Convergent Haplotype Association Tagging (CHAT)

Yuan Lin

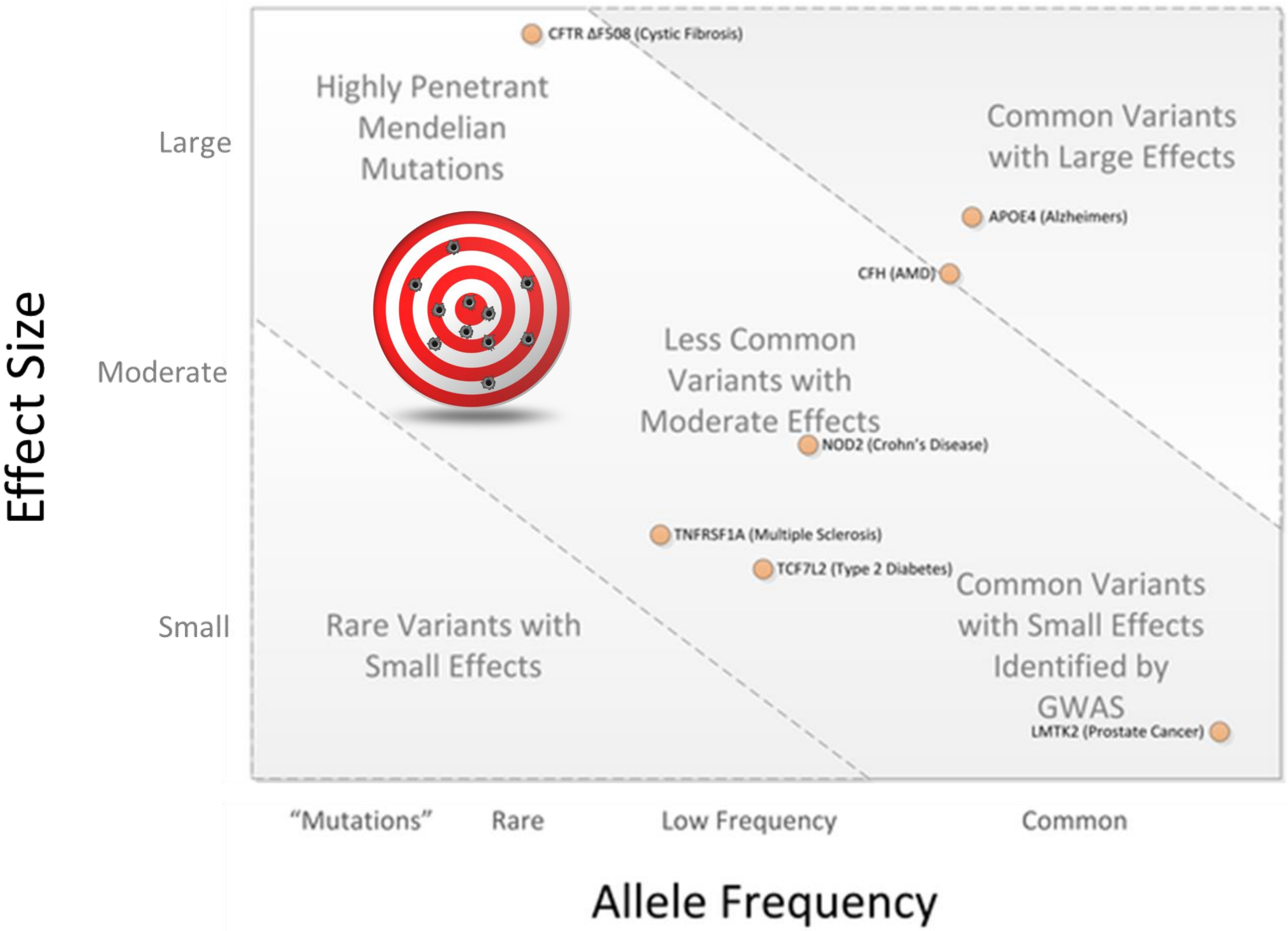
Post-doc Research Associate

@ Dr. Kirk Wilhelmsen’s Lab

Department of Genetics, School of Medicine

UNC at Chapel Hill

Targets: rare variants with moderate effects



Modified from Figure 1 in *Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12)*

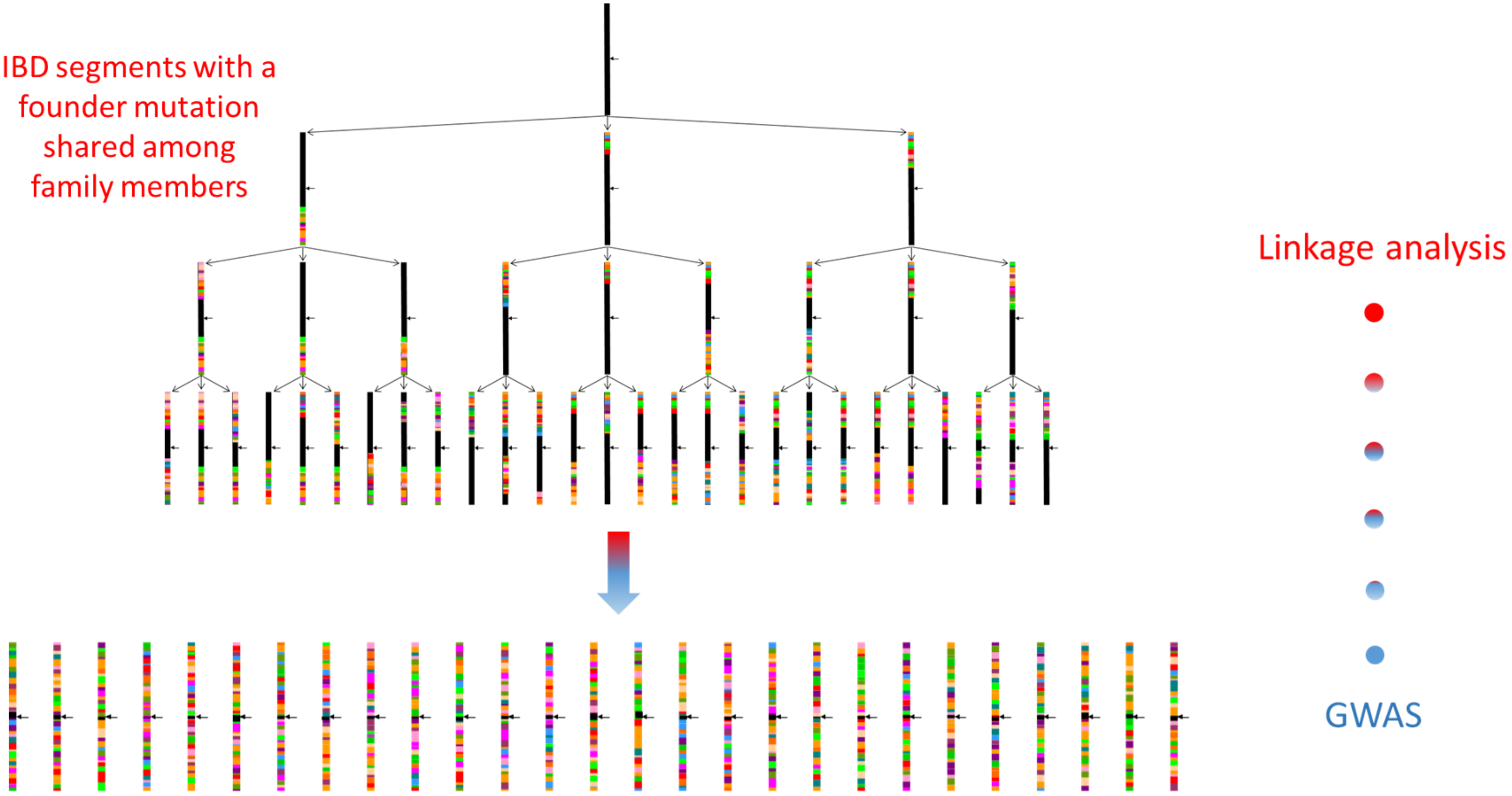
Association tagging using IBD segments

- IBD segments are chromosomal segments that share identical haplotypes because they are inherited from the same ancestor without disrupted by recombination

