“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)

Detect novel rare signals from existing case-control GWAS data: use IBD segments as genetic markers in association analysis

Yuan Lin

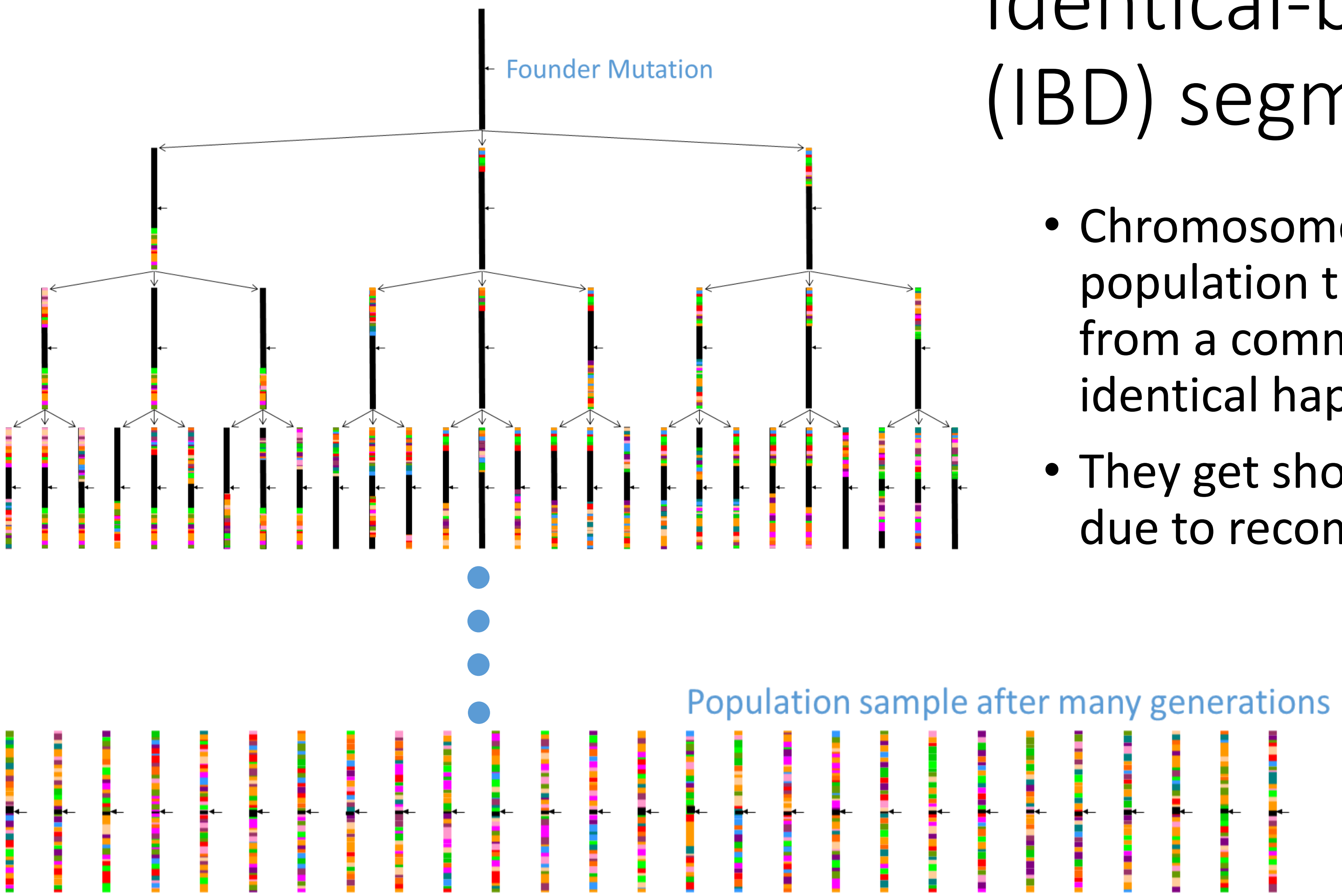
Post-doc Research Associate

@ Dr. Kirk Wilhelmsen’s Lab

Department of Genetics, School of Medicine

UNC at Chapel Hill

Identical-by-decent (IBD) segment



- Chromosome segments in current population that are descended from a common ancestor (display identical haplotypes).
- They get shorter over generations due to recombination events.

IBD-segment-based association mapping

- IBD mapping tests the association between phenotype similarity and the IBD sharing of some chromosome segments in “unrelated” individuals.
- Comparing with traditional linkage analysis, relatively short IBD segments in population samples allow for fine mapping, but these segments are not as accurately inferred due to the lack of pedigree information.
- IBD mapping targets at regions that contain some recent (relatively rare) mutations associated with the disease.

IBD-based association mapping (cont.)

- IBD mapping is different from haplotype-based association methods. It uses haplotypes to infer IBD sharing rather than directly tests their disease association.
- Due to its region focus, IBD mapping can be used to detect the effects of non-SNP variants (e.g., CNV), or the joint effect of multiple rare variants at a single locus (e.g., allele heterogeneity).
- IBD mapping is not a new idea, but not until recently can relatively short IBD segments be detected from GWAS-scale data with acceptable accuracy and efficiency.

