# IBD Mapping for the Discovery of Novel Susceptibility Loci in Large Disease Cohorts

CS262 03-03-2015

Daniel Newburger
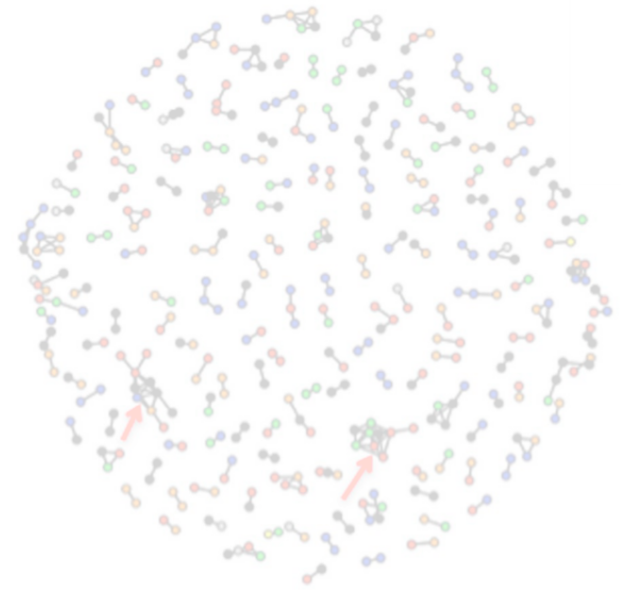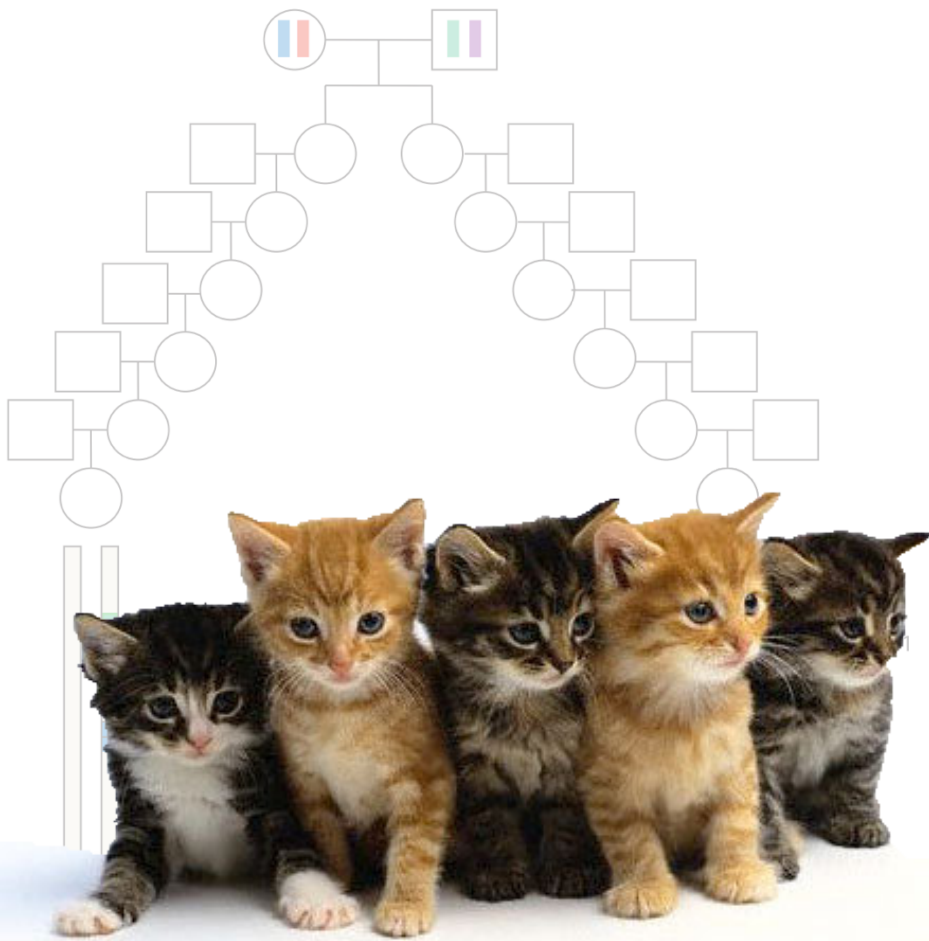
# The burden of complex disease

## Genetics
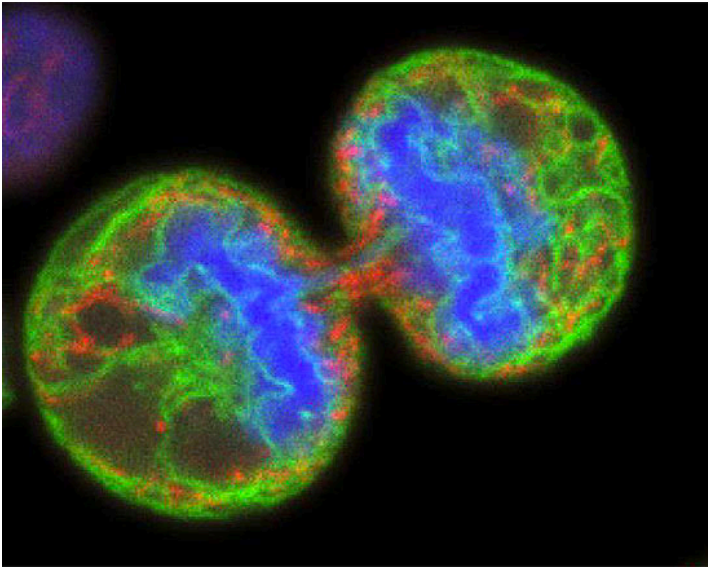


## Environment



**Complex Condition**: A condition caused by the interaction of multiple genes and environmental factors. Examples of complex conditions, which are also called multifactorial diseases, are cancer and heart disease.

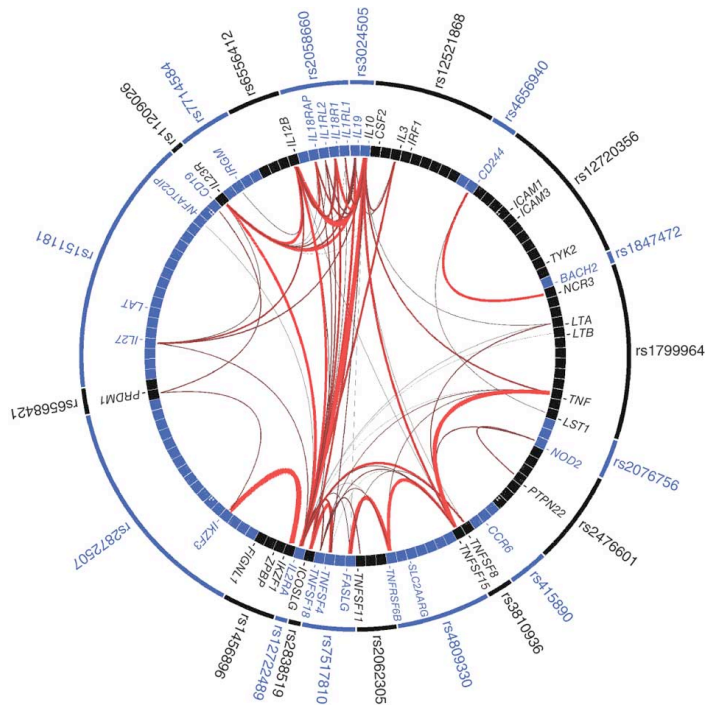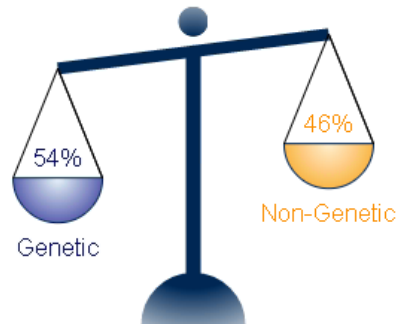**Heritability**: The proportion of interindividual differences (variance) in a trait that is the result of genetic factors
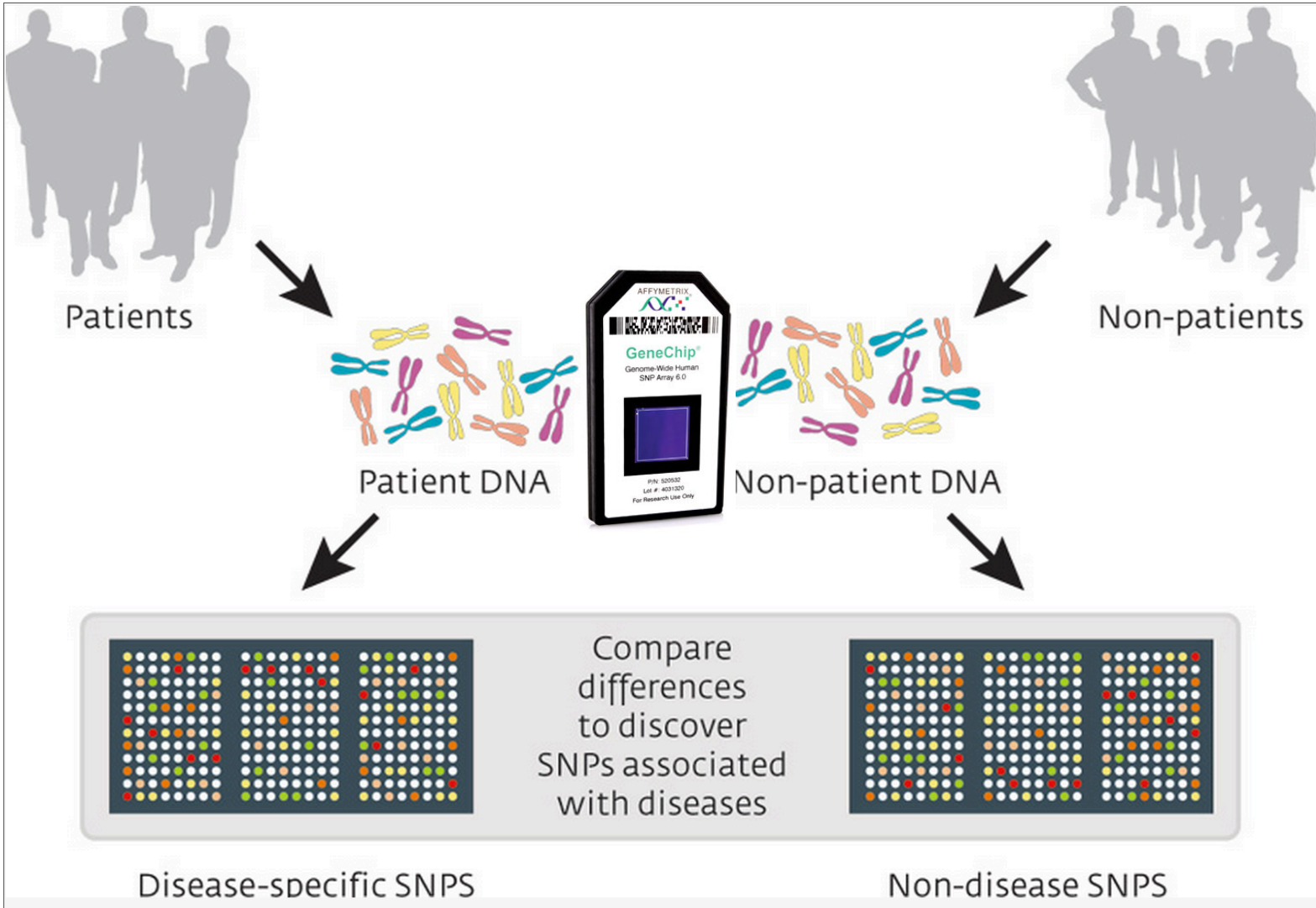
Definitions from Manolio, NEJM 2010

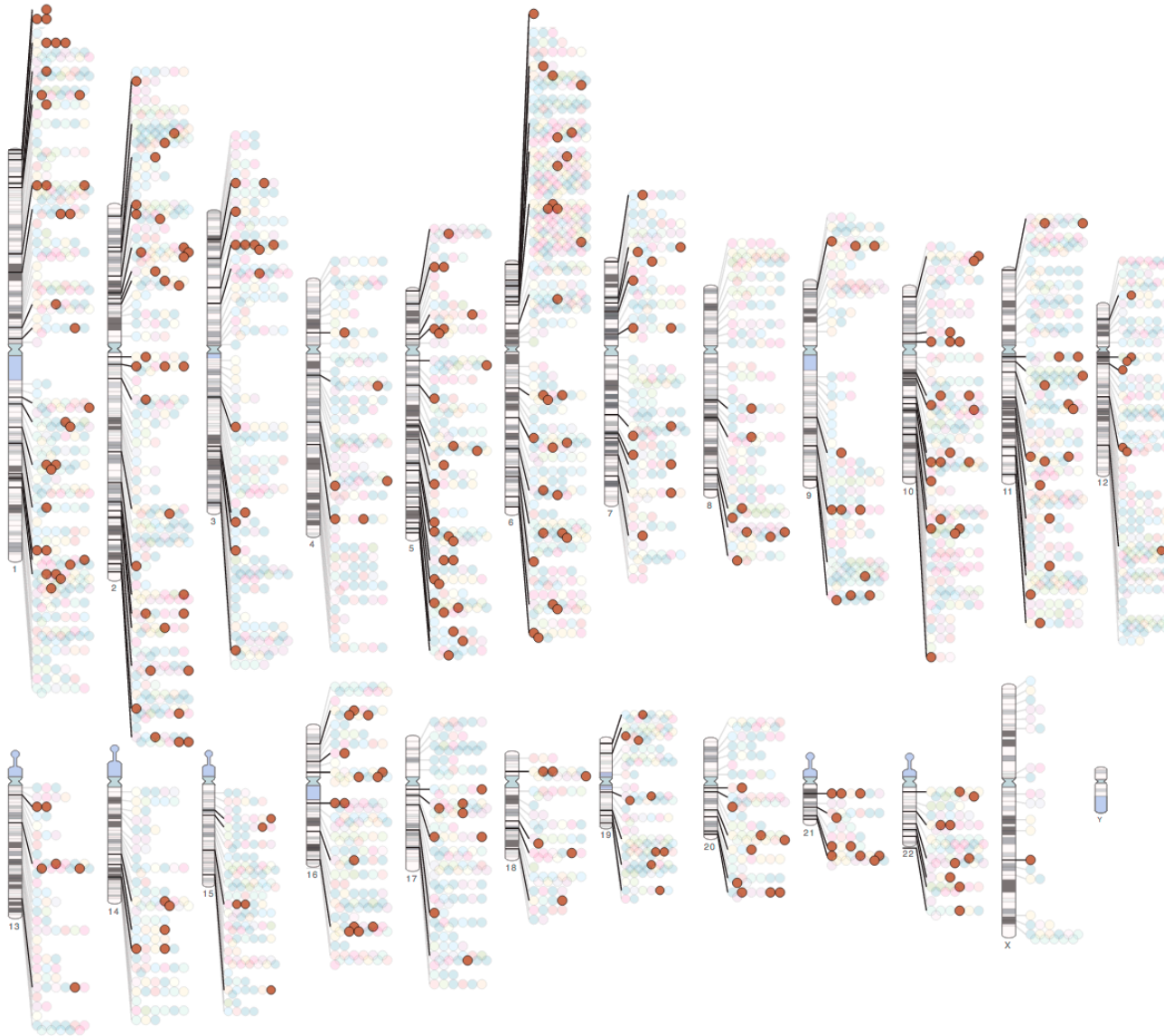# The burden of complex disease – Crohn's Disease
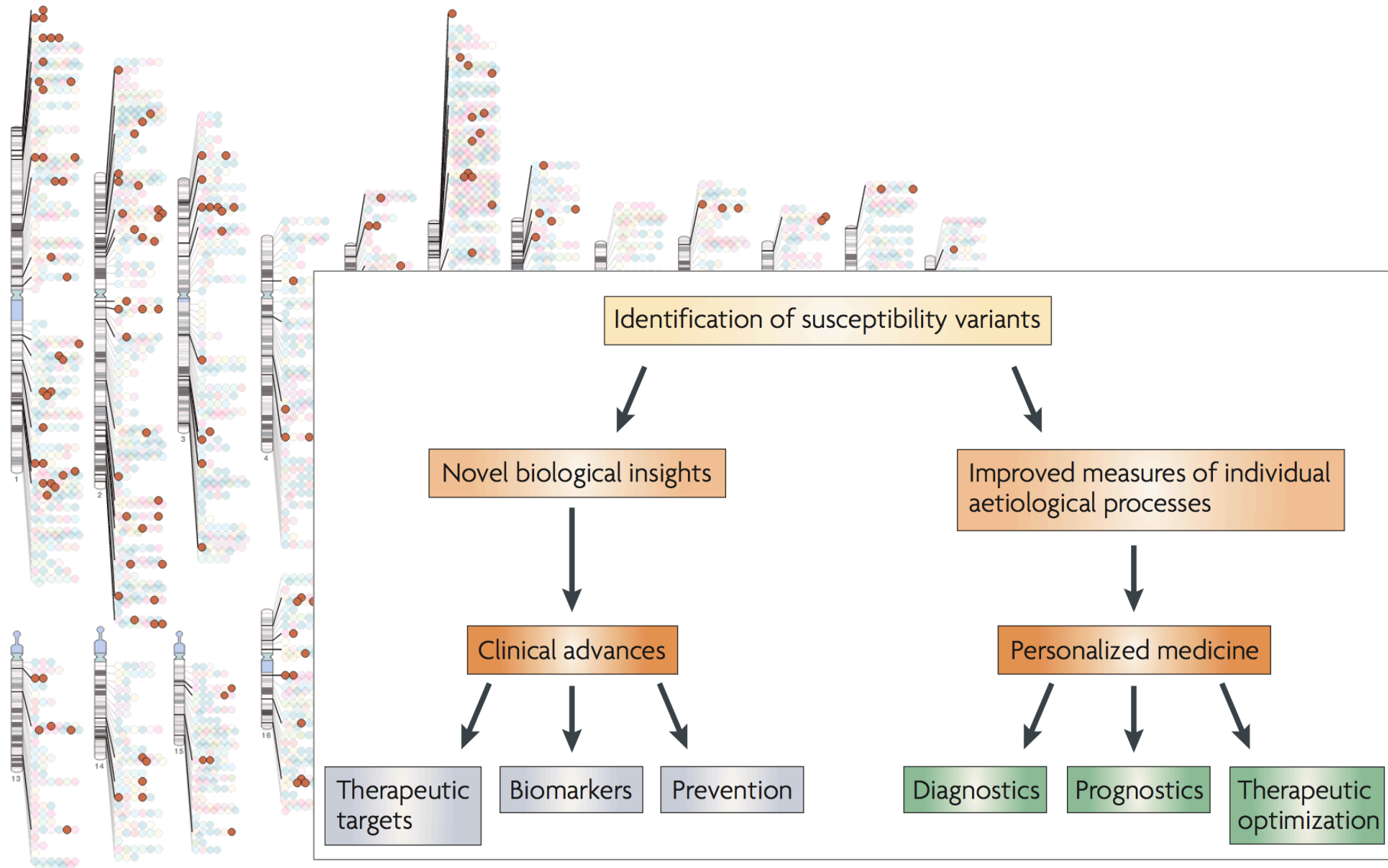
Genetics