- We can test the association between a set of individuals’ phenotype similarity and their co-inheritance of certain IBD segments

- Each segment tags genetic variants within that segment, including potential disease causal variants

Linkage analysis and GWAS represent two extremes



IBD segments with a founder mutation shared among family members

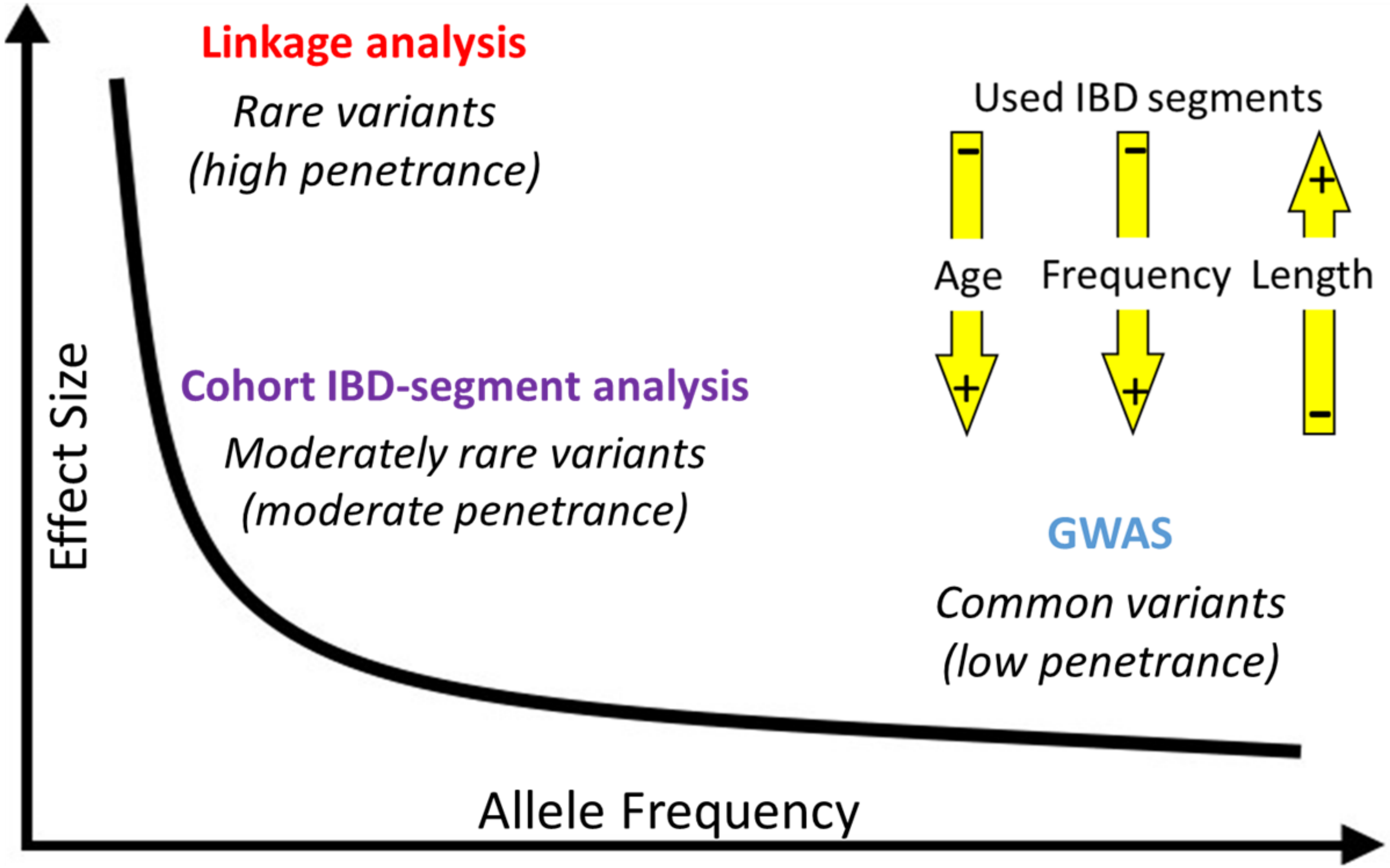
Linkage analysis

GWAS

Ancient IBD segments with the same mutation shared by population samples

Cohort IBD-segment analysis represents the middle way

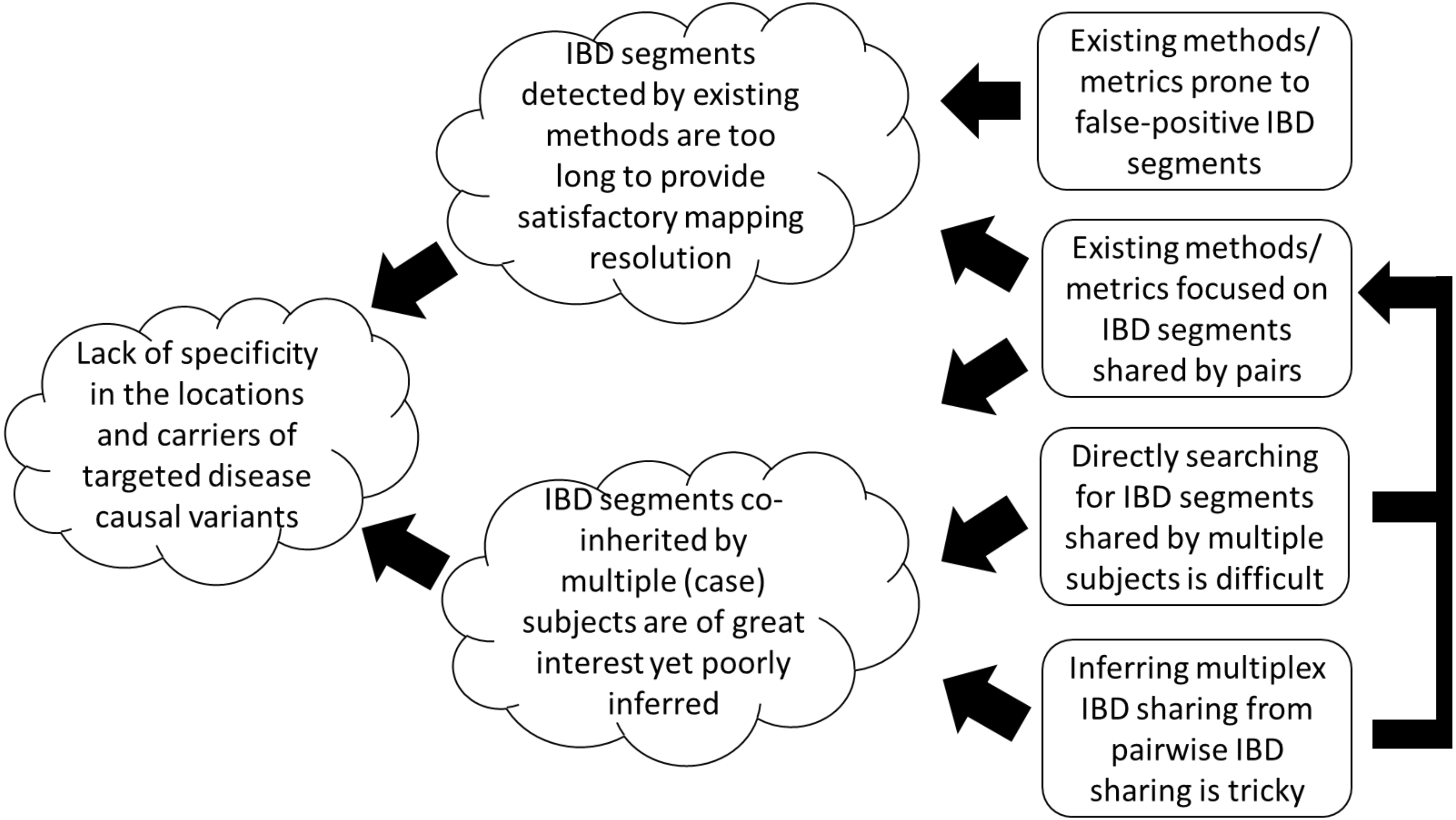
- Detect IBD sharing in unrelated population samples
- Distinguish from IBS sharing – sharing due to pure chance or strong LD in adjacent SNPs



