Graph-based IBD clustering and implications for CHAT

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

IBD segment detection

- Identity-by-descent (IBD) segments are chromosomal regions where two or more individuals inherit identical nucleotide sequences from the same most recent common ancestor (MRCA).
- Detected IBD segments have various downstream applications (Browning & Browning, 2012)
  - For example, IBD mapping, i.e., using detected IBD segments to determine genomic regions likely to harbor rare disease-susceptibility variants

Pairwise vs. multiway IBD detection

- Pairwise detection compares the similarity of two haplotypes at a time, while multiway detection compares multiple haplotypes simultaneously.
- To the end of IBD mapping, multiway detection may be advantageous over pairwise detection in identifying short IBD regions (< 1cM).
  - Existing pairwise methods mostly estimate IBD probability based on the length of sharing, so IBD segments must be long enough to be detectable.
  - Consider haplotypes *A, B* and *C* in a region not interrupted by recombination. If *A* and *B* are IBD and *A* and *C* are IBD, then by definition *B* and *C* are also IBD even though their IBD relation is undetectable via pairwise comparison.

Multiway detection: Solving conflicts in multiple IBD relations

- Essentially multiway detection needs to solve conflicting pairwise detection results, for example, haplotypes *A* and *B* are IBD, *A* and *C* are IBD, but *B* and *C* are not IBD.
- How to solve a conflict is not always straightforward, as IBD relations are estimated.
  - What would you do, if *A* and *B* are IBD with 81.2% probability, *A* and *C* are IBD with 49.8% probability, B and C are IBD with 13.2%?
- IBD relations are estimated independently.
- Shorter regions are subject to more uncertainty.

Multiway detection: Solving conflicts in multiple IBD relations (cont.)

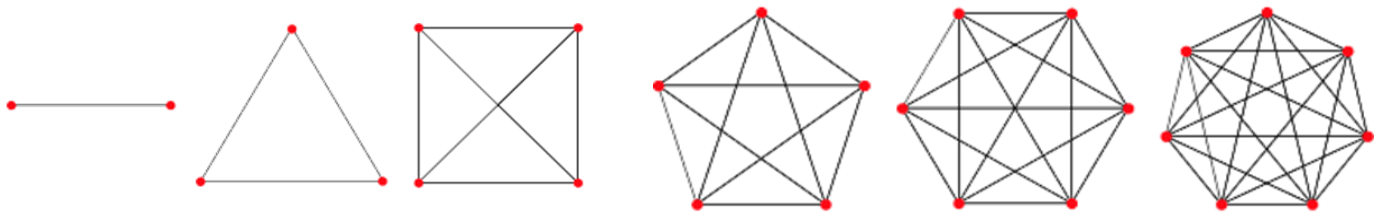
- Some multiway methods relies on building a global probabilistic framework and they gain accuracy at the expense of computational efficiency.
  - ***MCMC IBD finder*** (Moltke et al., 2011) detects IBD segments directly from multiple individuals. It needs tens of hours to process even small datasets (e.g., 20-30 individuals with 500 SNPs).
  - ***IBD-Groupon*** (He, 2013) and ***PIGS*** (Park et al., 2015) build on pairwise comparison results to improve efficiency, but still, IBD-Groupon cannot handle GWAS-scale datasets (e.g., 4000 haplotypes with 3000 SNPs) and PIGS uses sampling results to approximate exact results in practice.

Multiway detection: Solving conflicts in multiple IBD relations (cont.)

- Some methods utilize graph-based techniques.
  - IBD is not only about sequence similarity but also about the genealogical relationships among individuals.
  - Graph provides a global view of (pairwise) relational data.
- A graph is a set of linked nodes. The links (edges) can have weights.
- Given a set of *N* haplotypes at a locus**,** a typical **IBD graph** contains *N* nodes each representing a haplotype. Two nodes are linked if their corresponding haplotypes are IBD at the locus.

