CHAT: some experiment results

Yuan Lin
Post-doc Research Associate
@ Dr. Kirk Wilhelmsen's Lab
Department of Genetics, School of Medicine
UNC at Chapel Hill

Background

- CHAT is designed to find a group of cases that share one or more IBD segments that are likely to harbor rare disease-associated genetic variants.
- I will refer to a group of individuals who have a common IBD segment as an IBD cluster and a cluster of k individuals a **k-order cluster**.
- CHAT detects IBD clusters around each marker. Each cluster **tags** an IBD segment shared by **all** the members.
- **CHAT aims for higher-order clusters**, because they tend to tag shorter segments and thus provide higher mapping resolution.

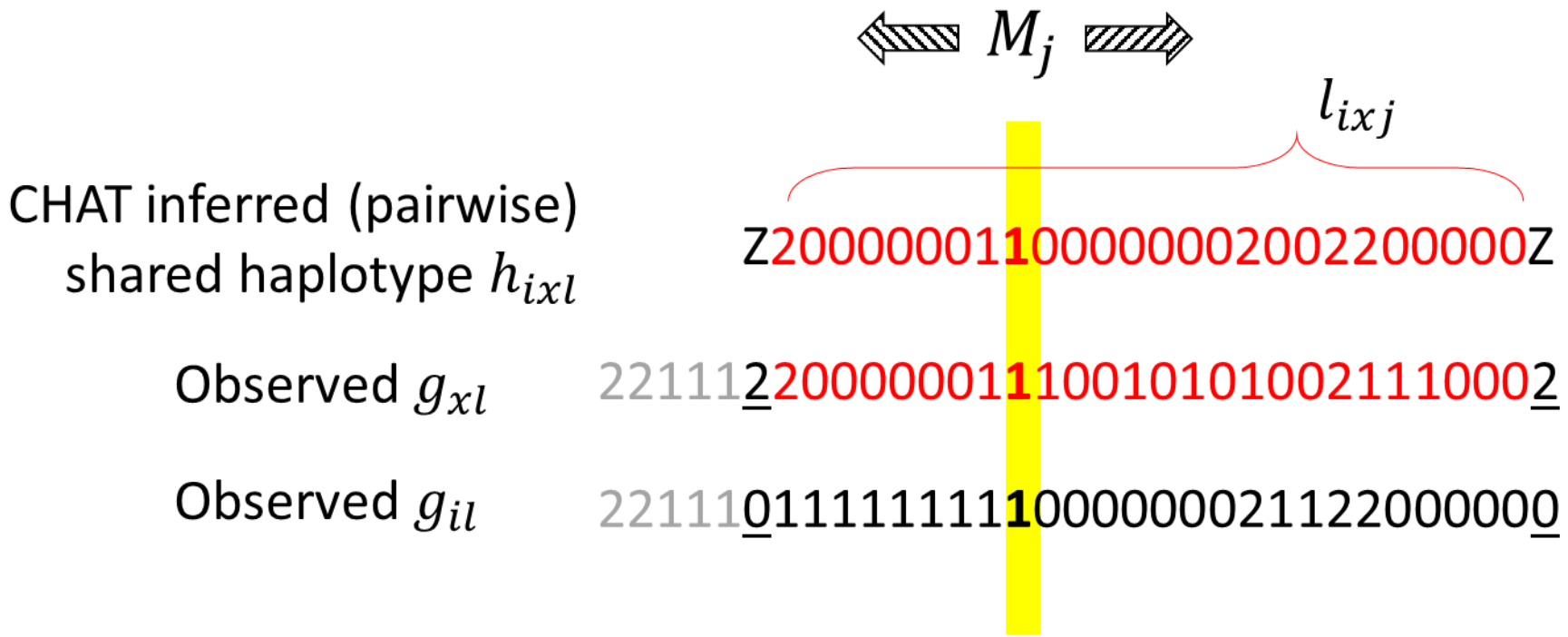
Background (cont.)

- However, high-order IBD clusters are increasingly difficult to detect.
- CHAT determines whether a shared segment is IBD or IBS (i.e., shared by chance) by evaluating genotype/haplotype similarity in that segment **statistically**.
- A major task of CHAT is to **accurately and efficiently detect high-order IBD clusters**.
- Then CHAT uses Fisher's exact test to evaluate the association of each IBD cluster with the disease. The segment tagged by a disease-associated cluster is likely to harbor causal variant(s).

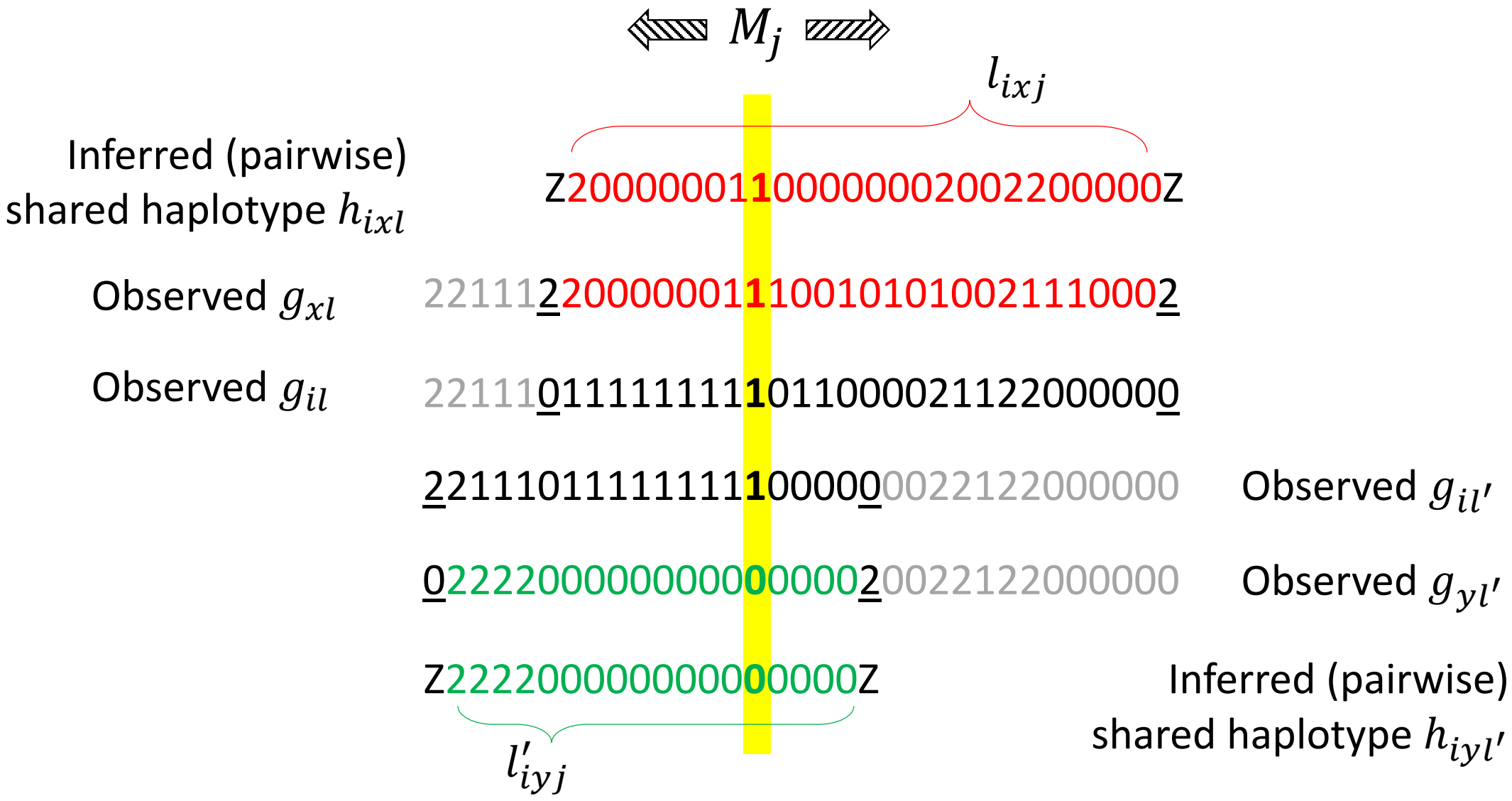
Nomenclature

- Subject Z_i, i = 1,…,n; genetic marker M_j, j = 1,…,m
- A chromosomal segment l_j starts from Marker M_j towards two directions containing adjacent markers
- The genotype/haplotype of Subject Z_i in Segment l: g_{il}/h_{il}
- Two or more subjects $\{Z_{i_1}, \dots, Z_{i_k}\}$ share a segment $l_{i_1 \dots i_k j}$ around M_j if in that segment their observed genotypes are compatible or their inferred haplotypes are the same*.