Promising results yet not many applications

- Compared to standard GWAS, cohort IBD-segment analysis
  - Show more power in mapping genomic regions containing multiple rare disease susceptibility variants (Browning & Thompson, 2012)
  - Find genome-wide significant regions missed by the former in both isolated and outbred populations (Gusev et al., 2011ab)
- Previous applications in mapping causal variants for various diseases
  - schizophrenia, multiple sclerosis, Parkinson’s disease, Crohn's disease, several types of cancer, and diastolic blood pressure (References omitted due to space limit)
  - the most recent paper published last year (Liu et al., 2016)

Methodological limitations may have impeded the usage



Existing methods/ metrics prone to false-positive IBD segments

Existing methods/ metrics focused on IBD segments shared by pairs

Directly searching for IBD segments shared by multiple subjects is difficult

Inferring multiplex IBD sharing from pairwise IBD sharing is tricky

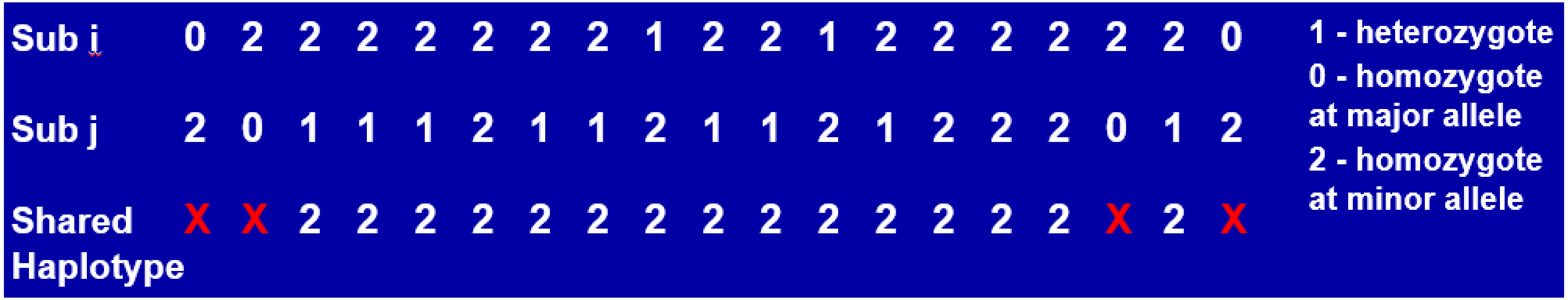
IBD segments detected by existing methods are too long to provide satisfactory mapping resolution

IBD segments co-inherited by multiple (case) subjects are of great interest yet poorly inferred

Lack of specificity in the locations and carriers of targeted disease causal variants

Existing methods prone to false long IBD segments

- Use length of (potential) haplotype sharing (or its variants) as the major (if not only) statistic to detect IBD segments
- Evaluate potential haplotype sharing via genotype compatibility
- Allow for (a small number of) incompatible genotypes

