Using DNA from many samples to distinguish pedigree relationships of close relatives

Amy L. Williams

@amythewilliams

February 24, 2020

Family History Technology Workshop

Cornell University

Massive datasets: Many close relatives / small pedigrees

>100,000 samples

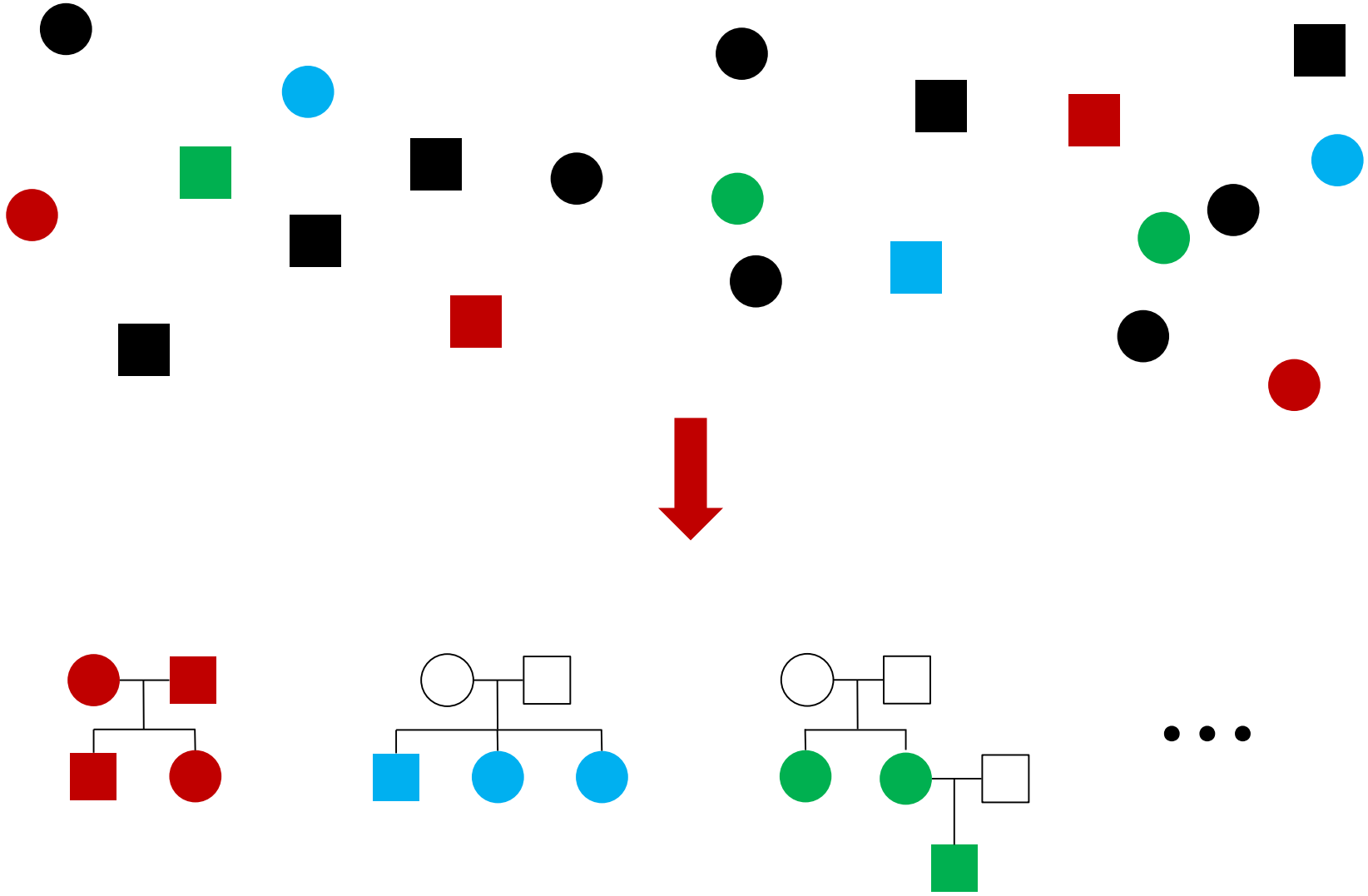
> 9 million samples

~500,000 samples

>14 million samples

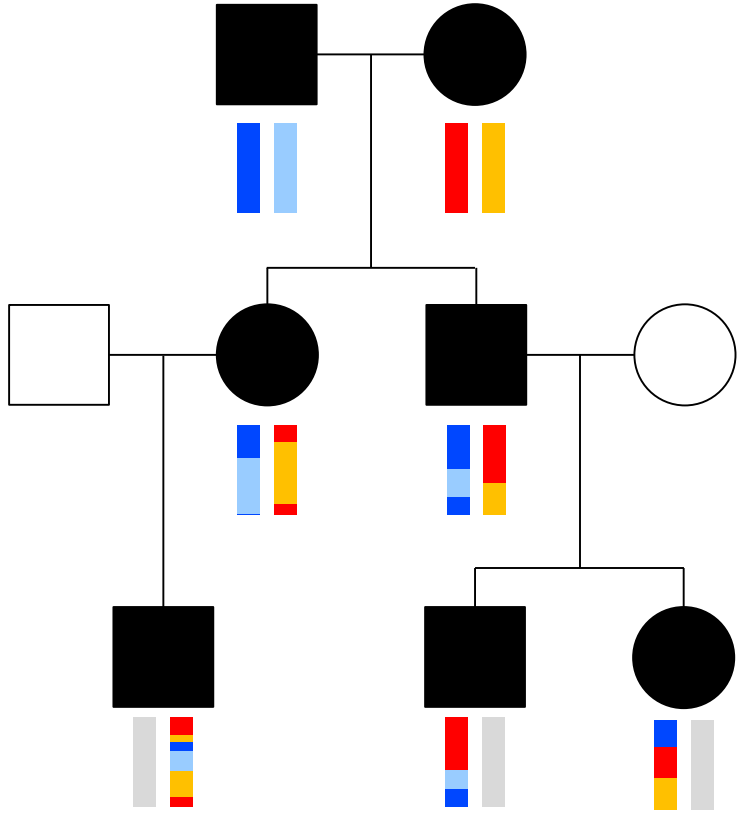
In dataset with n individuals, have $\binom{n}{2} = \frac{n(n-1)}{2} = \mathcal{O}(n^2)$ pairs