CHAT progressively infers haplotypes shared by multiple individuals



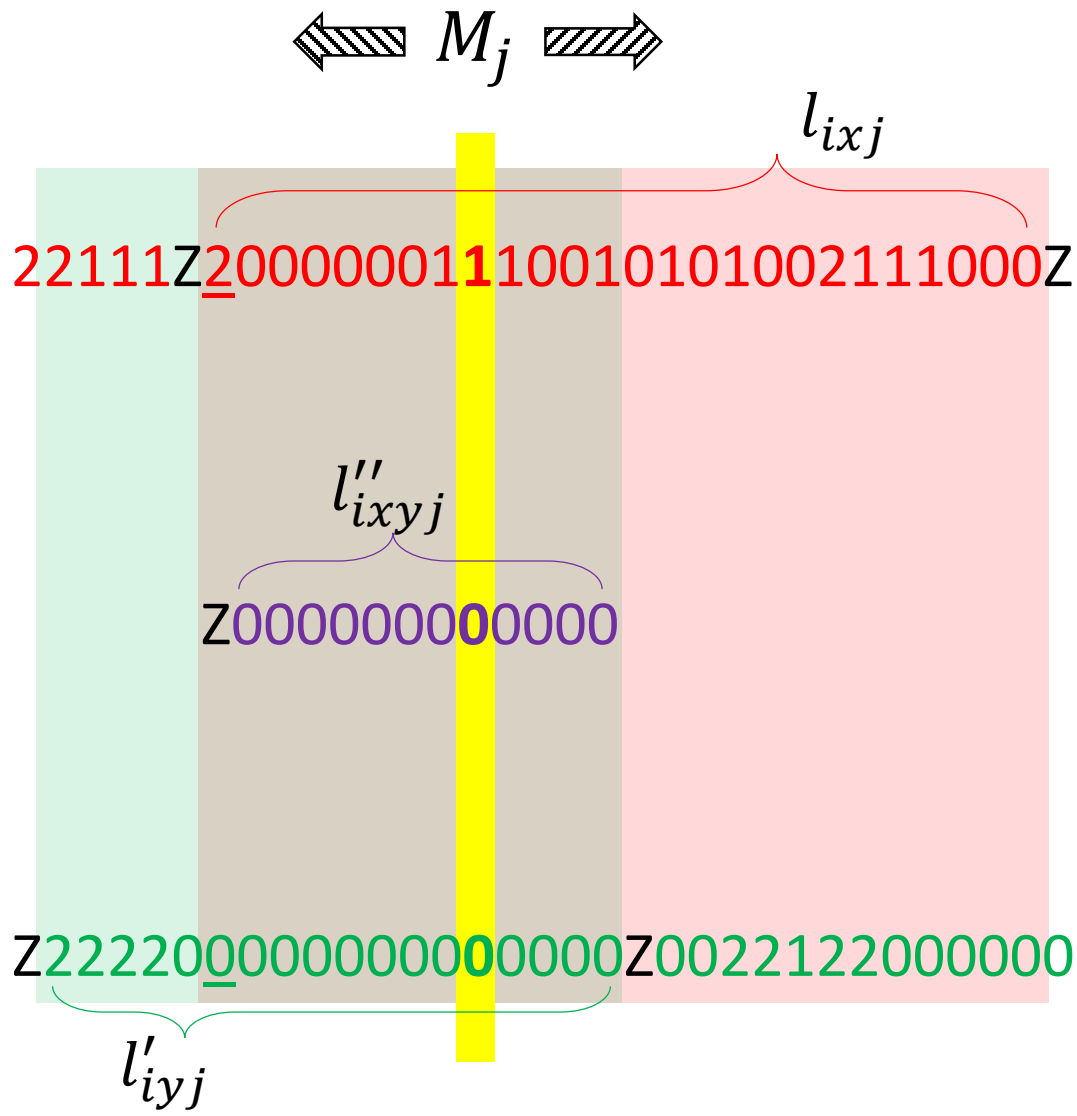
1 - heterozygote
0 - homozygote at major allele
2 - homozygote at minor allele
*Suppose we do not tolerate any mismatch, i.e., we assume no genotyping error


M_j
l_{ixj}
Inferred (pairwise) shared haplotype h_{ixl}
Z2000000110000000200220000Z
Observed g_{xl}
2211122000000011001010100211100002
Observed g_{il}
2211101111111110110000211220000000
2211101111111110000000022122000000
Observed $g_{il'}$
0222200000000000000200221220000000
Observed $g_{yl'}$
Z222200000000000000Z
Inferred (pairwise) shared haplotype $h_{iyl'}$
l'_{iyj}

Inferred haplotype shared by Z_i and Z_x around M_j

Inferred haplotype shared by Z_i , Z_x and Z_y around M_j

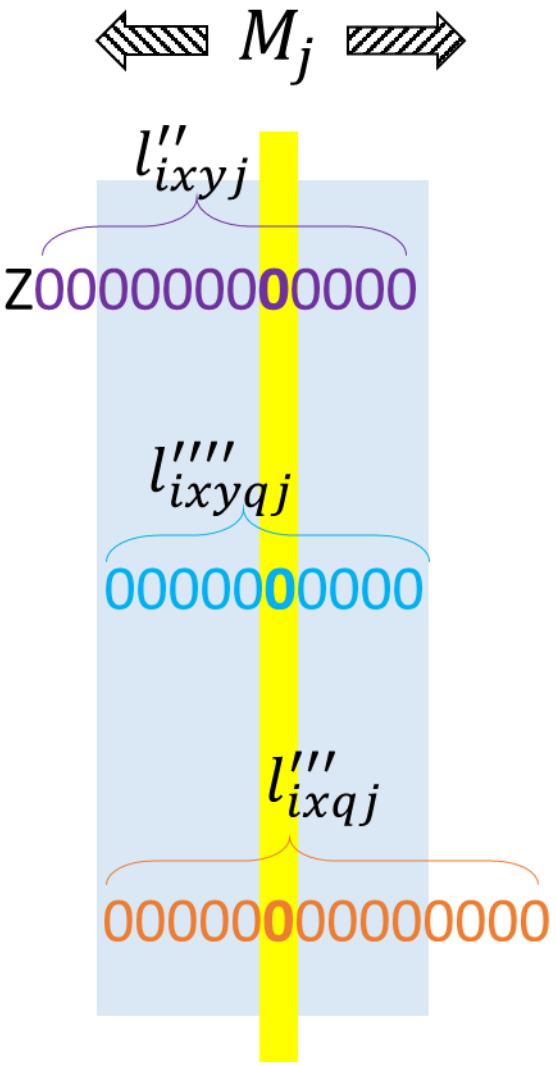
Inferred haplotype shared by Z_i and Z_y around M_j



Inferred haplotype shared by Z_i, Z_x and Z_y around M_j

Inferred haplotype shared by Z_i, Z_x, Z_y and Z_q around M_j

Inferred haplotype shared by Z_i, Z_x and Z_q around M_j



We could go on and on…

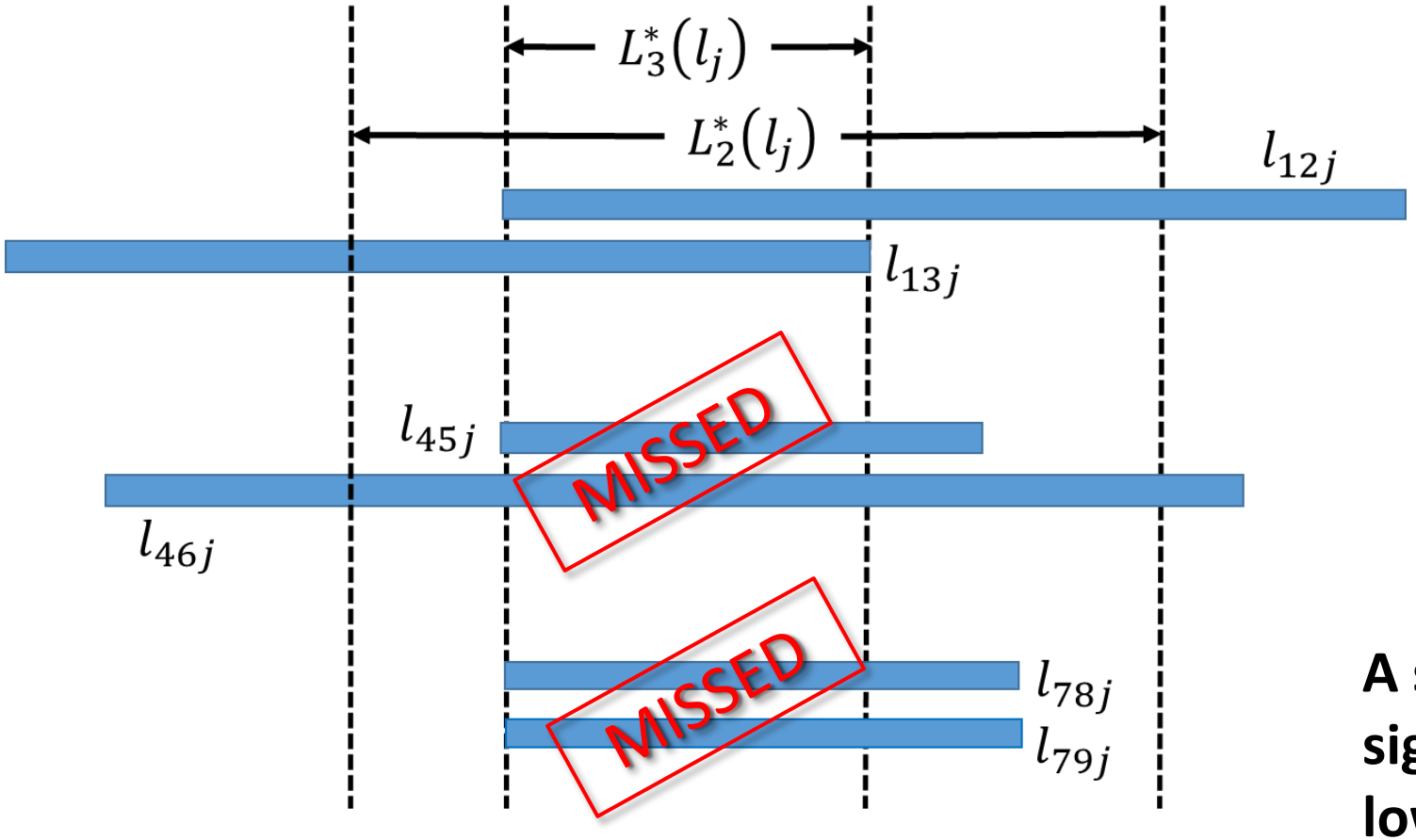
Are these shared segments IBD or IBS?

- CHAT answers this question statistically. The general strategy is to fit a null distribution of IBS sharing and to treat a specific sharing that has a significantly small p value as IBD.
- This question becomes increasingly difficult to answer when we search for higher-order IBD clusters.
  - More computational resources are needed to handle the rapid growing search space: Given *n* subjects, there are $\binom{n}{k}$ candidates for a k-order IBD cluster.
  - Better statistics and null models are needed to determine whether the increasingly short shared segments are IBD or IBS.
- How does CHAT handle these problems?

Incremental search of high-order IBD clusters

- Since an exhaustive search is computationally intensive, CHAT builds higher-order IBD clusters on established lower-order ones.
- In order for a three-subject group to be evaluated, there must be at least two established pairwise IBD relations (2-order IBD clusters) in that group.
  - Suppose l_{ixj} and l'_{iyj} are pairwise IBD segments around Marker M_j whereas l''_{iuj} is not. CHAT only evaluates the significance of l'''_{ixyj} being IBD.
- This strategy limits the search space by following promising paths. It improves efficiency at the cost of possible false negatives.

