NORMENT
Norwegian Centre for
Mental Disorders Research

Introduction to genetics of complex traits

Oleksandr Frei

MIPT-NORMENT, October 21st, 2019

UiO : University of Oslo
Oslo University Hospital
UNIVERSITY OF BERGEN
Haukeland University Hospital
Norwegian Centre of Excellence
The Research Council of Norway

Introduction to genetics of complex traits

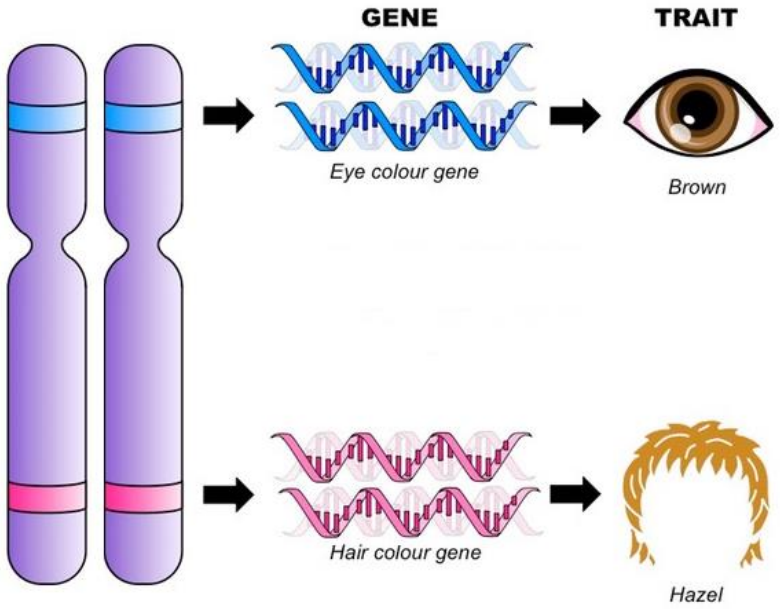
- **Genetics of Complex Traits**
  - Mendelian vs polygenic inheritance
  - Heritability, twin studies
  - Genome-wide association studies
- **Big Data samples**
  - GWAS consortia
  - TOP, MoBa (Norwegian samples)
  - UK Biobank, ABCD
- **Challenges and advanced in statistical analysis**
  - Correlation structure in the genotype matrix
  - Specific tools (conjFDR, MiXeR, MOSTest)

Introduction to genetics of complex traits

- **Genetics of Complex Traits**
  - Mendelian vs polygenic inheritance
  - Heritability, twin studies
  - Genome-wide association studies
- **Big Data samples**
  - GWAS consortia
  - TOP, MoBa (Norwegian samples)
  - UK Biobank, ABCD
- **Challenges and advanced in statistical analysis**
  - Correlation structure in the genotype matrix
  - Specific tools (conjFDR, MiXeR, MOSTest)

Genes and their alleles

- ~24,000 genes in humans
- Most genes exist in many forms called **alleles** (A or a)
- Our cells have two alleles for each gene, one from each parent (AA, Aa, aa)



Huntington’s disease

- Dominant inheritance

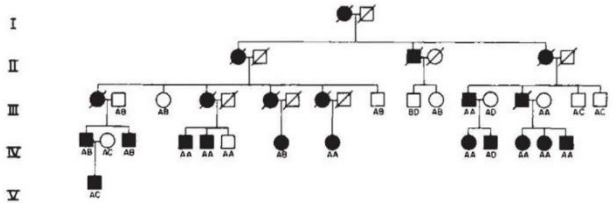


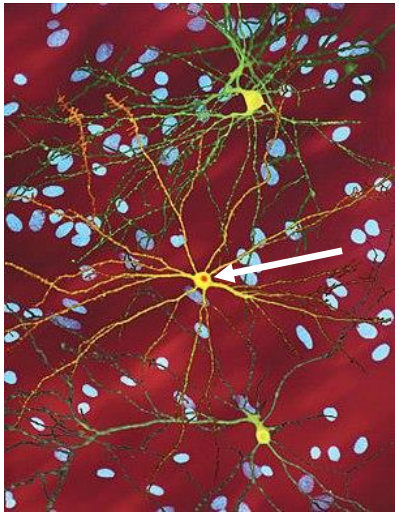
Fig. 1 Pedigree of an American Huntington’s disease family.

- Huntington Gene on chromosome 4

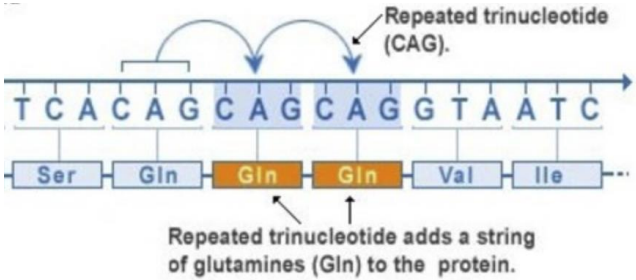
nature
International journal of science

Article | Published: 17 November 1983

A polymorphic DNA marker genetically linked to Huntington's disease



An edited microscopic image of medium spiny neurons(yellow) with nuclear inclusions (orange), which occur as part of the disease process

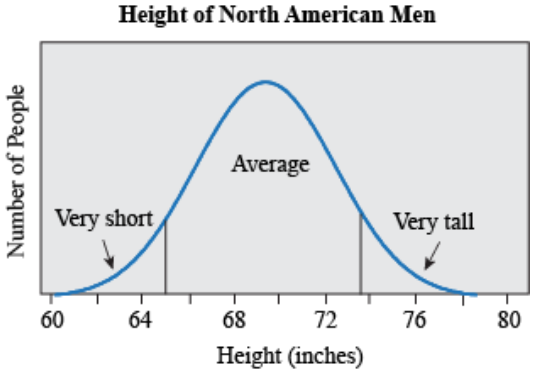


Quantitative genetics: complex traits

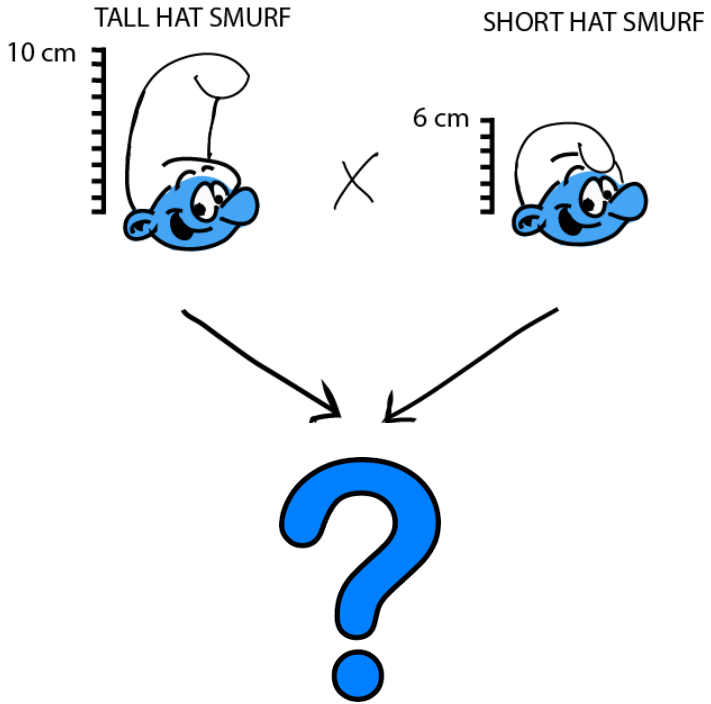
- Most traits are not discrete
- Height is a great example, can range from 74.6 cm (He Pingping) to 272 cm (Robert Pershing Wadlow)

- This also applies to discrete traits such as Schizophrenia diagnosis: Liability (or predisposition) to the disorder is a complex genetic trait