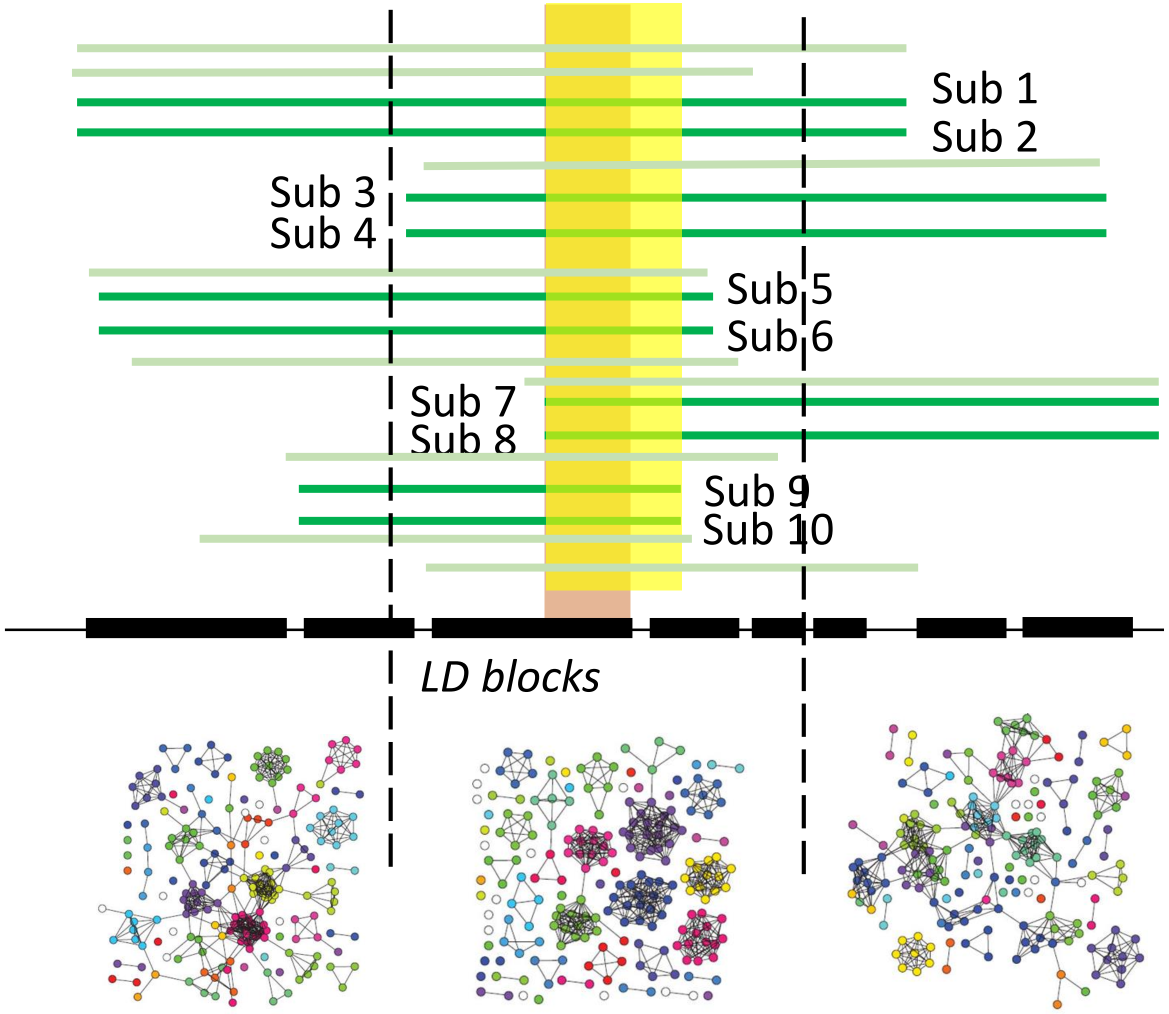
IBD mapping by use of pairwise IBD segments

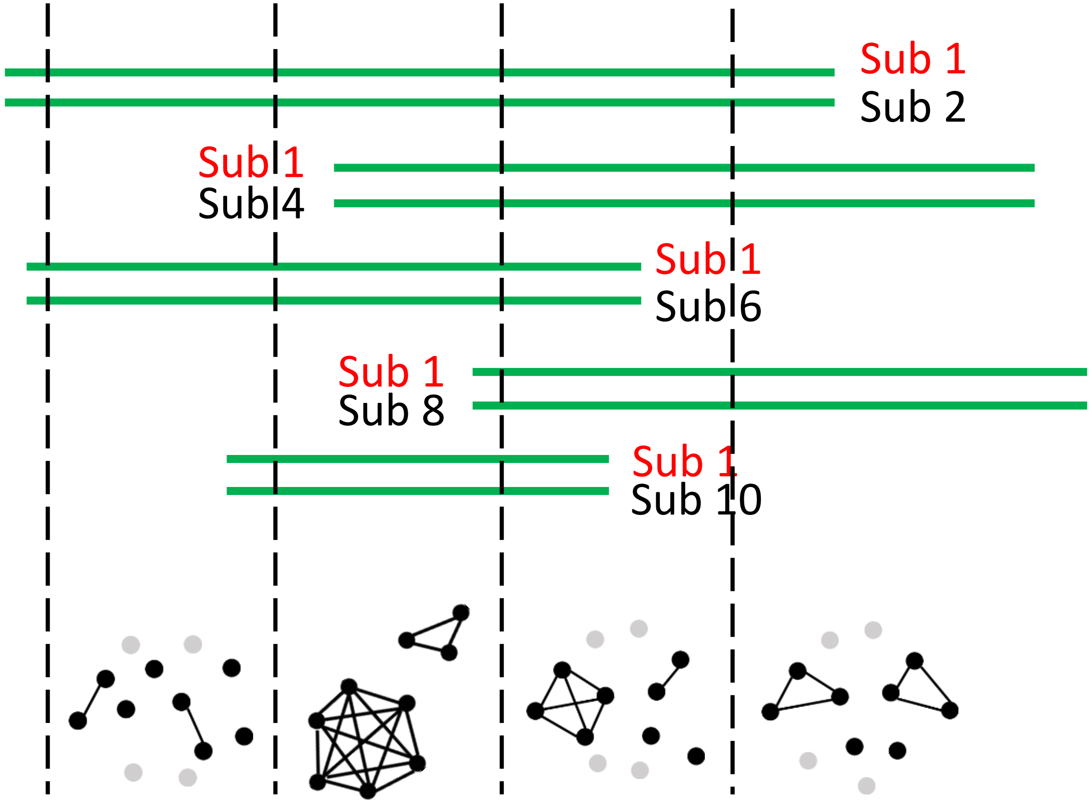
- At each SNP position, detect surrounding IBD segments between every pair of individuals.
  - Many available methods, such as **Refined IBD** (Browning & Browning, 2013) and **GERMLINE** (Gusev et al., 2009)
- Look for positions where there is an inflation in the number of IBD segments between case-case pairs, as opposed to case-control and control-control pairs.
- Test statistics are often normalized using the genome-wide average of each type of pairs