Goal: detect and reconstruct pedigrees using only DNA



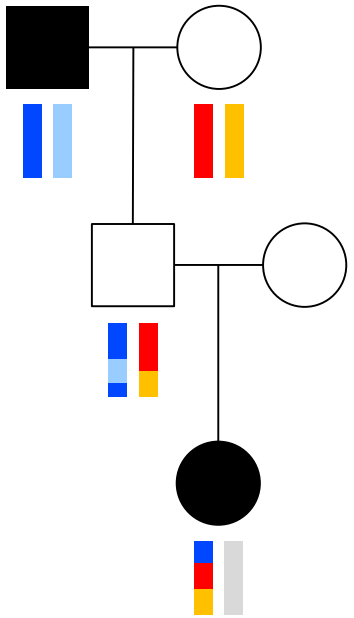
Signal: Identical by descent (IBD) sharing

- Close (and some distant) relatives share large regions identical by descent (IBD)
  - Represented here as same color
- Each generation, parents transmit random ½ of their genome to children
  - ➢ Relatives separated by M generations share average of $\frac{1}{2^M}$ of genome

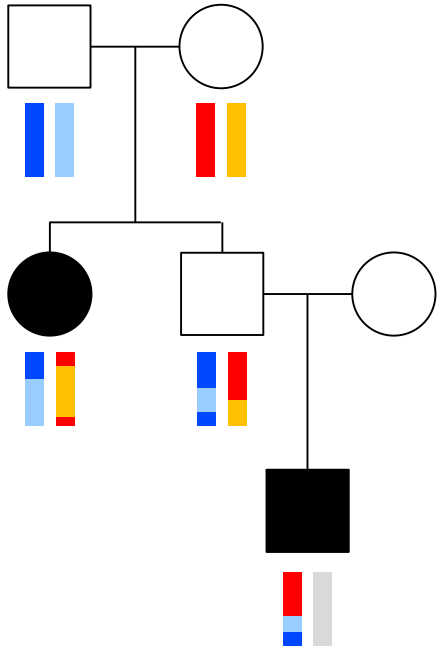


- Average IBD sharing fractions:
  - Full siblings: 50%, Aunt-nephew: 25%, First cousins: 12.5%

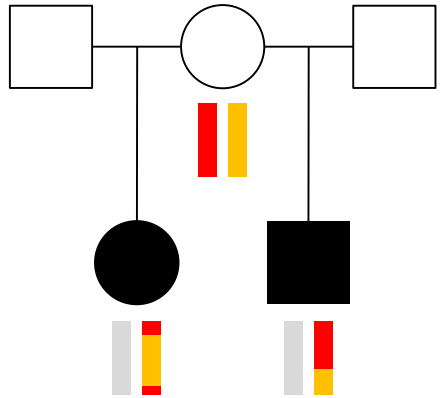
Second degree relatives: All share ~25% of genome IBD



Grandparent-grandchild (GP)



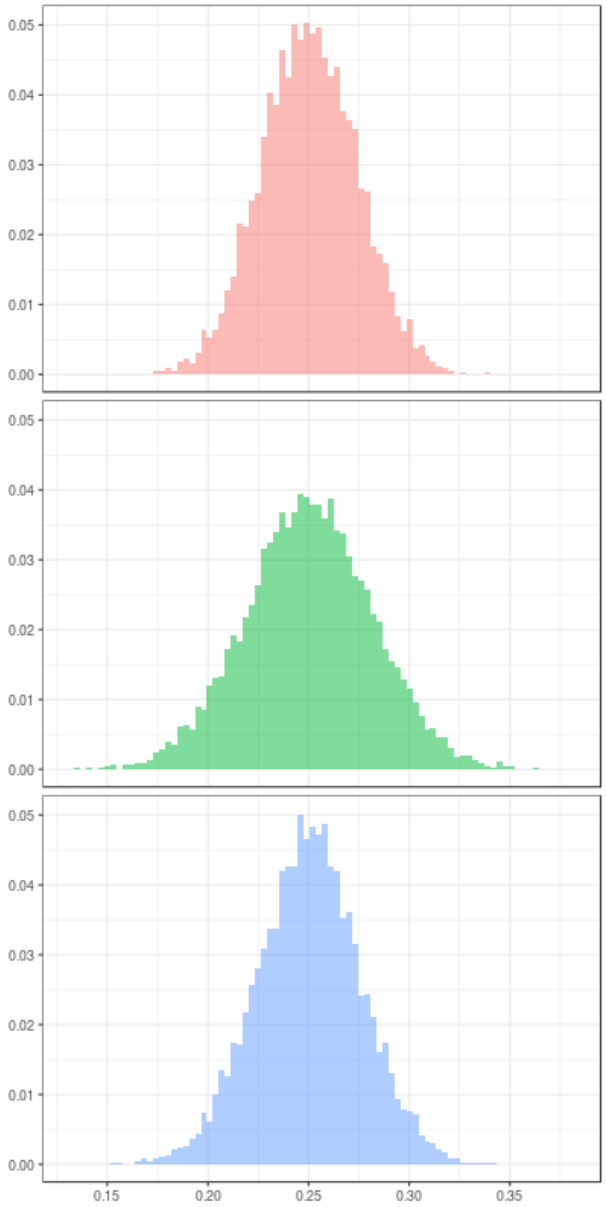
Avuncular (AV)

