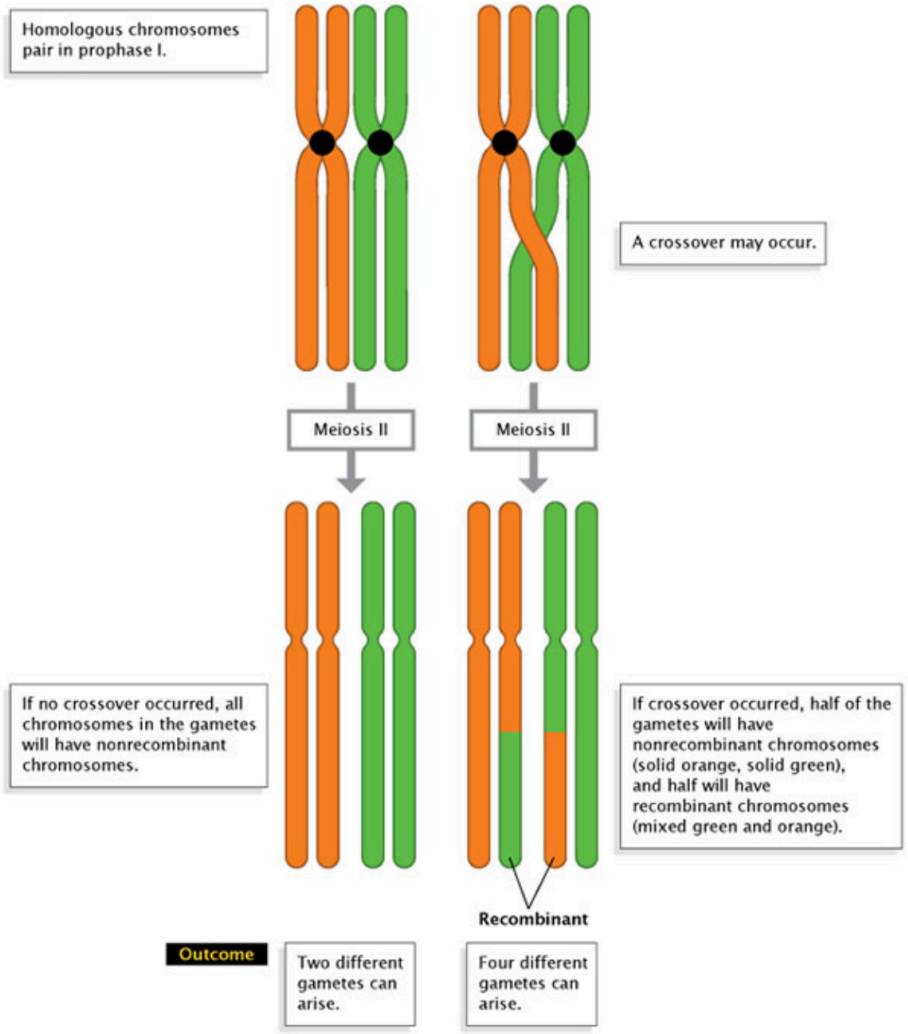
Module 10: Inheritance of Polygenic Diseases

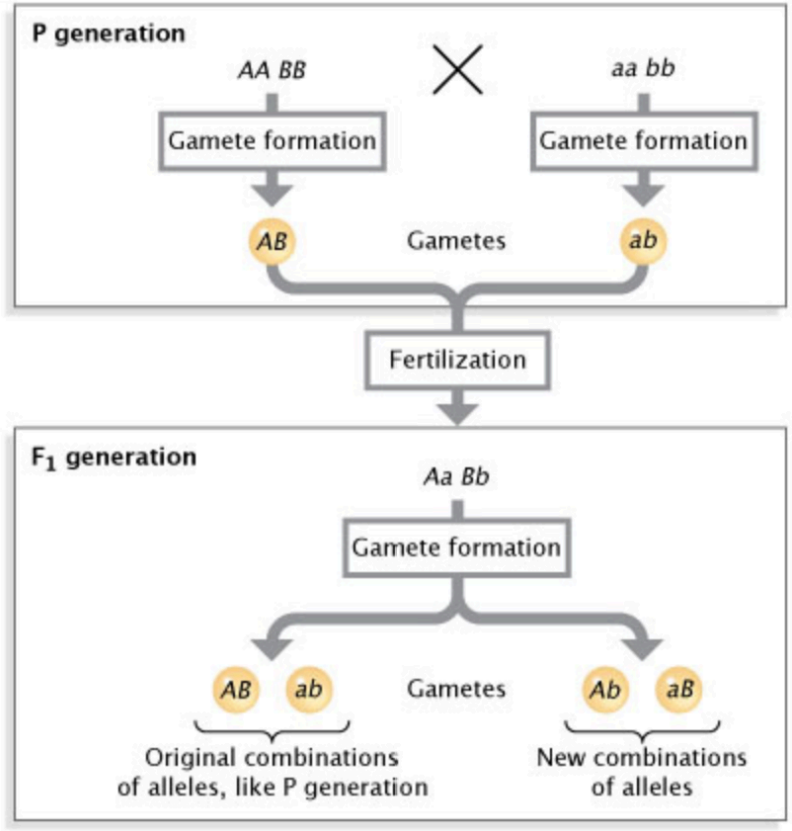
Key Concepts

- Linkage Analysis
- HWE and Allele Frequency
- Genome Wide Association Studies
- Applications

Recombination[1]

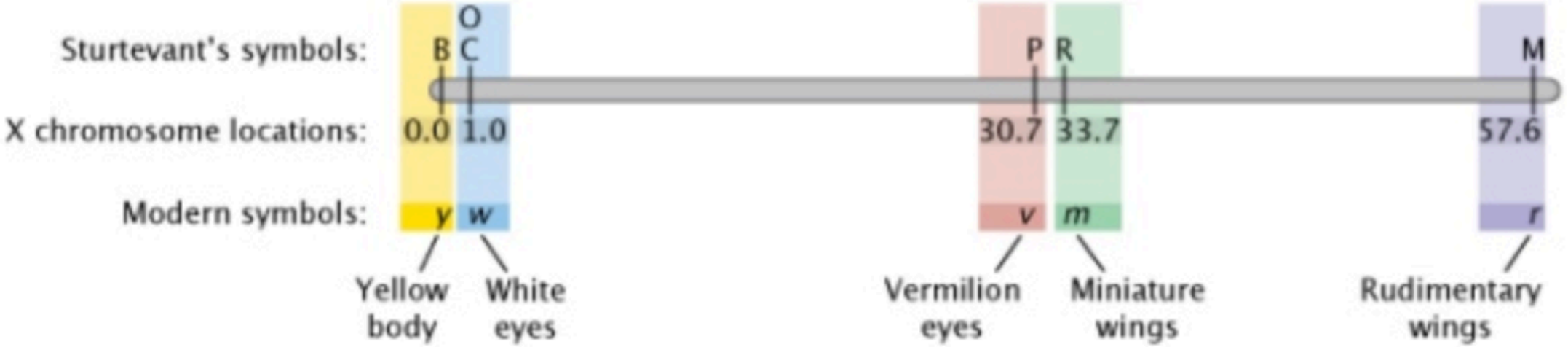


- Alleles recombine during crossover events in meiosis



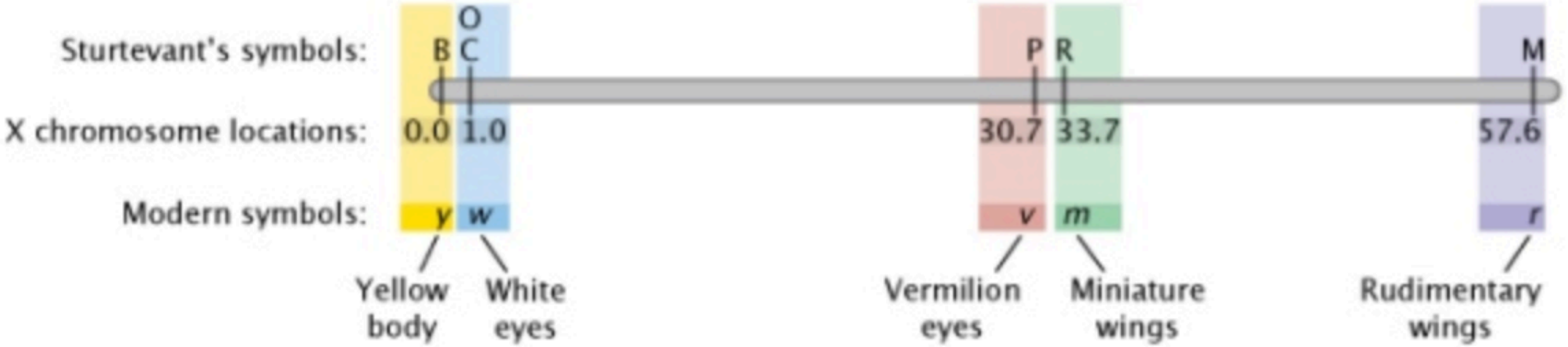
Genetic Linkage

- The frequency of recombination is related to the distance between the genes on a chromosome.[1]
- Violation of Mendel's law of independent assortment[1]



Genetic Linkage

- The **closer** two genes were to one another on a chromosome, the **greater** their chance of being inherited together.[1]
- Genes located **farther** away from one another on the **same** chromosome were **more likely to be separated** during recombination.[1]