Environment

# Association studies provide a first step towards understanding and treating complex disease

# Association studies provide a first step towards understanding and treating complex disease
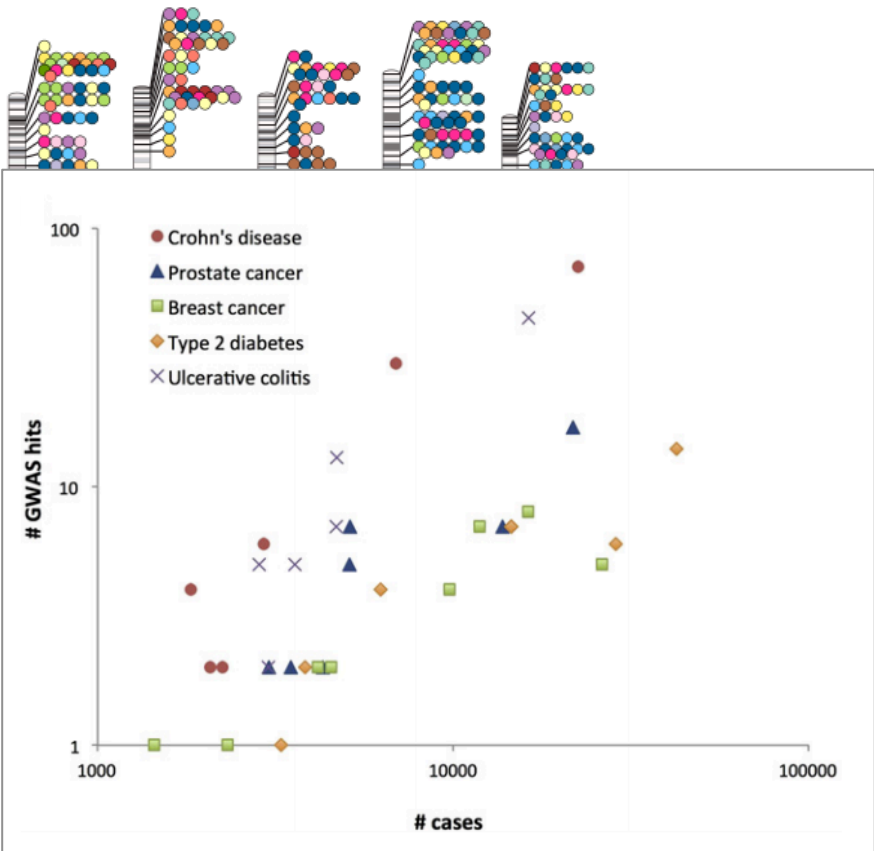
# Association studies provide a first step towards understanding and treating complex disease



Identification of susceptibility variants

Novel biological insights → Clinical advances → Therapeutic targets / Biomarkers / Prevention

Improved measures of individual aetiological processes → Personalized medicine → Diagnostics / Prognostics / Therapeutic optimization

# Genomic landscape of complex diseases



As of 2012, more than 140 studies with 10,000 or more individuals

Legend:
- Crohn's disease
- Prostate cancer
- Breast cancer
- Type 2 diabetes
- Ulcerative colitis

Axis labels: # GWAS hits (vertical), # cases (horizontal)

Visscher et al., AJHG 2012

# Missing heritability
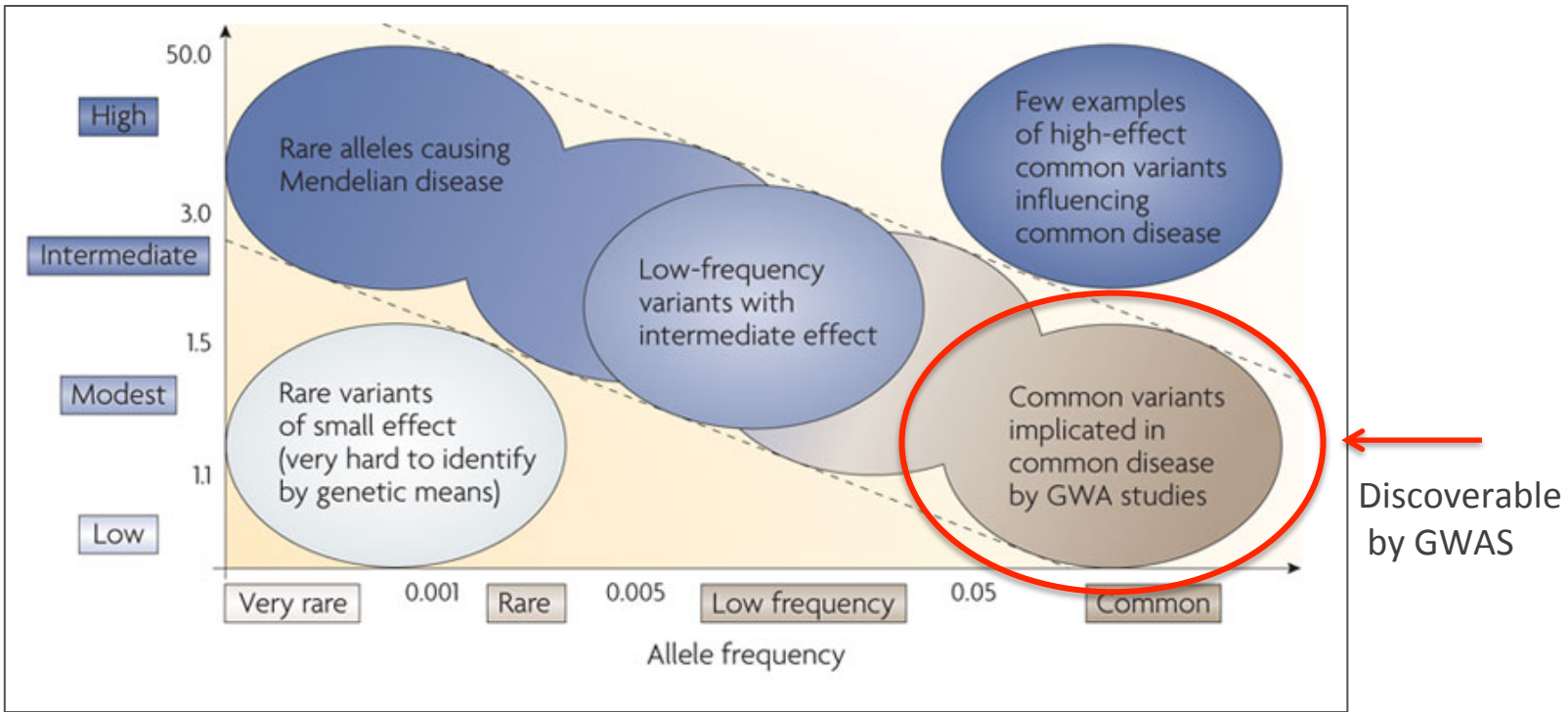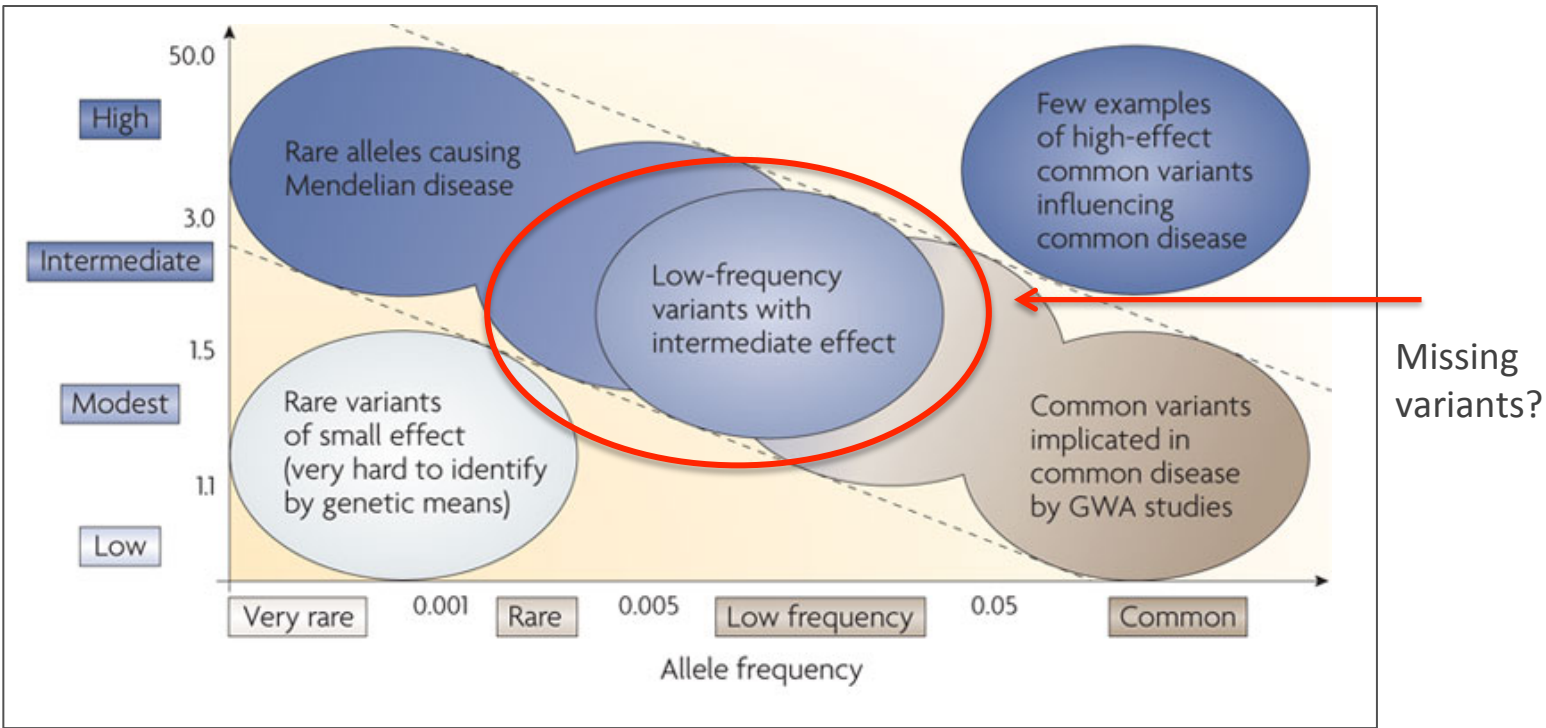
**Missing heritability and strategies for finding the underlying causes of complex disease**

Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore and Joseph H. Nadeau

**Finding the missing heritability of complex diseases**

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

**Table 1 | Estimates of heritability and number of loci for several complex traits**

| Disease | Number of loci | Proportion of heritability explained | Heritability measure |
|---|---|---|---|
| Age-related macular degeneration[72] | 5 | 50% | Sibling recurrence risk |
| Crohn's disease[21] | 32 | 20% | Genetic risk (liability) |
| Systemic lupus erythematosus[73] | 6 | 15% | Sibling recurrence risk |
| Type 2 diabetes[74] | 18 | 6% | Sibling recurrence risk |
| HDL cholesterol[75] | 7 | 5.2% | Residual* phenotypic variance |
| Height[15] | 40 | 5% | Phenotypic variance |
| Early onset myocardial infarction[76] | 9 | 2.8% | Phenotypic variance |
| Fasting glucose[77] | 4 | 1.5% | Phenotypic variance |

* Residual is after adjustment for age, gender, diabetes.

# Missing heritability

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]



Discoverable by GWAS

Manolio et al., Nature 2009

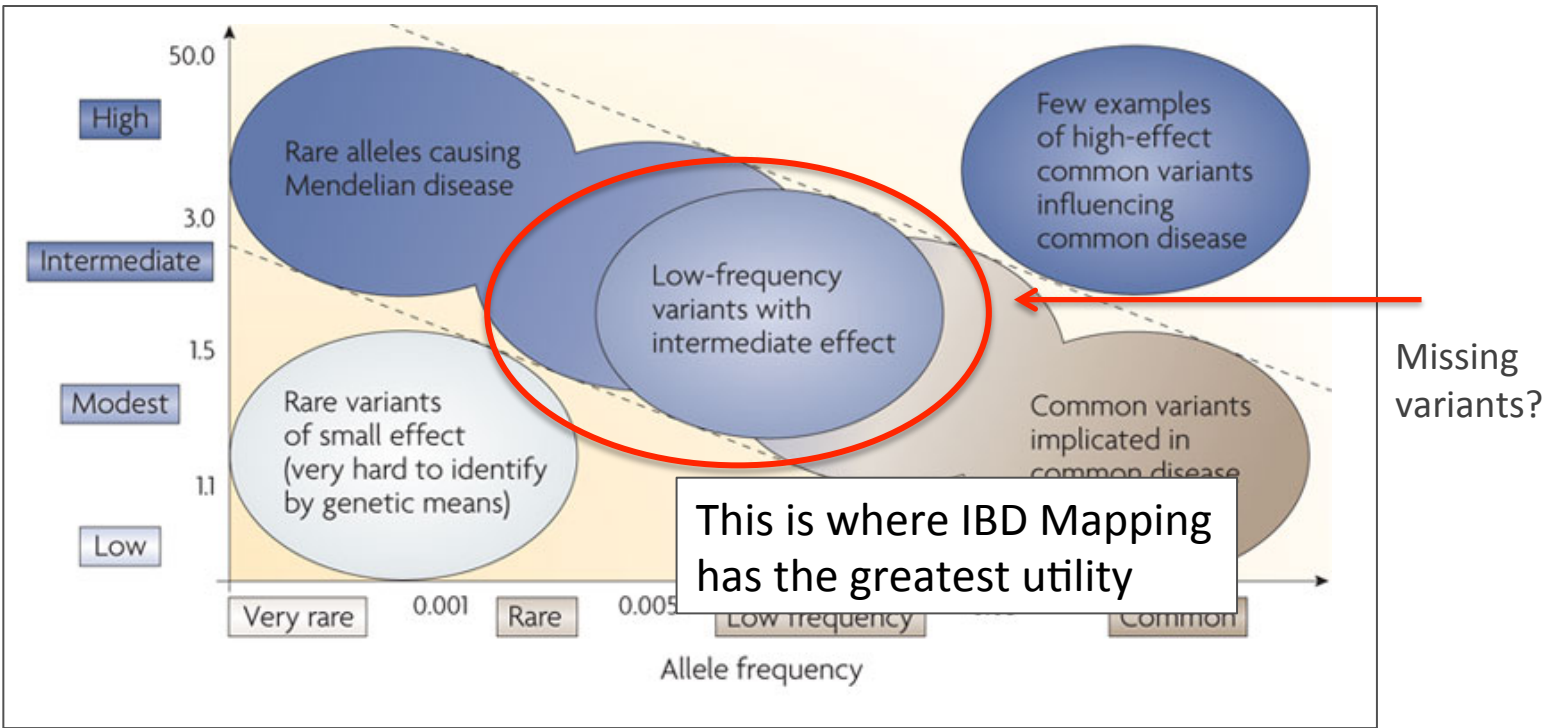# Missing heritability



Manolio et al., Nature 2009

# Missing heritability



**Missing heritability and strategies for finding the underlying causes of complex disease**

*Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore and Joseph H. Nadeau*

**Finding the missing heritability of complex diseases**

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Missing variants?

This is where IBD Mapping has the greatest utility

Manolio et al., Nature 2009

# How do association mapping studies work?

Case (i.e. patients)

Control (i.e. healthy individuals)

SNP X
G/A

G

Time

A

present

Population Size

AFFYMETRIX

GeneChip®
Genome-Wide Human
SNP Array 6.0

P/N: 520532
Lot #: 4031320
For Research Use Only

Genotype 1M
loci simultaneously

Standard GWAS Significance Testing

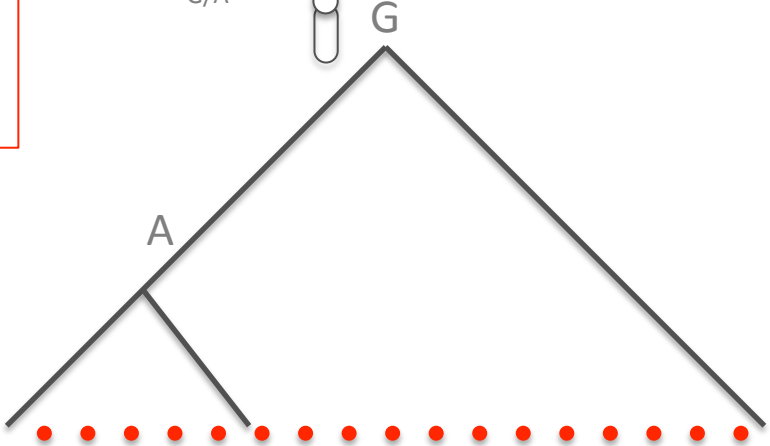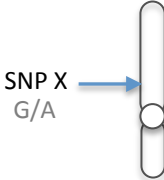# How do association mapping studies work?