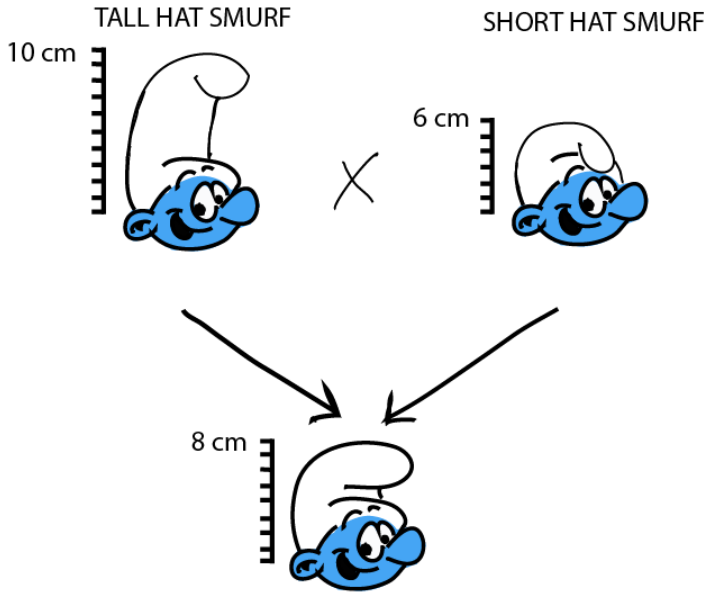
- Many genes involved (“polygenic”)
- Effects of the environment



Additive effects of multiple genes



Additive effects of multiple genes



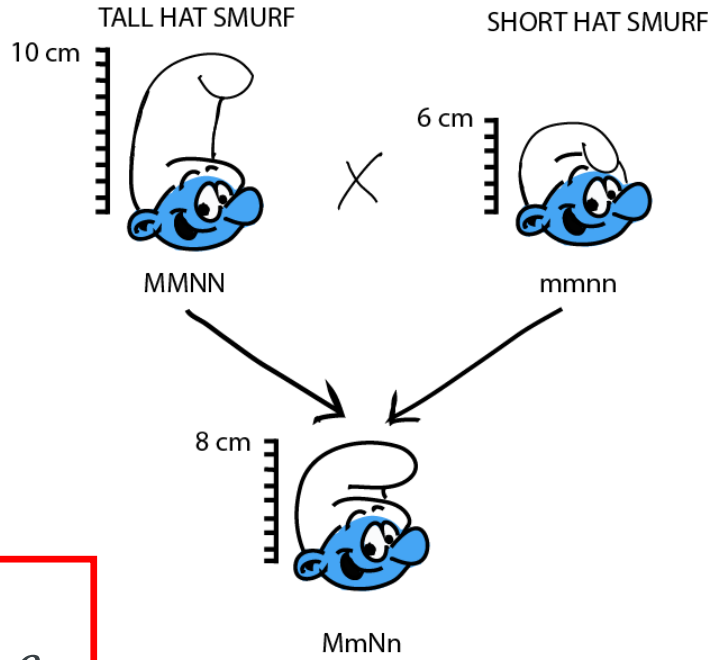
Additive effects of multiple genes

M: +3 cm

N: +2 cm

m: +2 cm

n: +1 cm

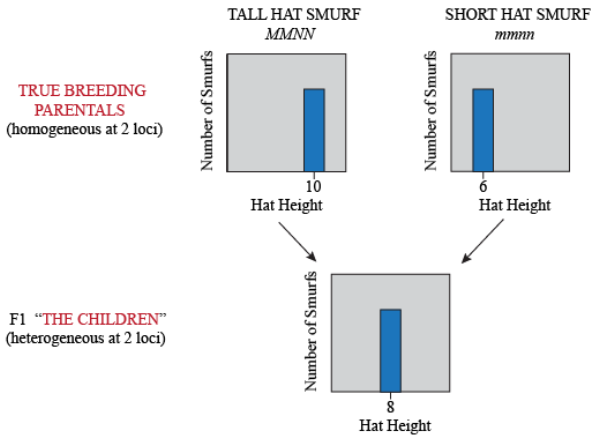


$$y_i = \sum_j w_{ij} u_j + e_i$$

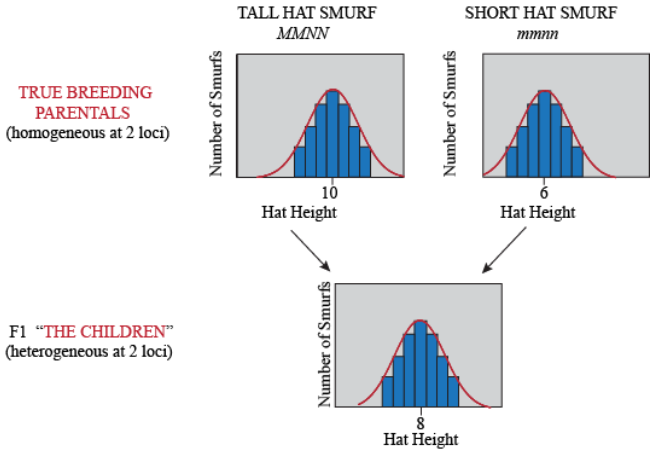
(simple additive genetic model)

Effects of the environment

Assume no environmental effects:



Adding the effect of environment:



Heritability

- Heritability represents the proportion of the phenotypic variation in a population that is explained by genetic factors
- For example, how is intelligence in people determined? Did someone with a high IQ just inherit it from her parents, or does stimulation while growing up make a difference too?

Broad-sense heritability:

$$V_P = V_G + V_E$$

$$H^2 = V_G/V_P$$

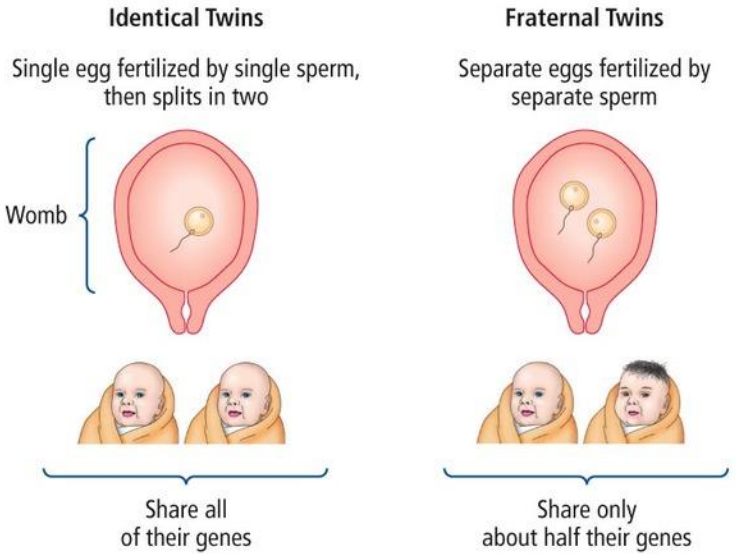
Narrow-sense heritability:

$$V_P = (V_A + V_D + V_I) + V_E$$

$$h^2 = V_A/V_P$$

Twin studies

- Monozygotic (MZ) twins were conceived in a single egg, which later split
- Dizygotic (DZ) twins were conceived when two or more eggs were fertilised at the same time

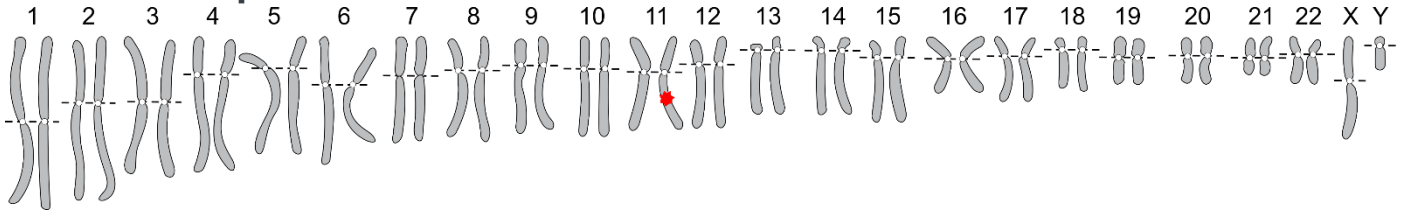


Genes or environment?

