Norwegian Institute of Public Health

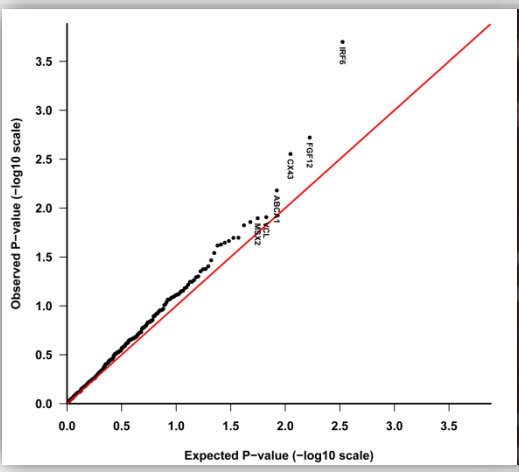
UNIVERSITY OF BERGEN

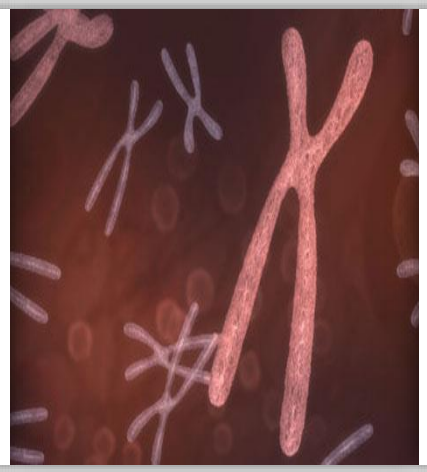
A Gentle Introduction to Genetic Epidemiology

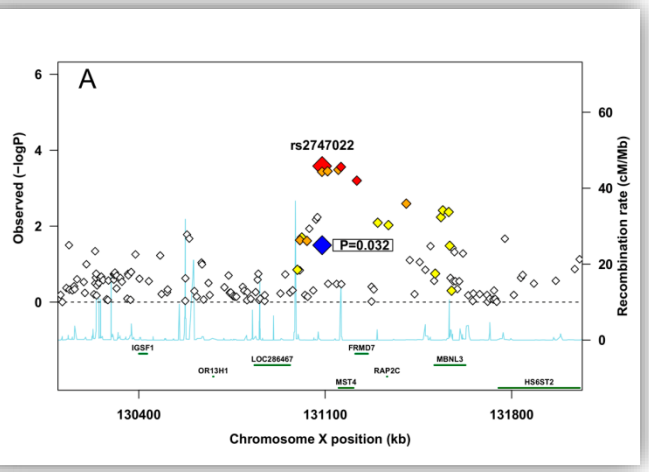
— Lecture 1, Parts I & 2 —

Anil Jugessur

Senior scientist, Norwegian Institute of Public Health, Oslo







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

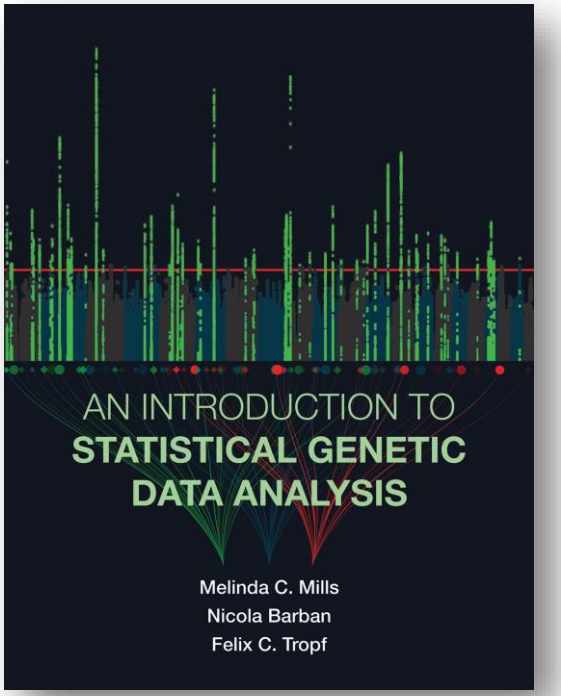
Lecture Outline

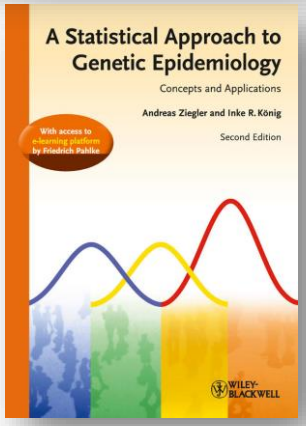
General introduction to genetic epidemiology (lecture I)

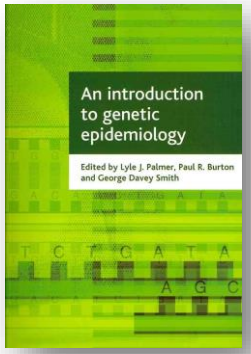
- Part I
  - What's a complex trait?
  - Genetic basis of complex traits
- Part II
  - Genetic approaches to studying complex traits
  - Candidate-gene analysis, GWAS, and GWAMA

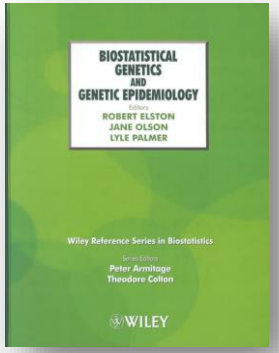
COURSE BOOK

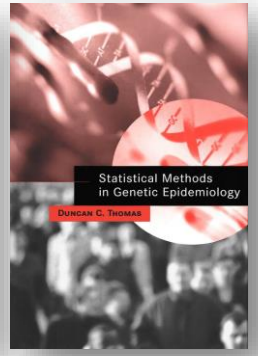
Course book

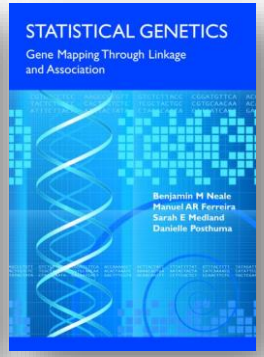


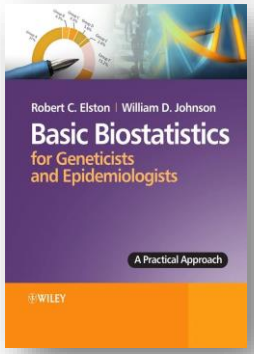


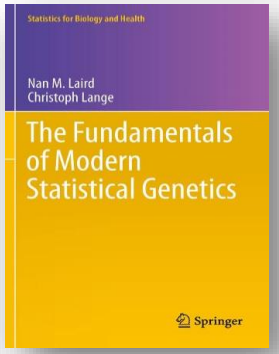


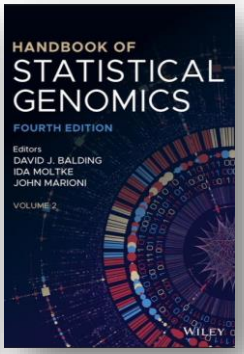


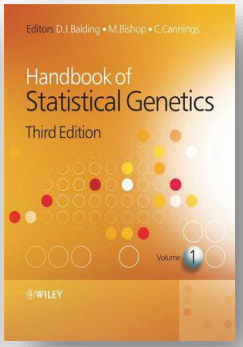








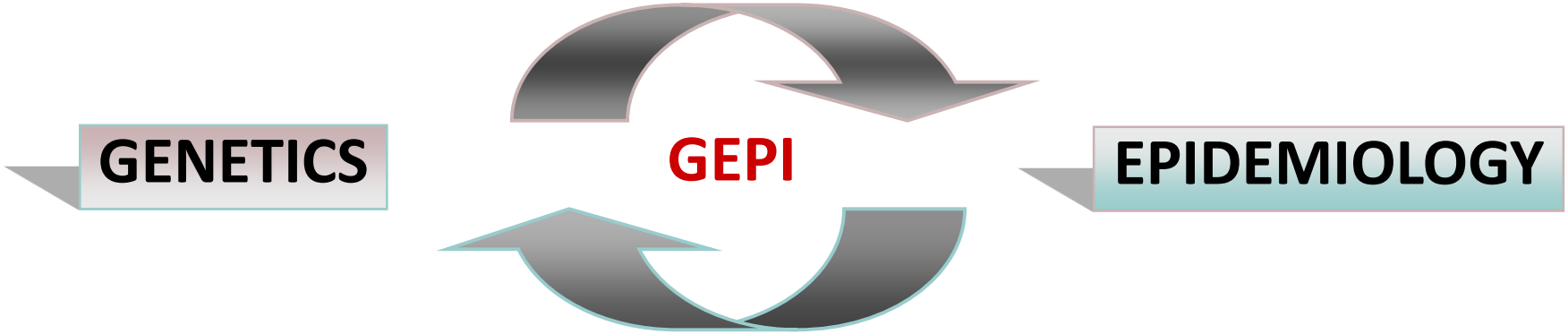




WHAT IS GENETIC EPIDEMIOLOGY?

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."

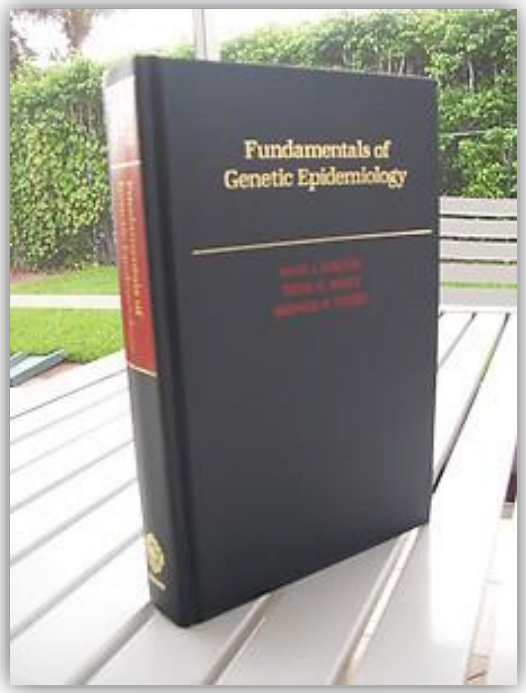
P. Phillippe (1982): "Genetic epidemiology studies the interaction between genetic and environmental factors at the origin of disease."

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."