IBD clusters and group-wise IBD segments

- Multiple (>2) individuals may inherit the same IBD segment. These individuals form a so-called **IBD cluster**.
- The IBD segment shared by a cluster of individuals tends to be shorter and rarer than any pairwise IBD segment from these individuals (given the same SNP position).
    - A group-wise IBD segment is the overlap of some pairwise IBD segments. (The reverse may not be true)
- Although group-wise IBD segments may serve the purpose of IBD mapping better, they are generally more difficult to detect.

Detection of group-wise IBD segments

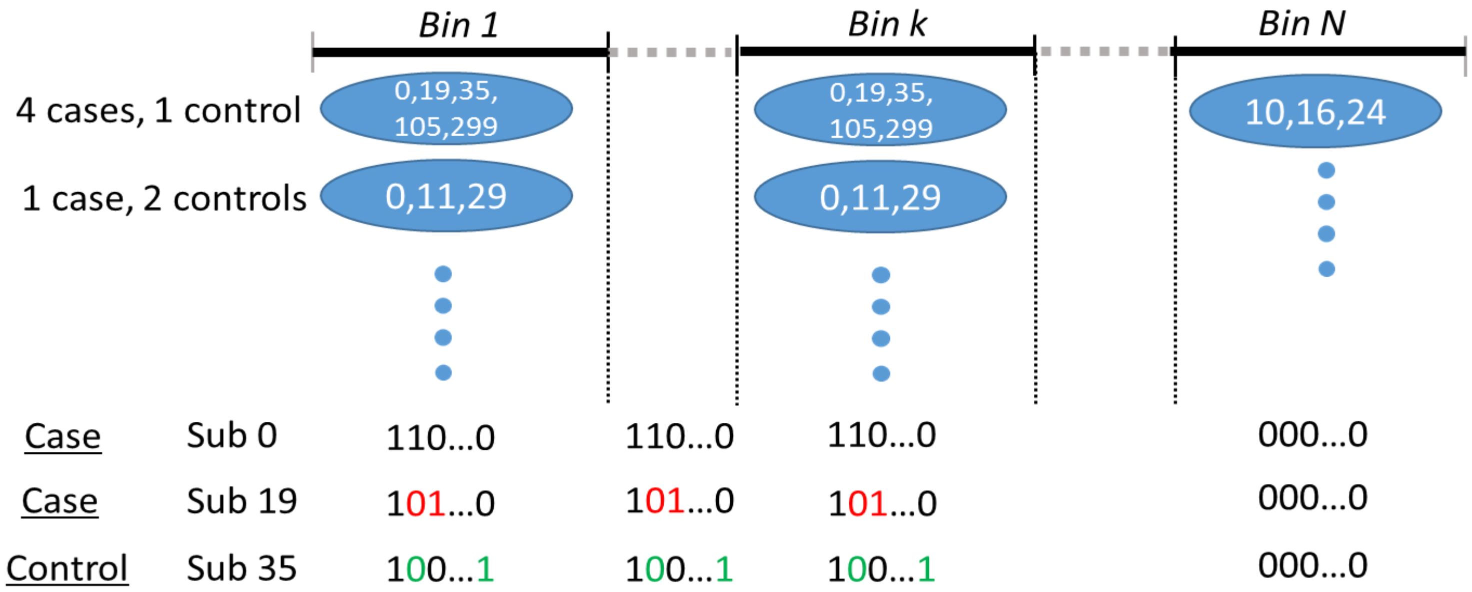




IBD mapping by use of IBD clusters

- At each SNP position (or a small region including a few SNPs), detect IBD clusters using previously identified pairwise IBD segments
  - Only a few methods suitable for GWAS data, such as **DASH** (Gusev et al., 2011), **EMI** (Qian, 2014), and **CHAT**
- Each cluster thus “tags” a group-wise IBD haplotype that may contain some rare risk variants.
- Test the disease association of IBD haplotypes by examining the case-control composition of corresponding clusters.
  - Chi-square or Fisher’s exact test

IBD mapping by use of IBD clusters (cont.)