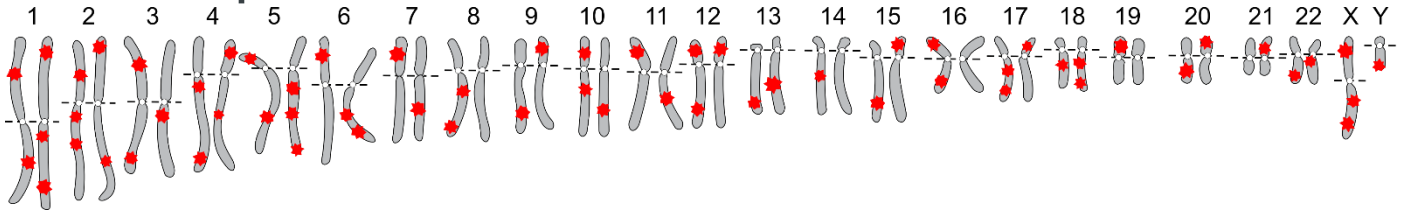
- A famous example would be the Genain quadruplets. Born in 1930, these identical (MZ) German sisters all developed schizophrenia, which suggests a "schizophrenic gene" is at work. The mother and father also had family histories of mental illness, which adds to the credibility of the theory that schizophrenia is at least partly due to genetical factors.

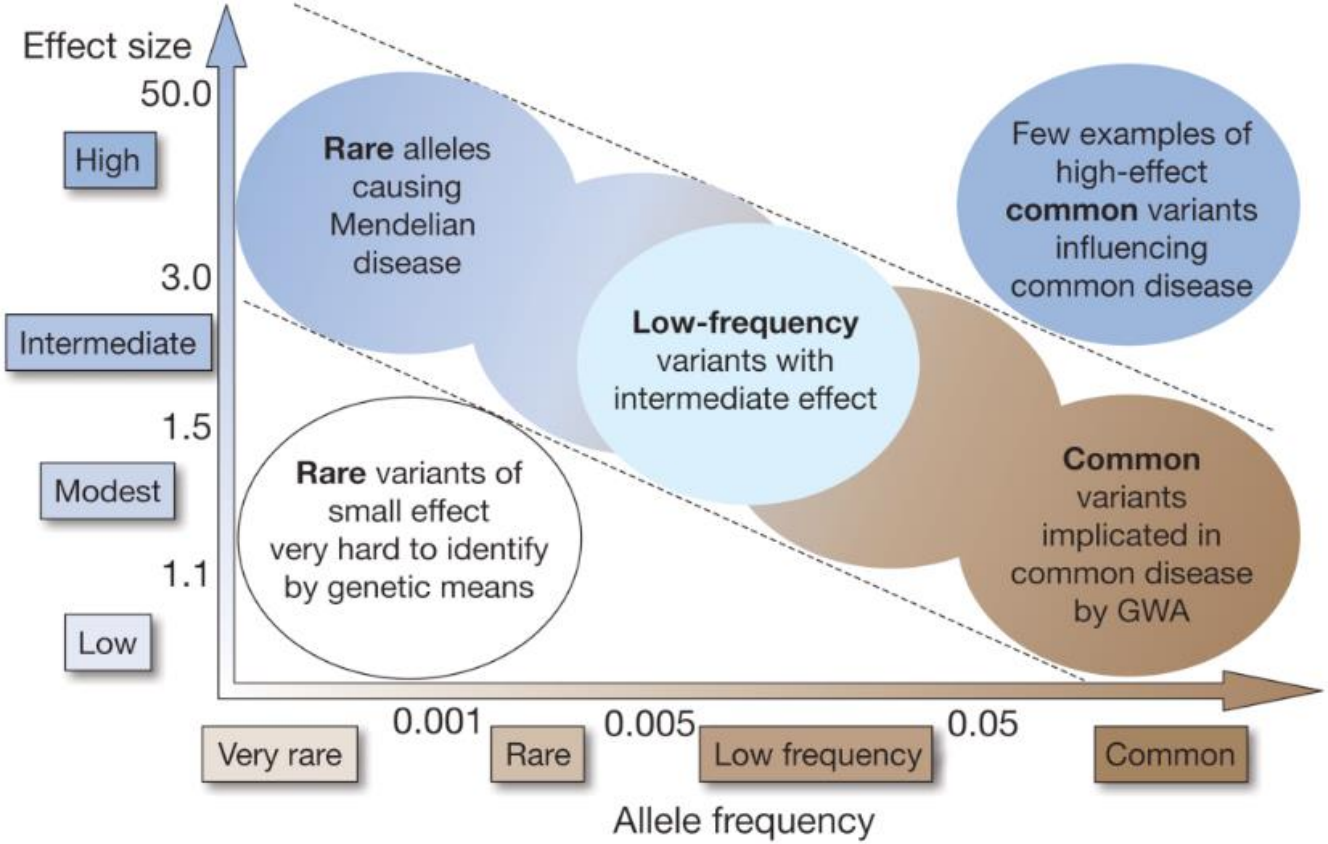
- Schizophrenia heritability: 80-90%
- Schizophrenia concordance rates:
  40-50% MZ twins
  4% DZ twins