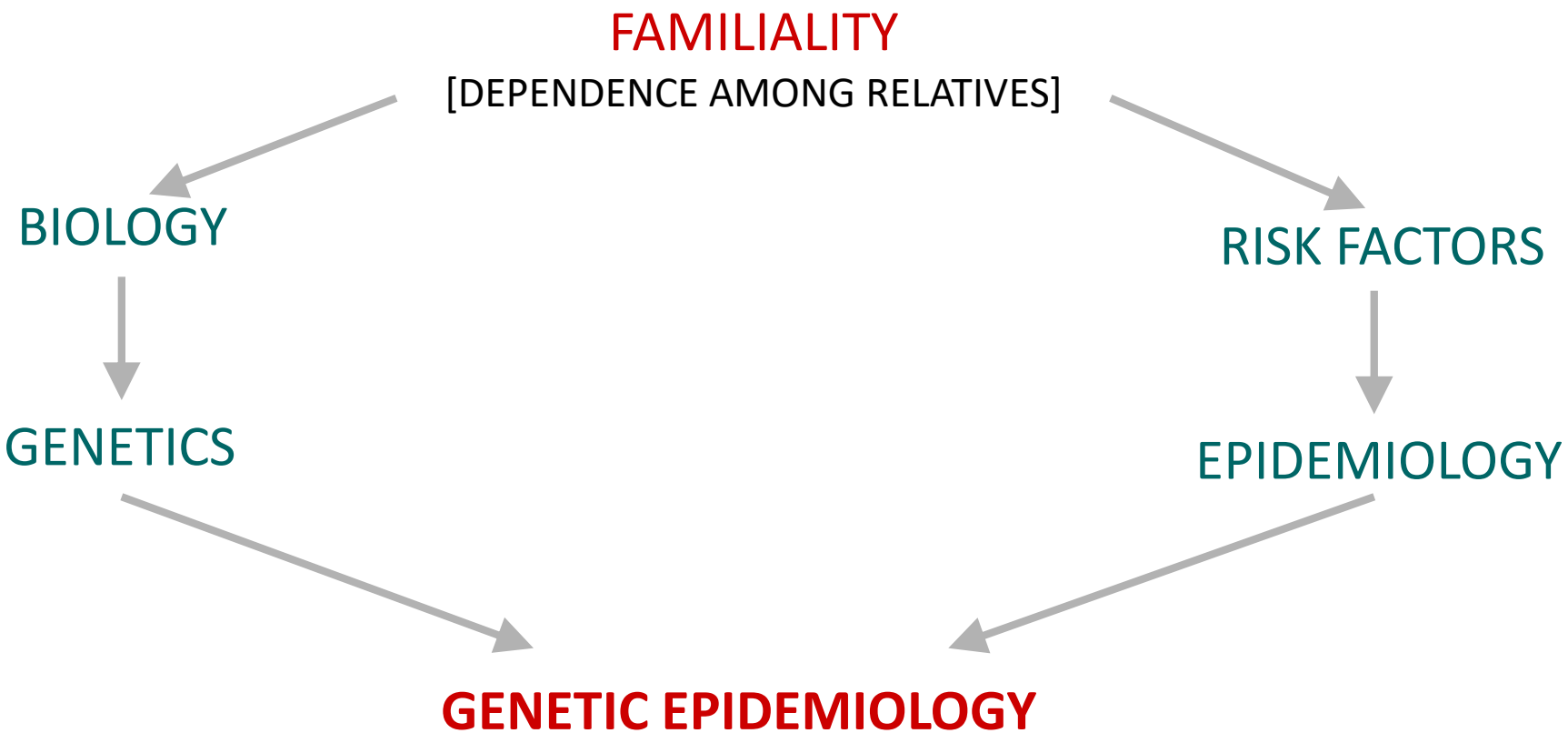
E.A. Thompson (1986a): "Genetic epidemiology is the analysis of the familial distributions of traits, with a view to understanding any possible genetic basis."



Prof of biostats @ WASH-U.

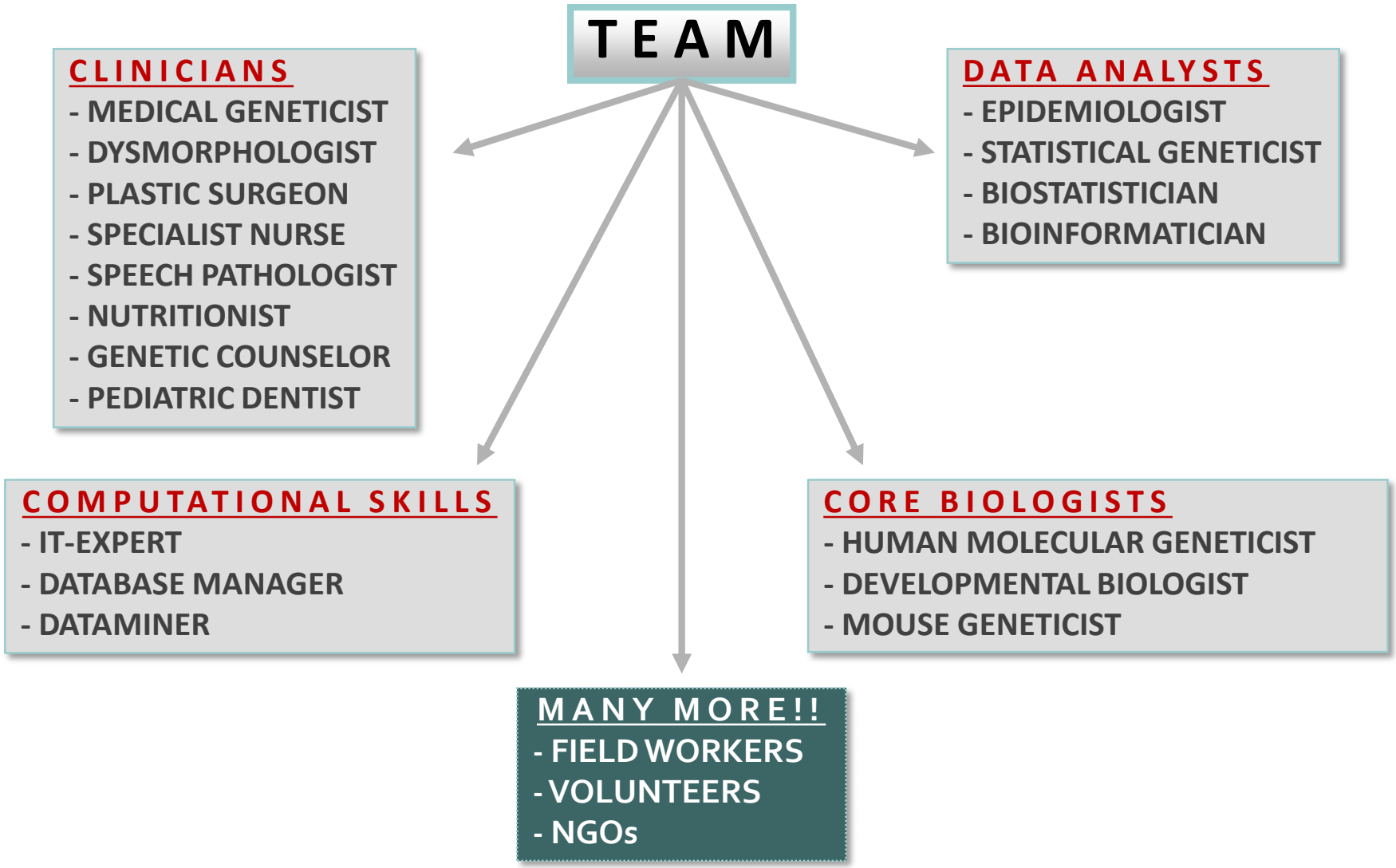
Source (book): «Fundamentals of Genetic Epidemiology», Eds. Khoury, Beaty and Cohen. Oxford Press, 1993.

Growing Convergence Of Different Fields

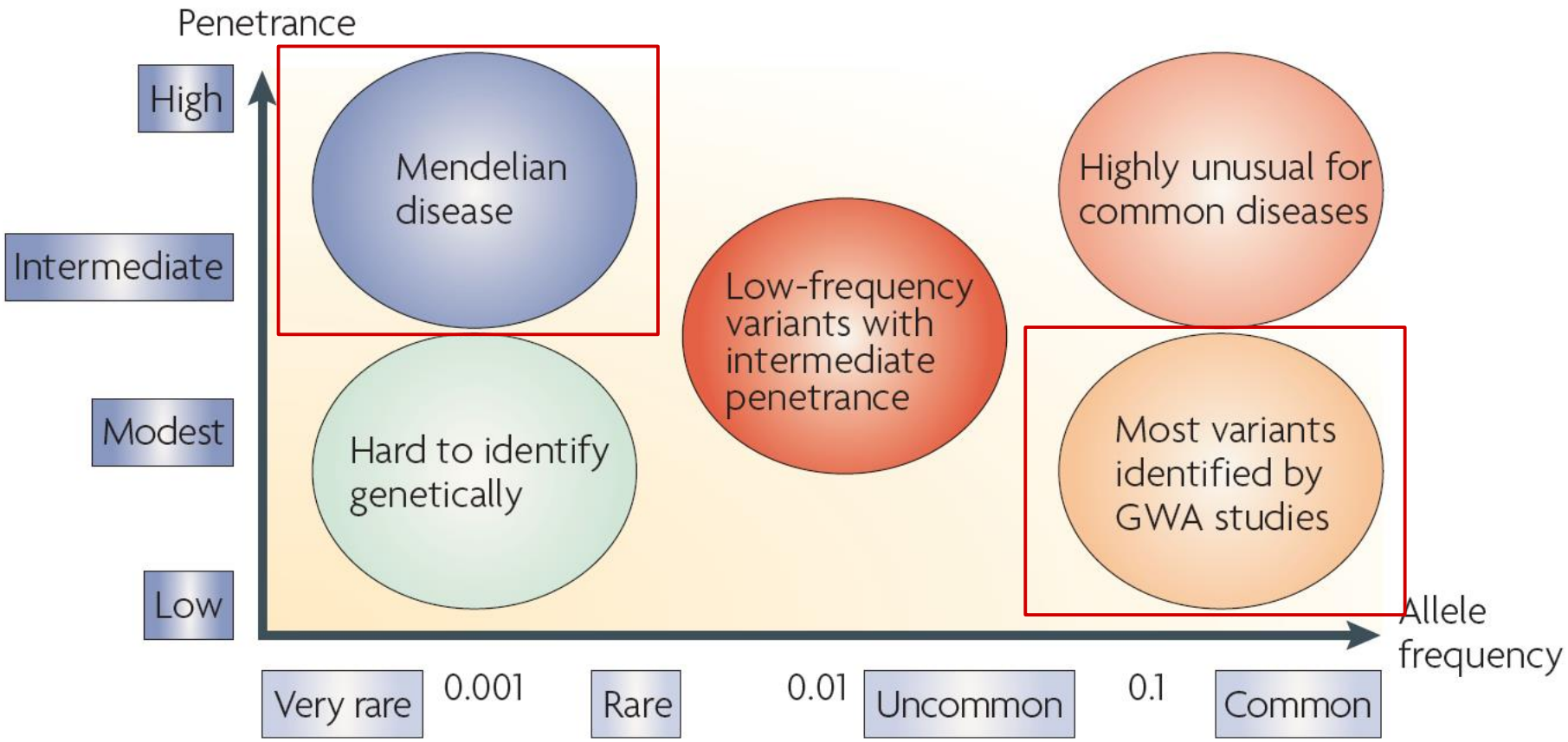


"Less divergence in terminology and methodology, and an increased conversation, collaboration and convergence across the fields."