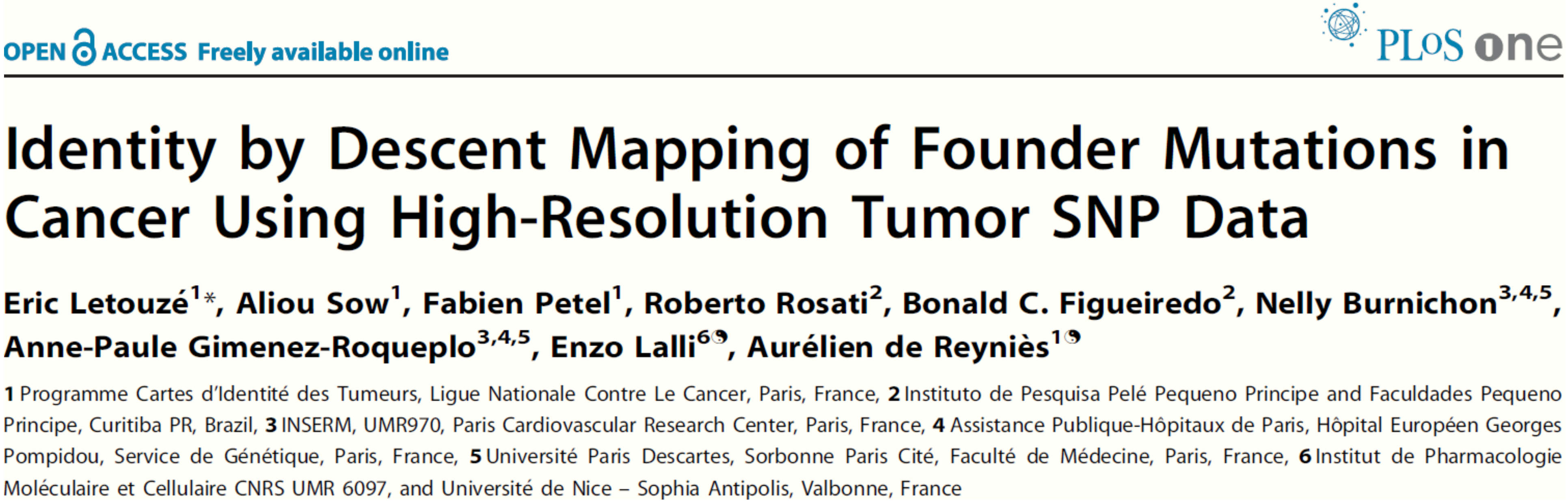


Case	Sub 0	110…0	110…0	110…0	000…0
Case	Sub 19	101…0	101…0	101…0	000…0
Control	Sub 35	100…1	100…1	100…1	000…0

- We can assign pseudo-genotypes to individuals based on their memberships to different clusters across the genome.
- ***Does it allow us to consider multiple clusters and incorporate covariates using statistical methods similar to those for GWAS?***

Application of IBD mapping

- Explore novel signals using already collected GWAS datasets
- ***What kind of data are mostly likely to produce interesting results?***



OPEN ACCESS Freely available online

PLoS one

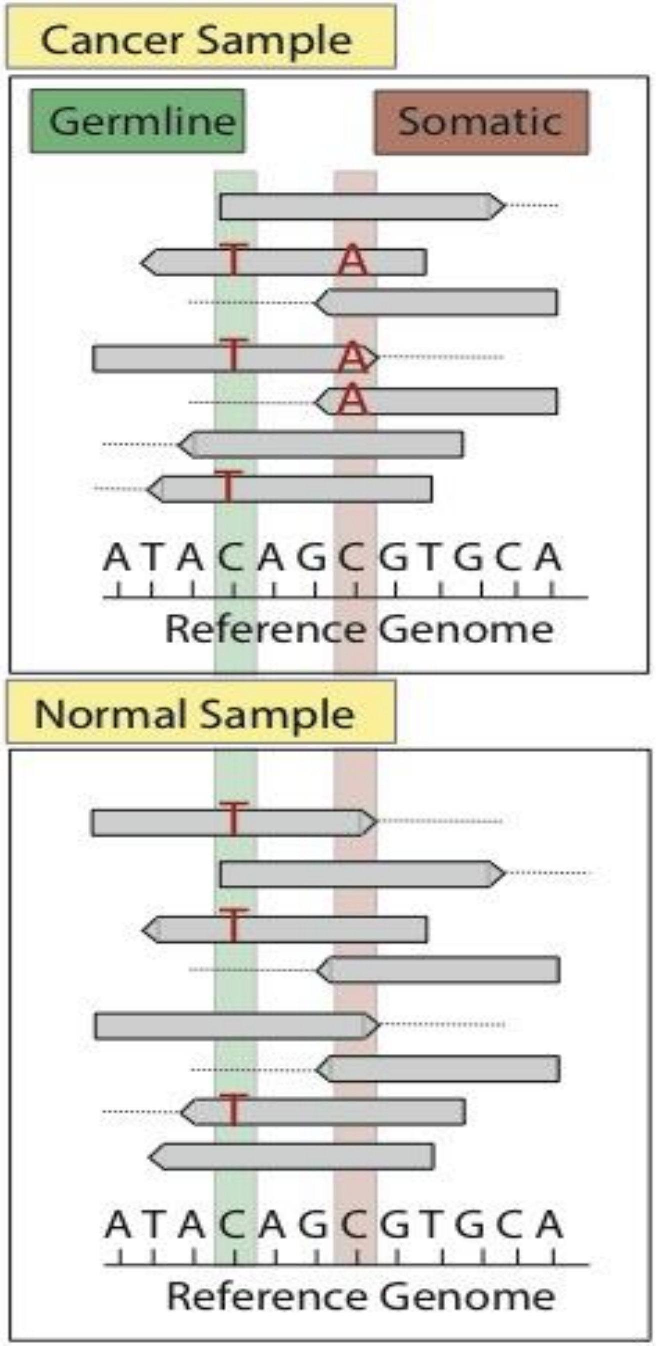
Identity by Descent Mapping of Founder Mutations in Cancer Using High-Resolution Tumor SNP Data