Case (i.e. patients)

Control (i.e. healthy individuals)

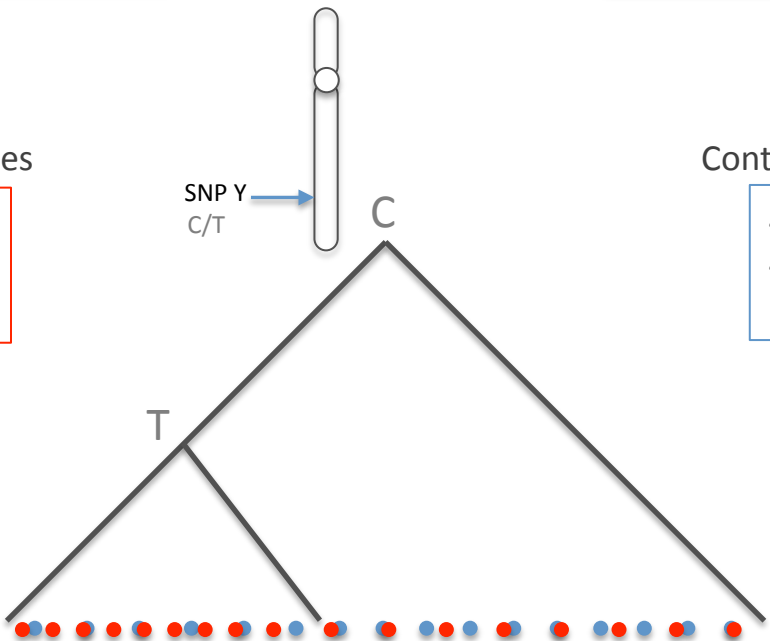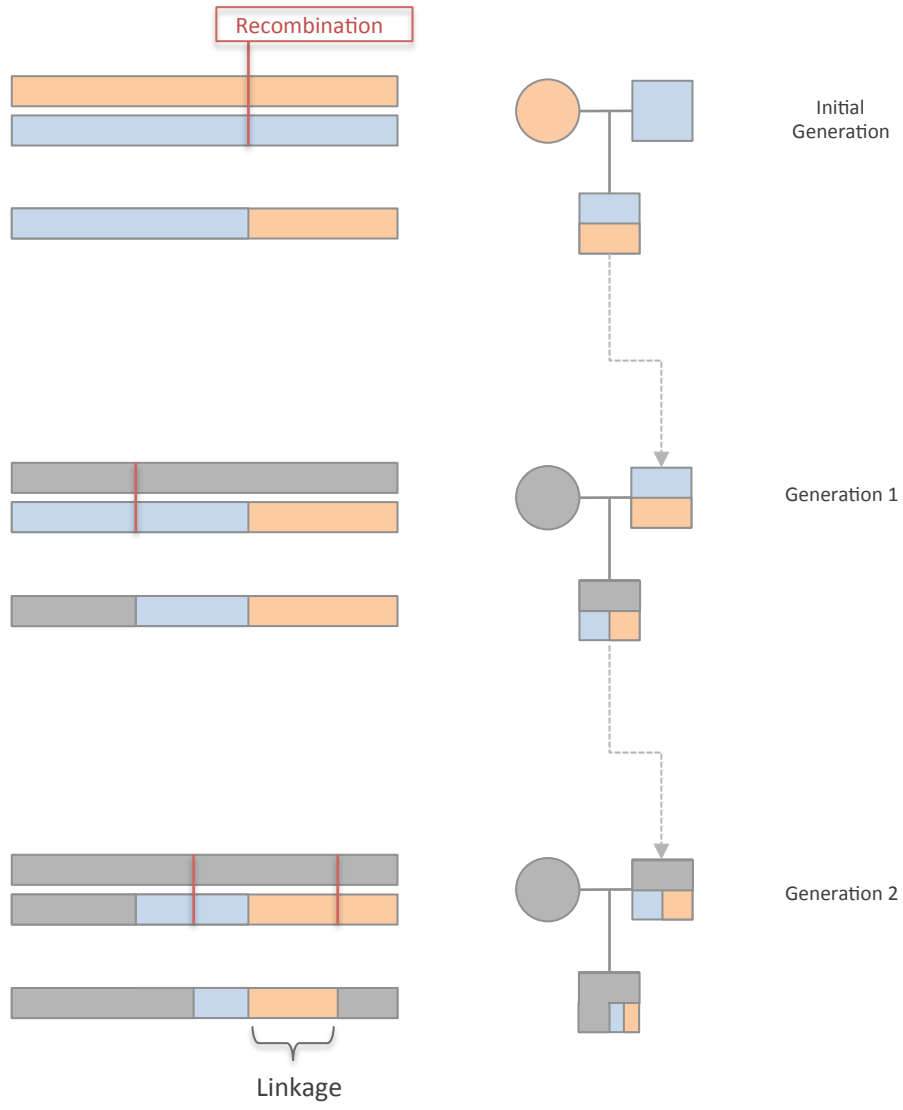Genotype 1M loci simultaneously

Case Alleles

64 GG
32 GA
4 AA

SNP X
G/A

G

A

Allele does not confer disease risk

# How do association mapping studies work?

Case (i.e. patients)

Control (i.e. healthy individuals)

Genotype 1M
loci simultaneously

Case Alleles

64 GG
32 GA
4 AA

SNP X
G/A

Control Alleles

63 GG
32 GA
5 AA

G

A
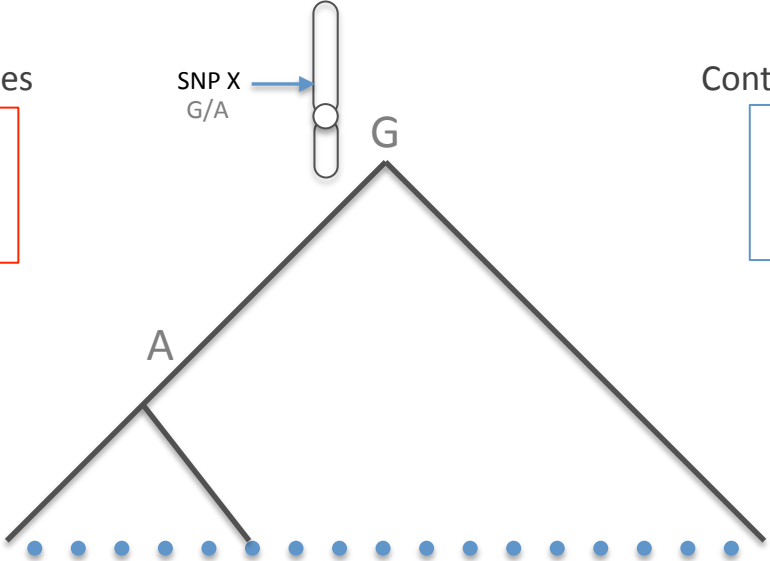
Allele does not confer disease risk

Test for significance

| 64 | 32 | 4 |
|----|----|---|
| 63 | 32 | 5 |

Chi-square p ≈ 0.90

Standard GWAS Significance Testing

AFFYMETRIX

GeneChip®
Genome-Wide Human
SNP Array 6.0

P/N 520532
Lot # 4031320
For Research Use Only

# How do association mapping studies work?



Case (i.e. patients)

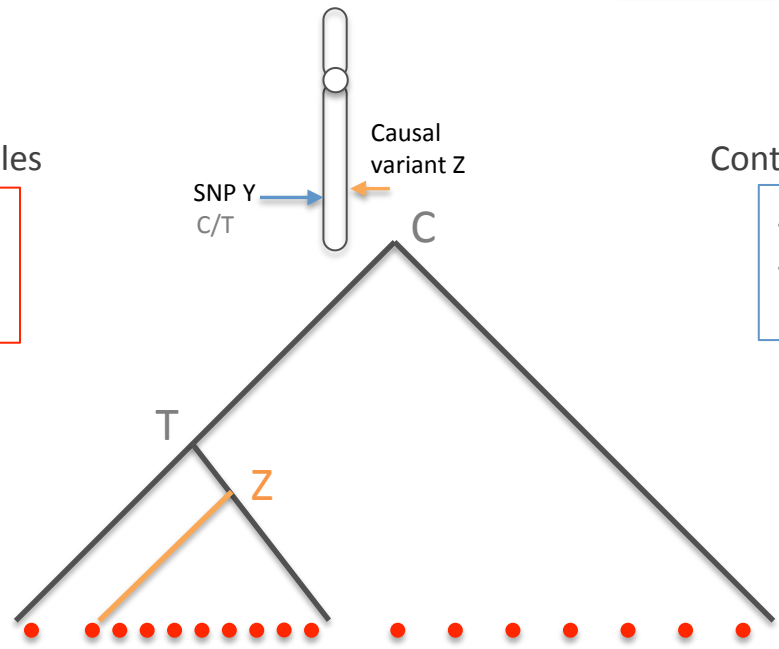Control (i.e. healthy individuals)

Genotype 1M
loci simultaneously

Case Alleles

36 CC
48 CT
16 TT

SNP Y
C/T

C

Control Alleles

49 CC
42 CT
9 TT

T

Allele _does_ confer disease risk

Test for significance

| 36 | 48 | 16 |
|----|----|----|
| 49 | 42 | 9  |

Chi-square p ≈ 0.01

Standard GWAS Significance Testing
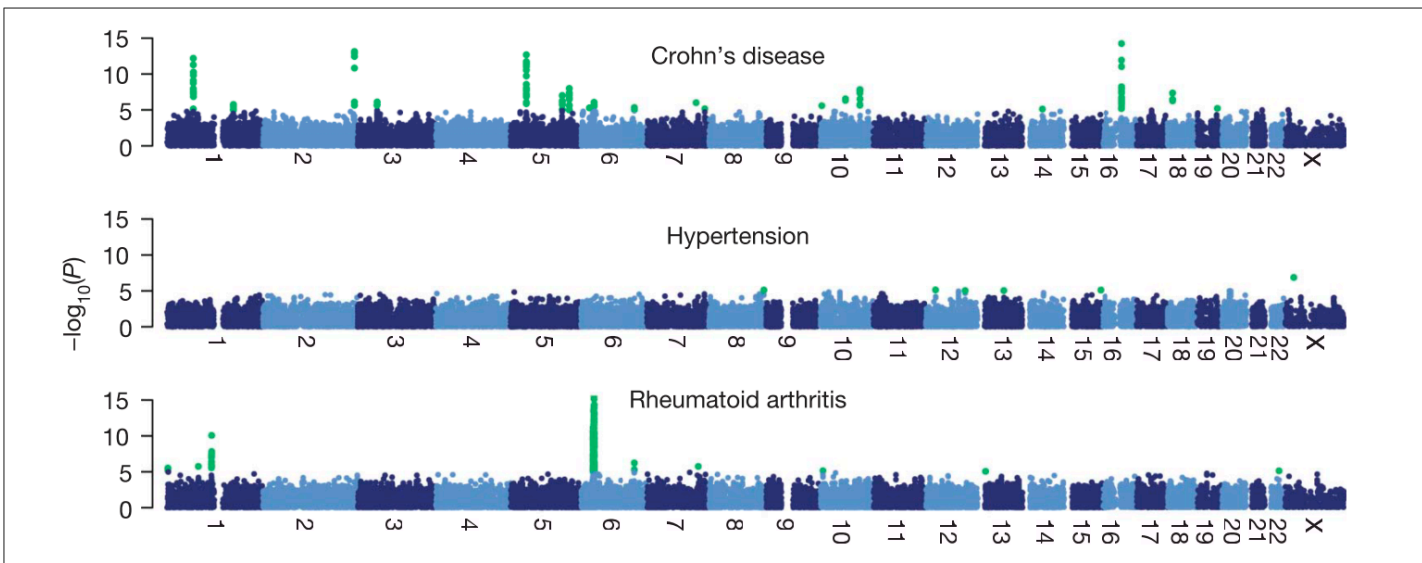
# How do association mapping studies work?

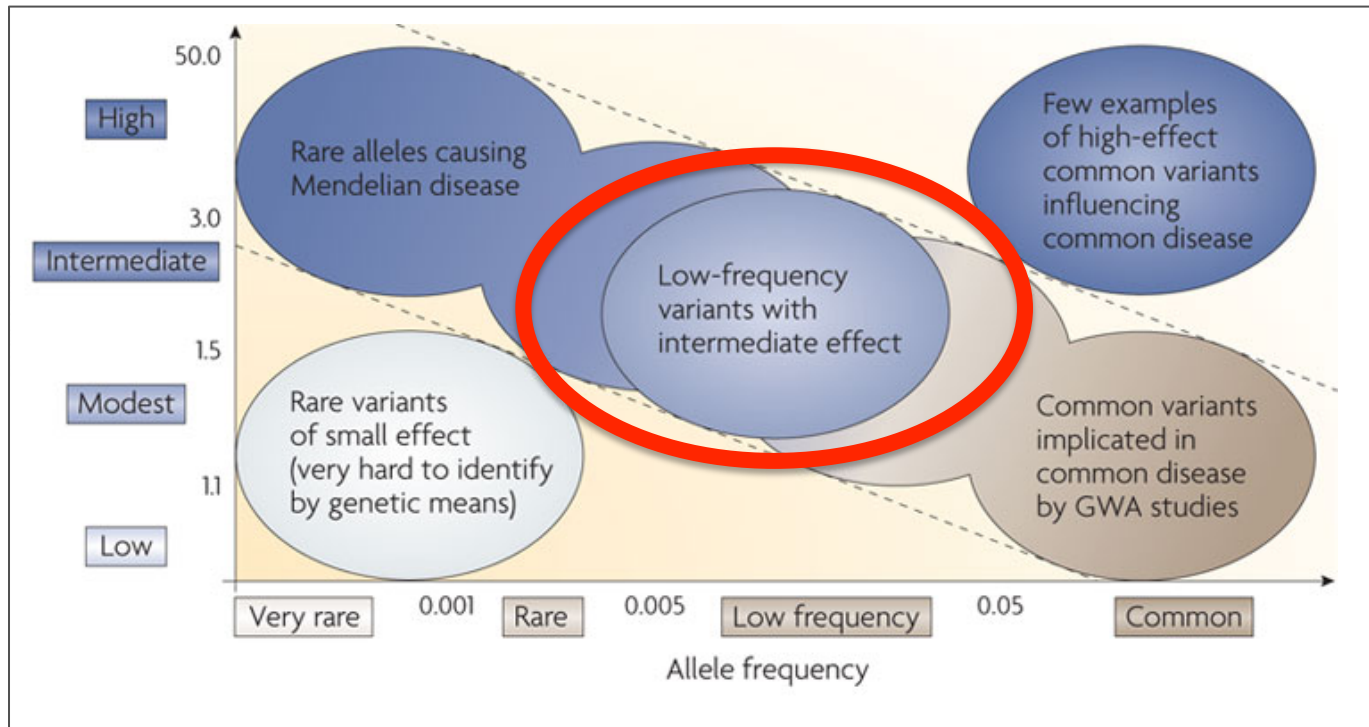Case (i.e. patients)

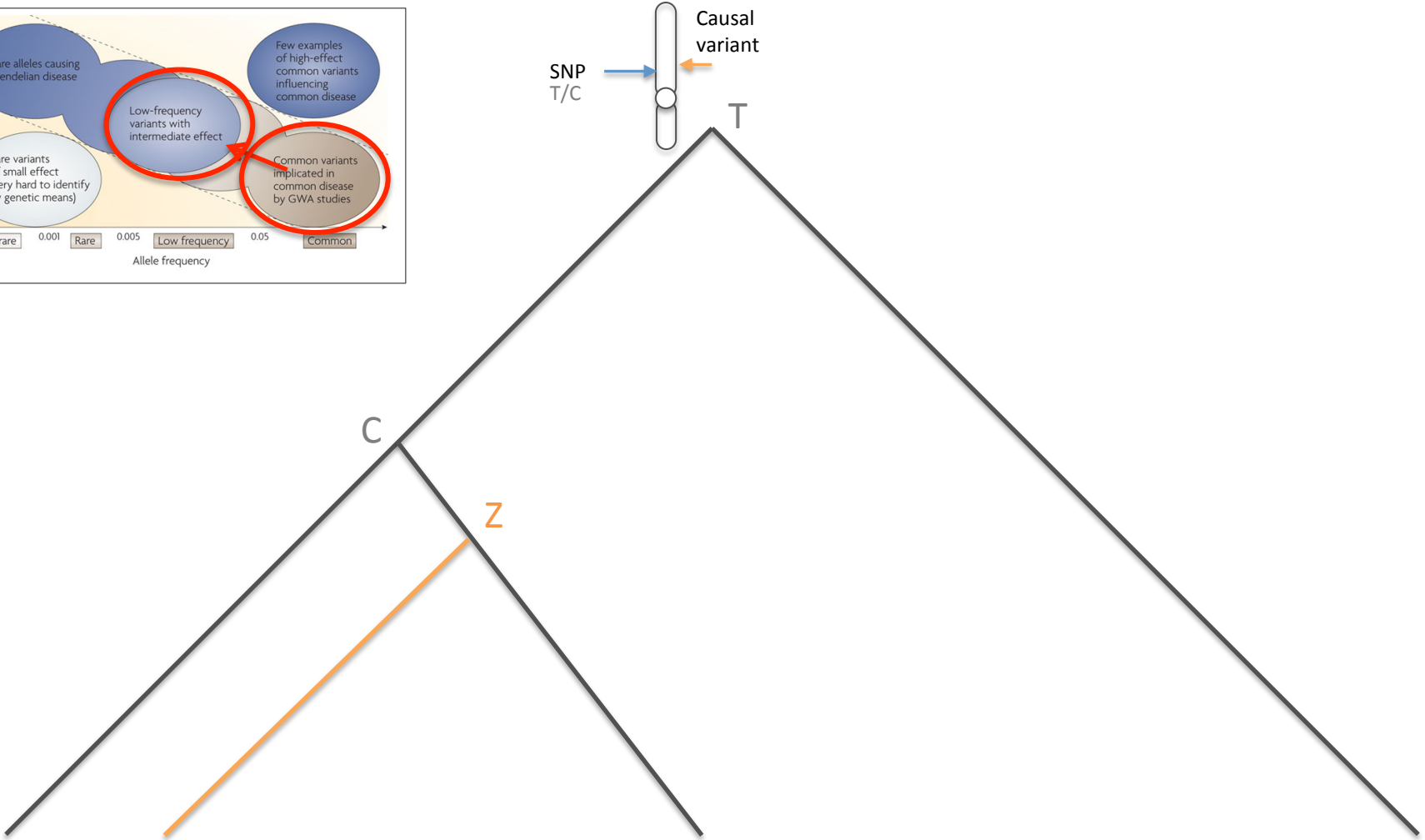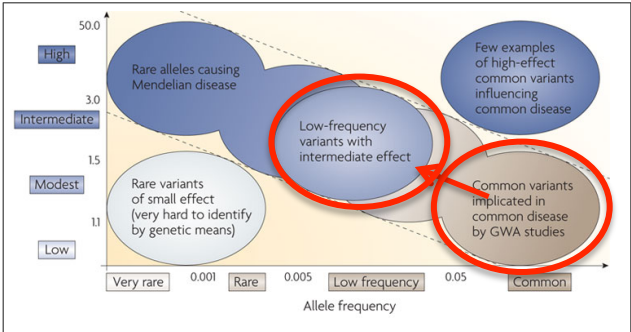Control (i.e. healthy individuals)