BUILDING A TEAM
– E.G. FOR A STUDY OF BIRTH DEFECTS –

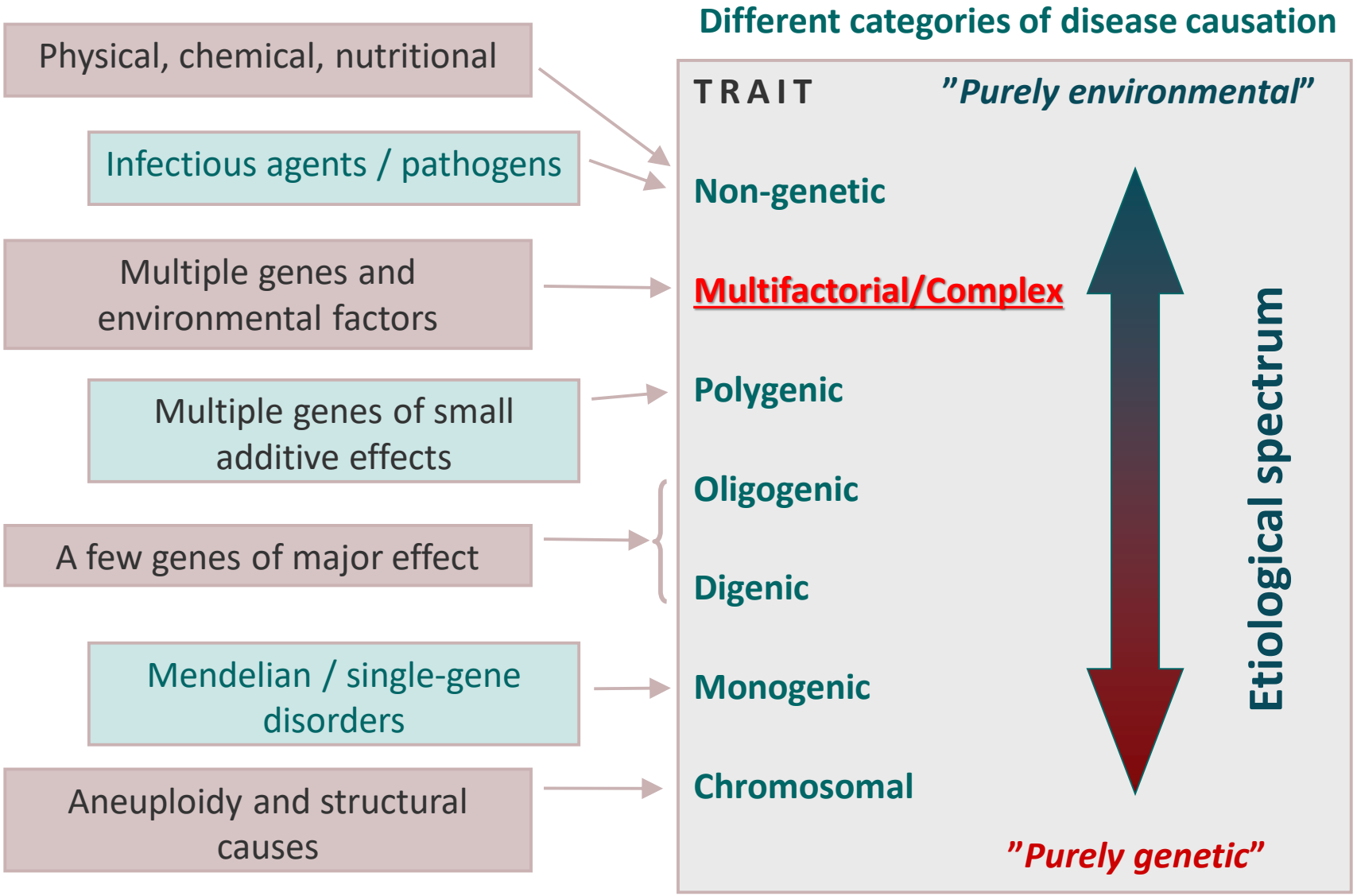


Mendelian disorder vs. Complex trait



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

- **Relatively common**
- **Heterogeneity at several levels:**
  - **Genetic heterogeneity:**
    - **«*locus*» and «*allelic*» heterogeneity**
- **Incomplete penetrance** ⇒ *not all individuals with mutant genotype express phenotype*
- **Variable expressivity** ⇒ *individuals with mutant genotype show range of phenotypes*
- **Effect of a gene can be masked by:**
  - **Phenocopies** ⇒ *environmentally-caused phenotype mirrors genetically-caused trait*
  - **Pleiotropy** ⇒ *mutations affect different traits or organs*
- **Complex interactions:**
  - **«gene-gene» and «gene-environment» interactions**
- **Stochastic effects** ⇒ *random or chance events; biological processes are error-prone!*

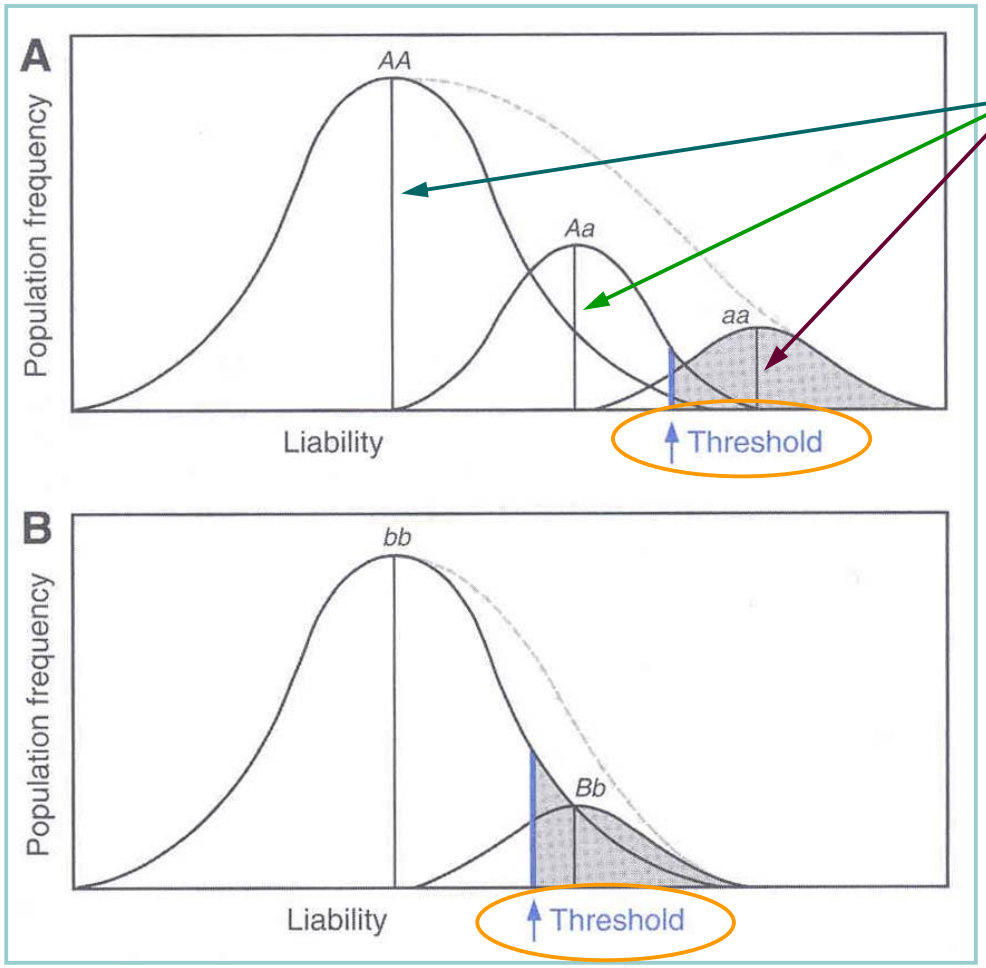
THE CONCEPT OF «LIABILITY»

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 attributable to genetic differences.

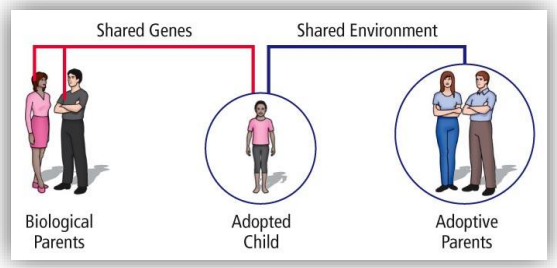
- Broad-sense vs. narrow-sense heritability
  - Broad-sense heritability is the proportion of variance in a phenotype (*Vp*) attributable to the total genetic variance (*Vg*). $H^2=Vg/Vp$, where $Vp=Vg+Ve$
  - Narrow-sense heritability is the proportion of *Vp* attributable to additive genetic variance (*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 of a single gene combine such that their combined effects on the phenotype equal 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?

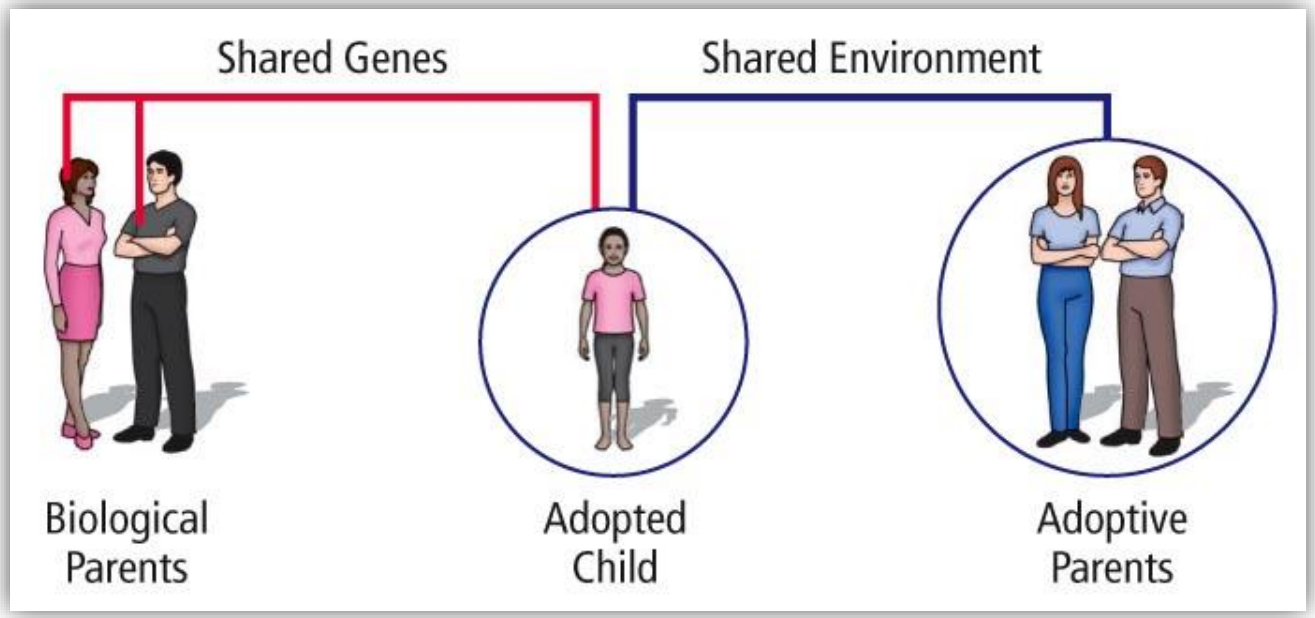
- Study whether the disease clusters in families:
  - Familial aggregation studies:
    - Relatives share a greater proportion of their alleles
      - Affected individuals will tend to cluster in families.
    - Reccurence risk measured as relative risk ratio (λ_r)
      - λ_r = **[risk to relatives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

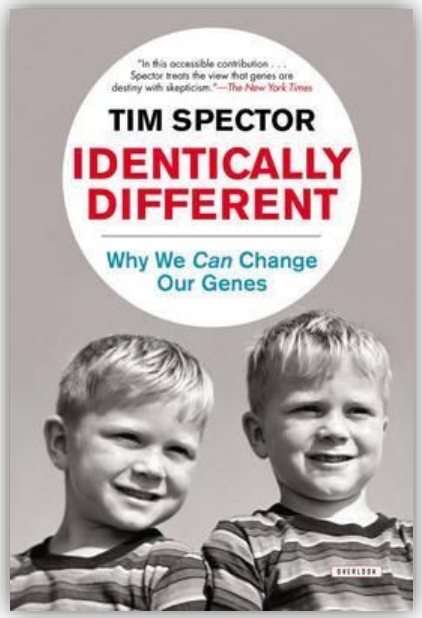
  - Adoption studies:
    - If a trait has a genetic influence, the risk of disease should be higher in biological relatives than in adopted relatives living in the same household.



