Inferring the genomes of mothers and fathers using genotype data from a set of siblings

Amy L. Williams

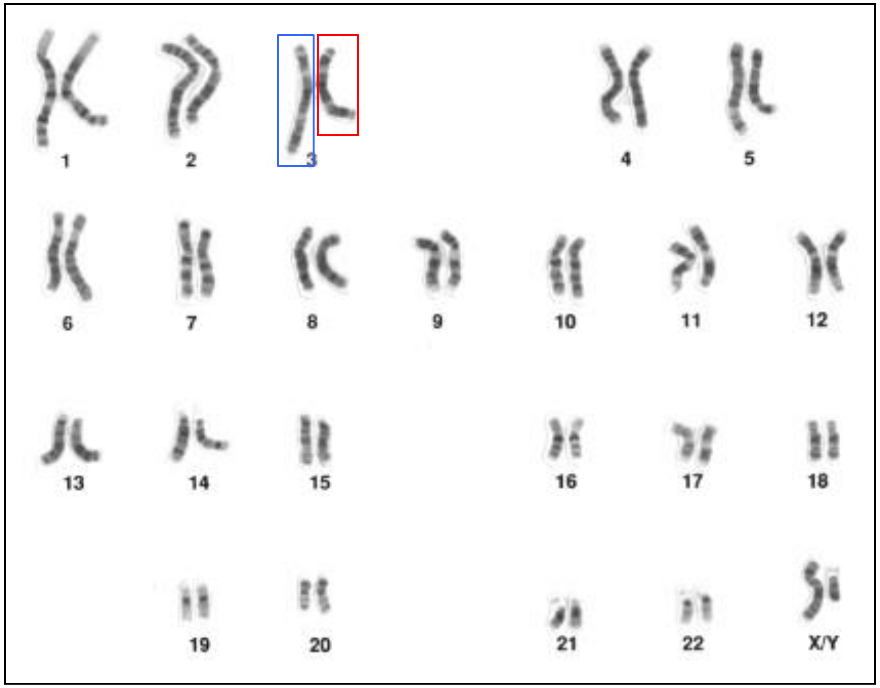
Cornell University

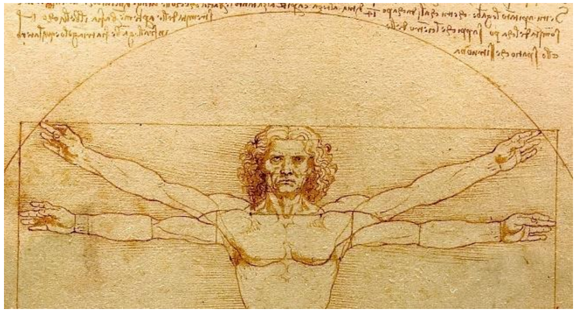
February 7, 2017

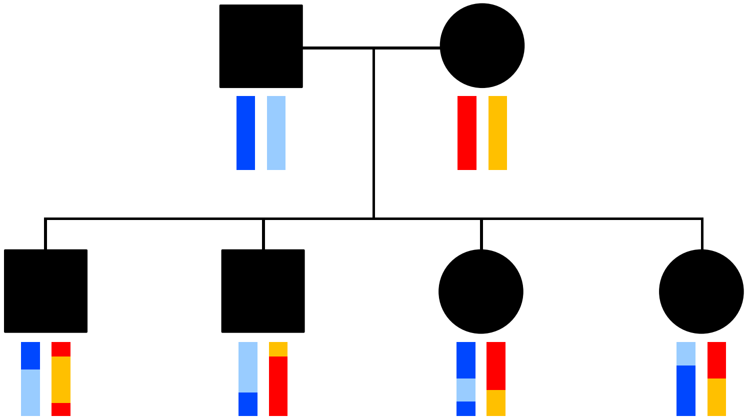
Family History Technology Workshop

Cornell University

Children inherit two chromosome copies: Mosaic of parents’ chromosomes



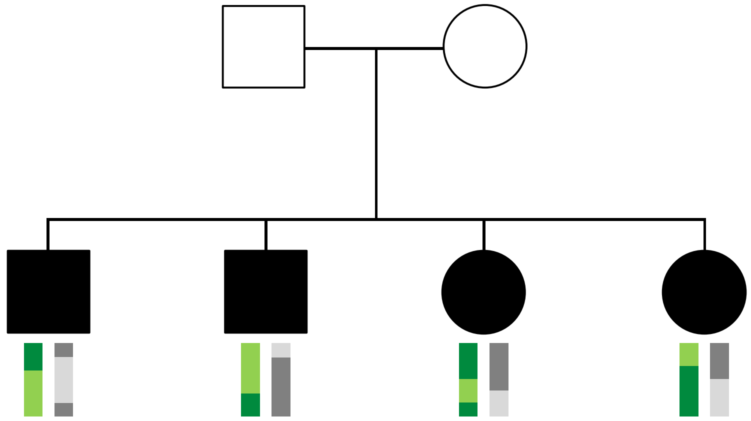




Squares and circles: males and females, respectively
Parents have line joining them and connected to children

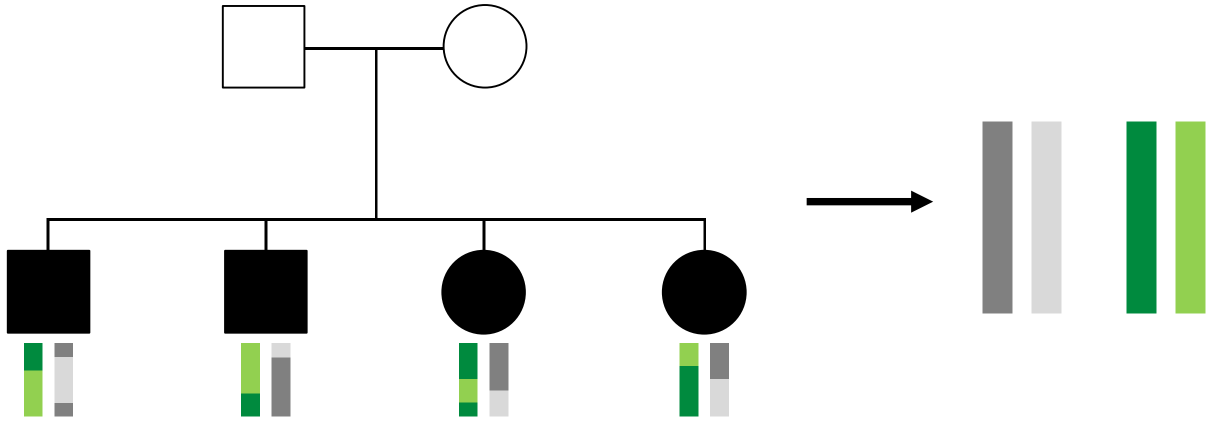
Can infer parents’ chromosomes from siblings … with a catch

- Color coding shown is not built into data
- Can get “color” by comparing siblings’ genomes: identical regions from same chromosome → same “color”



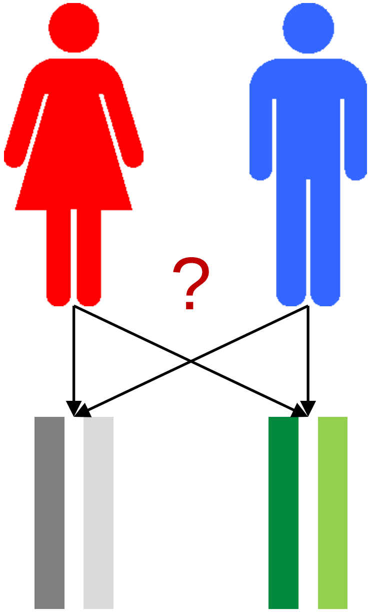
Can infer parents' chromosomes from siblings … with a catch

- Color coding shown is not built into data
- Can get "color" by comparing siblings' genomes: identical regions from same chromosome → same "color"
- Example: can find dark / light green chromosomes and dark / light grey chromosomes
  - Works by stitching together identical regions