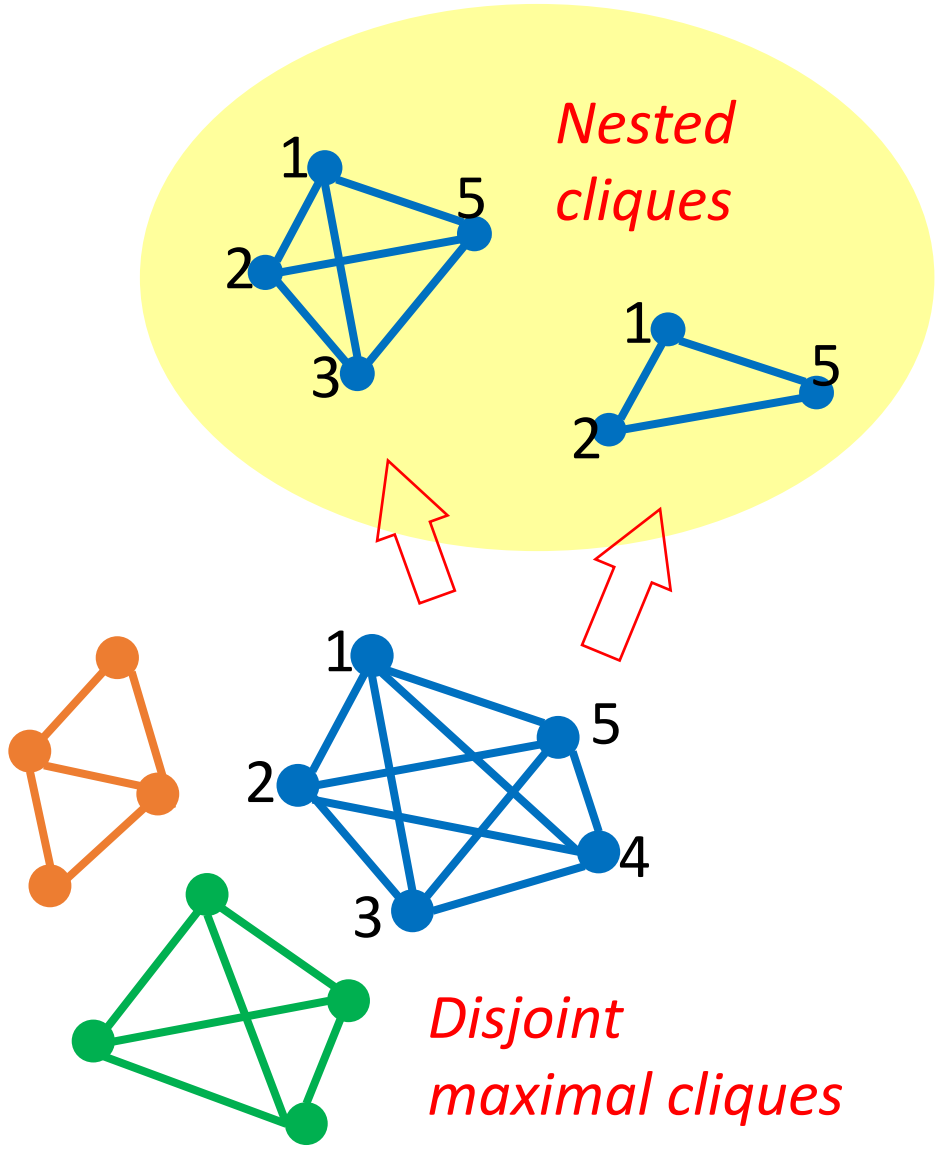
A graph model of IBD

- A key concept is the transitivity of IBD relations at a locus: If A $\overset{IBD}{\longleftrightarrow}$ B and A $\overset{IBD}{\longleftrightarrow}$ C, then B $\overset{IBD}{\longleftrightarrow}$ C. It means in the corresponding (undirected) graph, if Edge AB and AC exist, BC should also exist.
- Given that all and only true IBD relations were identified, every connected component of the corresponding IBD graph should be a ***clique***.



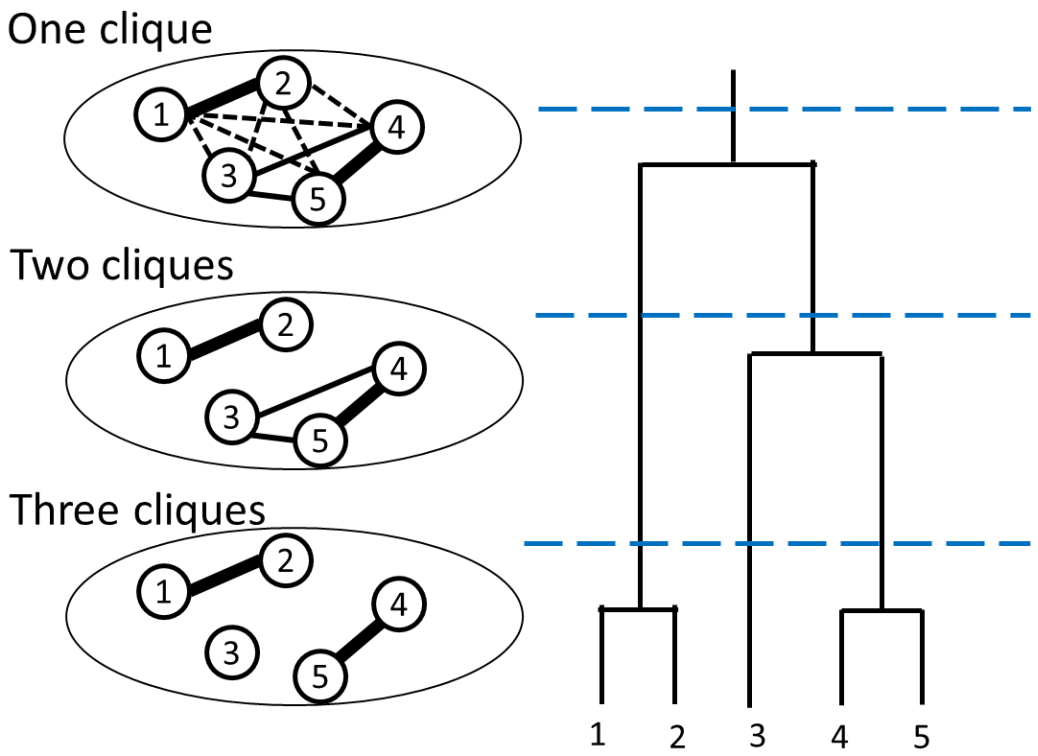
A graph model of IBD (cont.)

- Each clique represents a cluster of IBD haplotypes at a locus that have the same MRCA.
- If there are multiple such *IBD clusters* at the locus, the IBD graph will contain multiple disjoint cliques.
- Finding all IBD haplotype clusters is equivalent to finding all **maximal cliques** in an IBD graph.



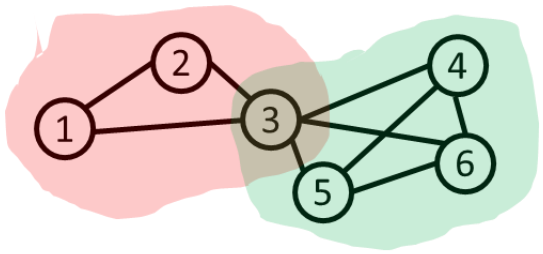
A graph model of IBD (cont.)

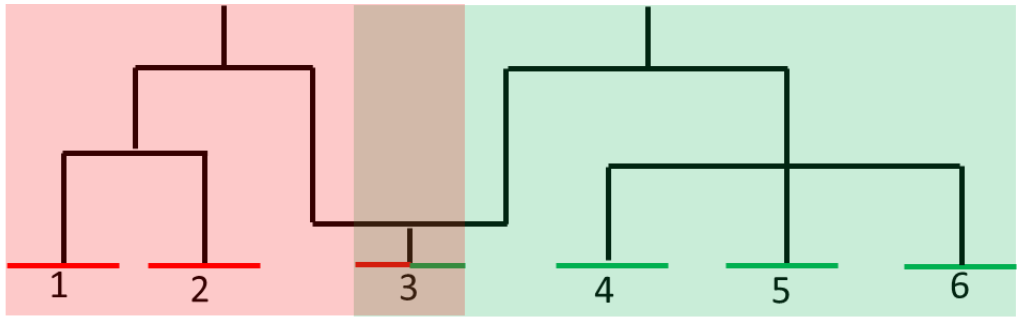
- The nested nature of cliques provides a way to model the hierarchical relations of IBD clusters.
- Depending on the generation of interest, the IBD graph of a genomic site contains either one clique or multiple disjoint cliques (assuming error-free IBD detection).



A graph model of IBD (cont.)