Suppose sharing is measured by the length of shared segments. Around Marker M_j, $L_2^*(l_j)$ and $L_3^*(l_j)$ is the critical length for detecting pairwise and triple IBD segments respectively. $L_2^*(l_j)$ is arguably longer than $L_3^*(l_j)$.

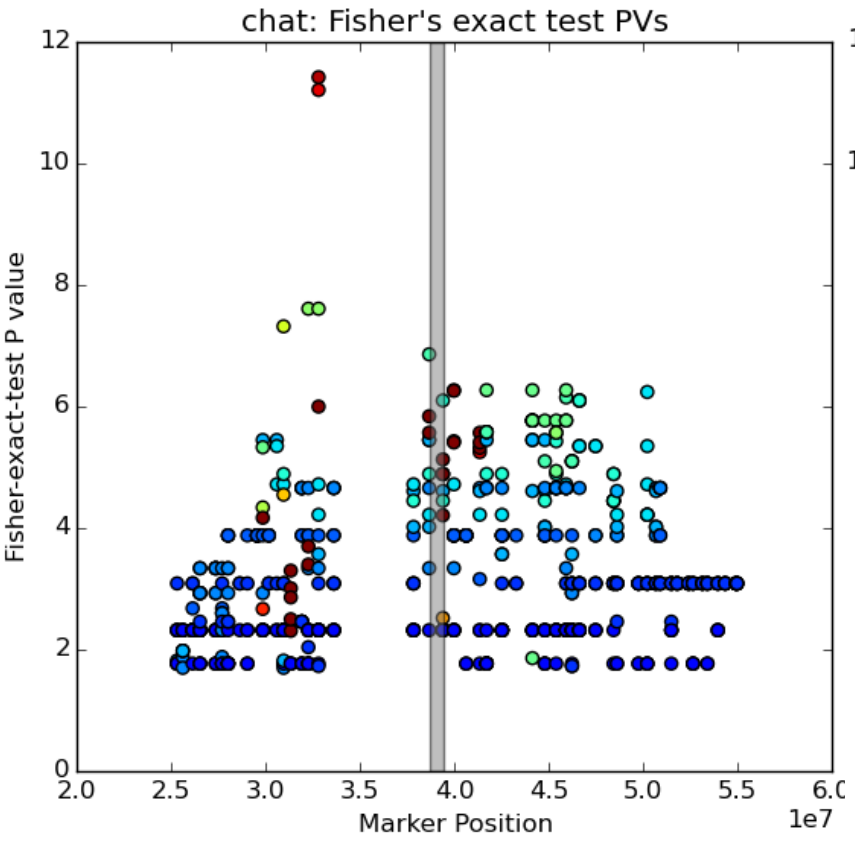


Suppose $\{Z_1, Z_2, Z_3\}$, $\{Z_4, Z_5, Z_6\}$, and $\{Z_7, Z_8, Z_9\}$ are all true IBD trios. l_{45j} , l_{78j}, and l_{79j} will be false negatives under $L_2^*(l_j)$.

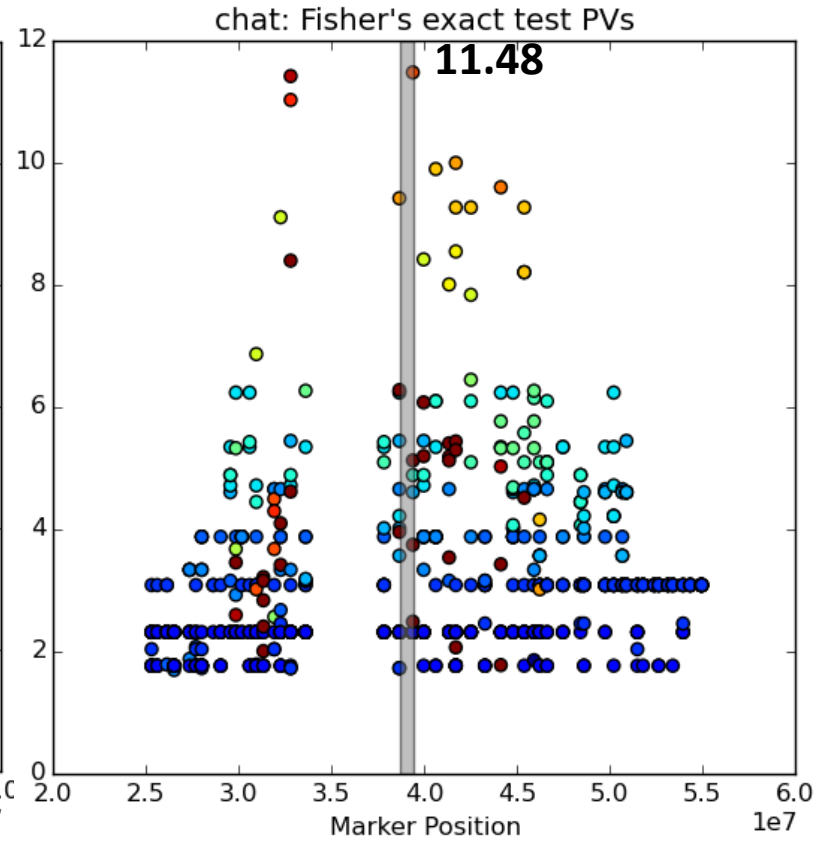
If we build IBD trios on existing IBD pairs, only l_{123j} will be detected, even though $l_{789j} > l_{456j} > l_{123j}$.

A solution is to reduce the significance threshold for selecting lower-order IBD clusters.

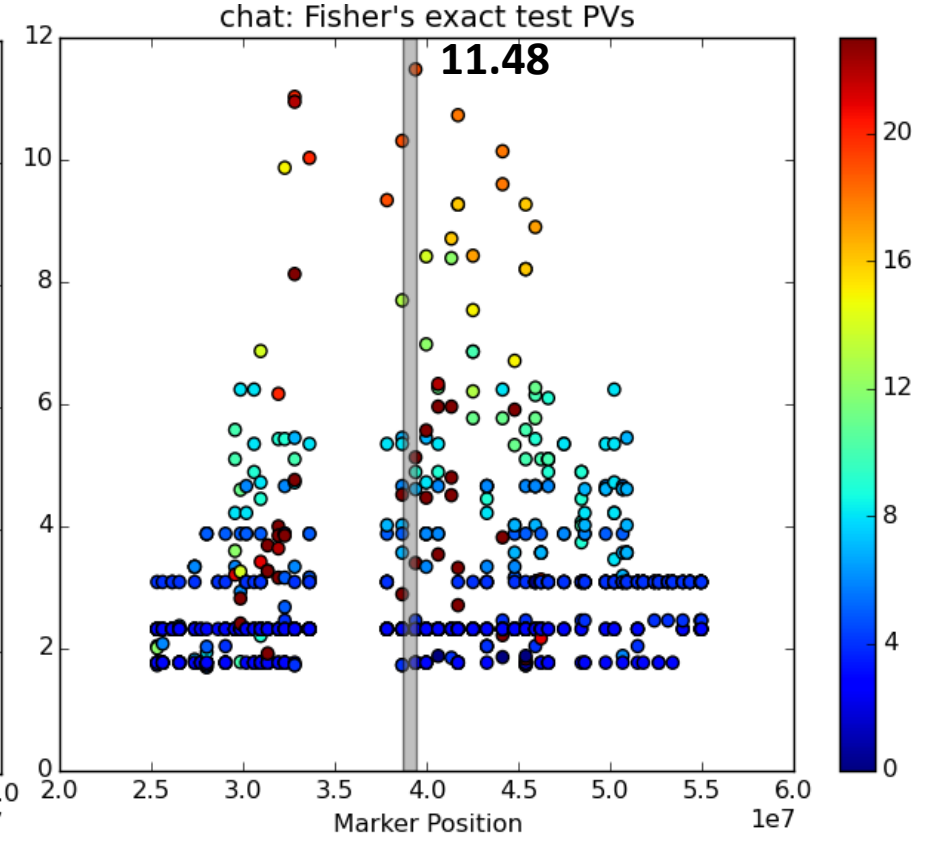
Strategy 1: IBD pairs identified using “Raw LSHs”



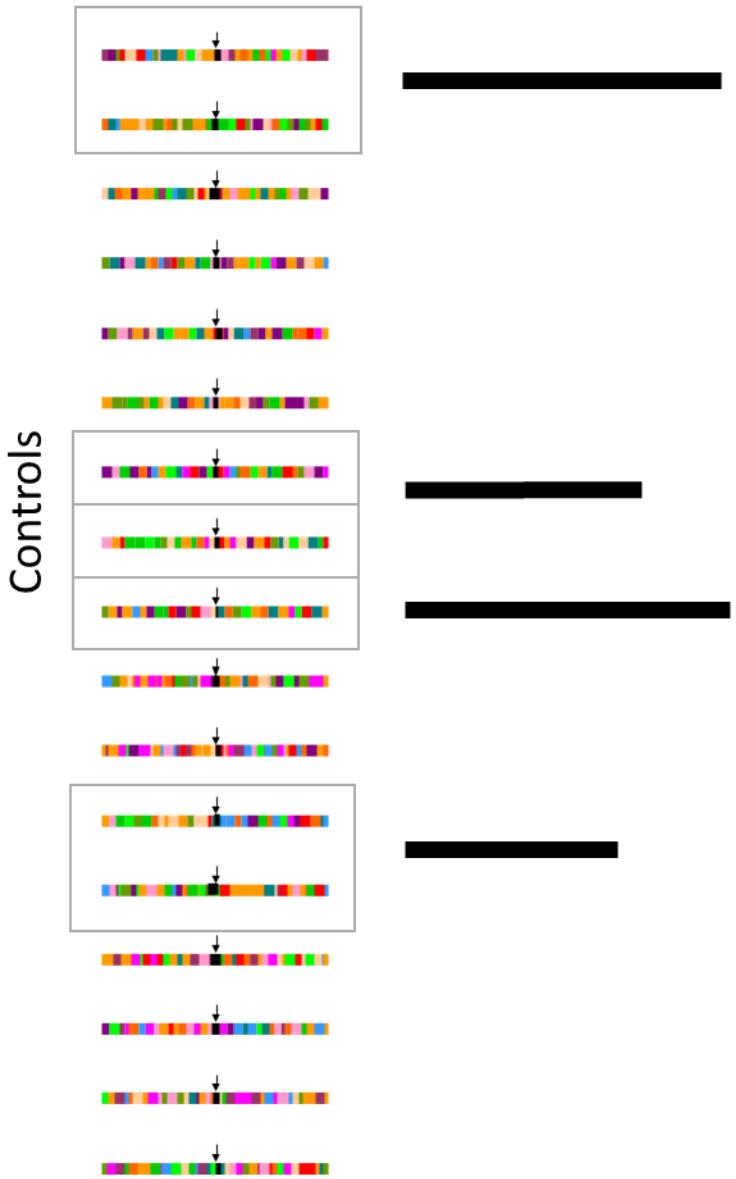
Search IBD trios using subject pairs with IBD probability = **0.9**