  - 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 is determined primarily by genetic mechanisms.
      - 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

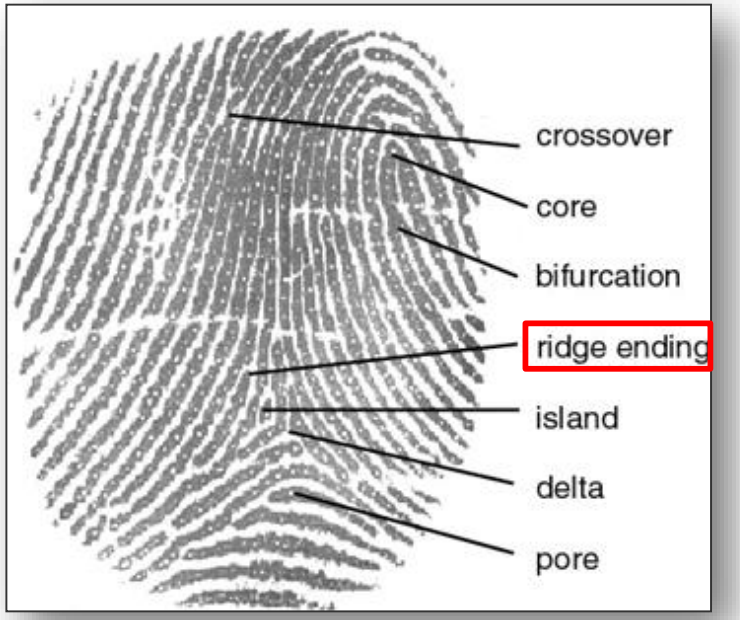
Usual to look at two types of correlations between relative pairs:

- «INTER»class correlation
  - Involves two different classes of relatives:
    - E.g. husband-wife, parent-offspring, brother-sister, grandparent-grandchild, etc.
    - Can distinguish between the members of a given (x,y) pair of relatives

- «INTRA»class correlation
  - Involves only a single class of relatives:
    - E.g. brother-brother, sister-sister, etc.
    - Cannot distinguish between the two members of a given (x,y) pair of relatives.

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**

How do we interpret these correlations?

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 parent-offspring
  - Don't see same degree of correlation in spouse group
    - Consistent with a lesser degree of genetic sharing between spouses.
- In **Dataset II**, higher spouse correlation may be due to shared spousal environment (some degree of 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:

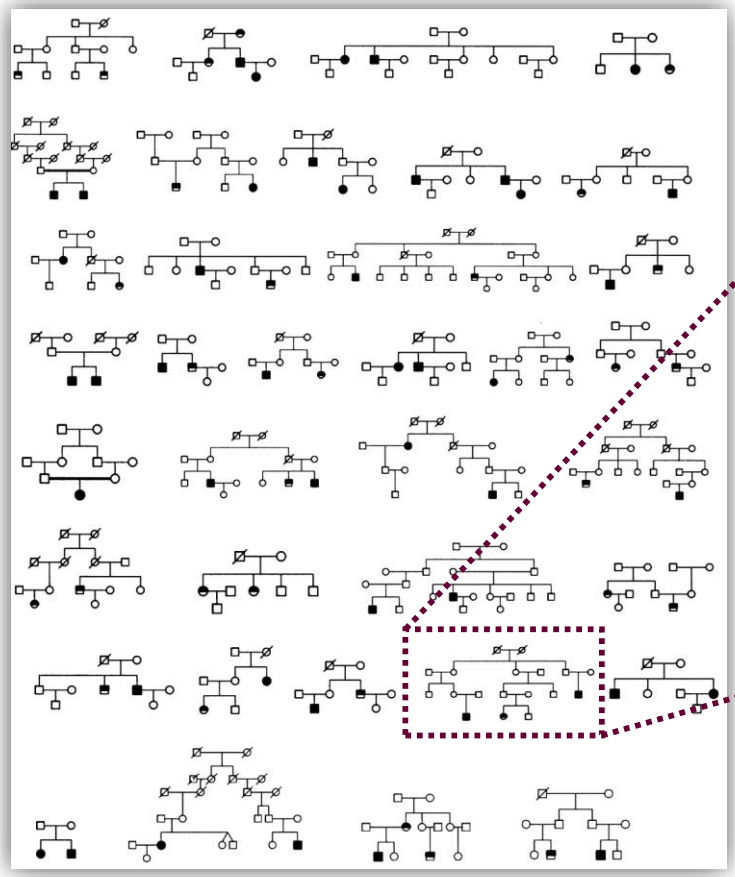
- **Linkage studies** in families with multiple affected members ('multiplex')
  - Test for cosegregation of a marker with the disease phenotype
    - To see if the genetic marker and disease gene are physically linked
  - Problematic for complex diseases because of a lack of multiplex families

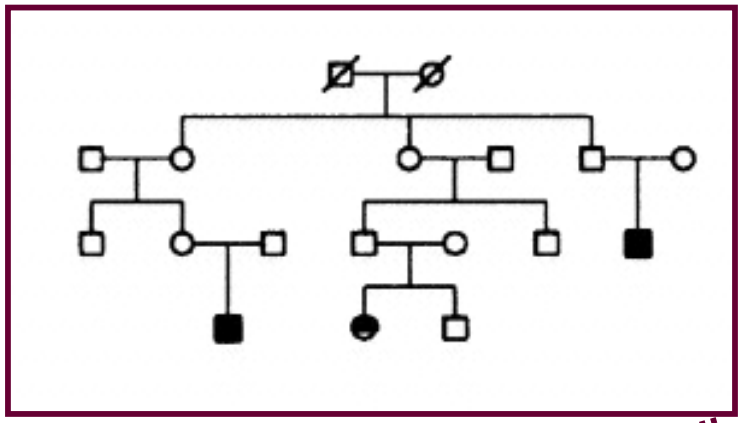
- **Allele-sharing studies** in affected relative pairs
  - Apply model-free methods on smaller subunits within multiplex families
  - «**Identity by descent**» (IBD) methods
    - Knowledge of transmission not required (non-parametric, or model-free)
    - Reasonably good power to detect genes of fairly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

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

Linkage Studies In Multiplex Families

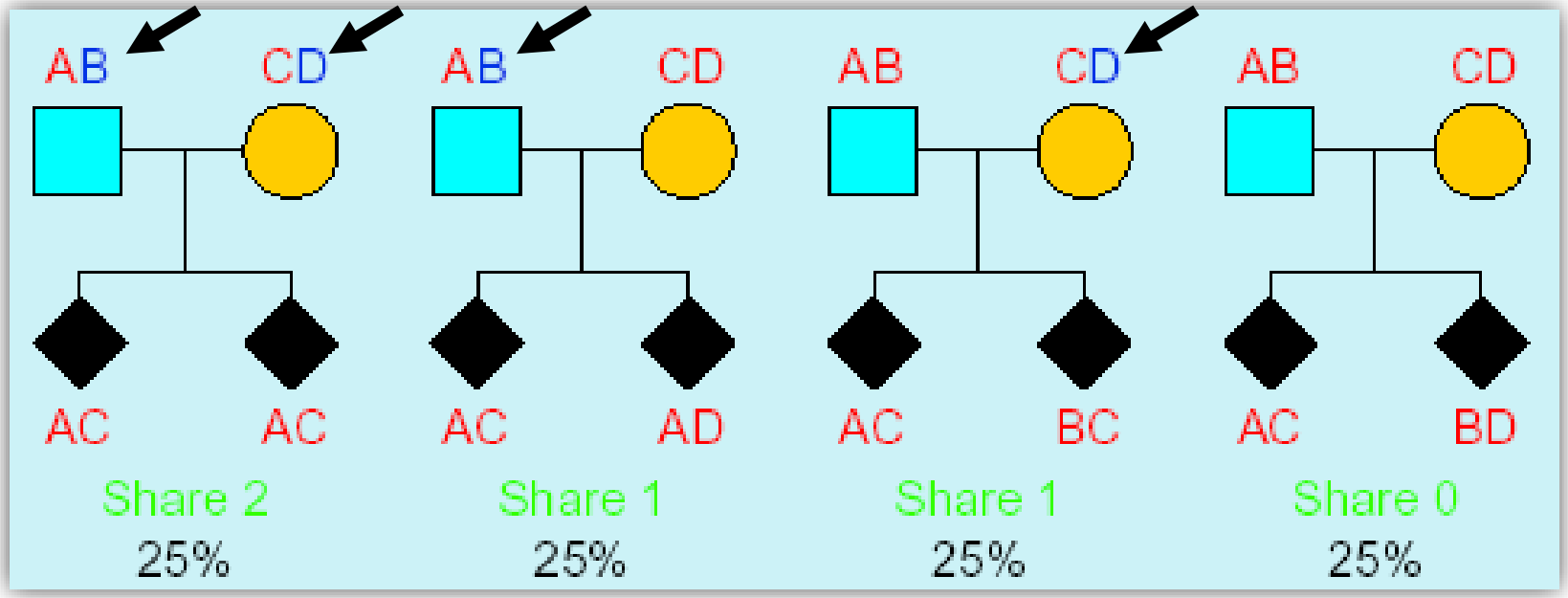
Genomewide linkage analyses can be performed using around 400 microsatellite markers distributed with an average spacing of 10 cM for genomewide coverage.





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 %)

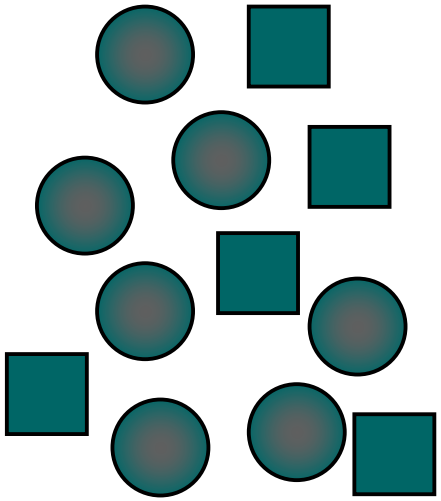
No. of parental alleles shared (% of Mendelian proportion)

Deviations from these expected proportions ⇒ evidence of linkage

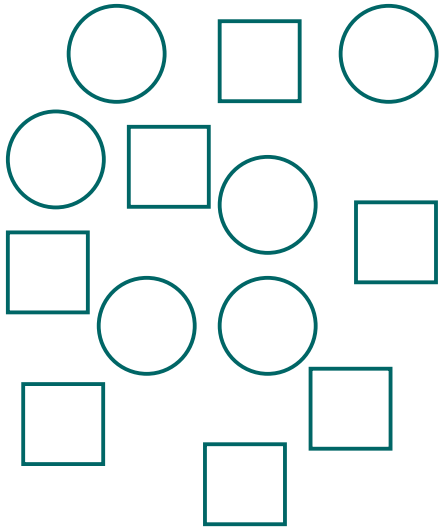
Source: Thomson and Esposito (1999). ***Trends Cell Biol***. 9(12):M17-20.