Genotype 1M
loci simultaneously
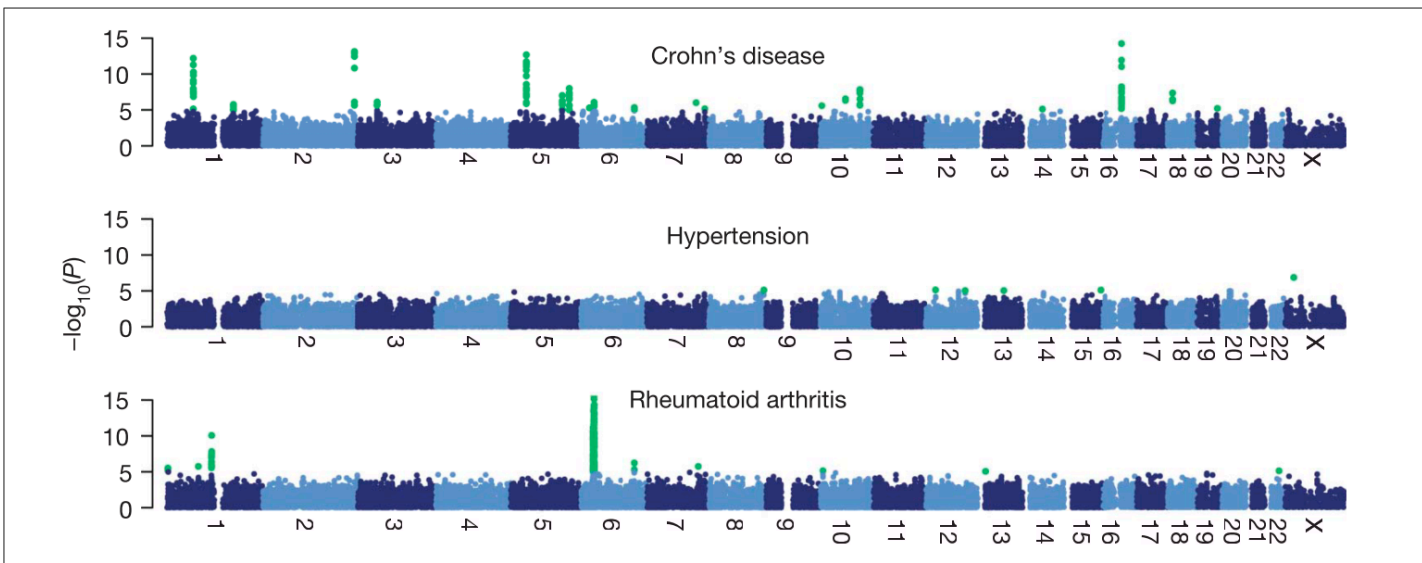


Welcome Trust Case Control Consortium, Nature 2007

Standard GWAS Significance Testing

# Linkage disequilibrium allows indirect associations



Initial
Generation

Generation 1

Generation 2

Linkage

# Linkage disequilibrium allows indirect associations



Adapted from Bush and Moore, PLoS Comp Bio 2012
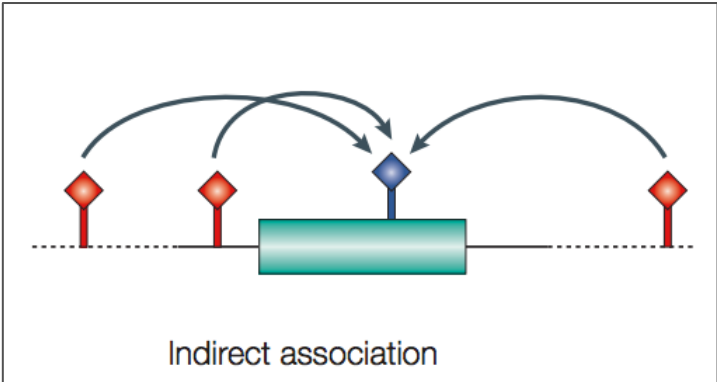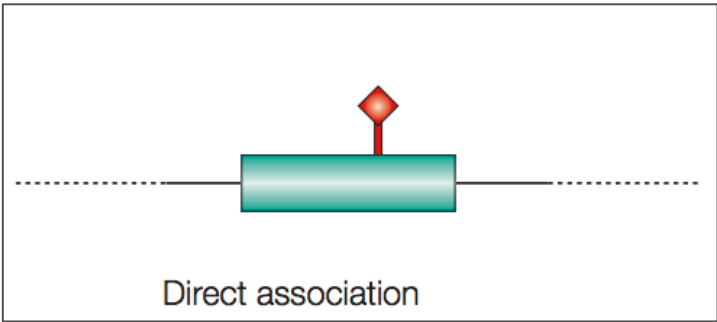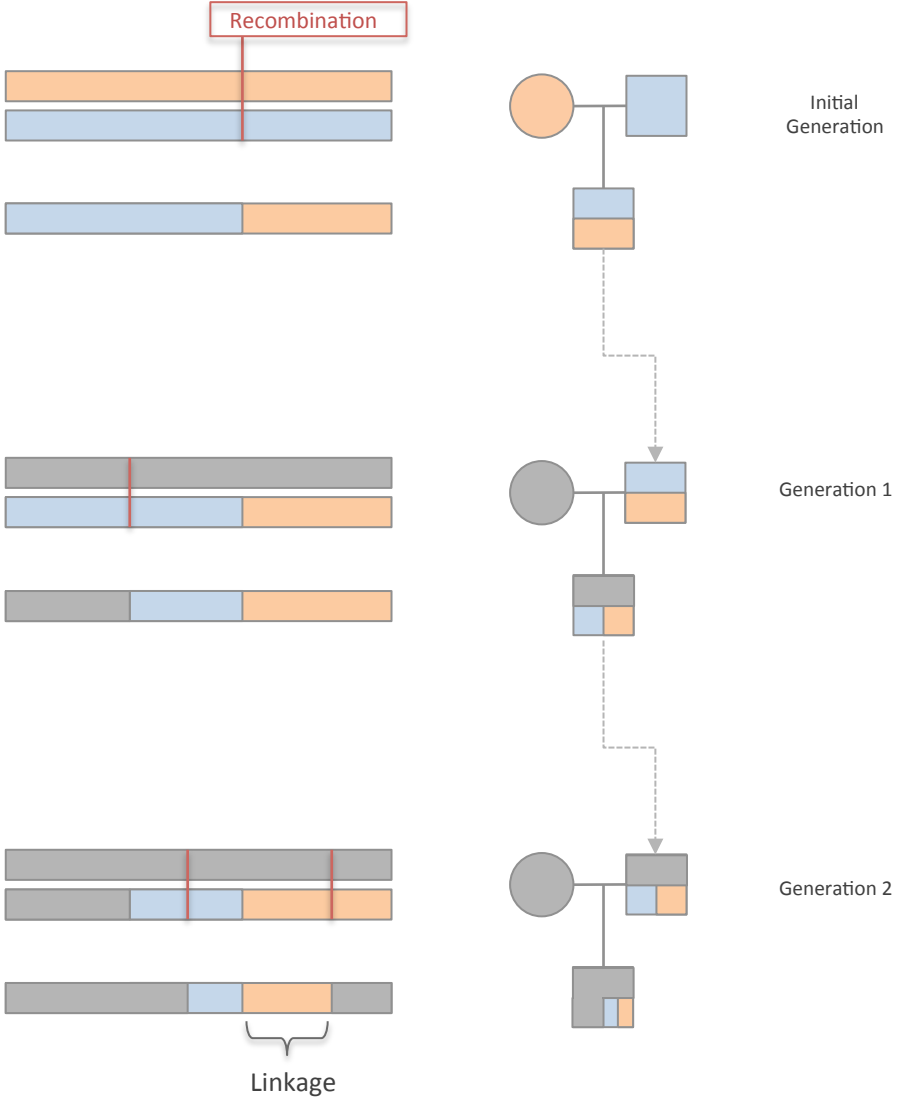
Hirschhorn and Daly, Nature Reviews Genetics 2005

# How do association mapping studies work?

Case (i.e. patients)

Control (i.e. healthy individuals)

Genotype 1M loci simultaneously

Case Alleles

36 CC
48 CT
16 TT

Causal variant Z

SNP Y
C/T

C

Control Alleles

49 CC
42 CT
9 TT

T

Z

Allele <u>does</u> confer disease risk

Test for significance

| 36 | 48 | 16 |
|----|----|----|
| 49 | 42 | 9  |

Chi-square p ≈ 0.01

Standard GWAS Significance Testing

# How do association mapping studies work?

Case (i.e. patients)

Control (i.e. healthy individuals)



Genotype 1M
loci simultaneously



Welcome Trust Case Control Consortium, Nature 2007

Standard GWAS Significance Testing

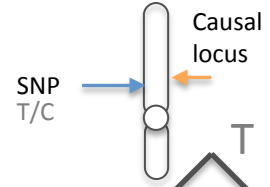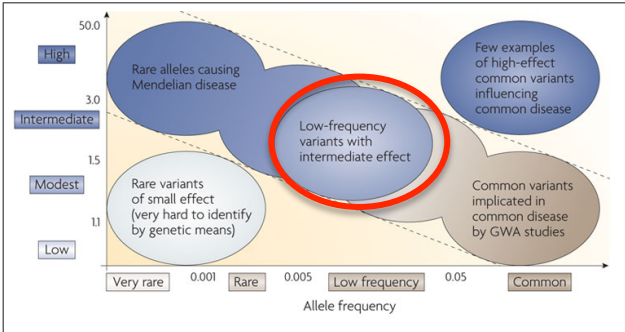# Why do we use identity by descent mapping?



- Traditional GWAS approaches no longer work in a scenario with recent (i.e. rare) causal variants.
- However, in a scenario where there are multiple rare variants of intermediate effect size at a genetic locus, we can use IBD mapping in place of the previously shown approaches.
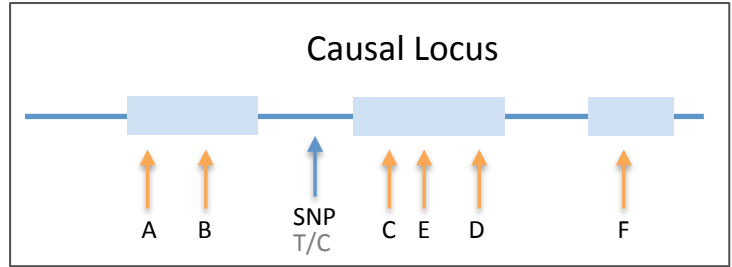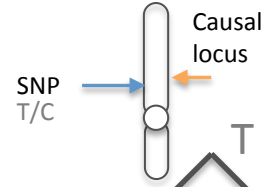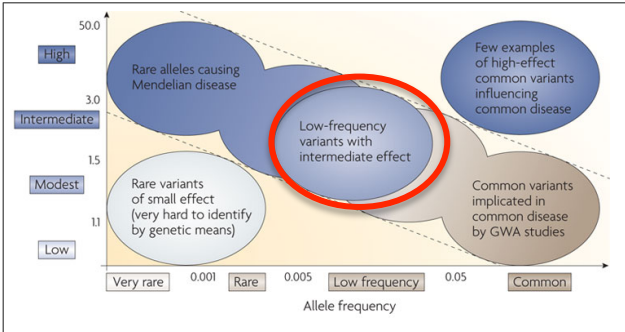
# Standard GWAS fails for rare variant diseases
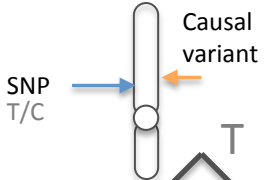
# Why do we use identity by descent mapping?



Causal locus

SNP
T/C

T

C

## Causal Locus

SNP
T/C

A  B  C  E  D  F

# Why do we use identity by descent mapping?



Causal locus

SNP T/C

Causal Locus

SNP T/C

A   B   C   E   D   F

T

C

Significance testing

|        | TT | CT | CC |
|--------|----|----|----|
| case   | 64 | 32 | 4  |
| control| 63 | 32 | 5  |

Chi-square p ≈ 0.90

A   C   F   B   D   E
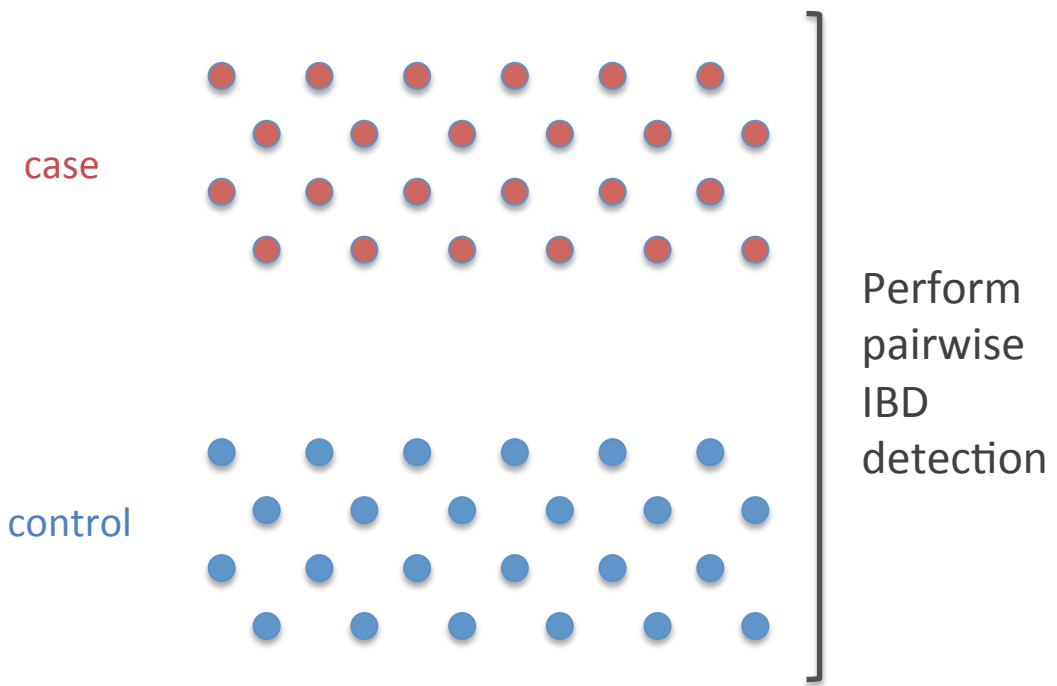
case
control
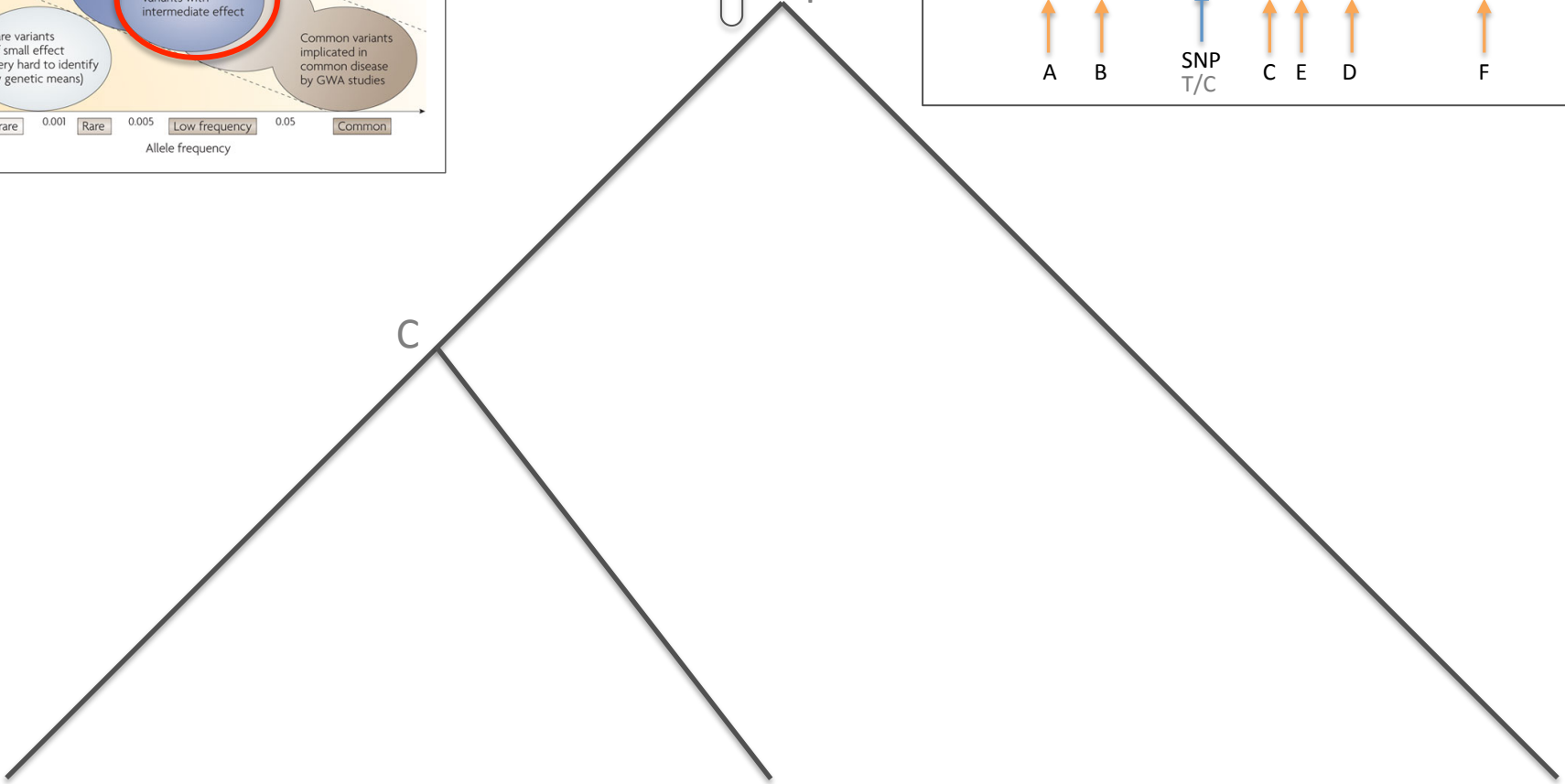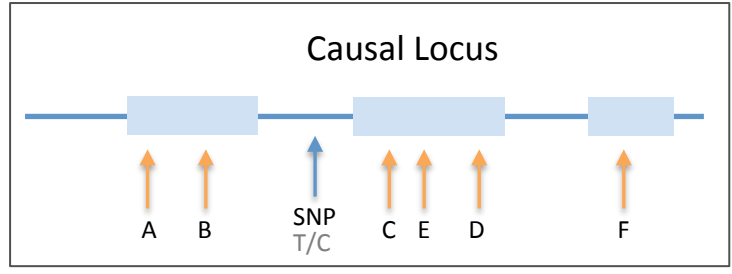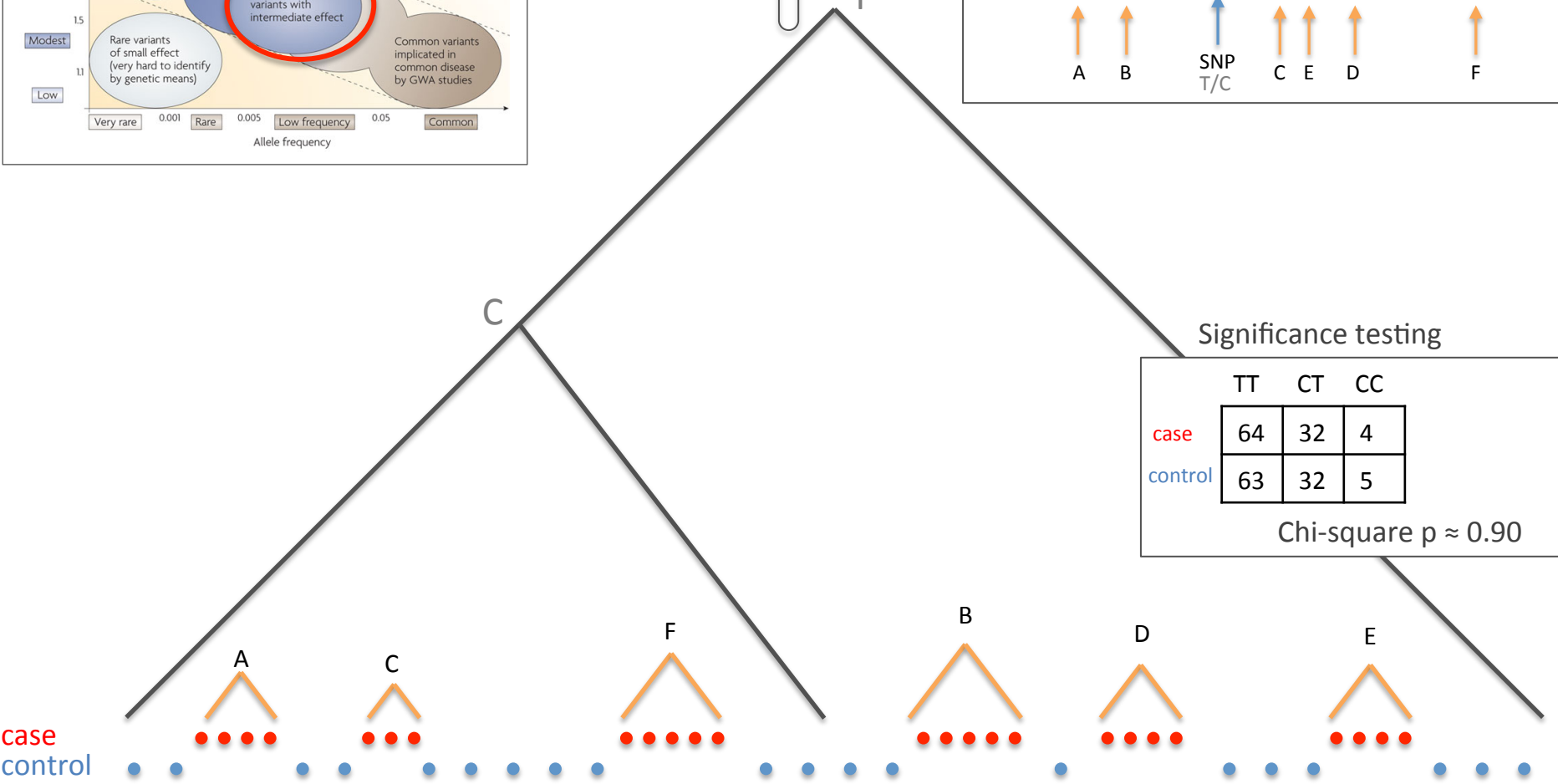
SNP frequency tests do not identity the indirect association

