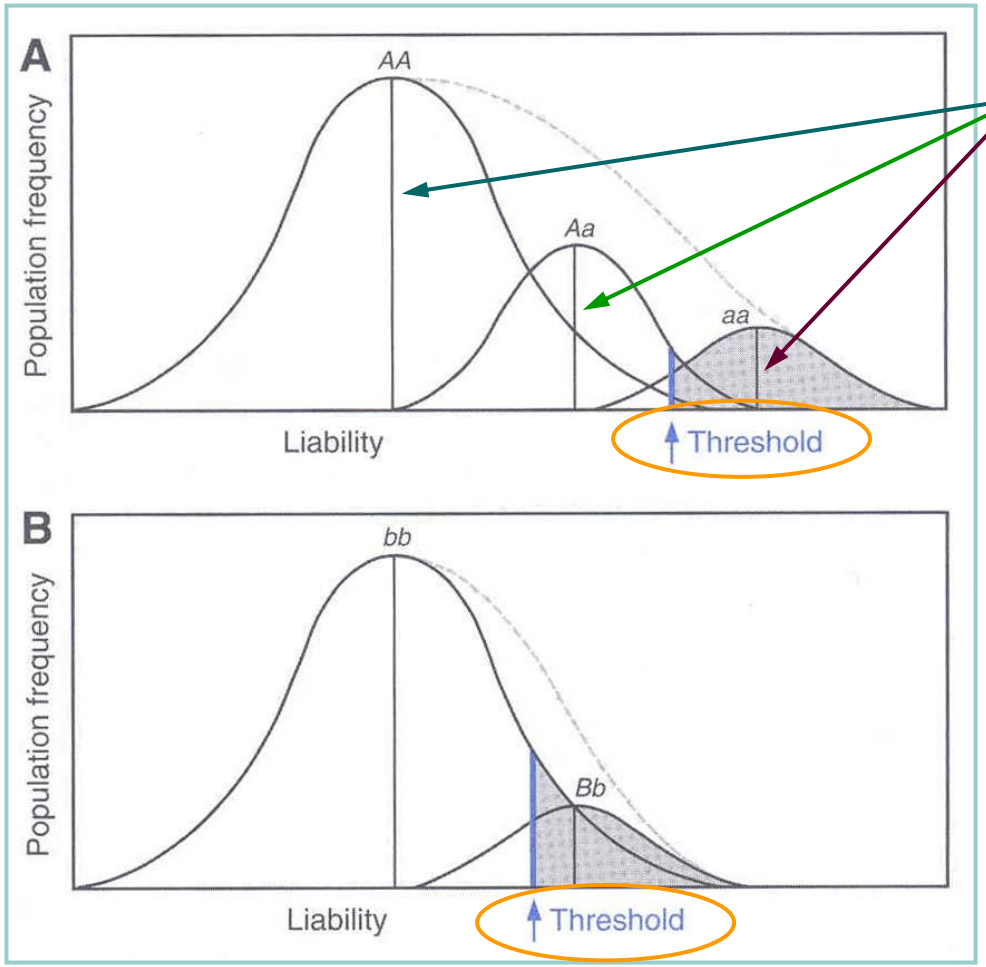
The Concept Of «Liability»

Liability is an underlying continuous variable comprising both genetic and non-genetic effects.

FIGURE: An idealized distribution of liability in individuals with various genotypes.

Recessive allele
'*a*' ↑ses liability

Dominant allele
'*B*' ↑ses liability



Mean liability for each genotype

Threshold = value in the liability that determines whether a disease will be expressed or not.

Anyone with liability greater than the threshold manifests the disease.

Source: Maroni (2001). In ***Molecular and genetic analysis of human traits.*** Blackwell Science, Inc.

The Concept Of «Heritability» - Ch. 1

- Heritability (H^2) is the proportion of phenotypic variance (*Vp*) attributable to genetic differences.

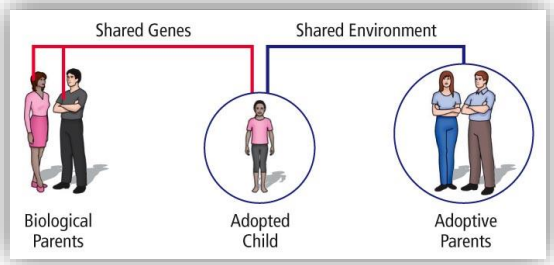
- Broad-sense vs. narrow-sense heritability
  - Broad-sense heritability is the proportion of variance in a phenotype (*Vp*) attributable to the total genetic variance (*Vg*). $H^2=Vg/Vp$, where $Vp=Vg+Ve$
  - Narrow-sense heritability is the proportion of *Vp* attributable to additive genetic variance (*Va*); i.e., H^2=Va/Vp

- Additive vs. non-additive genetic effects
  - Additive effects: 2 or more genes contribute to a phenotype, or when alleles in a single gene combine so that their combined effects on the phenotype are equal to the sum of their individual effects.
  - Non-additive effects can be dominance (*Vd*) or epistasis (*Vi*)
    - Dominance: The effect of one allele masks the effect of a second allele at the **same locus**; e.g., allele *A* dominates allele *a*.
    - Epistasis: An allele at one locus affects the expression of another allele at a **different locus**.

Is There A Genetic Basis To Complex Diseases?

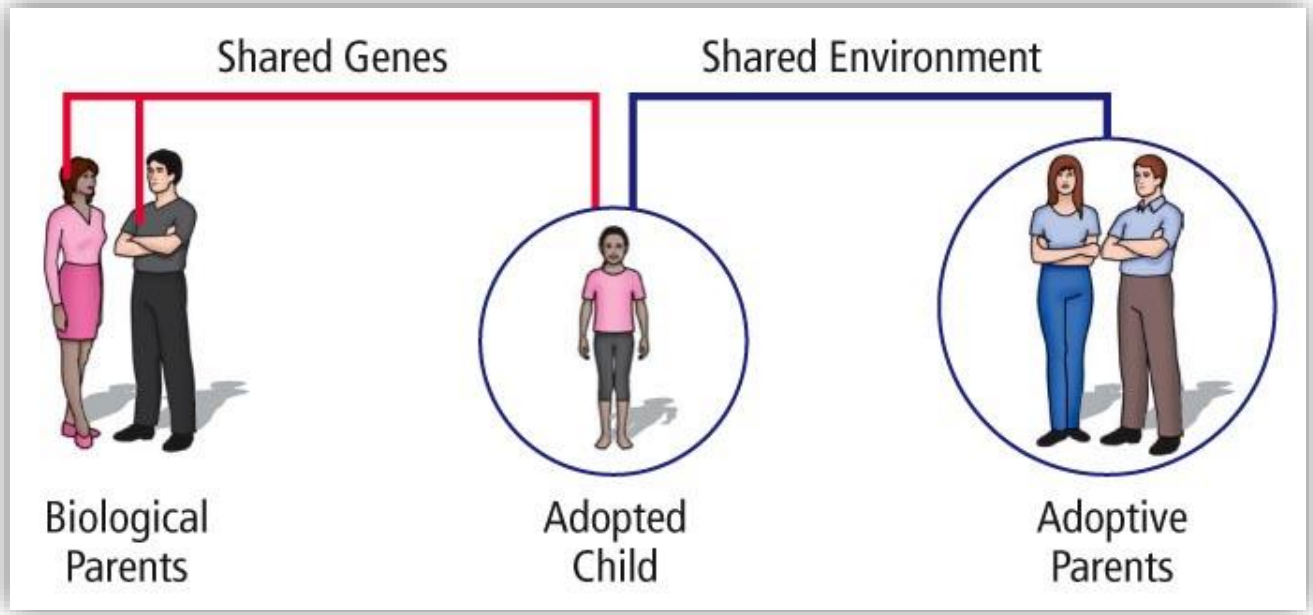
- Study whether the disease clusters in families:
  - Familial aggregation studies:
    - Relatives share a greater proportion of their alleles
      - Affected individuals will tend to cluster in families.
    - Reccurence risk measured as relative risk ratio (λ_r)
      - **λ_r = [risk to relatives of type *r*] ÷ [Population risk]**
    - Cannot establish that the disease is hereditary
      - Environmental factors could also cause this clustering!

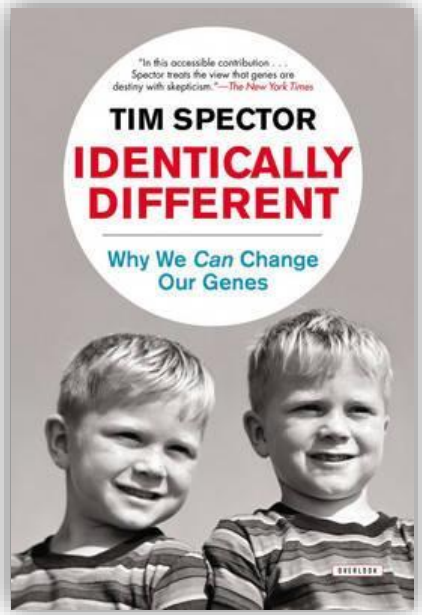
  - Adoption studies:
    - If a trait has a genetic influence, the risk of disease should be higher in biological relatives than in adopted relatives living in the same household.



  - Twin studies:
    - Compare concordance i MZ vs. DZ twins
      - If MZ twins show close to 100% concordance but DZ twins show significantly less: ⇒ the trait has a strong genetic basis.
      - If MZ twins shows moderate concordance (40-60%) but still significantly higher than DZ twins ⇒ both environmental and genetic components are likely involved in the disease.

Importance of Shared Environment!





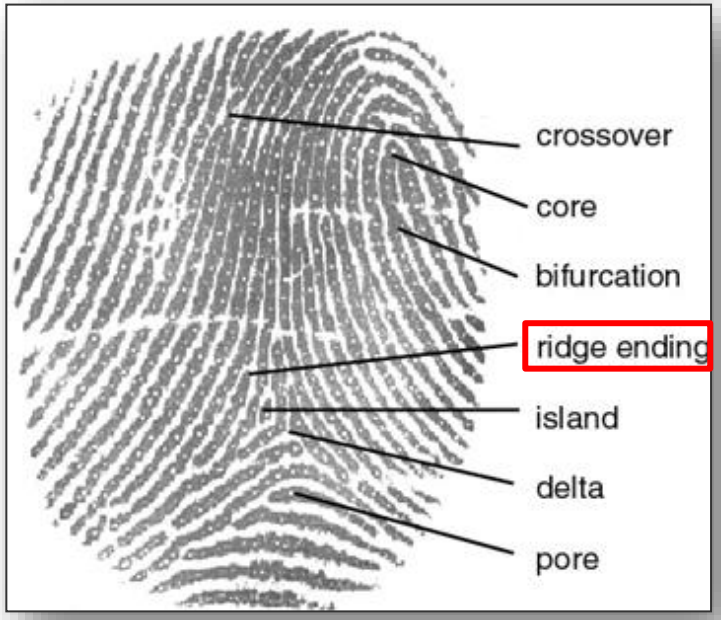
Assessing Evidence Of Familial Aggregation

Usual to look at two types of correlations between relative pairs:

- «INTER»class correlation
  - Involves two different classes of relatives:
    - E.g. husband-wife, parent-offspring, brother-sister, grandparent-grandchild, etc.
- «INTRA»class correlation
  - Involves only a single class of relatives:
    - E.g. brother-brother, sister-sister, etc.