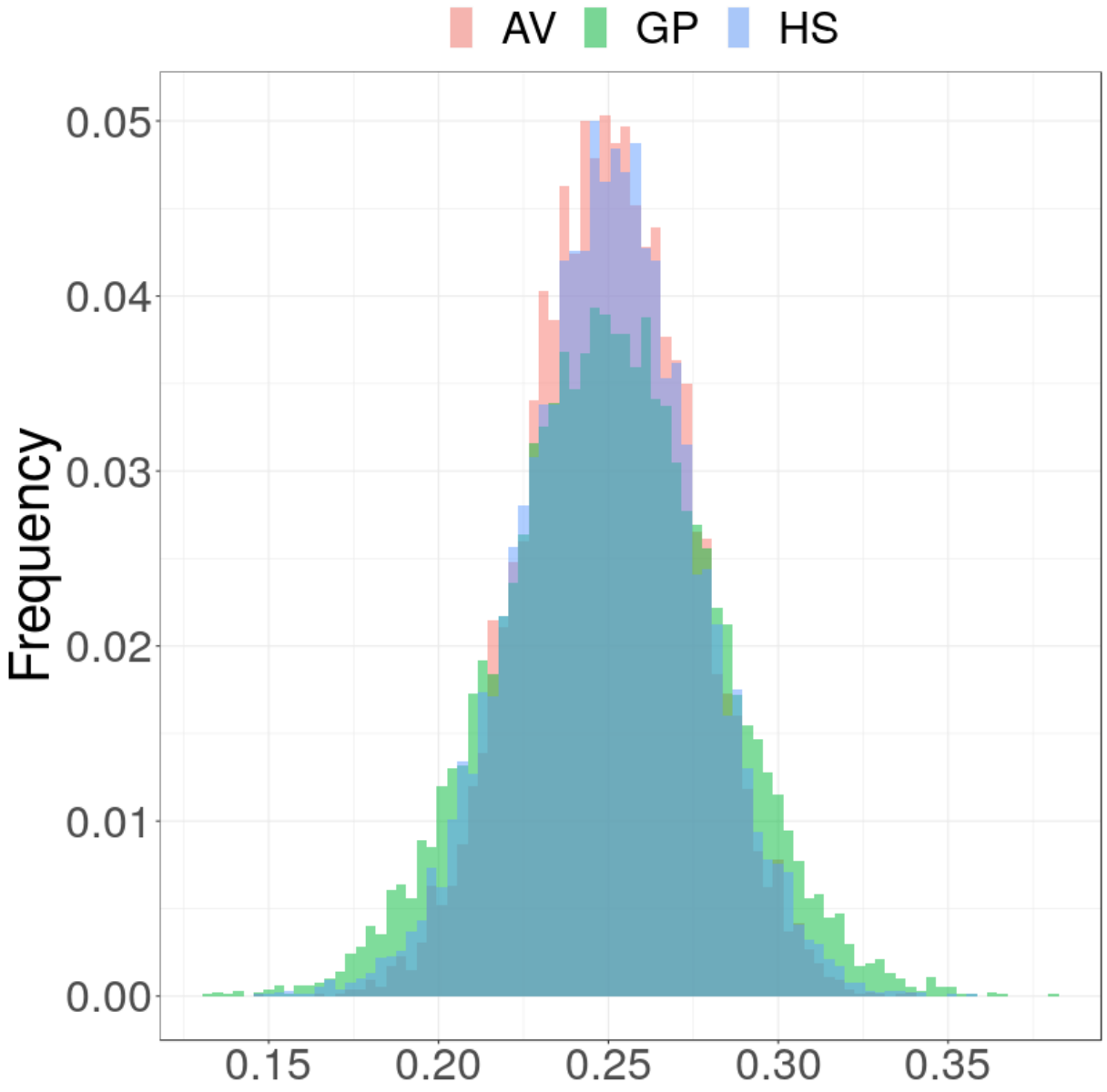


Half-sibling (HS)

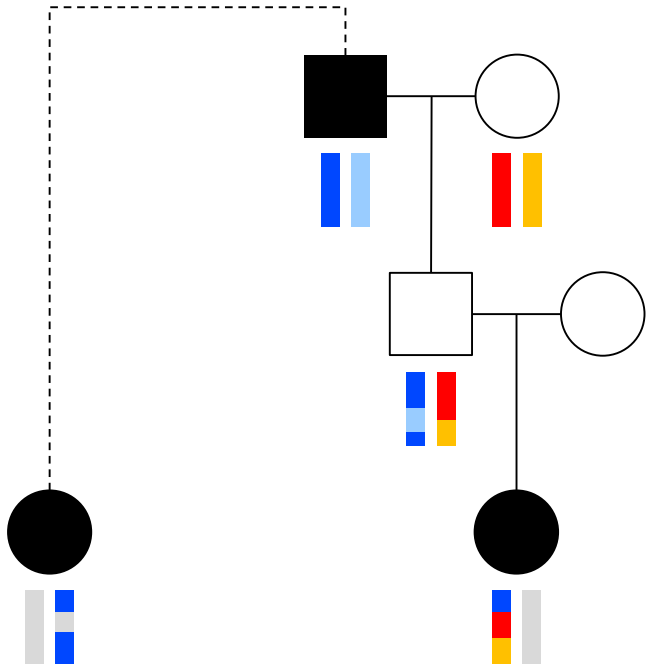
- Difficult to distinguish using only data from the pairs

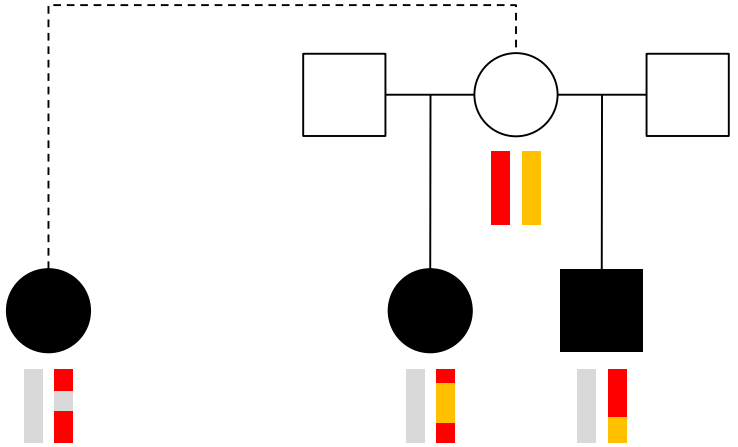
IBD sharing rates for these relationships heavily overlap