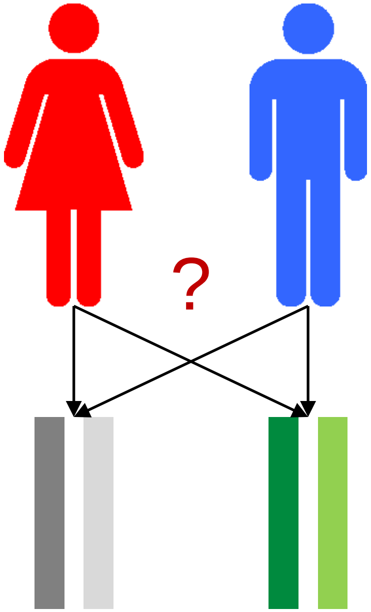
The catch: unclear which chromosome belongs dad / mom

- Can infer a pair of chromosomes that belongs to one parent
- But nothing indicates which chromosome is from dad / mom



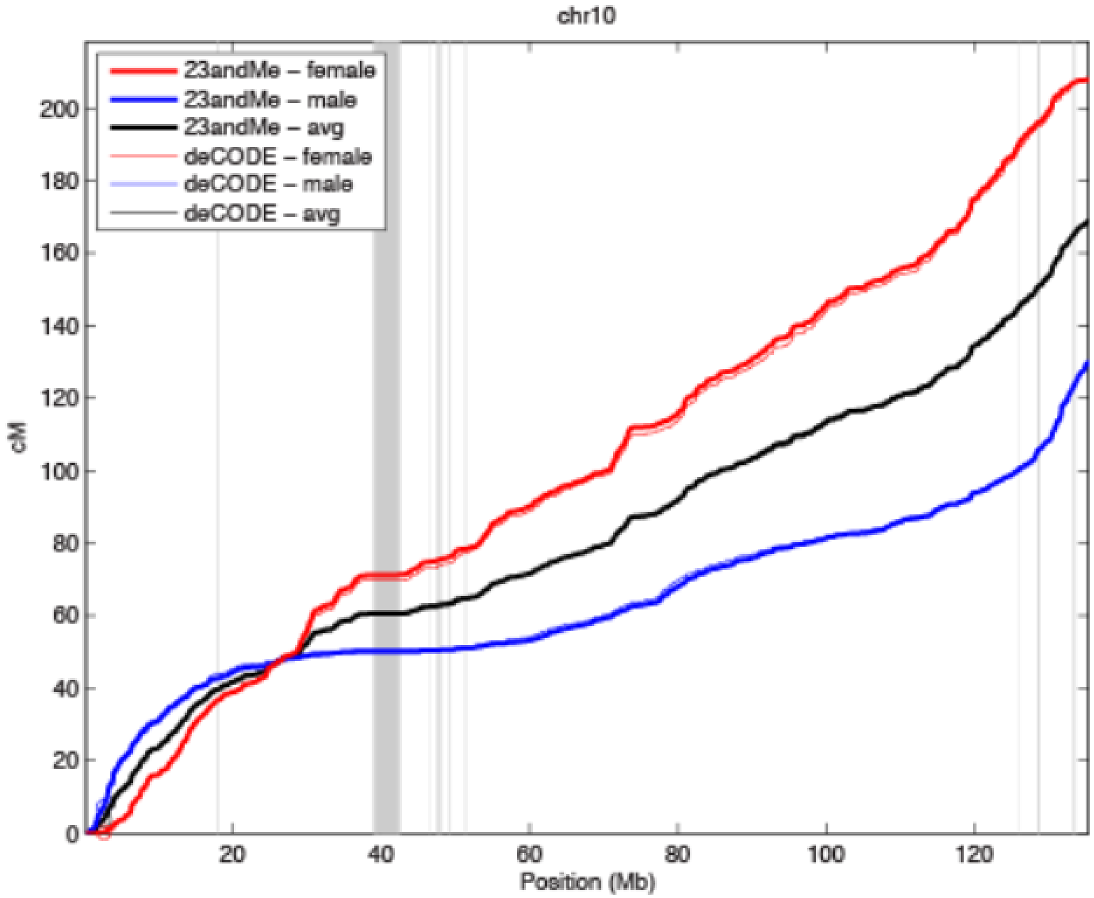
The catch: unclear which chromosome belongs dad / mom