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.

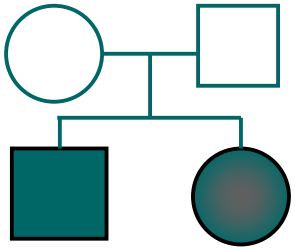


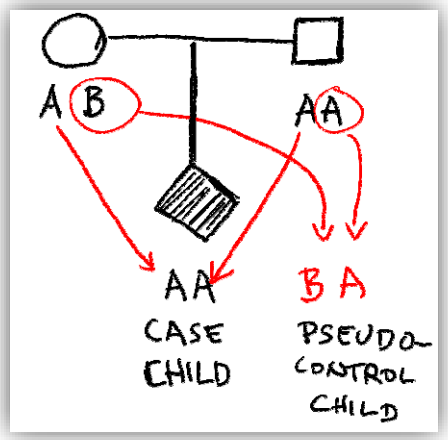
Case (desease)



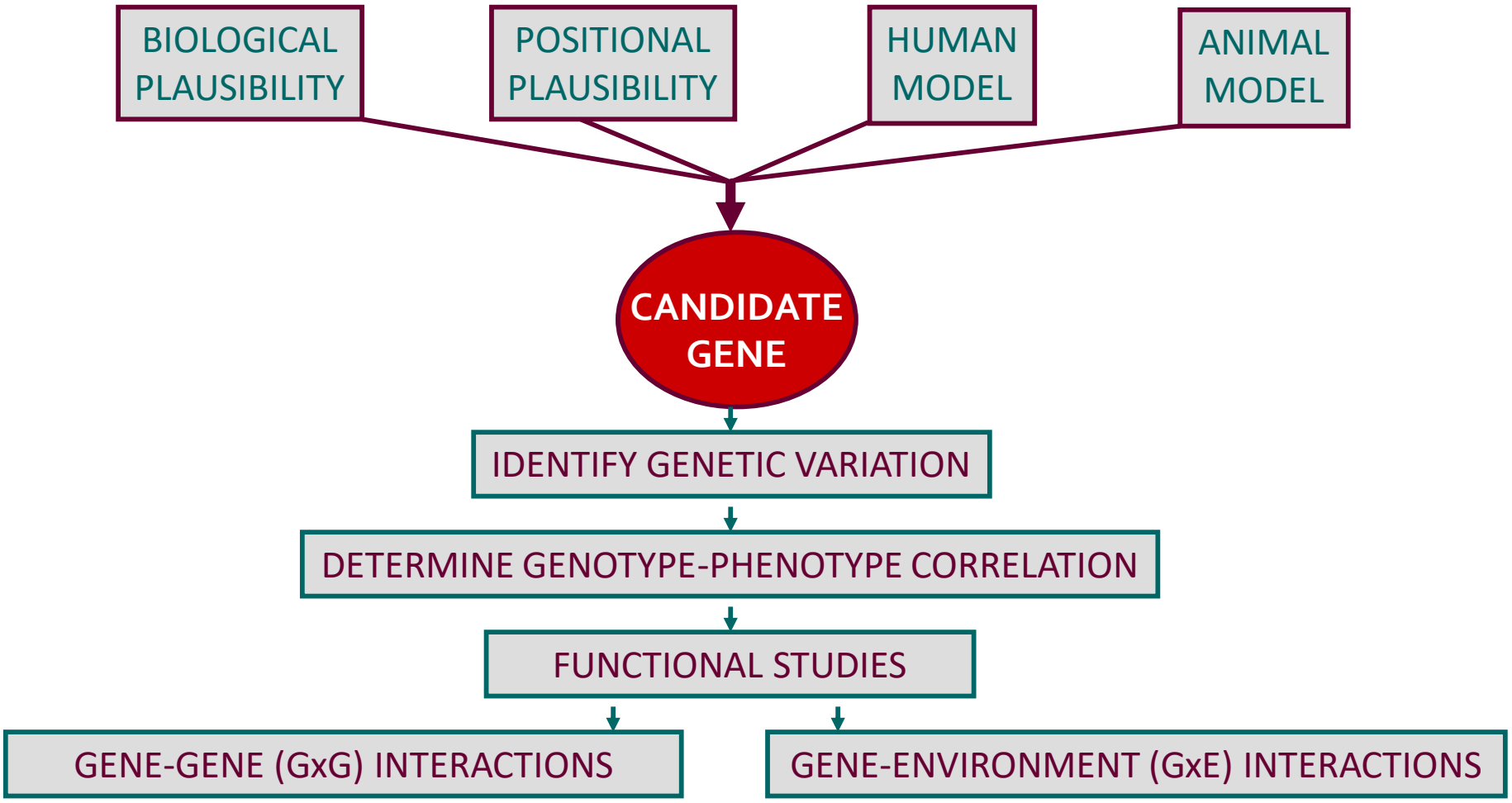
Control (healthy)

vs.



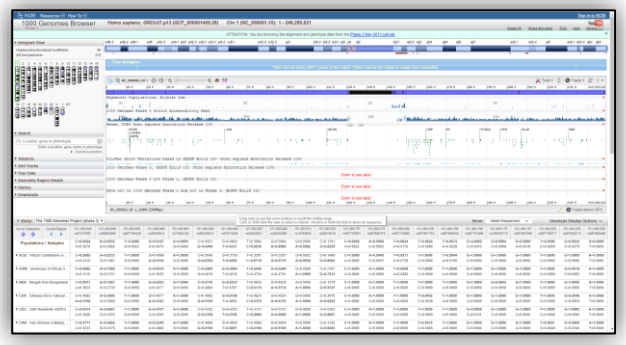


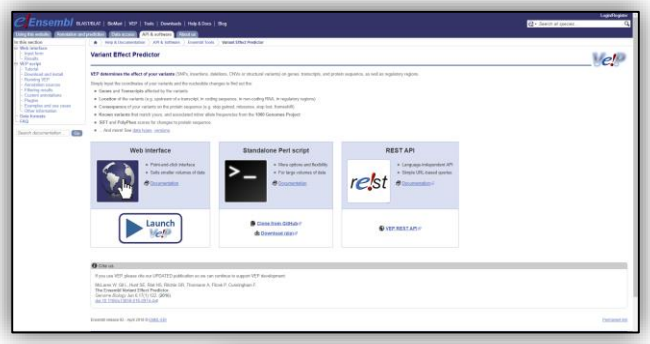
THE CANDIDATE-GENE APPROACH

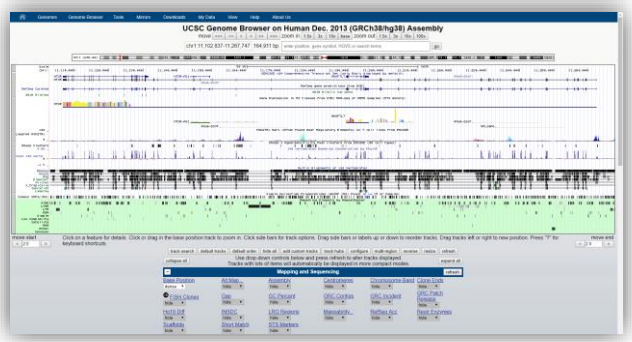


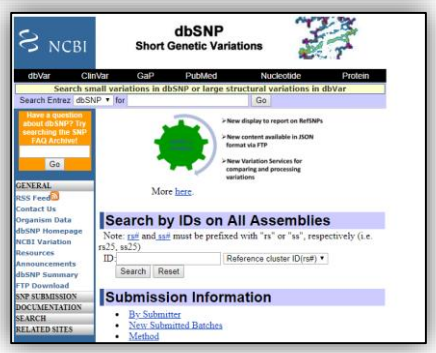
Selecting SNPs for candidate-gene analysis

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









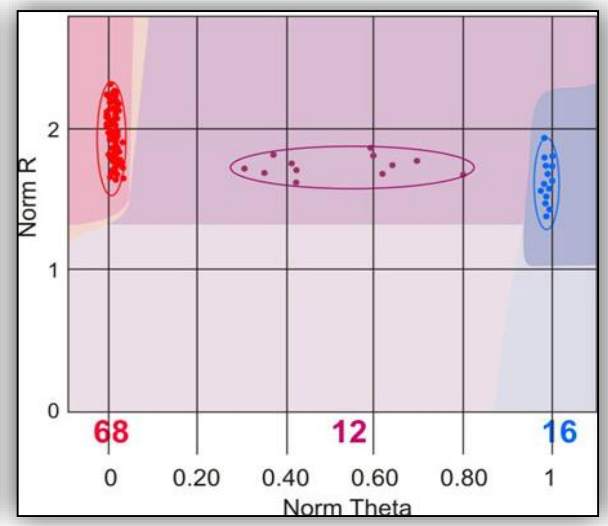
- **Criteria for prioritizing SNP selection:**
  - Prior association with the trait being studied
  - Minor allele frequency (MAF) of at least 5% to capture common variants
  - Preference for coding SNPs and SNPs in regulatory regions – functional!
  - SNPS with «haplotype-tagging» properties

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

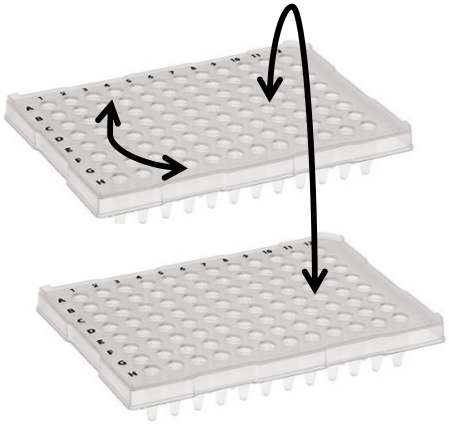
E.g. of genotype calling

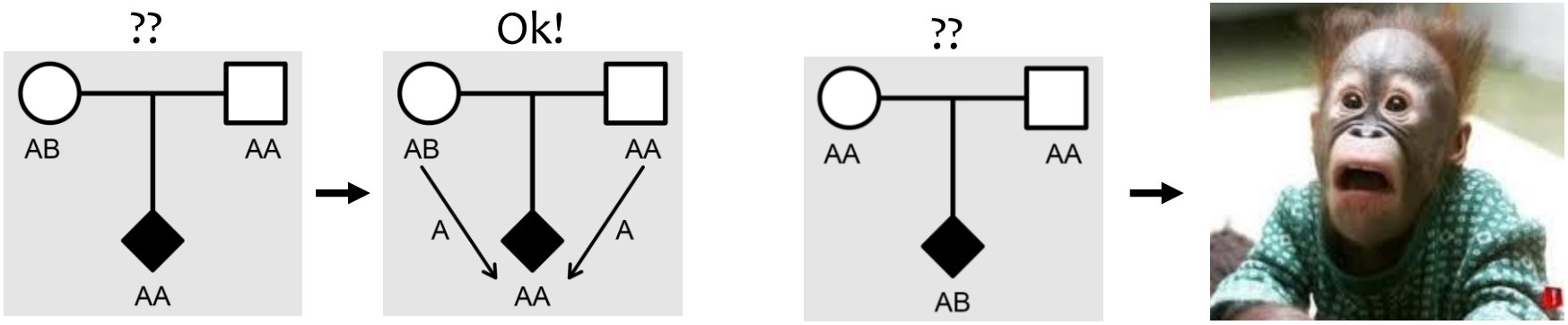


Genomics Core facility Oslo

Data Quality Control (Prelude to Øyvind Helgeland'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, latent population substructure, natural selection etc.
- **Screen for Mendelian errors 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 ☺