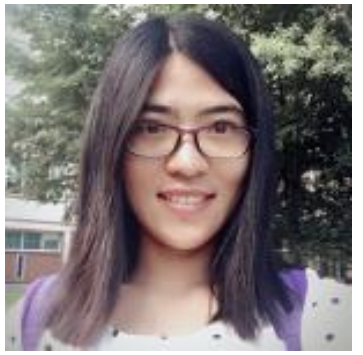


Idea: analyze IBD sharing of pair to other relatives





CREST: Classification of Relationship Types



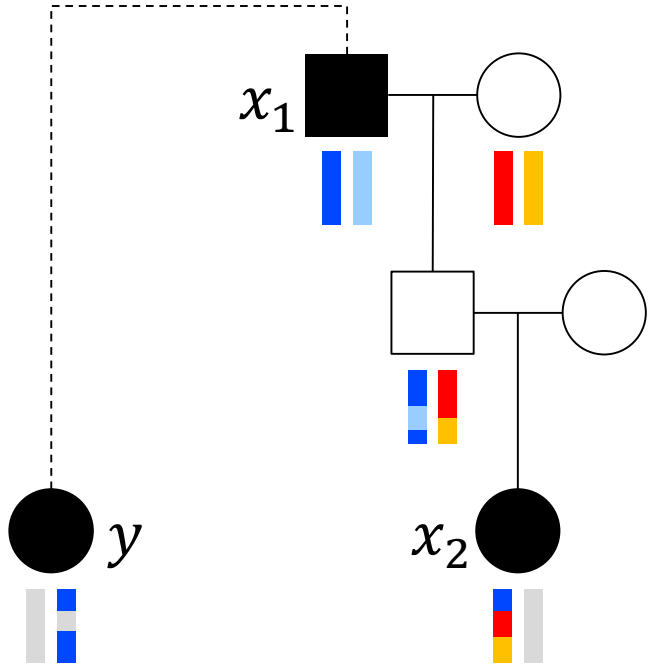
Ying Qiao

Jens Sannerud

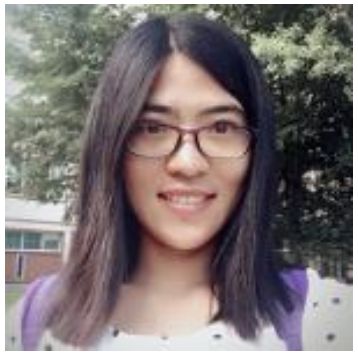
Approach: ratios of IBD sharing in three samples versus two

$$R_1 = \frac{Length\left(IBD_{x_1,y} \cap IBD_{x_2,y}\right)}{Length\left(IBD_{x_1,y}\right)}$$

$$R_2 = \frac{Length\left(IBD_{x_1,y} \cap IBD_{x_2,y}\right)}{Length\left(IBD_{x_2,y}\right)}$$



For GP, expect $R_1 = 1/4, R_2 = 1$

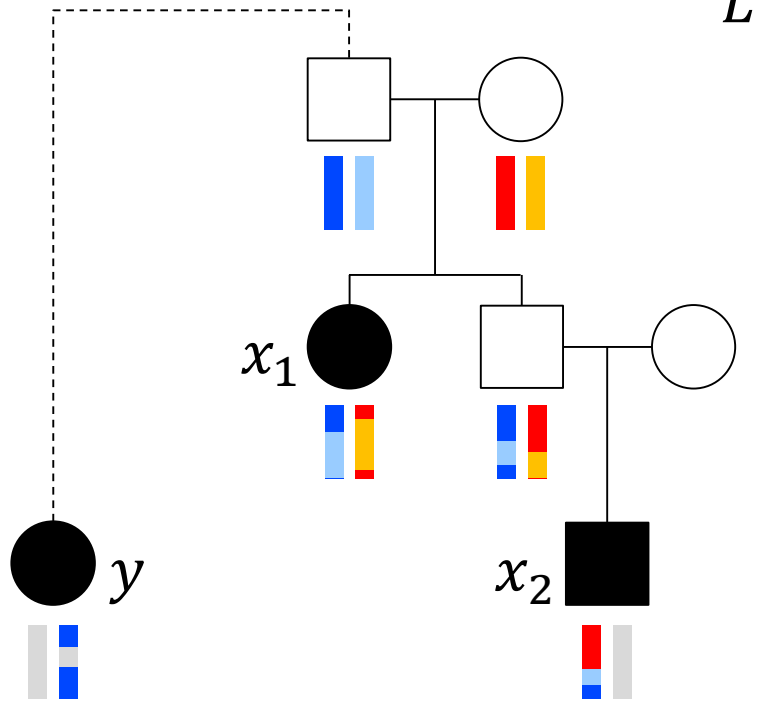


Ying Qiao

Approach: ratios of IBD sharing in three samples versus two

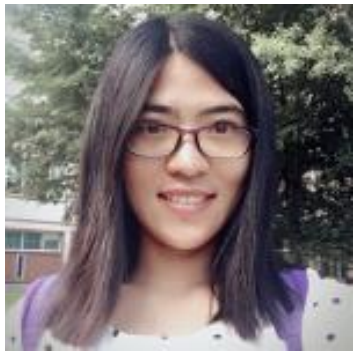
$$R_1 = \frac{Length\left(IBD_{x_1,y} \cap IBD_{x_2,y}\right)}{Length\left(IBD_{x_1,y}\right)}$$

$$R_2 = \frac{Length\left(IBD_{x_1,y} \cap IBD_{x_2,y}\right)}{Length\left(IBD_{x_2,y}\right)}$$



For GP, expect $R_1 = 1/4, R_2 = 1$

For AV, expect $R_1 = 1/4, R_2 = 1/2$

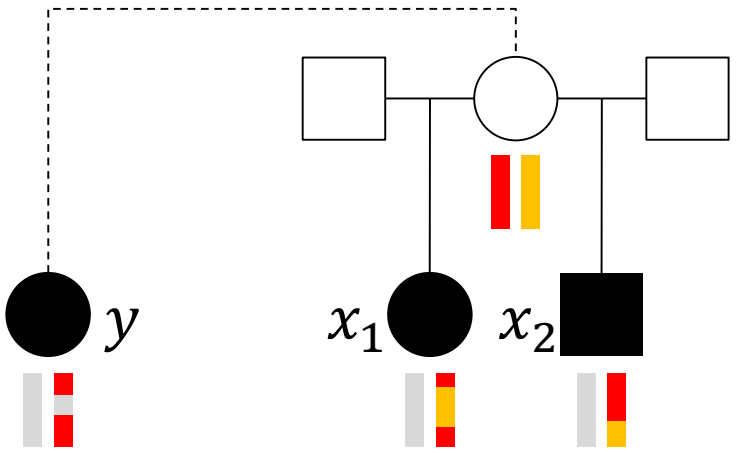


Ying Qiao

Approach: ratios of IBD sharing in three samples versus two

$$R_1 = \frac{Length\left(IBD_{x_1,y} \cap IBD_{x_2,y}\right)}{Length\left(IBD_{x_1,y}\right)}$$

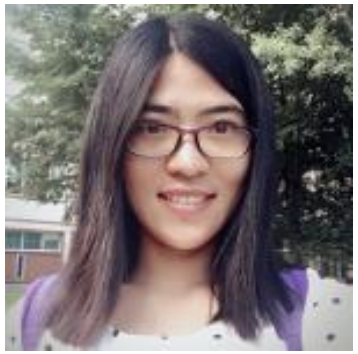
$$R_2 = \frac{Length\left(IBD_{x_1,y} \cap IBD_{x_2,y}\right)}{Length\left(IBD_{x_2,y}\right)}$$