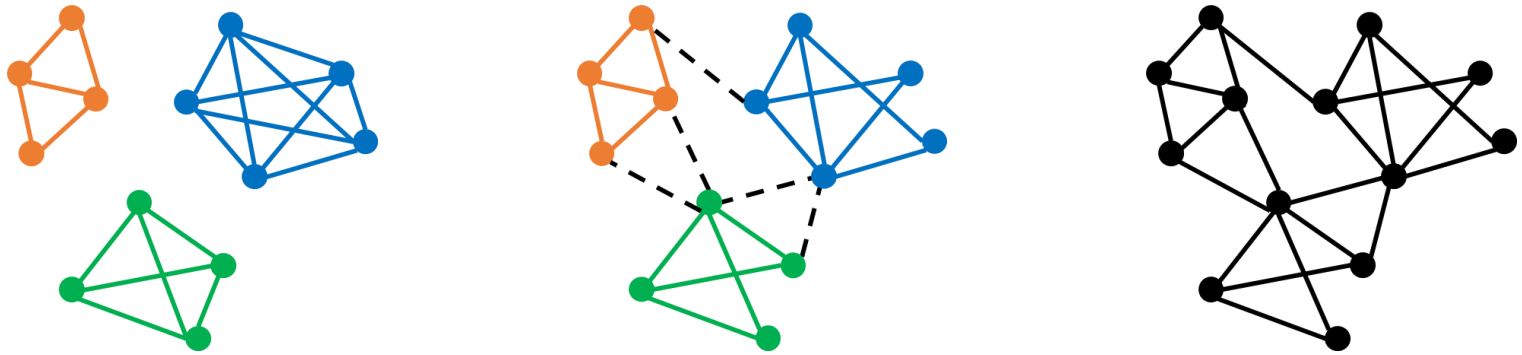
- While the genealogy at a genomic site corresponds to a coalescent tree, the genealogy of a genomic region with possible recombination may contain multiple trees.
- Thus, there can be overlapping cliques in the IBD graph of such a genomic region.





A graph model of IBD (cont.)

- Inaccurate or ambiguous pairwise detection results are missing or misplaced edges that lead to incomplete subgraphs in an IBD graph.
- Identifying true IBD clusters thus involves reconstructing clique(s) or dense enough subgraph(s), by removing/adding links that presumably represent false-positive/false-negative IBD relations.
- The problem is we don’t know the right answer.



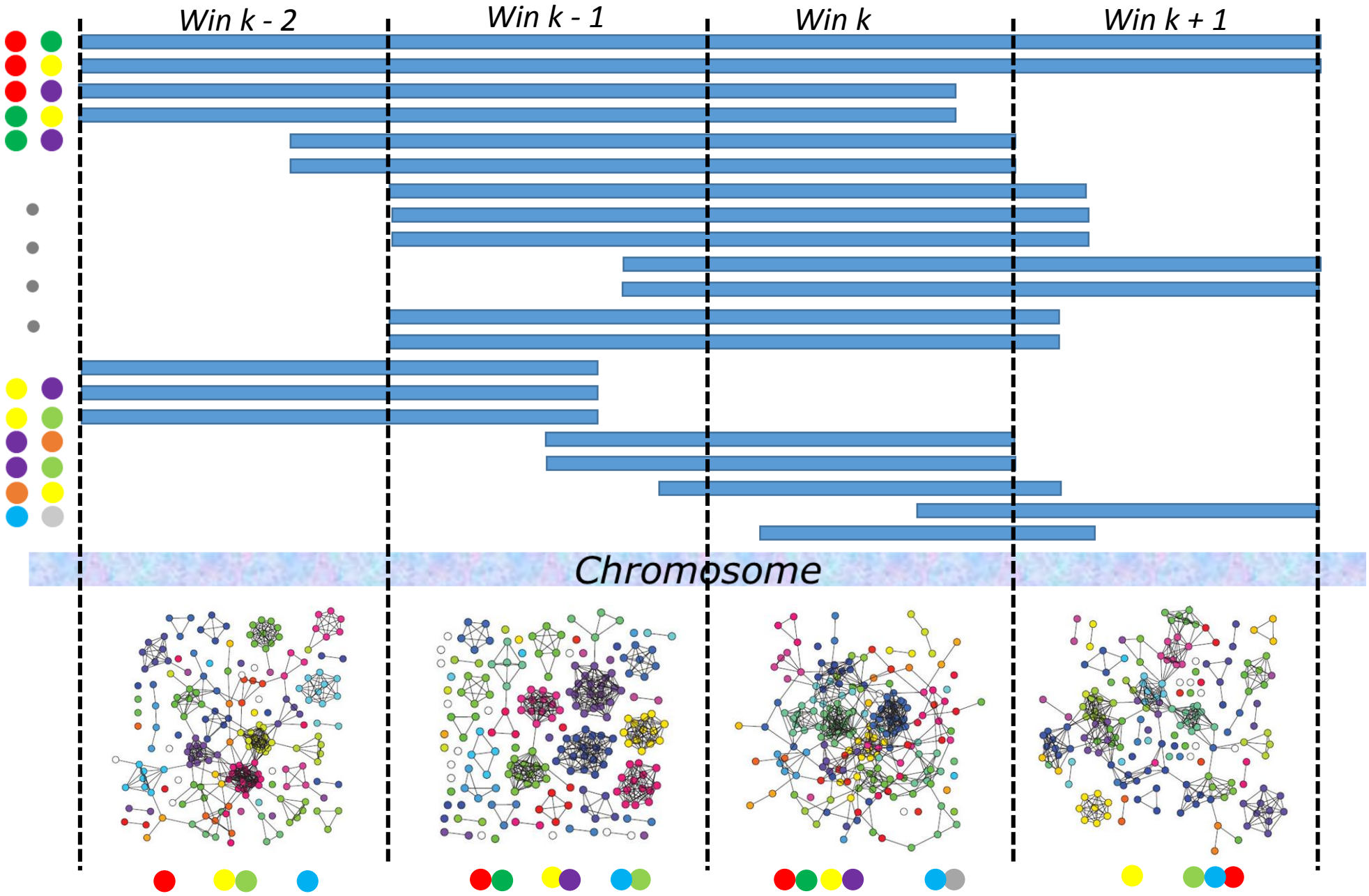
A graph model of IBD (cont.)

- However, if the errors are not pervasive, intra-clique areas should still be denser than inter-clique areas.
- Moreover, if we define the probability of an IBD relation (or some correlated measures) as the weight of the corresponding link, intra-clique links should be stronger than inter-clique links.
- Thus, **structural properties of the IBD graph** such as density and tie strength hint true cliques, even if the graph is “noisy” and “fuzzy”.

Graph-based multiway IBD detection

- Partition a graph based on its structural properties is an active research area known as community detection or graph clustering (Fortunato, 2010).
- Existing graph-based multiway IBD detection methods more or less borrow ideas/techniques from this area.
- ***DASH*** (Gusev et al., 2011) and ***EMI*** (Qian et al., 2014) are the only two methods that can handle genome-wide data of thousands of individuals. They are both graph-based methods.
- *IBD-Groupon* and *PIGS* also leverage graph representations of IBD.