Eric Letouzé[1]*, Aliou Sow[1], Fabien Petel[1], Roberto Rosati[2], Bonald C. Figueiredo[2], Nelly Burnichon[3,4,5], Anne-Paule Gimenez-Roqueplo[3,4,5], Enzo Lalli[6]¶, Aurélien de Reyniès[1]¶

1 Programme Cartes d'Identité des Tumeurs, Ligue Nationale Contre Le Cancer, Paris, France, 2 Instituto de Pesquisa Pelé Pequeno Principe and Faculdades Pequeno Principe, Curitiba PR, Brazil, 3 INSERM, UMR970, Paris Cardiovascular Research Center, Paris, France, 4 Assistance Publique-Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Service de Génétique, Paris, France, 5 Université Paris Descartes, Sorbonne Paris Cité, Faculté de Médecine, Paris, France, 6 Institut de Pharmacologie Moléculaire et Cellulaire CNRS UMR 6097, and Université de Nice – Sophia Antipolis, Valbonne, France

Background

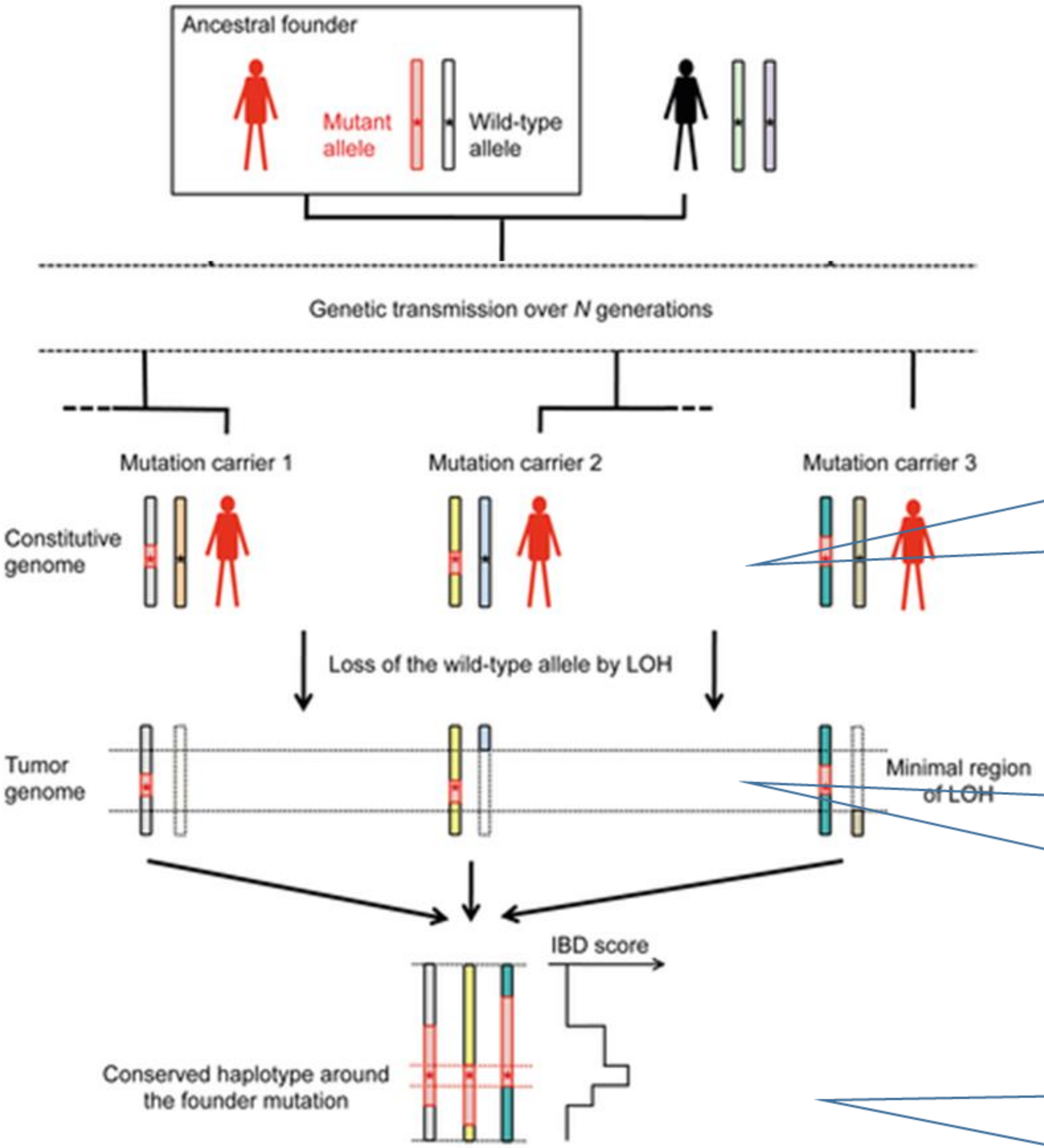
- A cancer patient has two types of single nucleotide variants (SNVs)
- Germline SNV (SNPs)
    - Inherited from parents
    - Present in all cells
- Somatic SNV
    - Acquired during one’s lifetime
    - Only present in tumor cells
    - Not passable to children



Background (cont.)

- Tumor suppressor genes (TSGs) are one type of genes often seen malfunctioning in tumor cells.
- TSGs act recessively. Both copies of a TSG must be lost or mutated for cancer to occur.
- An individual who inherits a TSG mutation is fine as long as the mutation is balanced by a wild-type allele.
- Cancer occurs when for some reason, the normal wild-type allele is lost in some cells at some time during that individual’s life (**loss of heterozygosity, LOH**)

1. Identify minimal LOH regions
2. Reconstruct tumor haplotypes in these regions
3. Detect disease-related IBD segments using tumor haplotypes



Premise 1: A founder mutation that increases carriers' susceptibility to a specific type of cancer is passed down generations.