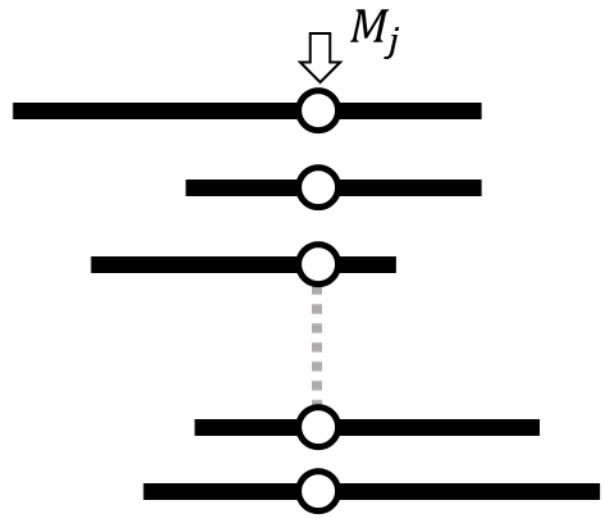


Search IBD trios using subject pairs with IBD probability = **0.5**



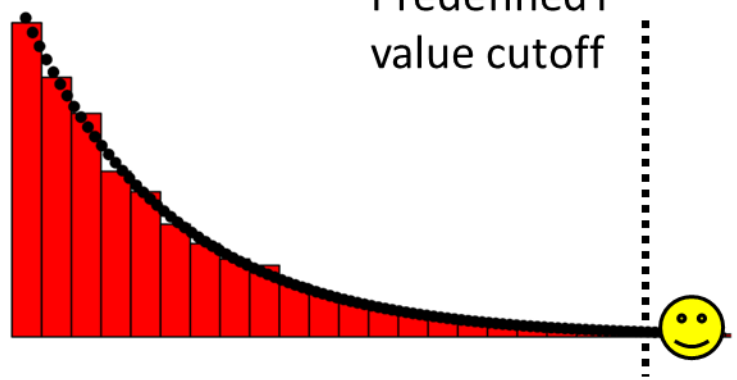
Search IBD trios using subject pairs with IBD probability = **0.1**





Get a sample of pairwise shared segments around M_j by comparing the genotypes of 10,000 randomly selected control pairs

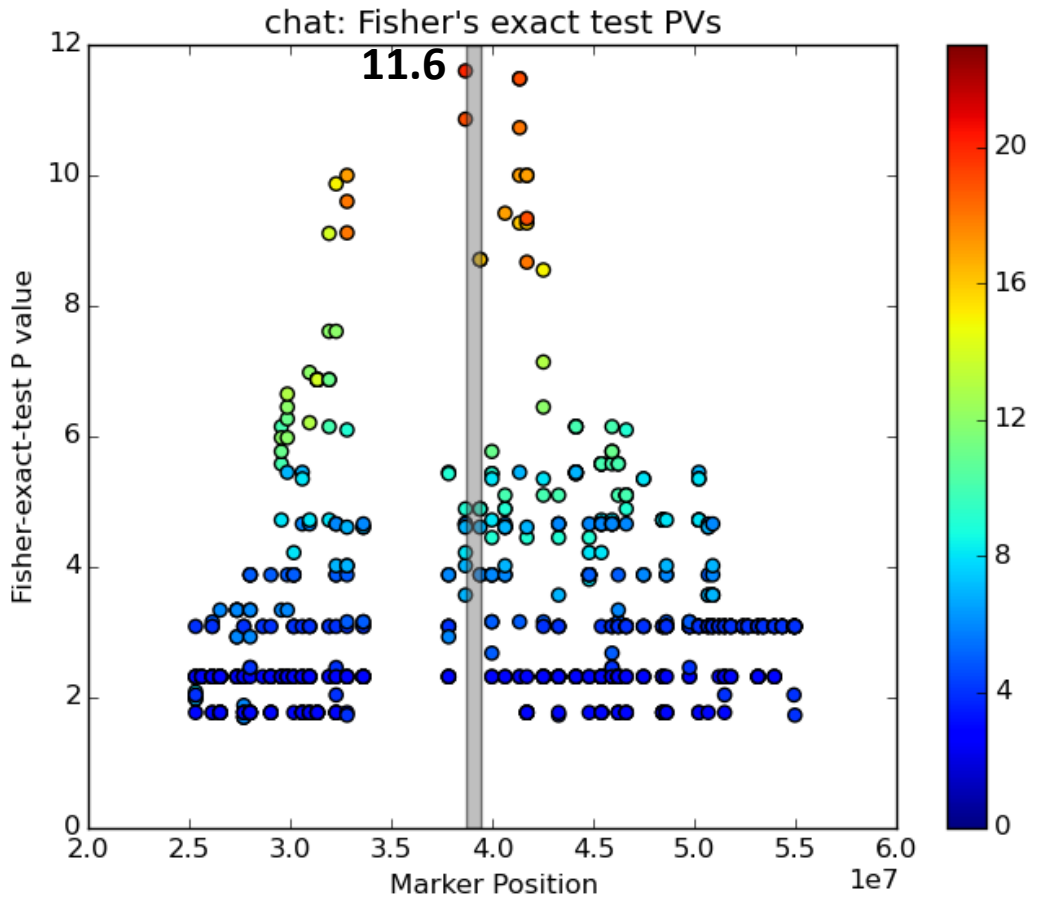
P_{IBS}
(1 - P_{IBD})



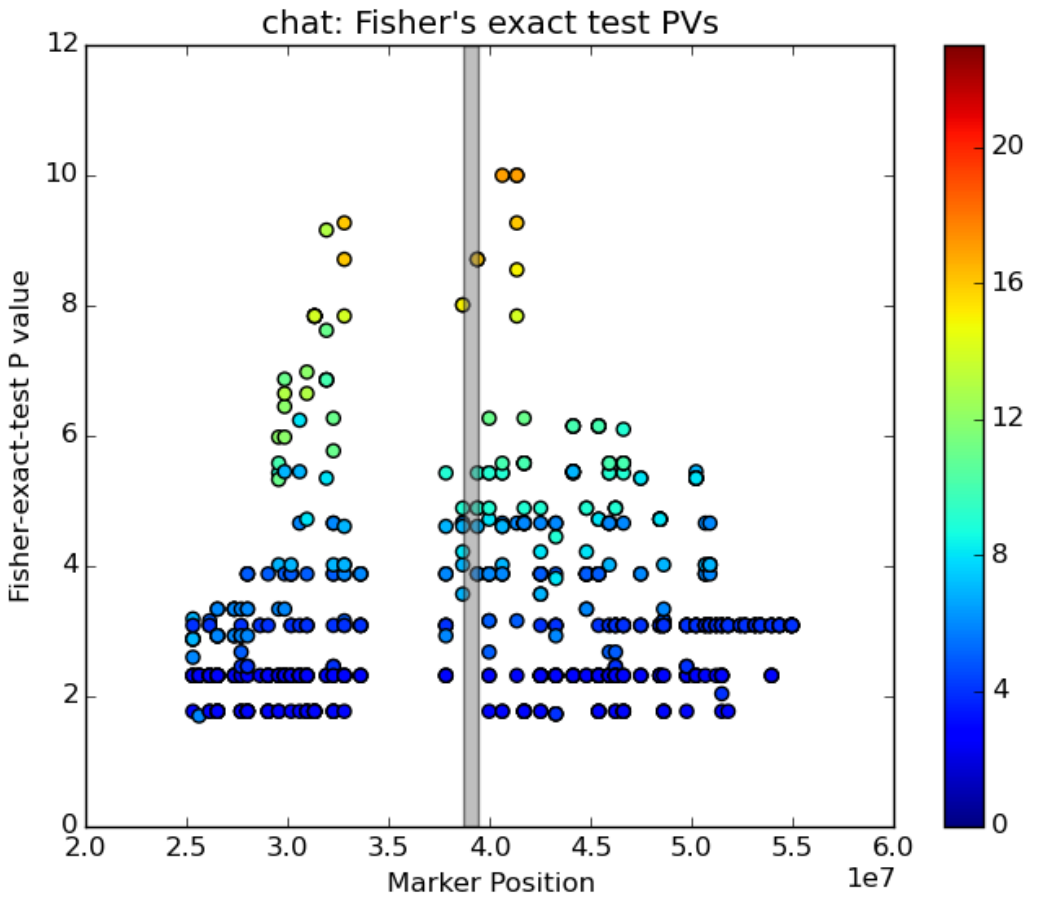
Fit a Gamma distribution using the length of these shared segments

Compare every pairwise shared segment against the distribution. Segments with p <= cutoff P are reported as a **raw LSHs**

Strategy 2: IBD pairs identified using subject-specific distributions of genotype sharing