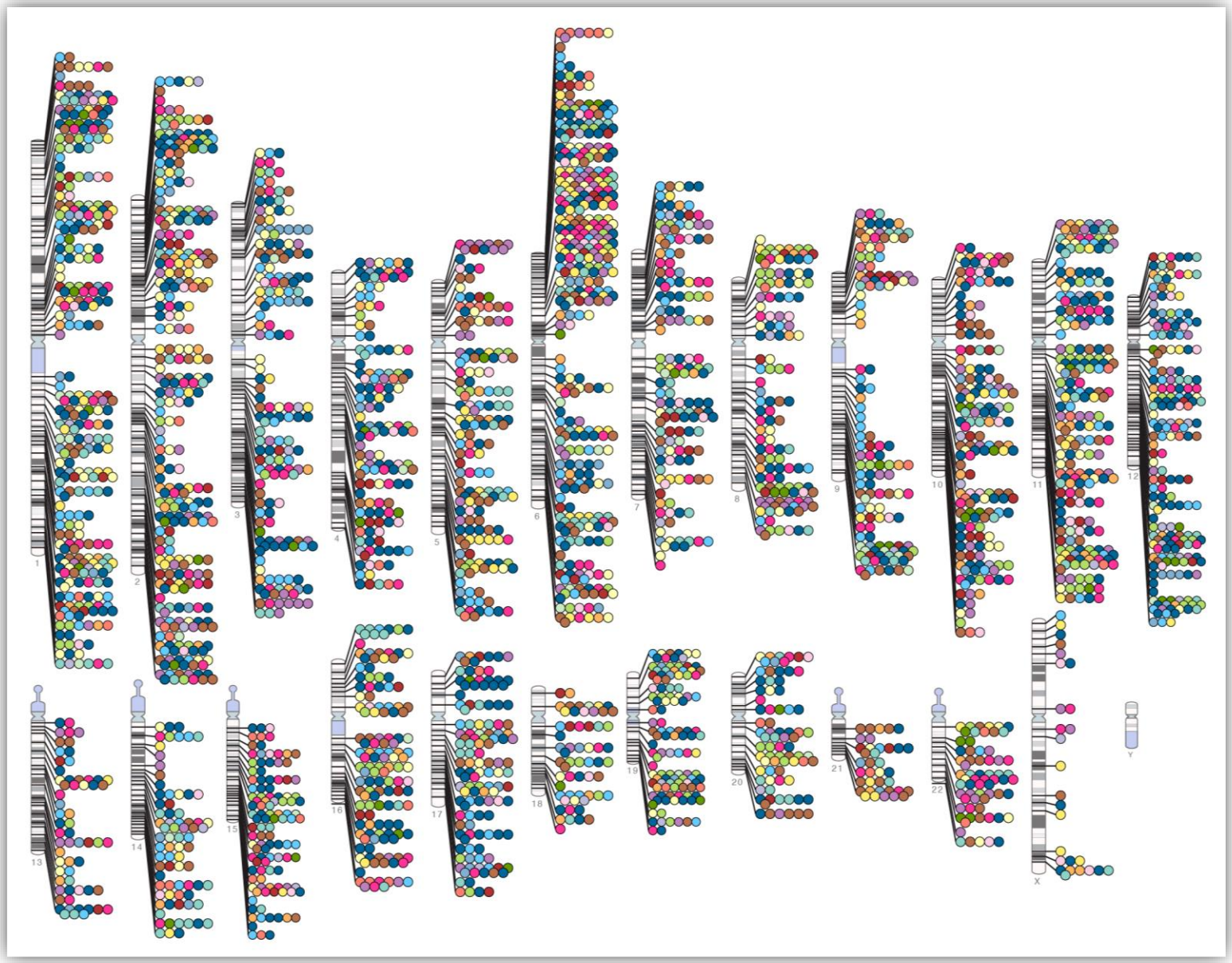
Signals in «gene deserts»:

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

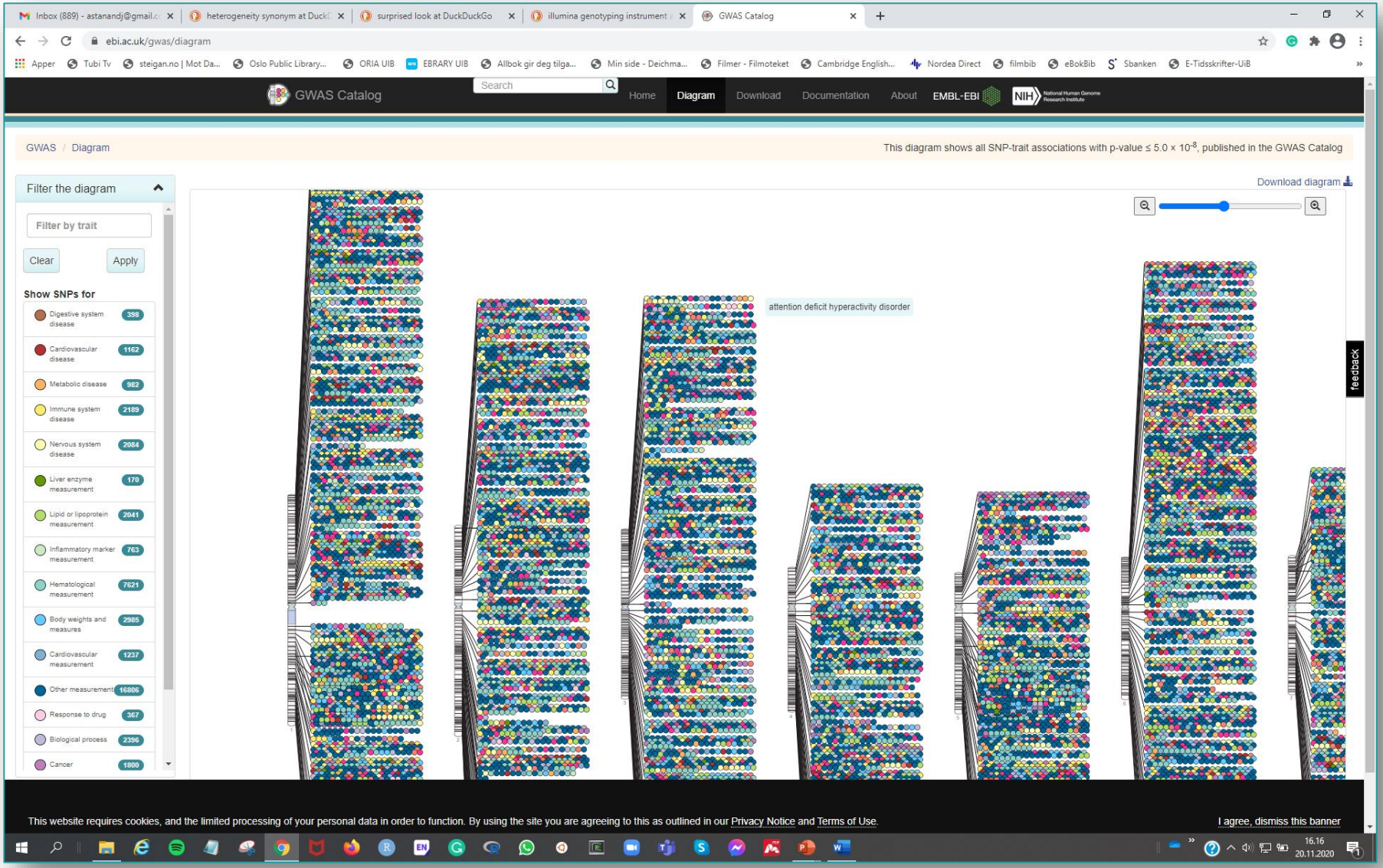
- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other 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