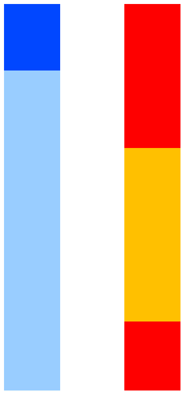
- Can infer a pair of chromosomes that belongs to one parent
- But nothing indicates which chromosome is from dad / mom



- In fact, each chromosome is independent
  - Not just 2 possibilities: $2^{22} > 4$ million possibilities
  - Only true for autosomes: X and Y chromosomes easier

Key insight: men / women produce different mosaic patterns





Y-axis unit is cM: centiMorgan

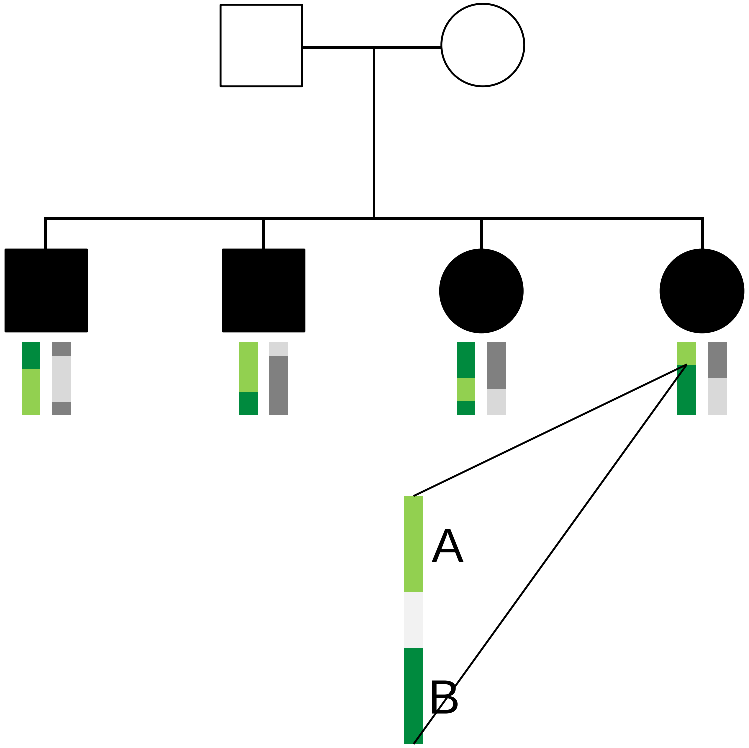
1 Morgan: interval with average of 1 crossover per generation

1 M = 100 cM

Campbell *et al.* (2015)

Step 1: locate crossovers using only siblings

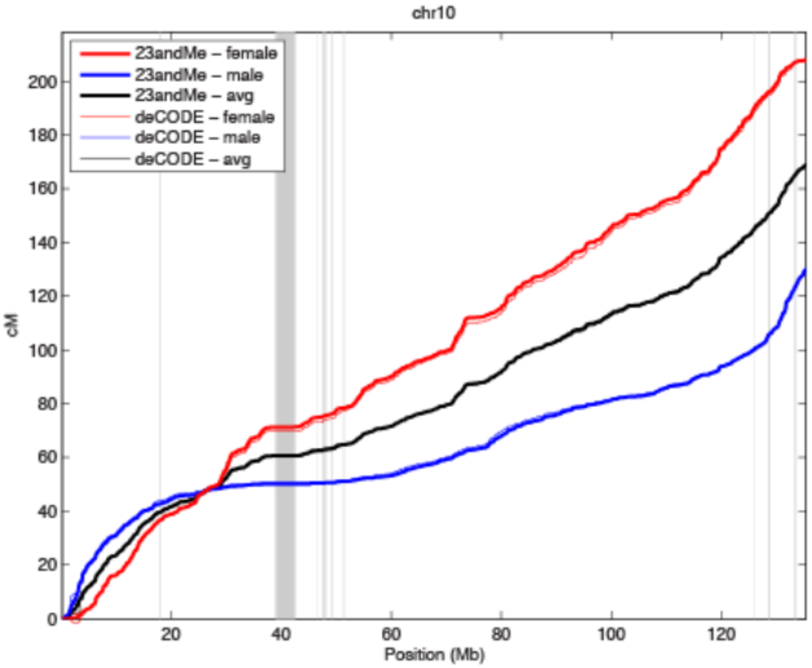
- Using hidden Markov model (HMM), can identify "colors" using only sibling data
  - Structured problem:
    - Four possible chromosomes
    - Two per parent
    - Each child inherits one from each parent at each position