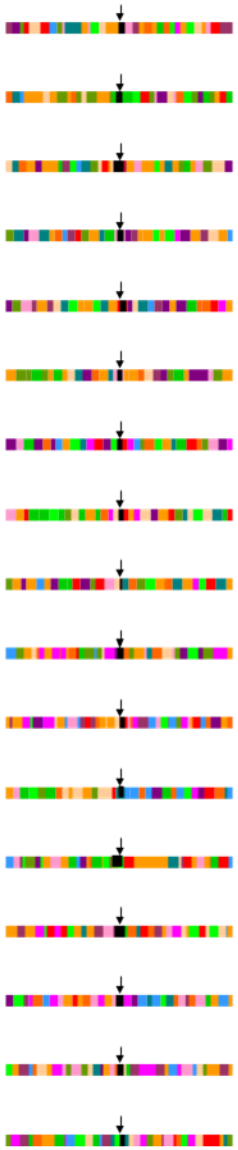


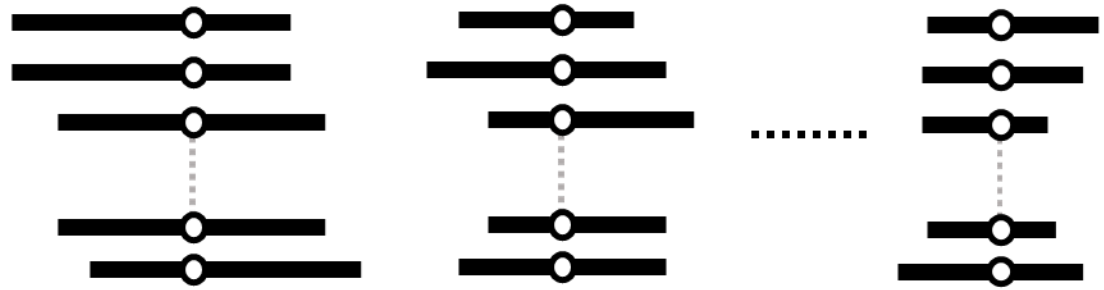
Search IBD trios in IBD pairs determined with threshold = 3



Search IBD trios in IBD pairs determined with threshold = 5 (fewer pairs)

Controls



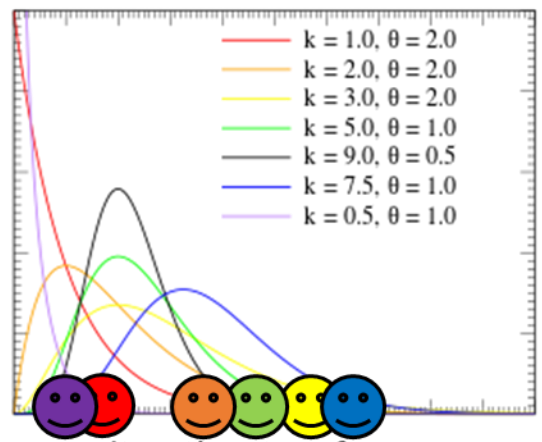


Pairwise shared segments l_{ic}, l_{id},…, l_{ik}

For each subject Z_i, get pairwise shared segments by comparing g_i with every control's genotypes around M_j

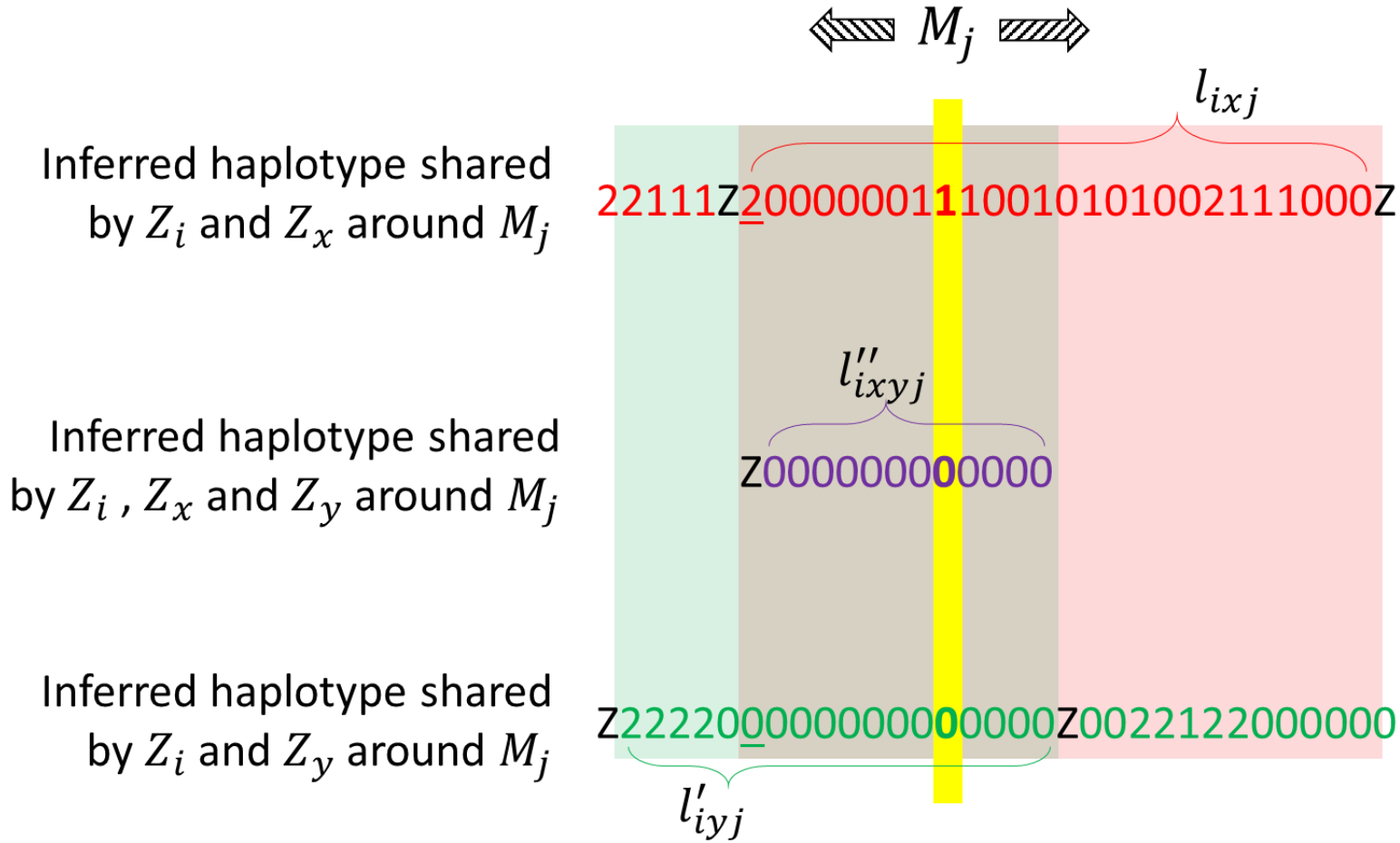
$P_{IBS|Z_i}$

$(1 - P_{IBD|Z_i})$



Fit a Gamma distribution for every subject using the length of that subject's shared segments

A pairwise shared segment, l_{xyj}, is reported as IBD when $-log_{10}\left[P_{IBS|Z_x}(l_{xyj})P_{IBS|Z_y}(l_{xyj})\right]$ exceeds a threshold



Besides short shared segments, CHAT has to detect high-order IBD clusters based on **inferred haplotypes**.

Given a potential triple shared segment l_{ixyj}, CHAT uses a test statistic that is **the maximal value of LD-weighted Pi-SMOR** calculated from the inferred haplotypes in pairwise shared segments l_{ixj} and l_{iyj}.

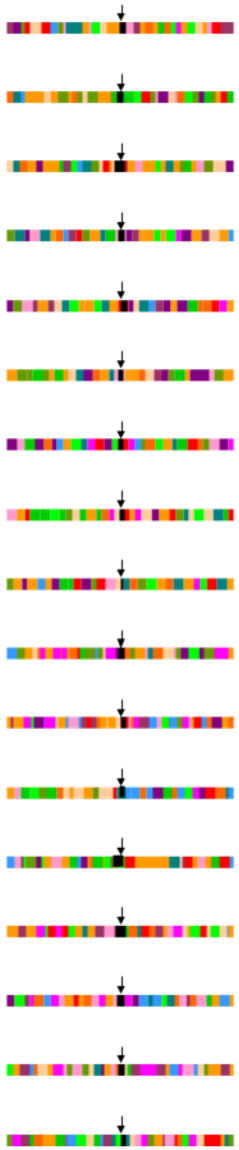
Advanced statistics and null models

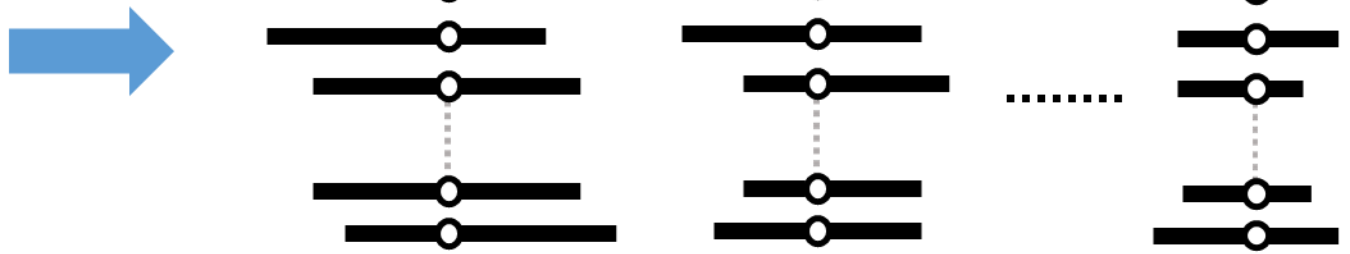
- **Pi-SMOR** is sum of the odds ratio of IBD against IBS at every single marker within the segment. It increases with not only the length a shared segment but also the number of rare variants in that segment, which is also a strong indicator of IBD sharing.
- The contribution of each marker (i.e., single marker odds ratio or SMOR) is adjusted with local LD to avoid over-representing certain regions.

Advanced statistics and null models (cont.)

- Pi-SMOR can be “fooled” by long range of heterozygous sites, which provides some interesting information (the presence of minor alleles) at the cost of high phase uncertainty.
- To avoid overestimating long range of heterozygotes, CHAT evaluates triple shared segment against **subject-specific distributions** of maximal LD-weighted Pi-SMOR.