Complex traits and GWAS

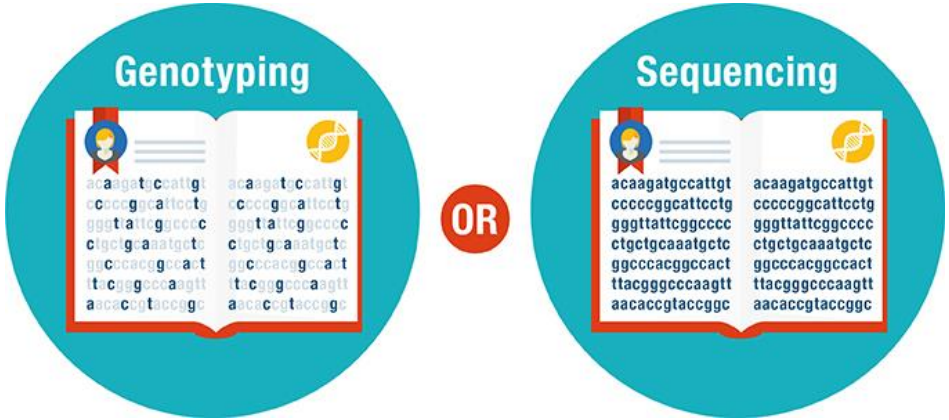


Complex traits and GWAS

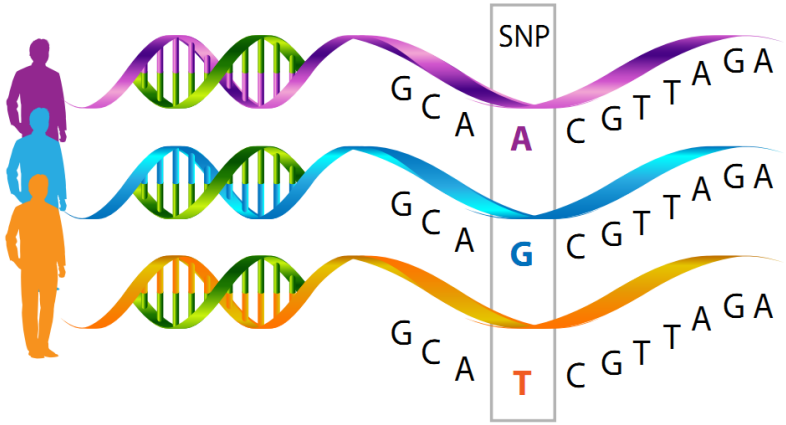



Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

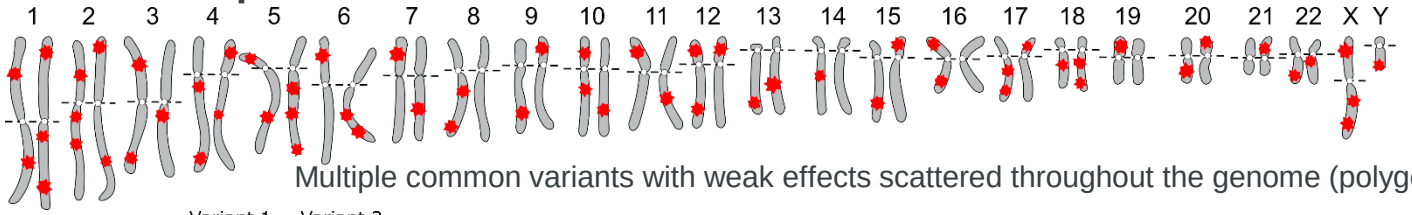
Sequencing and genotyping technologies



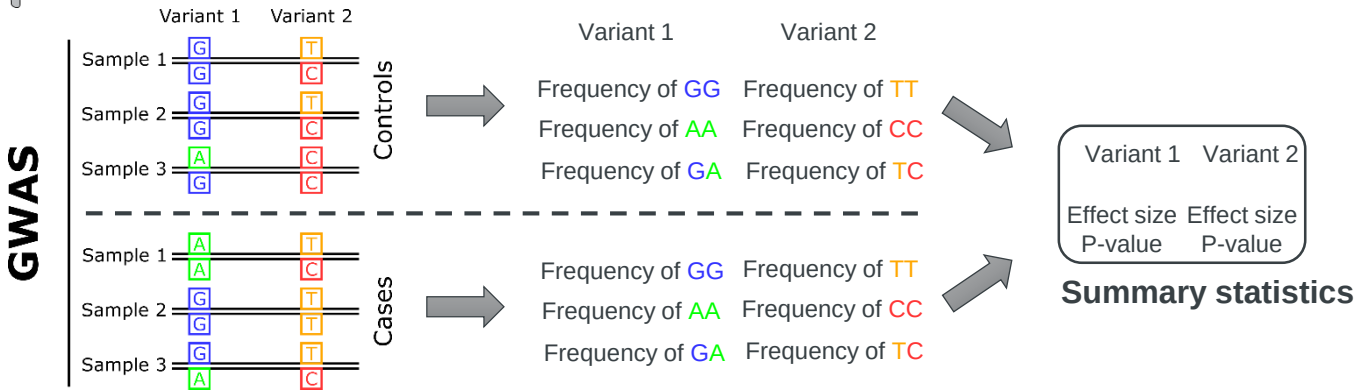
approx. 10 million common SNPs (frequency above 0.5%)



Complex traits and GWAS



Multiple common variants with weak effects scattered throughout the genome (polygenic)



Input data for GWAS analysis

Input:

N attributes (3-category), 1 binary class variable

M samples

	SNP_1	SNP_2	…	SNP_n	Class
$Patient_1$	1	1	…	0	1
$Patient_2$	0	2	…	1	0
$Patient_3$	1	0	…	2	1
…					
$Patient_M$	2	1	…	1	0

Output:

SNPs associated with diseases

1- order: $\{SNP_1\}, \{SNP_2\}, \{SNP_3\}\ldots$

2-order : $\{SNP_1, SNP_2\}\ldots$

3-order : $\{SNP_1, SNP_2, SNP_3\}\ldots$

Encoding:

- AA -> 0, Aa -> 1, aa -> 2
- Case-> 1, Control -> 0

Schizophrenia GWAS

ARTICLE

doi:10.1038/nature13595

Biological insights from 108 schizophrenia-associated genetic loci

Schizophrenia Working Group of the Psychiatric Genomics Consortium*

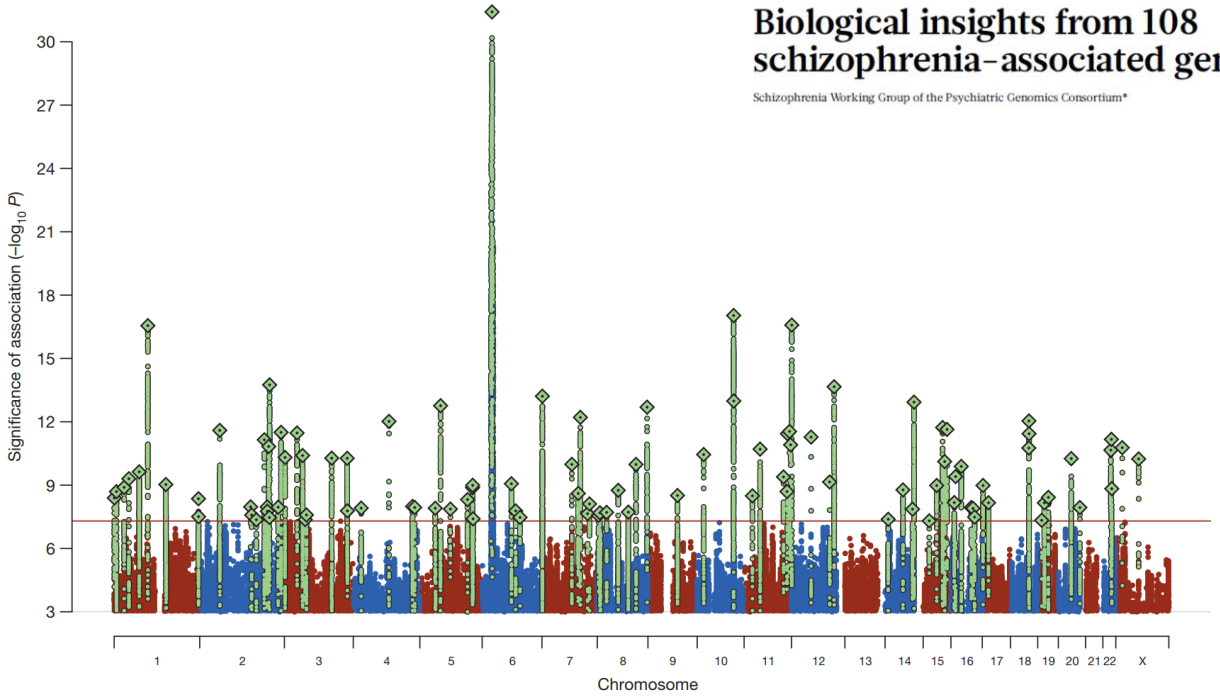


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal position and the *y* axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

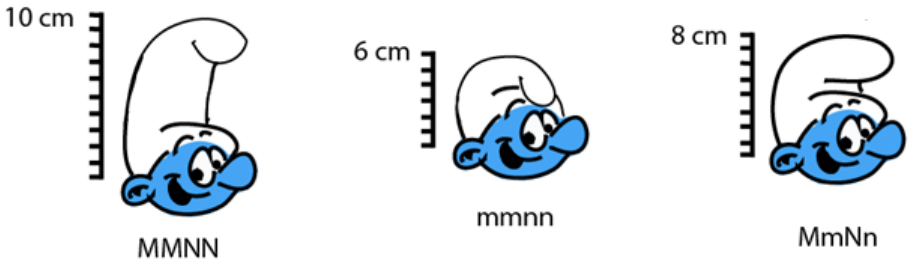
Supplementary Table 2: 128 genome-wide significant associations for schizophrenia