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

↓

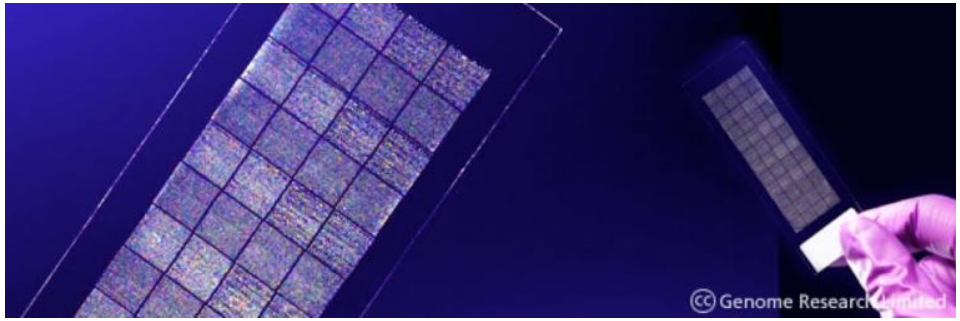
Sequencing / Genotyping

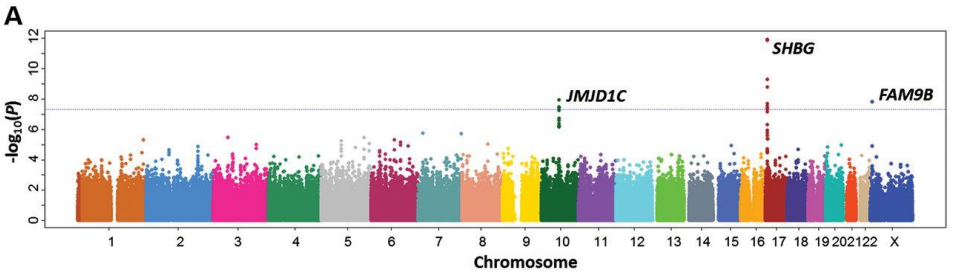
↓

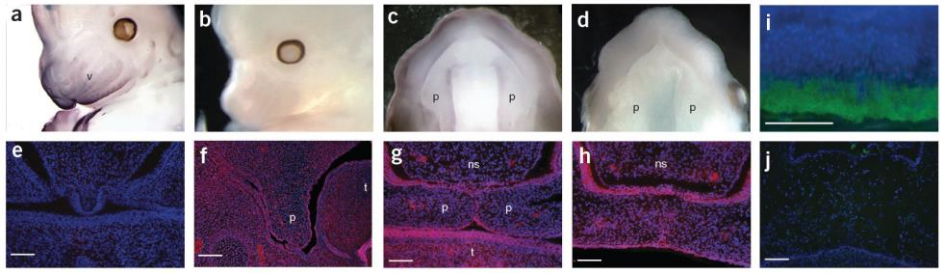
Functional studies

↓

Translational studies





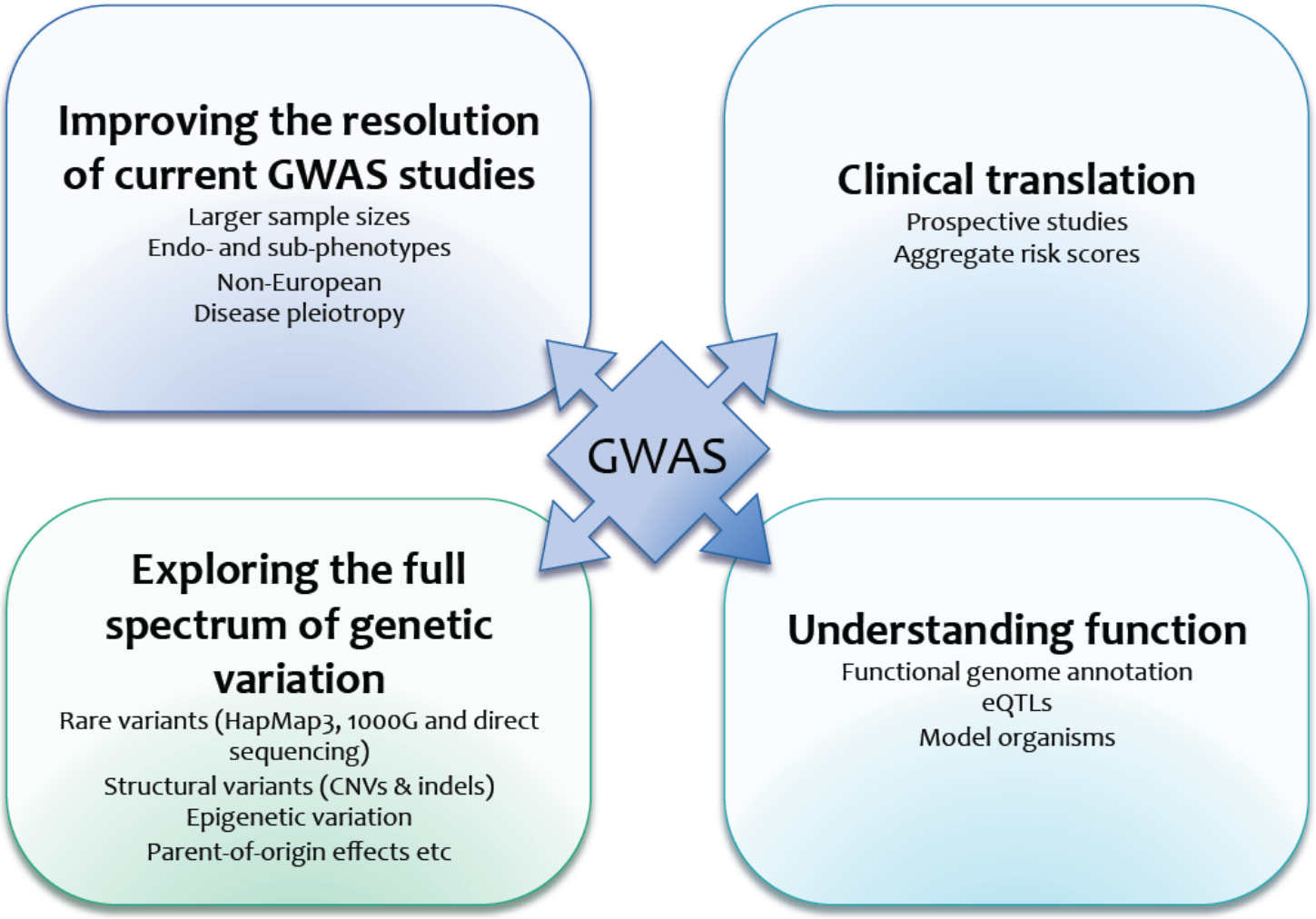


Hope: New therapies, improved diagnostics, better prevention, better public health, precision medicine.

GWAS – What Are The Criteria For Success?

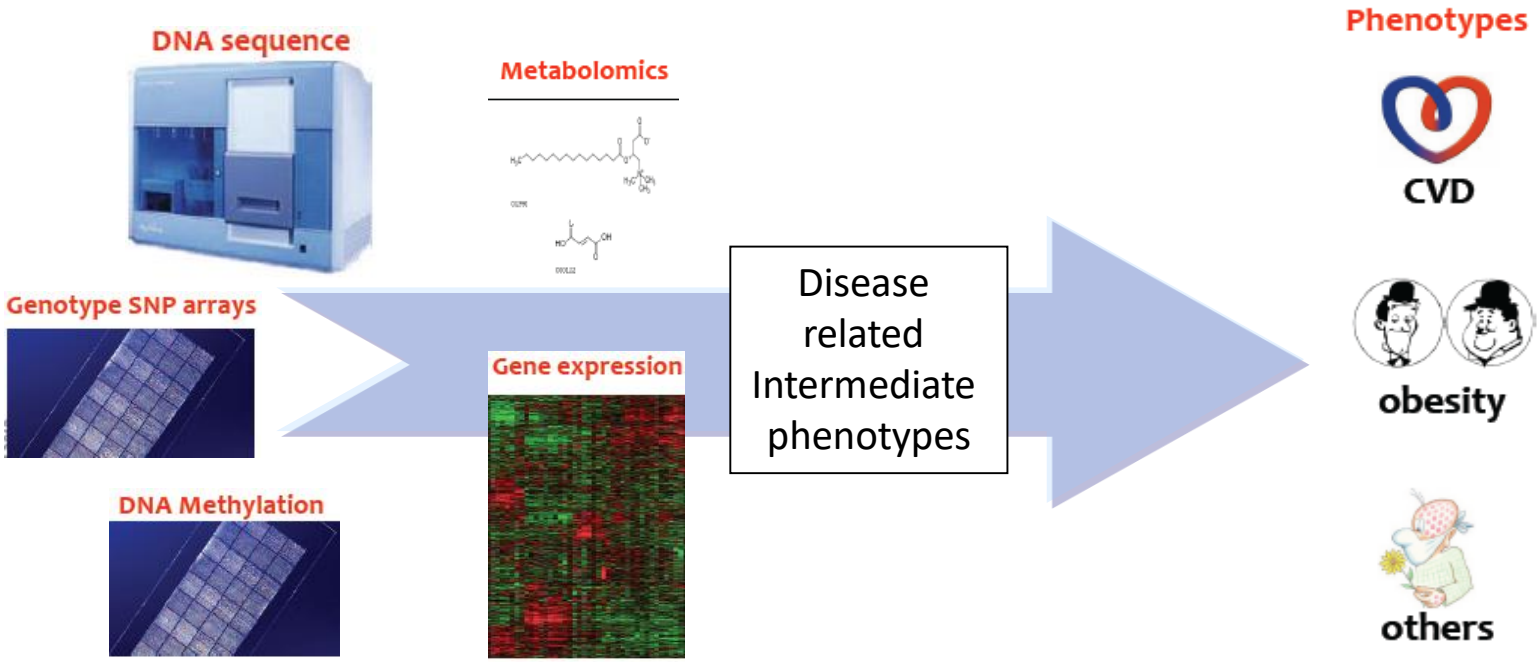
- Costs and availability of large samples are major limitations
  - Useful to conduct meta-analysis of summary data from multiple cohorts
- Strict quality control throughout the process (Øyvind Helgeland's Tuesday lecture) + Stringent significance thresholds + Importance of replication
- Data sharing 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 (Øyvind Helgeland's Tuesday lecture)
  - Analyze cohorts with clearly defined phenotypes and map Mendelian diseases



META-GWAS ANALYSES – A SHORT PRIMER

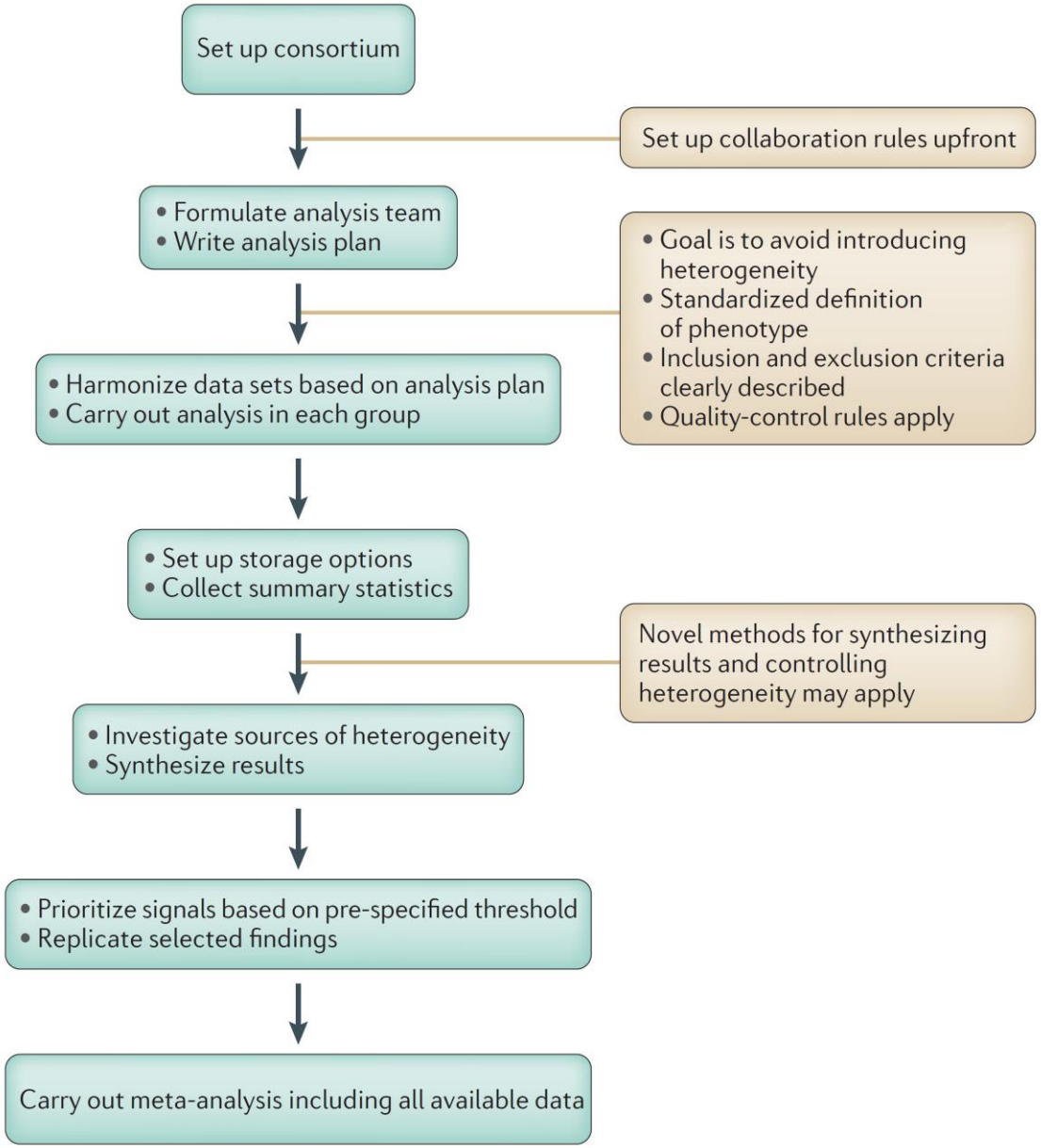
1) DISCOVERY PHASE

- 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

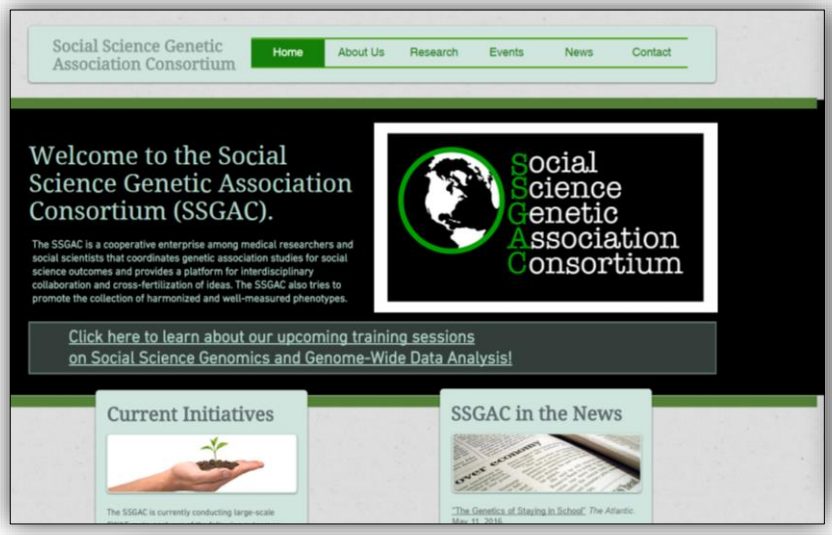
- Invite more cohorts for replication
  - Confirmation of original findings in discovery phase

