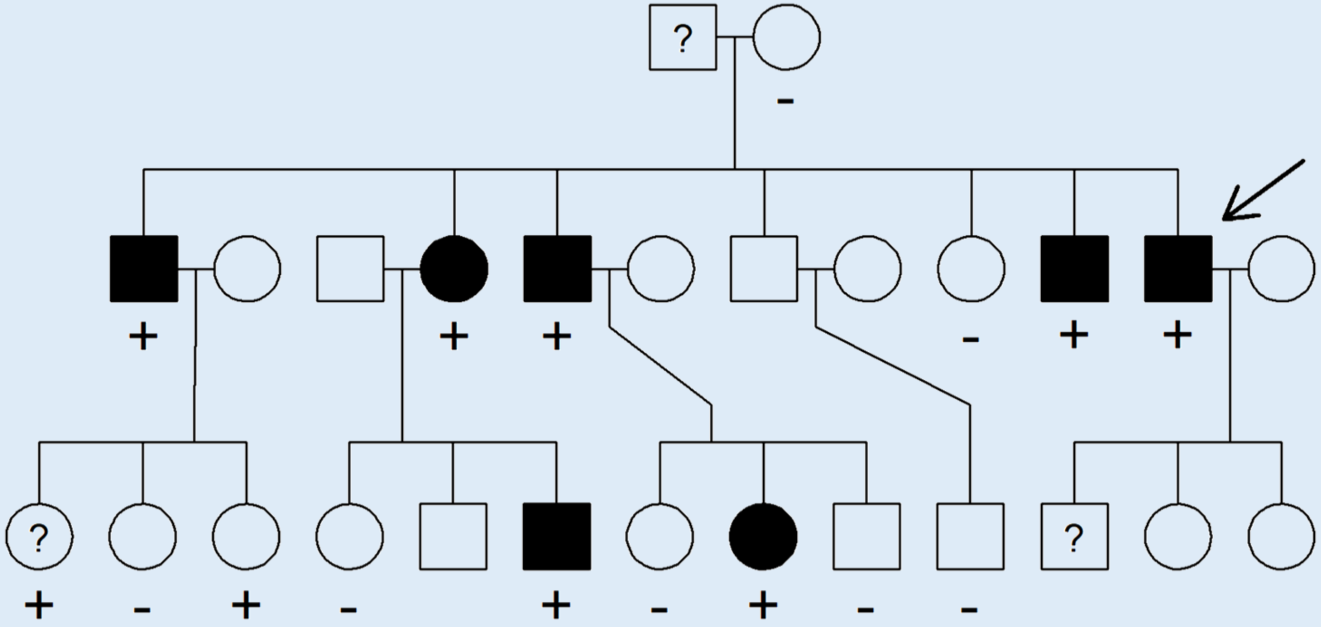


Linkage and segregation analysis in medical genetics

Statistical methods in genetic relatedness and pedigree analysis

NORBIS course, 13th – 17th of June 2022, Oslo

Magnus Dehli Vigeland

Outline

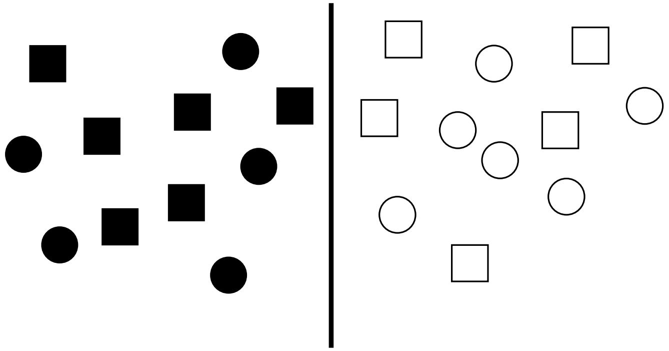
- Monogenic diseases: Inheritance patterns

- Traditional linkage analysis

- Segregation analysis in modern genetic diagnostics

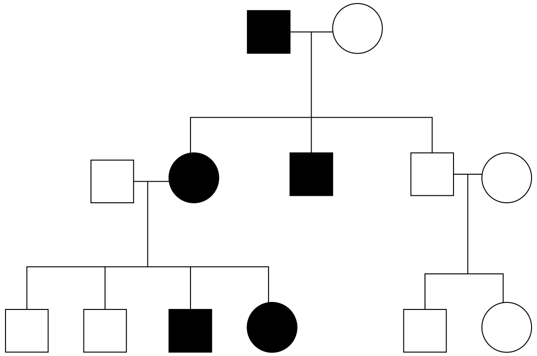
Finding the cause of genetic diseases: Two main approaches

Multifactorial disease



- Association analysis
- Case/control design
- Population based

Monogenic disease



- Linkage analysis
- Family design

Monogenic diseases: Inheritance patterns

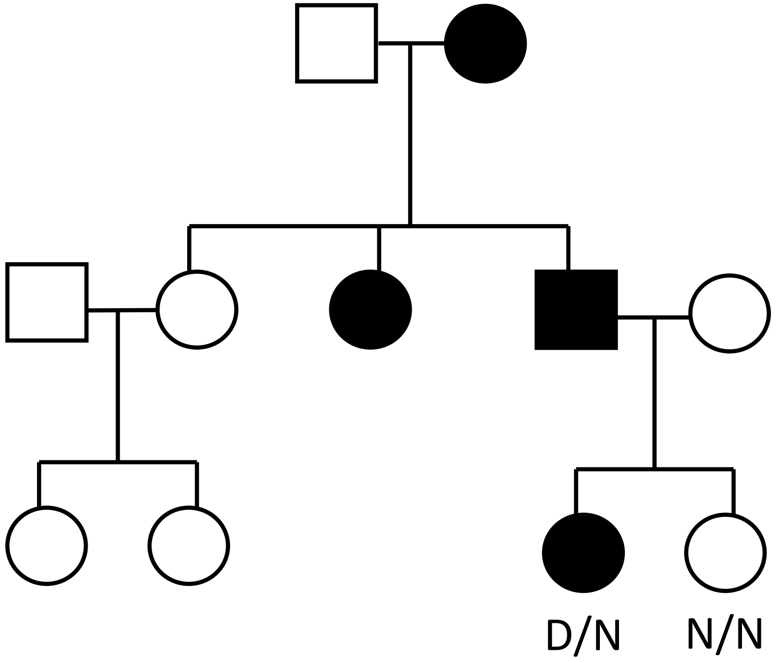
Autosomal dominant (AD) inheritance

Penetrance parameters

$f_0 = P(\text{aff} \mid NN)$

$f_1 = P(\text{aff} \mid DN)$

$f_2 = P(\text{aff} \mid DD)$



Penetrance values:

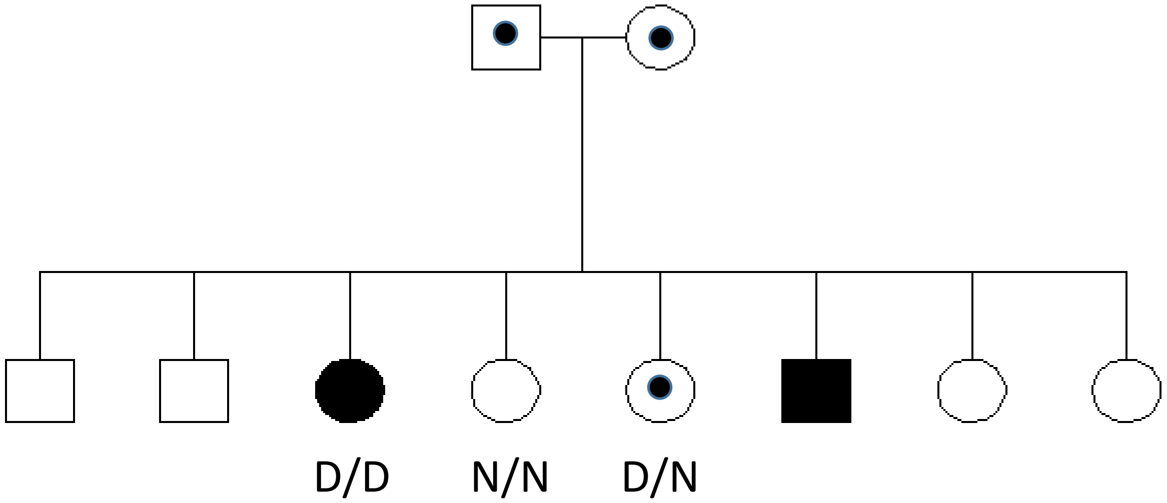
f_0 = 0

f_1 = 1

f_2 = 1

- ~50% affected children of an affected mother or father
- Affected male:female ratio is ~1
- Can be inherited from mother or father to both sons and daughters

Autosomal recessive (AR) inheritance



Penetrance values:

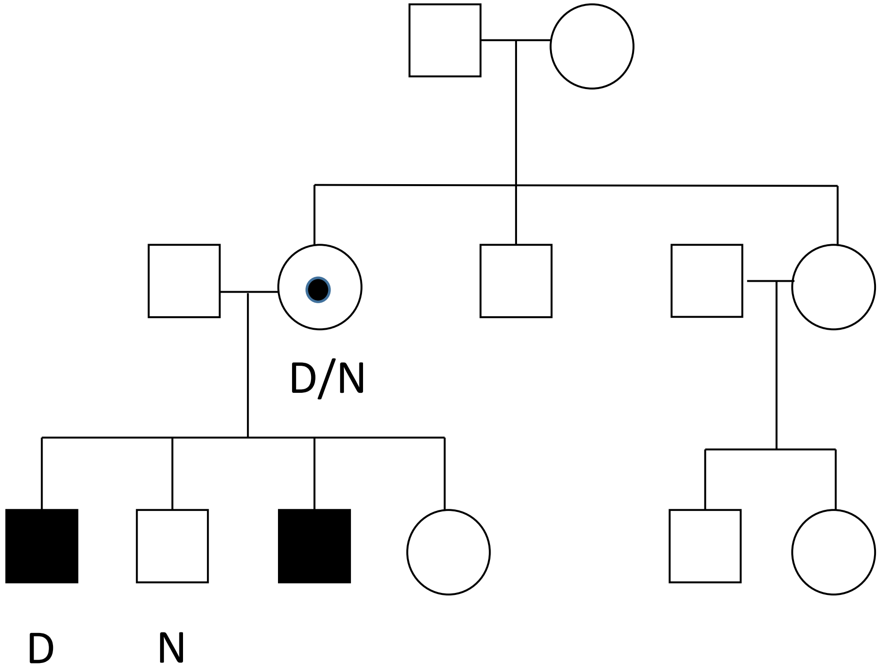
$f_0 = 0$

$f_1 = 0$

$f_2 = 1$

- Usually healthy parents
- ~25% of the children are affected
- Affected male:female ratio is ~1

X-linked recessive inheritance



Penetrance values:

Females:

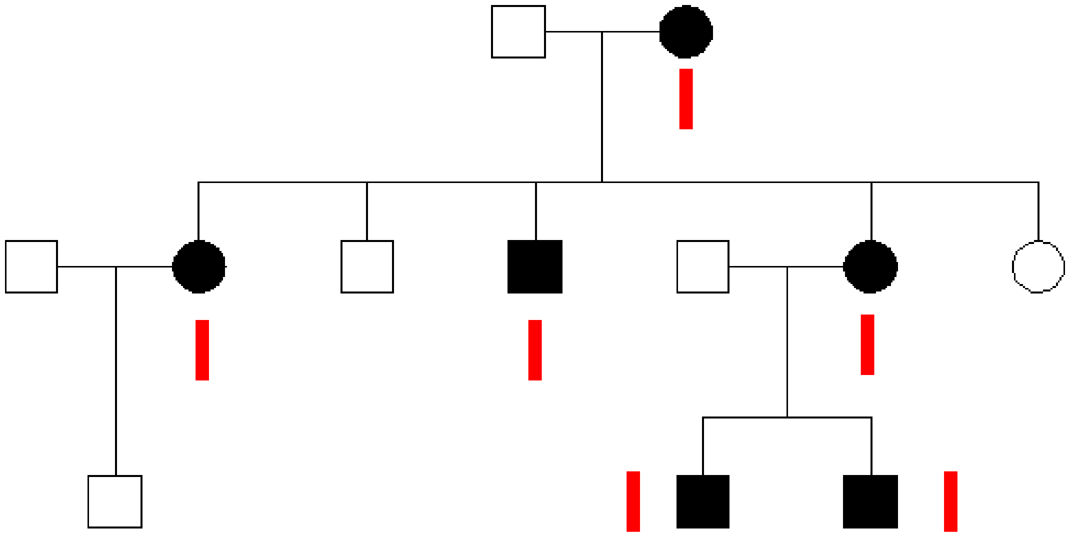
$f_0 = 0$
$f_1 = 0$
$f_2 = 1$

Males:

$f_0 = 0$
$f_1 = 1$

- Only males are affected
- Usually inherited through healthy females

Linkage analysis – the basic principle

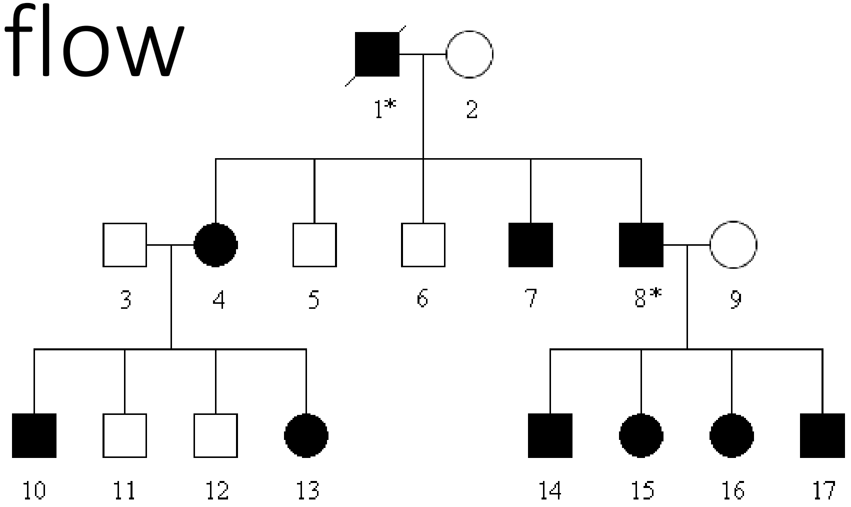


Compare

- the inheritance pattern of the disease
- IBD pattern across the genome

Goal: Identify the region harbouring the disease-causing locus

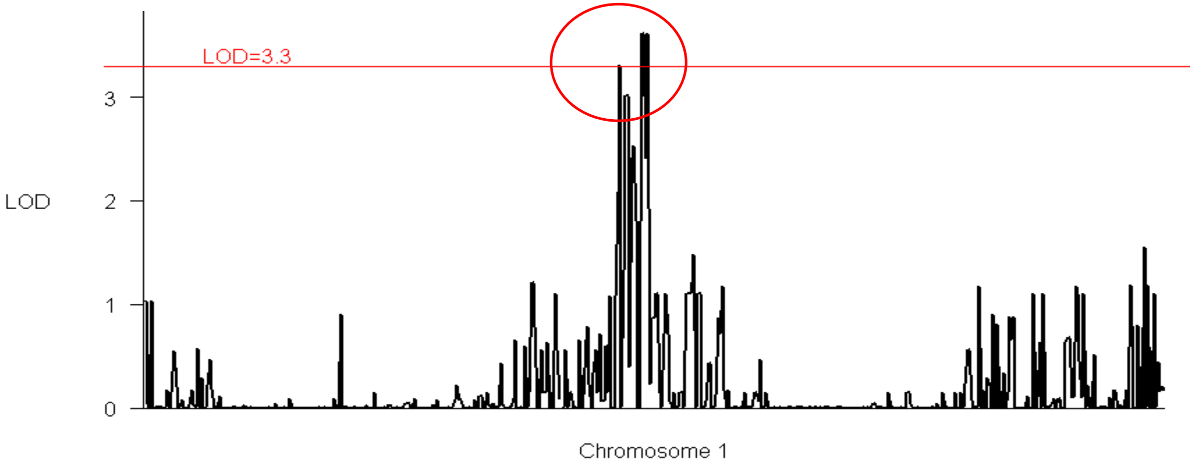
Linkage analysis - traditional workflow



1. Starting point: Large affected family

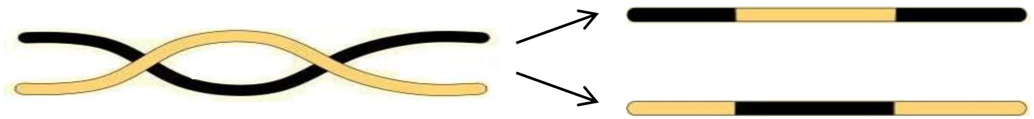
2. SNP genotyping

3. Parametric linkage analysis



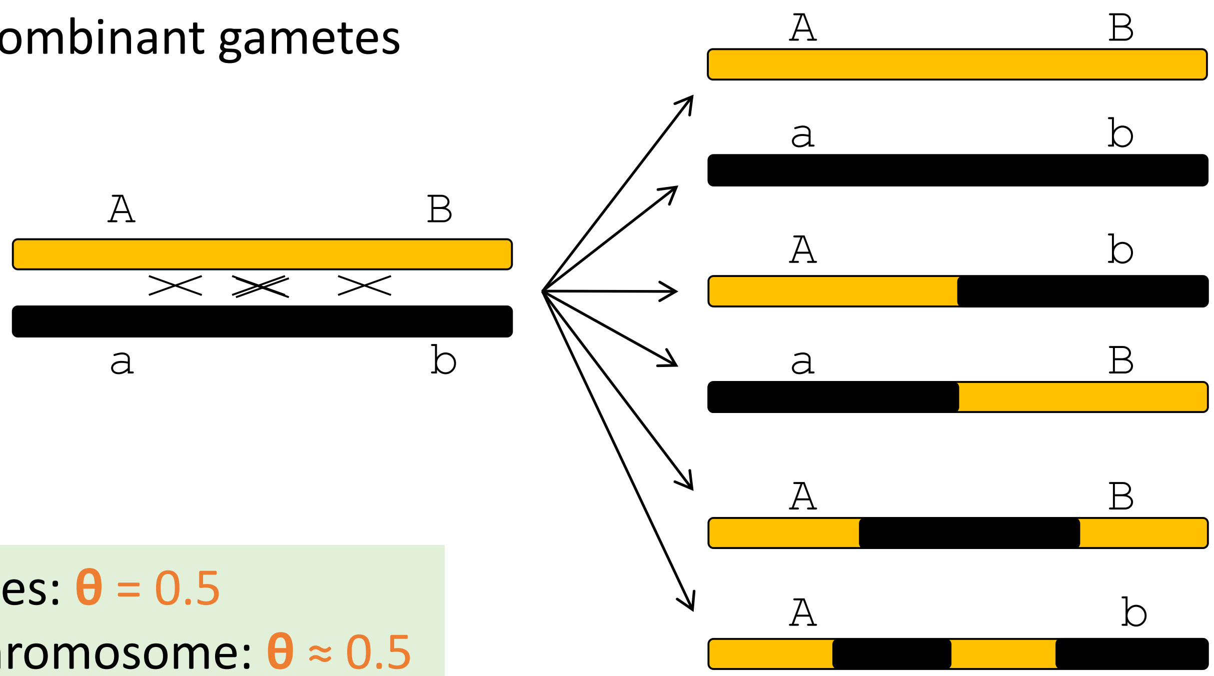
4. Sequence genes in linkage peak → identify causal mutation

