Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Shared haplotype estimation using rare variants to identify tracts of common ancestry

Supervised by

Gil McVean

UNIVERSITY OF OXFORD

Patrick K. Albers
DPhil student, Genomic Medicine and Statistics
Wellcome Trust Centre for Human Genetics
University of Oxford

Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Estimating the age of a rare mutation

(the time in generations since mutation gave rise to the allele)

This is attempted by inferring the

**haplotype region that is identical by descent surrounding the focal rare allele**

It is necessary to establish

**correct phasing around the focal rare allele**

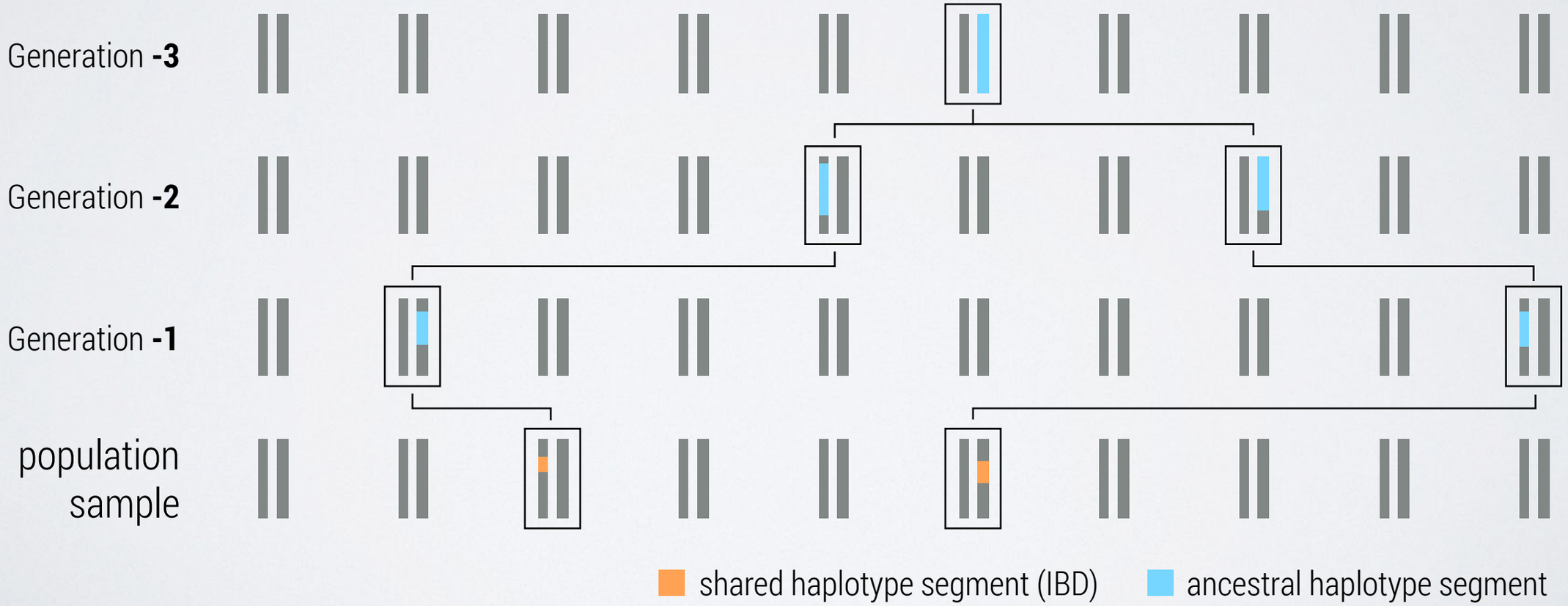
UNIVERSITY OF OXFORD

Patrick K. Albers
DPhil student, Genomic Medicine and Statistics
Wellcome Trust Centre for Human Genetics
University of Oxford

# Identity by decent **(IBD)**

- Individuals sharing a mutation are likely to have inherited this mutation from a **common ancestor**
- They must have also inherited the chromosomal **region surrounding the focal mutation**
- These individuals share a haplotype region **identical by descent (IBD)**
- The length of a IBD segment is **broken down by recombination**