An Example

Fingerprint data: count the number of ridges to explore degree of familiality.



2 scenarios:

- Dataset I:
  - **Parent-offspring** correlation: **0.48 ± 0.04**
  - **Sibling** correlation: **0.50 ± 0.04**
  - **Spouse** correlation: **0.05 ± 0.07**

- Dataset II:
  - **Parent-offspring** correlation **0.22 ± 0.01**
  - **Sibling** correlation: **0.39 ± 0.01**
  - **Spouse** correlation: **0.15 ± 0.02**

An Example – Contd…

- Dataset I:
  - **Parent-offspring** correlation: **0.48 ± 0.04**
  - **Sibling** correlation: **0.50 ± 0.04**
  - **Spouse** correlation: **0.05 ± 0.07**
- Dataset II:
  - **Parent-offspring** correlation **0.22 ± 0.01**
  - **Sibling** correlation: **0.39 ± 0.01**
  - **Spouse** correlation: **0.15 ± 0.02**

- Positive correlation coefficients suggest familial aggregation for this trait
- Strong degree of familiality in **Dataset I**.
  - Sibling correlation is slightly higher than parent-offspring correlation
    - Consistent with siblings sharing more of their environment than parents & offspring
  - Don't see same degree of correlation in the spouse group
    - Consistent with a less genetic sharing between spouses.
- In **Dataset II**, higher spouse correlation may be due to shared spousal environment (perhaps some assortative mating..?)
- Overall, there seems to be stronger environmental influences in Dataset II.

Lecture Outline

General introduction to genetic epidemiology (lecture I)

Part I

What's a complex trait?

Genetic basis of complex traits

- Part II
  - Genetic approaches to studying complex traits
  - Candidate-gene analysis, GWAS, and GWAMA

Genetic Approaches To Investigating A Complex Trait

Once we have found evidence for a genetic component:

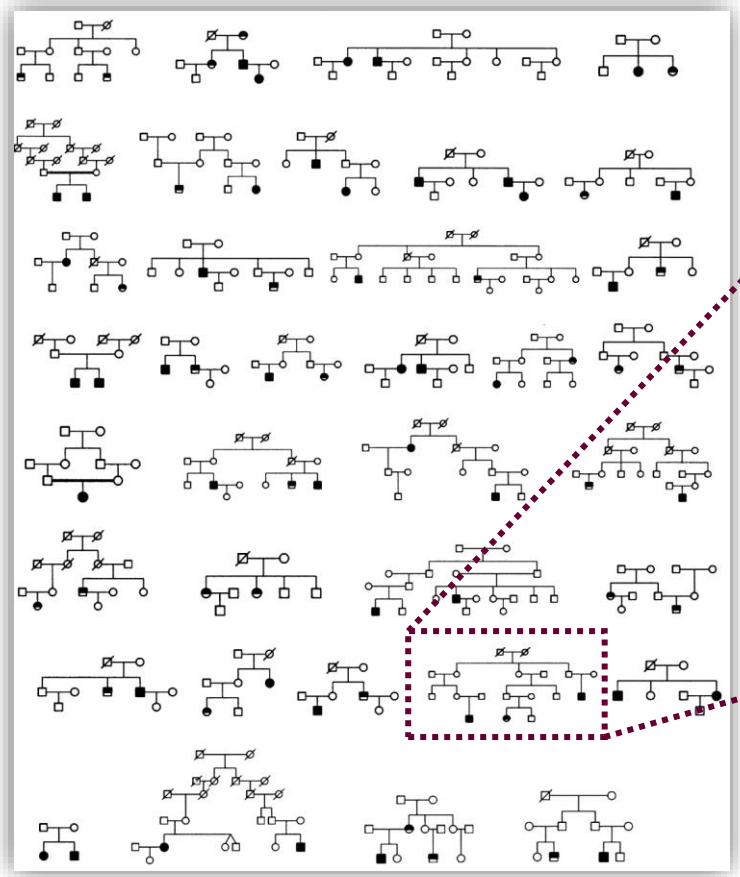
- **Linkage studies** in families with multiple affected members ('multiplex')
  - Test for cosegregation of a marker with the disease to see if the genetic marker and disease gene are physically linked
  - Problematic for complex diseases because of a lack of multiplex families

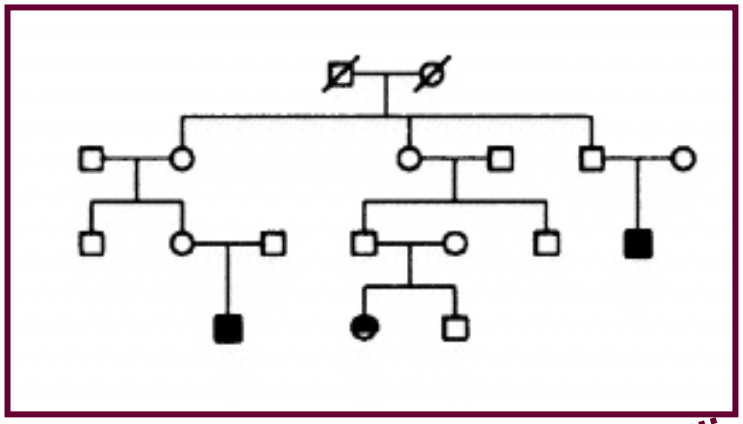
- **Allele-sharing studies** in affected relative pairs
  - Apply model-free methods on smaller subunits within multiplex families
  - «**Identity by descent**» (IBD) methods
    - Knowledge of transmission not required (non-parametric, or model-free)
    - Reasonable power to detect genes of fairly modest effects

- **Linkage disequilibrium** approaches
  - Exploit how genetic markers are correlated on chromosomes.

Linkage Studies In Multiplex Families

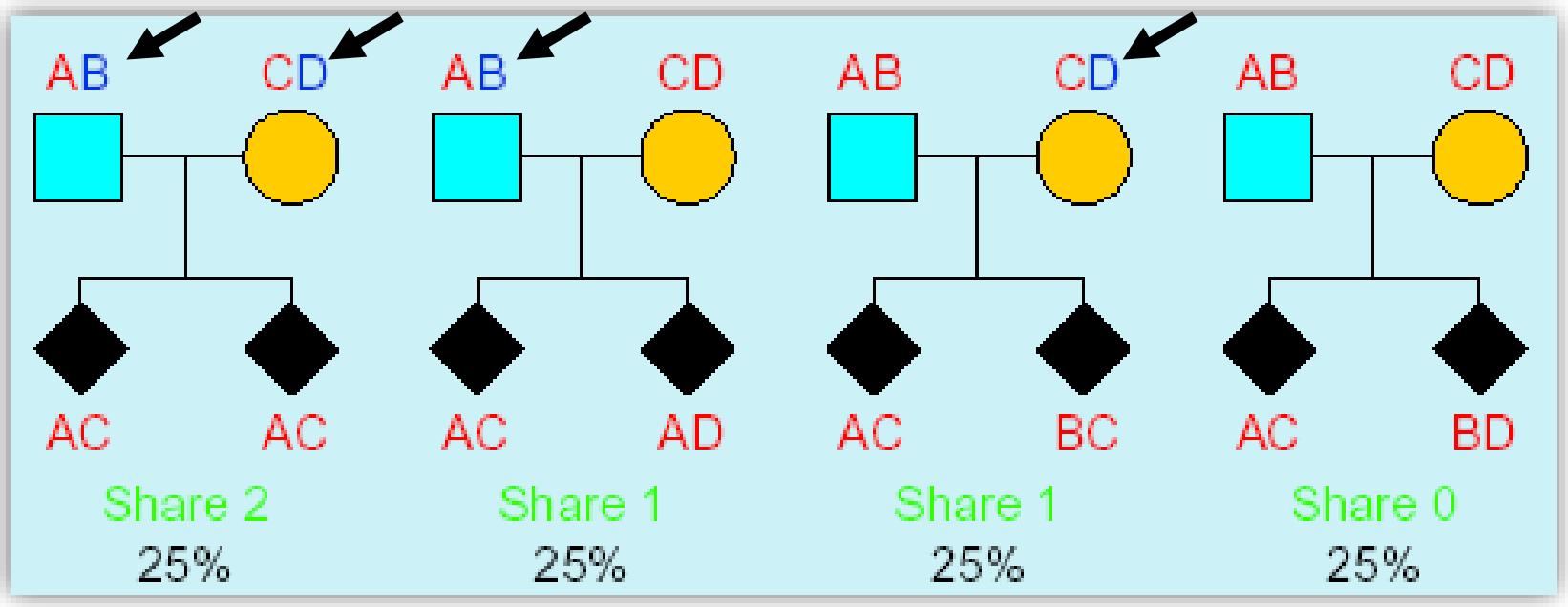
Genomewide linkage analyses can be performed using around 400 microsatellite markers distributed with an average spacing of 10 cM for genomewide coverage.





Allele-sharing Studies

Main idea: If affected pairs inherit a particular chromosomal fragment more often than would be expected by chance alone – this shows linkage!



2 (25 %) 1 (50 %) 0 (25 %)

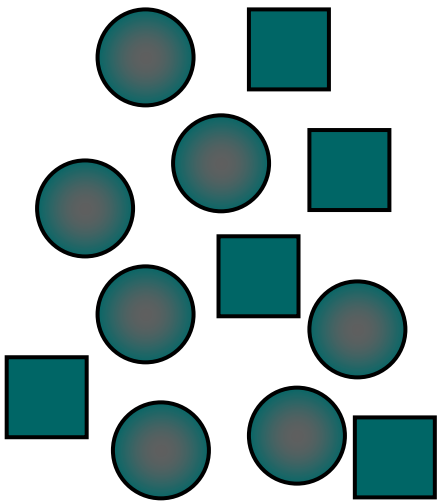
No. of parental alleles shared (% of Mendelian proportion)

Deviations from these expected proportions ⇒ evidence of linkage

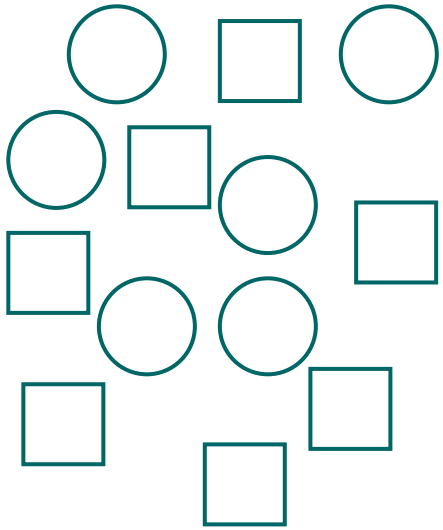
Source: Thomson and Esposito (1999). ***Trends Cell Biol***. 9(12):M17-20.

Linkage Disequilibrium (LD) Approaches

- Either case-control or family-based
  - Compare marker allele frequencies between a case and a control population
  - With family data, non-transmitted parental alleles are used as control alleles.
    - Test for deviations from the expected 50% transmission of an allele from parents to offspring.