Linkage Map

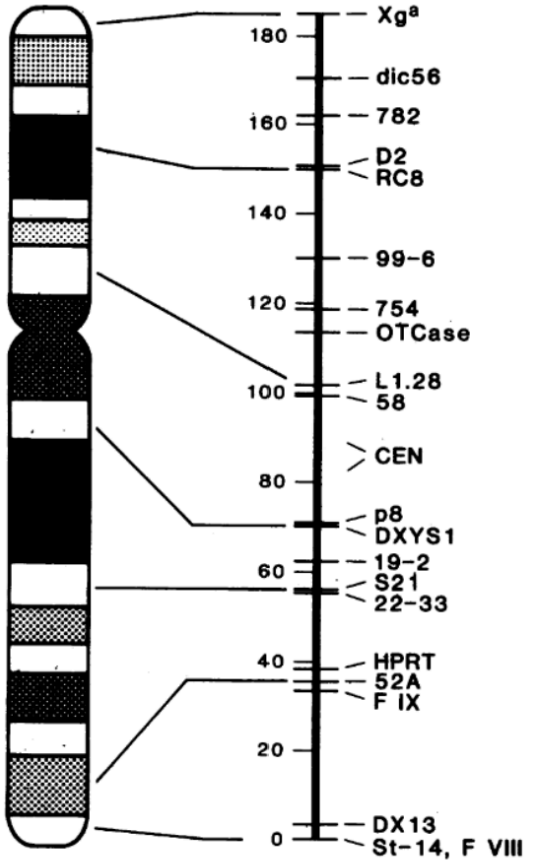
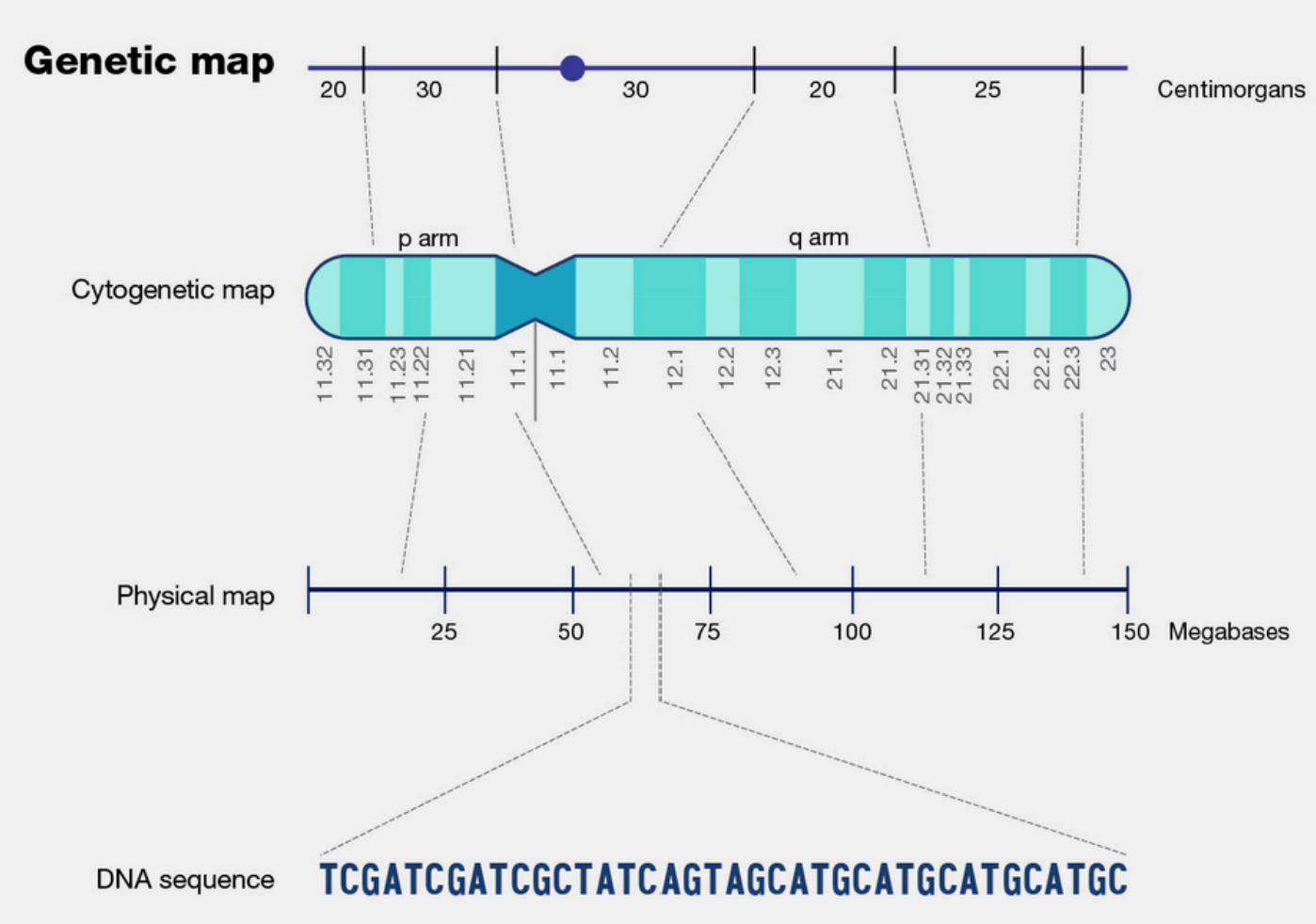


Fig. 1. Linkage map of the human X chromosome. Genetic distances are represented in recombination units, and physical locations of selected markers are indicated on the ideogram.

- **Linkage map**: (also called genetic map) the order and the linear distances between the linked genes[1]
- **Map unit** (mu): arbitrary unit of distance representing a recombination frequency of 1%[1]
  - renamed as **centimorgan (cM)**
- Completely linked genes = 0 mu
- Unlinked genes > 50 mu
- Linked genes < 50 mu

Drayna & White (1985)

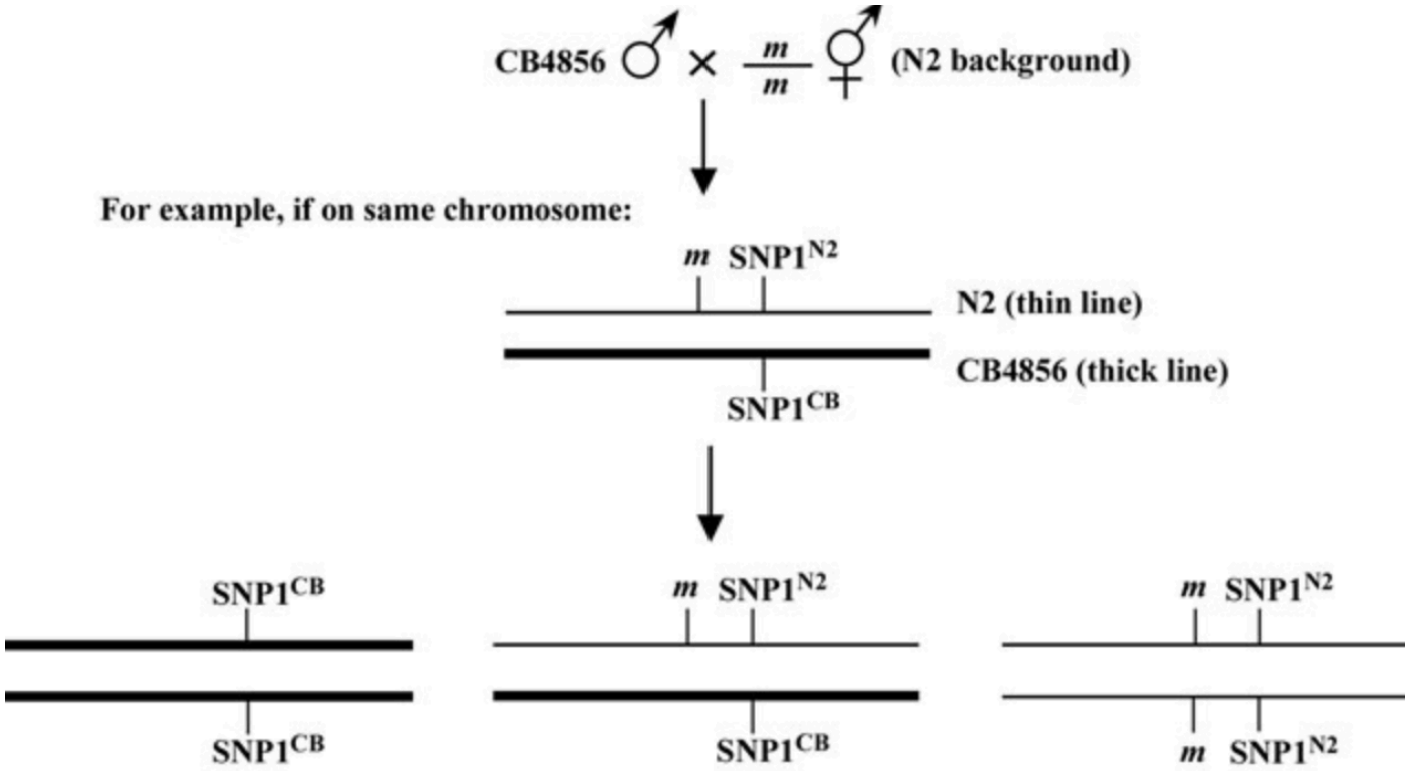
Linkage Map



- On average 1 cM corresponds to 1 million base pairs.[2]
- Linkage map is **NOT** the same as a cytogenetic map (karyotype)[3]

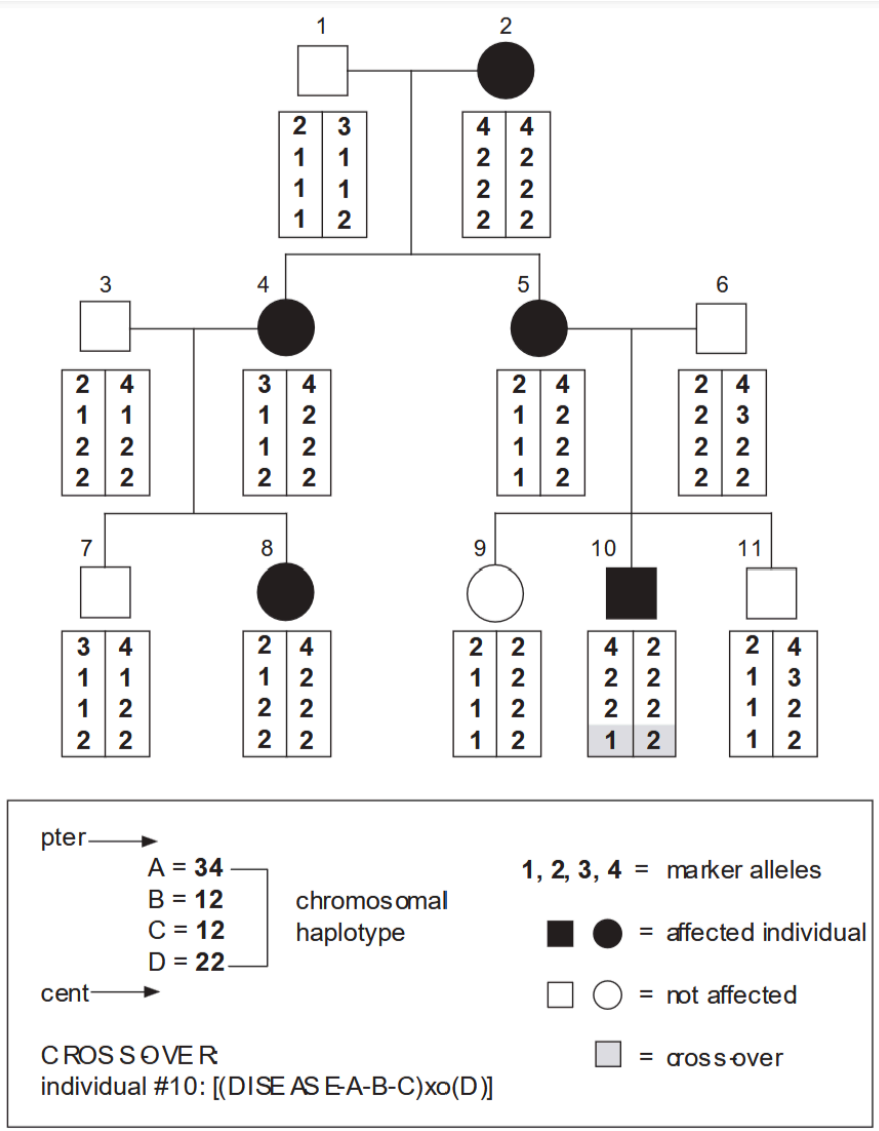
Two-point linkage

- Calculate the linkage between a putative disease locus and a single marker locus