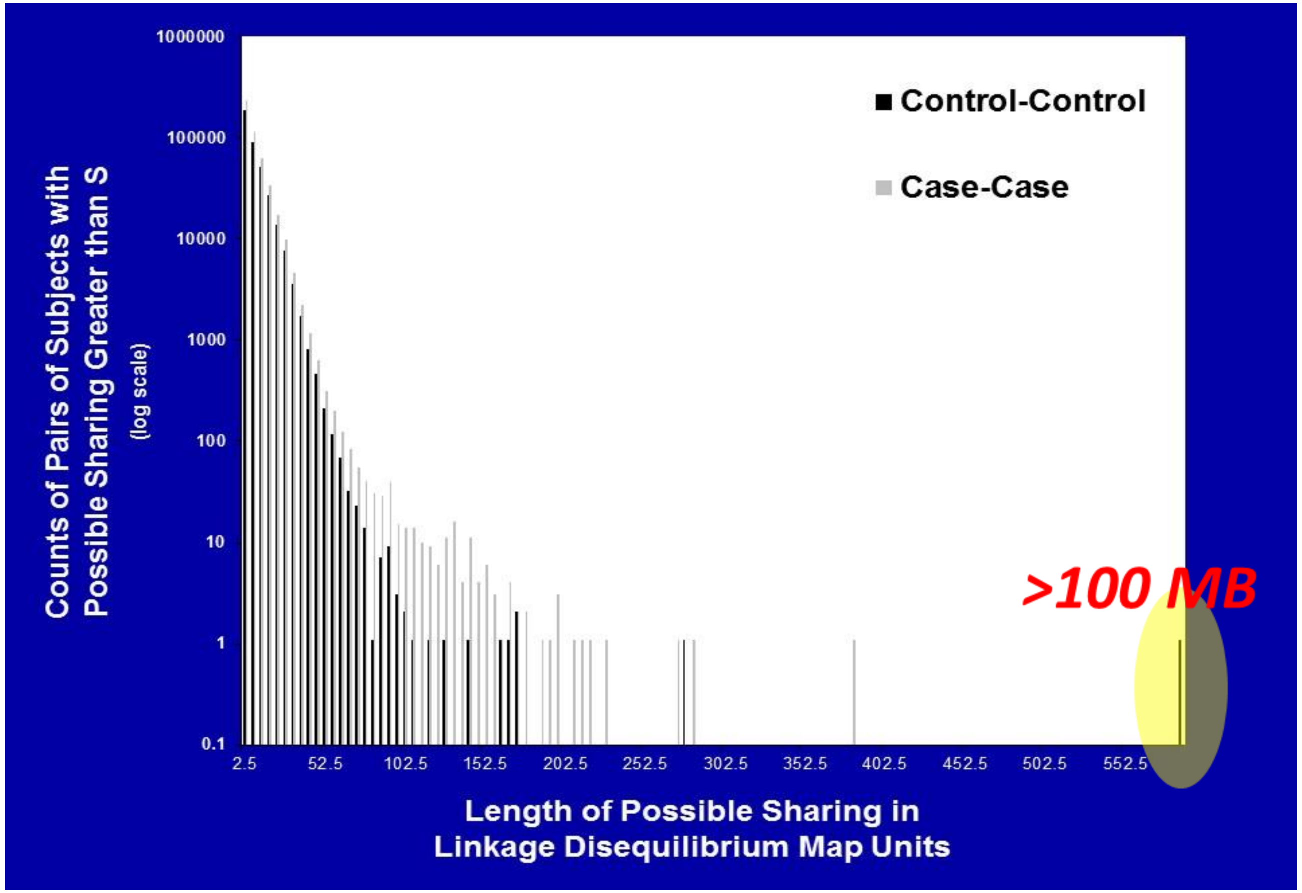


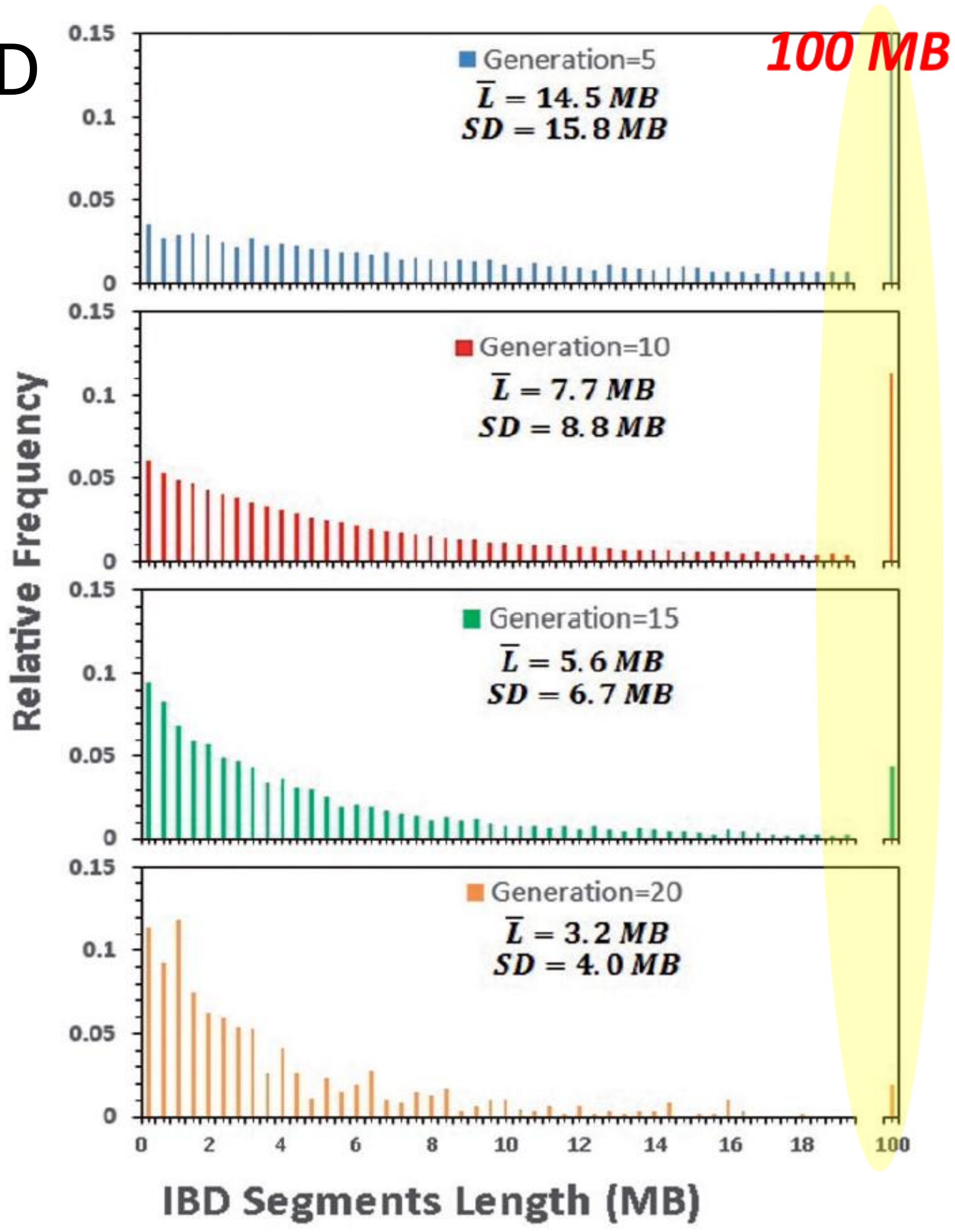
Sub i	0	2	2	2	2	2	2	2	1	2	2	1	2	2	2	2	2	2	0
Sub j	2	0	1	1	1	2	1	1	2	1	1	2	1	2	2	2	0	1	2
Shared Haplotype	X	X	2	2	2	2	2	2	2	2	2	2	2	2	2	2	X	2	X

1 - heterozygote
0 - homozygote at major allele
2 - homozygote at minor allele

Compatibility based haplotype sharing test claims mismatch only when the genotypes are homozygotes at opposite alleles