Rank	Index SNP	A12	Frq_{case}	$Frq_{control}$	Chr	Position	Combined		Discovery		Replication	
							OR (95% CI)	P	OR	P	OR	P
54	rs4648845	TC	0.533	0.527	1	2,372,401-2,402,501	1.072 (1.049-1.097)	8.7e-10	1.071	4.03e-9	1.088	8.85e-2
57	chr1_8424984_D	I2D	0.319	0.301	1	8,411,184-8,638,984	1.071 (1.048-1.095)	1.17e-9	1.071	2.03e-9	1.057	2.96e-1
65	rs1498232	TC	0.311	0.296	1	30,412,551-30,437,271	1.069 (1.046-1.093)	2.86e-9	1.072	1.28e-9	0.999	9.88e-1
50	rs11210892	AG	0.659	0.677	1	44,029,384-44,128,084	0.934 (0.914-0.954)	3.39e-10	0.933	4.97e-10	0.949	3.08e-1
22	rs12129573	AC	0.377	0.358	1	73,766,426-73,991,366	1.078 (1.056-1.101)	2.03e-12	1.072	2.35e-10	1.217	6.25e-5
107	rs76869799	CG	0.959	0.964	1	97,792,625-97,834,525	0.846 (0.798-0.897)	2.64e-8	0.850	1.44e-7	0.779	5.34e-2
2	rs1702294	TC	0.175	0.191	1	98,374,984-98,559,084	0.887 (0.865-0.911)	3.36e-19	0.891	2.79e-17	0.831	1.35e-3
52	rs140505938	TC	0.151	0.164	1	149,998,890-150,242,490	0.914 (0.888-0.940)	4.49e-10	0.913	9.34e-10	0.928	2.53e-1
120	rs6670165	TC	0.196	0.184	1	177,247,821-177,300,821	1.075 (1.047-1.103)	4.45e-8	1.074	1.16e-7	1.090	1.46e-1
121	rs7523273	AG	0.695	0.685	1	207,912,183-208,024,083	1.063 (1.040-1.087)	4.47e-8	1.062	1.61e-7	1.092	8.85e-2
101	rs10803138	AG	0.232	0.238	1	243,503,719-243,612,019	0.933 (0.911-0.956)	2.03e-8	0.932	1.79e-8	0.968	5.56e-1
68	rs77149735	AG	0.0225	0.0191	1	243,555,105-243,555,105	1.317 (1.202-1.444)	3.73e-9	1.329	4.4e-9	1.173	3.66e-1
119	rs14403	TC	0.207	0.222	1	243,639,893-243,664,923	0.934 (0.911-0.957)	4.42e-8	0.935	1.31e-7	0.920	1.53e-1
78	chr1_243881945_I	I2D	0.638	0.619	1	243,690,945-244,002,945	1.068 (1.045-1.092)	6.53e-9	1.066	3.11e-8	1.107	6.17e-2
30	rs11682175	TC	0.52	0.542	2	57,943,593-58,065,893	0.933 (0.914-0.952)	1.47e-11	0.928	2.54e-12	1.018	7.08e-1
117	rs75575209	AT	0.904	0.913	2	58,025,192-58,502,192	0.902 (0.869-0.936)	3.95e-8	0.896	1.01e-8	1.056	5.6e-1
80	rs3768644	AG	0.0967	0.101	2	72,357,335-72,368,185	0.904 (0.874-0.935)	7.39e-9	0.910	1.3e-7	0.765	2.15e-3
62	chr2_146436222_I	I2D	0.176	0.163	2	146,416,922-146,441,832	1.086 (1.057-1.116)	1.81e-9	1.084	1.07e-8	1.128	5.72e-2
95	chr2_149429178_D	I2D	0.955	0.961	2	149,390,778-149,520,178	0.857 (0.813-0.904)	1.59e-8	0.856	2.62e-8	0.880	2.97e-1
124	rs2909457	AG	0.568	0.593	2	162,798,555-162,910,255	0.944 (0.925-0.964)	4.62e-8	0.943	4.38e-8	0.971	5.36e-1
18	rs11693094	TC	0.44	0.458	2	185,601,420-185,785,420	0.929 (0.910-0.948)	1.53e-12	0.929	7.13e-12	0.918	7.64e-2
83	rs59979824	AC	0.322	0.337	2	193,848,340-194,028,340	0.937 (0.916-0.958)	8.41e-9	0.936	1.08e-8	0.959	4.32e-1
33	rs6434928	AG	0.635	0.643	2	198,148,577-198,835,577	0.929 (0.909-0.949)	2.06e-11	0.927	1.48e-11	0.969	5.36e-1
82	rs6704641	AG	0.819	0.805	2	200,161,422-200,309,252	1.081 (1.053-1.110)	8.33e-9	1.079	3.4e-8	1.123	8.1e-2
10	chr2_200825237_I	I2D	0.741	0.754	2	200,715,237-200,848,037	0.909 (0.887-0.932)	5.65e-14	0.906	1.78e-14	1.011	8.7e-1
87	rs11685299	AC	0.313	0.326	2	225,334,096-225,467,796	0.939 (0.919-0.959)	1.12e-8	0.937	1.11e-8	0.974	6.12e-1
23	rs6704768	AG	0.54	0.552	2	233,559,301-233,753,501	0.930 (0.911-0.949)	2.32e-12	0.929	3.15e-12	0.953	3.19e-1

GWAS aims to find genetic effects => PRS

Input (known):

y_i – phenotypes
w_{ij} – genotypes



Output (unknown):

$\hat{u}_j$ - genetic effects of allele substitution

M: +? cm

N: +? cm

m: +? cm

n: +? cm

$$\hat{y}_i = \sum_j w_{ij}\hat{u}_j + e_i$$

Polygenic Risk Scoring, or
Naïve Bayes classifier

Introduction to genetics of complex traits

- **Genetics of Complex Traits**
  - Mendelian vs polygenic inheritance
  - Heritability, twin studies
  - Genome-wide association studies
- **Big Data samples**
  - GWAS consortia
  - TOP, MoBa (Norwegian samples)
  - UK Biobank, ABCD
- **Challenges and advances in statistical analysis**
  - Correlation structure in the genotype matrix
  - Specific tools (conjFDR, MiXeR, MOSTest)

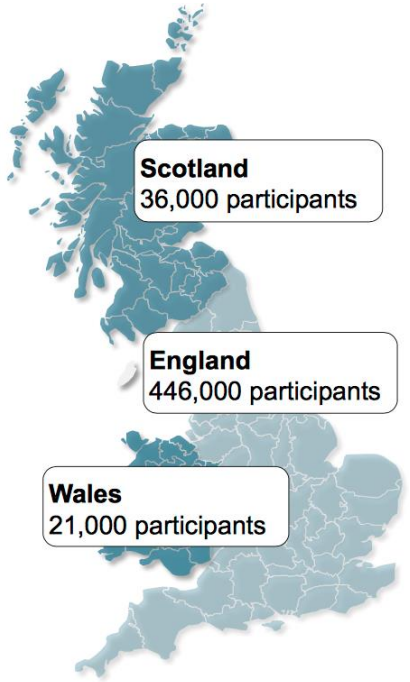
Big data NORMENT

- Clinical samples
  - Hospitals n=25k (Psychosis, dementia, ASD etc)
- Population genetics (prospective, registries):
  - MoBa n=240k (100k kids, parents)
  - HUSK n=36k (cohort)
  - HUNT n=70k (collaboration)
  - TromsøStudy n=35k(cohort, 2019)
- Total n=400k

- Other samples: 550k

NORMENT In-house Data

- N = 500k
- Aged 40-69 recruited between 2006 and 2010
- Genetic data (500k genotype, 50k whole-exome sequence)
- Registry-based information
- Extensive self-reported baseline data on lifestyl environment, personal & family medical history, includir detailed mental health questionnaire
- Brain imaging data ~45k