For GP, expect $R_1 = 1/4, R_2 = 1$

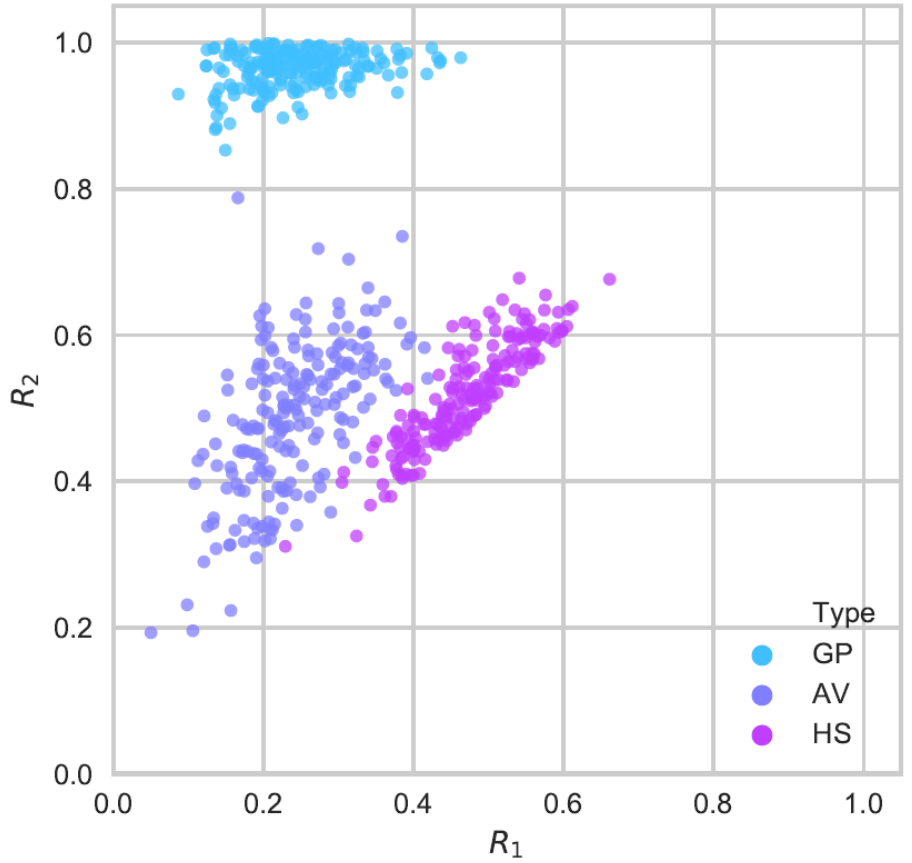
For AV, expect $R_1 = 1/4, R_2 = 1/2$

For HS, expect $R_1 = 1/2, R_2 = 1/2$



Ying Qiao

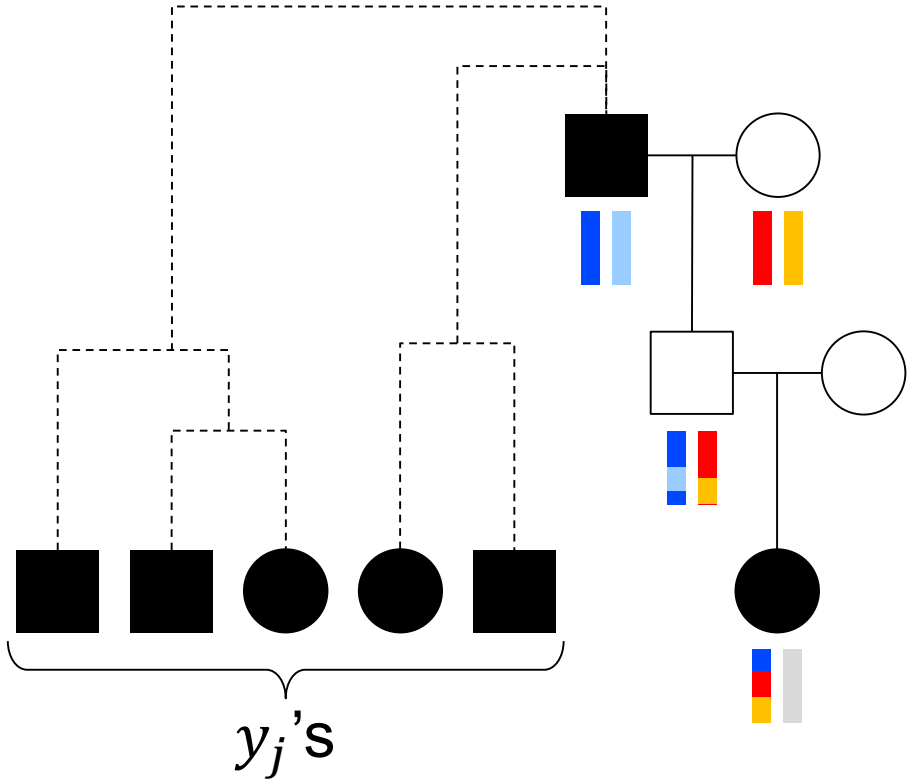
CREST uses kernel density estimators to infer relationships