- Breakpoints are **not directly observable** from genotype or haplotype data
- Identification of IBD regions is difficult because each chromosome represents a **mosaic of haplotypes** with different ancestry backgrounds
- Correct identification is confounded by regions that are **identical by state (IBS)**






Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Rare or low-frequency variants

(e.g. frequency < 0.05)

– represent the **majority of variation**
in the human genome

– are typically **young in age**
likely to have arisen recently through mutation

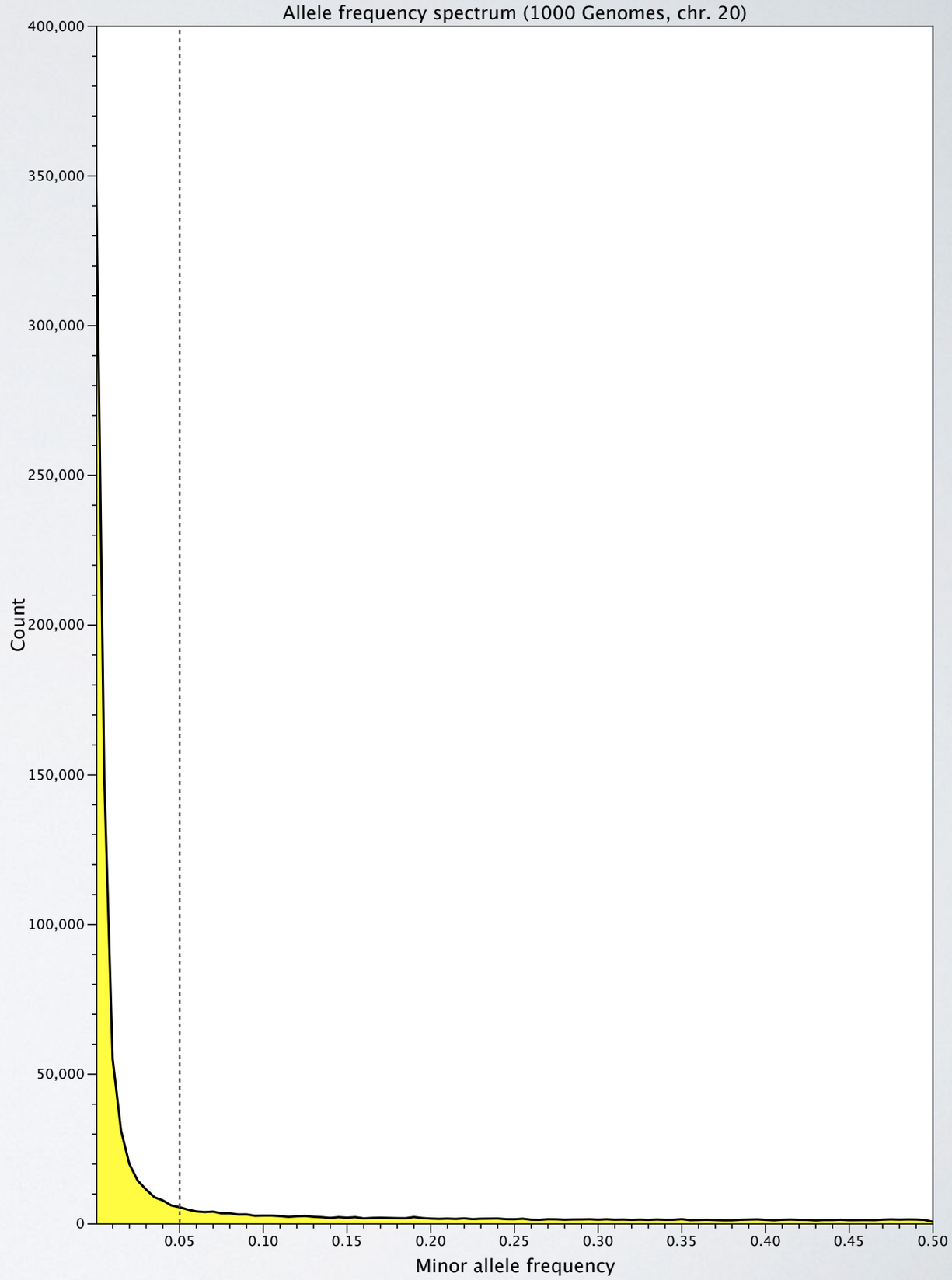
– are likely to be **population-specific**
due to likely recent origin