Existing methods focus on pairwise IBD segments which can be relatively long



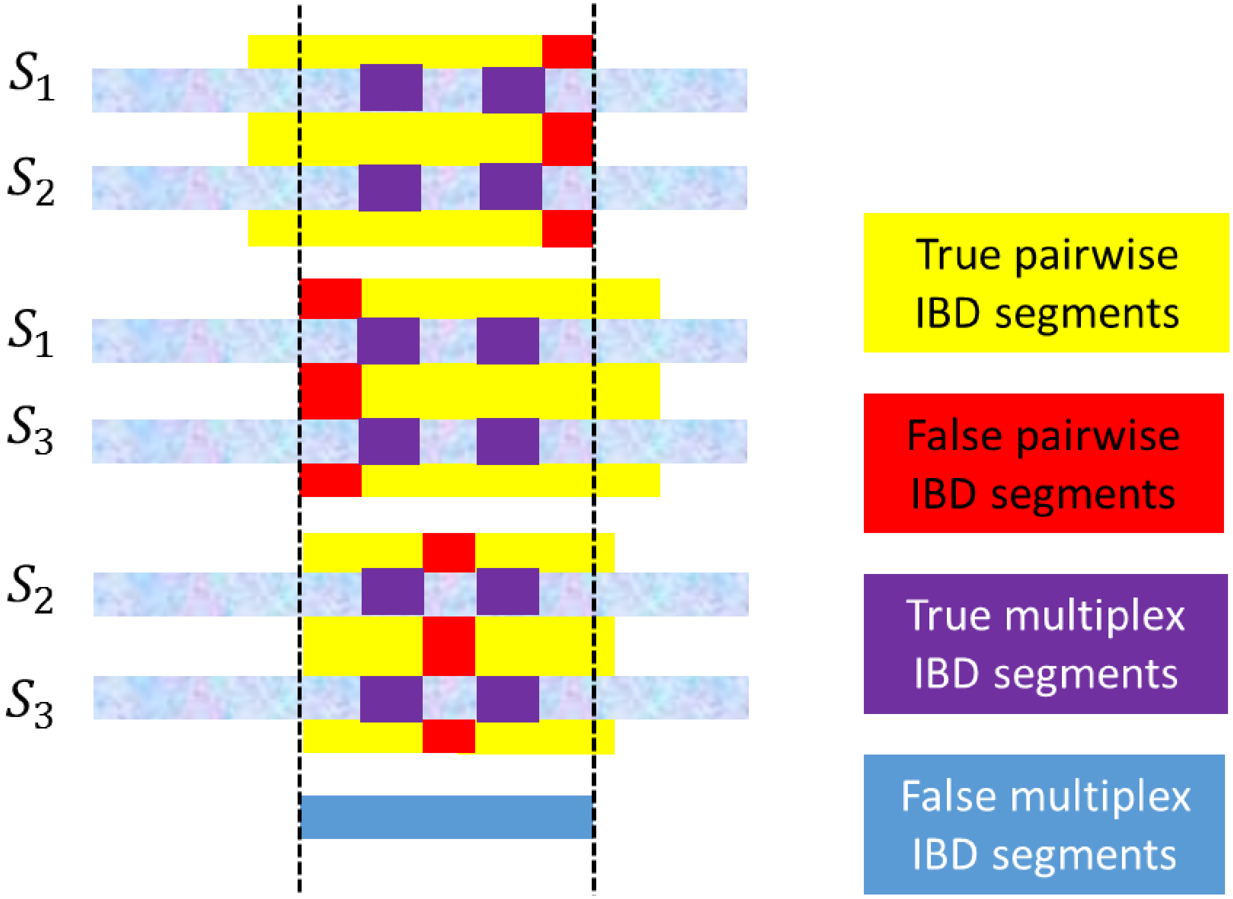


Directly searching for multiplex IBD sharing is challenging

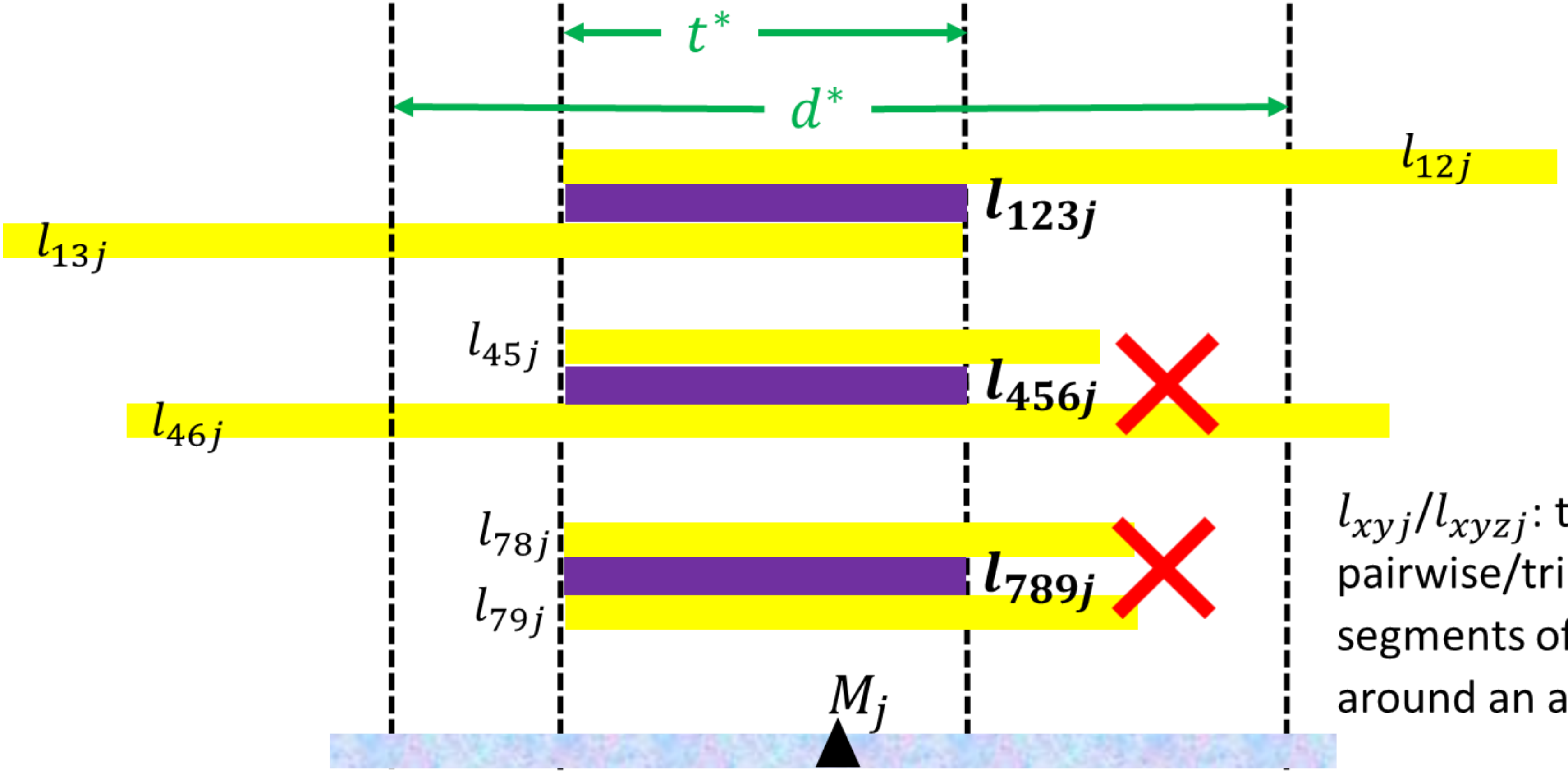
- Given a set of individuals, their multiplex IBD segments (if any) can be any part of but not necessarily the entire chromosomal regions where all individuals have compatible genotypes
- Effective statistics have yet to be developed
  - Multiplex IBD segments are shorter than any pairwise IBD segments shared among the same group of individuals
- Search space is huge
  - Given ***n*** subjects and ***m*** analysis points/regions on the genome, an exhaustive search through all subsets of subjects has complexity ~ $O(mn \cdot 2^n)$

Inferring multiplex IBD sharing from pairwise sharing is tricky

- Transitivity of IBD relation: if chromosome segments A and B are true IBD and B and C are true IBD at the same point *j* then A and C must be IBD at point *j* too.
- IBD transitivity allows us to infer multiplex IBD segments from the overlapping region of (some, not necessarily all) pairwise IBD segments


S_1
S_2
S_1
S_3
S_2
S_3
True pairwise IBD segments
False pairwise IBD segments
True multiplex IBD segments
False multiplex IBD segments

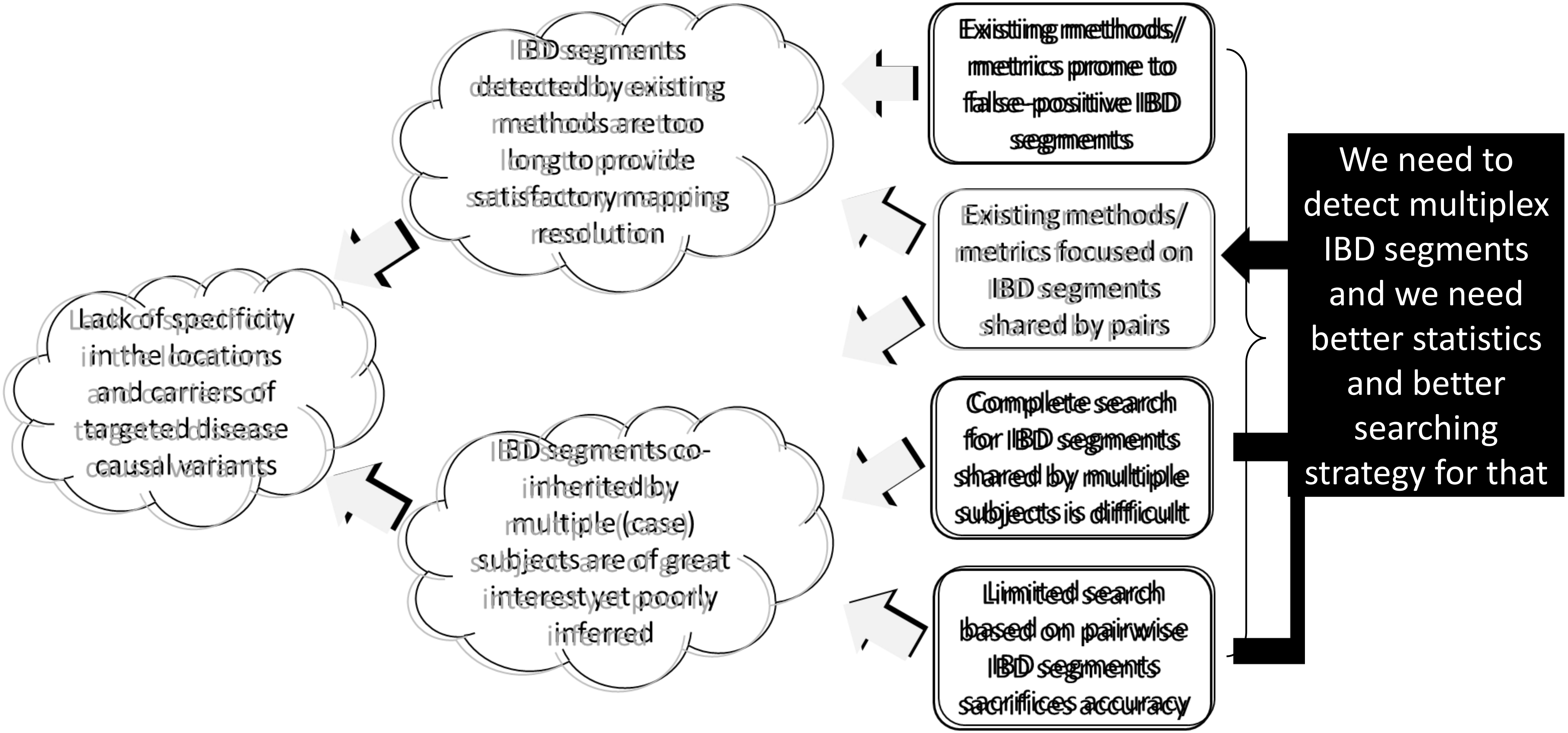
Inferring multiplex IBD sharing from pairwise IBD sharing is tricky



l_{xyj}/l_{xyzj}: true pairwise/triplex IBD segments of $\{S_x, S_y, S_z\}$ around an analysis point M_j