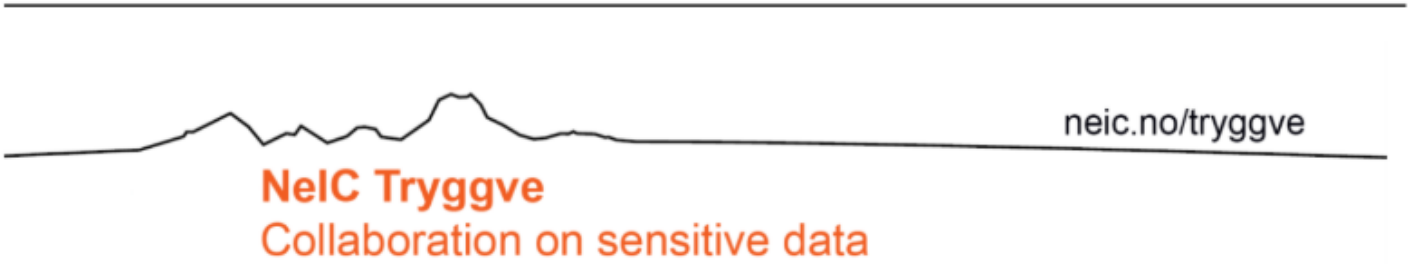


NORMENT In-house Data

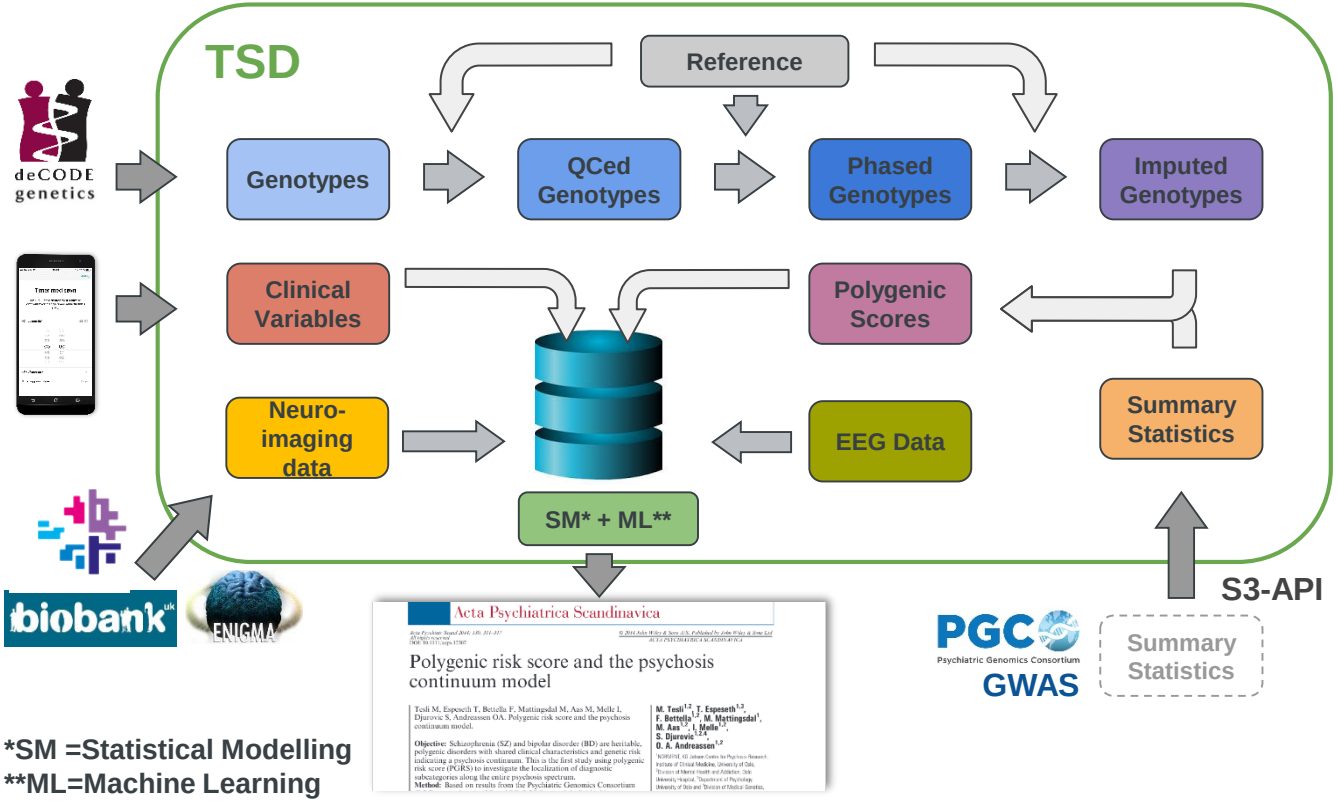
- N ~ 12k
- Aged 9-10 at recruitment
- Will be followed up through early adulthood
- Genetic data
- Extensive self-reported baseline data on lifestyle, environment, personal & family medical history, including detailed mental health questionnaire
- Brain imaging data - all
- Wait for phenotypes to develop - longitudinal, prospective

Data Sensitivity

- Most sensitive individual data
  - Individual identification
  - Privacy around mental health status
- All data stored and analyzed on secure computing cluster (TSD)
  - Individual level data anonymized and cannot be exported
- *TRYGGVE* - data sharing across Nordic countries



The NORMENT Big Data pipeline



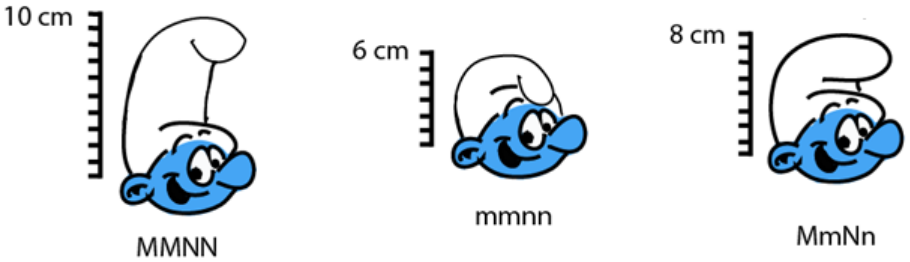
Introduction to genetics of complex traits

- **Genetics of Complex Traits**
  - Mendelian vs polygenic inheritance
  - Heritability, twin studies
  - Genome-wide association studies
- **Big Data samples**
  - GWAS consortia
  - TOP, MoBa (Norwegian samples)
  - UK Biobank, ABCD
- **Challenges and advances in statistical analysis**
  - Correlation structure in the genotype matrix
  - Specific tools (conjFDR, MiXeR, MOSTest)

What's so challenging?

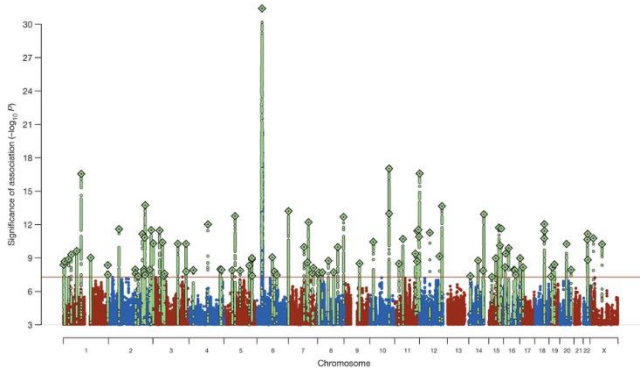
$$y = Gx + \epsilon$$

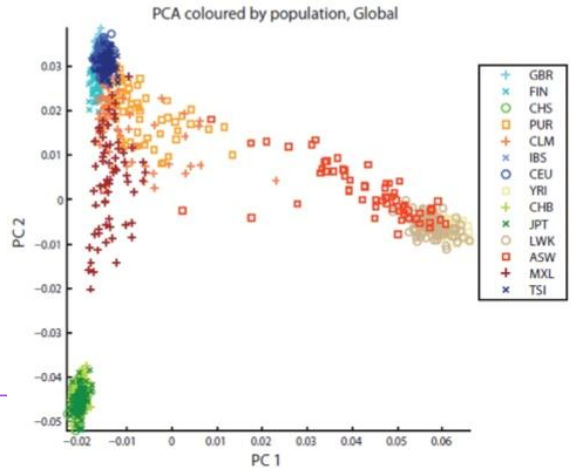
- y - phenotype vector
- G – genotype matrix
- x – unknown genetic effects
- ϵ – environmental noise



Challenges in GWAS analysis

- “BigData” scale:
  - 10^7 genetic variants (SNPs)
  - 10^6 individuals
- Correlation structure:
  - Relatedness among individuals
  - “Linkage disequilibrium” in SNPs
- Heterogeneity in GWAS cohorts
  - Different populations
  - Potential overlap between cohorts
- Statistical power
  - Small individual effects
  - ca. 1 000 000 independent tests, multiple hypothesis correction
    => typical p-value threshold $5x10^{-8}$