Linkage Analysis: Next Steps

- To identify the smallest region of the genome that should contain the disease gene, the minimum-candidate region (MCR), we can perform:
    i. Haplotype analysis or
    ii. Multipoint linkage analysis



Haplotype Analysis[7]

- **Haplotype**: the arrangement of alleles on sister chromatids
- Construct haplotyes by examining the ordered markers genotyped in an individual and arranging them by chromatid, on the basis of parental types

Haplotype Analysis[7]

- After summarizing all cross-over individuals in the study, the most likely region for the disease gene is the region where no cross-overs are found.

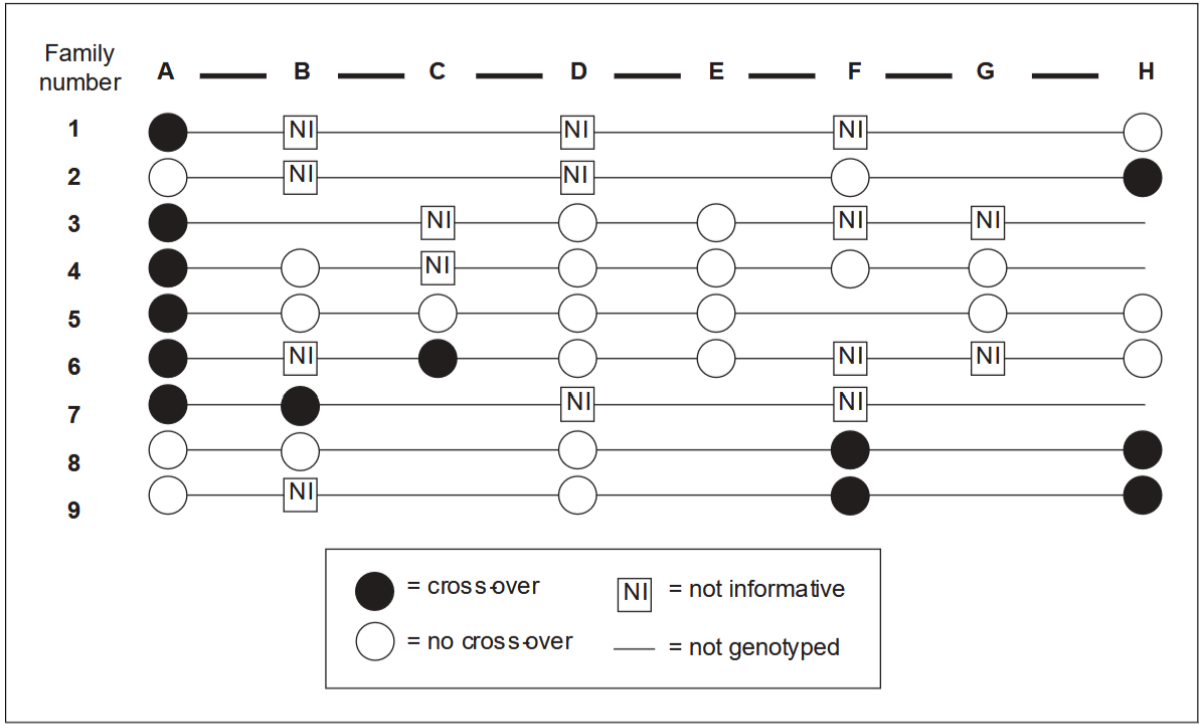


Figure 1.4.7 Method for summarizing all cross-over individuals in all family data used in a study. The most likely region for the disease gene is between markers C and F, the region where no cross-overs are found.

Multipoint analysis - Parametric[4]

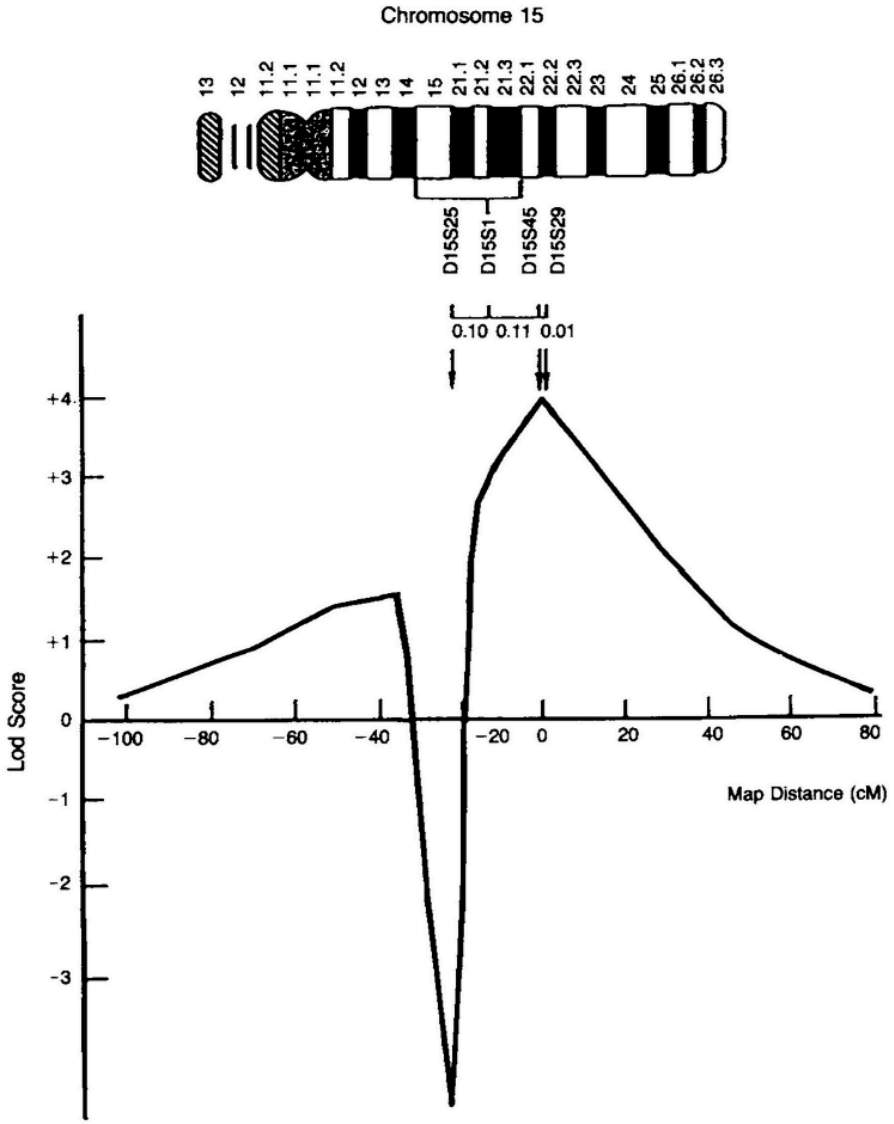
- Assume a model describing the probability of phenotype given genotype at the disease locus
- Calculate the likelihood ratio under the hypothesis that a disease gene is at x, versus the hypothesis that it is unlinked to x.

  LOD score: $Z(x) = \log_{10}\left(\frac{L(x)}{L(\infty)}\right)$ where

- $L(x)$ = Likelihood of the disease locus at position (x) on the marker map
  = the probability of **linkage** at a given recombination fraction (θ)
- $L(\infty)$ = Likelihood of the disease locus being infinitely far from a set of markers
  = the probability of **no linkage** (θ = 0.5)

How to use LOD score?[5]

- LOD score of 3 means the odds are 1,000:1 that the two genes are linked and therefore inherited together.
- **LOD score>3.3** → Strong evidence for linkage (genome-wide significance $p<0.05$)
    - Will still require replication with independent dataset
- 3.3>LOD score>1.86 → Suggestive linkage
- LOD score < 1.86 → Nominal linkage



Multipoint analysis

- **Multipoint LOD scores** generate a curve across a chromosome to identify the **likely position of a disease gene** (maximum of the curve) **given that the max. LOD score >=3.3**.[4]
- (Figure) Shows proof of linkage for Marfan Syndrome (autosomal dominant) in 5 families (LOD score = 3.92, θ = 0.0±0.11). The most probable location of the gene for the disease is D15S45 (LOD score = 3.32, θ = 0.0±0.12).[6]

LOD scoring algorithms[7]

Lander-Green algorithm

- Scales **linearly** with ***m*** **markers** , but **exponentially** with ***n*** **individuals** in the pedigree.
- More suitable for **smaller** pedigrees with **many markers**.
- Tool: MERLIN

Elston-Stewart algorithm

- Scales **linearly** with ***n*** **individuals**, but **exponentially** with ***m*** **markers**.
- More suitable for **large** pedigrees.
- Tool: FASTLINK (inbred pedigrees; multiple founder pairs), VITESSE (without loops, one founder pair)

