Identifying recurrent mutations in population-level sequencing data

Kelsey Johnson
SAGES
June 1st, 2018

What is a recurrent mutation?

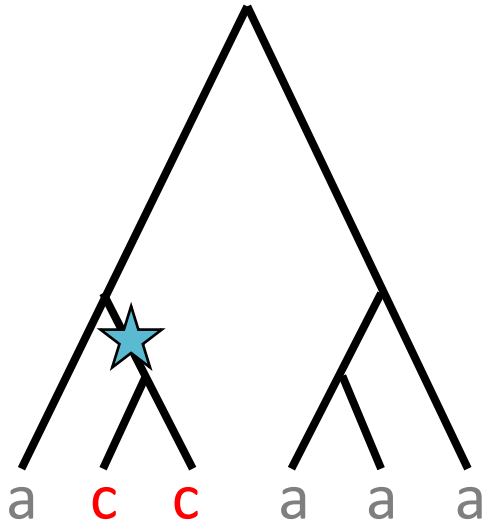
acggaagctag
acggaagctag
acggaagctag
acggaagctag
acggacgctag
acggacgctag
acggaagctag
acggaagctag

What is a recurrent mutation?

acggaagctag
acggaagctag
acggaagctag
acggaagctag
acggacgctag
acggacgctag
acggaagctag
acggaagctag

★ = mutation event

Identical by descent (IBD):

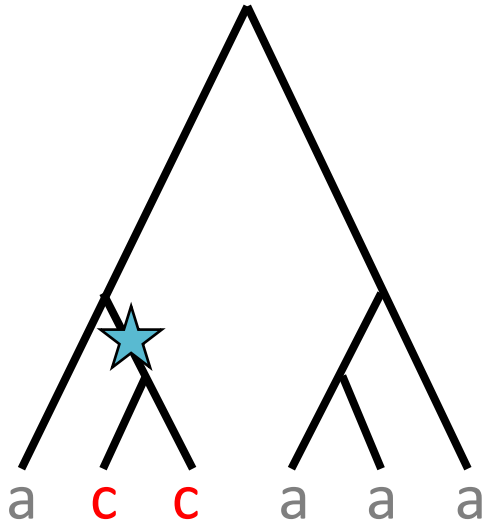


What is a recurrent mutation?

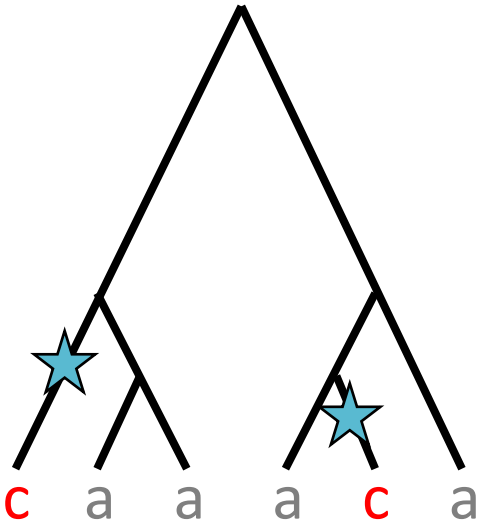
acggaagctag
acggaagctag
acggaagctag
acggaagctag
acggacgctag
acggacgctag
acggaagctag
acggaagctag

★ = mutation event

Identical by descent (IBD):



Recurrent:



Why care about recurrent mutations?

Recurrent mutations are a hallmark of some Mendelian diseases

Gene	Disease
CFTR	cystic fibrosis
SCN8A	epileptic encephalopathy
PKD1	polycystic kidney disease
FGFR1	Pfeiffer syndrome
FGFR3	achondroplasia
LMNA	Hutchinson–Gilford progeria syndrome