- Get location of crossovers as small window in genome
  - Example: between A and B variants

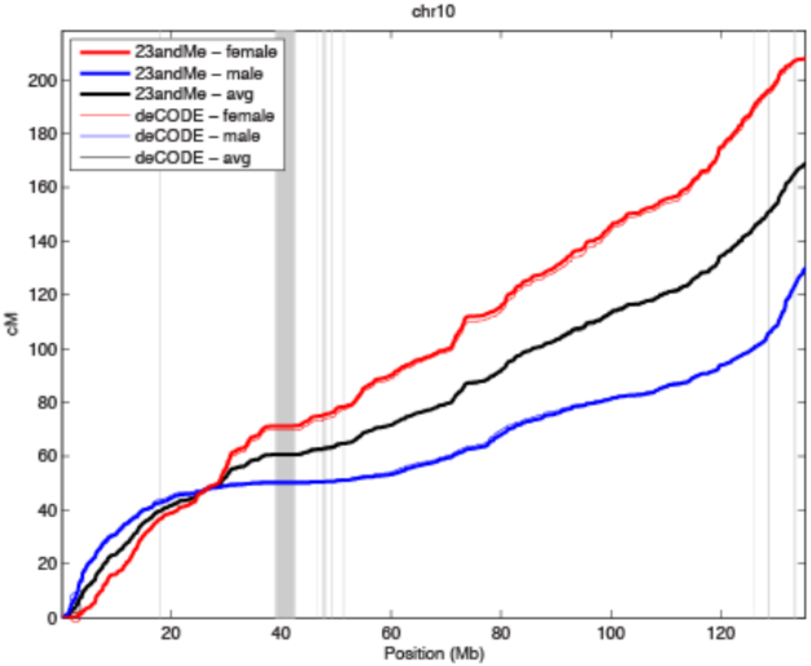
Step 2: define model of data

- Two features in data:
  - Number of transmitted crossovers per child
  - Windows in which crossovers occurred



Step 2: define model of data

- Two features in data:
  - Number of transmitted crossovers per child
  - Windows in which crossovers occurred



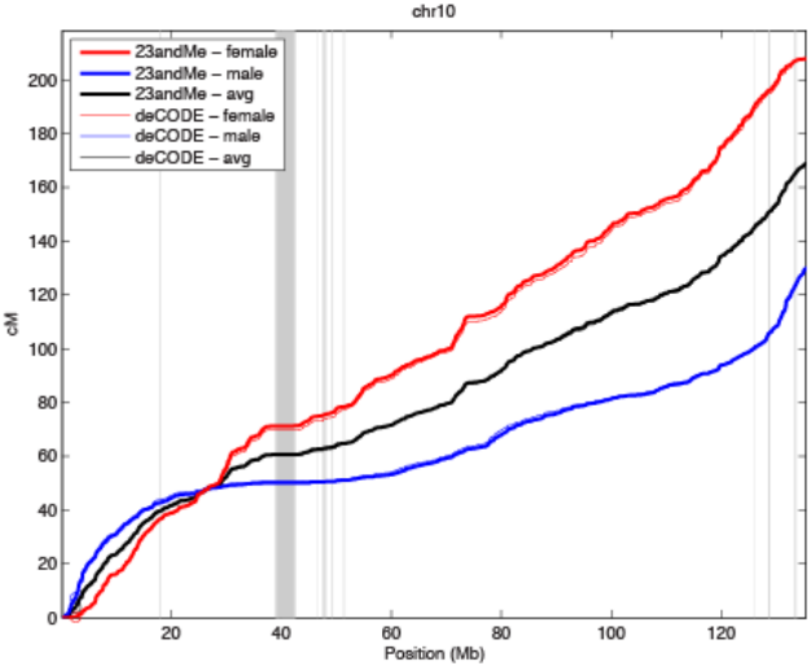
- Model for crossover number:

$$N \sim \mathrm{Pois}(T),$$

$T =$ chromosome length in Morgans male / female

Step 2: define model of data

- Two features in data:
  - Number of transmitted crossovers per child
  - Windows in which crossovers occurred