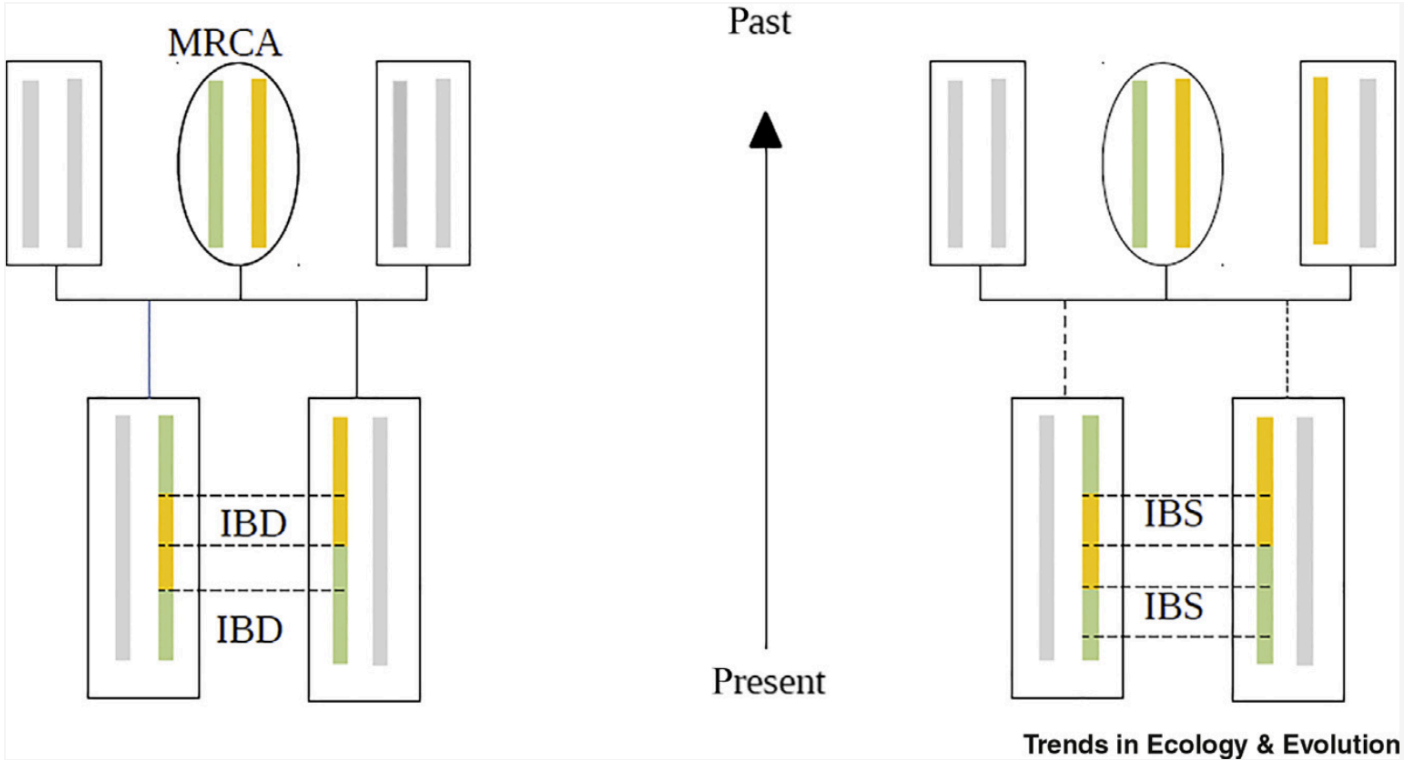
Multipoint analysis - Non-parametric[7]

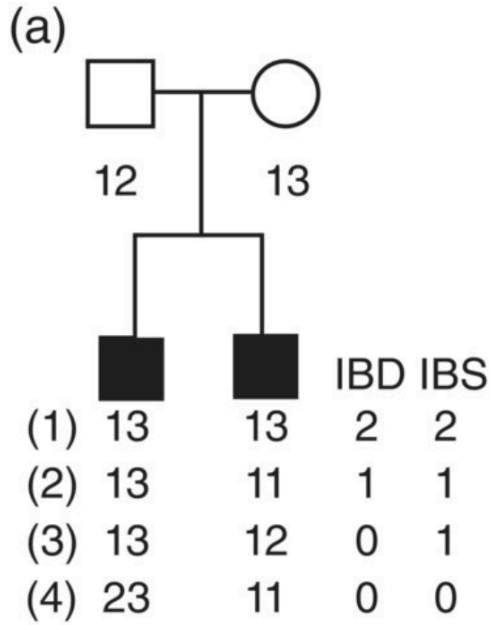
- Doesn't assume the mode of inheritance so it is suitable for complex traits
- **Assumptions:**
  - The trait has some genetic component.
  - The genes involved follow Mendelian laws of inheritance
- Significantly less than the power of parametric analysis
- **Use case**: Complex diseases with completely unknown inheritance model

Multipoint analysis - Non-parametric[7,8]

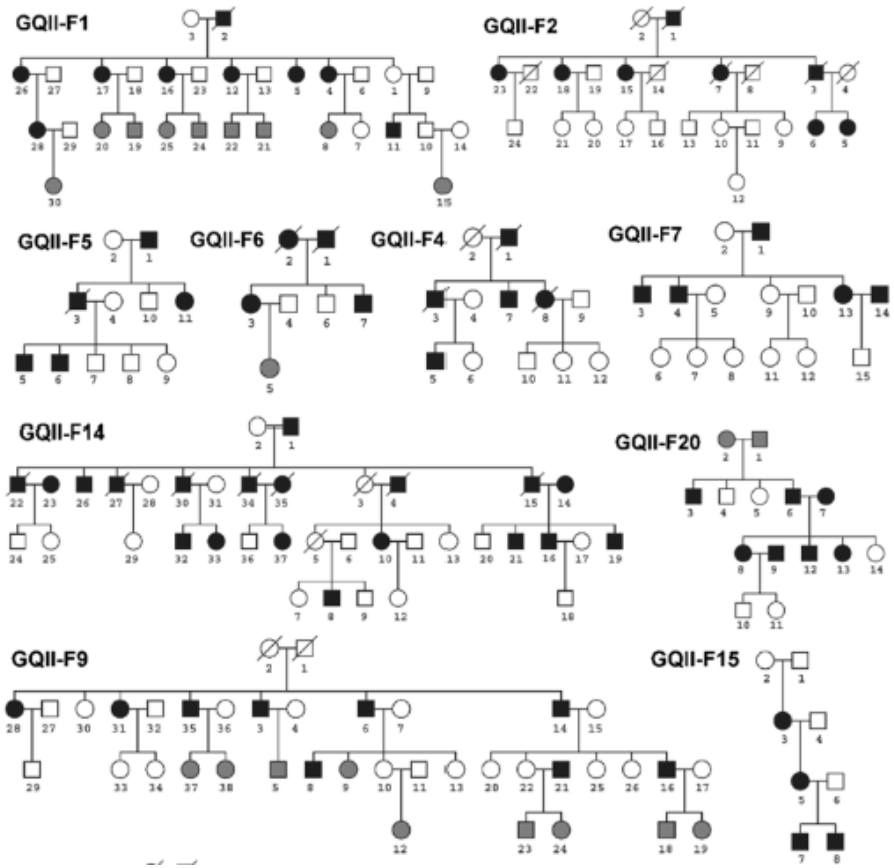
Quantifies the degree to which related individuals “share” alleles at the marker loci

1. Identity by state (IBS)
2. Identity by descent (IBD)





Genome-Wide Linkage Analysis



SCIENTIFIC REPORTS

OPEN

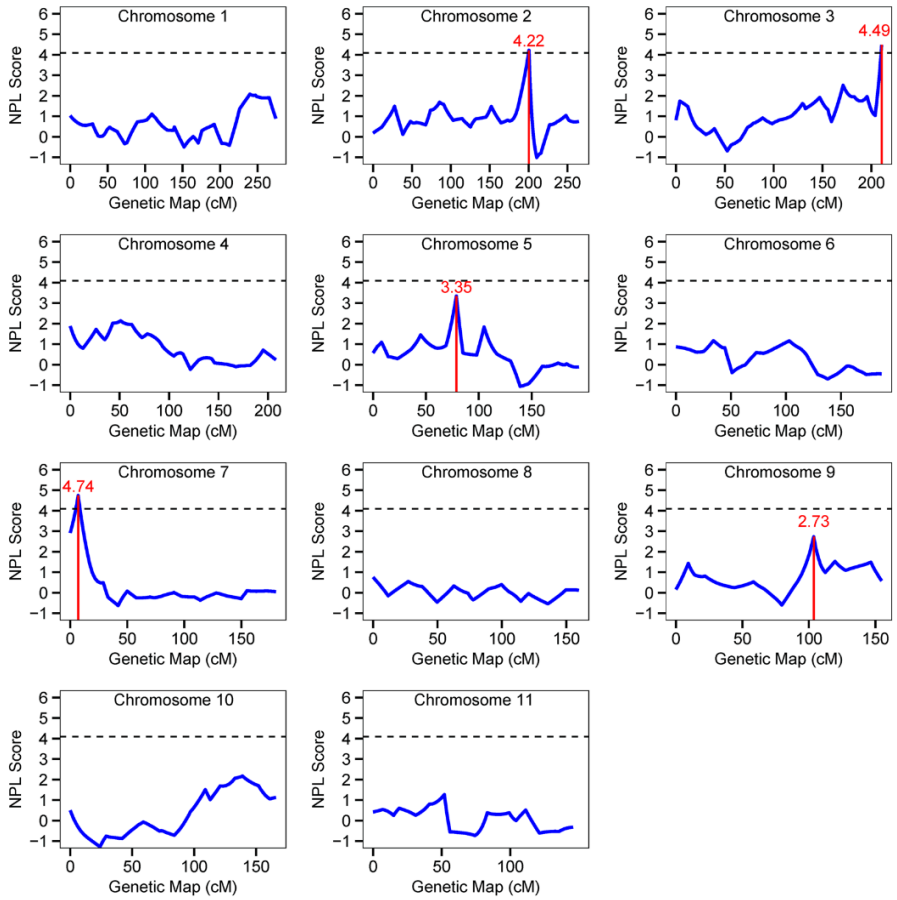
Genome-Wide Linkage Analysis of Large Multiple Multigenerational Families Identifies Novel Genetic Loci for Coronary Artery Disease