Recombination rate



- The recombination rate between two loci

  = average number of recombinant gametes



non-recombinant

non-recombinant

recombinant

recombinant

non-recombinant (!)

recombinant

Loci on different chromosomes: **θ** = 0.5
Loci far apart on the same chromosome: **θ** ≈ 0.5
Loci right next to each other: **θ** = 0

Definition: Two loci are *linked* if **θ** < 0.5
"On the same chromosome, not too far apart"

Crossover rate vs. recombination rate

Crossover rate (= *genetic distance*)

$$d = \mathrm{E}[\ \#\text{crossovers}\ ]$$

- Based on a fundamental property of the meiosis
- Statistically well-behaved
- But:

Hard to observe directly

Recombination rate

$$\theta = \mathrm{E}[\ \#\text{recombinant gametes}\ ]$$

- Not as intuitive
- Relative to markers
- But:

Easy to estimate using genotyping

Haldane's map function: $\boxed{\theta = \tfrac{1}{2}(1 - e^{-2d})}$

Hypothesis testing in linkage

- Hypotheses:

  H_0: $\theta = 0.5$ (no linkage)

  H_A: $\theta < 0.5$ (linkage)

θ = recombination rate between marker and disease

- For historical reasons the test statistic is

$$\text{LOD} = \log_{10} \frac{P(\text{data} \mid \theta = \theta)}{P(\text{data} \mid \theta = 0.5)}$$

LOD = "logarithm of the odds"

- Traditional significance thresholds:
  - Autosomal loci: LOD = 3 ($p \approx 0.0001$)
  - X-linked loci: LOD = 1.8

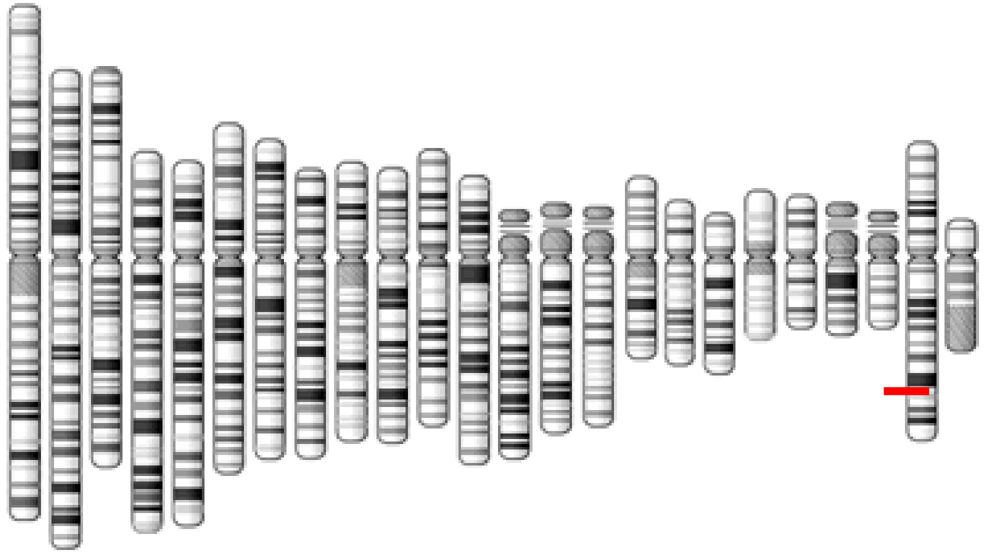
Why the LOD? Why not p<0.05?

- The common significance level $\alpha = 0.05$ is not used in linkage analysis
- Reason: Low *a priori* probability of H_A

H_0: $\theta = 0.5$ (no linkage)

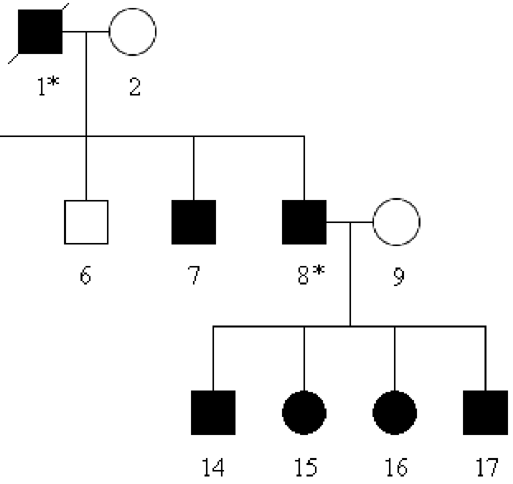
H_A: $\theta < 0.5$ (linkage)

$P(H_A)$ = P(random marker near the disease) $\approx 1/50$



© Clinical Tools, Inc.

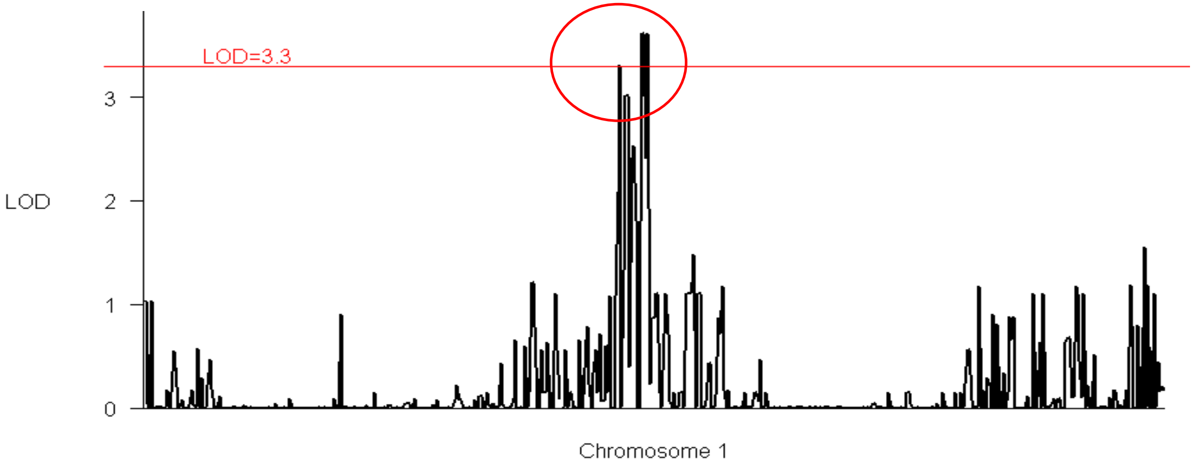
Linkage analysis - traditional workflow



1. Starting point: Large affected family

2. SNP genotyping

3. Parametric linkage analysis



4. Sequence genes in linkage peak → identify causal mutation

Modern approach (last decade)

1. Patient genome → identify the causal variant

1. Sequence

High-throughput sequencing

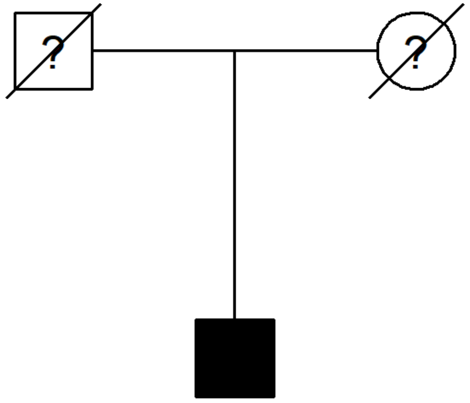
2. List of variants

Databases

- dbSNP / dbVAR
- 1000 Genomes / ExAC / gnomAD
- OMIM / ClinVar / HGMD
- + many others!