Models to estimate heritability from genotype data

- Mixed effects model

$$y = \boldsymbol{g} + \boldsymbol{X}\beta + \epsilon$$

Phenotype

Genetic component

$g \sim N\left(0, \boldsymbol{A}\sigma_g^2\right)$

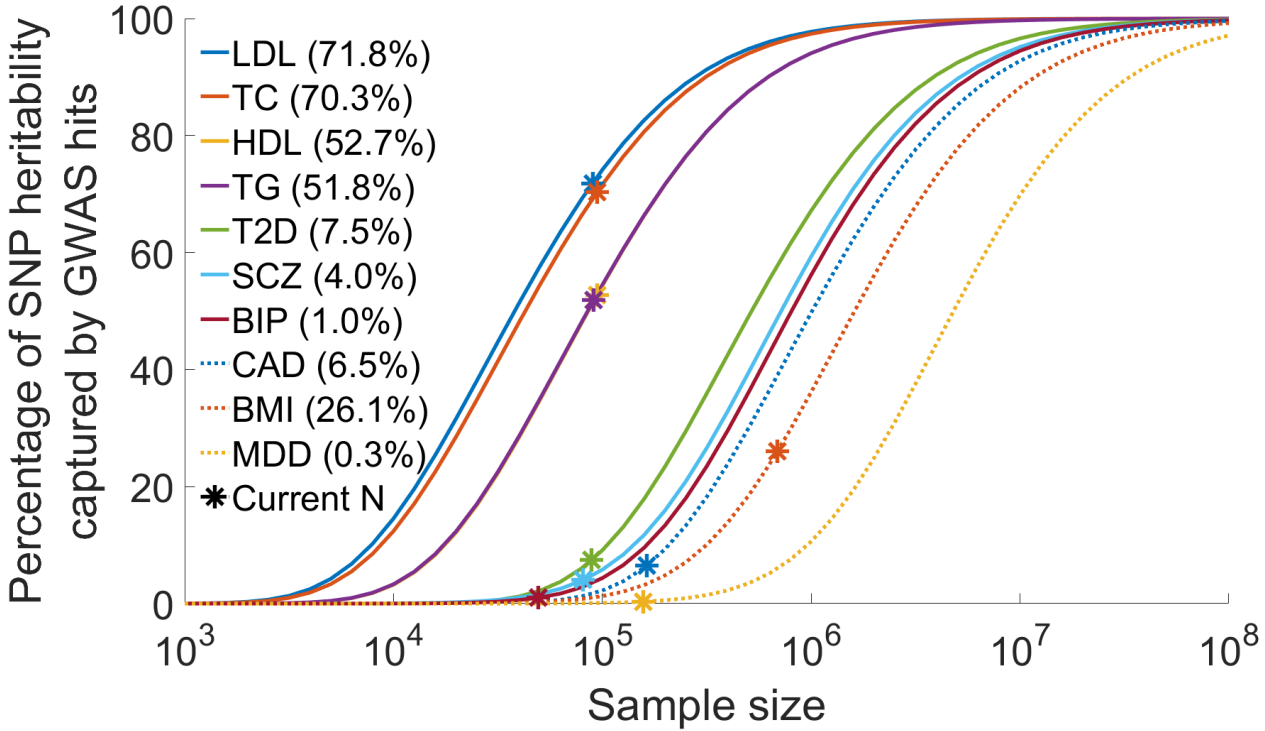
random effects,
$\boldsymbol{A}$ is derived from W

Clinical covariates
(Age, sex, …)
Fixed effects

Unmeasured contribution from the environment

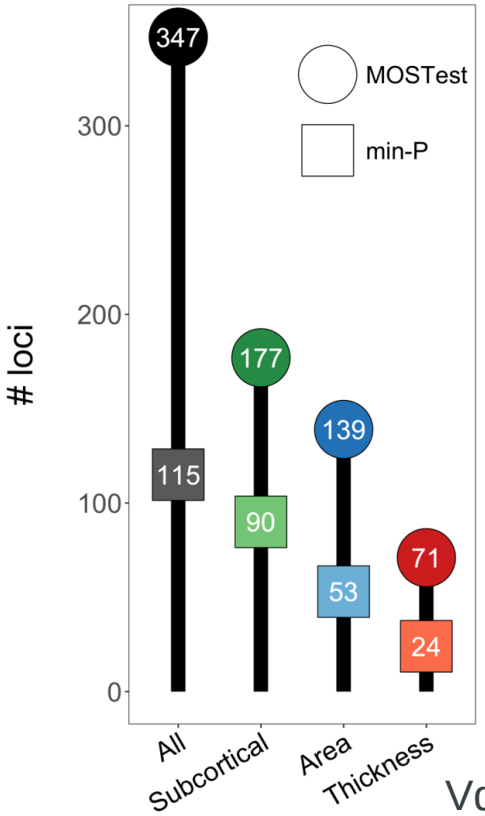
GCTA - Restricted maximum likelihood (RELM) [Patterson and Thompson, 1971] [Harville, 1974], Average Information (AI) inference [Arthur R. Gilmour, 1995]
BoltLMM – more efficient inference with Monte Carlo REML approximation, conjugate gradient iteration to solve mixed model equations, Variational Bayes iteration

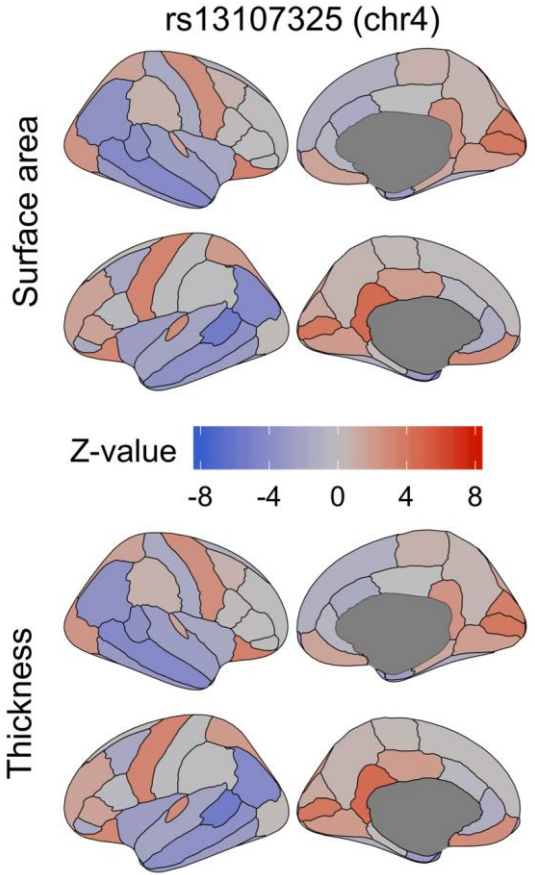
Polygenicity affects power to discover loci in GWAS



Holland et al. in review

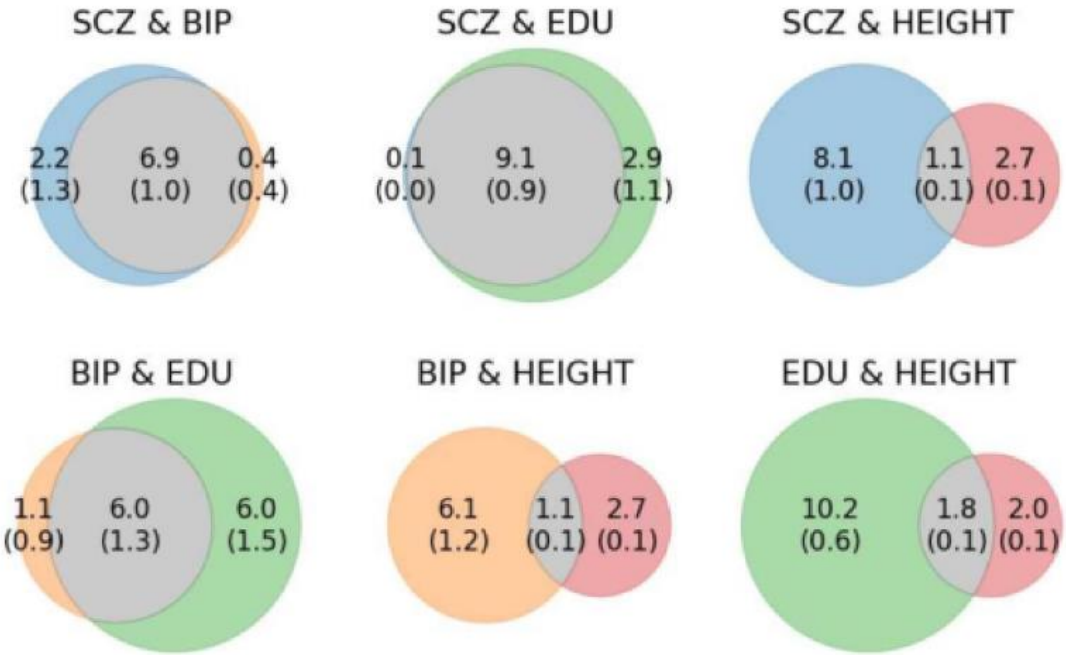
MOSTest – increase discovery (Multivariate Omnibus Statistical Test)