# What is identity by descent (IBD)?
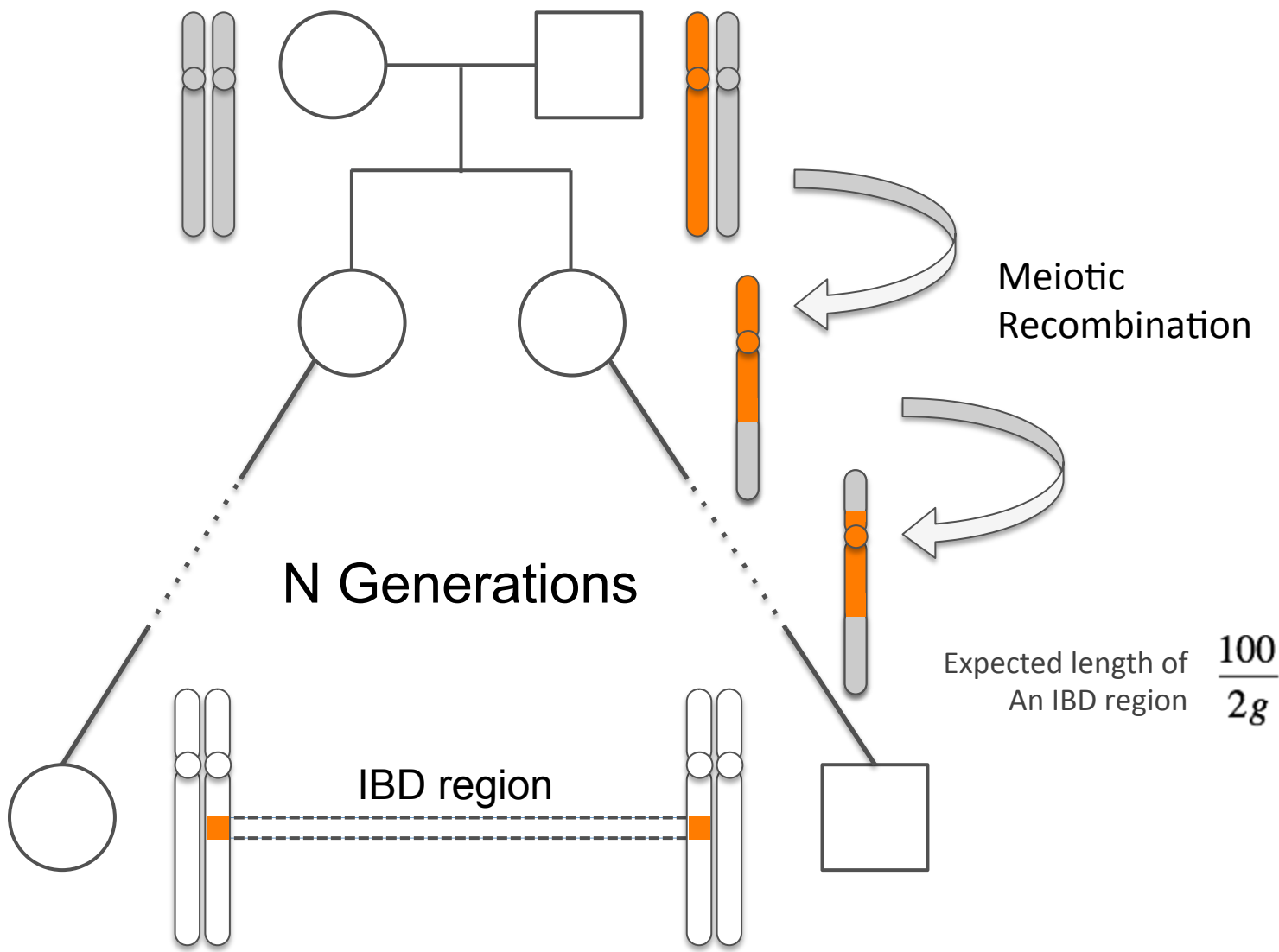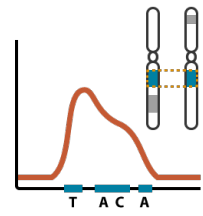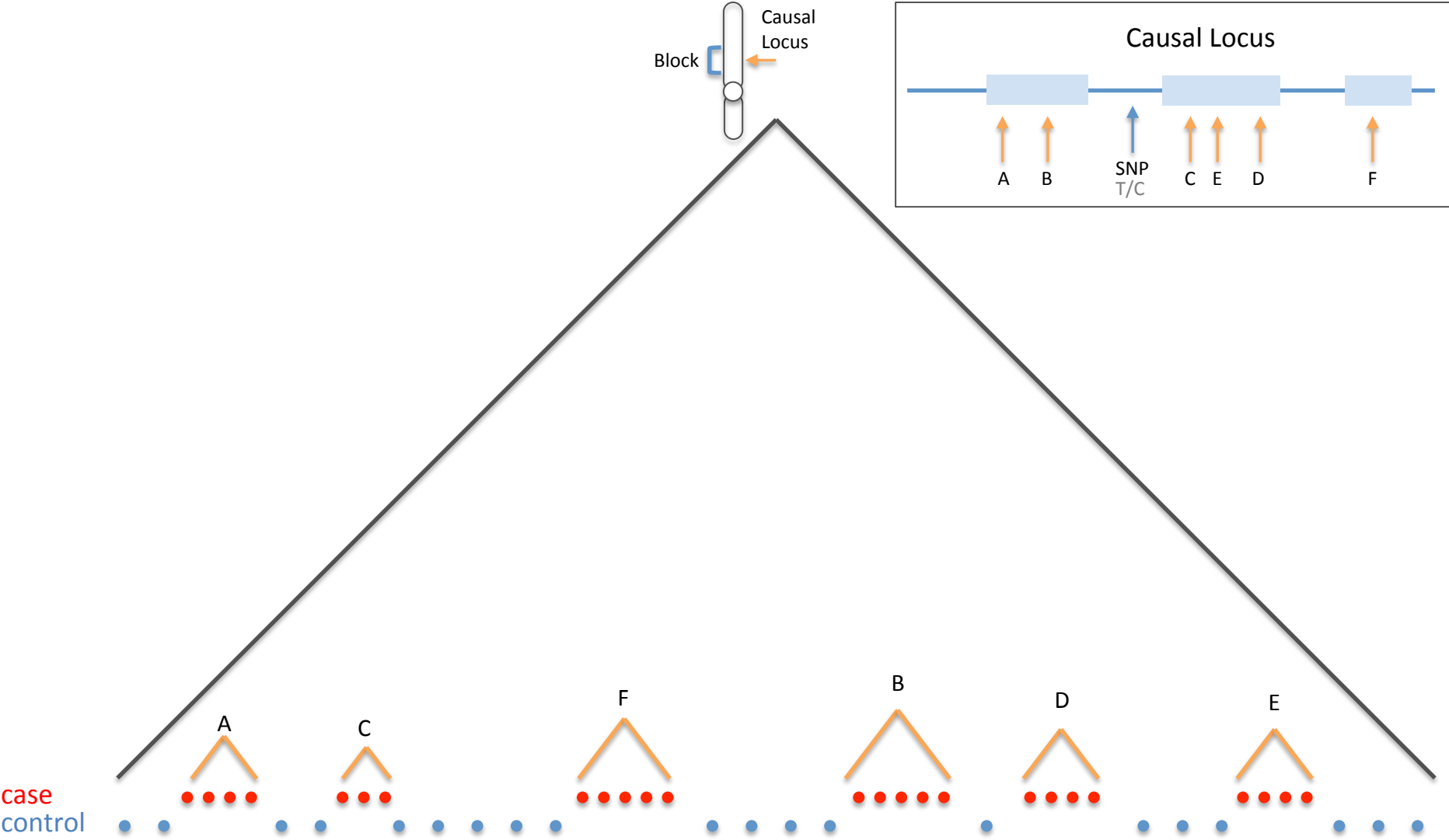


Meiotic Recombination

N Generations

IBD region

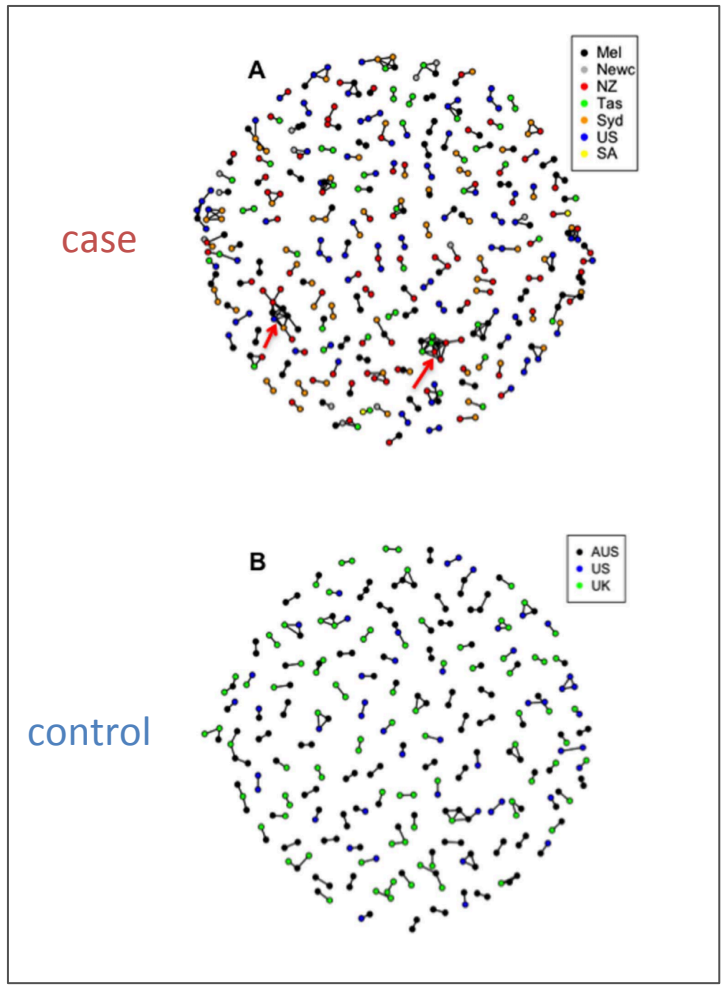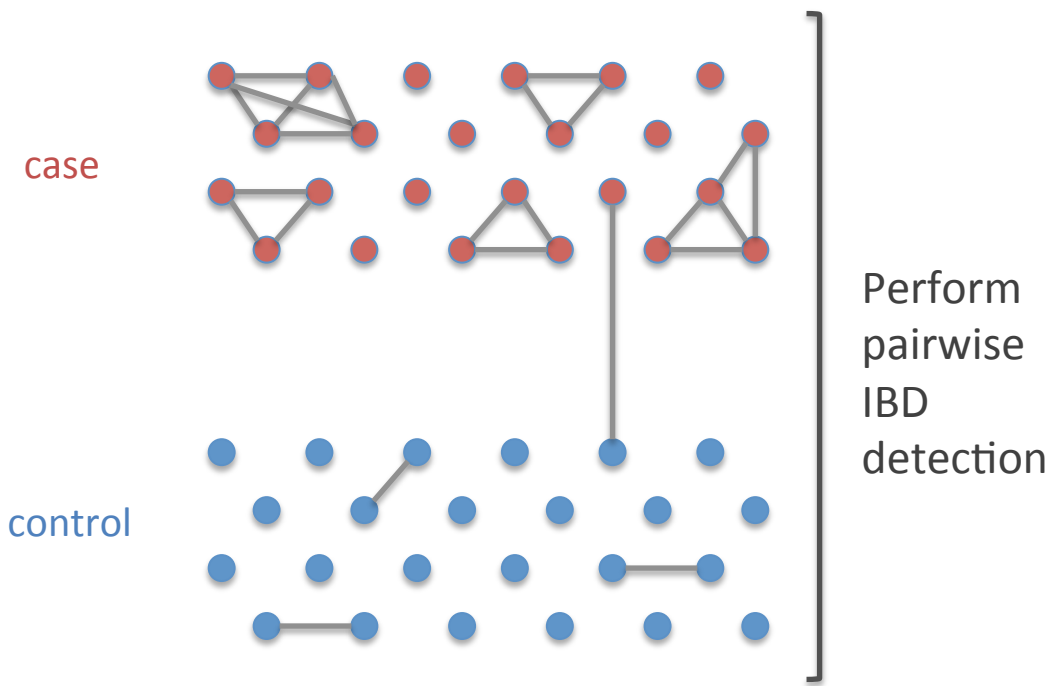Expected length of An IBD region $\dfrac{100}{2g}$