DASH & *EMI*: similar workflow

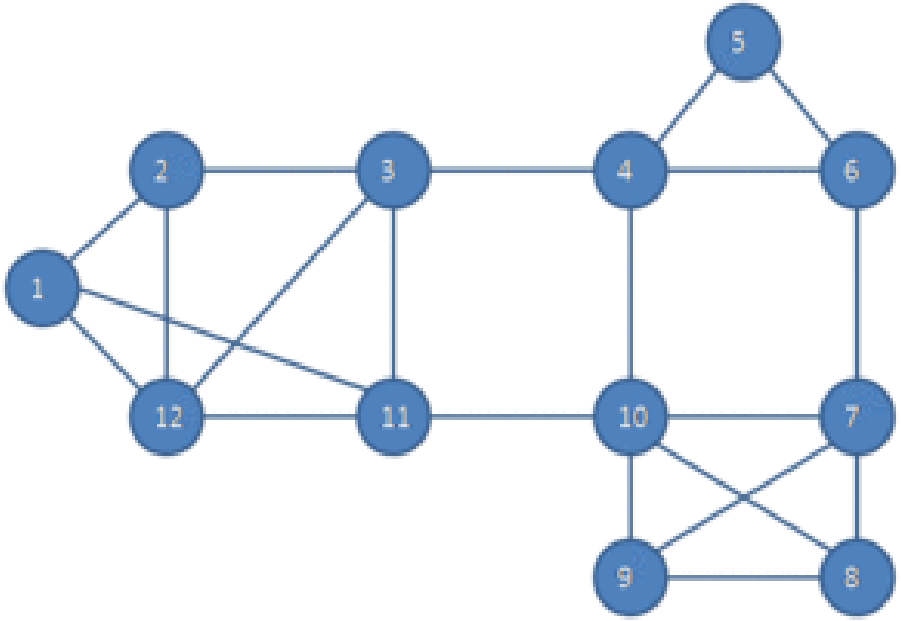


DASH vs. EMI: different clustering strategies

- As mentioned, IBD graphs have an intrinsic hierarchical structure.
- ***DASH*** works downwards the hierarchy. It starts from the largest connected component and divides the component into smaller and denser subgraphs, by removing putative false-positive edges.
- In contrast, ***EMI*** works upwards the hierarchy. It starts from two seed nodes and repeatedly adds qualified adjacent nodes to expand a dense subgraph.

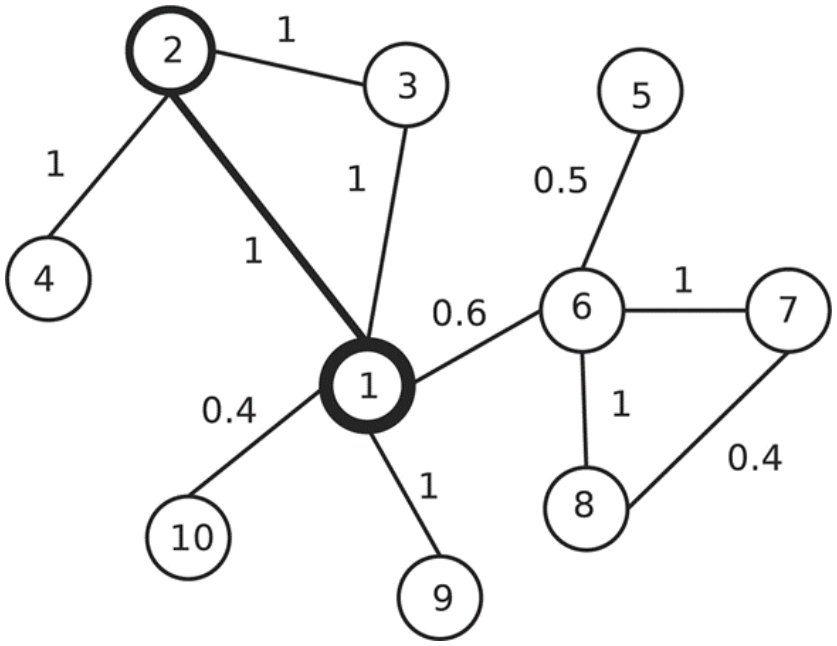
HCS (**H**ighly **C**onnected **S**ubgraphs) Clustering

- Hartuv & Shamir (2000)
- Work on similarity graphs
- Start from the global graph and identify minimum cut recursively
- Stop dividing a subgraph when its density reaches a predefined upper bound or when it has no links to cut.
- This algorithm is efficient when the graph is highly connected.