To prevent false-negative errors in triplex IBD detection, we may need to use loose inclusion criteria for pairwise IBD segments

Major problems and our solutions



Existing methods/ metrics prone to false-positive IBD segments

Existing methods/ metrics focused on IBD segments shared by pairs

Complete search for IBD segments shared by multiple subjects is difficult

Limited search based on pairwise IBD segments sacrifices accuracy

IBD segments detected by existing methods are too long to provide satisfactory mapping resolution

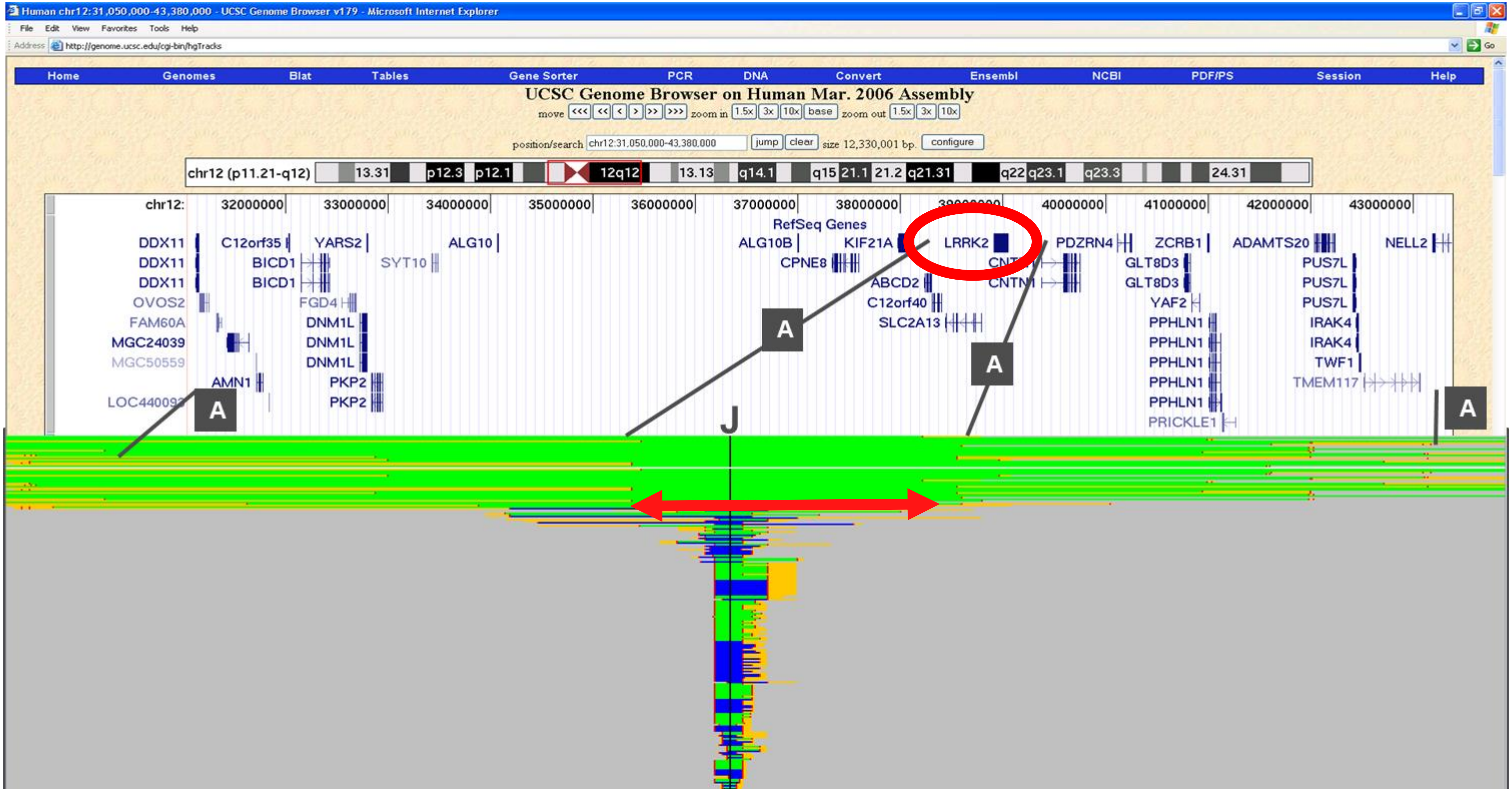
IBD segments co-inherited by multiple (case) subjects are of great interest yet poorly inferred

Lack of specificity in the locations and carriers of targeted disease causal variants

We need to detect multiplex IBD segments and we need better statistics and better searching strategy for that

CHAT (Convergent Haplotype Association Tagging)

- CHAT targets at rare variants with moderate effects that neither GWAS nor linkage analysis can effectively find
- CHAT relies on detecting short IBD segments shared by small (but significant) subsets of population samples to tag these variants
- Our statistic detects IBD sharing among three subjects (trios) based on their genotypes with high specificity.
- Our strategy for continuously searching subsets of individuals (larger than trios) that are plausibly IBD utilizes graph theory and is relatively efficient


Human chr12:31,050,000-43,380,000 - UCSC Genome Browser v179 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://genome.ucsc.edu/cgi-bin/hgTracks
Go
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
UCSC Genome Browser on Human Mar. 2006 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr12:31,050,000-43,380,000 jump clear size 12,330,001 bp. configure
chr12 (p11.21-q12)
13.31 p12.3 p12.1 12q12 13.13 q14.1 q15 21.1 21.2 q21.31 q22 q23.1 q23.3 24.31
chr12: 32000000 33000000 34000000 35000000 36000000 37000000 38000000 39000000 40000000 41000000 42000000 43000000
RefSeq Genes
DDX11
DDX11
DDX11
OVOS2
FAM60A
MGC24039
MGC50559
LOC440093
C12orf35
BICD1
BICD1
AMN1
YARS2
SYT10
FGD4
DNM1L
DNM1L
DNM1L
PKP2
PKP2
ALG10
ALG10B
CPNE8
KIF21A
ABCD2
C12orf40
SLC2A13
LRRK2
CNTN1
CNTN1
PDZRN4
ZCRB1
GLT8D3
GLT8D3
YAF2
PPHLN1
PPHLN1
PPHLN1
PPHLN1
PPHLN1
PRICKLE1
ADAMTS20
PUS7L
PUS7L
PUS7L
IRAK4
IRAK4
TWF1
TMEM117
NELL2
A
A
A
A
J