- Model for crossover number:

$$N \sim \mathrm{Pois}(T),$$

$T =$ chromosome length in Morgans male / female

- Probability of crossover in window length l Morgans:

$$L \sim \mathrm{Exp}(1)$$

$$P(L \leq l) = 1 - \exp(-l)$$

  - In general, l differs between males / females

Step 3: infer male / female origin can treat each child independently

- Data are sets of crossovers inherited by n children:

  $X_1 = (X_{11}, X_{12}, \ldots X_{1n})$

  $X_2 = (X_{21}, X_{22}, \ldots, X_{2n})$

  $X_{pc} = \{w_{pc1}, w_{pc2}, \ldots\}, p \in \{1,2\}, c$ child number

  w_{pcj} indicate window in which crossover j occurred

- Want to compute the following (and the opposite)

  $P(X_1, X_2 | S_1 = F, S_2 = M)$

Step 3: infer male / female origin can treat each child independently

- Data are sets of crossovers inherited by n children:

$$X_1 = (X_{11}, X_{12}, \dots X_{1n})$$
$$X_2 = (X_{21}, X_{22}, \dots, X_{2n})$$
$$X_{pc} = \{w_{pc1}, w_{pc2}, \dots\}, p \in \{1,2\}, c \text{ child number}$$

w_{pcj} indicate window in which crossover j occurred

- Want to compute the following (and the opposite)

$$P(X_1, X_2 | S_1 = F, S_2 = M) = P(X_1 | S_1 = F) P(X_2 | S_2 = M)$$

Step 3: infer male / female origin
can treat each child independently

- Data are sets of crossovers inherited by n children:

  $X_1 = (X_{11}, X_{12}, \ldots X_{1n})$
  $X_2 = (X_{21}, X_{22}, \ldots, X_{2n})$
  $X_{pc} = \{w_{pc1}, w_{pc2}, \ldots\}, p \in \{1,2\}, c$ child number

  w_{pcj} indicate window in which crossover j occurred

- Want to compute the following (and the opposite)

  $P(X_1, X_2 | S_1 = F, S_2 = M) = P(X_1 | S_1 = F) P(X_2 | S_2 = M)$

- Can break into terms for each child:

$$P(X_1 | S_1 = M) = \prod_{c=1}^{n} P(X_{1c} | S_1 = M)$$

Step 3: probabilities for each child use number, locations of crossovers

- Can now apply model and get different probabilities of male / female origin for each crossover

$$P(X_{1c}|S_1 = M) = P\left(N_{S_1} = |X_{1c}|\right) \times \prod_{w_{1cj} \in X_{1c}} P\left(L \le Rec\left(w_{1cj}, S_1\right)\right)$$

$Rec(w, S)$: probability of crossover in window w in $S \in \{M, F\}$

Results

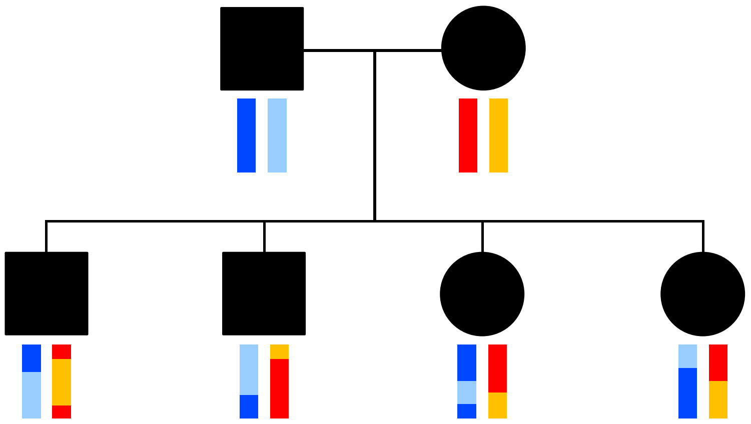
- Data: San Antonio Family Studies
  - Total: 2,490 genotyped samples, 80 pedigrees
  - Analyzed 69 families, 3 to 12 children
    - Include data for both parents to check accuracy
  - Genotypes from 888,748 SNPs (variants)
- In 1,518 chromosomes, posterior probabilities of correct configuration:

	Full model	Poisson	Crossover windows
> 0.5	1,515	1,099	1,513
> 0.9	1,513	372	1,511

One issue… currently finding crossovers with parent data

- These results based on finding crossovers with parent data
  - Is cheating, but will fix soon

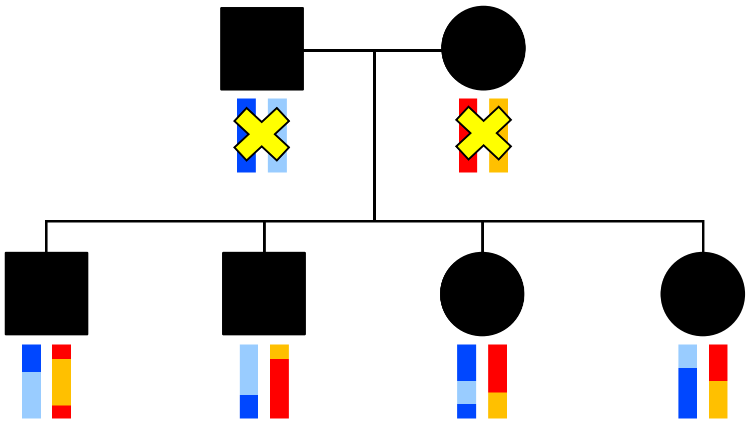
- For > 8 children should generally do this well
  - Basically perfect results



One issue… currently finding crossovers with parent data

- These results based on finding crossovers with parent data
  - – Is cheating, but will fix soon

- For > 8 children should generally do this well
  - ➢ Basically perfect results



- Fewer siblings: some portions of genome will be ambiguous
  - – But substantial parts will not be

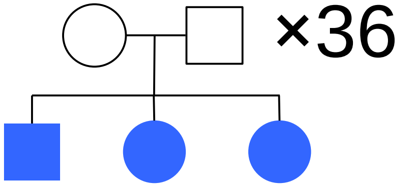
➢ Will have accuracy results for only siblings in coming weeks

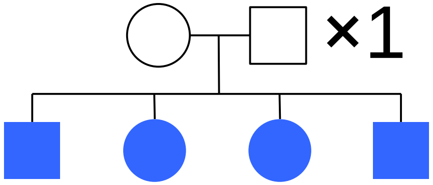
Applications: large datasets

- Used new method Attila to identify pedigrees in large cohorts

biobank uk

152,095 samples

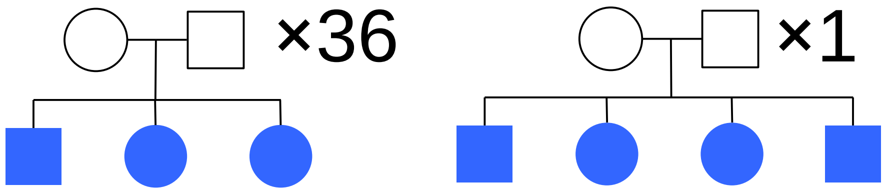




Applications: large datasets

- Used new method Attila to identify pedigrees in large cohorts

152,095 samples



- Why not get DNA from everyone in the world?
    1. Find siblings
    2. Infer parents' genomes
    3. Repeat 1 & 2 for many generations

Acknowledgements

Sayantani Basu-Roy

Ryan O’Hern

Funding: Cornell University — Alfred P. Sloan FOUNDATION

Cornell seed grant
Meinig Family Investigator Award

Postdoc and graduate student openings