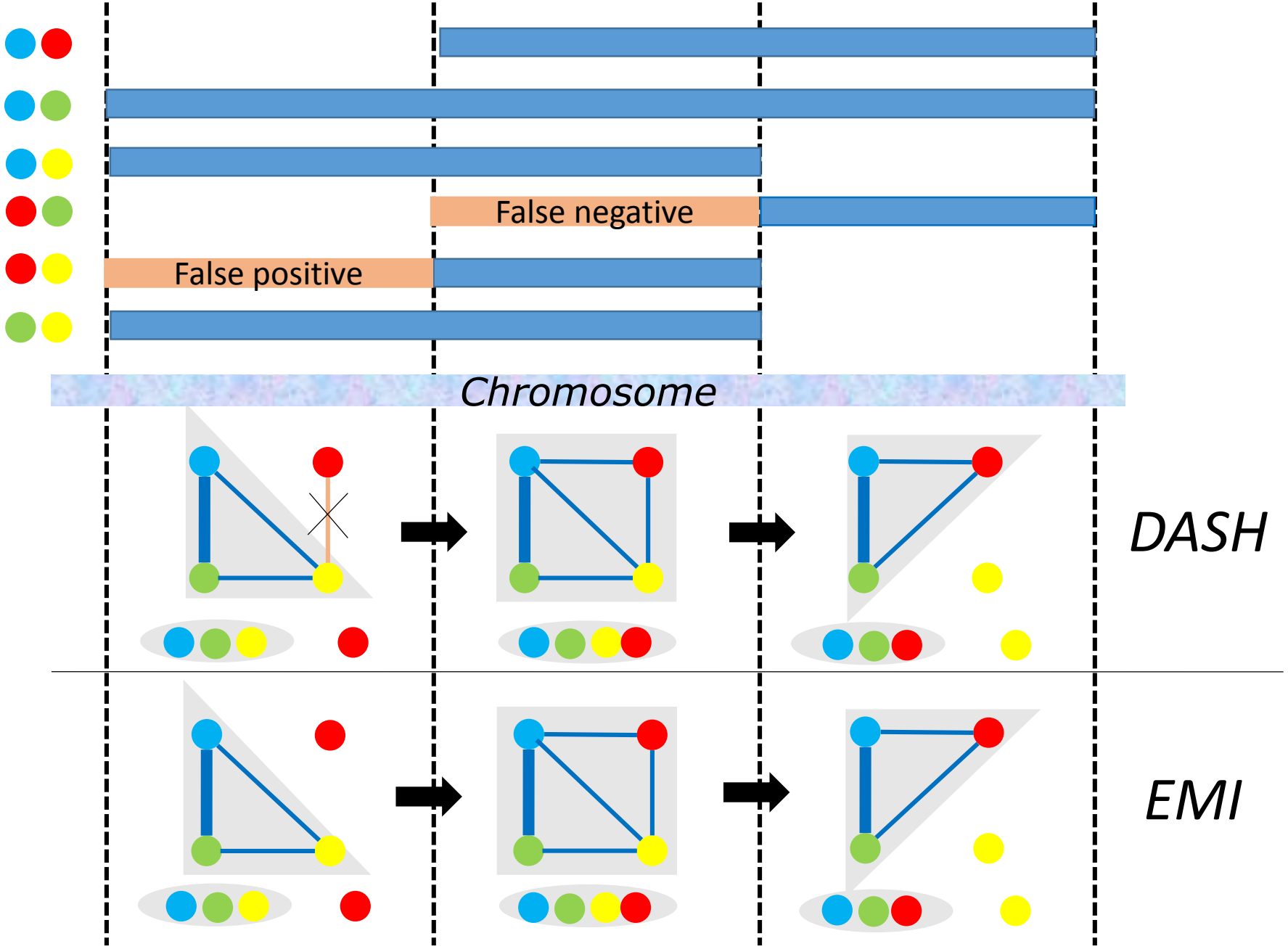


SPICi (**S**peed and **P**erformance **I**n **C**luster**i**ng)

- Jiang & Singh (2010)
- Expand a subgraph from seed nodes and by adding qualified adjacent nodes
- Stop when the density of the subgraph reaches a lower bound
- Output the subgraph and remove its links to the rest of the graph
- Search for another subgraph in the remaining graph until all non-isolated nodes are clustered.
- Apply computationally efficient data structures

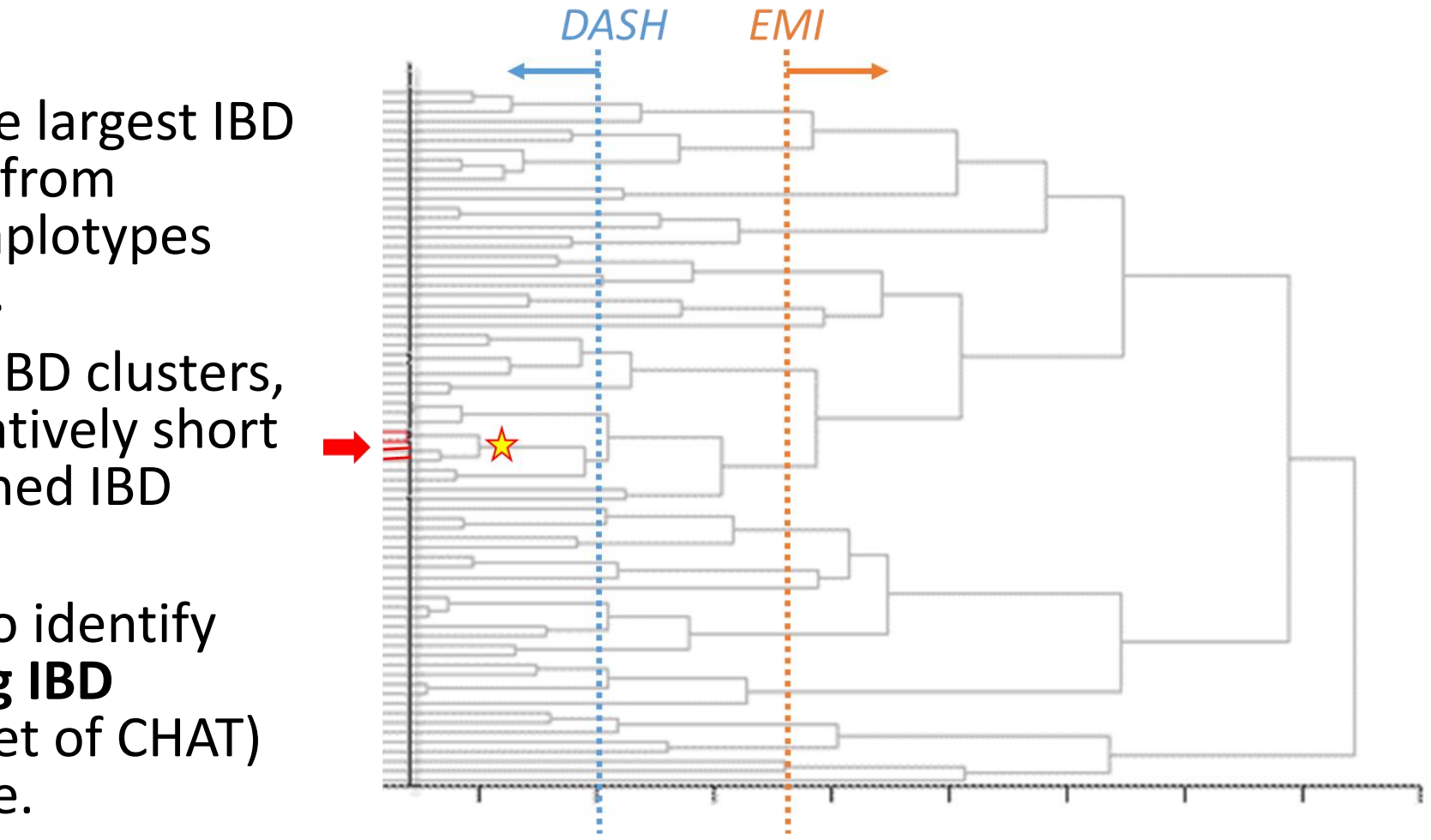


DASH & *EMI*: different clustering strategies (cont.)



DASH vs. EMI: different starting points (possible)

- DASH aims to find the largest IBD clusters, but it starts from relatively long IBD haplotypes (GERMLINE outputs).
- EMI can cover small IBD clusters, but it starts from relatively short IBD haplotypes (Refined IBD outputs)
- Thus, their abilities to identify **small clusters of long IBD haplotypes** (the target of CHAT) are still unclear to me.



The graph elements in IBD-Groupon

- IBD-Groupon constructs an HMM across the genome to identify the most likely IBD clusters. It splits pairwise IBD haplotypes into *maximum IBD chunks* and creates an HMM state for each chunk.
- To figure out the values of an HMM state, IBD-Groupon builds an IBD graph for each chunk and identifies all maximal cliques in that graph.
- IBD graphs are built on the results of an external pairwise IBD detection algorithm called fastIBD (Browning & Browning, 2011).
- IBD-Groupon only eliminates putative false-positive links. It does not add links to the IBD graph.