Trained kernel density estimators (KDEs) using simulated data

Features: R_1, R_2

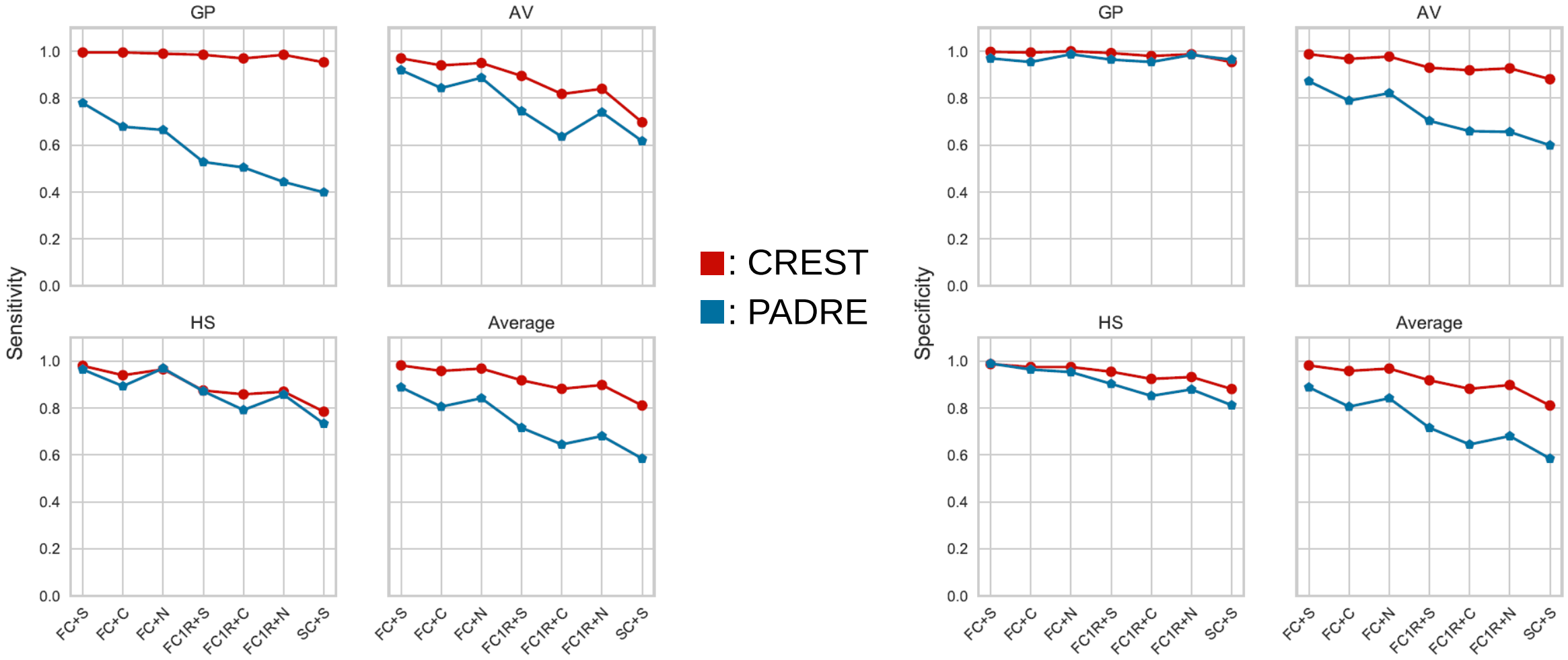
Can combine multiple relatives by taking union of IBD sharing



$$R_i = \frac{Length\left(\left(\bigcup_j IBD_{x_1,y_j}\right) \cap \left(\bigcup_j IBD_{x_2,y_j}\right) \cap IBD_{x_1,x_2}\right)}{Length\left(\bigcup_j IBD_{x_i,y_j}\right)}$$

CREST highly sensitive, highly specific

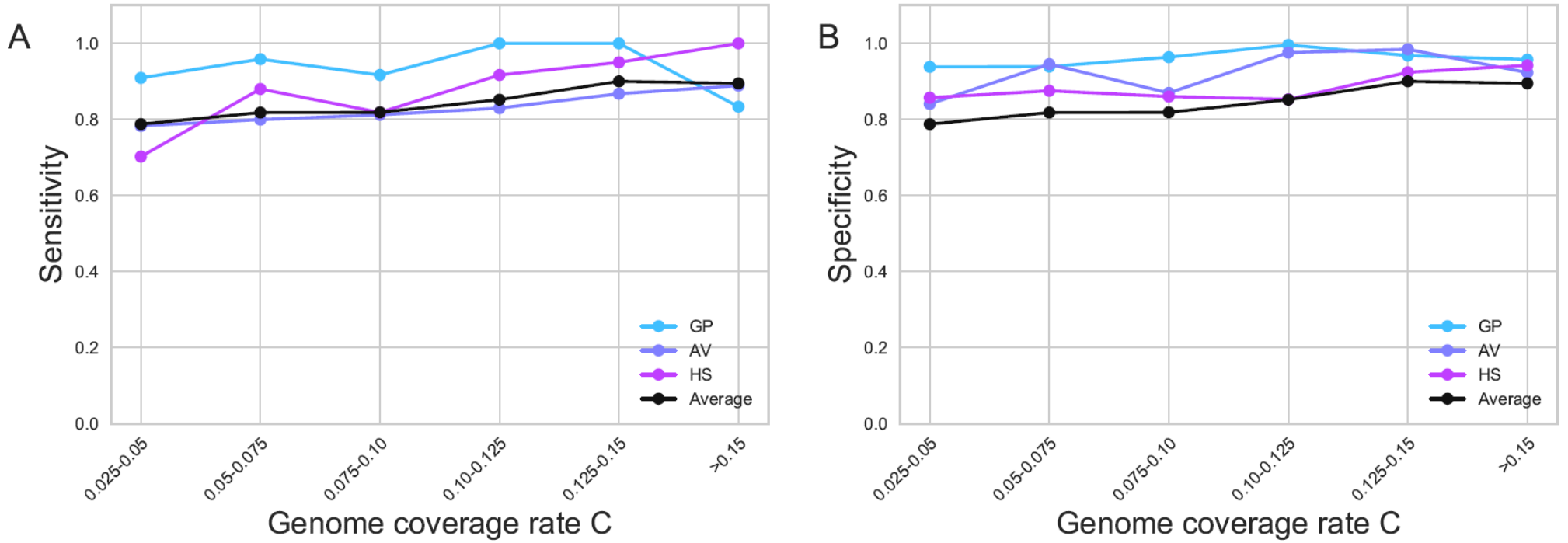
Ran PADRE, CREST on 200 replicates of various pedigree structures



Qiao, Sannerud et al. (in revision, 2019)

CREST infers relative types in Generation Scotland data

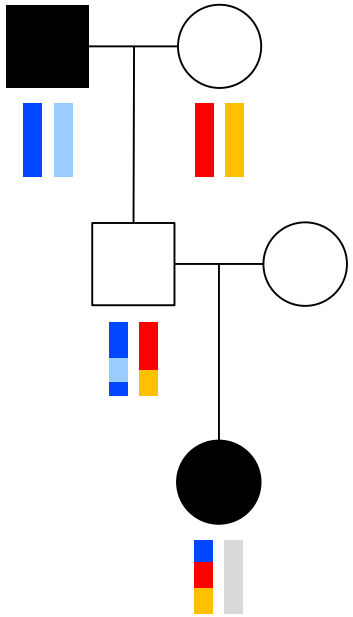
Generation Scotland data:
205 GP, 1,949 AV, and 121 HS pairs with at least one mutual relative