– are informative for **relatedness**
due to a likely common ancestor

– they are likely to sit in
**long shared haplotype segments**
because recombination had
**less time to break down**
the ancestral full length of IBD

Rare variants are useful to

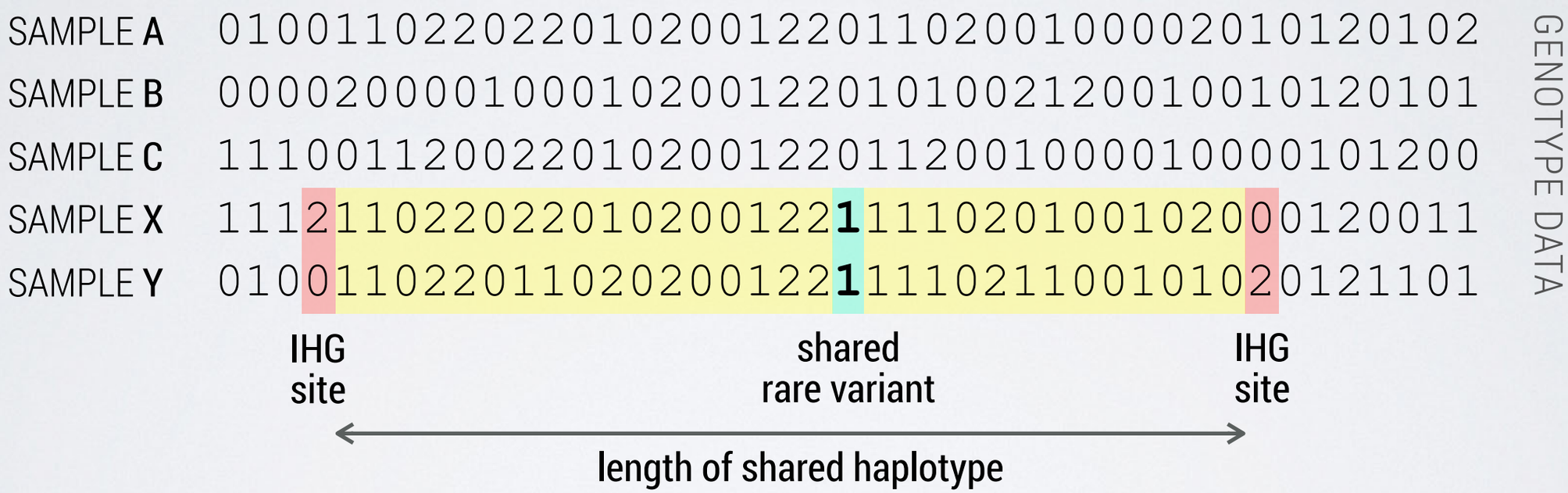
- identify individuals that are likely to share a recent common ancestor
- tag haplotypes of shared ancestry
- resolve recent demographic history



# Shared haplotype **boundaries**

**Mathieson & McVean (2014)** described a method to estimate the length of shared haplotypes for variants shared by **2** individuals, i.e. doubletons or **$f_2$ variants**

The length of the shared haplotype is delimited by **inconsistent homozygote genotypes** (IHG)



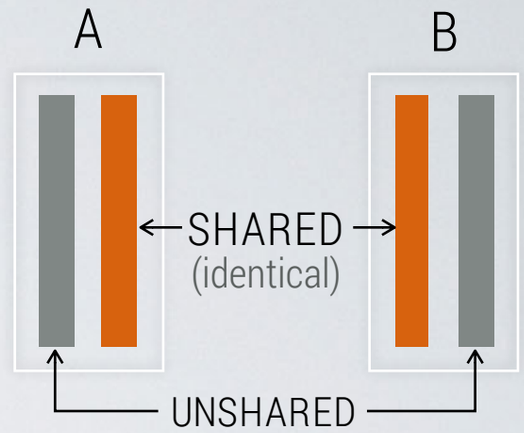

Inferred breakpoints
- provide a first indication for IBD boundaries
- can only be explained by recombination (or genotyping error)