Premise 2: The germline mutation resides in the minimal (overlapping) LOH region of all cases' tumor genome.

Premise 3: Current IBD detection methods can identify the IBD haplotype around that mutation at least in some cases.

FounderTracker

- A score for each pairwise IBD segment: the log-transformed probability of observing two identical haplotypes by chance, assuming all SNPs were independent.

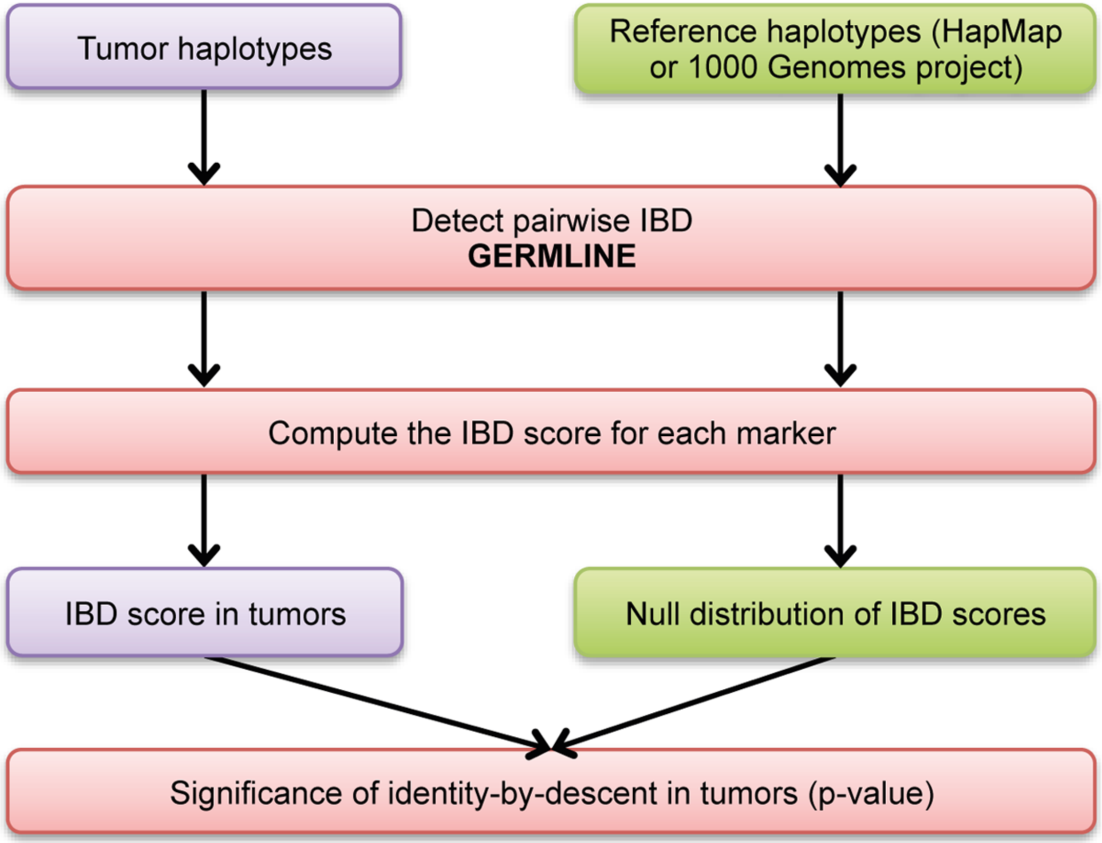
$$S_H = -\log 10\left(\prod_{i/H_i=A}(1-Bfreq_i)^2 \times \prod_{i/H_i=B} Bfreq_i^2\right)$$

- A IBD score for SNP marker i: sum of the S_H of every pairwise IBD segment containing that SNP, assuming these segments are independent.

$$IBDscore_i = \sum_{i \in H} S_H$$

- A null distribution of IBD scores for SNP i, fitted using that SNP's IBD scores in all reference sets, assuming they follows a Gumbel distribution

$$P_i(X \le x) = e^{-e^{-(x-\mu_i)/\beta_i}}, \text{ mode } \mu_i, \text{ variance } \pi^2 {\beta_i}^2/6$$


Tumor haplotypes
Reference haplotypes (HapMap or 1000 Genomes project)
Detect pairwise IBD
GERMLINE
Compute the IBD score for each marker
IBD score in tumors
Null distribution of IBD scores
Significance of identity-by-descent in tumors (p-value)

- Thus, the probability for SNP i to have an IBD score higher than that from the set of tumor haplotypes (final p-value) is