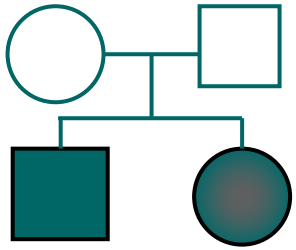


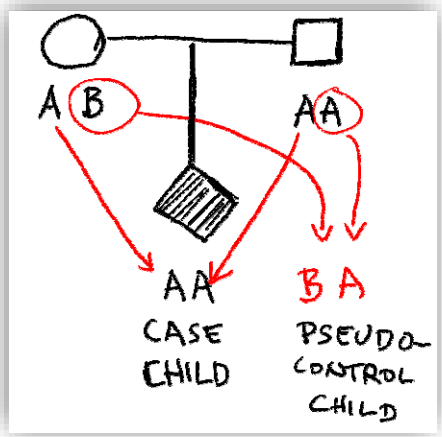
Case (desease)



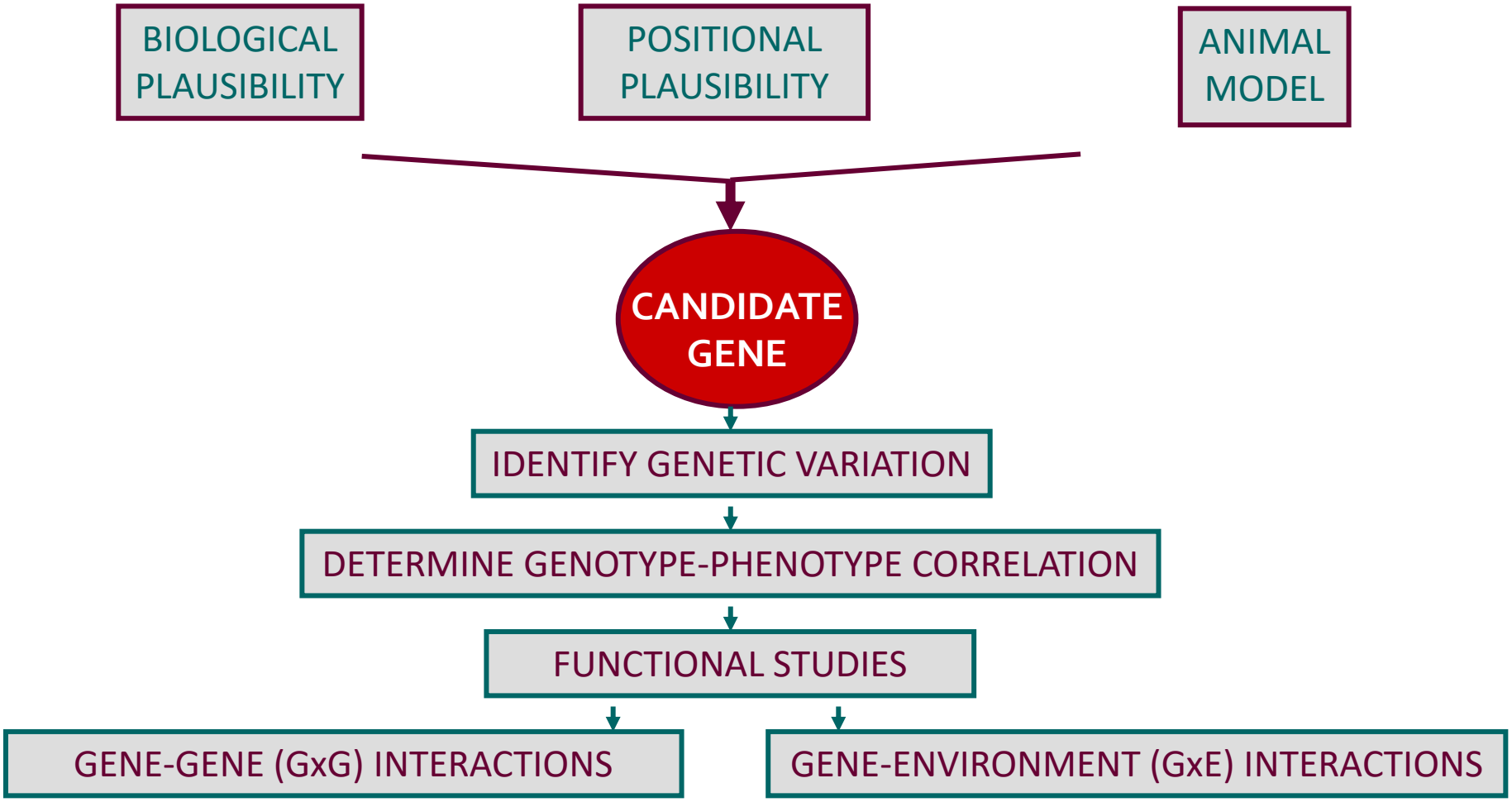
Control (healthy)

vs.



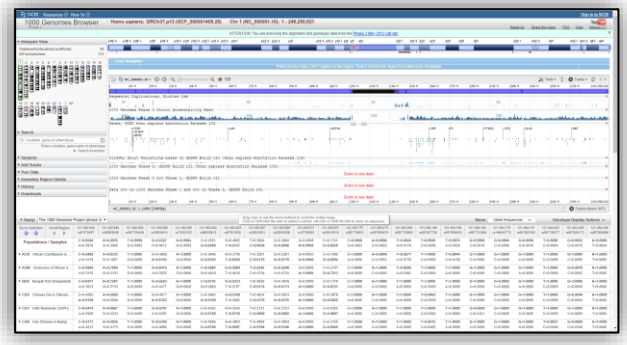


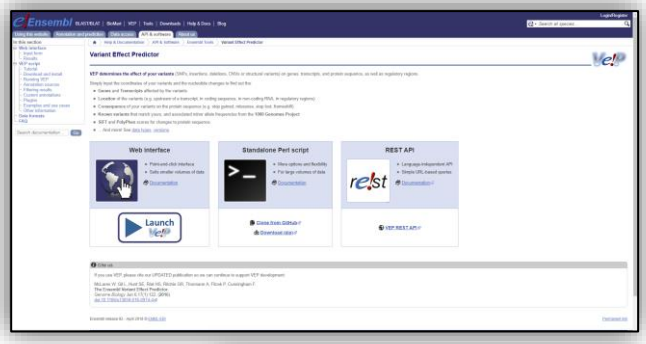
The Candidate-Gene Approach

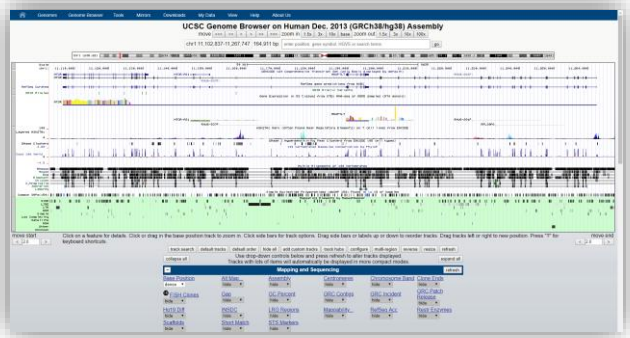


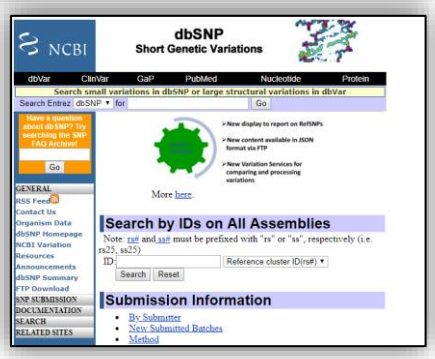
Selecting SNPs for candidate-gene analysis

- **Databases for selection/evaluation of SNPs:**
  - 1000 Genomes, *e!Ensembl,* UCSC's genome browser, and dbSNP, etc..









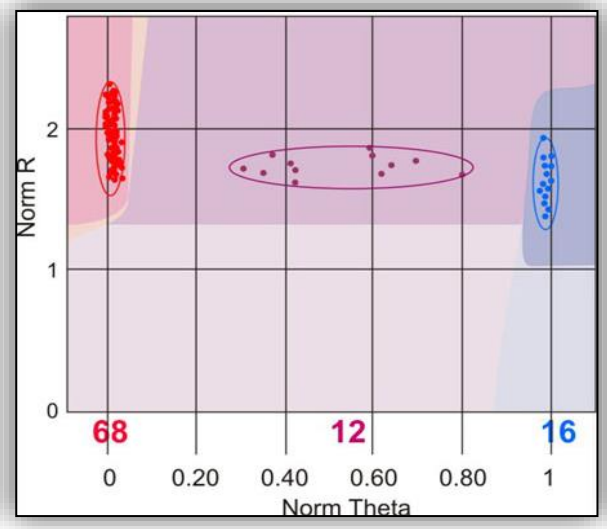
- **Criteria for prioritizing SNP selection:**
  - Prior association with the trait being studied
  - Minor allele frequency (MAF) of at least 5% to capture common variants
  - Preference for coding SNPs and SNPs in regulatory regions – functional!
  - SNPS with «haplotype-tagging» properties

SNP custom-assay and genotyping

- **SNP assays can be designed by ILLUMINA™**
  - A customized full panel of X number of SNPs in Y number of candidate genes.
- **Outsource the genotyping (and QC) to a core facility: e.g Microarray facility (Oslo), Sanger Institute (UK), DeCode genetics (Iceland), etc..**

Illumina iScan system

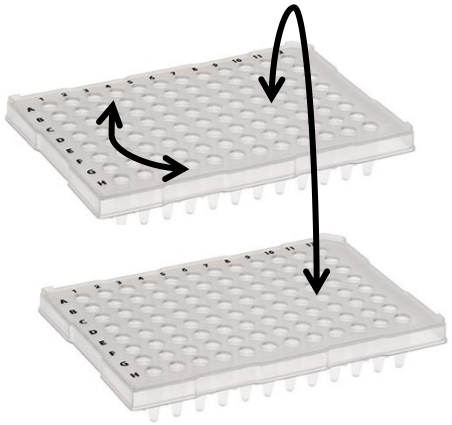
E.g. of genotype calling

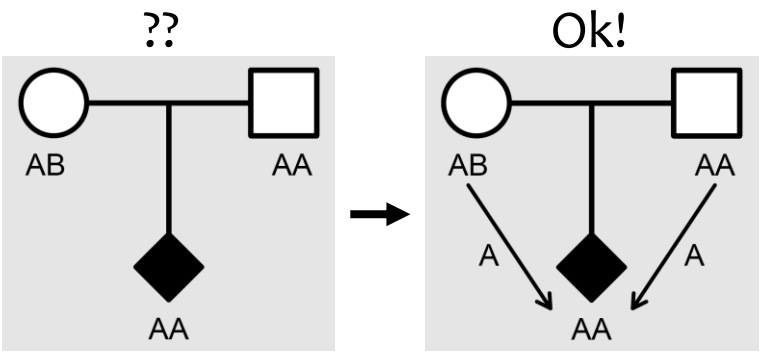


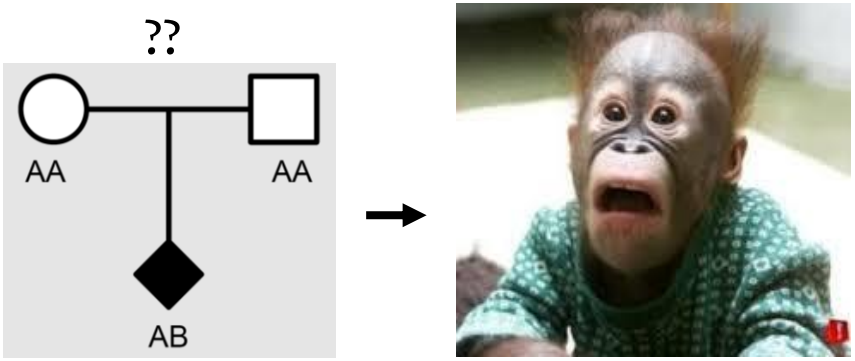
Genomics Core facility Oslo

Data Quality Control (Prelude to Marc's lecture on Tuesday)

- **Assess within/between plate genotype reproducibility**
  ⇒ SNP is deemed to have failed if <95% of samples generate a genotype at the locus
- **Exclude all SNPs with MAF <1%.**
  ⇒ Low statistical power in association analysis
- **Remove all SNPs that show deviation from HWE.**
  ⇒ Systematic genotyping errors, sample mix-ups, latent population substructure, or a biological effect (e.g., natural selection).
- **Screen for Mendelian inconsistencies within families.**
  ⇒ Sample switches or misidentified paternity/maternity







Genome-wide Asssociation Studies – Ch.4

- Hypothesis-free (agnostic) compared to candidate-gene approach
  - Looks for association across the entire genome using high-resolution SNP arrays (0.5-2.5 mill).