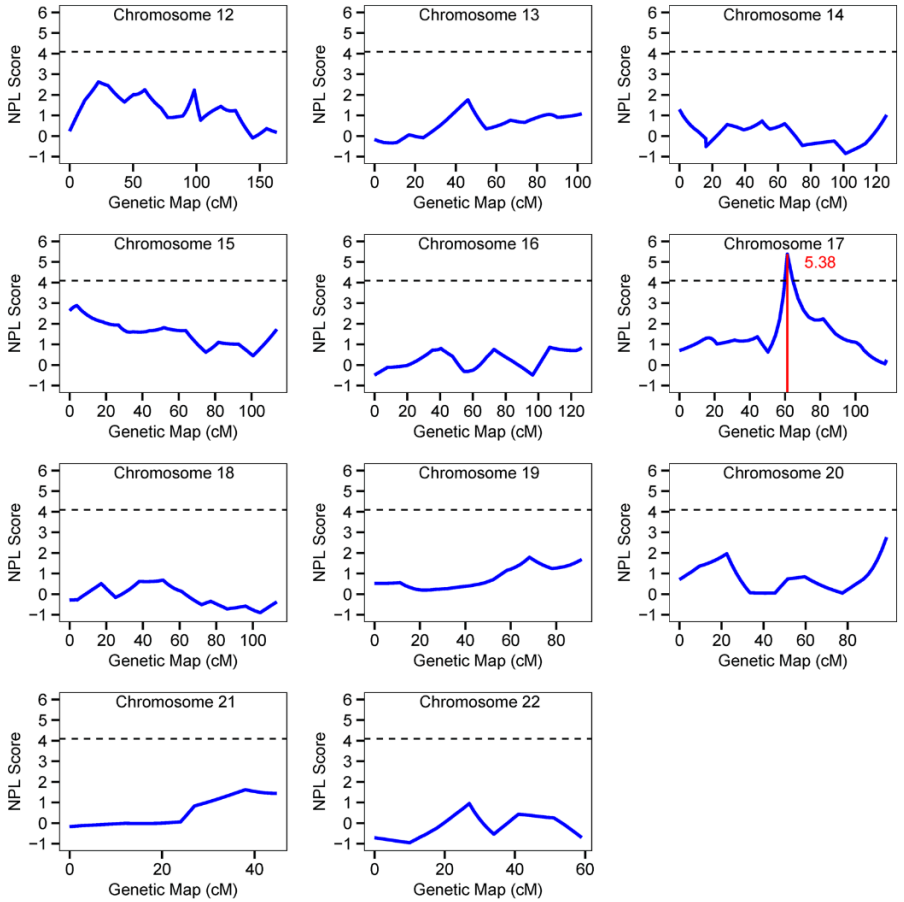
Yang Guo[1,2], Fan Wang[1,2], Lin Li[1,2], Hanxiang Gao[1,3], Stephen Arckacki[1,2], Isabel Z. Wang[1,4], John Barnard[5], Stephen Ellis[6], Carlos Hubbard[6], Eric J. Topol[7], Qiuyun Chen[1,2] & Qing K. Wang[1,2,6,8]

- 24 large and multigenerational families with 433 family members
- All family members were genotyped with markers spaced by every 10 cM

Genome-Wide Linkage Analysis

- The threshold of significant linkage is NPL ≥4.08





Genome-Wide Linkage Analysis

CAD Locus (Genetic Map[a])	Genomic Region[b]	RefSeq Genes[c]	Genes Related to Cardiovascular diseases[d]
17q21.2 (56.9–83.1 cM)	34.40–57.50 Mb	514	*CCL3, CCL4, CCL3L3, CCL4L1, CCL4L2, CCL3L1, TADA2A, PLXDC1, TCAP, PNMT, PGAP3, ERBB2, IKZF3, CSF3, MED24, THRA, NR1D1, CCR7, KRT12, KRT20, GAST, HAP1, JUP, FKBP10, CNP, KCNH4, HCRT, STAT5B, STAT5A, STAT3, ATP6V0A1, MLX, RAMP2, WNK4, BECN1, AOC3, BRCA1, SOST, PYY, G6PC3, HDAC5, GRN, ITGA2B, FZD2, ADAM11, GJC1, CCDC103, GFAP, HEXIM1, MAP3K14, CRHR1, MAPT, WNT3, GOSR2, MYL4, ITGB3, MRPL10, PNPO, MIR10A, UBE2Z, GIP, IGF2BP1, B4GALNT2, ZNF652, NGFR, ITGA3, PDK2, SGCA, COL1A1, XYLT2, CACNA1G, LUC7L3, NME1, MMD, AKAP1, MPO, MIR142*
7p22.2 (1.4 –13.0 cM)	0.88–7.25 Mb	87	*GPER1, MAFK, NUDT1, GNA12, SDK1, ACTB, AIMP2, EIF2AK1, RAC1*
2q33.3 191.9–202.4 cM	177.33–192.47 Mb	88	*NFE2L2, PDE11A, RBM45, TTN, CCDC141, ZNF385B, ITGA4, NEUROD1, PDE1A, FRZB, ITGAV, ZSWIM2, CALCRL, TFPI, COL3A1, COL5A2, SLC40A1, PMS1, MSTN, STAT1, STAT4*
3q29 (206.5–216.0 cM)	188.96–193.86 Mb	31	*TP63, CLDN16, UTS2B, HRASLS, OPA1*
5q13.2 (74.9–80.0 cM)	66.68–71.63 Mb	39	*PIK3R1, CCNB1, OCLN, SMN2, SMN1, NAIP, MCCC2, CARTPT*
9q22.33(103.6–105.3 cM)	101.72–104.22 Mb	20	*TGFBR1, NR4A3, INVS*

Applications of Linkage Analysis in Clinical Genomics

- Historically used in **family-based studies** to identify **Mendelian disease genes**.
- Previously required **sequencing candidate regions**, but now NGS allows rapid follow-up of suggestive linkage signals.

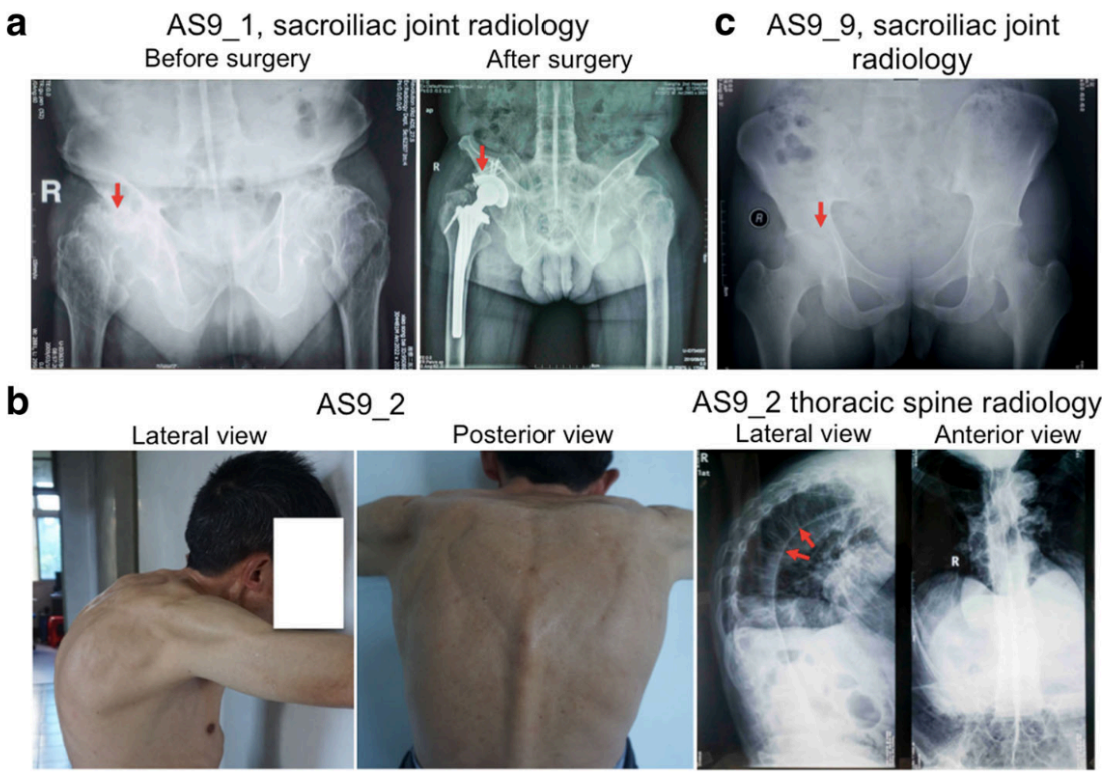
RESEARCH ARTICLE Open Access

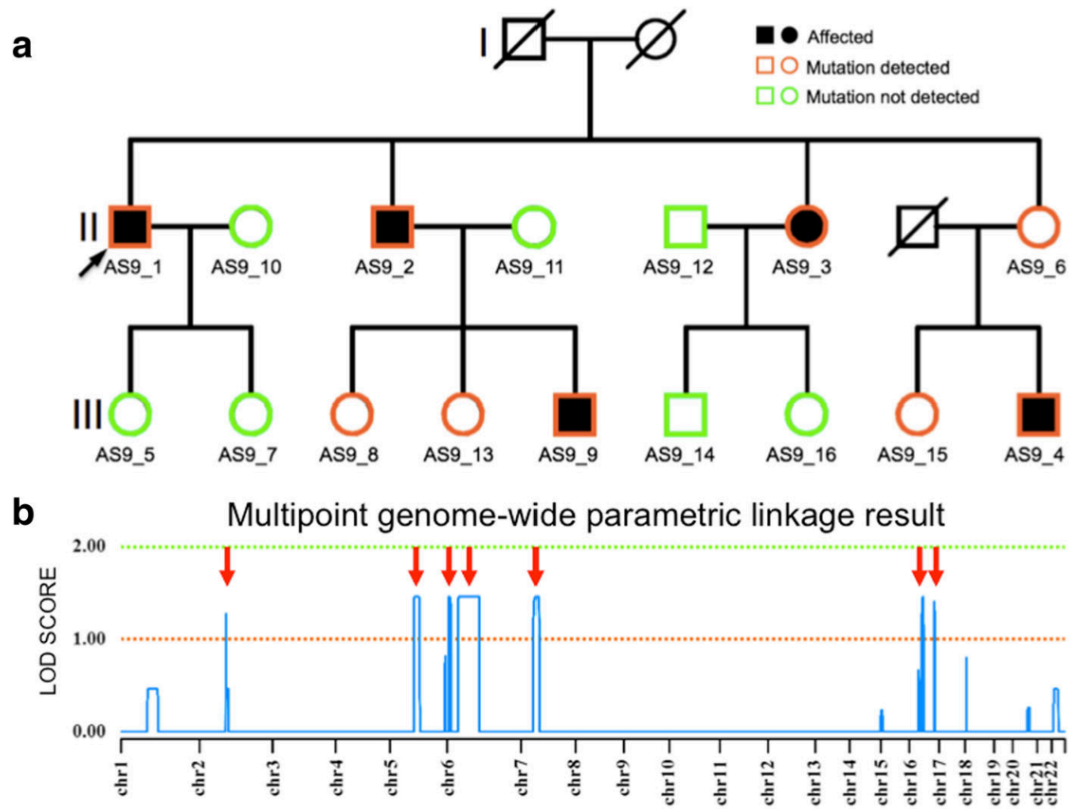
Identification of *ANKDD1B* variants in an ankylosing spondylitis pedigree and a sporadic patient

Zhiping Tan[1,2*], Hui Zeng[1,2], Zhaofa Xu[3], Qi Tian[3], Xiaoyang Gao[3], Chuanman Zhou[3], Yu Zheng[3], Jian Wang[1,2], Guanghui Ling[4], Bing Wang[5], Yifeng Yang[1,2] and Long Ma[3*]

Background: Ankylosing spondylitis (AS) is a debilitating autoimmune disease affecting tens of millions of people in the world. The genetics of AS is unclear. Analysis of rare AS pedigrees might facilitate our understanding of AS pathogenesis.