Interpretation guidelines

Segregation: Motivating example



- Autosomal dominant disease
- Suspected DNA variant identified:
  - Rare/novel
  - Predicted damaging
  - But...no previous connection with disease

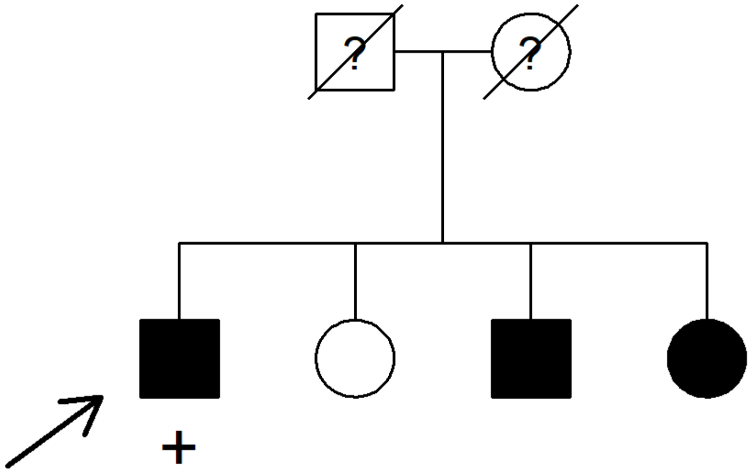
Classification: VUS

(Variant of uncertain significance)

What to do?

- ~~Replicate in unrelated patient~~
- ~~Functional studies~~
- Segregation?

Segregation: Motivating example



+ Confirmed carrier
– Confirmed noncarrier
↗ Proband/index patient

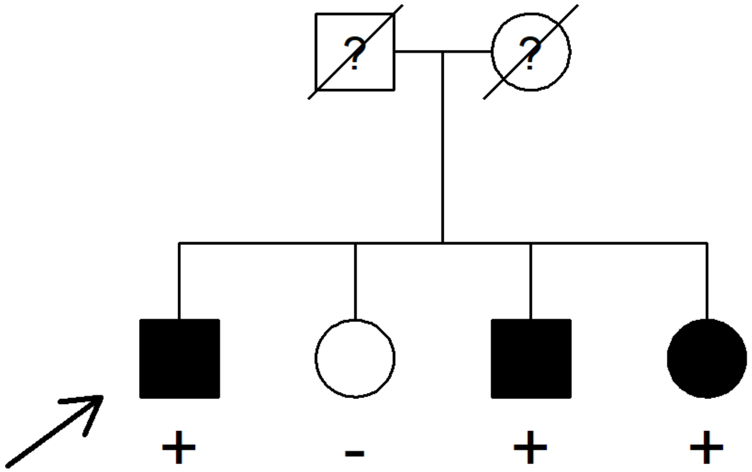
- Autosomal dominant disease
- Suspected DNA variant identified:
  - Rare/novel
  - Predicted damaging
  - But...no previous connection with disease

Classification: VUS
(Variant of uncertain significance)

What to do?

- ~~Replicate in unrelated patient~~
- ~~Functional studies~~
- Segregation?

Segregation: Motivating example



- Autosomal dominant disease
- Suspected DNA variant identified:
  - Rare/novel
  - Predicted damaging
  - But...no previous connection with disease

Classification: VUS
(Variant of uncertain significance)

What to do?

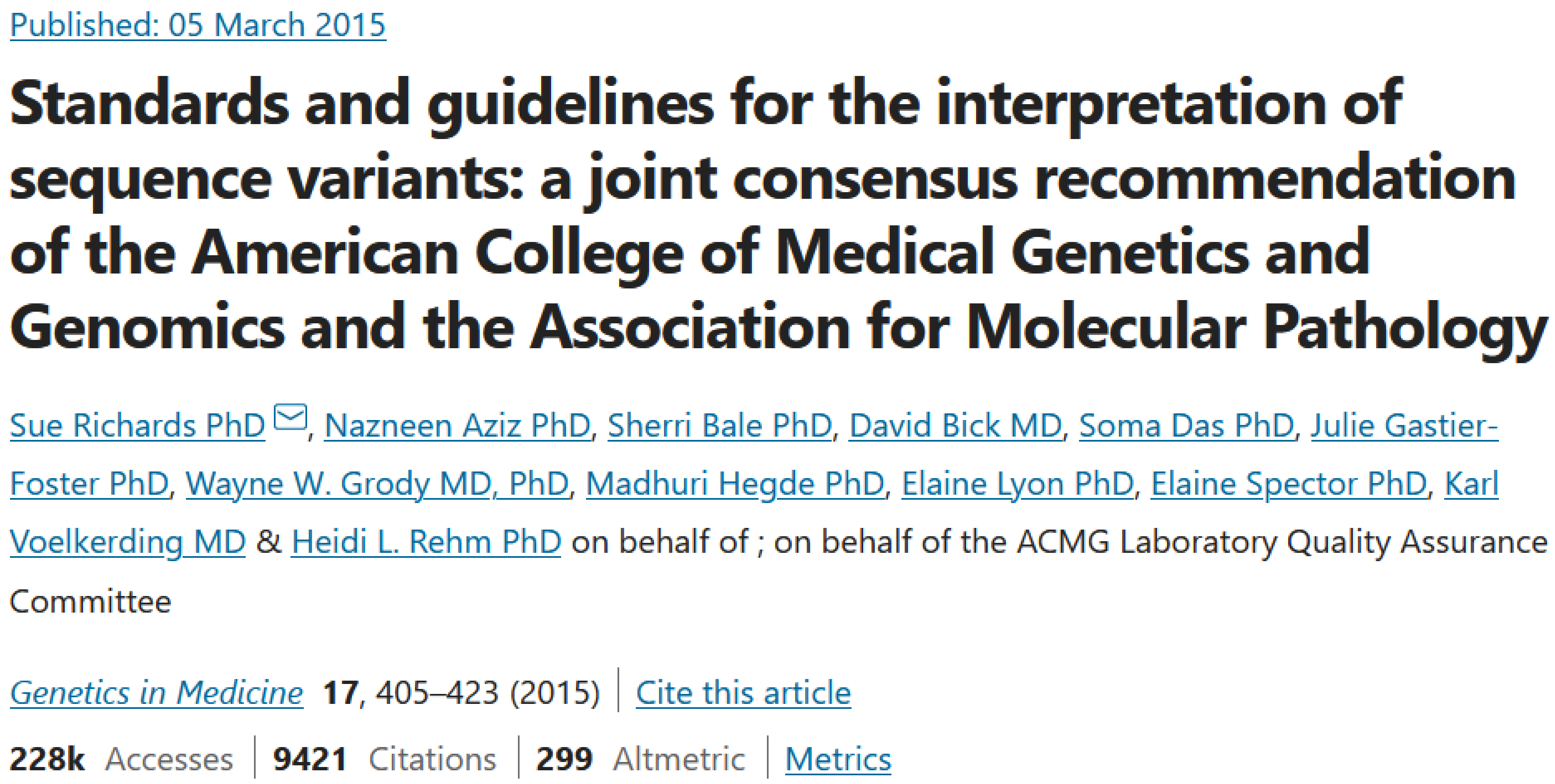
- ~~Replicate in unrelated patient~~
- ~~Functional studies~~
- Segregation?

Co-segregation supports pathogenicity!

But **how much**?

+ Confirmed carrier
– Confirmed noncarrier
↗ Proband/index patient

The ACMG framework for variant interpretation



Published: 05 March 2015

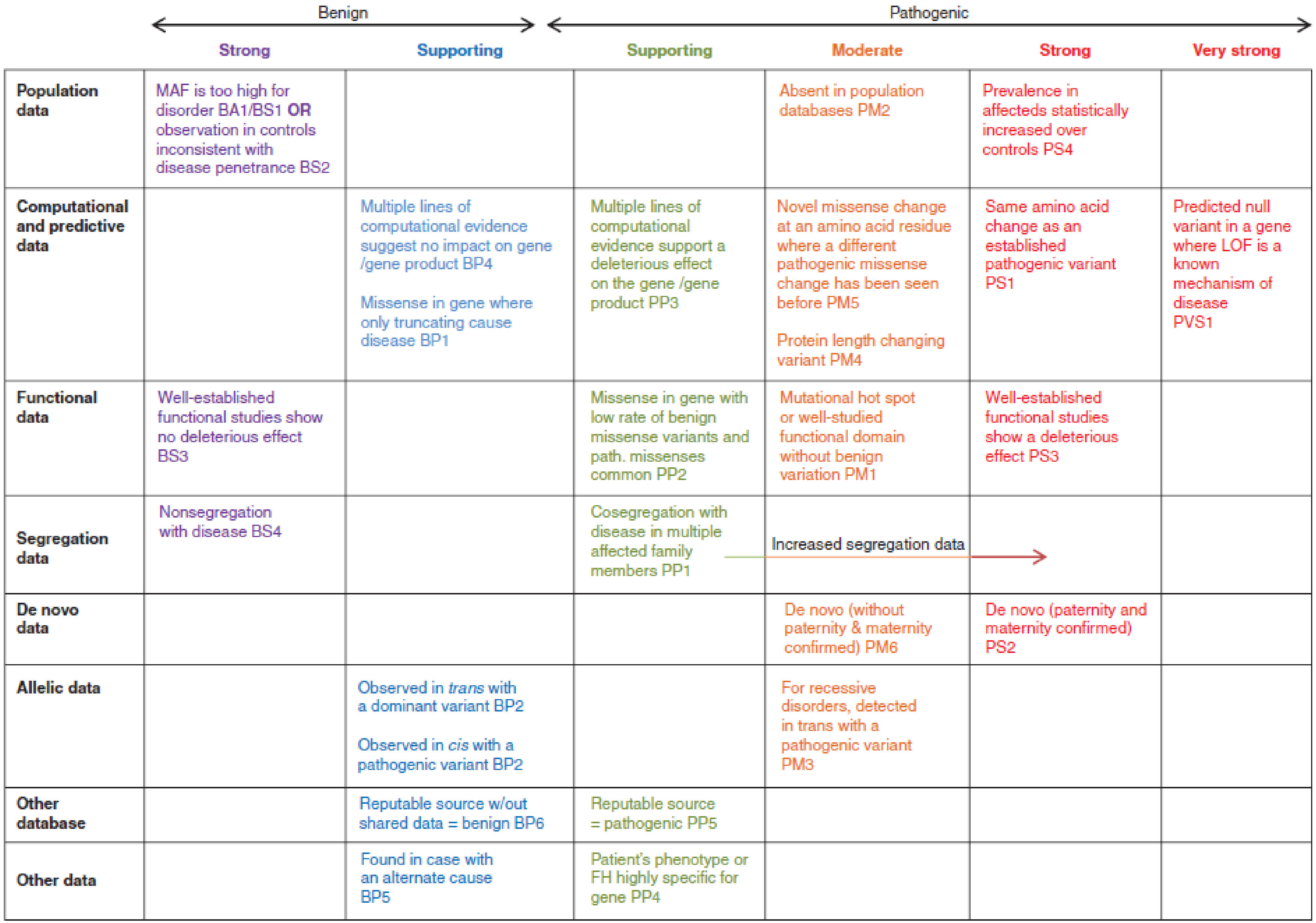
Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards PhD, Nazneen Aziz PhD, Sherri Bale PhD, David Bick MD, Soma Das PhD, Julie Gastier-Foster PhD, Wayne W. Grody MD, PhD, Madhuri Hegde PhD, Elaine Lyon PhD, Elaine Spector PhD, Karl Voelkerding MD & Heidi L. Rehm PhD on behalf of ; on behalf of the ACMG Laboratory Quality Assurance Committee