# Why do we use identity by descent mapping?



IBD Mapping

# Identity by descent mapping captures rare variant loci



Causal Locus

Block

case
control

case

control

Perform
pairwise
IBD
detection

IBD Mapping

# Identity by descent mapping captures rare variant loci



case

control

case

control

Perform
pairwise
IBD
detection

Lin et al., PLOS ONE 2013

IBD Mapping

# Identity by descent mapping captures rare variant loci



Block

Causal Loci

case
control

case

control
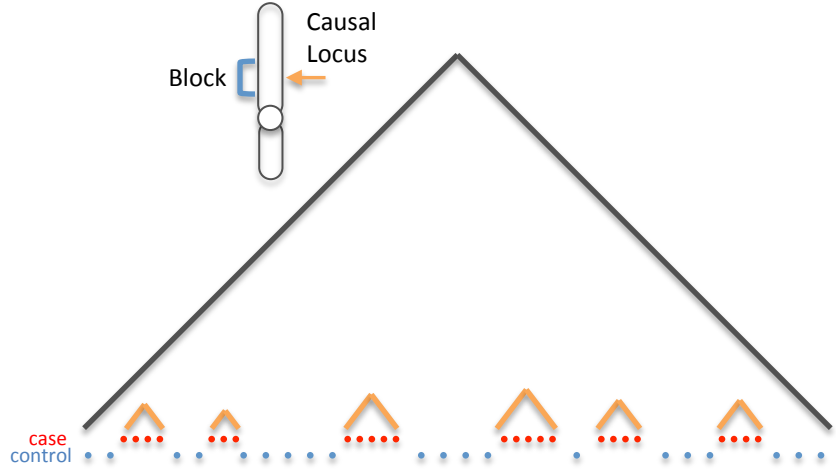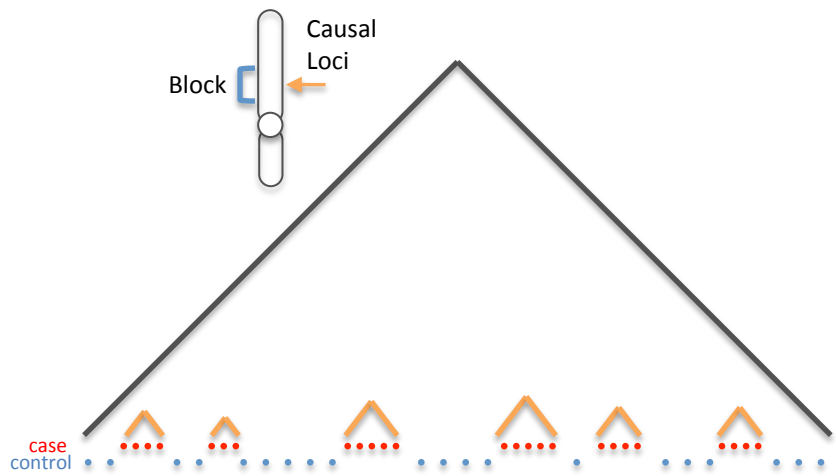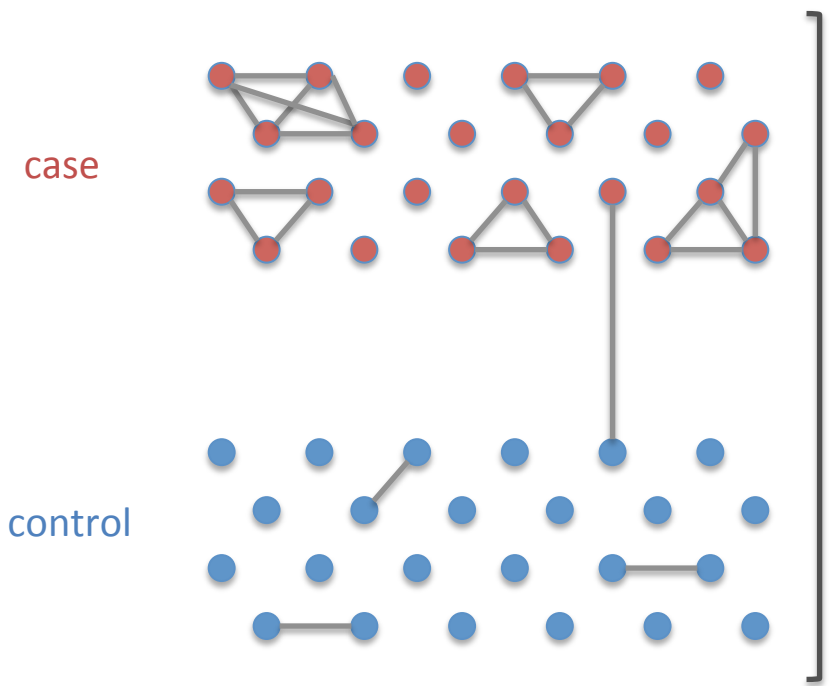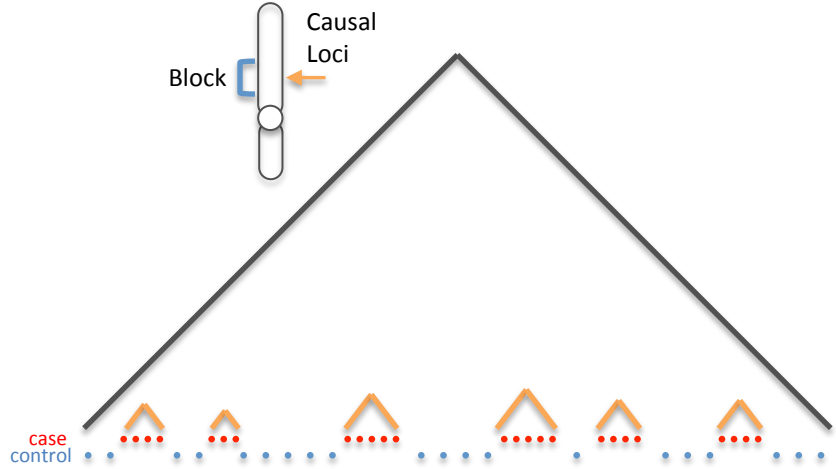
Perform pairwise IBD detection

Pairwise mapping statistic

$$\frac{\#\ \bullet\!\!-\!\!\bullet}{\left(\dfrac{\#\,\bullet}{2}\right)}\ -\ \frac{\#\ \bullet\!\!-\!\!\bullet}{\left(\dfrac{\#\,\bullet}{2}\right)}$$

IBD Mapping

# Identity by descent mapping captures rare variant loci



Block

Causal Loci

case
control

case

control

Perform pairwise IBD detection

-log P
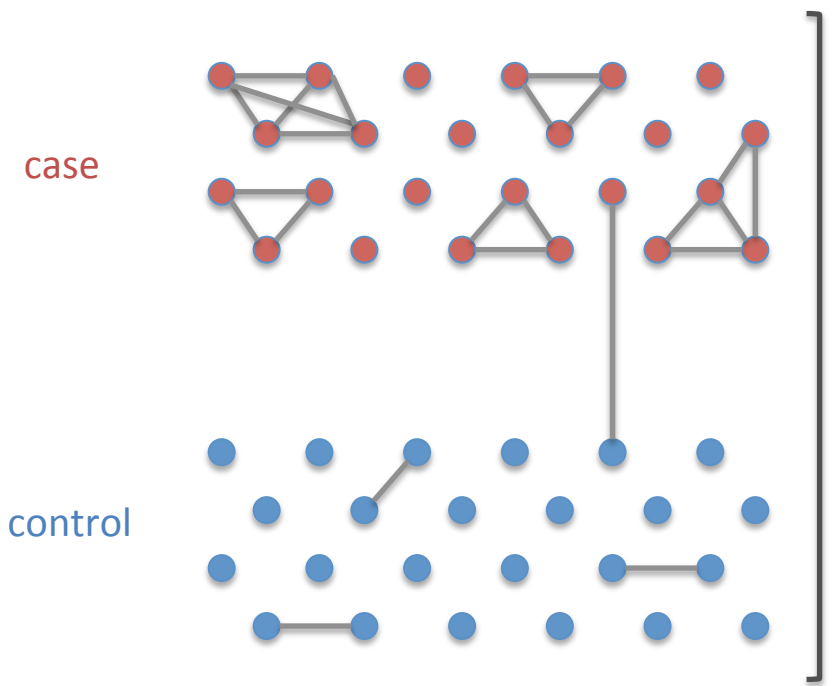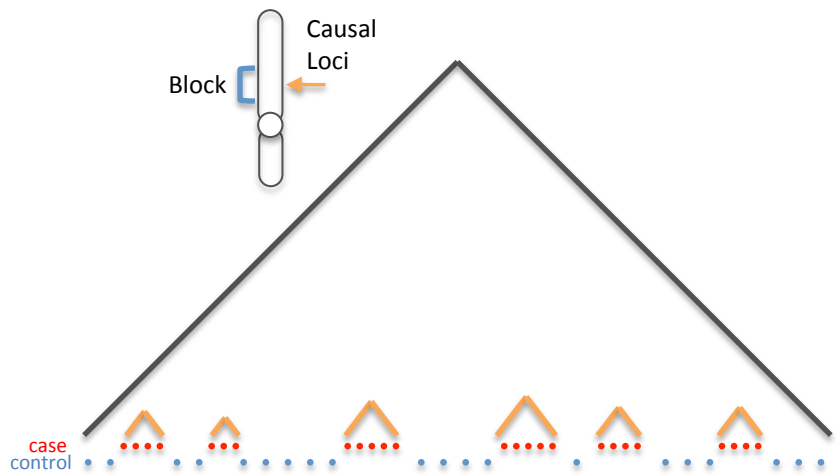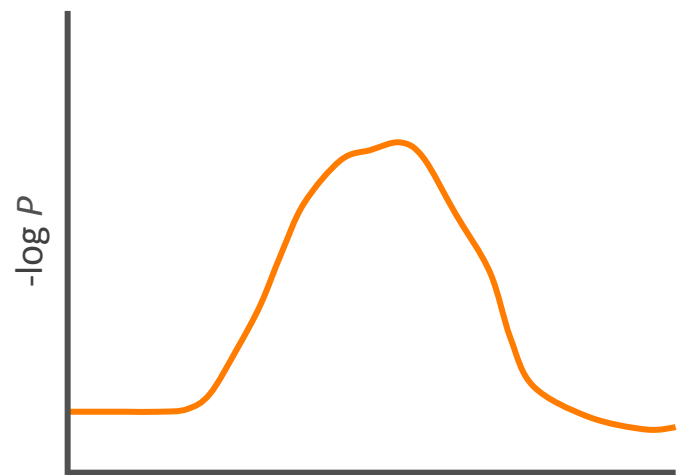
Causal Loci

Permutation testing

Pairwise mapping statistic

$$\frac{\# \bullet\!\!-\!\!\bullet}{\binom{\#\bullet}{2}} - \frac{\# \bullet\!\!-\!\!\bullet}{\binom{\#\bullet}{2}}$$

IBD Mapping

# Why doesn't everyone use IBD mapping?

## Accuracy



Fifth-cousins: Chromosome 1

IBD
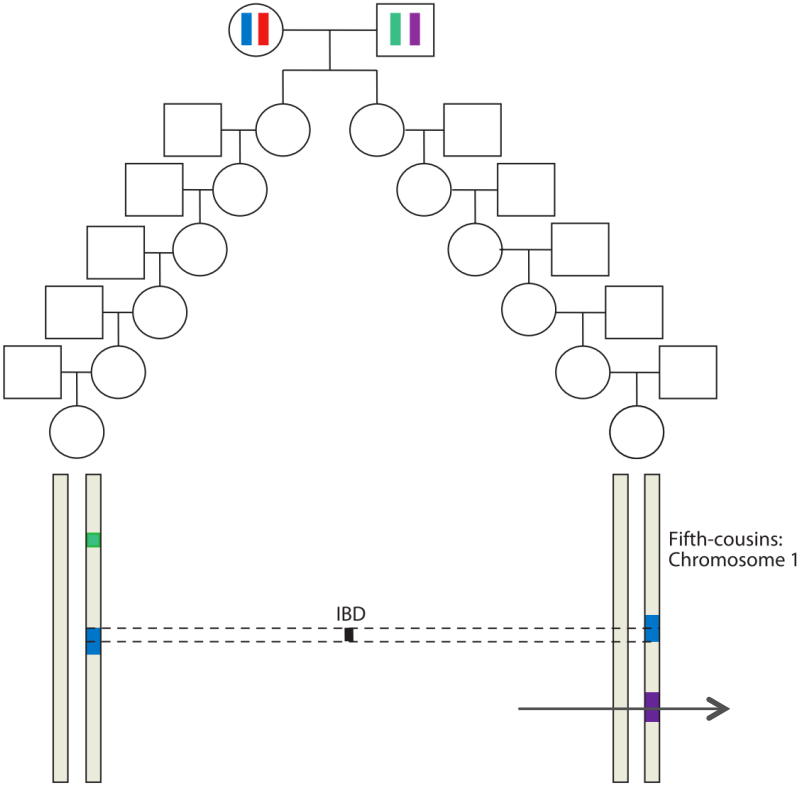
## Speed

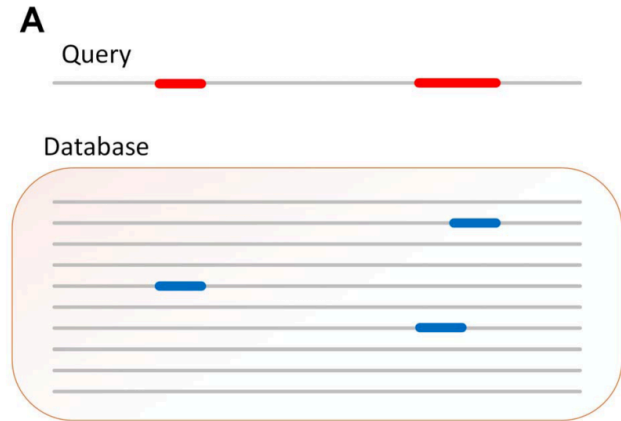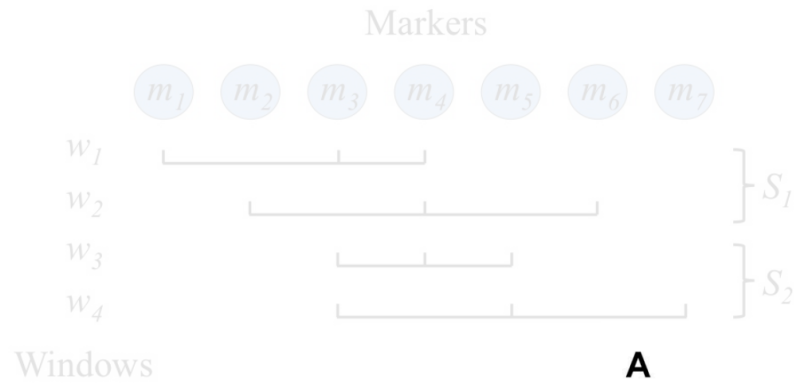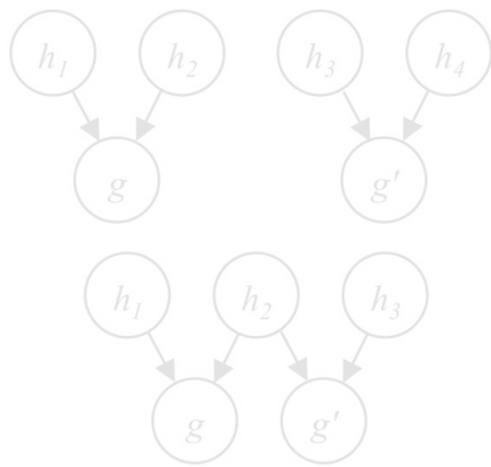| IBD Detection | | IBD Mapping | | Cohort size |
|---|---|---|---|---|
| FastIBD | + | GraphIBD | | <10K |
| GERMLINE | + | GraphIBD | | Insufficient accuracy |
| SpeeDB & Parente2 | + | GraphIBD | | >30K (upwards of 2 orders of magnitude speed increase) |

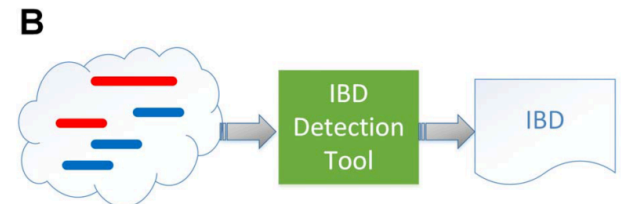# Parente2 and SpeeDB present an opportunity for applying IBD Mapping to large disease cohorts

Parente2: A fast and accurate method for detecting identity by descent

Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]
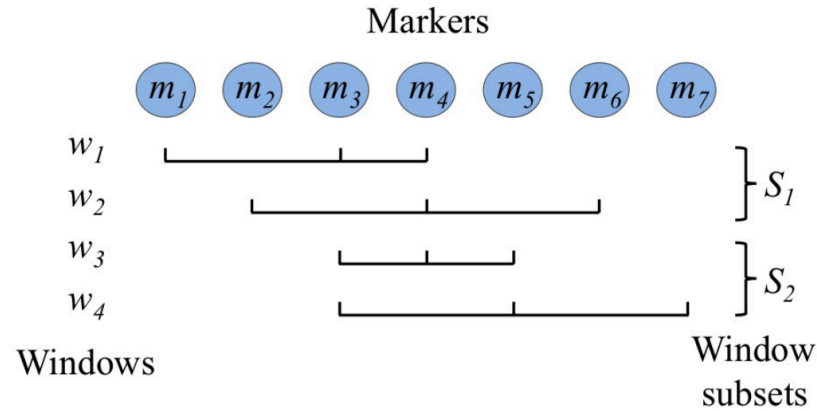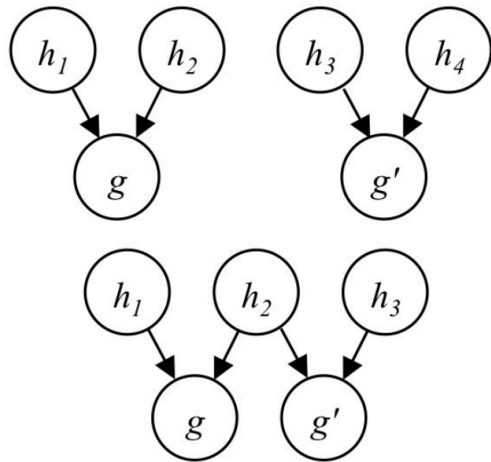
# Parente2 and SpeeDB present an opportunity for applying IBD Mapping to large disease cohorts

Parente2: A fast and accurate method for detecting identity by descent

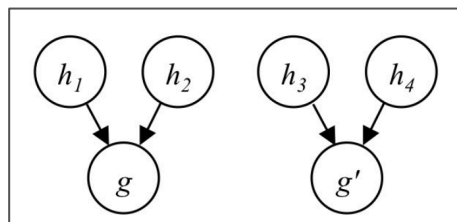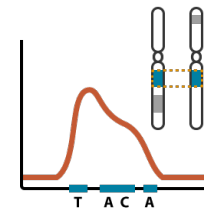Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]

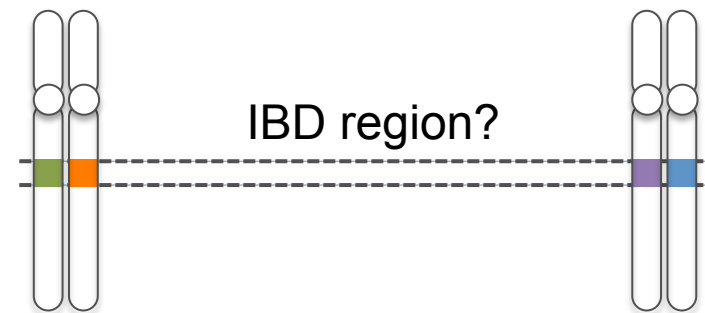## An Effective Filter for IBD Detection in Large Data Sets

Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

Non-IBD individuals

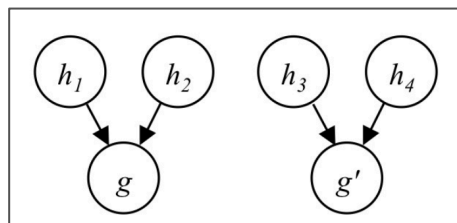IBD region?