- What have we learnt?
  - Many association signals are not in genes previously thought to be associated with the disease.
  - Some associations are in areas that weren't even known before.

⇒ Provide new insights into biology and disease mechanism ☺

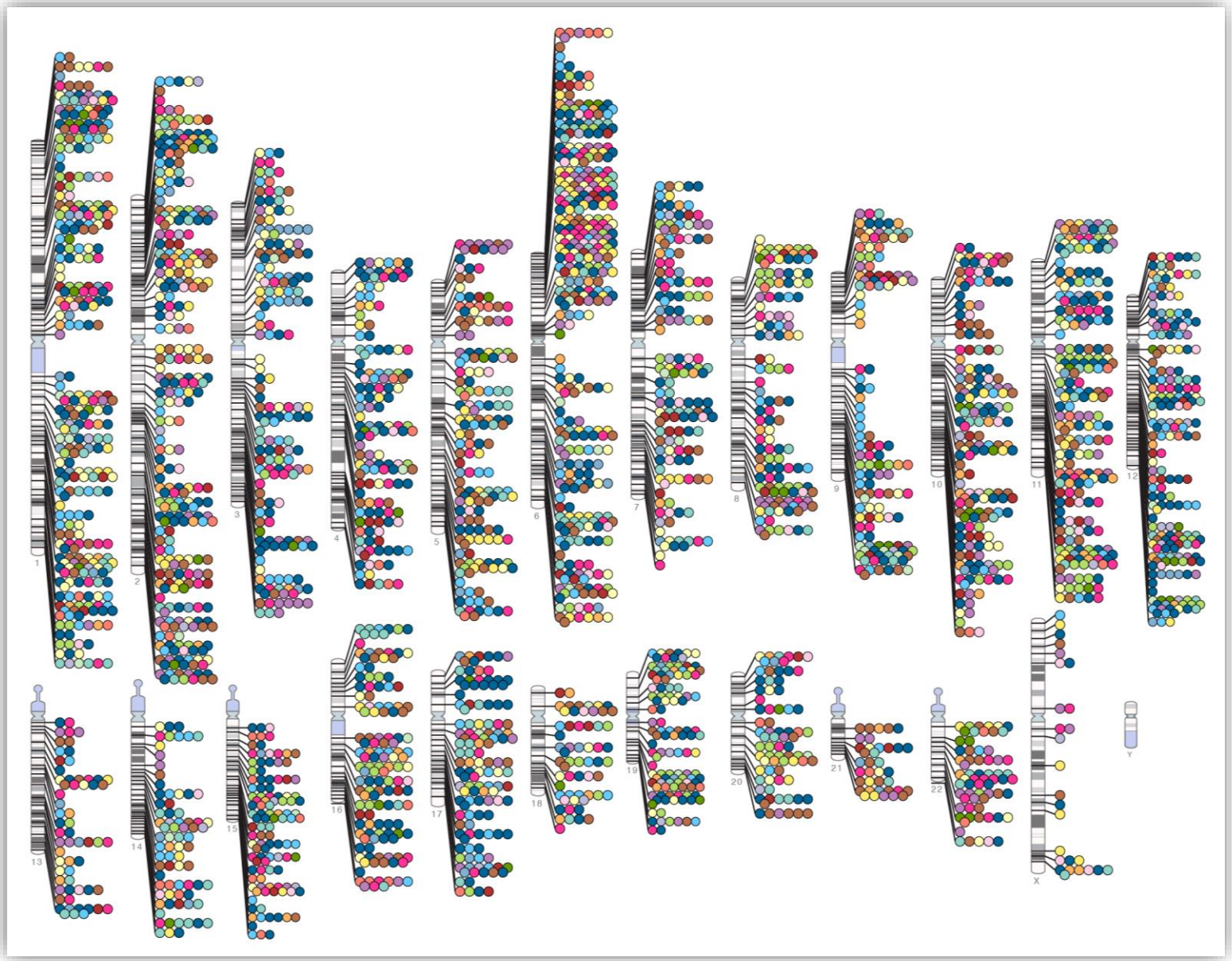
Signals in «gene deserts»:

Prostate cancer; CL/P	8q24
Crohn's disease	5p13.1; 1q31.2; 10p21

Signals in common (pleiotropy):

Diabetes/CHD/Melanoma	*CDKN2A/2B*
Prostate/breast/colon cancers; CL/P	8q24
Crohn's disease/Psoriasis	*IL23R*
Crohn's disease/T1DM	*PTPN2*

Published GWAS through Dec 2012 at $p \leq 5X10^{-8}$ for 17 trait categories



17 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 measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

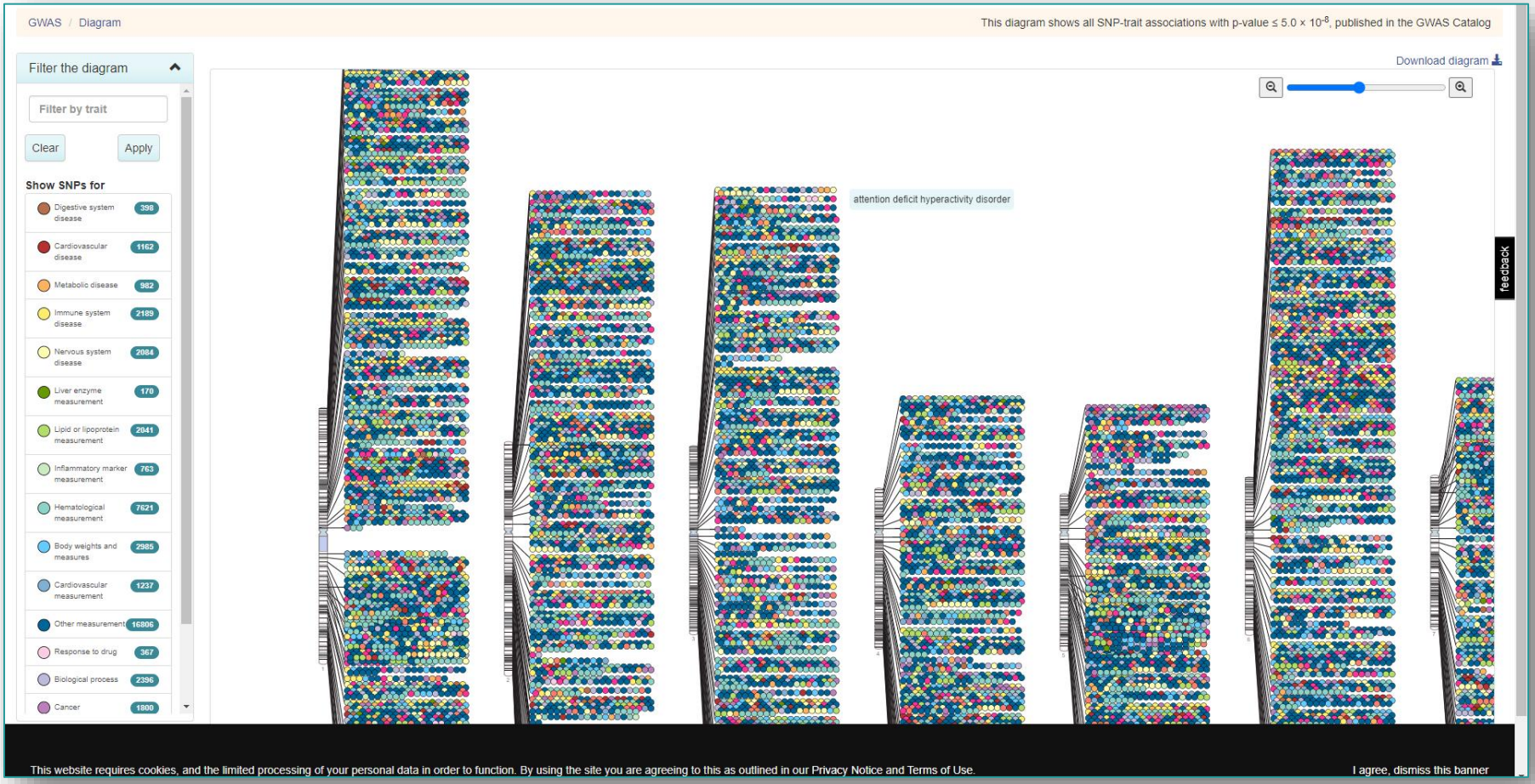
National Human Genome Research Institute

EMBL-EBI

NHGRI GWA Catalog at www.genome.gov/GWAStudies
Other useful sites that catalog GWAS (interactive): https://www.ebi.ac.uk/gwas/

Interactive GWAS catalog at EBI

Interactive diagram shows all SNP-trait associations with genome-wide significant p-value ≤ 5.0×10^{-8}



NHGRI GWA Catalog at www.genome.gov/GWAStudies
Other useful sites that catalog GWAS (interactive): https://www.ebi.ac.uk/gwas/diagram

TYPICAL GWAS WORKFLOW (CH. 4, P 79)

Initial GWAS – «Discovery sample»