Genetics in Medicine **17**, 405–423 (2015) | Cite this article

228k Accesses | **9421** Citations | **299** Altmetric | Metrics

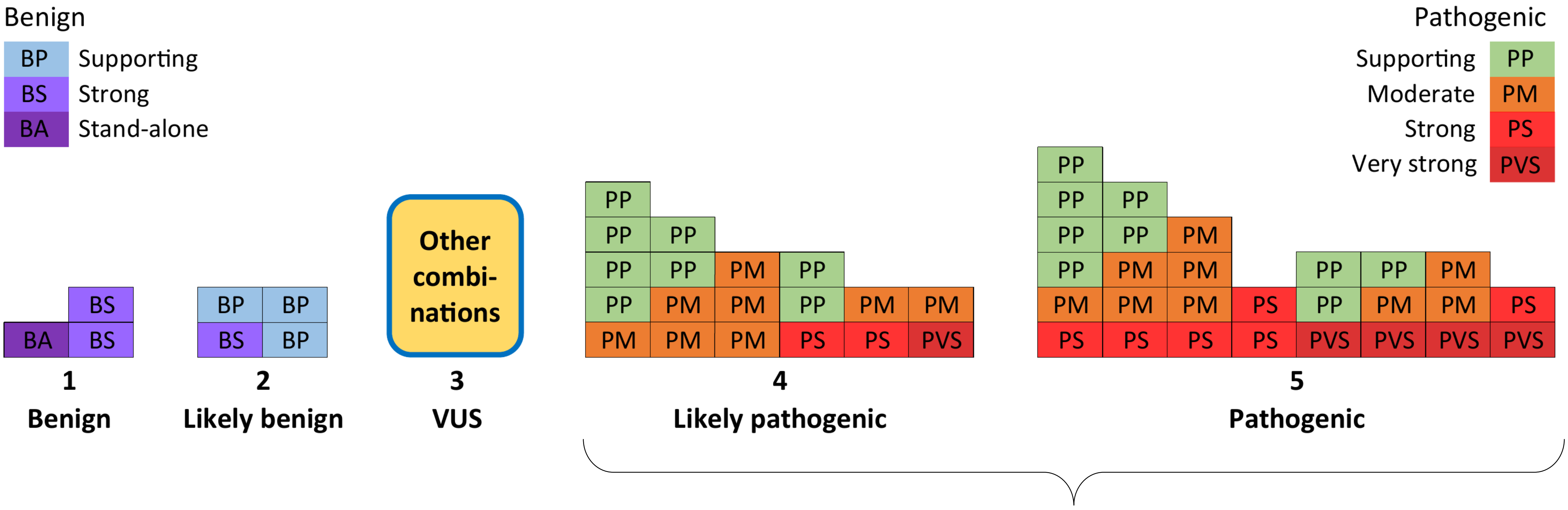
ACMG evidence framework



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Source: Richards et al., (2015)

ACMG classification rules



Benign

BP	Supporting
BS	Strong
BA	Stand-alone

Pathogenic

Supporting	PP
Moderate	PM
Strong	PS
Very strong	PVS

BA

BS
BS

1
Benign

BP
BS

BP
BP

2
Likely benign

Other combi-nations

3
VUS

PP
PP
PP
PP
PM

PP
PP
PM
PM

PM
PM
PM

PP
PP
PS

PM
PS

PM
PVS

4
Likely pathogenic

PP
PP
PP
PP
PM
PS

PP
PP
PM
PM
PS

PM
PM
PM
PS

PS
PS

PP
PP
PVS

PP
PM
PVS

PM
PM
PVS

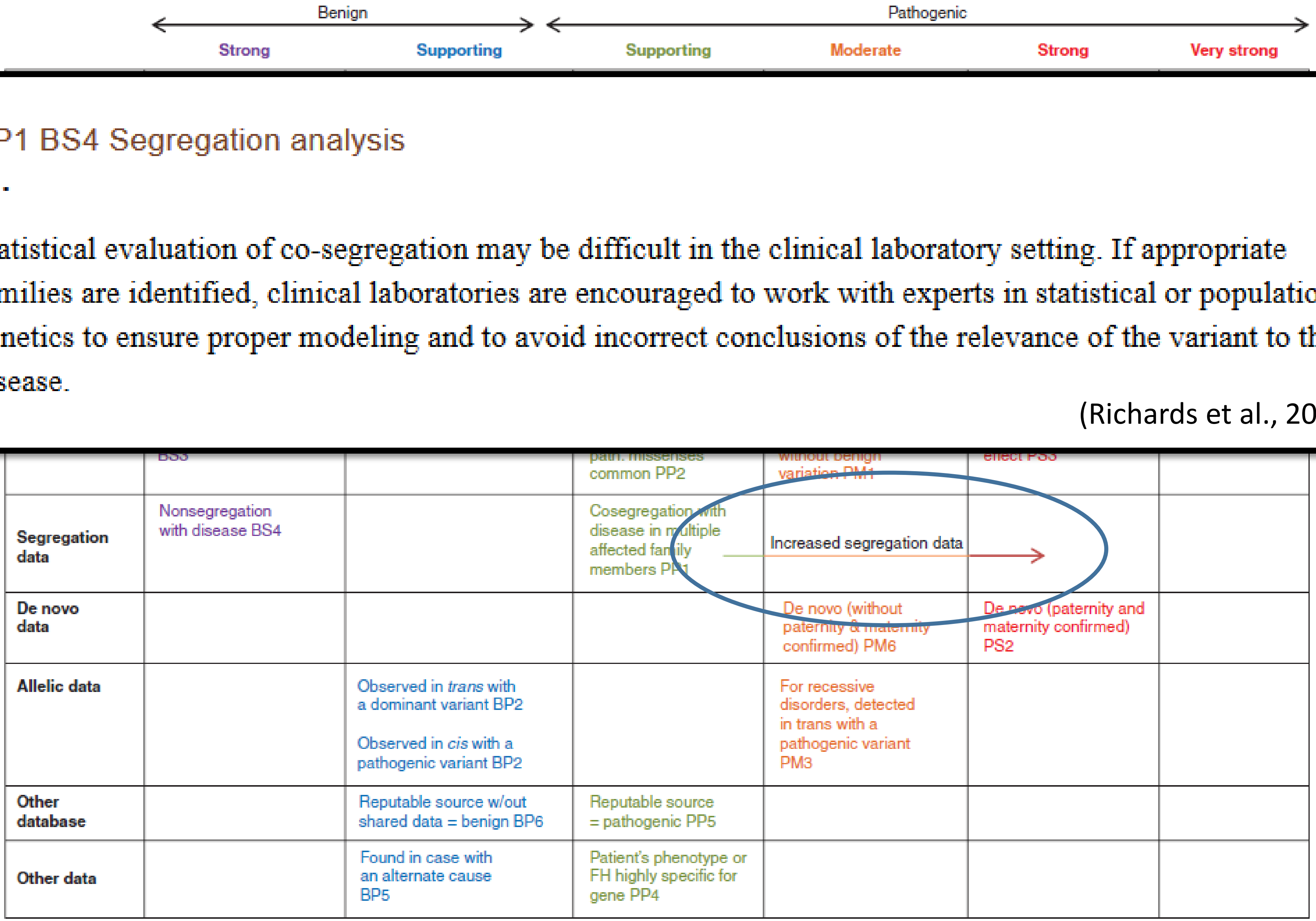
PS
PVS

5
Pathogenic

= genetic diagnosis

⇒ access to treatment & follow-up of patient and family

ACMG evidence framework



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
	BS3		path. missenses common PP2	without benign variation PM1	effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

PP1 BS4 Segregation analysis

…

Statistical evaluation of co-segregation may be difficult in the clinical laboratory setting. If appropriate families are identified, clinical laboratories are encouraged to work with experts in statistical or population genetics to ensure proper modeling and to avoid incorrect conclusions of the relevance of the variant to the disease.