Controls

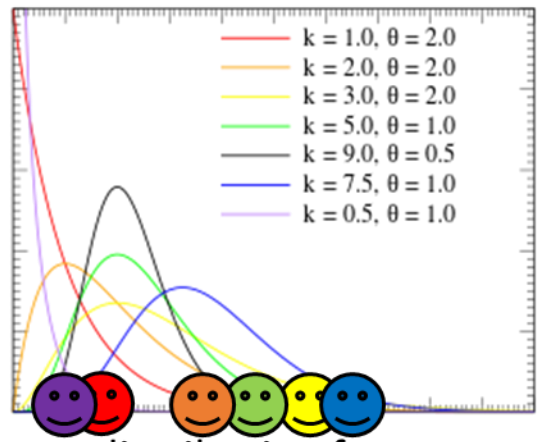




Triple shared segments
$l_{icd}, l_{ief}, \ldots, l_{iuv}$

For each subject Z_i, get triple shared segments by comparing h_{ix} and h_{iy} where {x,y} represents one of 10,000 randomly selected control pairs

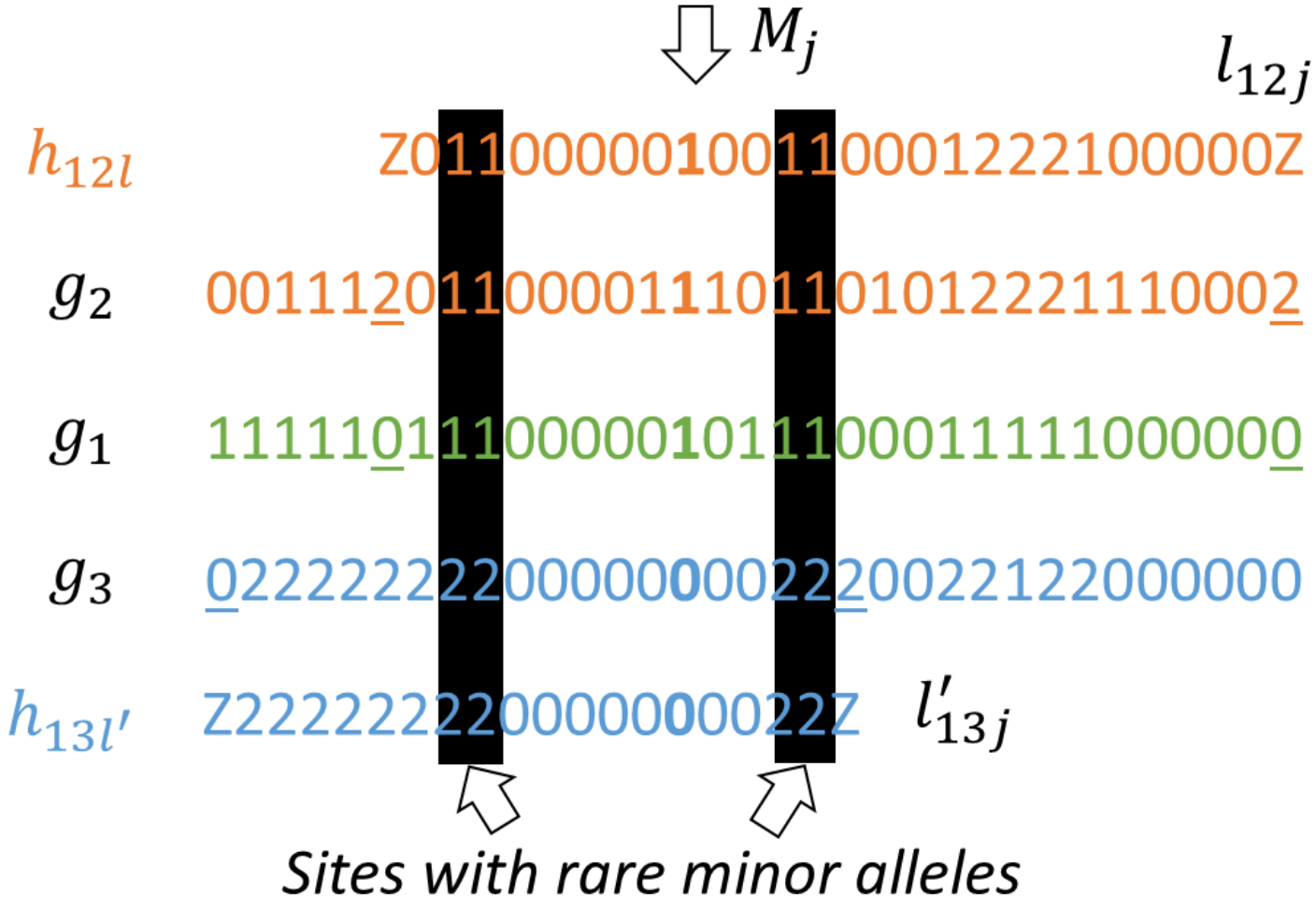
$P_{IBS|Z_i}$
$(1 - P_{IBD|Z_i})$



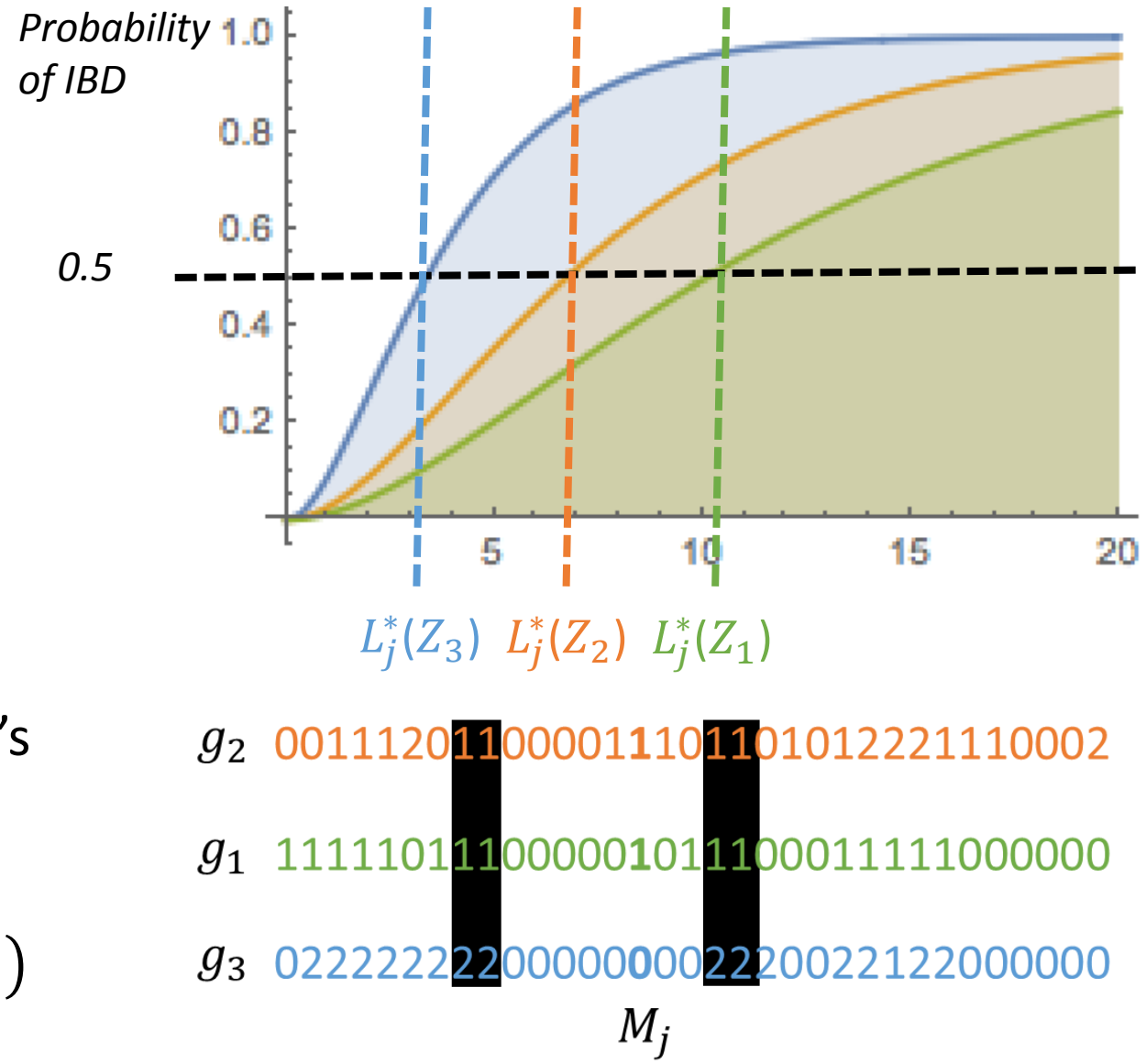
Fit a Gamma distribution for every subject using the length of that subject's shared segments

A triple shared segment, l_{wxyj}, is reported as IBD when the genometric mean of $-log_{10}[P_{IBS|Z_w}(l_{wxyj})]$, $-log_{10}[P_{IBS|Z_x}(l_{wxyj})]$, and $-log_{10}\left[P_{IBS|Z_y}(l_{wxyj})\right]$ exceeds a threshold