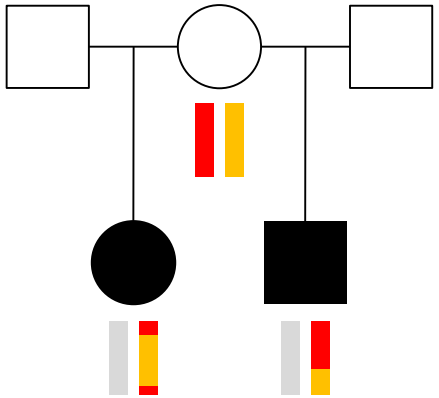
Given data equivalent to one first cousin (10% of genome covered by IBD regions), CREST's sensitivity is 0.99 in GP, 0.86 in AV, and 0.95 in HS pairs

Qiao, Sannerud et al. (in revision, 2019)

Secondary aim: infer whether relatives are paternal or maternal

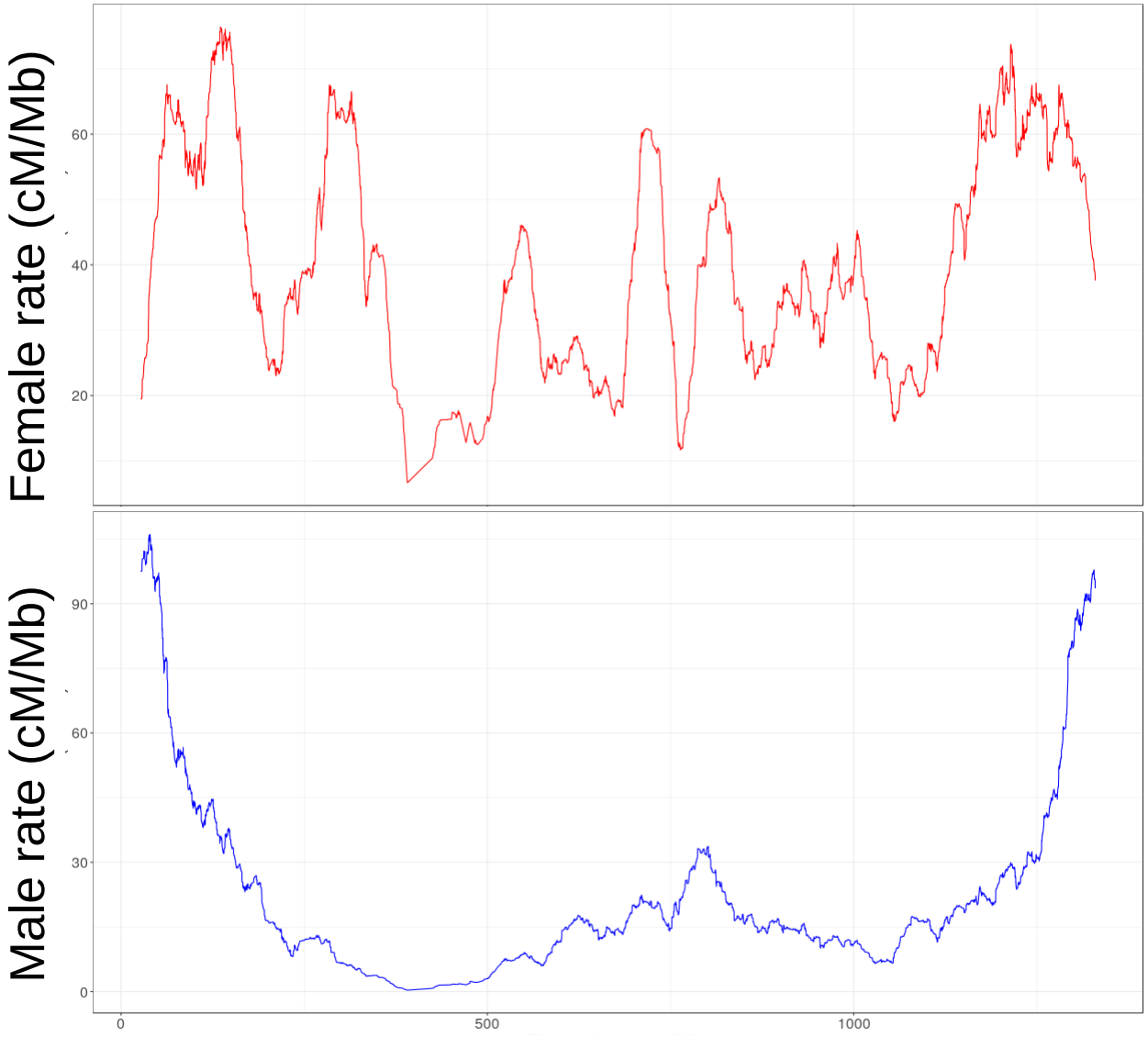


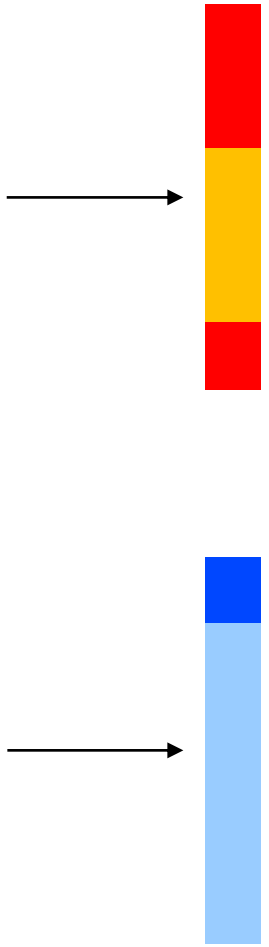
Paternal
Grandparent



Maternal
Half-siblings

Key insight: males / females have different crossover locations





Data from human chromosome 10

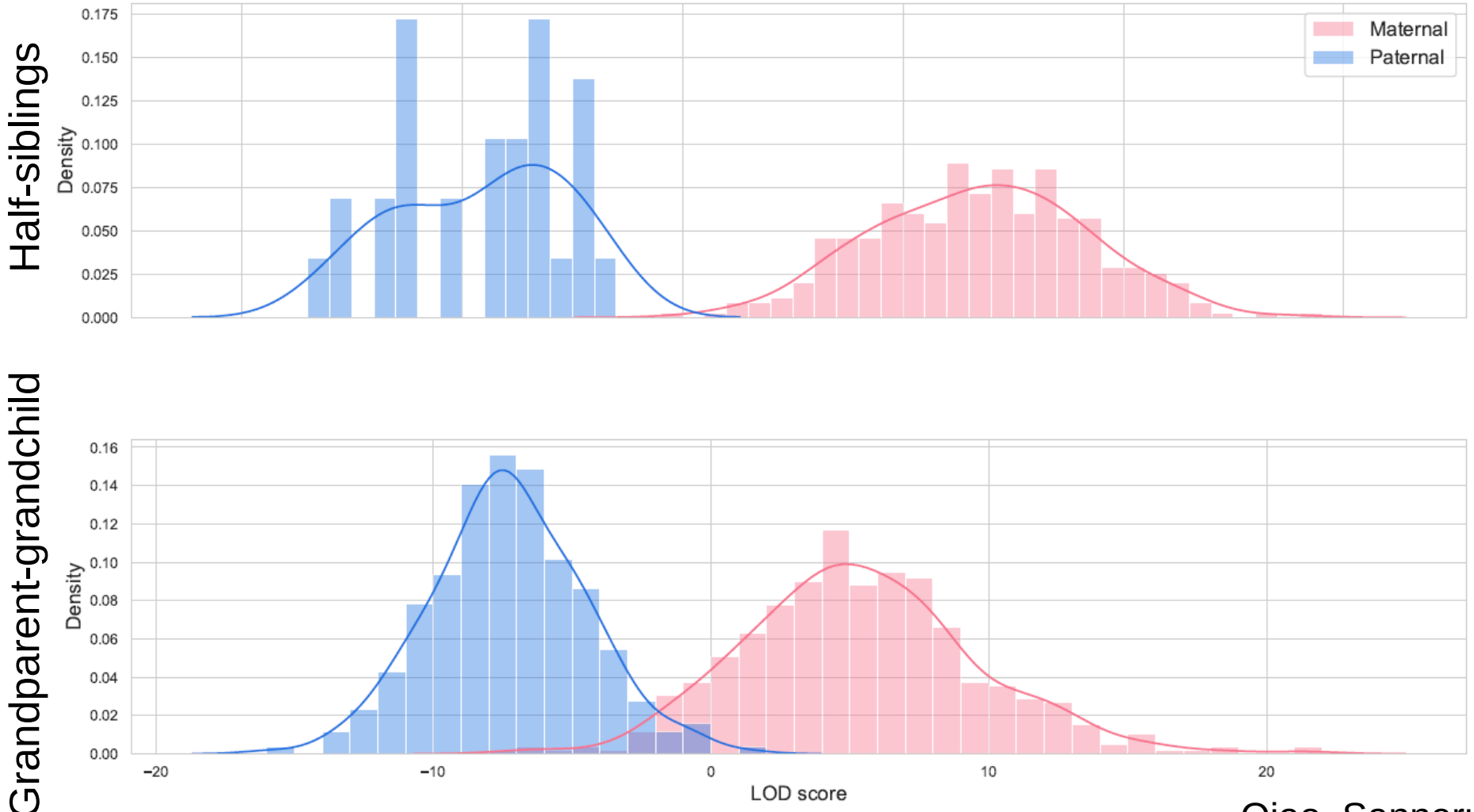
Average number of crossovers:

- Females: 2.04
- Males: 1.27

Genetic map from Bhérer et al. (2017)

CREST infers maternal / paternal type in Generation Scotland

Analyzed all 848 GP and 381 HS pairs in Generation Scotland



Using $LOD = 0$ as boundary:

- 99.7% of HS
- 93.5% of GP

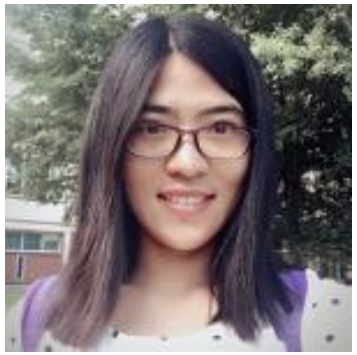
Inferred correctly

Qiao, Sannerud et al. (in revision, 2019)

Conclusions

- CREST classifies second degree relationship types
  - Enabled by multi-way IBD sharing