(Richards et al., 2015)

Source: Richards et al., (2015)

How to **quantify** segregation evidence?

- 2003: Thompson et al. (Am J Hum Genet)

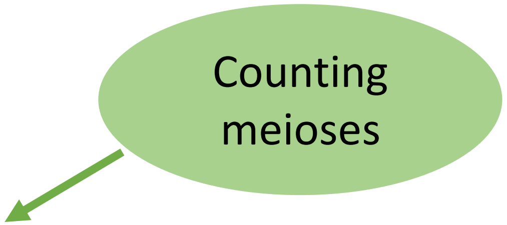
  A full-likelihood method for the evaluation of causality of sequence variants from family data

Bayes factor

- 2008: Bayrak-Toydemir et al (Exp Mol Pathol)

  Likelihood ratios to assess genetic evidence for clinical significance of uncertain variants: Hereditary hemorrhagic telangiectasia as a model

----



Counting meioses

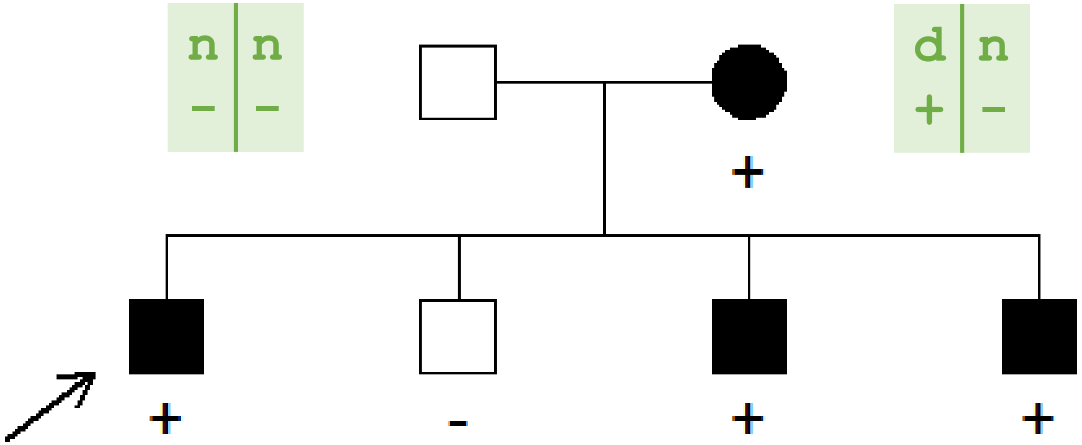
- 2016: Jarvik & Browning (Am J Hum Genet)

  Consideration of cosegregation in the pathogenicity classification of genomic variants

Jarvik & Browning (2016)

1. Segregation score based on counting: $N = (1/2)^m$

m = number of meioses informative for cosegregation



$m = 4$

(moderate evidence)

2. Suggested ACMG thresholds

- $N \leq \frac{1}{32} \approx 0.03$: **strong** $(m \geq 5)$
- $N \leq \frac{1}{16} \approx 0.06$: **moderate** $(m \geq 4)$
- $N \leq \frac{1}{8} = 0.125$: **supportive** $(m \geq 3)$

Jarvik & Browning

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact *BP4* Missense when only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7* In-frame indels in repeat w/out known function *BP3*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		N ≤ 1/8 if 1 family N ≤ 1/4 if > 1 family	N ≤ 1/16 if 1 family N ≤ 1/8 if > 1 family	N ≤ 1/32 if 1 family N ≤ 1/16 if > 1 family	
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Jarvik & Browning (2016)

- Advantage
  - Works well in **simple cases**

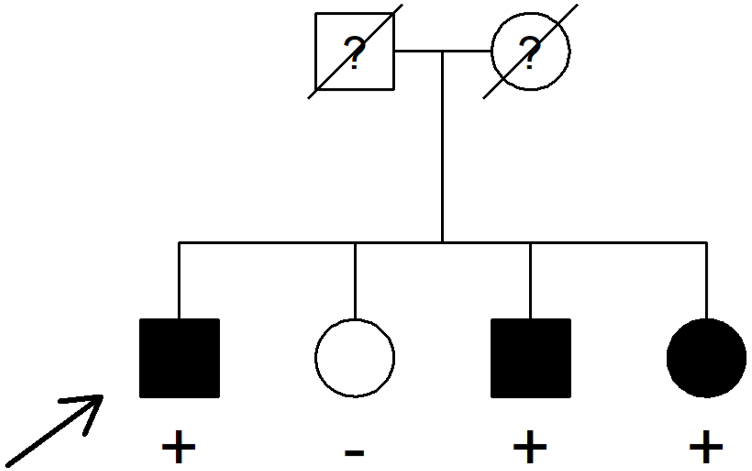
- What are **simple cases**?
  - no phenocopies
  - complete penetrance
  - allele entered pedigree only once
  - "everyone" genotyped

- Disadvantage:
  - Confusing and potentially inaccurate in other cases

Disease model (f_0, f_1, f_2)

$$f_0 = P(\text{affected} \mid \text{genotype} = \text{n/n})$$
$$f_1 = P(\text{affected} \mid \text{genotype} = \text{d/n})$$
$$f_2 = P(\text{affected} \mid \text{genotype} = \text{d/d})$$

- f_0 = phenocopy rate
- f_1 = penetrance



"[Thompson's method is more accurate], but requires training and tools." (J&B 2016)

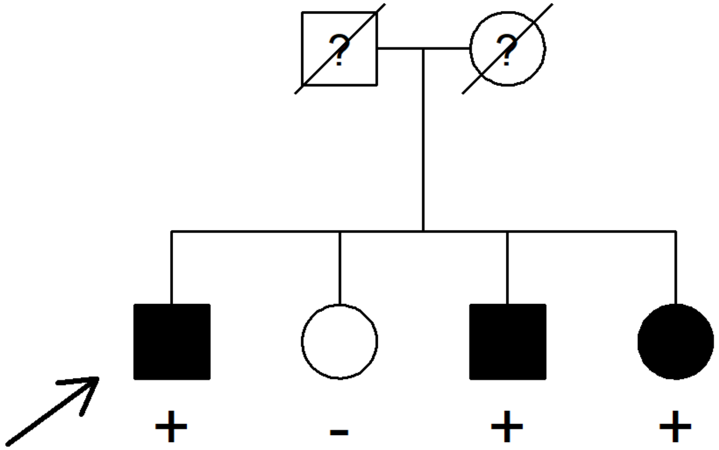
The Bayes factor approach

- Recall, for models M_1 and M_2:

$$\frac{P(M_1 \mid \text{data})}{P(M_2 \mid \text{data})} = \frac{P(\text{data} \mid M_1)}{P(\text{data} \mid M_2)} \cdot \frac{P(M_1)}{P(M_2)}$$

$$\textit{Posterior odds} = \textit{Bayes factor} \cdot \textit{Prior odds}$$

- Our models
  - $\mathcal{C}$: variant is causal
  - $\bar{\mathcal{C}}$: variant is not causal
- Our data
  - Carrier statuses $\boldsymbol{g}$
  - Affection statuses $\boldsymbol{a}$

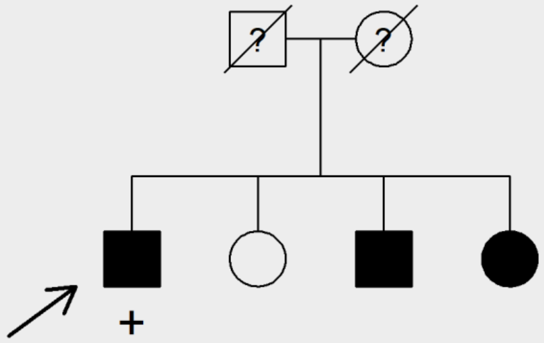


Fixed parameters
- Pedigree
- Allele freqs
- Disease model
- +++

The Bayes factor approach (2)

$\boldsymbol{a}$ = affection status vector
$\boldsymbol{g}$ = genotype vector
g_0 = proband genotype

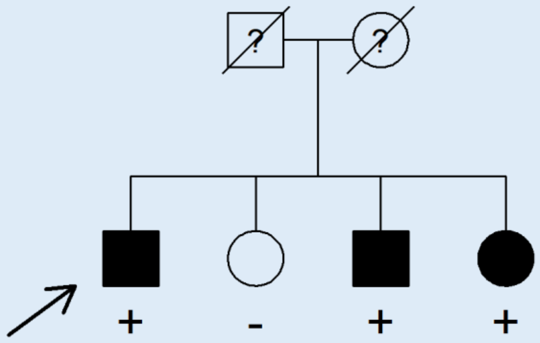
Original situation



$$\frac{P(\mathcal{C} \mid g_0, \boldsymbol{a})}{P(\overline{\mathcal{C}} \mid g_0, \boldsymbol{a})} = B_0 \cdot \frac{P(\mathcal{C})}{P(\overline{\mathcal{C}})}$$

$$\text{where } B_0 = \frac{P(g_0,\boldsymbol{a}|\mathcal{C})}{P(g_0,\boldsymbol{a}|\overline{\mathcal{C}})}$$