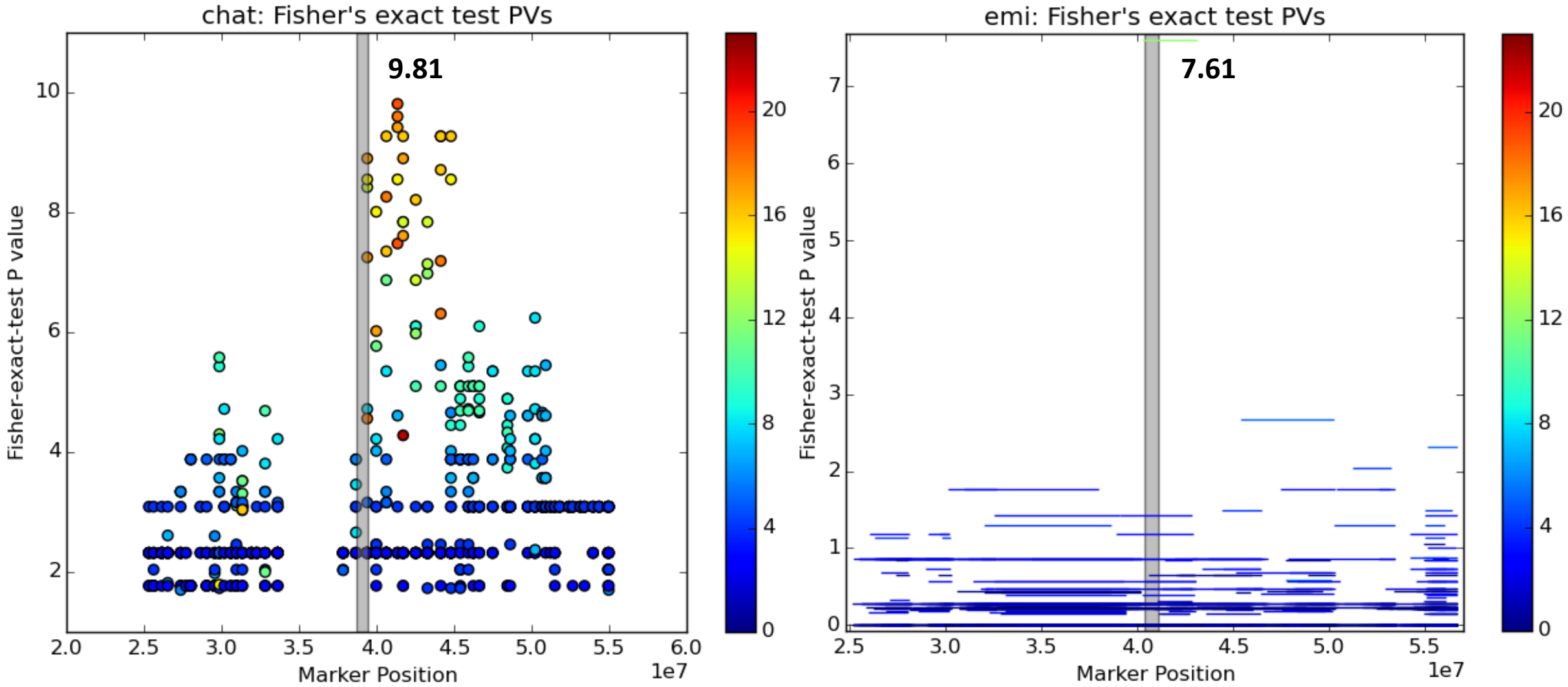
Two pairwise shared segments l_{12j} and l'_{13j} span 26 and 18 consecutive markers respectively. Which one do you think is more likely to be IBD?



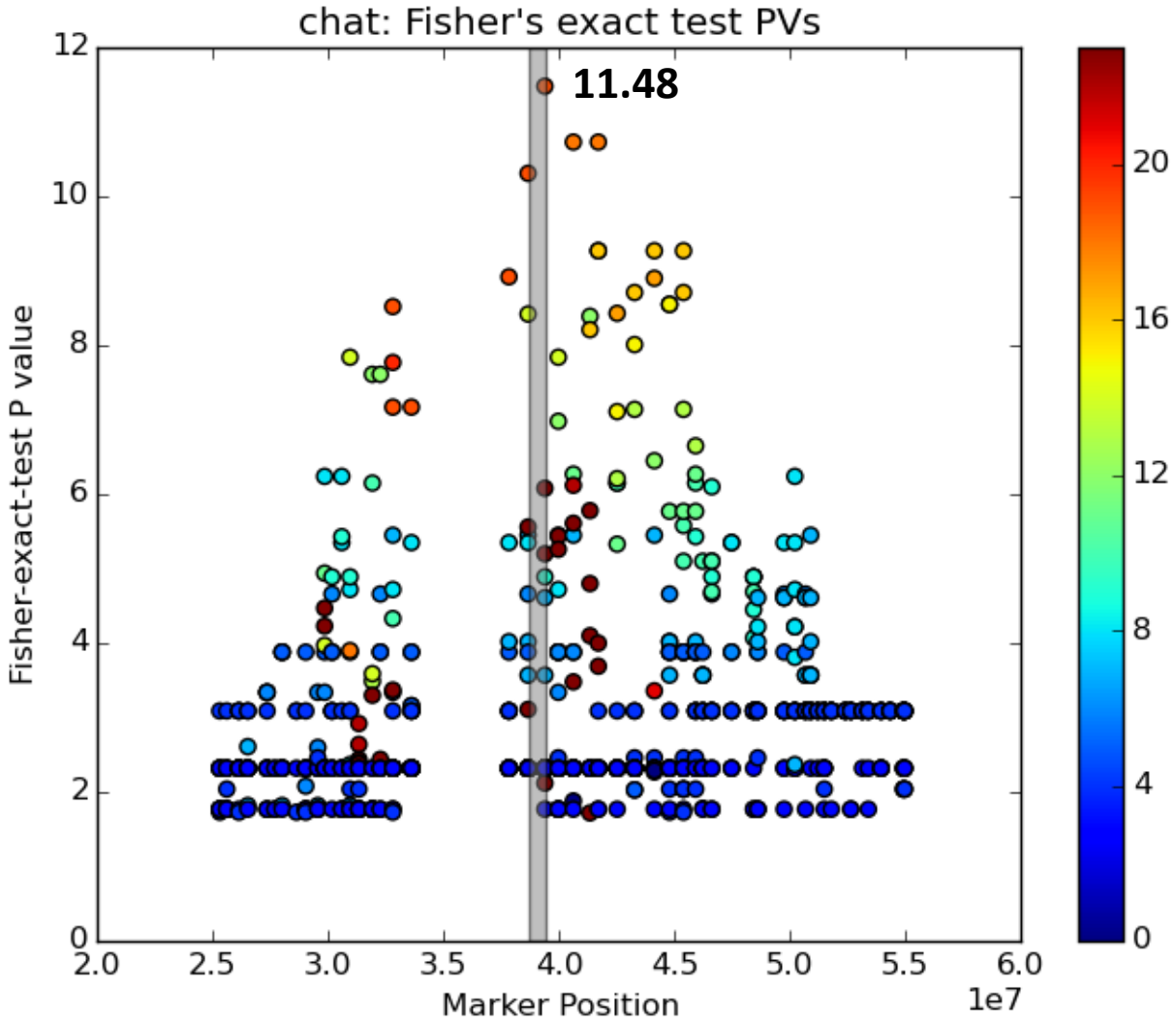
- Given genotypes g_1, g_2 and g_3, the average length of pairwise sharing between Z_1 and any other subject in the dataset around Marker M_j is probably greater than that for Z_2 and both greater than that for Z_3.
- The critical length (L_j^*) for determining whether a shared segment around M_j is IBD or IBS could vary with subjects.
- In general, L_j^* increases with the number of consecutive common SNPs in a subject's genotypes around M_j.
- In this case, it is likely that $P_{IBD}(l'_{13j}|g_3) > P_{IBD}(l_{12j}|g_2) > P_{IBD}(l_{12j}|g_1) > P_{IBD}(l'_{13j}|g_1)$