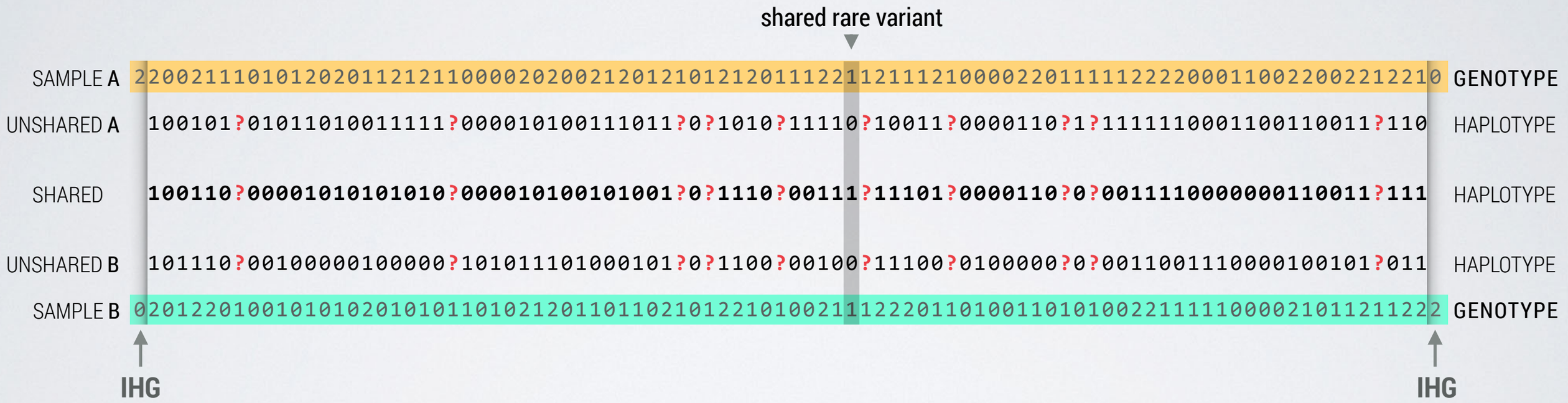

Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Shared haplotype **estimation**

A genotype represents two unphased haplotypes, one from each chromosome

In a subsample of **2** individuals that share a particular rare variant
there are **1 shared haplotype** and **2 unshared haplotypes**




Available haplotypes in the set of genotypes are distributed so that
- the common haplotype is shared
- the remaining haplotype is unshared