Vd Meer, Frei et al. in subm

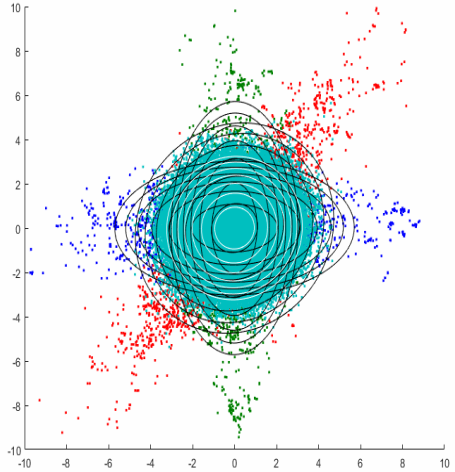
Polygenic overlap – mathematical modeling



nature COMMUNICATIONS

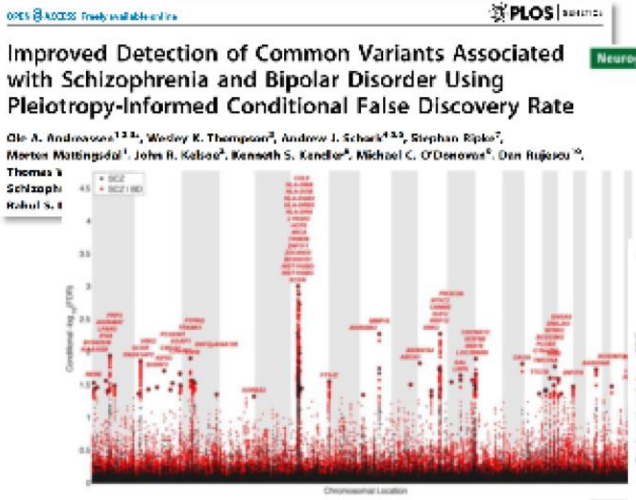
Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation

Oleksandr Frei, Dominic Holland, Olav B. Smeland, Alexey A. Shadrin, Chun Chieh Fan, Steffen Maeland, Kevin S. O'Connell, Yunpeng Wang, Srdjan Djurovic, Wesley K. Thompson, Ole A. Andreassen & Anders M. Dale

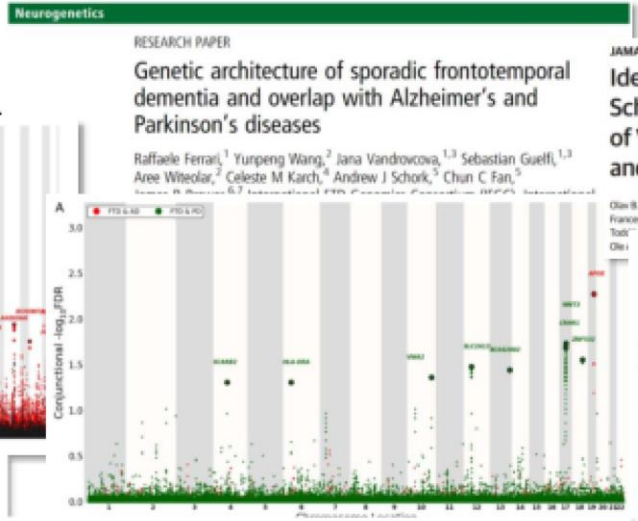


https://github.com/precimed/mixer

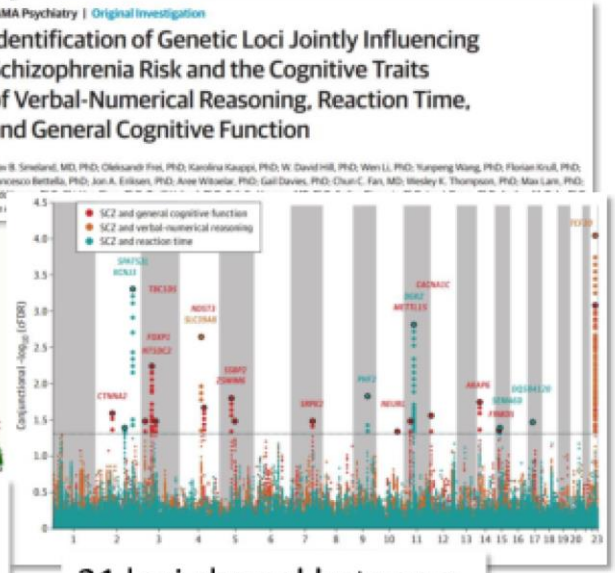
Extensive genetic overlap
Conditional/conjunctional False Discovery Rate



51 novel SCZ loci
11 novel BIP loci
14 loci shared between SCZ and BIP



8 loci shared between FTD and PD
1 shared locus between FTD and AD
13 novel FTD loci



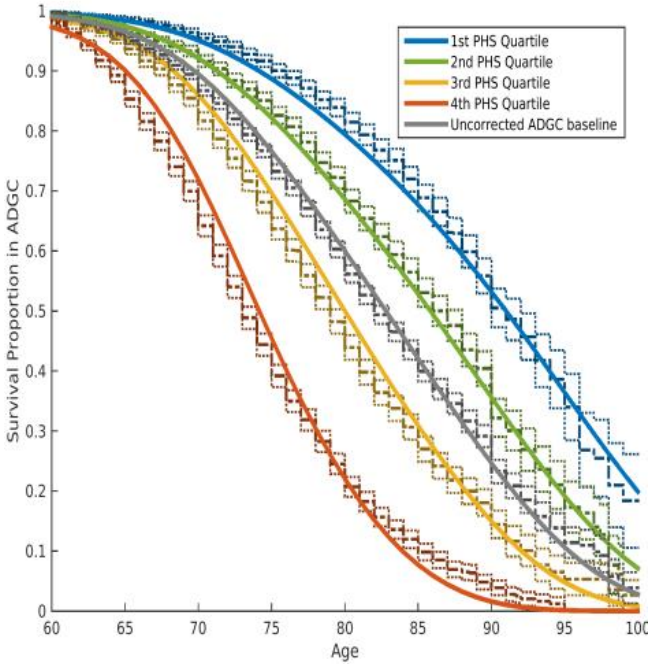
21 loci shared between SCZ and cognitive traits

Discovery of shared genomic loci using the conditional false discovery rate approach

Review

First Online: 13 September 2019

Polygenic prediction disease onset



Age	Population Baseline*	PHS 1 percentile (95% CI)	PHS 20th percentile (95% CI)	PHS 80th percentile (95% CI)	PHS 99th percentile (95% CI)	*APOE* ε4+ (95% CI)	*APOE* ε4- (95% CI)
60	0.08	0.03 (0.02,0.04)	0.05 (0.03,0.07)	0.14 (0.08, 0.21)	0.48 (0.27, 0.69)	0.19 (0.18, 0.20)	0.06 (0.06, 0.7)
65	0.17	0.05 (0.03,0.08)	0.1 (0.06, 0.15)	0.29 (0.17, 0.42)	0.97 (0.56,1.39)	0.38 (0.36, 0.40)	0.13 (0.12, 0.13)
70	0.35	0.11 (0.06,0.16)	0.21 (0.12,0.3)	0.6 (0.35, 0.85)	1.98 (1.13, 2.84)	0.78 (0.74, 0.82)	0.26 (0.25, 0.27)
75	0.71	0.22 (0.13,0.32)	0.43 (0.25,0.62)	1.22 (0.71, 1.73)	4.03 (2.3, 5.77)	1.58 (1.51, 1.66)	0.53 (0.52, 0.55)
80	1.44	0.45 (0.26,0.64)	0.88 (0.51,1.25)	2.47 (1.44, 3.51)	8.2 (4.67, 11.73)	3.22 (3.06, 3.38)	1.08 (1.05, 1.11)
85	2.92	0.92 (0.54,1.31)	1.8 (1.04, 2.55)	5.03 (2.92, 7.15)	16.68 (9.5, 23.86)	6.55 (6.23, 6.87)	2.2 (2.13, 2.27)
90	5.95	1.88 (1.09,2.66)	3.65 (2.12, 5.18)	10.24 (5.94, 14.53)	33.93 (19.33, 48.54)	13.33 (12.68, 13.98)	4.48 (4.34, 4.61)
95	12.1	3.82 (2.22,5.42)	7.43 (4.32, 10.54)	20.82 (12.09, 29.56)	69.02 (39.32, 98.72)	27.11 (25.79, 28.43)	9.1 (8.83, 9.38)

NORMENT
Norwegian Centre for
Mental Disorders Research

Thank you

NORMENT
Oslo University Hospital HF
Division of Mental Health and Addiction
Psychosis Research Unit/TOP
Ullevål Hospital, building 49
P.O. Box 4956 Nydalen
N-0424 Oslo
Norway

Kevin O'Connell - kevin.oconnell@medisin.uio.no
Oleksandr Frei - oleksandr.frei@medisin.uio.no
Ole Andreassen - ole.Andreassen@medisin.uio.no