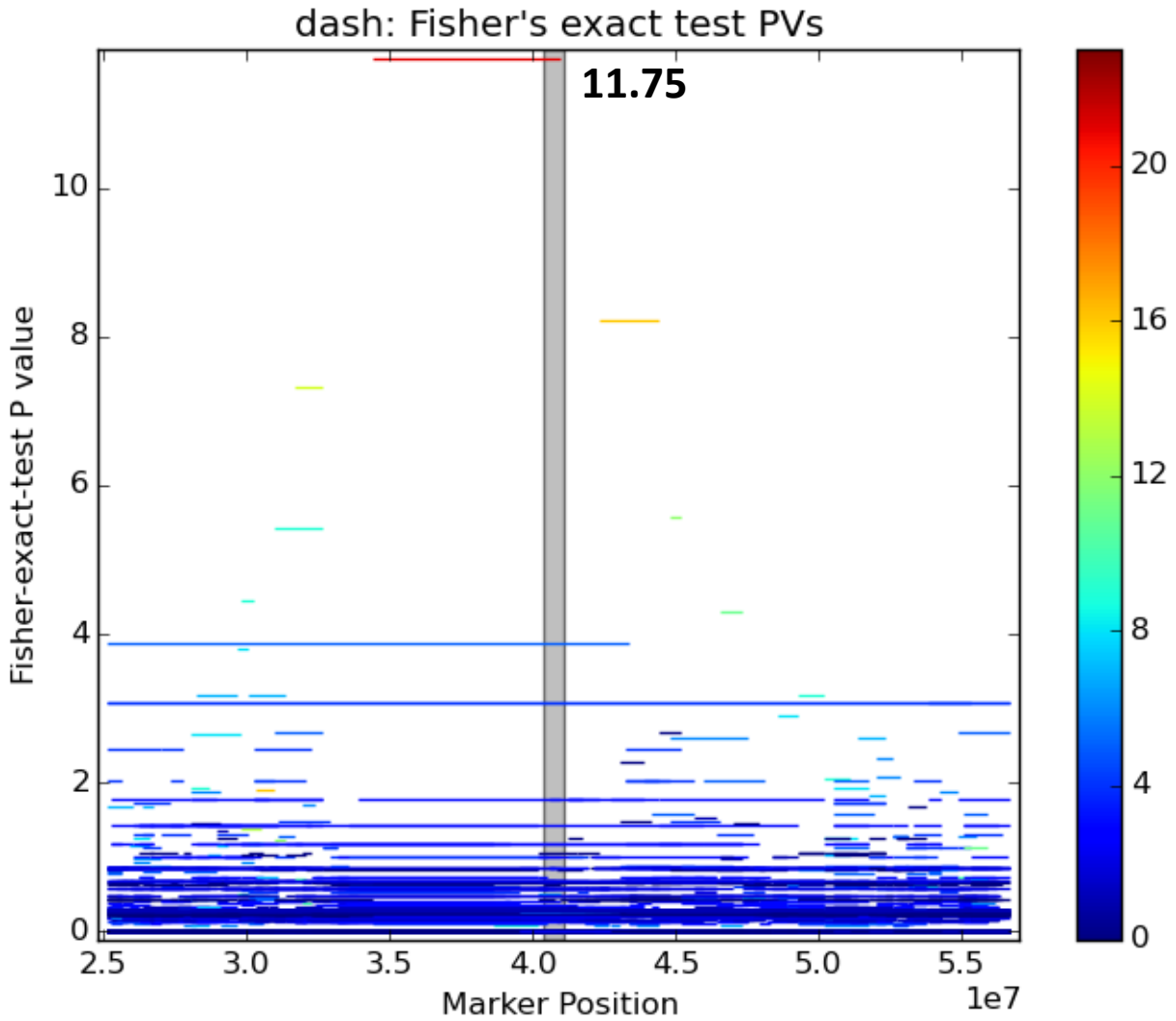


CHAT Vs EMI: using IBD pairs detected by Refined IBD



CHAT Vs DASH: using IBD pairs detected by Germline

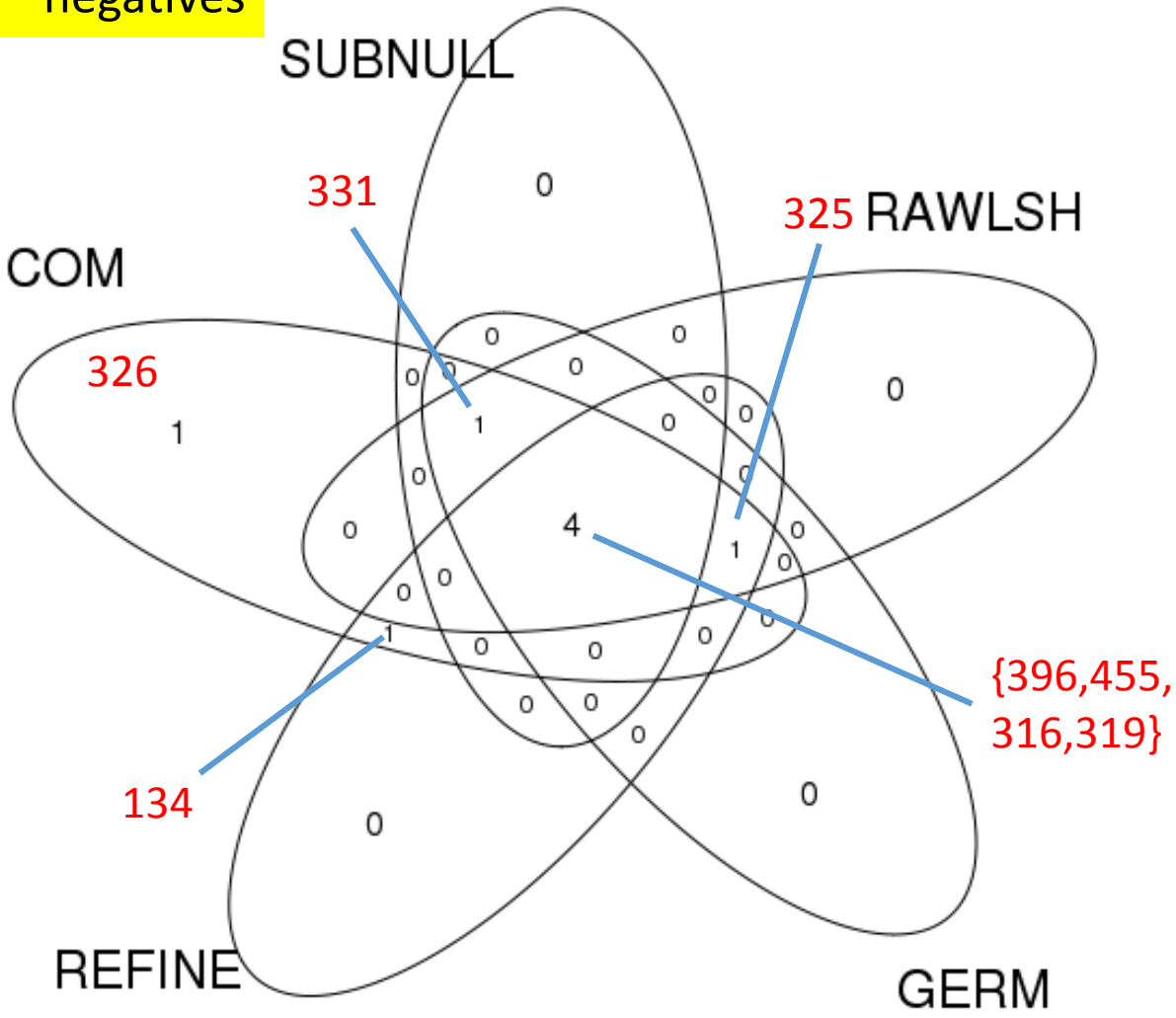




Lrrk2 mutation carriers that tend to be excluded from the finally detected IBD cluster

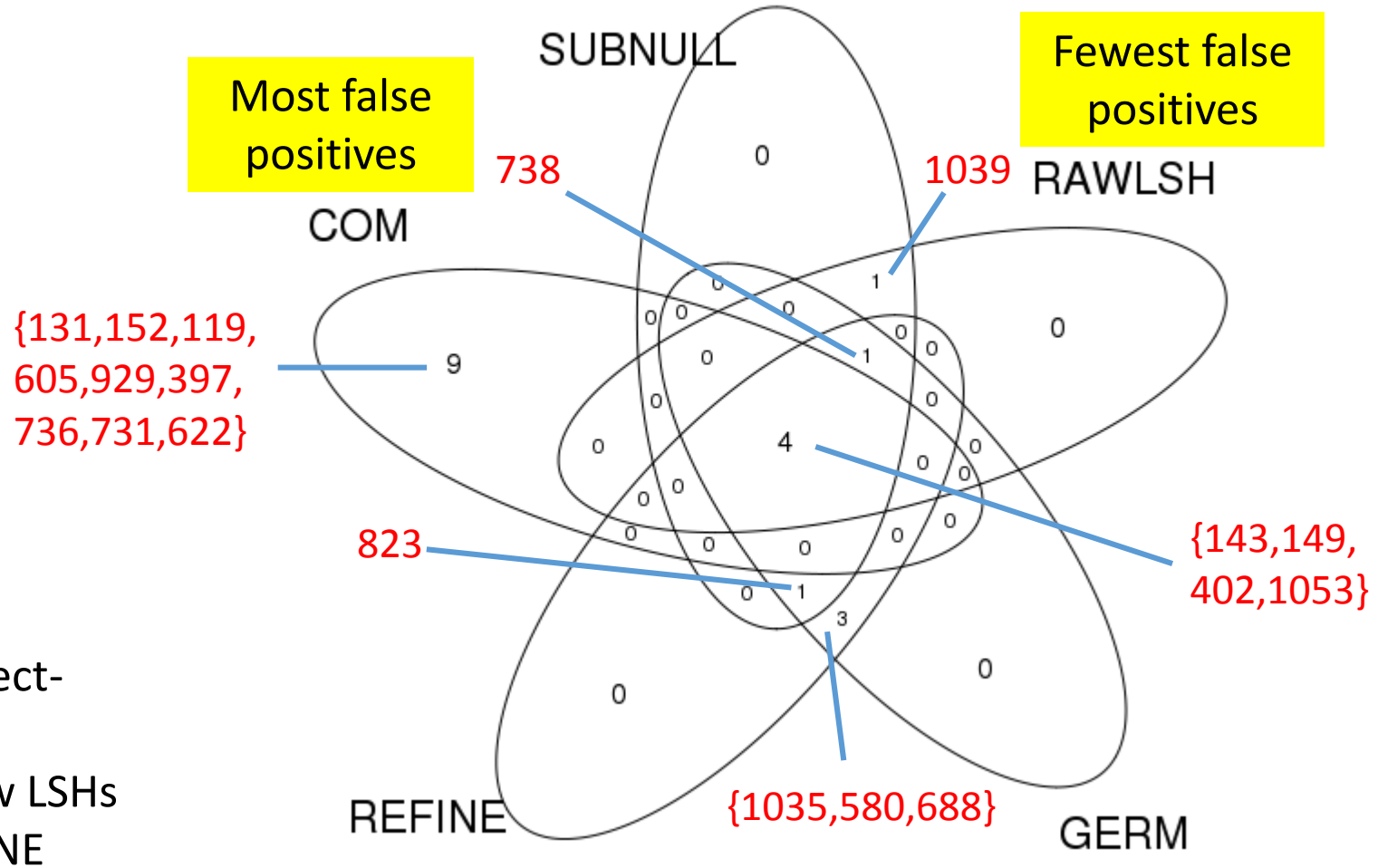
Fewest false negatives

Most false negatives



COM: Exhaustive search
SUBNULL: IBD pairs identified via subject-specific null distributions
RAWLSH: IBD pairs identified from Raw LSHs
GERM: IBD pairs identified by GERMLINE
REFINE: IBD pairs identified by Refined IBD

Subjects that are not known Lrrk2 mutation carriers yet tend to be included in the finally detected IBD cluster



COM: Exhaustive search
SUBNULL: IBD pairs identified via subject-specific null distributions
RAWLSH: IBD pairs identified from Raw LSHs
GERM: IBD pairs identified by GERMLINE
REFINE: IBD pairs identified by Refined IBD

Efficiency

- Using raw LSHs as the sources of IBD pairs seems most efficient regarding CHAT's workload and the final result.

CHAT's workload in building IBD trios on IBD pairs from different sources (measured by the number of comparison jobs CHAT needs to do)

Raw LSH p=0.1	Raw LSH p=0.3	Raw LSH p=0.9	Subject GT thresh=0.3	Subject GT thresh=0.5	Germline	Refined IBD
5219	5206	5200	5414	5399	5543	5305

Efficiency (cont.)

- Fitting subject-specific distributions of sharing is time demanding.
- When the maximal LD-weighted Pi-SMOR is the test statistic, fitting subject-specific distributions does not provide useful null models for genotype sharing as it does for haplotype sharing.

Thank you!
Questions?