Parente2: A fast and accurate method for detecting identity by descent

Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]

# Parente2 performs fast and accurate IBD detection
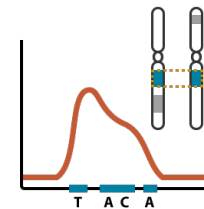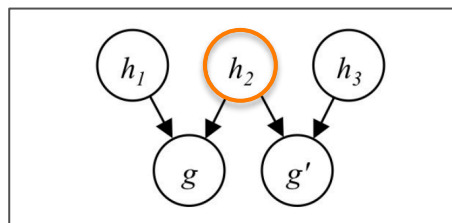


Non-IBD individuals

IBD individuals

IBD region?

Parente2: A fast and accurate method for detecting identity by descent

Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]
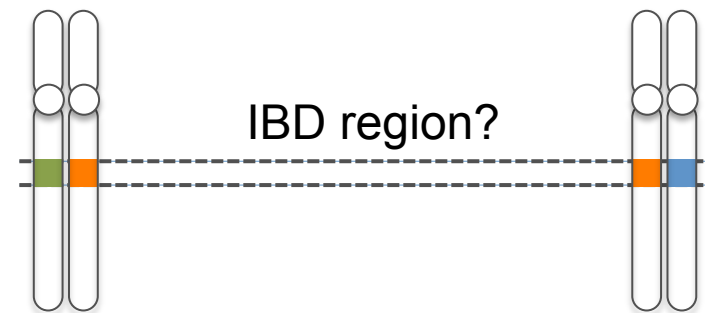
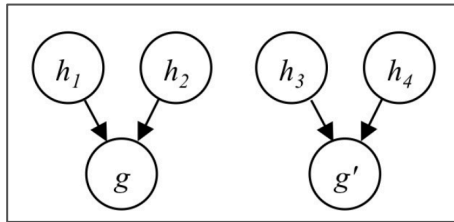# Parente2 performs fast and accurate IBD detection



Non-IBD individuals

IBD individuals

IBD region?

$$p_{M_{\overline{IBD}}}(g_{i,j}, g_{i',j}) = \sum_{h_j^1, h_j^2, h_j^3, h_j^4} p(g_{i,j}|h_j^1, h_j^2)p(g_{i',j}|h_j^3, h_j^4)p(h_j^1)p(h_j^2)p(h_j^3)p(h_j^4)$$

$$p_{M_{IBD}}(g_{i,j}, g_{i',j}) = \sum_{h_j^1, h_j^2, h_j^3} p(g_{i,j}|h_j^1, h_j^2)p(g_{i',j}|h_j^2, h_j^3)p(h_j^1)p(h_j^2)p(h_j^3)$$

Parente2: A fast and accurate method for detecting identity by descent

Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]

# Parente2 performs fast and accurate IBD detection
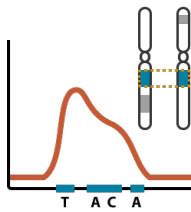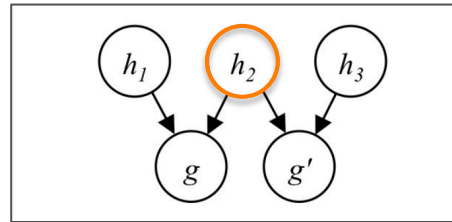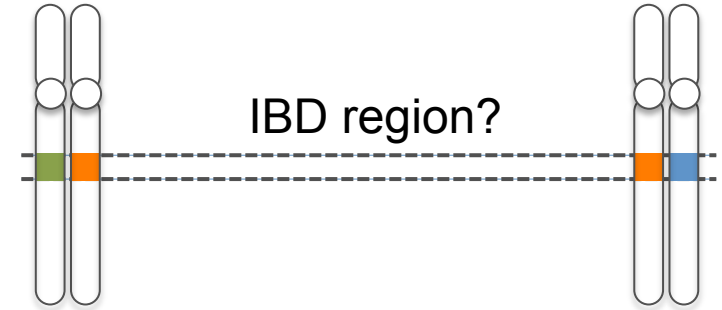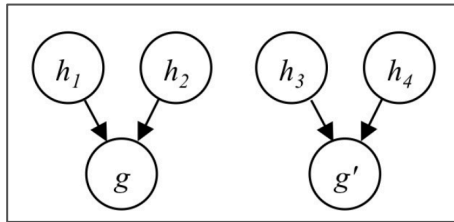


Non-IBD individuals

IBD individuals

IBD region?

Markers

Windows

Parente2: A fast and accurate method for detecting identity by descent

Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]

# Parente2 performs fast and accurate IBD detection
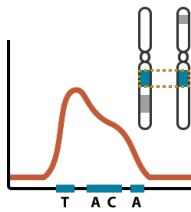


Non-IBD individuals

IBD individuals

IBD region?

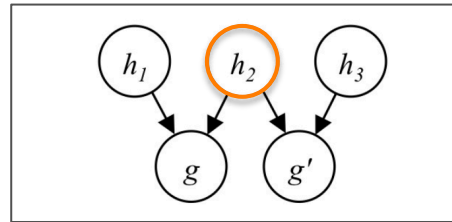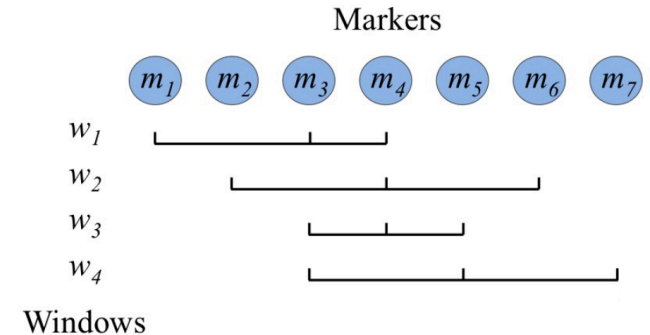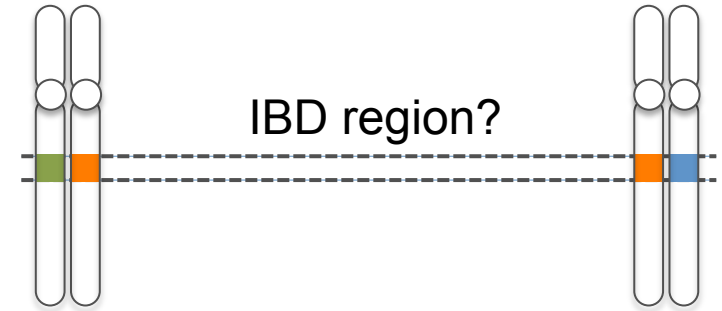Related individuals
Unrelated individuals

Markers

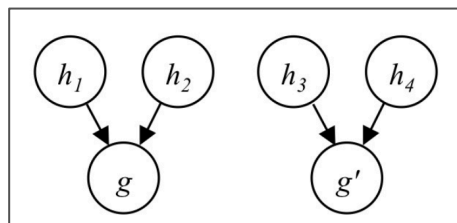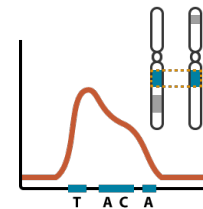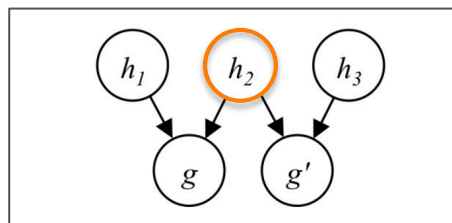Windows

Parente2: A fast and accurate method for detecting identity by descent

Jesse M. Rodriguez[*,1,2], Sivan Bercovici[*,1], Lin Huang[1], Roy Frostig[1], and Serafim Batzoglou[†,1]

# SpeeDB efficiently filters candidate IBD regions

I₁  

I₂

IBD region?

| SNV | I$_1$ geno | I$_2$ geno | IBD? |
|-----|-----------|-----------|------|
| S$_1$ | | | |
| S$_2$ | | | |
| S$_3$ | | | |
| S$_4$ | | | |

**An Effective Filter for IBD Detection in Large Data Sets**

Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

# SpeeDB efficiently filters candidate IBD regions



I$_1$

I$_2$

IBD region?

| SNV | I$_1$ geno | I$_2$ geno | IBD? |
|-----|-----------|-----------|------|
| s$_1$ | AA | AA | |
| s$_2$ | | | |
| s$_3$ | | | |
| s$_4$ | | | |

**An Effective Filter for IBD Detection in Large Data Sets**

Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

# SpeeDB efficiently filters candidate IBD regions



$I_1$

$I_2$

IBD region?

A

A

| SNV | $I_1$ geno | $I_2$ geno | IBD? |
|-----|-----------|-----------|------|
| $s_1$ | AA | AA | compatible |
| $s_2$ | | | |
| $s_3$ | | | |
| $s_4$ | | | |

# SpeeDB efficiently filters candidate IBD regions

$I_1$

$I_2$

IBD region?

| SNV | $I_1$ geno | $I_2$ geno | IBD? |
|-----|------------|------------|------|
| $S_1$ | AA | AA | compatible |
| $S_2$ | AT | AA | |
| $S_3$ | | | |
| $S_4$ | | | |

**An Effective Filter for IBD Detection in Large Data Sets**

Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

# SpeeDB efficiently filters candidate IBD regions



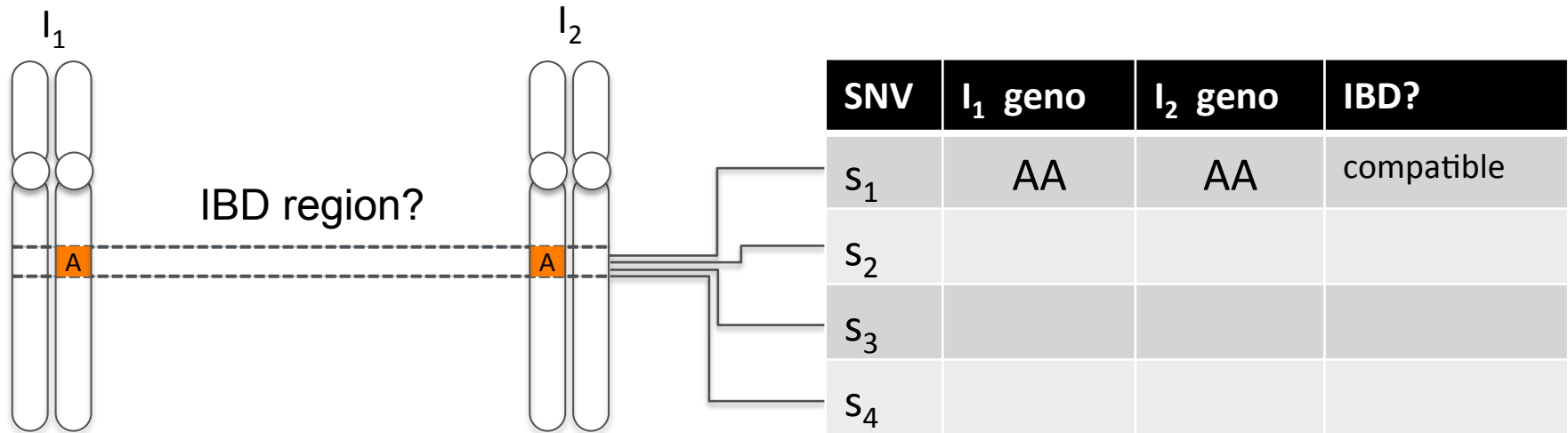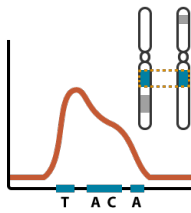| SNV | $I_1$ geno | $I_2$ geno | IBD? |
|-----|-----------|-----------|------|
| $s_1$ | AA | AA | compatible |
| $s_2$ | AT | AA | compatible |
| $s_3$ | | | |
| $s_4$ | | | |

I$_1$

I$_2$

IBD region?

**An Effective Filter for IBD Detection in Large Data Sets**

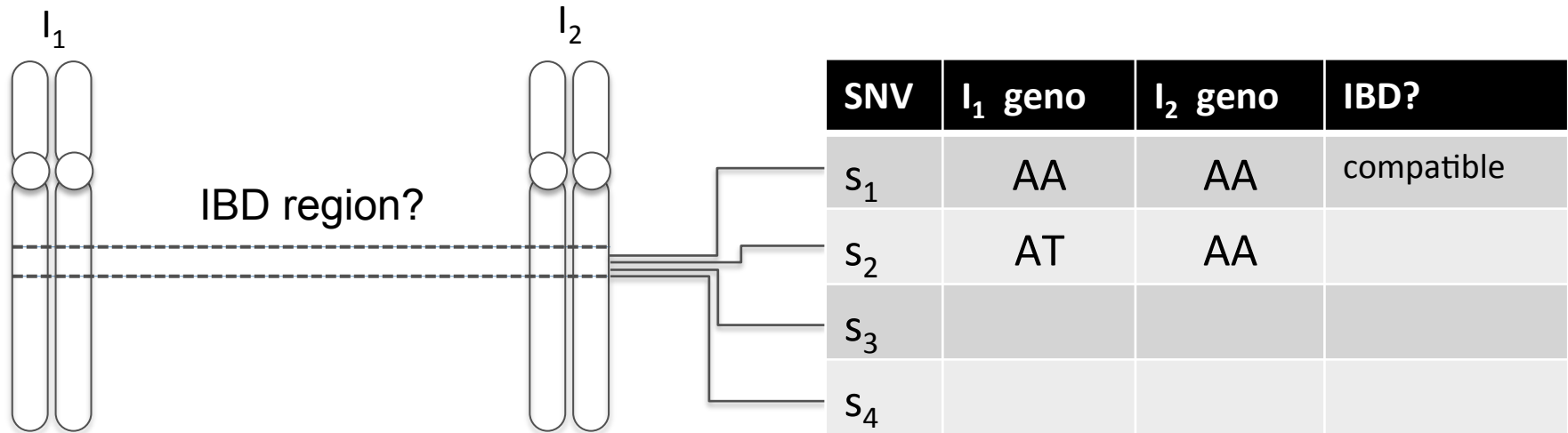Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

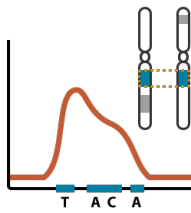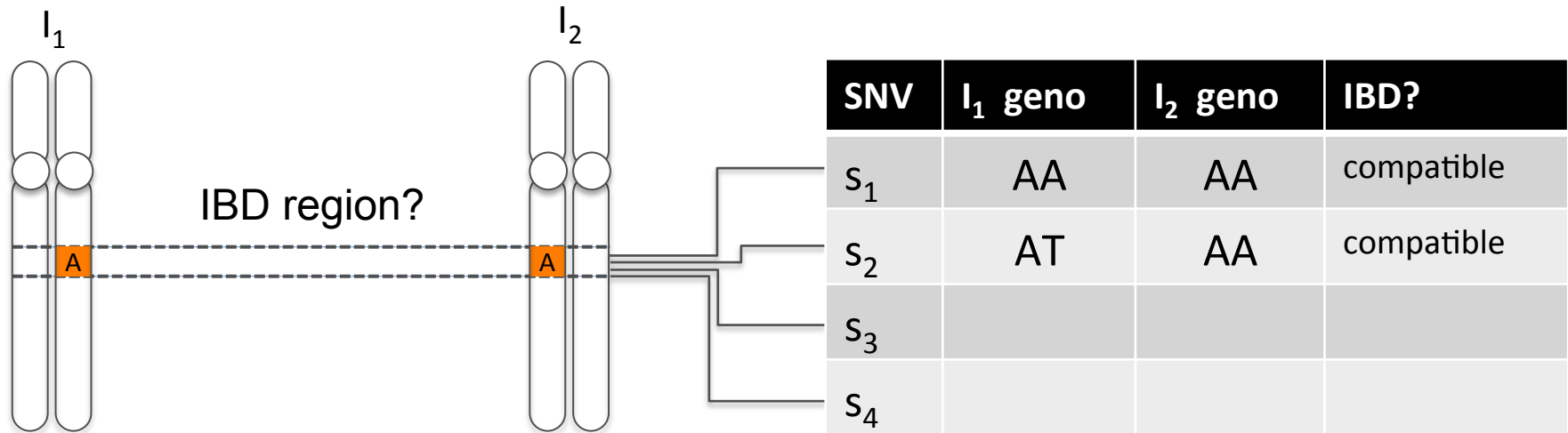# SpeeDB efficiently filters candidate IBD regions

I₁

I₂

IBD region?

| SNV | I₁ geno | I₂ geno | IBD? |
|-----|---------|---------|------|
| $s_1$ | AA | AA | compatible |
| $s_2$ | AT | AA | compatible |
| $s_3$ | AT | AT | compatible |
| $s_4$ | | | |

# SpeeDB efficiently filters candidate IBD regions

$I_1$

$I_2$

IBD region?

| SNV | $I_1$ geno | $I_2$ geno | IBD? |
|-----|-----------|-----------|------|
| $s_1$ | AA | AA | compatible |
| $s_2$ | AT | AA | compatible |
| $s_3$ | AT | AT | compatible |
| $s_4$ | AA | TT | |