Methods: We used genome-wide linkage analysis and whole-exome sequencing in combination with variant co-segregation verification and haplotype analysis to study an AS pedigree and a sporadic AS patient.


a
AS9_1, sacroiliac joint radiology
Before surgery
After surgery
c
AS9_9, sacroiliac joint radiology
b
AS9_2
Lateral view
Posterior view
AS9_2 thoracic spine radiology
Lateral view
Anterior view


a
Affected
Mutation detected
Mutation not detected
I
II
III
AS9_1
AS9_10
AS9_2
AS9_11
AS9_12
AS9_3
AS9_6
AS9_5
AS9_7
AS9_8
AS9_13
AS9_9
AS9_14
AS9_16
AS9_15
AS9_4
b
Multipoint genome-wide parametric linkage result
LOD SCORE
2.00
1.00
0.00
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22

Population Genetics

The quantitative study of the distribution of genetic variation in populations and of how the frequencies of genes and genotypes are maintained or change over time both within and between populations.

- HWE and Allele Frequency
- Genome Wide Association Studies

Hardy-Weinberg Equilibrium

The genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors.

Assumes:

- Large population
- Random mating
- Allele frequencies remain constant over time because there are no mutations, natural selection, nonrandom mating, genetic drift, and gene flow.

Hardy-Weinberg Equilibrium

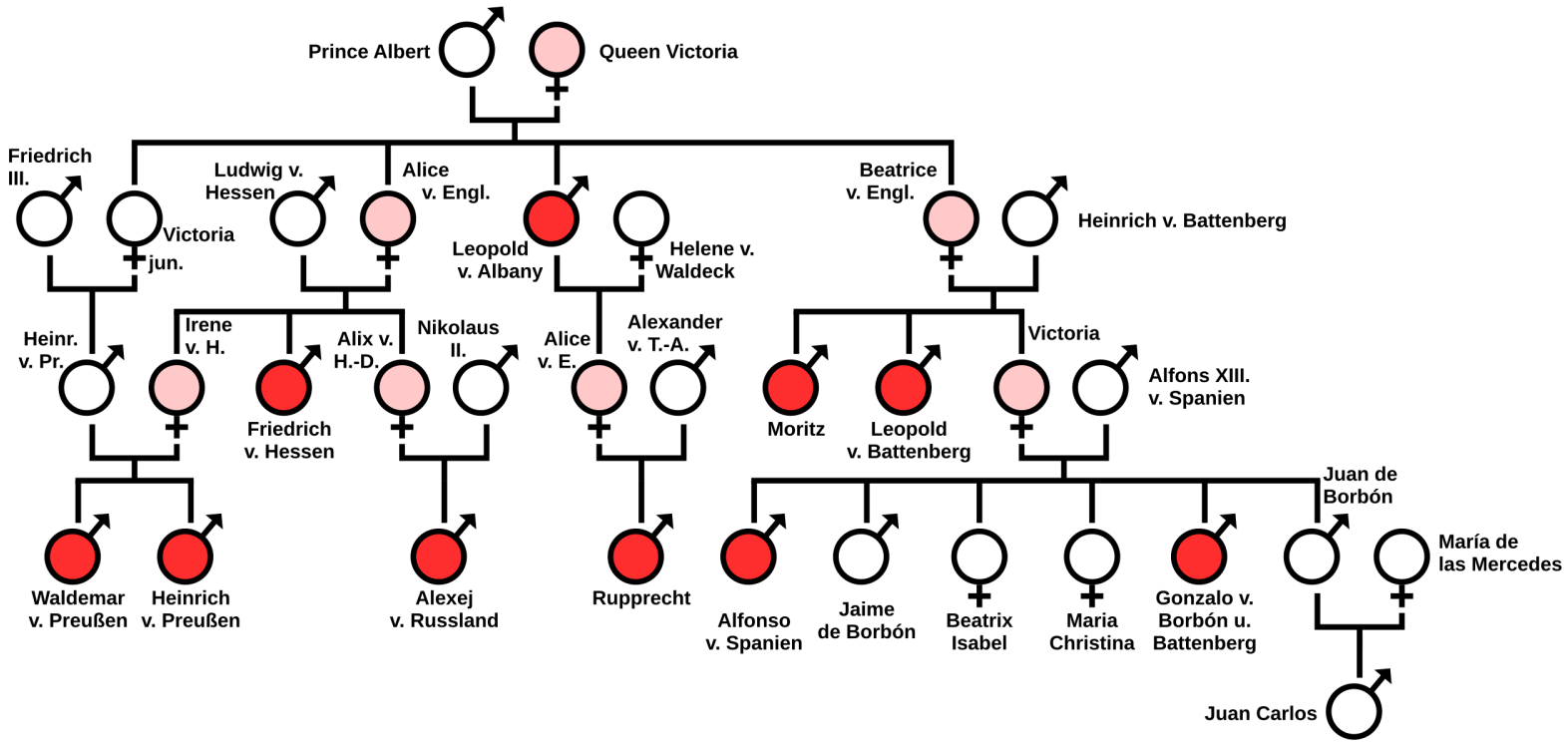
$$p^2 + 2pq + q^2 = 1$$

where:

- p and q are **allele frequencies**.
- p^2, $2pq$, and q^2 are **genotype frequencies**.

Factors That Disturb HWE: Nonrandom Mating

In human populations, nonrandom mating may occur due to: stratification, assortative mating, consanguinity.

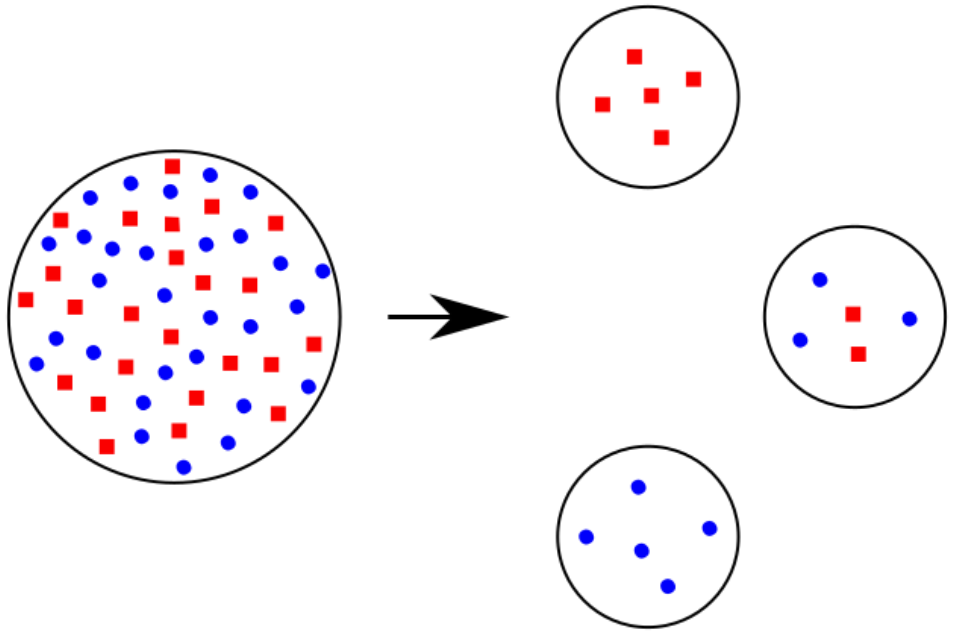


Factors That Disturb HWE: Mutation

- Selection pressure on mutations determines the allele frequencies in the population.
- Fitness is the outcome of the joint effects of survival and fertility.
- When a genetic disorder limits reproduction so severely that the fitness is zero (i.e., $s = 1$), it is thus referred to as a **genetic lethal**.

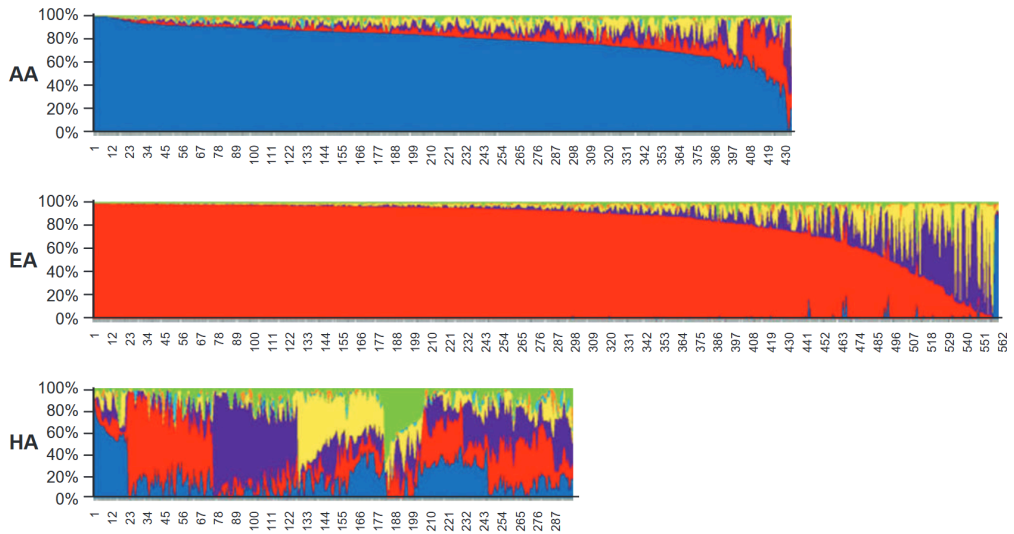
Factors That Disturb HWE: Genetic Drift

- Random effects of environment or other chance occurrences can change the frequency of the disease allele when the population is small.
- **Founder effect**: the loss of genetic variation that occurs when a new population is established by a very small number of individuals from a larger population



Factors That Disturb HWE: Migration and Gene Flow

- Migration can change allele frequency by the process of **gene flow** (the slow diffusion of genes across a barrier).
- **Genetic admixture**: the genes of migrant populations with their own allele frequencies are gradually merged into the gene pool of the population into which they have migrated.