**?** = heterozygous genotypes

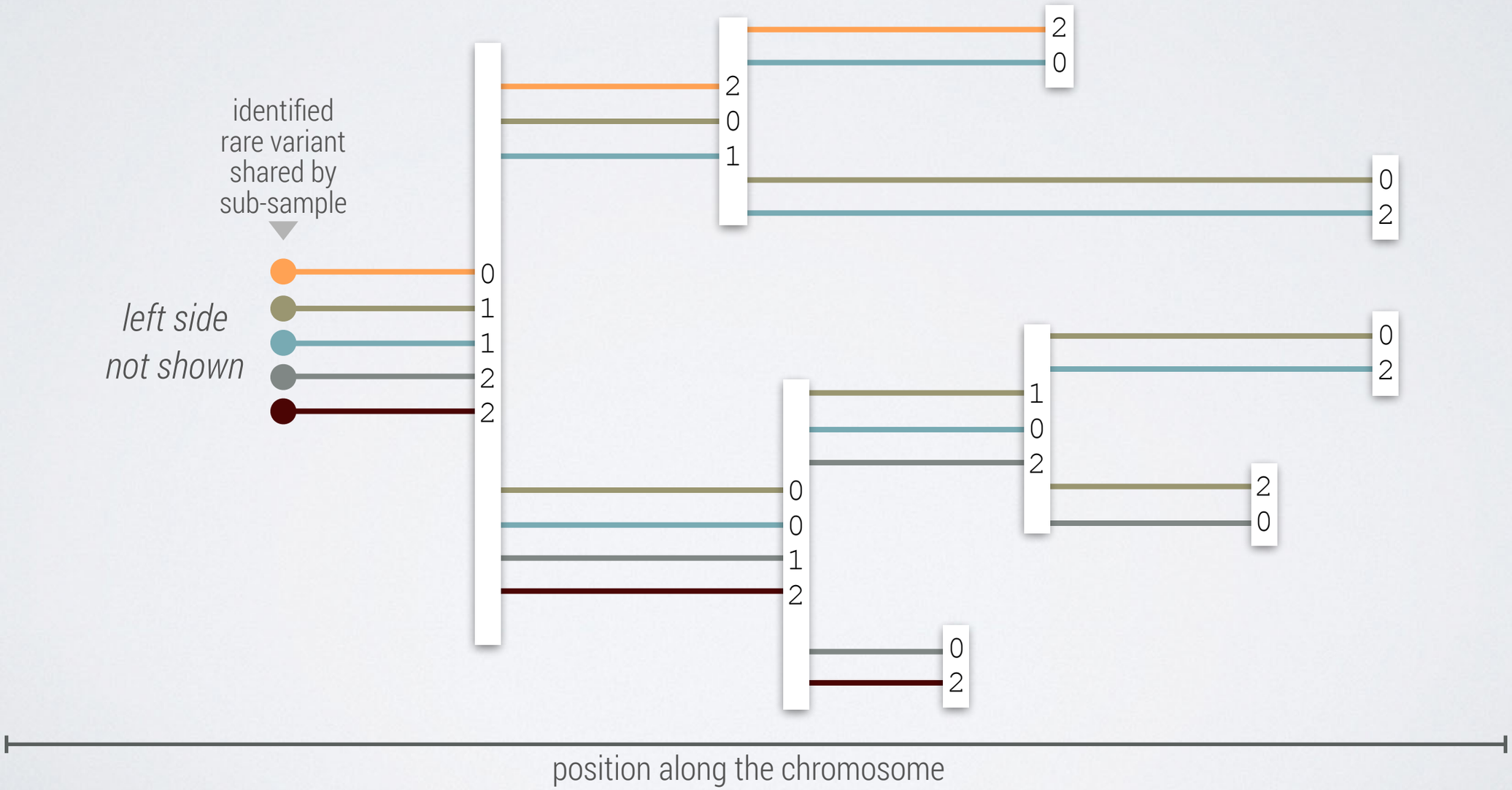


Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Shared haplotype **structures**

Generalising the method by **Mathieson & McVean (2014)**
for variants shared by **k** individuals, i.e. **$f_k$ variants**