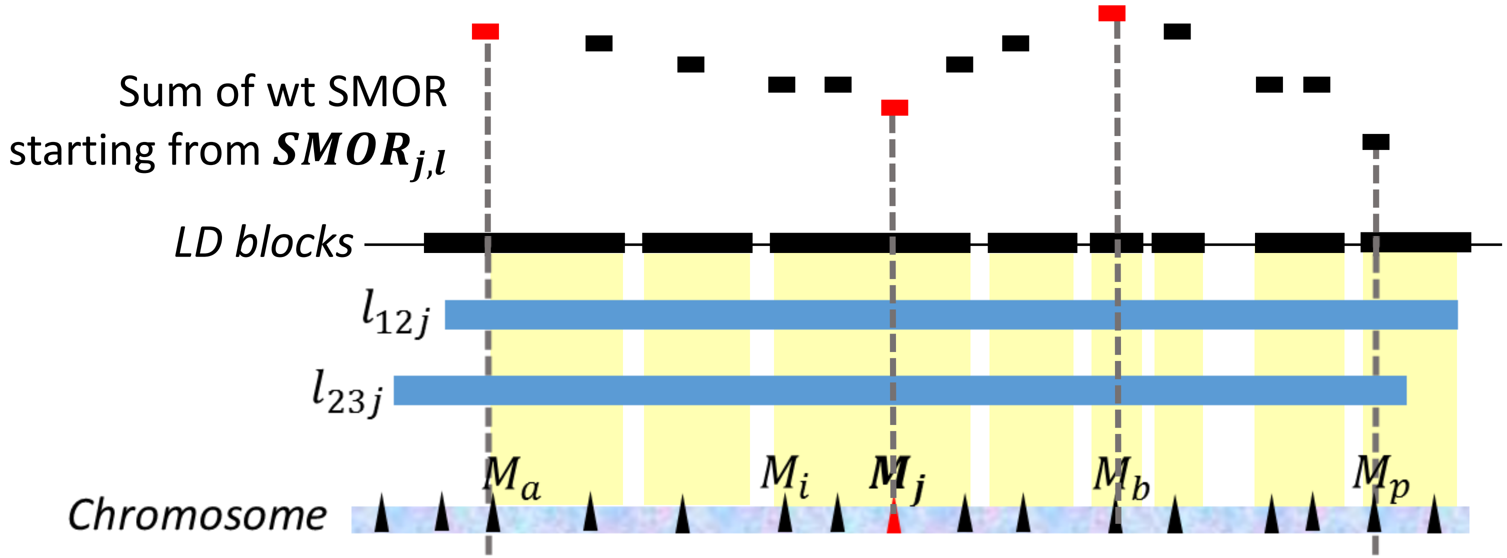
CHAT (Convergent Haplotype Association Tagging)

- CHAT targets at rare variants with moderate effects that neither GWAS nor linkage analysis can effectively find
- CHAT relies on detecting short IBD segments shared by small (but significant) subsets of population samples to tag these variants
- Our statistic detects IBD sharing among three subjects (trios) based on their genotypes with high specificity.
- Our strategy for continuously searching subsets of individuals (larger than trios) that are plausibly IBD utilizes graph theory and is relatively efficient

Our statistic for trio IBD sharing: *maxSumWtSMOR*

Trio $l = \{S_1, S_2, S_3\}$ has two detected pairwise IBD segments l_{12j} and l_{23j} around an analysis point j which share (inferred) haplotypes from Marker a to Marker p

Do the trio share an IBD segment around Marker j?



Single marker odds ratio of IBD against IBS

$$\boldsymbol{SMOR_{i,l}} = \log \frac{Prob(GTS_{i,l} \mid IBD_{i,l})}{Prob(GTS_{i,l})}$$

Weighted SMOR

$$f(LD(i, [i-w, i+w])) \cdot \boldsymbol{SMOR_{i,l}} \ \ (f \leq 1)$$

Max sum of wt SMOR for trio l at analysis point j

$$\sum_{i=a}^{j} wt\boldsymbol{SMOR_{i,j,l}} + \sum_{i=j+1}^{b} wt\boldsymbol{SMOR_{i,j,l}}$$

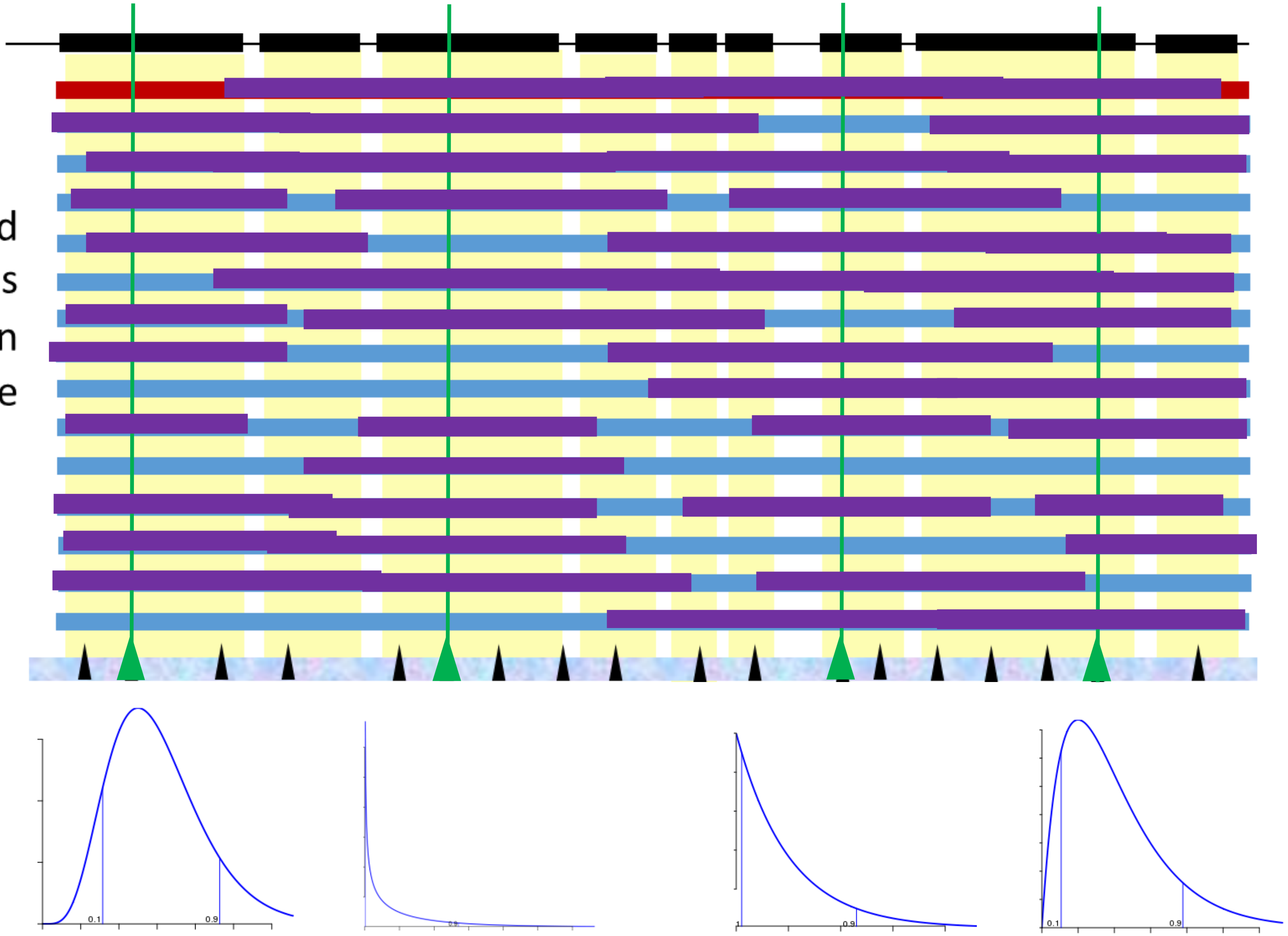
Closet value

Pre-calculated SMOR scores

- We first created a database to get all possible actual genotype combinations given three identical observed genotypes, assuming they or IBD or IBS
- We then calculated a series of SMOR scores using these combinations and a range of allele frequencies and genotyping error rates