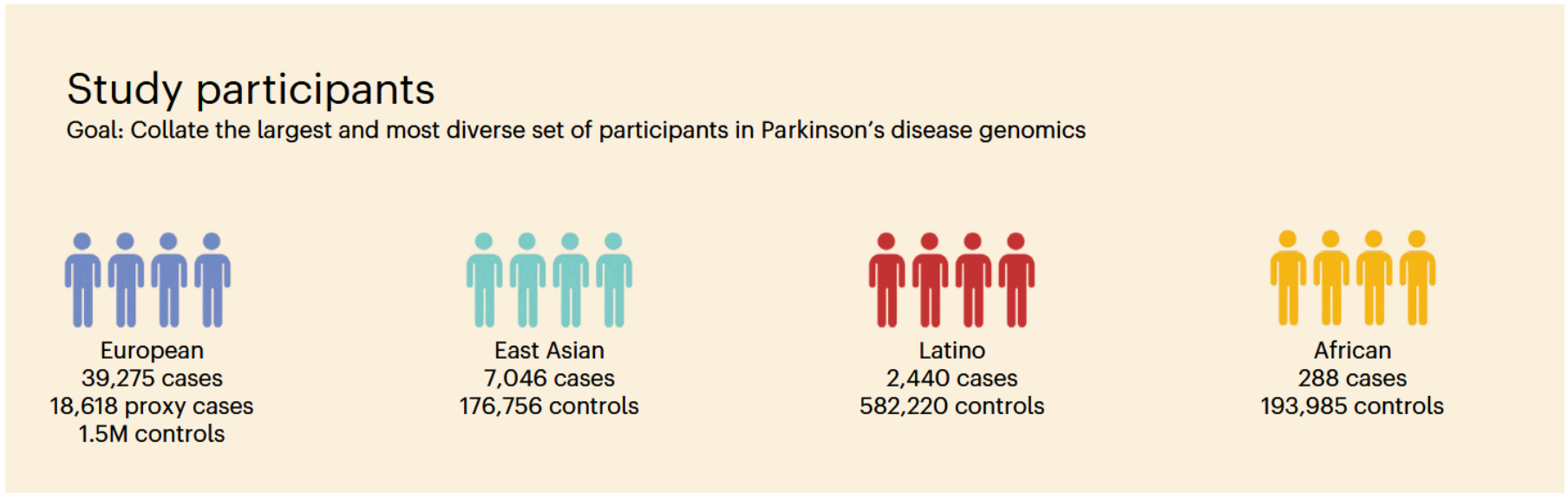


nature genetics

Article

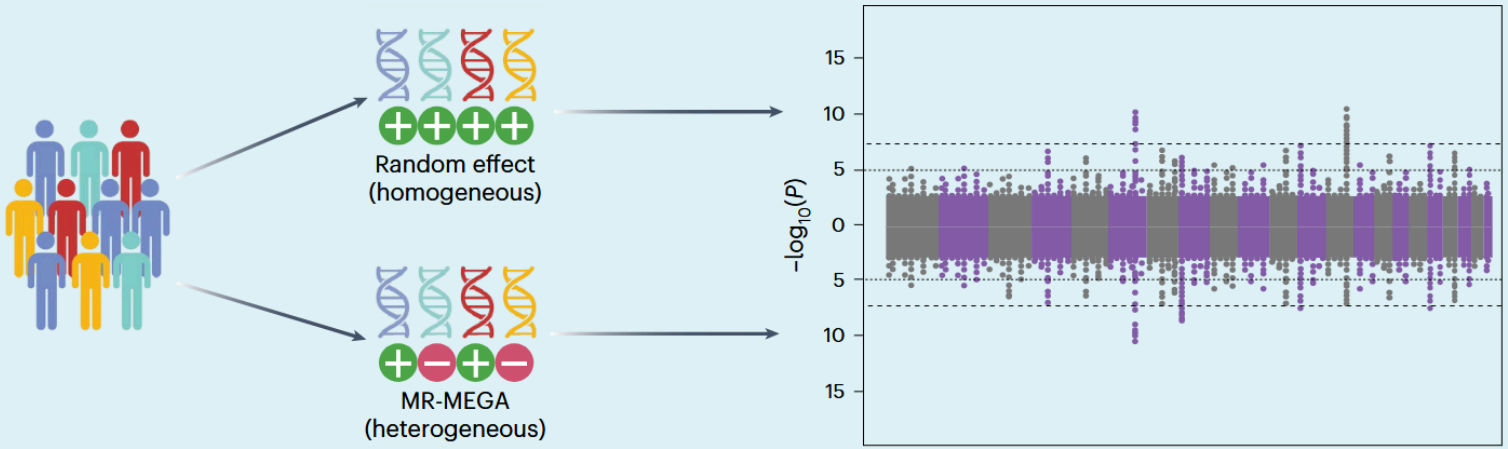
https://doi.org/10.1038/s41588-023-01584-8

Multi-ancestry genome-wide association meta-analysis of Parkinson's disease



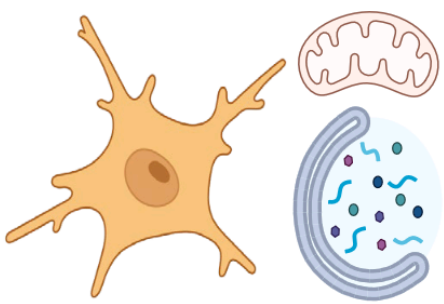
Multiancestry genome-wide meta-analysis

Goal: Identify common SNPs that are associated with Parkinson's disease risk that are applicable across different ancestries

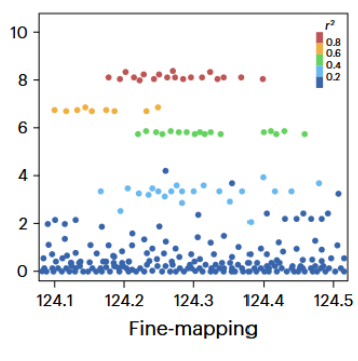


Downstream analyses

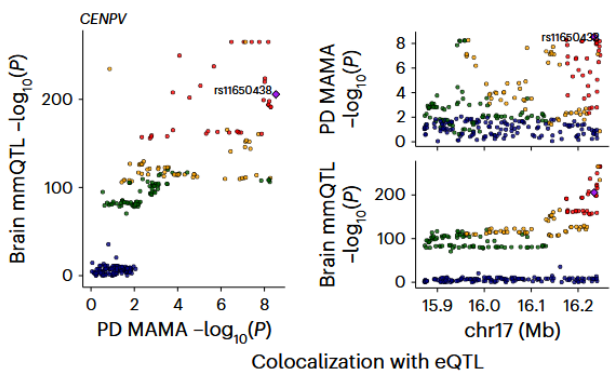
Goal: Interpret the meta-analysis results and identify potential targets and biological mechanisms