After additional genotyping



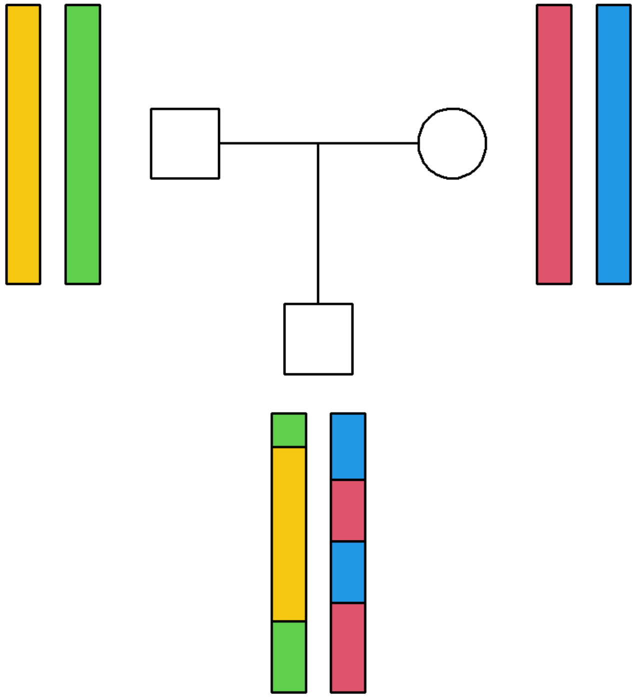
$$\frac{P(\mathcal{C} \mid \boldsymbol{g}, \boldsymbol{a})}{P(\overline{\mathcal{C}} \mid \boldsymbol{g}, \boldsymbol{a})} = B_1 \cdot \frac{P(\mathcal{C})}{P(\overline{\mathcal{C}})}$$

$$\text{where } B_1 = \frac{P(\boldsymbol{g},\boldsymbol{a}|\mathcal{C})}{P(\boldsymbol{g},\boldsymbol{a}|\overline{\mathcal{C}})}$$

Contribution of segregation analysis:

$$\frac{B_1}{B_0} = \frac{P(\boldsymbol{g}, \boldsymbol{a} \mid \mathcal{C})}{P(\boldsymbol{g}, \boldsymbol{a} \mid \overline{\mathcal{C}})} \Bigg/ \frac{P(g_0, \boldsymbol{a} \mid \mathcal{C})}{P(g_0, \boldsymbol{a} \mid \overline{\mathcal{C}})}$$

Parametrization of models $\mathcal{C}$ and $\bar{\mathcal{C}}$



- $\bar{\mathcal{C}}$ (variant independent of disease allele)
  - Unlinked $\rho = 0.5$
  - Linkage equilibrium $r = 0$

- $\mathcal{C}$ (variant ~~is~~ inseparable from the disease allele)
  - Complete linkage $\rho = 0$
  - Complete LD $r = 1$

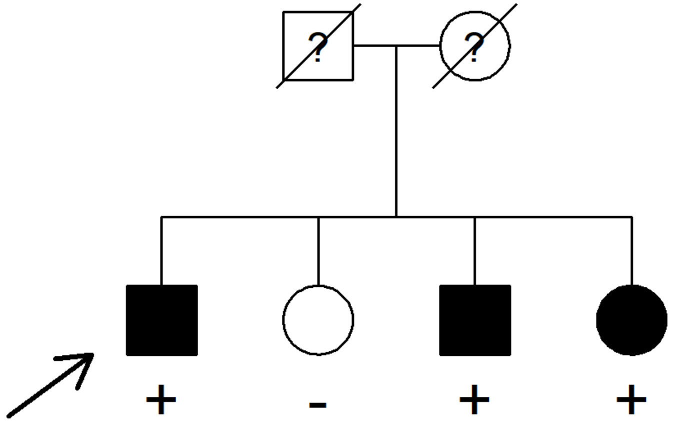
ρ = recombination rate
r = linkage disequilibrium

*Thompson et al. AJHM, 2003

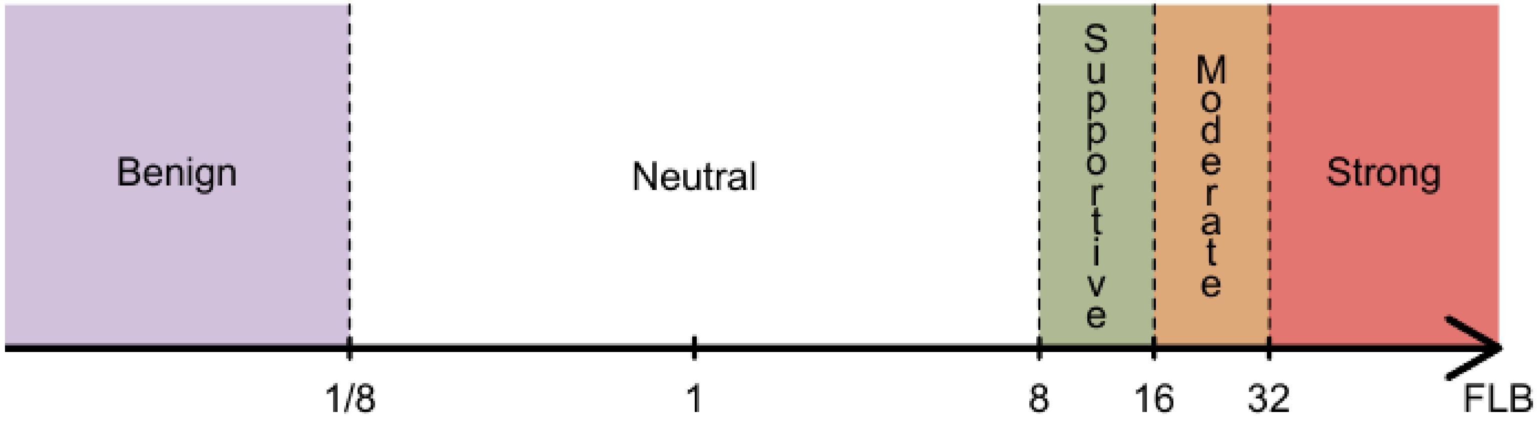
The full-likelihood Bayes factor (FLB)*

$$\boxed{\mathrm{FLB} = \frac{P(\boldsymbol{g}, \boldsymbol{a} \mid \rho = 0, r = 1)}{P(\boldsymbol{g}, \boldsymbol{a} \mid \rho = \frac{1}{2}, r = 0)} \Big/ \frac{P(g_0, \boldsymbol{a} \mid \rho = 0, r = 1)}{P(g_0, \boldsymbol{a} \mid \rho = \frac{1}{2}, r = 0)}}$$

- FLB = "LOD score adjusted for proband"
- FLB = 1/N (as given by J & B, under ideal conditions)



$\boldsymbol{a}$ = affection status vector
$\boldsymbol{g}$ = genotype vector
g_0 = proband genotype
ρ = recombination rate
r = linkage disequilibrium



Computing FLB

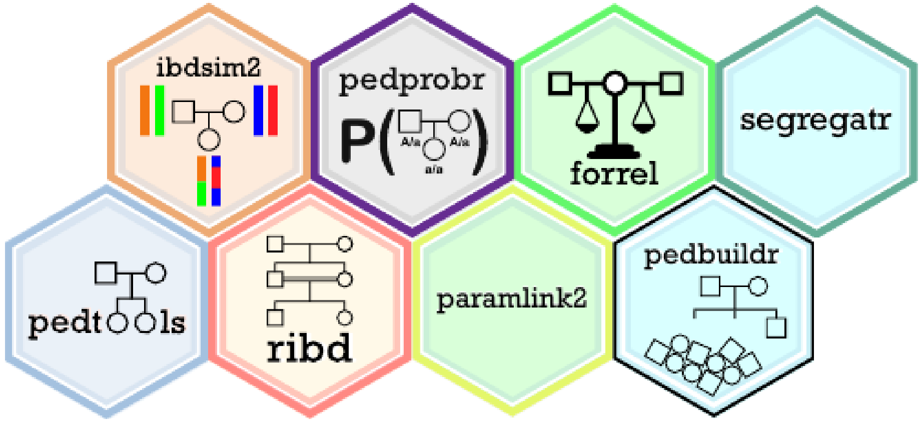
$$\mathrm{FLB} = \frac{P(\boldsymbol{g}, \boldsymbol{a} \mid \rho = 0, r = 1)}{P(\boldsymbol{g}, \boldsymbol{a} \mid \rho = \frac{1}{2}, r = 0)} \Big/ \frac{P(g_0, \boldsymbol{a} \mid \rho = 0, r = 1)}{P(g_0, \boldsymbol{a} \mid \rho = \frac{1}{2}, r = 0)}$$