Estimate region and subject-specific null distributions

Genotypes of S and sampled controls $\{S_{co\_k}\}$ in a given chromosome



$$\text{Trio } l_1 = \{S_{co\_1}, S, S_{co\_2}\}$$
$$\text{Trio } l_2 = \{S_{co\_2}, S, S_{co\_3}\}$$
$$\text{Trio } l_3 = \{S_{co\_1}, S, S_{co\_3}\}$$
$$\vdots$$
$$\text{Trio } l_n = \{S_{co\_1}, S, S_{co\_k}\}$$

$\boldsymbol{Gamm_{j,S}}$ is empirically estimated using $\boldsymbol{maxSumWtSMOR_{j,l_i}}$ $(0 \leq i \leq n)$

Evaluate the significance of trio IBD sharing

Trio $l = \{S_1, S_2, S_3\}$ share an IBD segment around Marker $\boldsymbol{j}$ if

$$\sqrt[3]{\prod\nolimits_{i\in\{1,2,3\}}\left(-\log_{10} P_{j,S_i}(\mu)\right)} \geq C$$

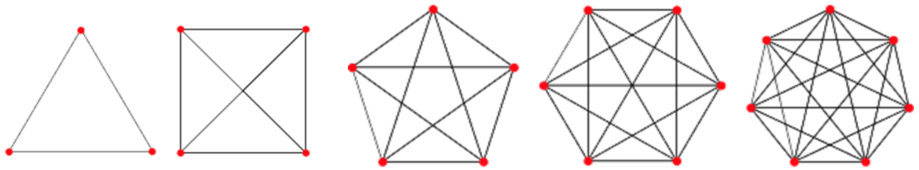
where C is a predefined fixed threshold, $\mu = maxSumWtSMOR_{j,l}$, $P_{j,S_i}(\mu)$ is the p value of μ given by $Gamm_{j,S_i}$,

- *maxSumWtSMOR* increases with the length of sharing and the rarity of genotypes in the segment. SMOR is weighted down based on local LD to not inflate μ by over-representing some regions
- Region and subject specific null distributions prevent spurious IBD sharing caused by long sequence of heterozygotes
- Geometric mean ensures that the inclusion of even one single subject that is not IBD with the rest of the trio will be easily detected.

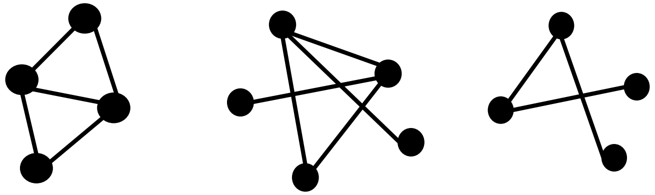
Our searching strategy: graph-based trio extension

- IBD transitivity dictates that every IBD subset should form a clique in this graph.
- False pairwise IBD detection results lead to missing or wrong edges and eventually subgraphs that violate IBD transitivity.
- Triangles are building blocks of cliques (size >=3). Open triads are building blocks of semi cliques.

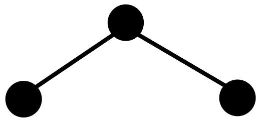
K-vertex Cliques (*K* >=3)



Semi-cliques



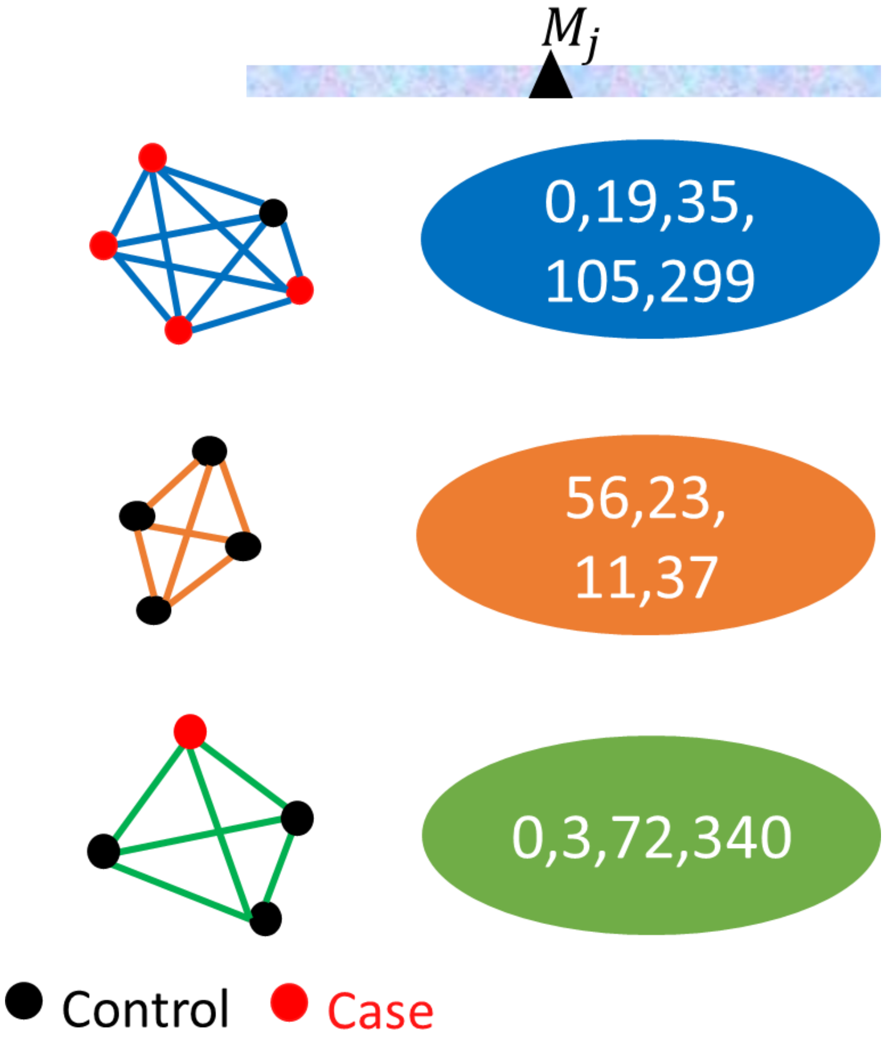
Open triad



Our searching strategy: graph-based trio extension

- After examining
- CHAT searches for IBD subsets with > 3 members by examining open triads and merging triangles with at least one overlapping edge.

Test the disease association: nominal p



Fisher's exact test

Nominal P

	In-set	Out-set
Case	4	$N_1 - 4$
Control	1	$N - N_1 - 1$

$$10^{-9}$$

	In-set	Out-set
Case	0	N_1
Control	4	$N - N_1 - 4$

$$0.8$$

	In-set	Out-set
Case	1	$N_1 - 1$
Control	3	$N - N_1 - 3$

$$10^{-5}$$

Test the disease association: adjusted p

- CHAT uses a permutation approach as IBD subsets at the same or nearby analysis point/region tend to have similar members and thus are not independent
- To reduce computational burden and speed up the procedure, CHAT models the null distribution of most significant P-values with extreme value distributions (EVD)
  - It chooses among EVD family members and learns distribution parameters from 1000 permutations using maximum likelihood estimation.
  - The cumulative area under the curve of the fitted EVD density function is used to estimate small adjusted P-values at any significance level.

Biased urn sampling based permutation tests

- Instead of randomly shuffling disease status among subjects to create permutation sets, CHAT applies a biased urn sampling procedure (Epstein et al., 2012)
- It preserves the confounding structure in the original data set as well as the numbers of cases and controls in permuted data sets
- Allow CHAT to handle more general population (by including PC covariates) and an arbitrary number of categorical and continuous covariates

THANK YOU

"If one surely believes that the disease has an underlying genetic basis that is at least partially shared among affected individuals … then, presumably, this means that the cases will be somewhat more closely related to each other, on average, than they are to control individuals." (Voight & Pritchard, 2005)