The graph elements in PIGS

- PIGS (probabilistic IBD graph sampling) imports (independent) IBD probabilities obtained by pairwise methods and updates each of them conditional on all others.
  - The global probabilistic framework is modeled as a fully connected weighted graph *G*, whose nodes represent haplotypes and edge weight represents the IBD probabilities between corresponding haplotypes.
  - A set of unweighted **transitive** graphs *gi* are induced from *G*, each representing a valid scenario of IBD clustering on haplotypes in *G*. PIGS calculates a conditional probability $P(gi|G)$ for each graph.
  - The IBD probability for Haplotypes *A* and *B* is updated as $\sum P(gk|G) / \sum P(gi|G)$, where *gk* is a transitive graph in which *A* and *B* are connected.
  - A new set of *gi* are recreated based on the updated edge weight in *G.*
- For efficiency, *gi* is sampled rather than enumerated.

Hierarchical graph clustering

- Identify multiple levels of clusters by iteratively grouping nodes with high similarity or removing links between nodes with low similarity.
- Need to define the similarity between one node/cluster and the other.

✓No prior knowledge on the number and size of the clusters is required.

✓Reveal the hierarchical structure in the data with multiple levels of resolution

× Results are sensitive to the similarity measure.

× Cannot choose a globally optimal level of clusters. The decision is often arbitrary.

× Nodes with only one neighbors are often mistaken as isolated clusters.

× Agglomerative clustering does not scale well.

× Cannot discover partially overlapped clusters.

Implications for CHAT

- CHAT divides the genome into segments of a fixed length (0.5 LDU). We will refer to these segments as **bin**s. They are equivalent to the windows in DASH and EMI.
- A major step in CHATv1.0 is to look for so called “CHATSets” in each bin. This step is equivalent to finding IBD clusters based on an IBD graph. A **CHATSet** represents an IBD cluster.

CHATSet building process

Pairwise detection results as LSHs

Explore merging

{Subject A, B}

LSH_AB

A: Bin i, i+1, …, i+k

LSH_AB

LSH_A#

LSH_AC

Randomly pick two to phase

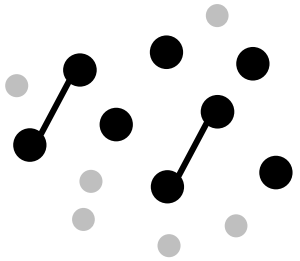
ASB ANSB ASC ANSC BSA BNSA CSA CNSA

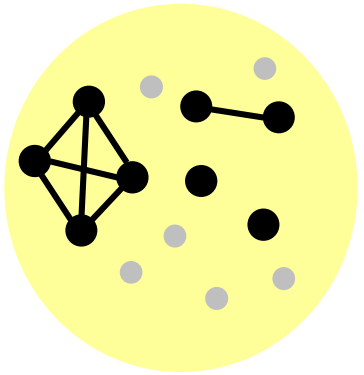
Merged successfully?

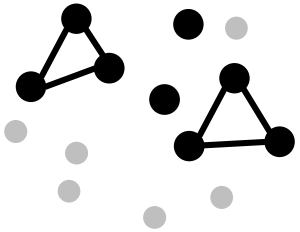
ASB (ASC) ANSB (ANSC) BNSA CNSA BSA CSA

OR

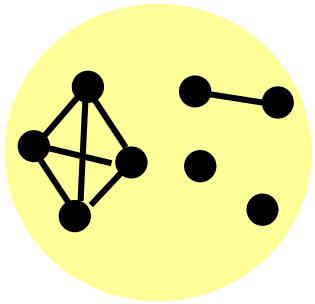
ASB (ANSC) ANSB (ASC) CSA BSA BNSA CNSA





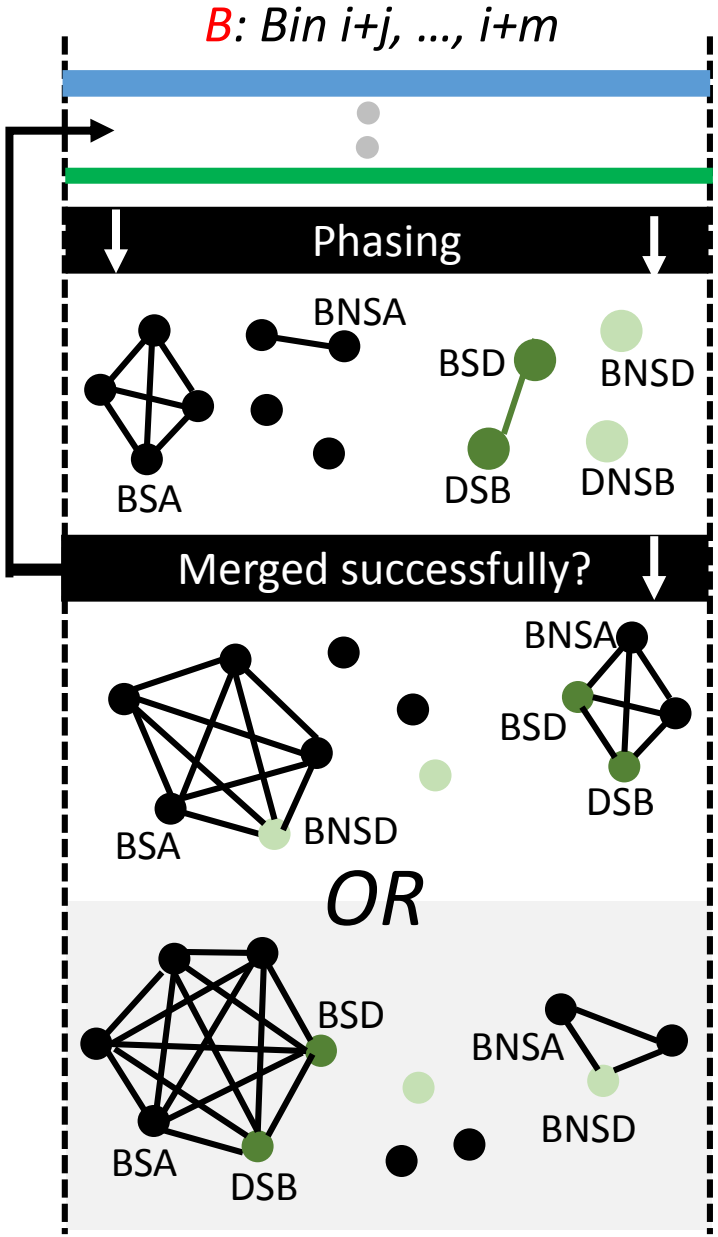


CHATSet building process (cont.)