**An Effective Filter for IBD Detection in Large Data Sets**

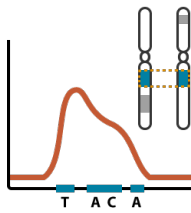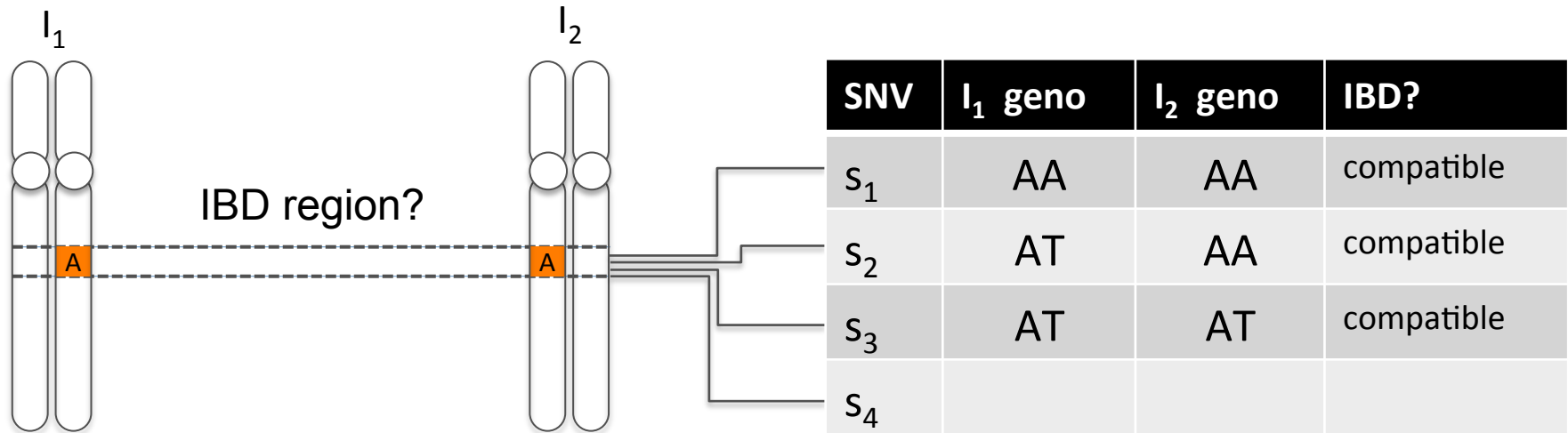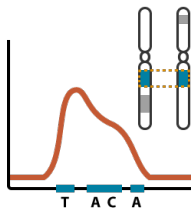Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

# SpeeDB efficiently filters candidate IBD regions



| SNV | $I_1$ geno | $I_2$ geno | IBD? |
|-----|-----------|-----------|------|
| $s_1$ | AA | AA | compatible |
| $s_2$ | AT | AA | compatible |
| $s_3$ | AT | AT | compatible |
| $s_4$ | AA | TT | conflict |

**An Effective Filter for IBD Detection in Large Data Sets**

Lin Huang*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou

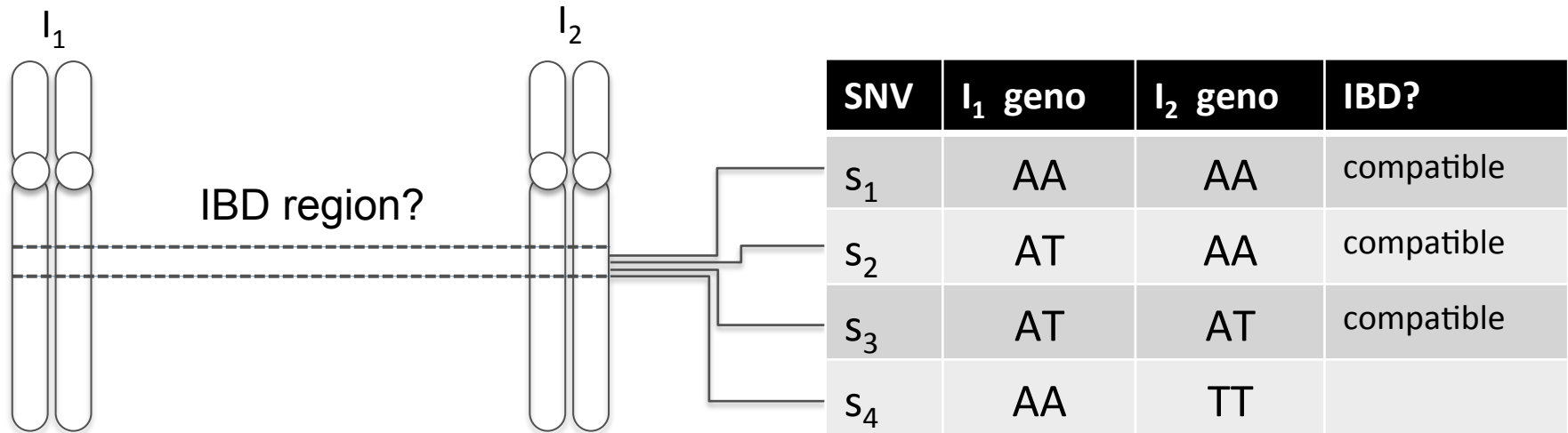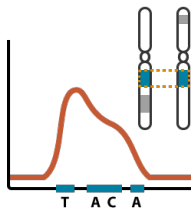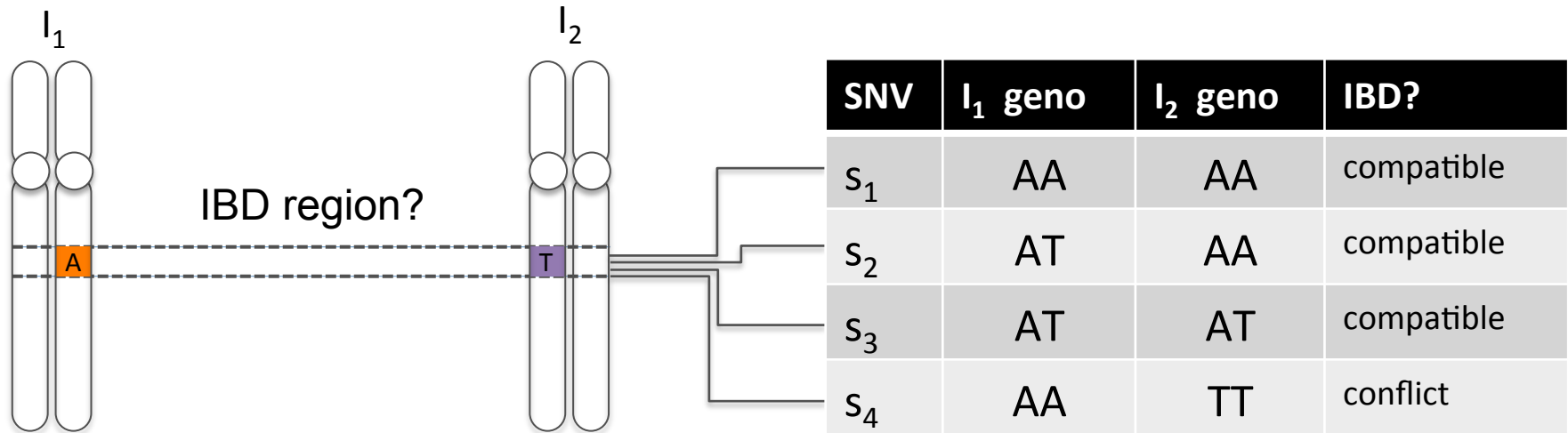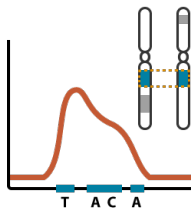# SpeeDB efficiently filters candidate IBD regions



I₁   I₂

IBD region?

| SNV | I$_1$ geno | I$_2$ geno | IBD? |
|-----|-----------|-----------|------|
| S$_1$ | AA | AA | compatible |
| S$_2$ | AT | AA | compatible |
| S$_3$ | AT | AT | compatible |
| S$_4$ | AA | TT | conflict |

**An Effective Filter for IBD Detection in Large Data Sets**

**Lin Huang\*, Sivan Bercovici, Jesse M. Rodriguez, Serafim Batzoglou**

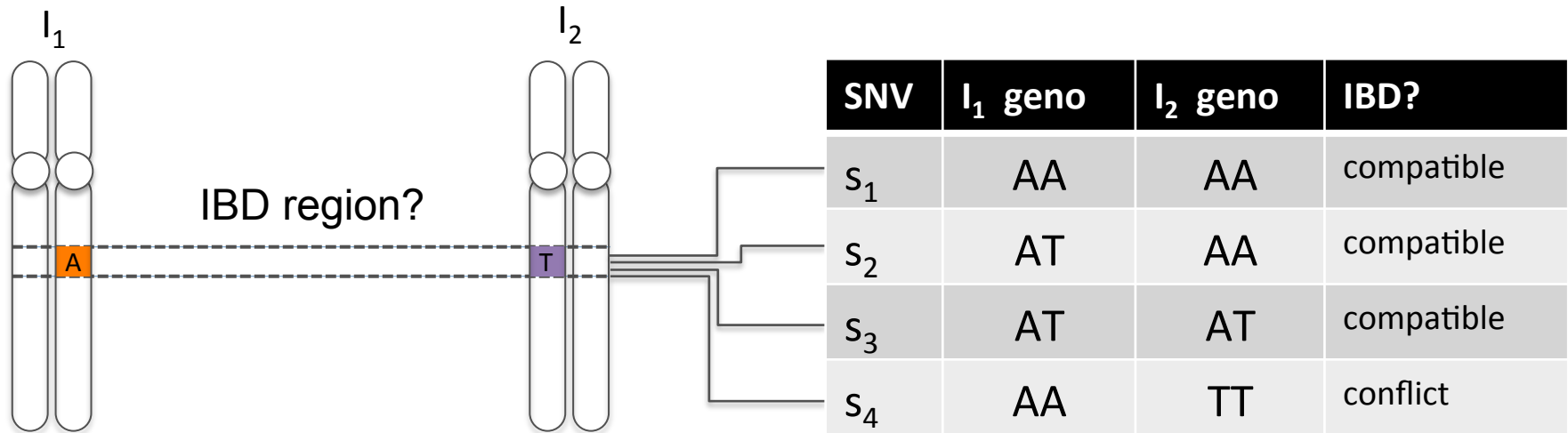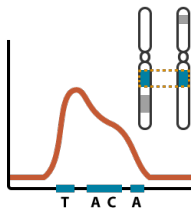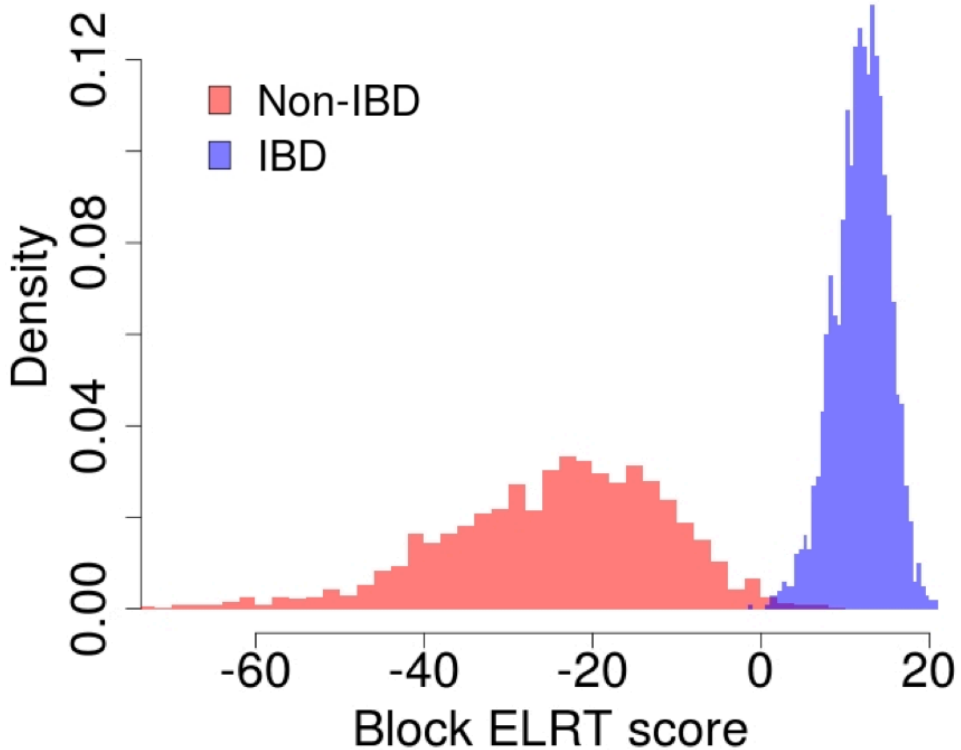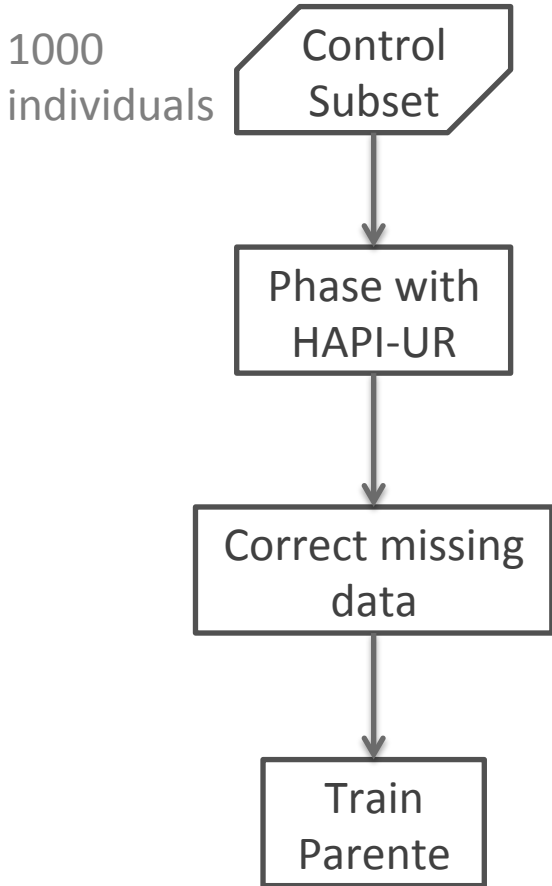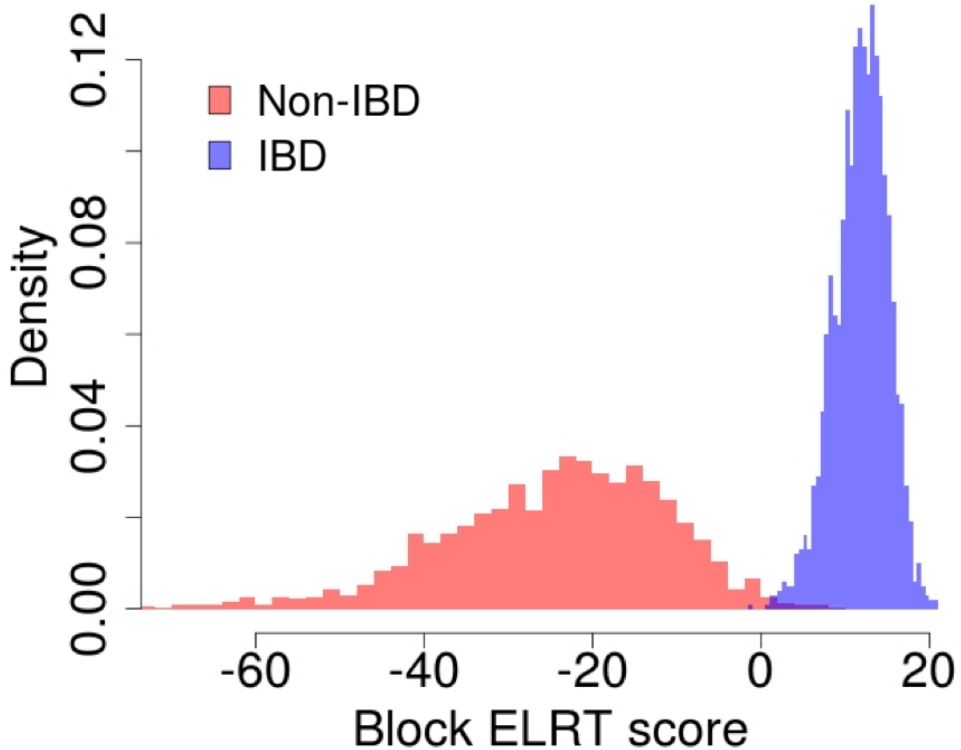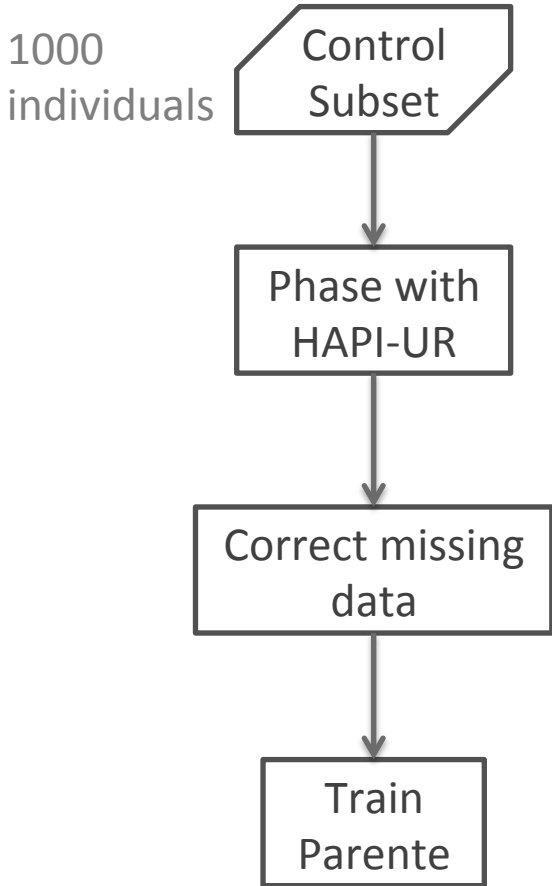# IBD Detection Pipeline

**Training**

1000 individuals → Control Subset

↓

Phase with HAPI-UR

↓

Correct missing data

↓

Train Parente



Legend:
- Non-IBD
- IBD

X-axis: Block ELRT score
Y-axis: Density

# IBD Detection Pipeline

**Training**

1000 individuals → Control Subset

Control Subset → Phase with HAPI-UR

Phase with HAPI-UR → Correct missing data

Correct missing data → Train Parente



Density vs Block ELRT score

Legend:
- Non-IBD (red)
- IBD (blue)

# IBD Detection Pipeline

**Training**

**Test**

Control Subset

↓

Phase with HAPI-UR

↓

Correct missing data

↓

Train Parente on haplotypes

Cohort minus training controls

↓

Correct missing data

↓

Filter IBD candidates with SpeeDB

↓

Parente Inference on genotypes

↓

Analyze!

# IBD Detection Pipeline – SpeeDB and Parente Parallelism



Cohort .geno

Partition (3)

1.geno    2.geno

3.geno

SpeeDB

1.dbs — 1.query     1_1    2_2

2.dbs — 2.query     1_2    2_3

3.dbs — 3.query     1_3    3_3

Cohort minus
training controls

Correct missing
data

Filter IBD candidates
with SpeeDB

Parente Inference
on genotypes

Analyze!

**A**

Query

Database

**B**

IBD
Detection
Tool

IBD

# IBD Detection Pipeline – SpeeDB and Parente Parallelism

# Next steps



As of 2012, more than 140 studies with 10,000 or more individuals