$$p_i = 1 - P_i(X \le tumorIBDscore_i)$$

Reference set construction

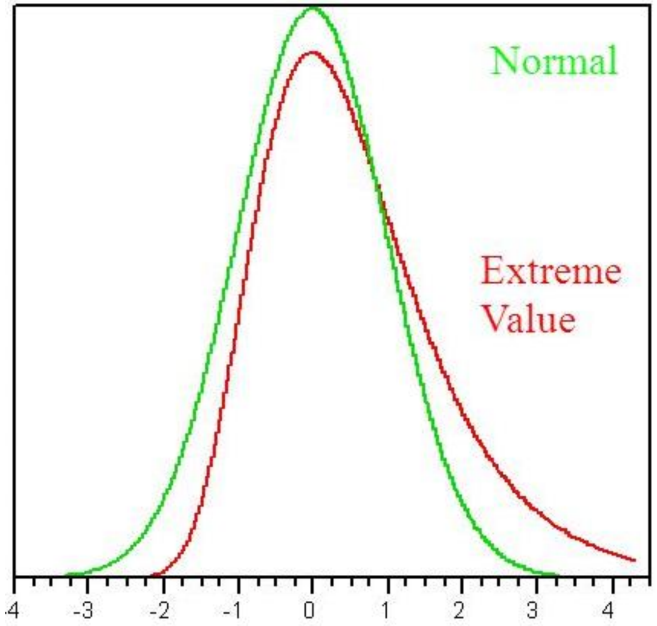
- Better use the haplotypes of some controls collected from the same population to improve accuracy
- The size of the reference pool should be big enough comparing to the sample size. Otherwise, there could be “artificial” long haplotype sharing in reference sets and even identical reference sets.

Assumption of independence

- The calculations are conducted by assuming markers are independent (no LD) and pairwise IBD segments around the same marker are independent.
- The FDR adjustment procedure they used assumes independence among all the tests (Benjamin & Hochberg, 1995).
- Would it be more reasonable to do an LD pruning of all the markers and then only use markers that are low in LD?

Null distribution estimation

- Given a specific SNP position, if there are *M* (a sufficiently large number) pairwise IBD segments detected by GERMLINE, the largest S_H of all segments or the sum of *K* (*K*<<*M*) largest S_H would follow an extreme value distribution such as Gumbel (Dudbridge & Koeleman, 2004).
- Using either the largest S_H or the sum of *K* largest S_H in each reference set to fit null distributions of the test statistic seems more reasonable to me than the current way (i.e. use the sum of all S_H)



IBD haplotype simulation

- In every simulated tumor dataset, a haplotype of length *x* implanted in *y*% samples.
  1. Randomly select a chromosome and a genomic region of length *x* on that chromosome.
  2. Randomly select 100 samples from a pool of haplotypes obtained from 1000 Genomes.
  3. Copy the first sample’s haplotype at the region selected in Step 1 to *y* samples randomly.
  4. Adjust the length of the copied haplotype in each sample.
- Significant SNPs detected inside/outside the implanted haplotype are treated as true/false positives.

IBD haplotype simulation (cont.)

- The authors seem to implicitly assume that there would be no other IBD haplotypes besides the implanted one. However, the original haplotype samples are not “IBD-free”. Thus, some “false positives” may actually be true.
- ***What are more appropriate ways to simulate IBD haplotypes?***
  1. Simulate true IBD haplotypes. Generate DNA sequences from scratch and keep ancestry information to infer true IBD haplotypes.
  2. Implant IBD haplotypes into real data as this study did, but before that destroy latent IBD segments (>= certain length).

Approach 1

- In coalescent simulation, the **ancestral recombination graph (ARG)** describes coalescent events across the genome and IBD partitions.
- We can then implant a founder mutation and know exactly which IBD segment in current samples harbor that mutation.