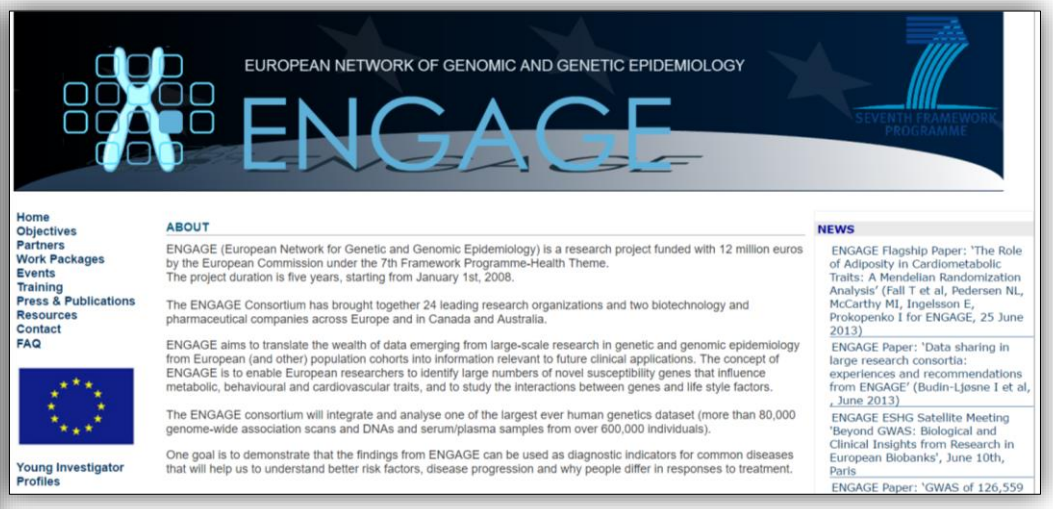
Stages in a Meta-GWAS Analysis



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

Examples of Consortia





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

STANDARD OPERATION PROTOCOL

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)
  - Filters to be applied before imputation (SNP call >95%, HWE p >10e-6, MAF >5%)
- Analysis
  - 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.

EXAMPLE OF A META-GWAS

1) Trait proposed for meta-GWAS: «Aggressive behavior»

- Background of the proposed Meta-GWAS analysis
  - **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: 1000 subjects for at least one rater at 1 age.
  - 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 gender, XXY genotype, 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
- 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…

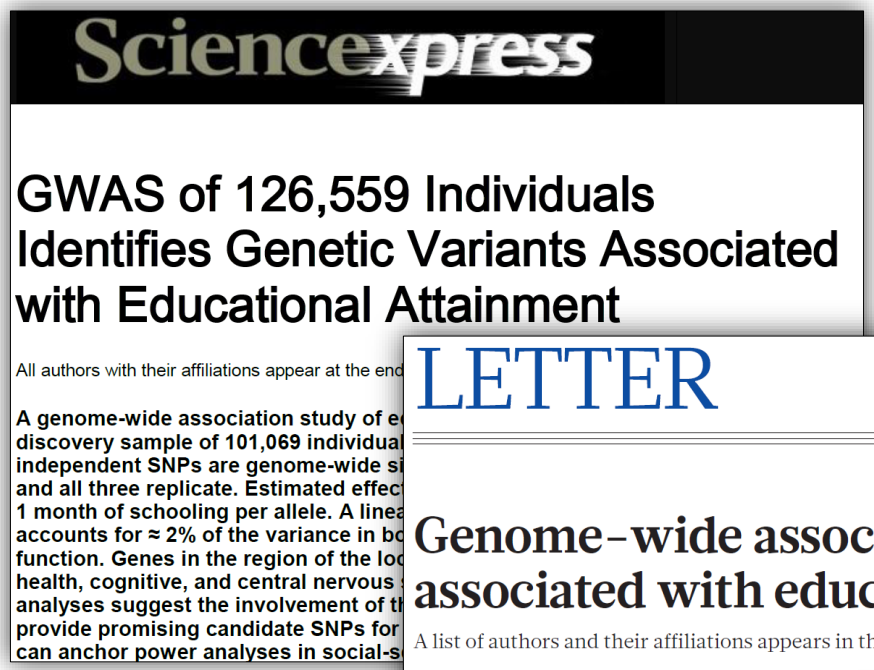
Uploading data

- Use 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 meta-GWAS
- Software: METAL or GWAMA

A Few Examples…



Sciencexpress

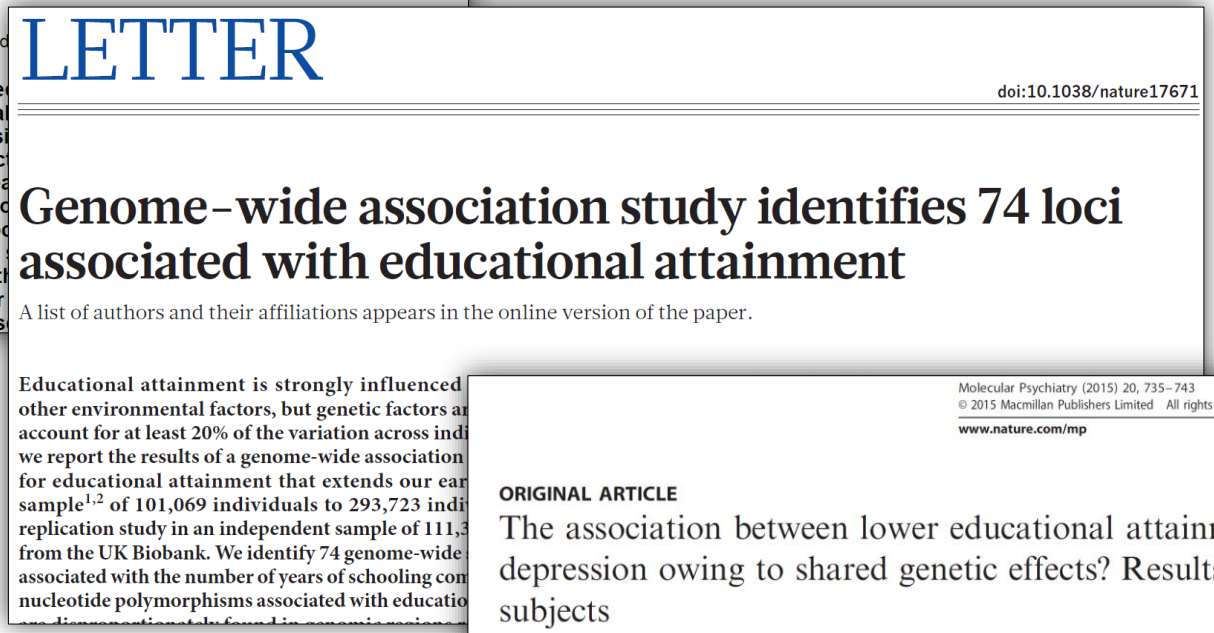
GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

All authors with their affiliations appear at the end

A genome-wide association study of e
discovery sample of 101,069 individual
independent SNPs are genome-wide si
and all three replicate. Estimated effec
1 month of schooling per allele. A linea
accounts for ≈ 2% of the variance in bo
function. Genes in the region of the loc
health, cognitive, and central nervous
analyses suggest the involvement of th
provide promising candidate SNPs for
can anchor power analyses in social-s

N=101,069 inds

N=293,723 inds



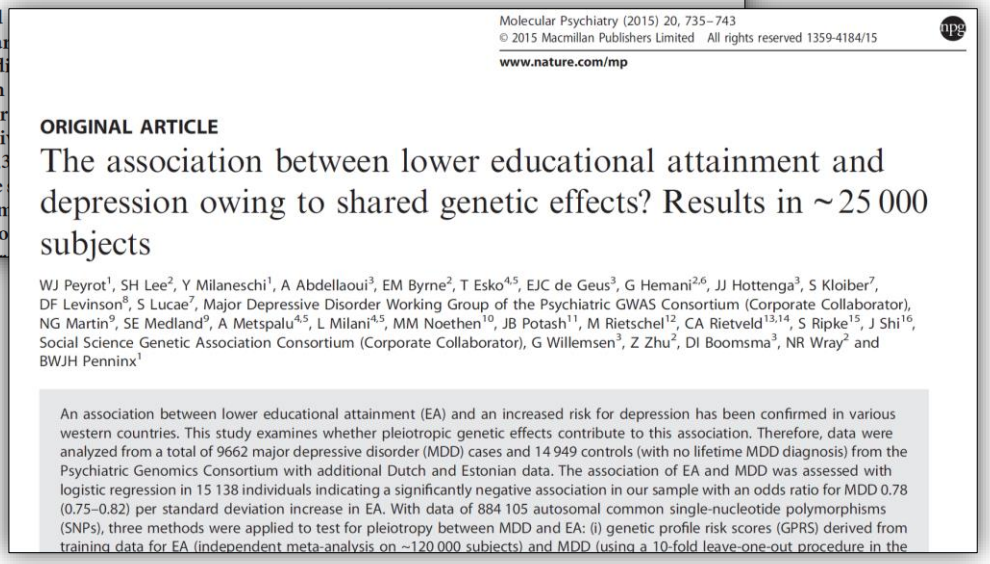
LETTER

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
other environmental factors, but genetic factors ar
account for at least 20% of the variation across indi
we report the results of a genome-wide association
for educational attainment that extends our ear
sample[1,2] of 101,069 individuals to 293,723 indi
replication study in an independent sample of 111,3
from the UK Biobank. We identify 74 genome-wide
associated with the number of years of schooling con
nucleotide polymorphisms associated with educatio



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 Rietveld[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]

An association between lower educational attainment (EA) and an increased risk for depression has been confirmed in various western countries. This study examines whether pleiotropic genetic effects contribute to this association. Therefore, data were analyzed from a total of 9662 major depressive disorder (MDD) cases and 14 949 controls (with no lifetime MDD diagnosis) from the Psychiatric Genomics Consortium with additional Dutch and Estonian data. The association of EA and MDD was assessed with logistic regression in 15 138 individuals indicating a significantly negative association in our sample with an odds ratio for MDD 0.78 (0.75–0.82) per standard deviation increase in EA. With data of 884 105 autosomal common single-nucleotide polymorphisms (SNPs), three methods were applied to test for pleiotropy between MDD and EA: (i) genetic profile risk scores (GPRS) derived from training data for EA (independent meta-analysis on ~120 000 subjects) and MDD (using a 10-fold leave-one-out procedure in the

N=25,000 inds

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