↓

Replication / Fine-mapping

↓

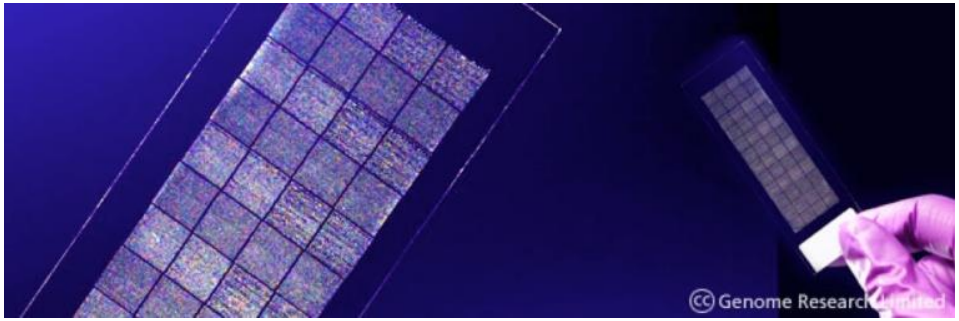
Sequencing / Genotyping

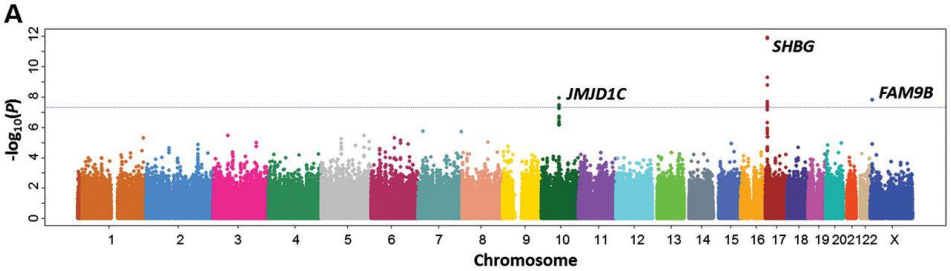
↓

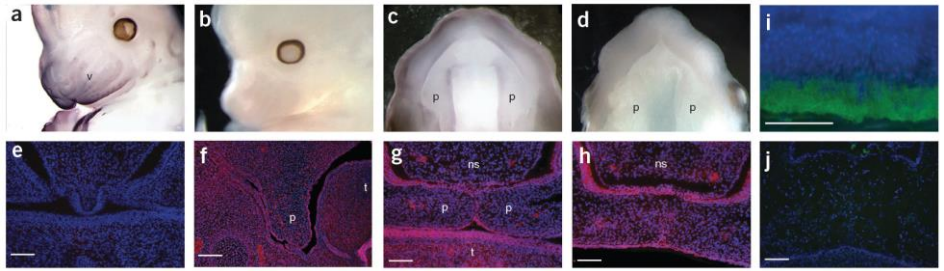
Functional studies

↓

Translational studies







- Most of the GWAS findings so far have not led to any major clinical applications.
- HOPE -- New therapies, improved diagnostics, better prevention, better public health, & precision medicine.

GWAS – What Are The Criteria For Success?

- Costs and availability of large samples are major limitations
  - Useful to meta-analyze summary statistics from multiple cohorts (GWAMA)

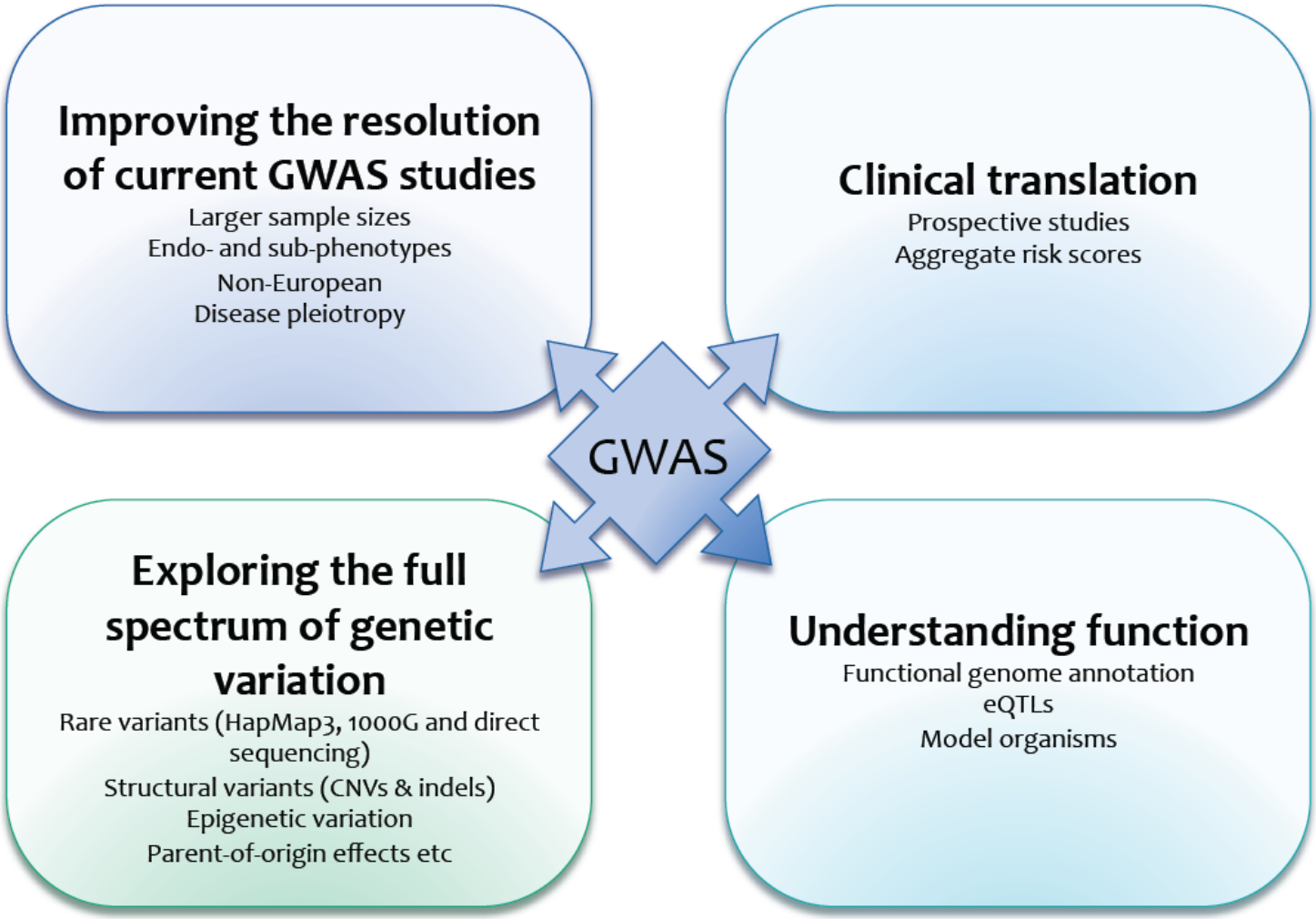
- Strict quality control throughout the process (Marc Vaudel's Tuesday lecture) + Stringent significance thresholds + Importance of replication

- Data sharing between several research groups is an effective way of increasing power to find new genes and loci.
  - But control for confounders is even more important when using data from different cohorts participating in a large consortium

- Disease heterogeneity is a problem.
  - The more narrowly/precisely the phenotype is defined, the better the odds for identifying a causal variant (but not always!)

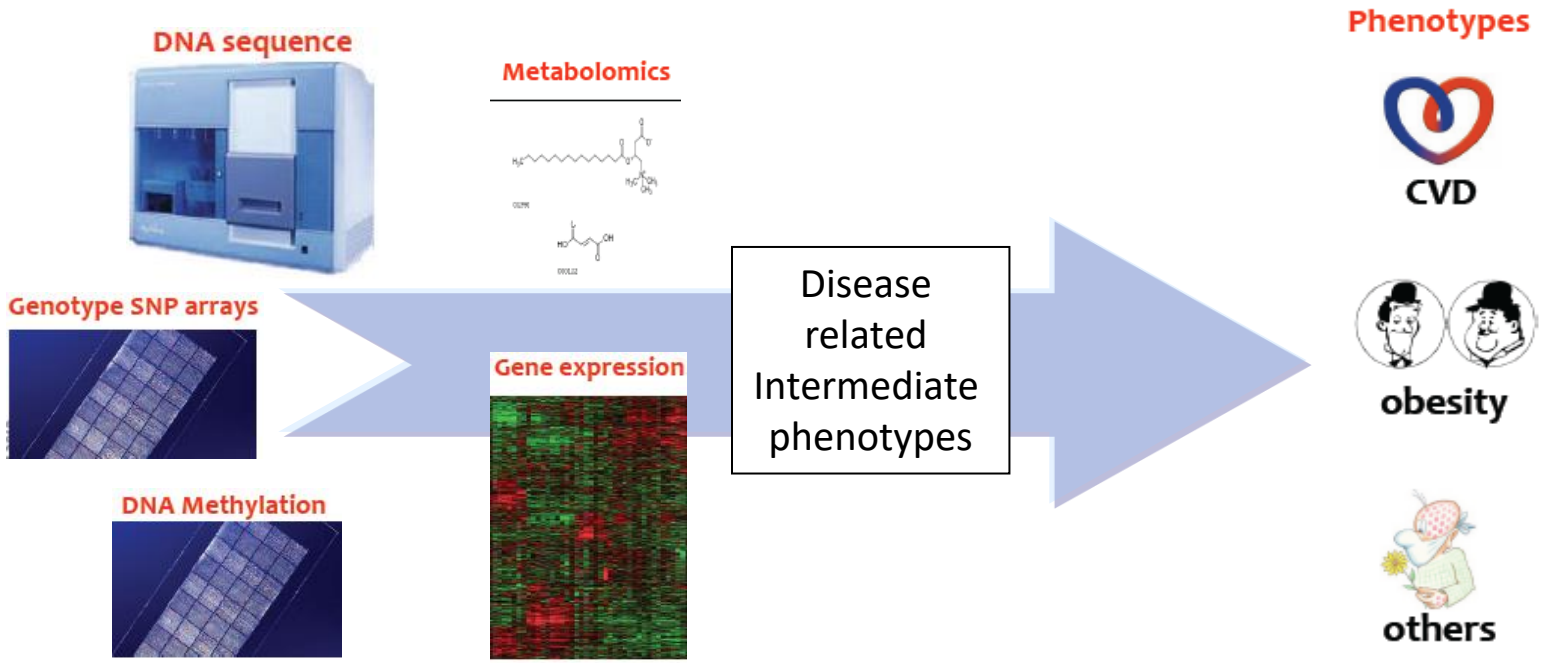
- Current methods are not well developed to identify rare variants (MAF <1%) that are perhaps associated with higher disease penetrance.

What Can We Do?