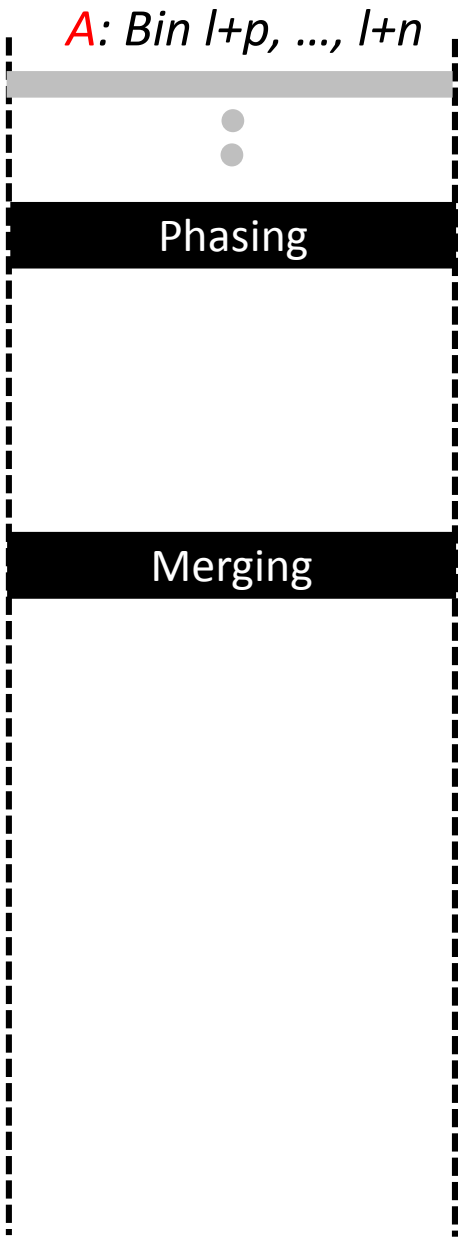
{A, B, C}

Explore merging



{A, B, C, D}

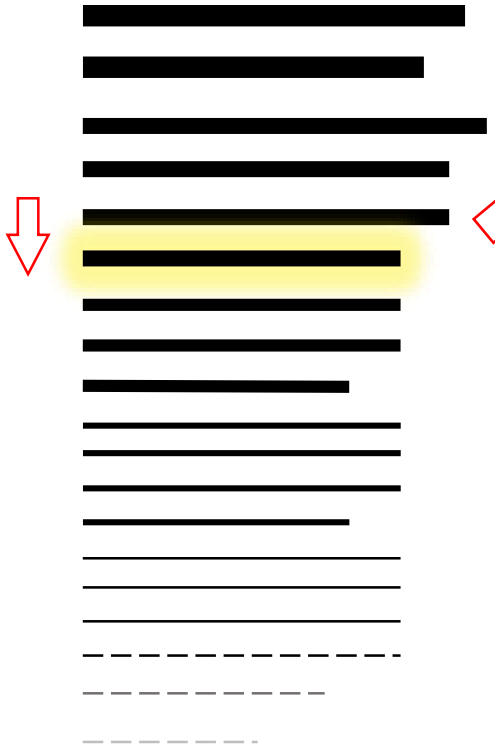
Explore merging with updated consensus haplotype



{A, B, C, D}

Explore merging

CHATSet building process (cont.)



Pairwise detection results as LSHs

{Subject P, Q}

LSH_PQ

P: Bin #, …

Phasing

Merging

{P, Q, …}

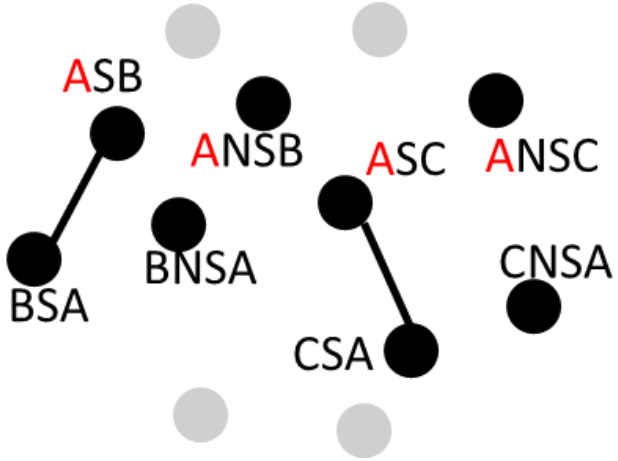
Q: Bin #, …

Phasing

Merging

CHATSet building process

- CHAT iterates through ordered pairwise detection results and updates the IBD graphs in all involved windows by adding **the strongest edges**.
- Meanwhile it keeps checking whether in some windows there are >= 2 pairwise IBD segments involving the same individual, i.e., an IBD graph like

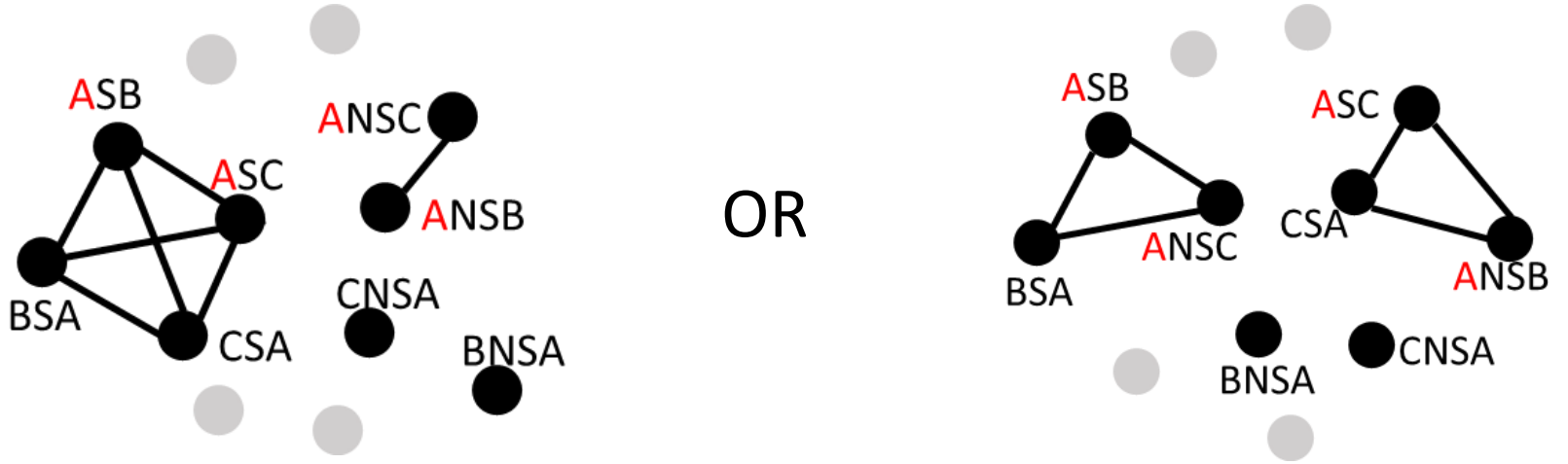


CHATSet building process (cont.)