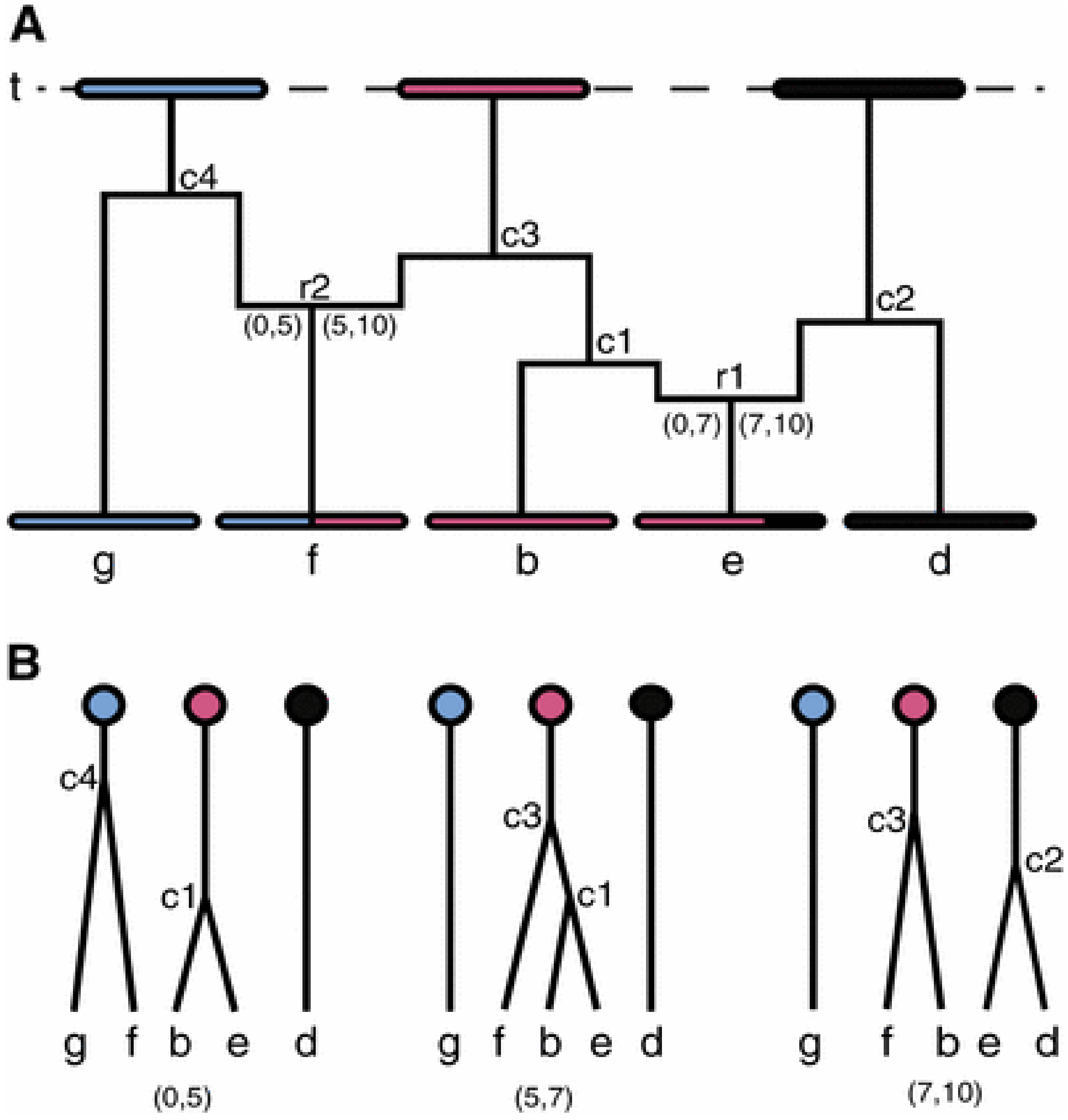
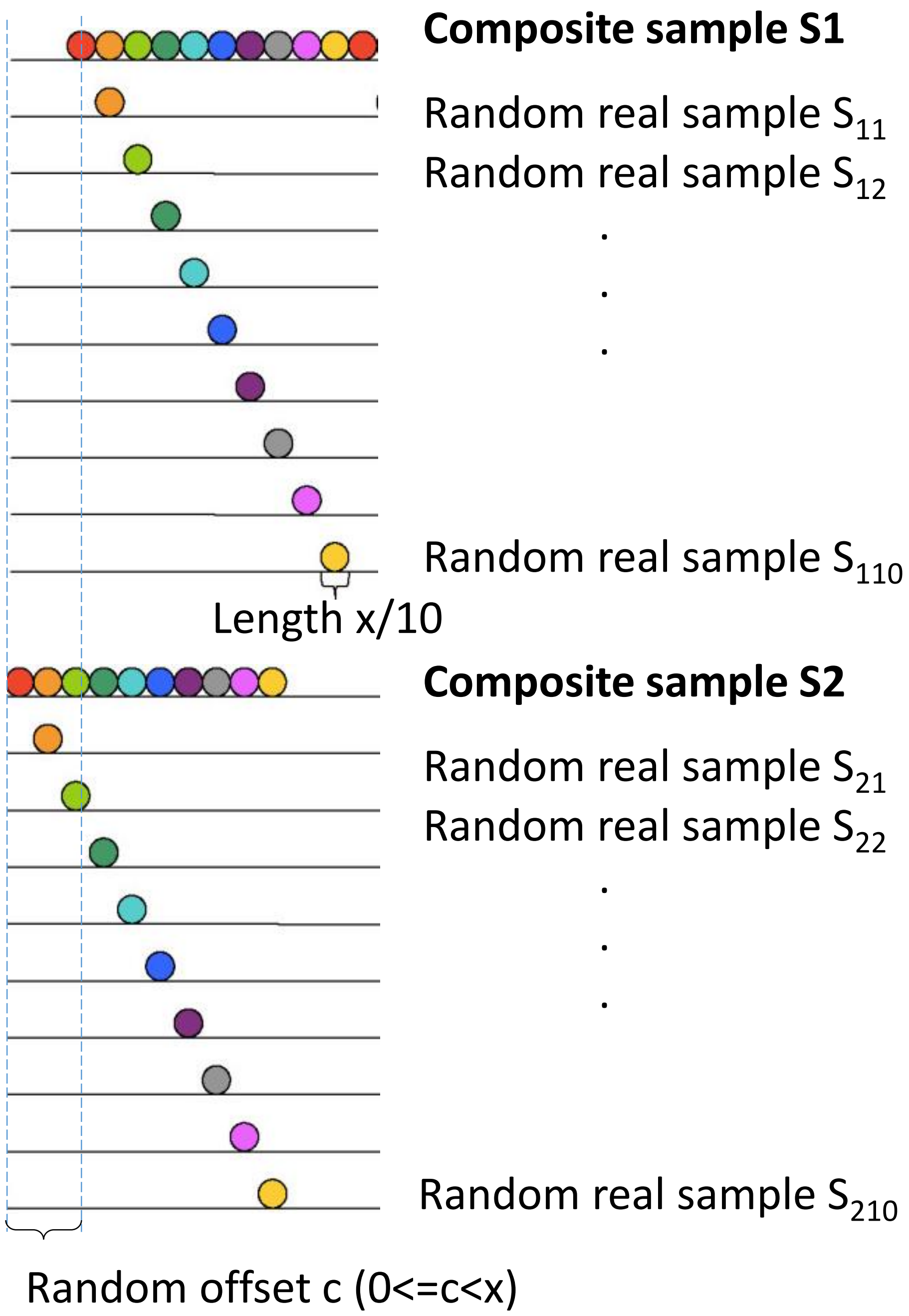


Figure 1 in (Thompson, 2013)

Approach 2

- We can greatly reduce latent IBD haplotypes in existing data by constructing composite samples from real samples.
- Then, any IBD segment of length >= x/10 detected among composite samples can be considered as a false positive



Composite sample S1

Random real sample S_{11}

Random real sample S_{12}

.
.
.

Random real sample S_{110}

Length x/10

Composite sample S2

Random real sample S_{21}

Random real sample S_{22}

.
.
.

Random real sample S_{210}

Random offset c (0<=c<x)