Whole Genome/Exome Sequencing

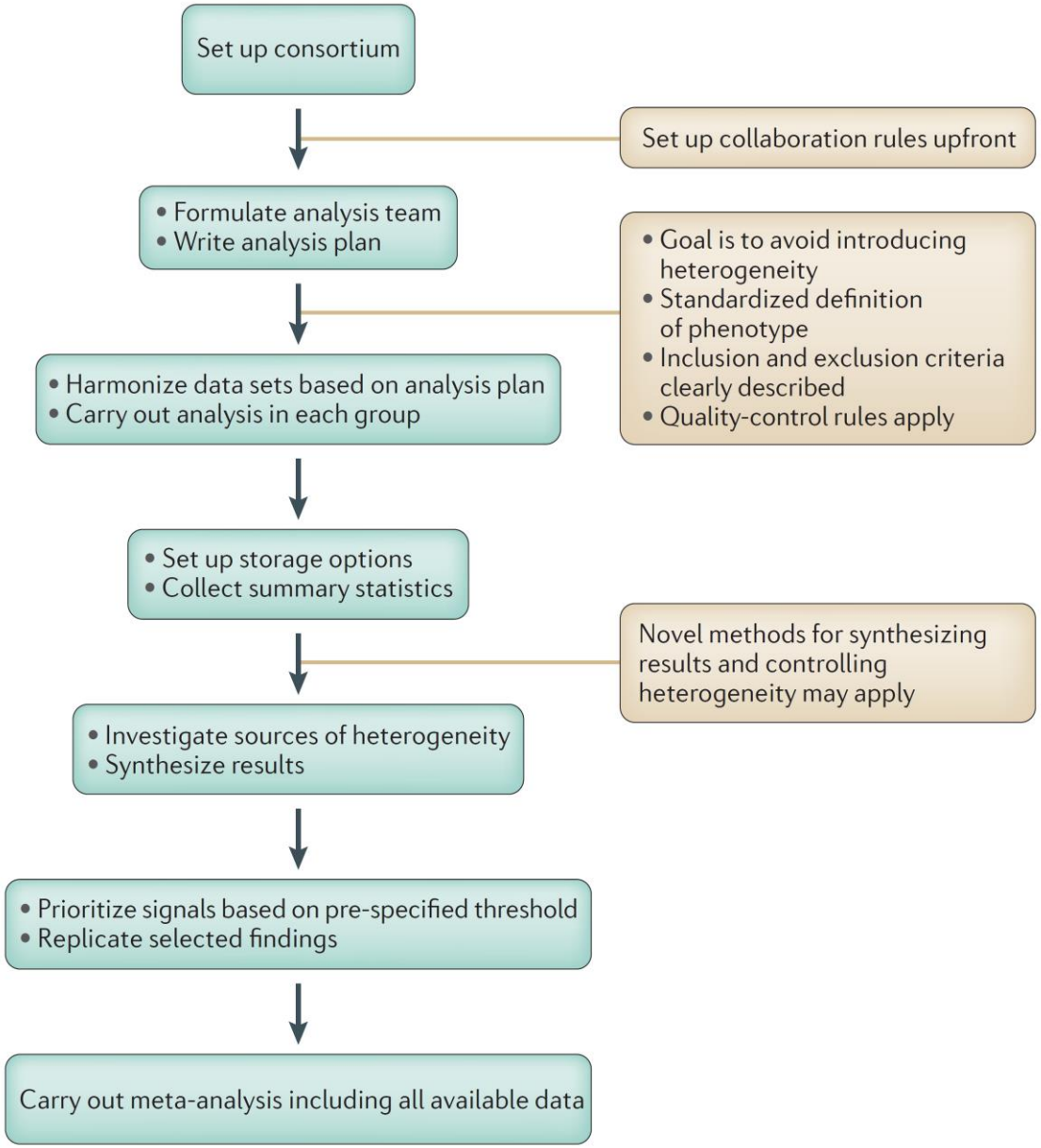
- **Two main objectives:**
  - Build a comprehensive catalog of genetic variation containing both common and rare genetic variants
  - Test these variants for association with disease.
- **Potential applications:**
  - Sequence based imputations in GWAS data (Marc Vaudel's Tuesday lecture)
  - Analyze cohorts with clearly defined phenotypes and map Mendelian diseases



META-GWAS ANALYSES – A SHORT PRIMER

- 1) DISCOVERY PHASE
  - Analyze GWAS results from different cohorts (Consortia)
    - Increase statistical power through increasing sample size
    - Beware of heterogeneity (PCA, stratified analyses, inflation, QQ plot)
  - Analysis of data at an aggregated level, i.e. not individual-level data.
    - Many Ethic Committees have an issue with sharing individual-level data.
    - Meta-GWAS analysis offers a good compromise
  - Different cohorts perform genome-wide imputation using the same imputation panel to harmonize genotype data across cohorts
    - Harmonizes genotyping platforms (standardization)
    - Lots more SNPs to analyze ⇒ More statistical power
- 2) REPLICATION PHASE
  - Invite more cohorts for replication
    - Confirmation of original findings in discovery phase

Stages in a Meta-GWAS Analysis



Reference: Evangelou & Ioannidis. Meta-analysis for genome-wide association studies and beyond. *Nature Reviews Genetics* | Vol 14 | June 2013 | 379

Examples of Consortia

Social Science Genetic Association Consortium

Home | About Us | Research | Events | News | Contact

Welcome to the Social Science Genetic Association Consortium (SSGAC).

The SSGAC is a cooperative enterprise among medical researchers and social scientists that coordinates genetic association studies for social science outcomes and provides a platform for interdisciplinary collaboration and cross-fertilization of ideas. The SSGAC also tries to promote the collection of harmonized and well-measured phenotypes.

Social Science Genetic Association Consortium

Click here to learn about our upcoming training sessions on Social Science Genomics and Genome-Wide Data Analysis!

Current Initiatives

The SSGAC is currently conducting large-scale

SSGAC in the News

"The Genetics of Staying in School" The Atlantic

EUROPEAN NETWORK OF GENOMIC AND GENETIC EPIDEMIOLOGY

ENGAGE

SEVENTH FRAMEWORK PROGRAMME

Home
Objectives
Partners
Work Packages
Events
Training
Press & Publications
Resources
Contact
FAQ

Young Investigator Profiles

ABOUT

ENGAGE (European Network for Genetic and Genomic Epidemiology) is a research project funded with 12 million euros by the European Commission under the 7th Framework Programme-Health Theme.
The project duration is five years, starting from January 1st, 2008.

The ENGAGE Consortium has brought together 24 leading research organizations and two biotechnology and pharmaceutical companies across Europe and in Canada and Australia.

ENGAGE aims to translate the wealth of data emerging from large-scale research in genetic and genomic epidemiology from European (and other) population cohorts into information relevant to future clinical applications. The concept of ENGAGE is to enable European researchers to identify large numbers of novel susceptibility genes that influence metabolic, behavioural and cardiovascular traits, and to study the interactions between genes and life style factors.

The ENGAGE consortium will integrate and analyse one of the largest ever human genetics dataset (more than 80,000 genome-wide association scans and DNAs and serum/plasma samples from over 600,000 individuals).

One goal is to demonstrate that the findings from ENGAGE can be used as diagnostic indicators for common diseases that will help us to understand better risk factors, disease progression and why people differ in responses to treatment.

NEWS

ENGAGE Flagship Paper: 'The Role of Adiposity in Cardiometabolic Traits: A Mendelian Randomization Analysis' (Fall T et al, Pedersen NL, McCarthy MI, Ingelsson E, Prokopenko I for ENGAGE, 25 June 2013)

ENGAGE Paper: 'Data sharing in large research consortia: experiences and recommendations from ENGAGE' (Budin-Ljøsne I et al, , June 2013)

ENGAGE ESHG Satellite Meeting 'Beyond GWAS: Biological and Clinical Insights from Research in European Biobanks', June 10th, Paris

ENGAGE Paper: 'GWAS of 126,559

Cohorts for Heart and Aging Research in Genomic Epidemiology

CHARGE Consortium

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium was form phenotyped longitudinal cohort studies.

Its founding member cohorts include:

- Age, Gene, Environment, Susceptibility Study -- Reykjavik
- Atherosclerosis Risk in Communities Study
- Cardiovascular Health Study
- Framingham Heart Study
- Rotterdam Study

Additional core cohorts include:

- Coronary Artery Risk Development in Young Adults
- Family Heart Study
- Health, Aging, and Body Composition Study
- Jackson Heart Study
- Multi-Ethnic Study of Atherosclerosis

EAGLE Consortium

The EArly Genetics and Lifecourse Epidemiology (EAGLE) Consortium is a consortium of pregnancy and birth cohorts that aims to collaborate to investigate the genetic basis of phenotypes in antenatal and early life and childhood.

EAGLE covers a broad range of pathways and phenotypes, and will integrate closely with the DOHaD (developmental origins of health and disease) community.

All participating cohorts (1958 British Birth Cohort; ALSPAC; CHOP; COPSAC; DBC; Exeter Family Study; Generation R; HBCS; LISA+; MoBa; NTR; NFBC 66; Project Viva; Raine) have GWAS data available by July 1st 2009.

EAGLE working groups and leaders are listed below:

- **Antenatal Growth** (Vincent Jaddoe and Craig Pennell)

EAGLE
EArly Genetics & Lifecourse Epidemiology Consortium

Social Science Genetic Association Consortium

Home | About Us | Research | PGI Repository | News | Events | Contact

Welcome to the Social Science Genetic Association Consortium (SSGAC).

Social Science Genetic Association Consortium

The SSGAC is a cooperative enterprise among medical researchers and social scientists that coordinates genetic association studies for social science outcomes and provides a platform for interdisciplinary collaboration and cross-fertilization of ideas. The SSGAC also tries to promote the collection of harmonized and well-measured phenotypes.

Recent Events: Russell Sage Foundation Summer Institute in Social-Science Genomics, 2021

Click here to learn about the 2018 Polygenic Prediction and its Application in Social Science Conference

Current Initiatives

The SSGAC is currently conducting:

- Within-family GWAS of multiple phenotypes
- Ongoing updates of the Polygenic Index Repository
- Developing methods for multi-ancestry genetic analysis

Please contact us if you are interested in joining these initiatives.

MORE

Data

To locate and download summary data from past studies of the SSGAC, click the link below.

MORE

SSGAC in the News

"Many Genes Play a Role in Educational Attainment, Enormous Genetic Study Finds"
The New York Times. July 23, 2018

"Why Study the Genetics of Staying in School?"
The Atlantic. July 23, 2018.

MORE

Check us out on Twitter: @thessgac

Social Science Genetic Association Consortium

Home | About Us | Research | PGI Repository | News | Events | Contact

Team

Co-Founders

- Daniel Benjamin
  University of California, Los Angeles (UCLA)
- David Cesarini
  New York University
- Philipp Koellinger
  Vrije Universiteit Amsterdam

Steering Committee

- Daniel Benjamin
  University of California, Los Angeles (UCLA)
- David Cesarini
  New York University
- Philipp Koellinger
  Vrije Universiteit Amsterdam
- Patrick Turley
  University of Southern California (USC)

STANDARD OPERATION PROTOCOL (SOP)

1) STANDARD OPERATING PROTOCOL (SOP) in «Discovery Phase»

- Background of the proposed Meta-GWAS analysis (GWAMA)
  - Goals of the initiative
- Trait definition and instructions for phenotype harmonization
  - A detailed definition of the trait (not all cohorts have same measures)
  - Eligibility and sample inclusion/exclusion criteria
- Genotypes and imputation
  - Imputation with chosen panel (HapMap Phase II CEU Panel, 1000 Genomes, HRC)
  - Filters to be applied <u>before</u> imputation (SNP call >95%, HWE p >10e-6, MAF >5%)
- Analysis details
  - Specification of models to be used in the analysis
  - Linear regression/Logistic regression, Include PCA for correcting for stratification
- Results file formats
  - Format to report GWAS results from individual cohorts

Reporting Of Results

Variable name (case sensitive!!)	Description
SNPID	SNP ID as rs number
Chr	Chromosome number (1-22).
position	physical position for the reference sequence (indicate build 35/36 in readme file)
coded_all	Coded allele, also called modelled allele (in example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G)
noncoded_all	The other allele
strand_genome	+ or -, representing either the positive/forward strand or the negative/reverse strand of the human genome reference sequence; to clarify which strand the coded_all and noncoded_all are on
Beta	Beta estimate from genotype-phenotype association, at least 5 decimal places – 'NA' if not available
SE	Standard error of beta estimate, to at least 5 decimal places – 'NA' if not available
Pval	*p*-value of test statistic, here just as a double check – 'NA' if not available
AF_coded_all	Allele frequency for the coded allele – 'NA' if not available
HWE_pval	Exact test Hardy-Weinberg equilibrium *p*-value -- only directly typed SNPs, NA for imputed
callrate	Genotyping call rate after exclusions
n_total	Total sample with phenotype and genotype for SNP
imputed	1/0 coding; 1=imputed SNP, 0=if directly typed
used_for_imp	1/0 coding; 1=used for imputation, 0=not used for imputation
oevar_imp*	Observed divided by expected variance for imputed allele dosage -- NA otherwise
avpostprob**	Average posterior probability for imputed SNP allele dosage (applies to best-guess genotype imputation)

* oevar_imp is called r^2 in Mach, proper_info in Impute and R^2 in Beagle.
** avpostprob is called Quality in Mach, certainty in Impute and Beagle does not give this statistic.