- If that happens, CHAT tries to cluster certain nodes, because at a single window we should have

$$\begin{cases} ASB \equiv ASC \\ ANSB \equiv ANSC \end{cases} \text{ OR } \begin{cases} ASB \equiv ANSC \\ ANSB \equiv ASC \end{cases}$$

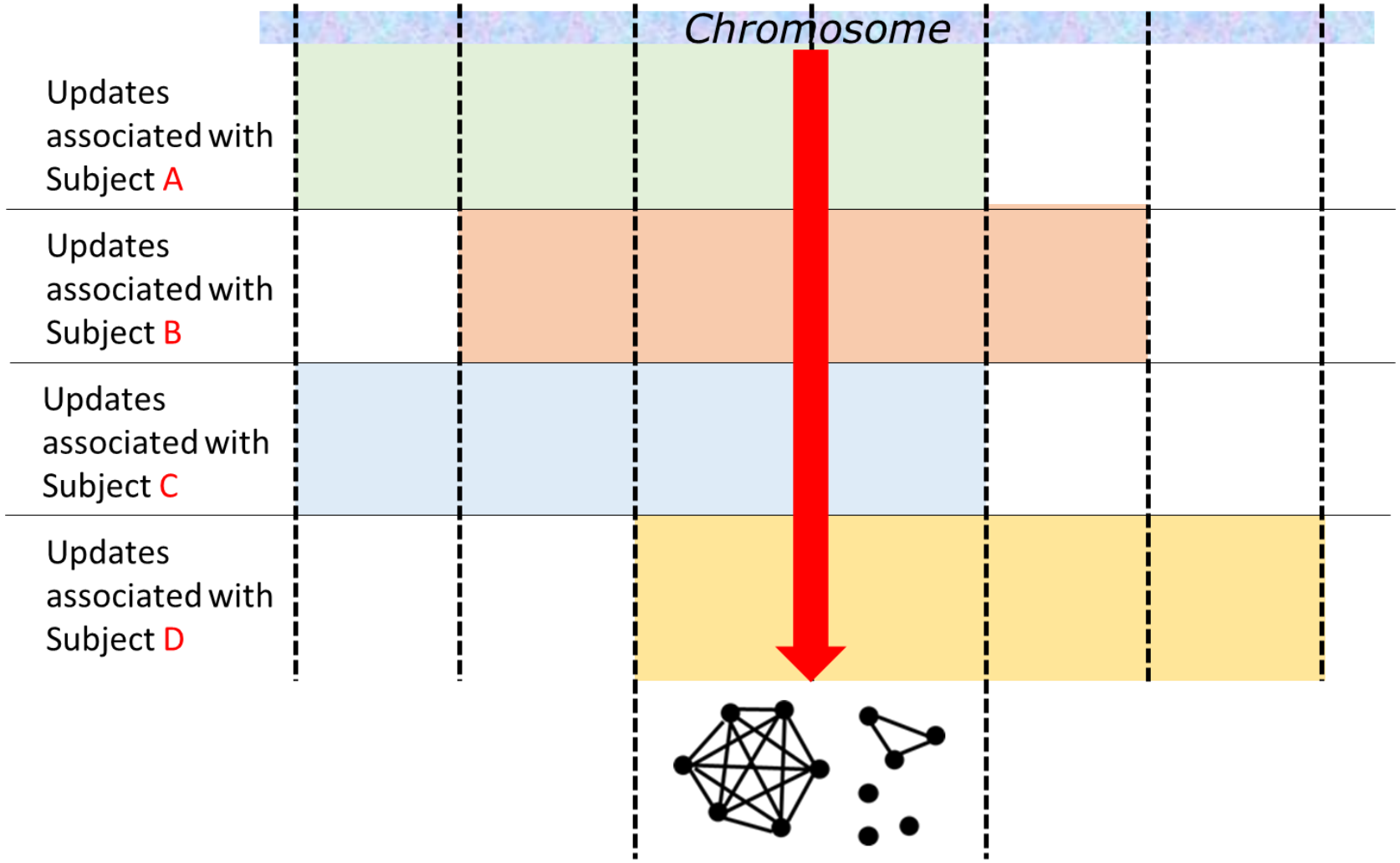
Successful clustering changes IBD graphs **in all relevant windows** to



CHATSet building process (cont.)

- Given successful clustering, CHAT checks every other subject also involved (i.e., B and C) to see whether they can, as the first subject, bring in more haplotypes to further expand existing clusters.
- This process ends when no more new subject were added. Then CHAT goes back to add strong edges base on pairwise results, until all results are processed.
- Notably, the examination of each subject may lead to updates of IBD graphs at multiple windows, but only at overlapping windows will the expansion happen.

Continuous expansion at overlapping bins



Difference from EMI

Is CHAT more vulnerable to errors due to 2 and 3?

1. Seed selection

   EMI selects one node with the highest strength, i.e., (the haplotype of) one subject that is in strong IBD with many other subjects. CHAT selects a few nodes that may indicate one long IBD haplotype. Thus, CHAT probably has fewer false positives (good for its purpose).

2. Density of the forming clusters

   CHAT assumes every cluster is a clique. Thus, once a node is added, a k-clique will automatically become a (k+1)-clique

3. Expansion of clusters

   EMI considers one candidate's IBD relations with ALL current members, while CHAT consider it's relation with only one current member

References

- Browning, B. L., and S. R. Browning, (2011) A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88: 173–182.
- Browning, S. R., and B. L. Browning. (2012) Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. 46: 617–633.
- Browning, B. L., and S. R. Browning. (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194.2: 459-471.
- Fortunato, Santo. "Community detection in graphs." *Physics reports* 486.3 (2010): 75-174.
- Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler et al., 2009 Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19: 318–326.
- Gusev, A. et al. (2011) DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet., 88, 706–717.
- ences[edit]
- Hartuv, E.; Shamir, R. (2000), A clustering algorithm based on graph connectivity, Information Processing Letters 76 (4-6): 175–181
- He, D. (2013) IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. Bioinformatics, 29, i162–i170.
- Hochreiter S (2013). HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic acids research, 41(22):202.
- Jiang,P. and Singh,M. (2010) SPICi: a fast clustering algorithm for large biological networks. Bioinformatics, 26, 1105–1111.
- Moltke,I. et al. (2011) A method for detecting IBD regions simultaneously in multiple individuals–with applications to disease genetics. Genome Res., 21, 1168–1180.
- Park et al. (2015) PIGS: improved estimates of identity-by-descent probabilities by probabilistic IBD graph sampling. Bioinformatics, 16(Suppl 5):S9.
- Qian Y et al., (2014). Efficient clustering of identity-by-descent between multiple individuals. Bioinformatics, 30(7):734-922.