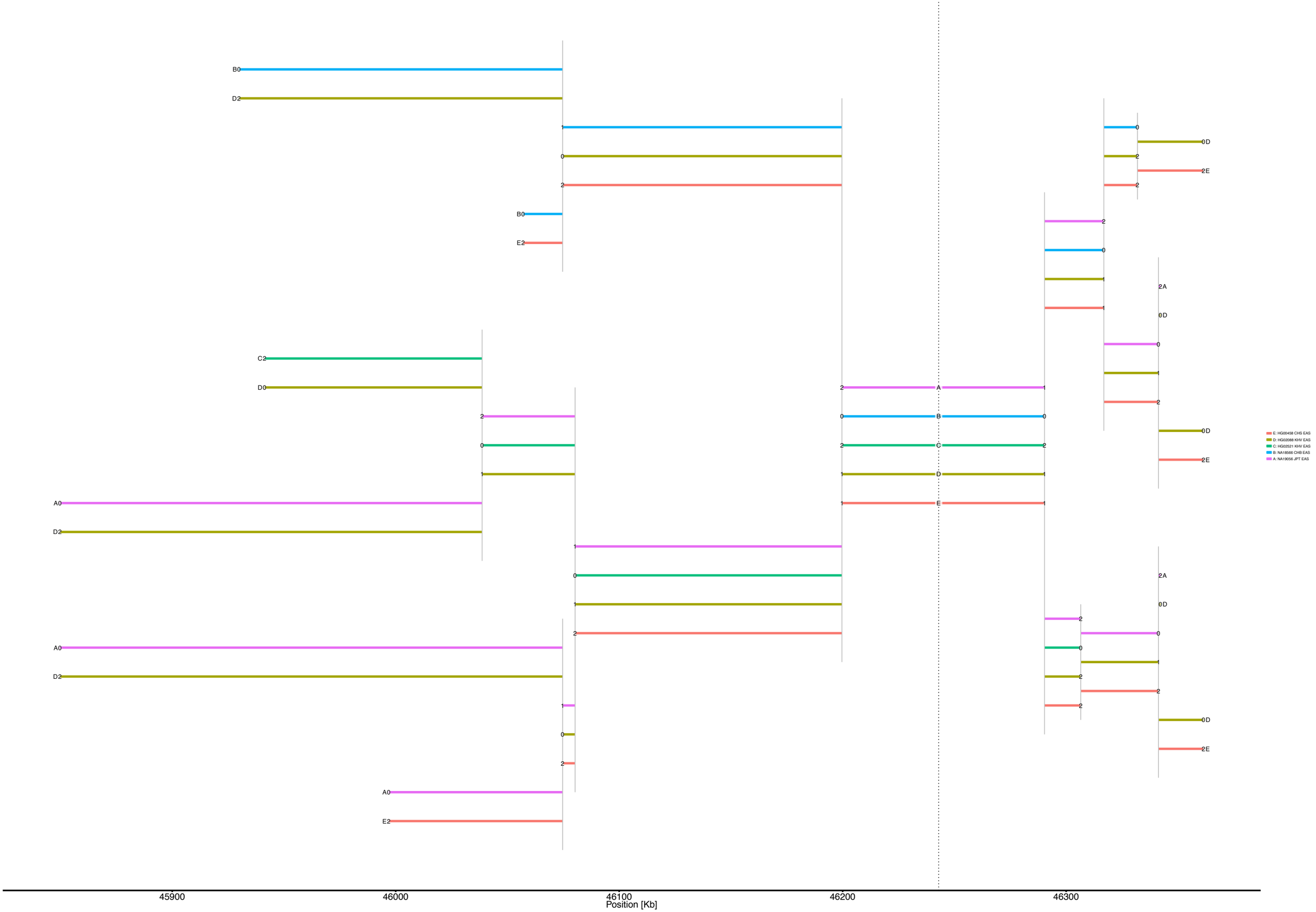
A **shared haplotype structure** is constructed by
searching **for all paths** in the subsample sharing a focal mutation
that are **sharing any haplotype** along the chromosome






Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# f 5






Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

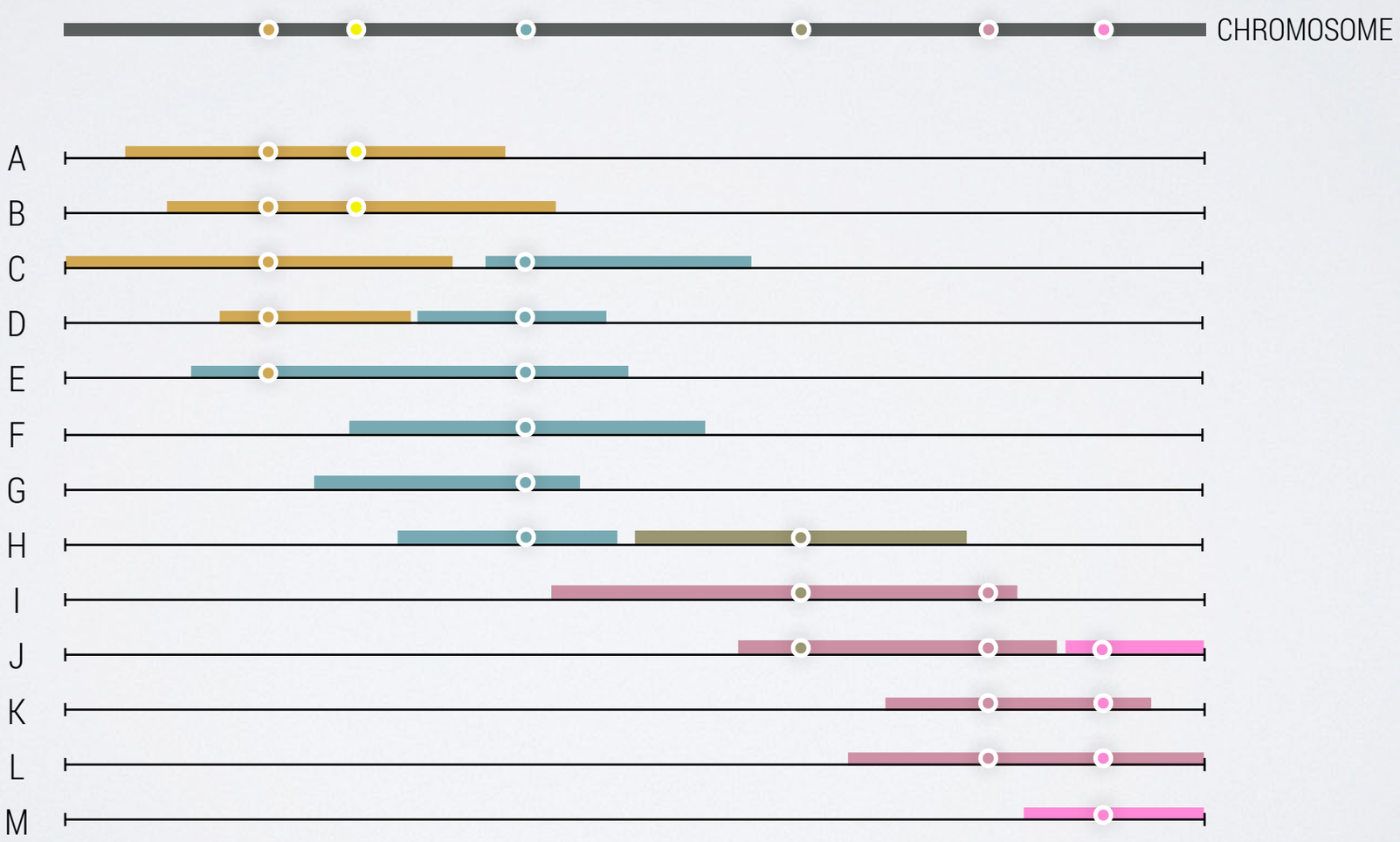
# Shared haplotype **phasing**

More information about shared haplotypes is available from **$f_k$ variants nearby** that sit on the **same haplotype**

- Overlapping ranges of other segments inform *shared haplotype estimation*
- Rare variants nested within another shared variant subsample inform *mutational timing*
- Sharing patterns across different shared variant subsamples inform *estimation of demographic history*