Gene ontology enrichment

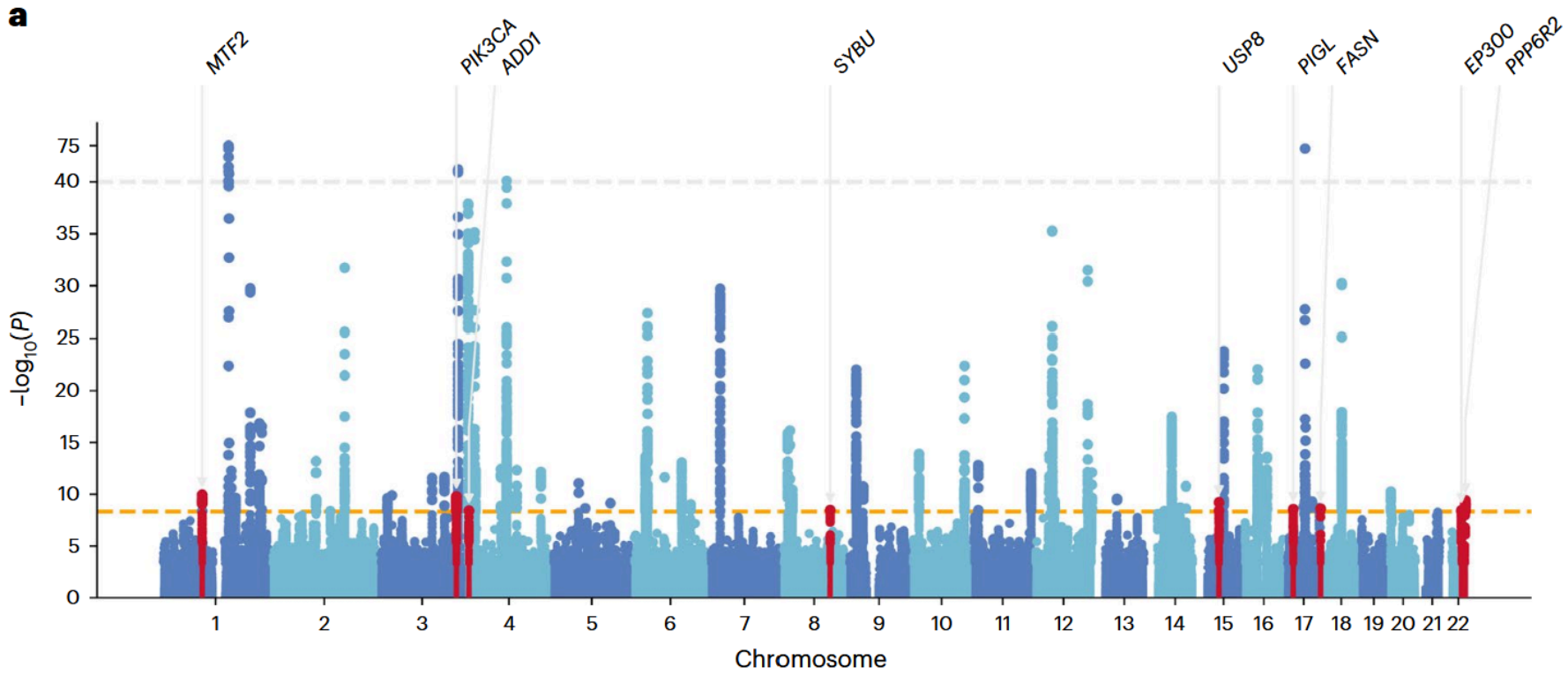


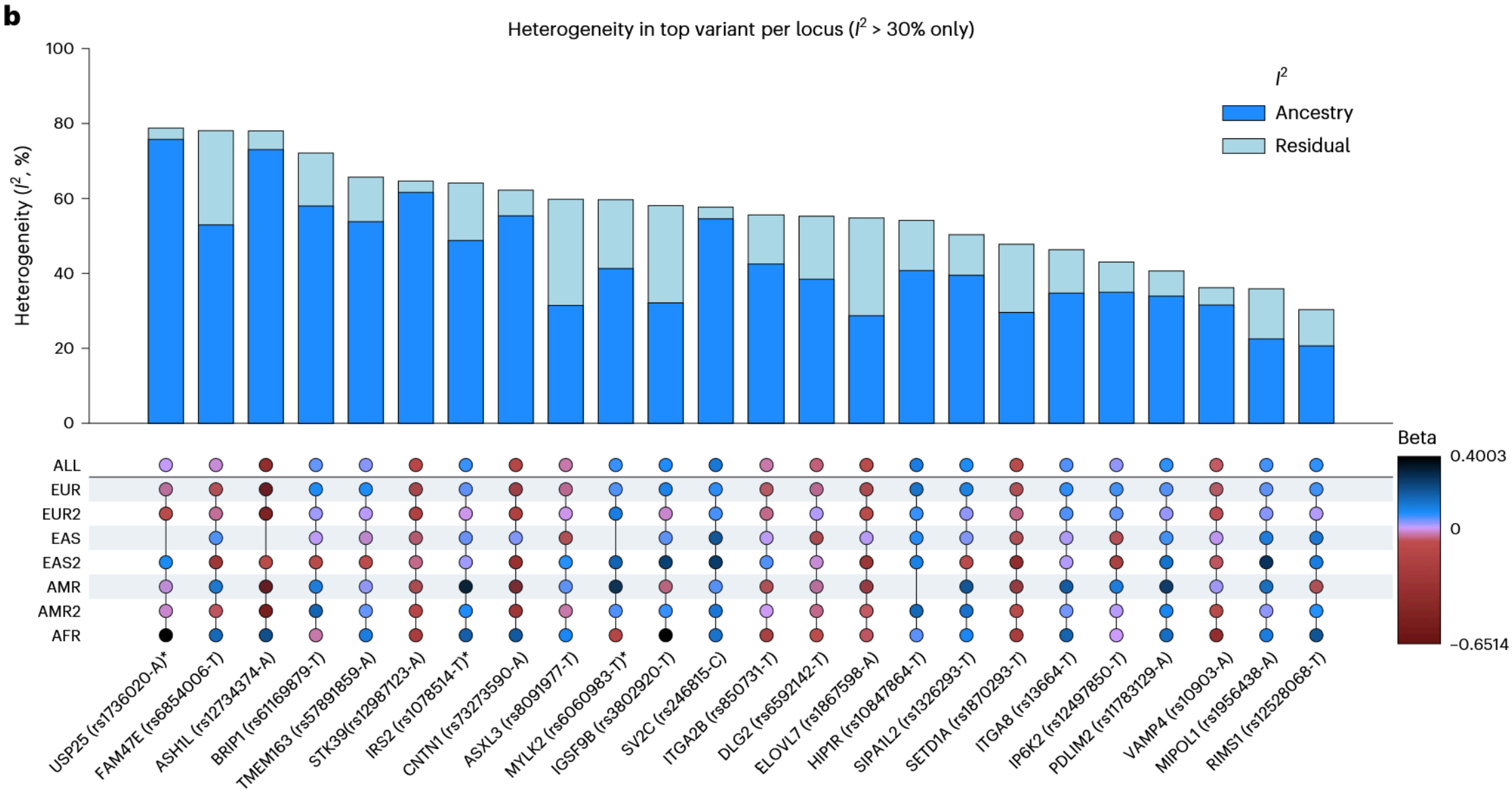
Fine-mapping



Colocalization with eQTL

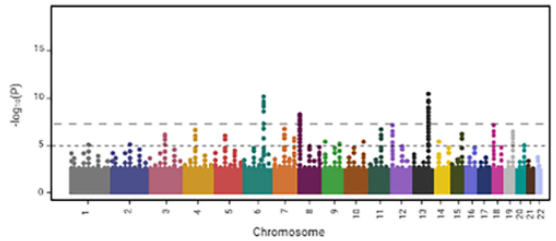
Manhattan Plot



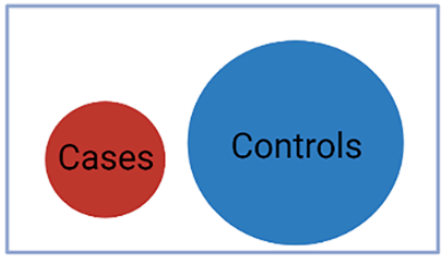

b
Heterogeneity in top variant per locus (I^2 > 30% only)
Heterogeneity (I^2, %)
100
80
60
40
20
0
I^2
Ancestry
Residual
Beta
0.4003
0
−0.6514
ALL
EUR
EUR2
EAS
EAS2
AMR
AMR2
AFR
USP25 (rs1736020-A)*
FAM47E (rs6854006-T)
ASH1L (rs12734374-A)
BRIP1 (rs61169879-T)
TMEM163 (rs57891859-A)
STK39(rs12987123-A)
IRS2 (rs1078514-T)*
CNTN1 (rs73273590-A)
ASXL3 (rs8091977-T)
MYLK2 (rs6060983-T)*
IGSF9B (rs3802920-T)
SV2C (rs246815-C)
ITGA2B (rs850731-T)
DLG2 (rs6592142-T)
ELOVL7 (rs1867598-A)
HIP1R (rs10847864-T)
SIPA1L2 (rs1326293-T)
SETD1A (rs1870293-T)
ITGA8 (rs13664-T)
IP6K2 (rs12497850-T)
PDLIM2 (rs11783129-A)
VAMP4 (rs10903-A)
MIPOL1 (rs1956438-A)
RIMS1 (rs12528068-T)

Locus	Number of significant SNPs	Nominated variant	CHR:BP:A1:A2	Nearest gene	Known PD gene ± 1 MB	Functional consequence
11	6	rs57891859	2:135464616:A:G	*TMEM163*	*TMEM163*	intronic
19	926	rs34311866	4:951947:C:T	*TMEM175*	*TMEM175*	exonic
23	1483	rs356182	4:90626111:A:G	*SNCA*	*SNCA*	ncRNA intronic
24	121	rs13117519	4:114369065:T:C	*CAMK2D*	*CAMK2D*	intergenic
45	1371	rs10847864	12:123326598:G:T	*HIP1R*	*HIP1R*	intronic
60	1	rs55818311	19:2341047:C:T	*SPPL2B*	*LSM7*	ncRNA exonic

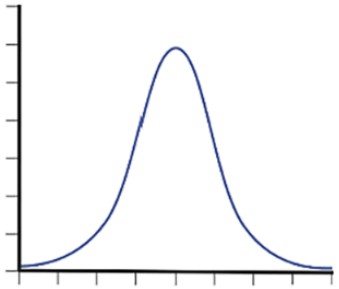
Clinical Application of GWAS



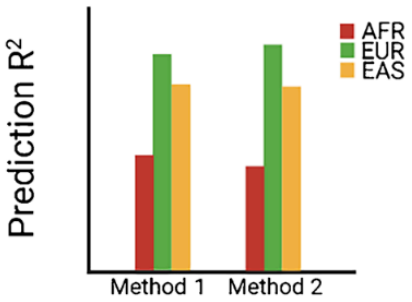
I. Obtain base data of betas/odds ratios from GWAS summary statistics.



II. Train/develop PRS based on independent data using method(s) of choice and weights from GWAS summary statistics.



III. Tune or calibrate PRS based on independent data.

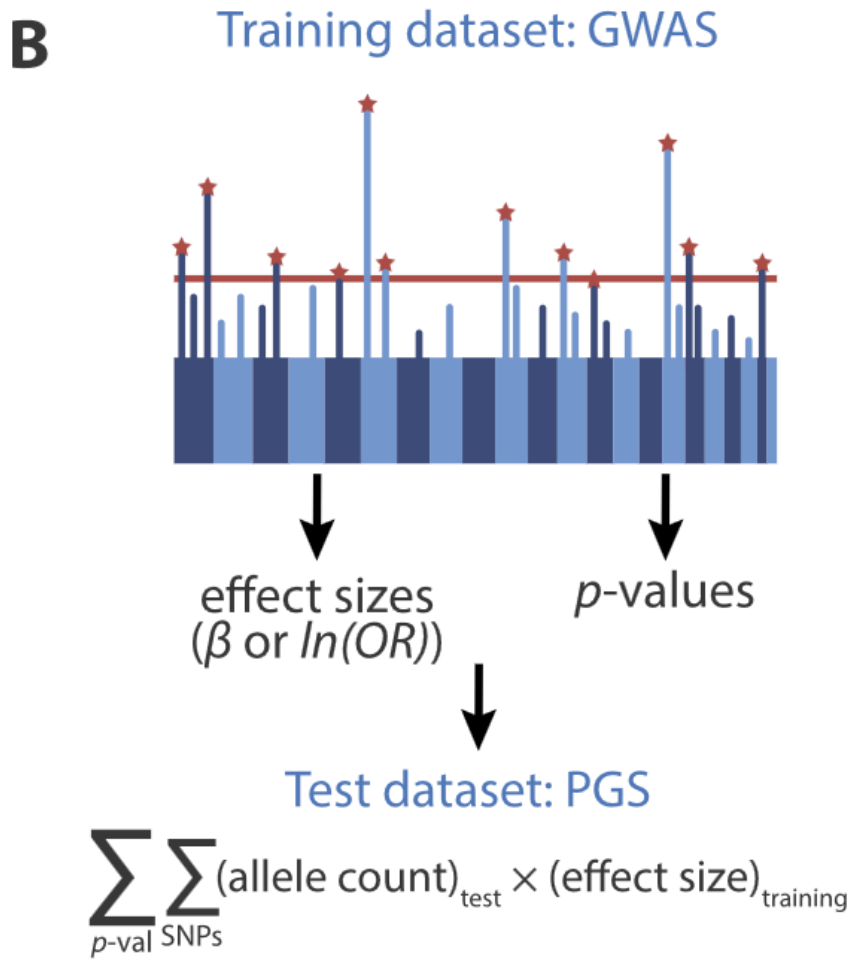


IV. Validate PRS in additional independent cohorts. Assess PRS prediction accuracy.

V. Use PRS to identify individuals at high risk.

VI. Relay information to clinician for decision-making and clinical management.

Polygenic Risk Score



- Sum of the weighted allele counts of independent SNPs found to be associated with a trait or disorder in a GWAS
- For quantitative traits, the weights are the linear regression effect sizes, i. e., the β coefficients.
- For case/control phenotypes, the weights are the natural logarithm of the odds ratio, ln(OR).