PDB 289 (Study of prematurity – PI Bo Jacobsson)

	A	B	C	D	E	F	G	H	I	J	K	L
1	**SAMPLE INFORMATION**											
2												
3	Country					Norway						
4	Sampling scheme					Population-based, nested case-control	e.g., family-based, clinically-selected (please specify selected phenotype), etc.					
5	SNP chip					Illumina 660W quad						
6	Pre-imputation QC											
7		Marker filters:										
8			MAF >			0.5	5% recommended					
9			Call rate >			95%	95% recommended					
10			HWE exact test at p >			0.001	10E-06 recommended					
11		Removed subjects with:										
12			Overall call rates <			98%						
13	Imputation & association procedure											
14		Imputation software				PLINK 1.07	please specify version number					
15		Reference sample				HAPMAP II CEU	HapMap phase II CEU recommended					
16		NCBI build				NCBI 36.2	e.g. NCBI 36.2					
17		Association software				PLINK 1.07	please specify version number					
18	Study contacts (name, email):											
19		Data analyst:				Ronny Myhre, Astanand Jugessur and Håkon Gjessing						
20		Primary contact:				Bo Jacobsson (PI; bo.jacobsson@obgyn.gu.se) and Astanand Jugessur (astanand.jugessur@fhi.no)						
21		Other contact(s):				Per Magnus						
22	Additional notes					x	e.g., non-standard covariates included in analyses (please specify)					
23												
24	**SAMPLE DEMOGRAPHICS**											
25						Females	Controls	sPTD cases				
26	N					1338	678	660				
27	Age at reporting											
28		Mean				28.7	28.9	28.4				
29		St. Dev.				3.5	3.6	3.6				
30		Range				14 (20-34)	14 (20-34)	14 (20-34)				
31	Birth Year											
32		Mean				1974.7	1974.4	1975				
33		St. Dev.				3.8	3.4	4.0				
34		Range				20 (1966-1986)	19 (1966-1985)	20 (1966-1986)				
35	Race (N per category)						Note that only individuals of European heritage should be included in the analysis					
36		American Indian or Alaska Native				na						
37		Asian				na						
38		Native Hawaiian or Other Pacific Islander				na						
39		Black				na						
40		White				na						
41	Ethnicity (N per category)											
42		Hispanic or Latino				na						
43		Not Hispanic or Latino				na						

Example Of A GWAMA

1) Trait proposed for a GWAMA: «Aggressive behavior»

- SOP describes the goal of the proposed GWAMA
  - **Goal:** large-scale meta-GWAS on Aggressive behavior
  - **Merit:** Findings will help identify to what extent the effect of the SNP(s) changes with age, instrument, or the rater of the behavior.
- Trait definition and instructions for phenotype harmonization
  - Phenotype data at different ages (3 to 18 yrs) and as rated by different raters (parental, self and/or teacher ratings) to be included in a single analysis
  - **Instruments:** A variety of psychometric instruments (e.g. CBCL, SDQ, ASR, YSR)
  - Sample size threshold for inclusion: at least 1000 subjects.
  - Limit analyses to subjects of European ancestry.
- Genotypes and imputation
  - Imputation with chosen panel (1000 Genomes)
  - **Software for imputation:** IMPUTE, MACH, MINIMAC or BEAGLE.
  - Filters to be applied before imputation (SNP call >95%, HWE p >10e-6, MAF >5%)
- Analysis
  - For cohorts providing a single phenotype measure: Run the GWA using linear Reg.
  - **Covariates:** sex, Z-score of age at time of assessment, Age^2 (Z-transformed, then squared), the first 5 PCs, Study-specific covariates (study site, batch effects etc.)

Example Of A Meta-GWAS – Contd…

1) Instructions for genotype handling (pre-imputation QC):

- Exclude SNPs with:
  - MAF <1%
  - SNP call rate <95%
  - Failure of HWE exact test at p<1e-6
  - Poor clustering on visual inspection of intensity plots.
  - Wrong sex, aberrant genotype (XXY), known 1st or 2nd degree relatives in sample

2) Imputation:

- Use 1000 genomes Phase I release and coordinates as used in GRCh37
- Imputation software: IMPUTE or MACH
- Use servers for imputation: Michigan imputation server or Sanger Institute in UK
- Provide per-SNP quality indicators (proper_info in IMPUTE, r^2.hat in MACH)

3) Analysis:

- Perform association test using MACH2QTL or SNPTEST

Example Of A Meta-GWAS – Contd…

Uploading data

- To a secure server using secure transfer protocol (sftp)
  - Download and Install an sftp software; e.g. Filezilla or WinScp
  - Upload a «README.txt» file with a brief description of data uploaded, the date, the human genome reference sequence used for strand reference, and scale of Beta estimates.
  - Prepare a file named «STUDY.PHEN.DATE.txt»
    - Study=Cohort, PHEN=phenotype, Date=DDMMYYYY (date file was prepared)

Meta-analysis:

- Usually done by the lead analysts from the cohort(s) initiating this GWAMA
- Software: METAL or GWAMA

A Few Examples…

Papers on EA based including the MoBa dataset (PDB 289)

2013 (N=101,069 individuals; 3 GW sig SNPs; 2% variance)

Sciencexpress Reports

GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

All authors with their affiliations appear at the end of this paper.

A genome-wide association study of educational attainment was conducted in a discovery sample of 101,069 individuals and a replication sample of 25,490. Three independent SNPs are genome-wide significant (rs9320913, rs11584700, rs4851266), and all three replicate. Estimated effects sizes are small ($R^2 \approx 0.02\%$), approximately 1 month of schooling per allele. A linear polygenic score from all measured SNPs accounts for ≈ 2% of the variance in both educational attainment and cognitive function. Genes in the region of the loci have previously been associated with health, cognitive, and central nervous system phenotypes, and bioinformatics analyses suggest the involvement of the anterior caudate nucleus. These findings provide promising candidate SNPs for follow-up work, and our effect size estimates can anchor power analyses in social-science genetics.

2015 (N=293,723 individuals; 74 GW sig SNPs)

LETTER

doi:10.1038/nature17671

Genome-wide association study identifies 74 loci associated with educational attainment

A list of authors and their affiliations appears in the online version of the paper.

Educational attainment is strongly influenced by social and other environmental factors, but genetic factors are estimated to account for at least 20% of the variation across individuals[1]. Here we report the results of a genome-wide association study (GWAS) for educational attainment that extends our earlier discovery sample[1,2] of 101,069 individuals to 293,723 individuals, and a replication study in an independent sample of 111,349 individuals from the UK Biobank. We identify 74 genome-wide significant loci associated with the number of years of schooling completed. Single-nucleotide polymorphisms associated with educational attainment are disproportionately found in genomic regions regulating gene expression in the fetal brain. Candidate genes are preferentially

Our meta-analysis identified 74 approximately independent genome-wide significant loci. For each locus, we define the 'lead SNP' as the SNP in the genomic region that has the smallest P value (Supplementary Information section 1.6.1). Figure 1 shows a Manhattan plot with the lead SNPs highlighted. This includes the three SNPs that reached genome-wide significance in the discovery stage of our previous GWAS meta-analysis of educational attainment[1]. The quantile–quantile (Q–Q) plot of the meta-analysis (Extended Data Fig. 1) exhibits inflation ($\lambda_{GC} = 1.28$), as expected under polygenicity[3].

Extended Data Fig. 2 shows the estimated effect sizes of the lead SNPs. The estimates range from 0.014 to 0.048 standard deviations per allele (2.7 to 9.0 weeks of schooling), with incremental R^2 in the

2015

Molecular Psychiatry (2015) 20, 735–743
© 2015 Macmillan Publishers Limited All rights reserved 1359-4184/15
www.nature.com/mp

ORIGINAL ARTICLE

The association between lower educational attainment and depression owing to shared genetic effects? Results in ~25 000 subjects

WJ Peyrot[1], SH Lee[2], Y Milaneschi[1], A Abdellaoui[3], EM Byrne[2], T Esko[4,5], EJC de Geus[3], G Hemani[2,6], JJ Hottenga[3], S Kloiber[7], DF Levinson[8], S Lucae[7], Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium (Corporate Collaborator), NG Martin[9], SE Medland[9], A Metspalu[4,5], L Milani[4,5], MM Noethen[10], JB Potash[11], M Rietschel[12], CA Rieteveld[13,14], S Ripke[15], J Shi[16], Social Science Genetic Association Consortium (Corporate Collaborator), G Willemsen[3], Z Zhu[2], DI Boomsma[3], NR Wray[2] and BWJH Penninx[1]

2018 (N=1.1 mill individuals; 1271 GW sig SNPs; 11-13% variance)

nature genetics

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > nature genetics > articles > article

Article | Published: 23 July 2018

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

James J. Lee, Robbee Wedow, … David Cesarini + Show authors

Nature Genetics **50**, 1112–1121 (2018) | Cite this article

64k Accesses | **727** Citations | **2042** Altmetric | Metrics

Abstract

Here we conducted a large-scale genetic association analysis of educational attainment in a sample of approximately 1.1 million individuals and identify 1,271 independent genome-wide-significant SNPs. For the SNPs taken together, we found evidence of heterogeneous effects across environments. The SNPs implicate genes involved in brain-development processes and neuron-to-neuron communication. In a separate analysis of the X chromosome, we identify 10 independent genome-wide-significant SNPs and estimate a SNP heritability of around 0.3% in both men and women, consistent with partial dosage compensation. A joint (multi-phenotype) analysis of educational attainment and three related cognitive phenotypes generates polygenic scores that explain 11–13% of the variance in educational attainment and 7–10% of the variance in cognitive performance. This prediction accuracy substantially increases the utility of polygenic scores as tools in research.