SAMPLE **X**

sharing **$f_k$ variants** with any of the other samples

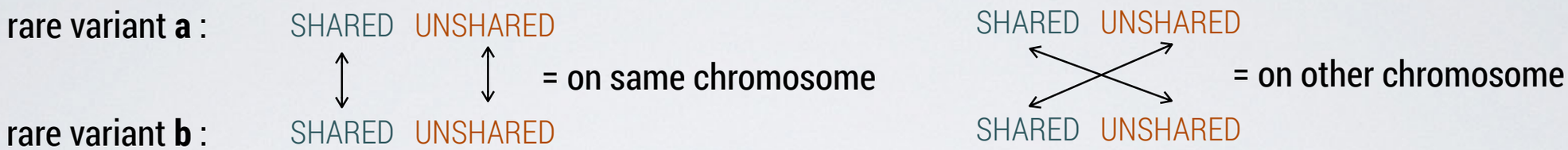




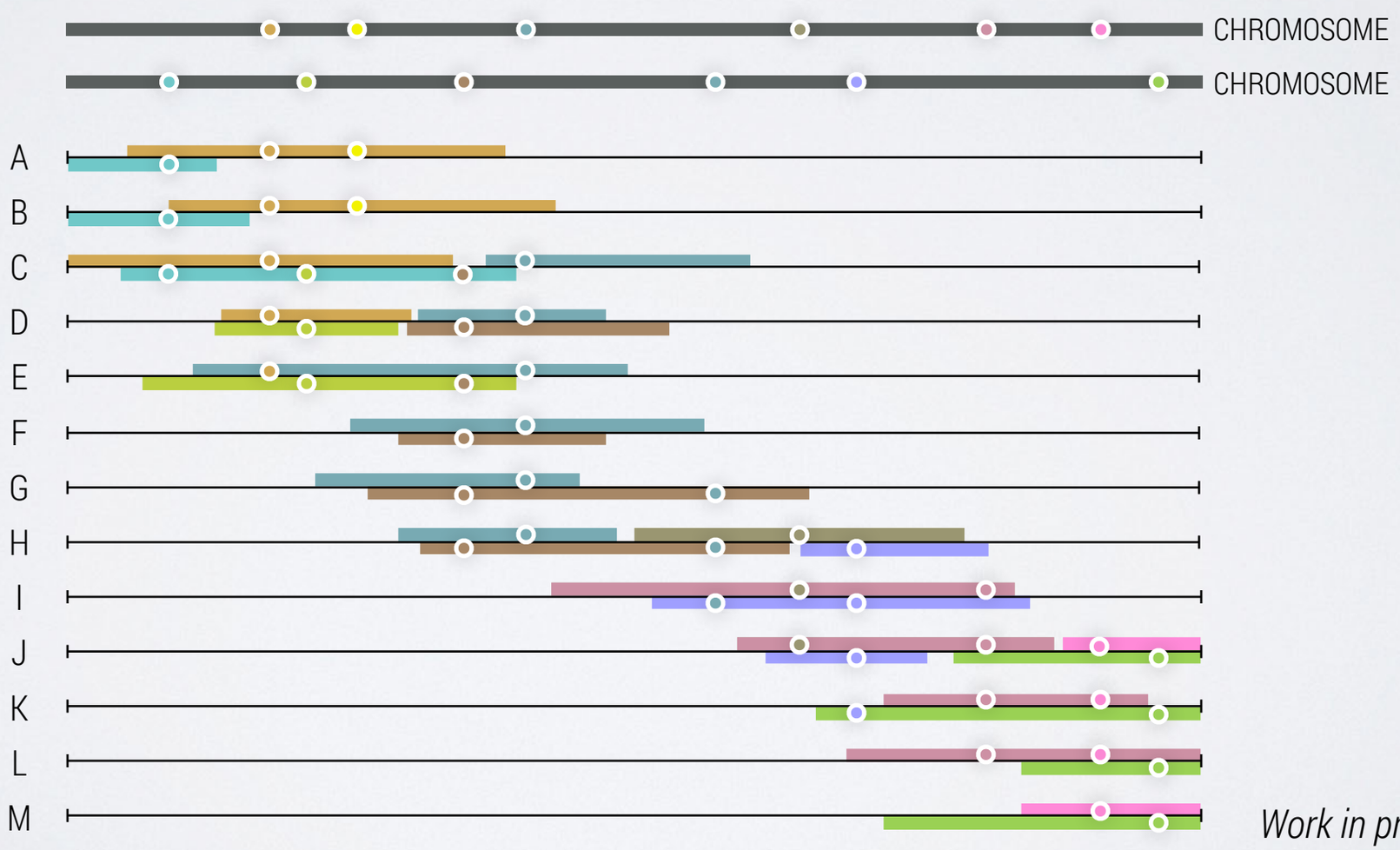

Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Shared haplotype **phasing**

Haplotype phase can be estimated for each individual based on
**match and mismatch** between estimated **shared and unshared haplotypes**




SAMPLE **X** sharing **$f_k$ variants** with any of the other samples




*Work in progress:*
*Hidden Markov Models (HMM)*



Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

# Quantitative Genomics

29 May 2015

Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

## Conclusion

A **new phasing algorithm** around rare mutations is a necessary step for detecting shared haplotype segments, which …

+ becomes more accurate for phasing of and around rare variants
+ becomes more accurate for densely typed samples and larger sample sizes
– but may not stretch the whole chromosome

The **age of a particular rare variant** can be estimated through …

– its allele frequency (insufficient by itself)
+ length of its shared haplotype segment (*recombination clock*)
+ diversity along the segment in samples sharing the allele (*mutation clock*)
+ local relatedness structures (*genealogical tree*)
– but estimates may have very large confidence intervals

UNIVERSITY OF OXFORD

Patrick K. Albers
DPhil student, Genomic Medicine and Statistics
Wellcome Trust Centre for Human Genetics
University of Oxford

Patrick K. Albers pkalbers@well.ox.ac.uk @pkalbers

Thank you for your attention

# Shared haplotype estimation using rare variants to identify tracts of common ancestry

Supervised by

Gil McVean

UNIVERSITY OF OXFORD

Patrick K. Albers
DPhil student, Genomic Medicine and Statistics
Wellcome Trust Centre for Human Genetics
University of Oxford