- All terms are **pedigree likelihoods:** $P(genotypes \mid pedigree;\ \theta)$

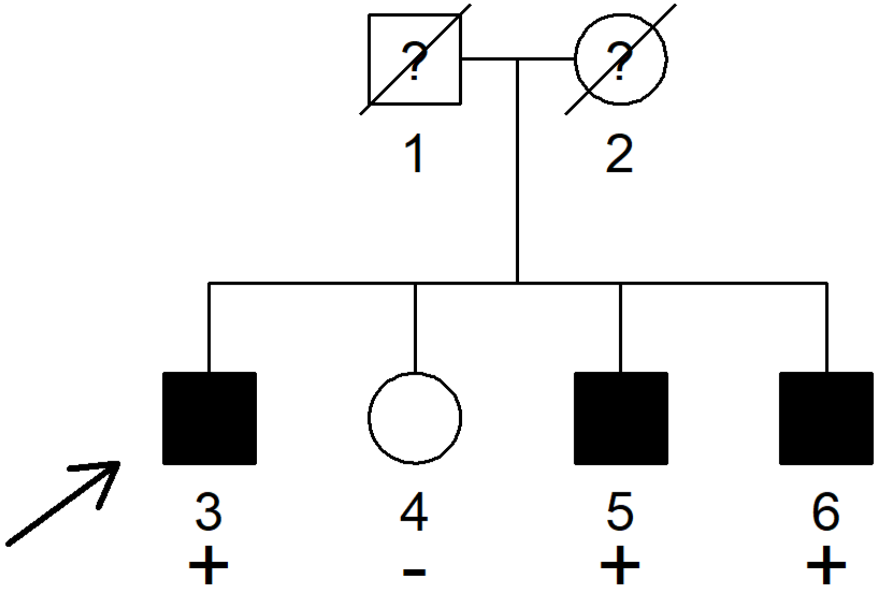
- Implementations notoriously cumbersome

- Old software still prevailing (require bioinformatic training):
  - LINKAGE, FastLink, Allegro, GeneHunter (Elston-Stewart algorithm)
  - MERLIN (Lander-Green algorithm)

- R: pedsuite/segregatr



segregatr: Segregation analysis

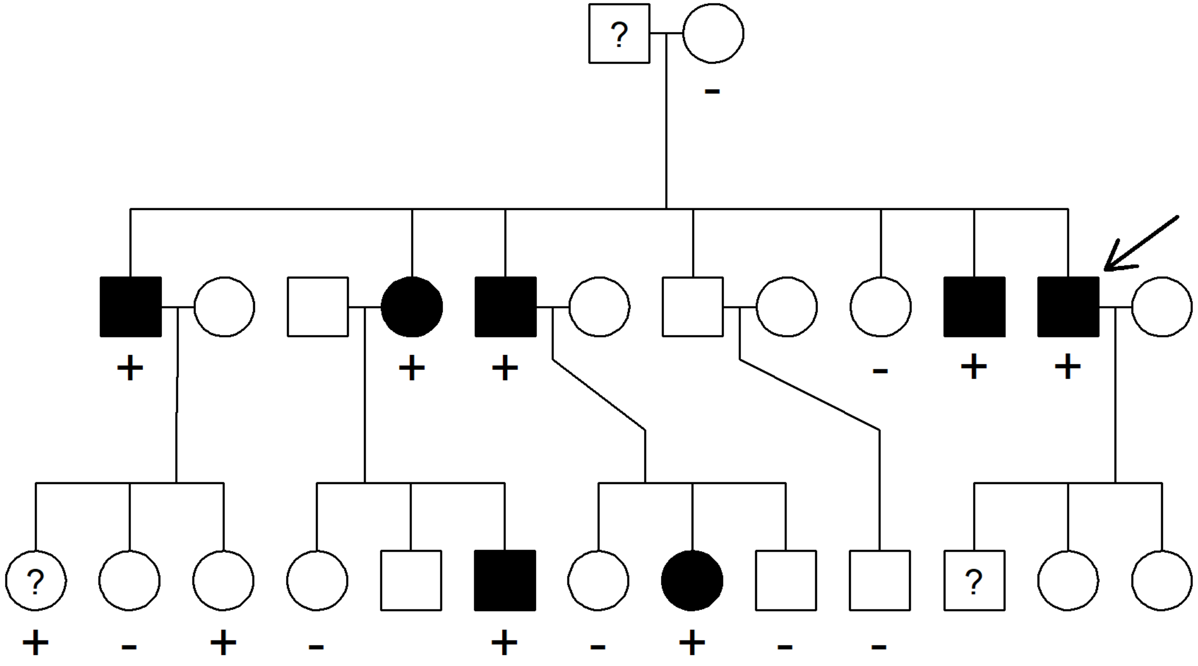


Not enough to qualify as
supportive evidence

```
➢ library(segregatr)

➢ x = nuclearPed(nch = 4, sex = c(1,2,1,1))

➢ FLB(x,
      aff = c(3, 5, 6),
      unknown = c(1, 2),
      proband = 3,
      carrier = c(3, 5, 6),
      noncarrier = 4,
      freq = 0.00001,
      penetrances = c(0.01, 0.9, 0.9))

[1] 7.107237
```

segregatr: Segregation analysis

Example from research



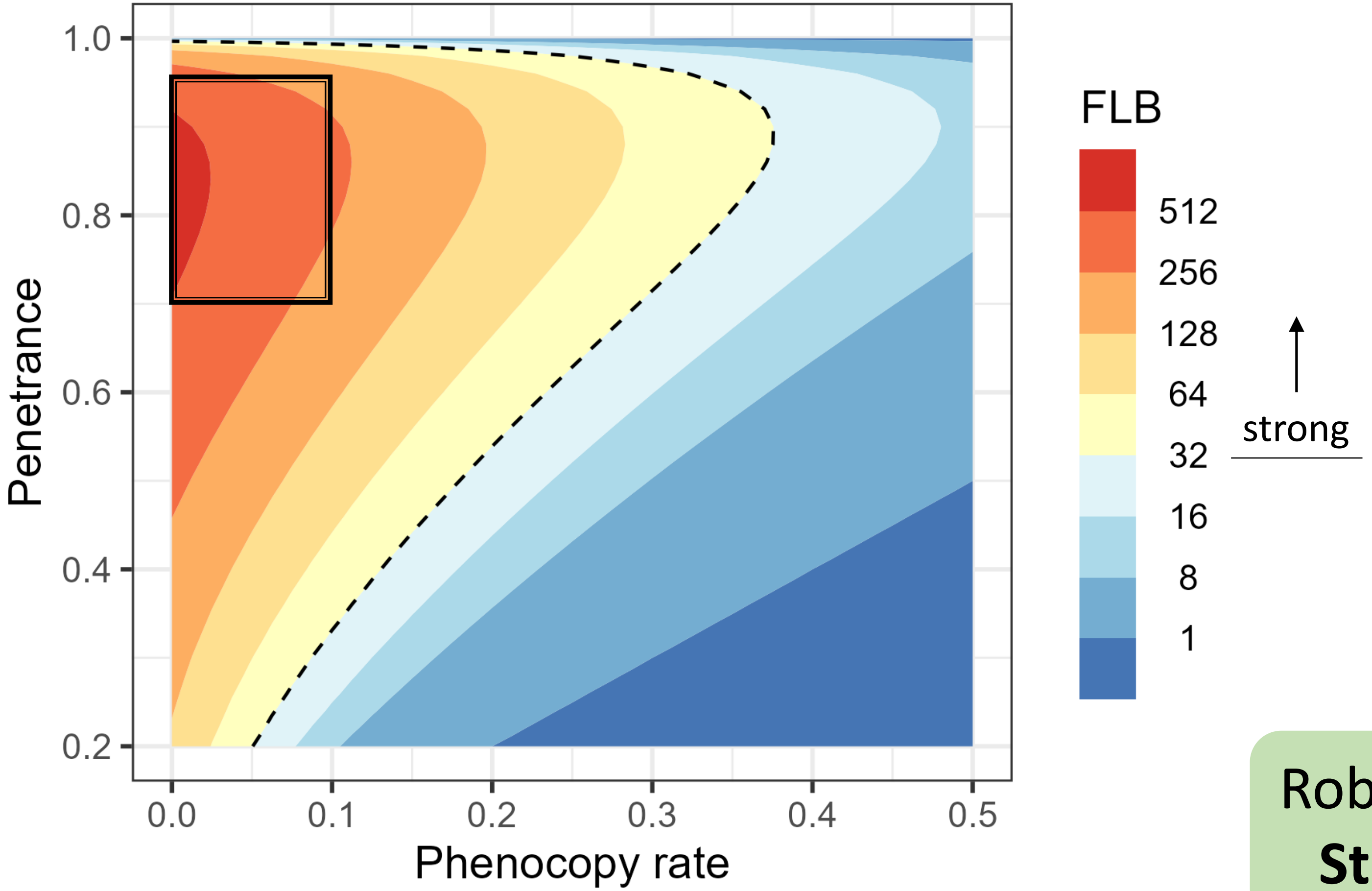
- Heritable thoracic aortic aneurysms with dissections (HTAAD)

- Potential cause: *SMAD3*:c.XXXG>A
- Classified as VUS

- Challenges:
  - Reduced penetrance
  - Phenocopies

Case report: *The use of segregation analysis in interpretation of sequence variants in SMAD3*
A Ratajska, MD Vigeland, KV Wirgenes, K Krohg-Sørensen, B Paus

Contours of FLB



Robust conclusion: **Strong** evidence

Resulted in genetic diagnosis

Now try the exercises!