- Male / female crossovers reveal the paternal / maternal type of half-siblings and grandparent-grandchild pairs
- Can apply to pedigree reconstruction: other methods subject to ambiguities for second degree pairs
- Preliminary results indicate CREST also applies to third degree pairs

Acknowledgements



Ying Qiao

Jens Sannerud

Generation Scotland
Caroline Hayward
Archie Campbell

Alfred P. Sloan FOUNDATION

NIH National Institute of General Medical Sciences

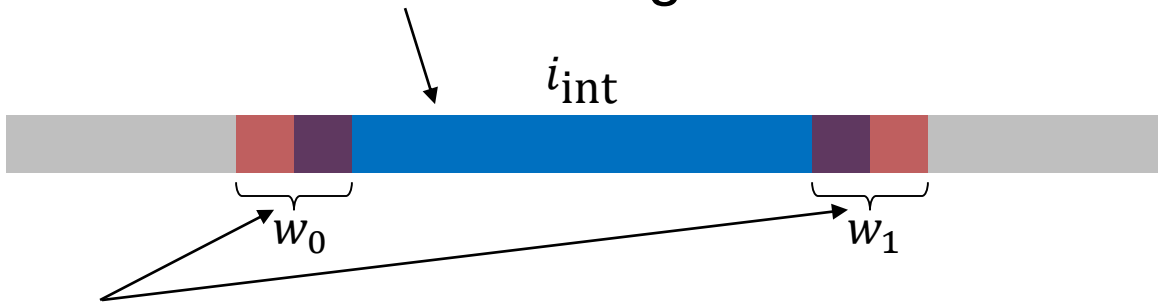
Nancy E. and Peter C. Meinig

Cornell University

Approach: IBD segment ends approximate crossover locations

- Model IBD segments as regions flanked by two crossovers

  No-crossover interval: interior of IBD segment



  Locations of crossovers: window surrounding IBD segment ends

- For each IBD segment i, likelihood of parent being $S \in \{F, M\}$ is

$$P(i|S) = P(w_0|S) \cdot P\left(i_{\text{int}}\middle|S\right) \cdot P(w_1|S)$$

- Taking all IBD segments to be independent, we compute

$$LOD = \log_{10} \frac{\prod_i P(i|F)}{\prod_i P(i|M)}$$

Jens Sannerud