2022 (N= ~3 mill individuals; 3952 GW sig SNPs; 12-16% of variance)

nature genetics ARTICLES https://doi.org/10.1038/s41588-022-01016-z

Check for updates

OPEN

Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals

Aysu Okbay[1,197,198], Yeda Wu[2], Nancy Wang[3], Hariharan Jayashankar[3], Michael Bennett[3], Seyed Moeen Nehzati[4], Julia Sidorenko[2], Hyeokmoon Kweon[1], Grant Goldman[3], Tamara Gjorgjieva[3], Yunxuan Jiang[5], Barry Hicks[5], Chao Tian[5], David A. Hinds[5], Rafael Ahlskog[6], Patrik K. E. Magnusson[7], Sven Oskarsson[6], Caroline Hayward[8], Archie Campbell[9,10], David J. Porteous[9,10,11], Jeremy Freese[12], Pamela Herd[13], 23andMe Research Team*, Social Science Genetic Association Consortium*, Chelsea Watson[4], Jonathan Jala[4], Dalton Conley[14], Philipp D. Koellinger[1,15], Magnus Johannesson[16], David Laibson[17], Michelle N. Meyer[18], James J. Lee[19], Augustine Kong[20], Loic Yengo[2,198], David Cesarini[3,21,22,198], Patrick Turley[23,24,198], Peter M. Visscher[2,198], Jonathan P. Beauchamp[25,198], Daniel J. Benjamin[3,4,26,198] and Alexander I. Young[4,26,197,198]

We conduct a genome-wide association study (GWAS) of educational attainment (EA) in a sample of ~3 million individuals and identify 3,952 approximately uncorrelated genome-wide-significant single-nucleotide polymorphisms (SNPs). A genome-wide polygenic predictor, or polygenic index (PGI), explains 12–16% of EA variance and contributes to risk prediction for ten diseases. Direct effects (i.e., controlling for parental PGIs) explain roughly half the PGI's magnitude of association with EA and other phenotypes. The correlation between mate-pair PGIs is far too large to be consistent with phenotypic assortment alone, implying additional assortment on PGI-associated factors. In an additional GWAS of dominance deviations from the additive model, we identify no genome-wide-significant SNPs, and a separate X-chromosome additive GWAS identifies 57.

2016

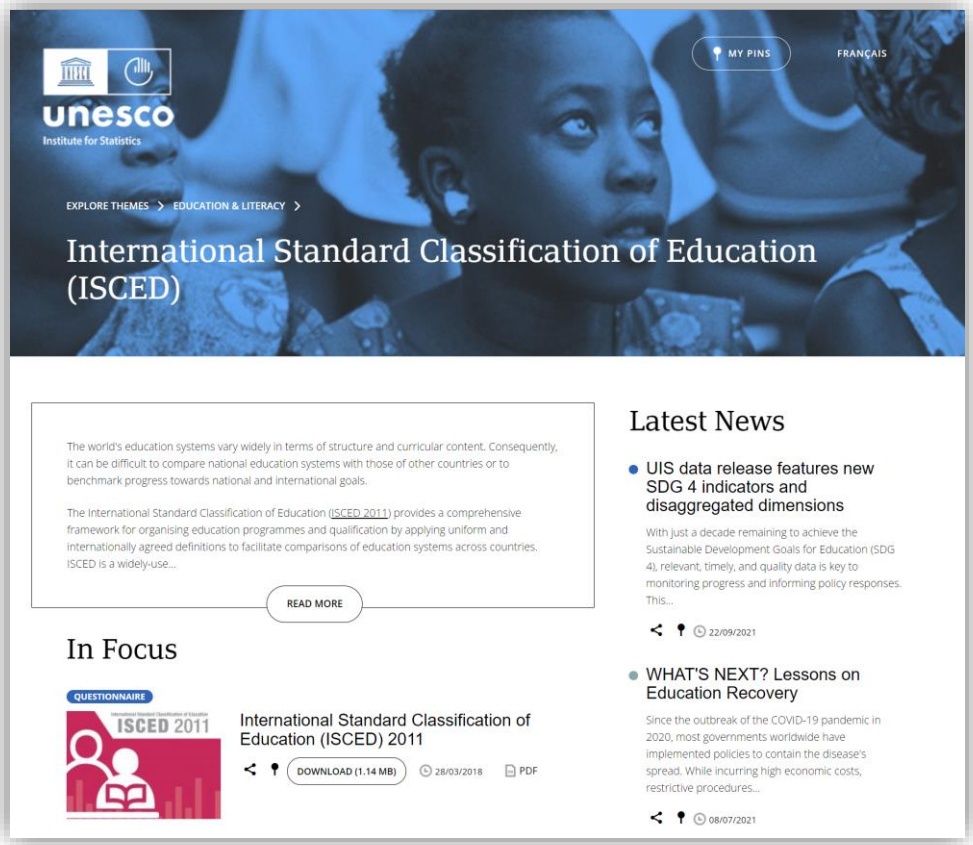
CrossMark click for updates

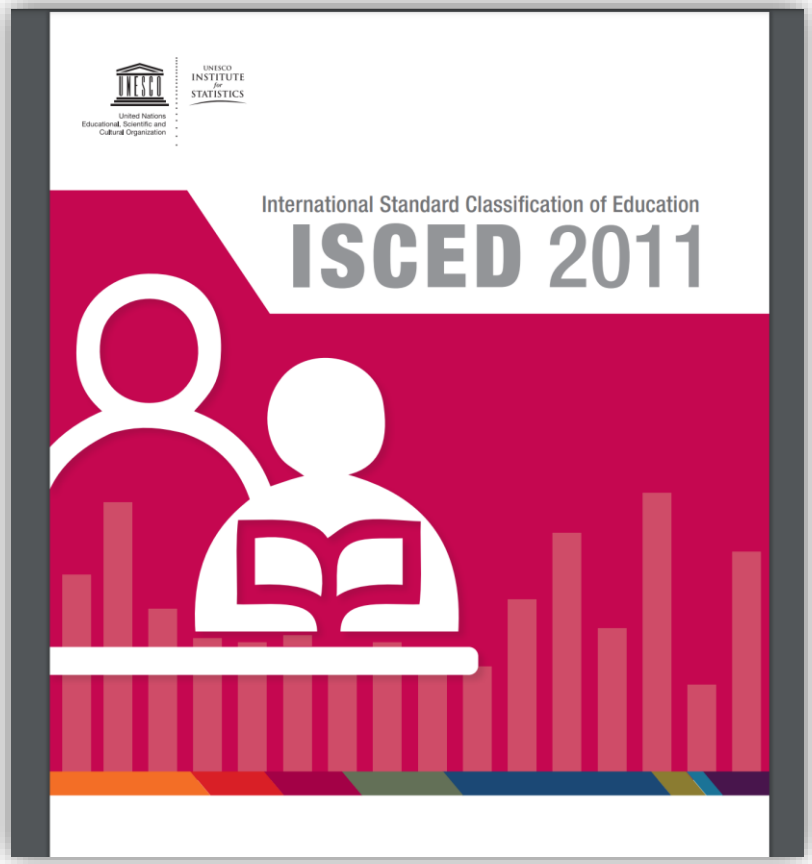
Genetic variants linked to education predict longevity

Riccardo E. Marioni[a,b,c,1,2], Stuart J. Ritchie[a,d,1], Peter K. Joshi[e,1], Saskia P. Hagenaars[a,d,f], Aysu Okbay[g,h,i], Krista Fischer[j], Mark J. Adams[f], W. David Hill[a,d], Gail Davies[a,d], Social Science Genetic Association Consortium[3], Reka Nagy[k], Carmen Amador[k], Kristi Läll[j], Andres Metspalu[l], David C. Liewald[a,d], Archie Campbell[b], James F. Wilson[e,k], Caroline Hayward[k], Tõnu Esko[l,m,n], David J. Porteous[a,b], Catharine R. Gale[a,d,o,4], and Ian J. Deary[a,d,4]

[a]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom; [b]Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom; [c]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia; [d]Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom; [e]Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh EH16 4UX, United Kingdom; [f]Division of Psychiatry, University of Edinburgh, Edinburgh EH8 9YL, United Kingdom; [g]Department of Applied Economics, Erasmus School of Economics, Erasmus University, 3062 PA Rotterdam, The Netherlands; [h]Department of Epidemiology, Erasmus Medical Center, 3015 CE Rotterdam, The Netherlands; [i]Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam 3062 PA, The Netherlands; [j]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia; [k]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom; [l]Institute of Mathematical Statistics, University of Tartu, Tartu 50409, Estonia; [m]Broad Institute, Cambridge, MA 02142; [n]Department of Endocrinology, Children's Hospital Boston, Boston, MA 02115; and [o]Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, Southampton SO17 1BJ, United Kingdom

http://uis.unesco.org/en/topic/international-standard-classification-education-isced





nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01016-z

Check for updates

OPEN

Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals

Aysu Okbay[1,197,198] ✉, Yeda Wu[2], Nancy Wang[3], Hariharan Jayashankar[3], Michael Bennett[3], Seyed Moeen Nehzati[4], Julia Sidorenko[2], Hyeokmoon Kweon[1], Grant Goldman[3], Tamara Gjorgjieva[3], Yunxuan Jiang[5], Barry Hicks[5], Chao Tian[5], David A. Hinds[5], Rafael Ahlskog[6], Patrik K. E. Magnusson[7], Sven Oskarsson[6], Caroline Hayward[8], Archie Campbell[9,10], David J. Porteous[9,10,11], Jeremy Freese[12], Pamela Herd[13], 23andMe Research Team*, Social Science Genetic Association Consortium*, Chelsea Watson[4], Jonathan Jala[4], Dalton Conley[14], Philipp D. Koellinger[1,15], Magnus Johannesson[16], David Laibson[17], Michelle N. Meyer[18], James J. Lee[19], Augustine Kong[20], Loic Yengo[2,198], David Cesarini[3,21,22,198], Patrick Turley[23,24,198], Peter M. Visscher[2,198] ✉, Jonathan P. Beauchamp[25,198], Daniel J. Benjamin[3,4,26,198] ✉ and Alexander I. Young[4,26,197,198] ✉

We conduct a genome-wide association study (GWAS) of educational attainment (EA) in a sample of ~3 million individuals and identify 3,952 approximately uncorrelated genome-wide-significant single-nucleotide polymorphisms (SNPs). A genome-wide polygenic predictor, or polygenic index (PGI), explains 12–16% of EA variance and contributes to risk prediction for ten diseases. Direct effects (i.e., controlling for parental PGIs) explain roughly half the PGI's magnitude of association with EA and other phenotypes. The correlation between mate-pair PGIs is far too large to be consistent with phenotypic assortment alone, implying additional assortment on PGI-associated factors. In an additional GWAS of dominance deviations from the additive model, we identify no genome-wide-significant SNPs, and a separate X-chromosome additive GWAS identifies 57.

Main findings of 2022 paper

- N=~3 mill individuals
- 3952 GW significant SNPs identified
- GW polygenic predictor (PGI) explains 12-16% of EA variance
- The PGI contributes to risk prediction for 10 diseases

Lecture Outline

General introduction to genetic epidemiology (lecture I)

- Part I
  - What's a complex trait?
  - Genetic basis of complex traits
- Part II
  - Genetic approaches to studying complex traits
  - Candidate-gene analysis, GWAS, and GWAMA