Recurrent mutations are used to identify genes associated with complex disease

nature COMMUNICATIONS

ARTICLE

Received 15 Sep 2014 | Accepted 16 Oct 2014 | Published 24 Nov 2014

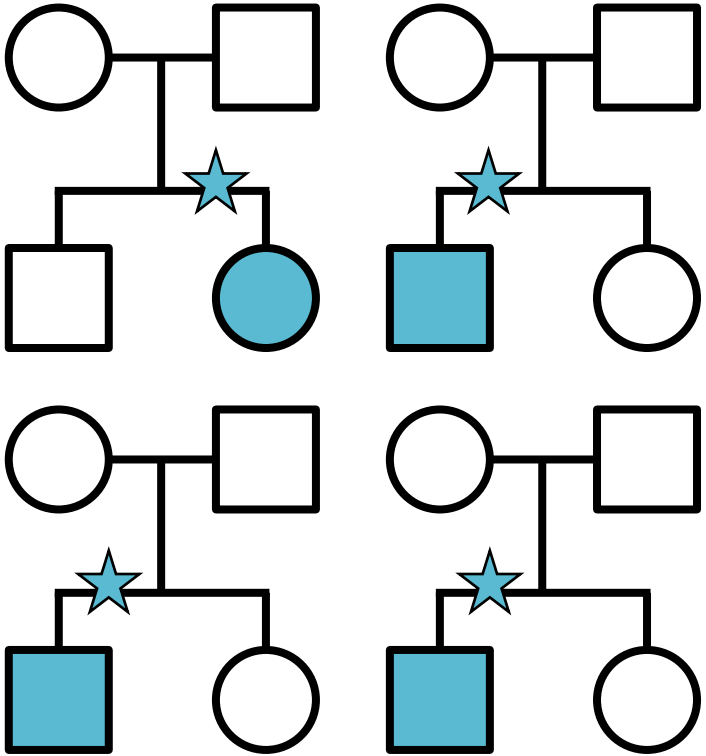
DOI: 10.1038/ncomms6595

Recurrent *de novo* mutations implicate novel genes underlying simplex autism risk

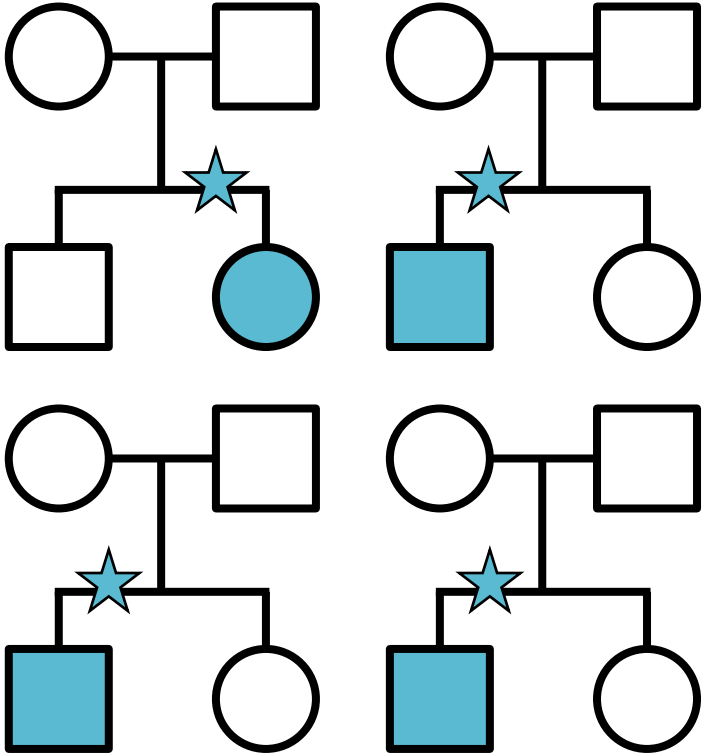
B.J. O'Roak[1,†,*], H.A. Stessman[1,*], E.A. Boyle[1], K.T. Witherspoon[1], B. Martin[1], C. Lee[1], L. Vives[1], C. Baker[1], J.B. Hiatt[1], D.A. Nickerson[1], R. Bernier[2], J. Shendure[1] & E.E. Eichler[1,3]

These studies rely on family-based sequencing to identify recurrent mutations

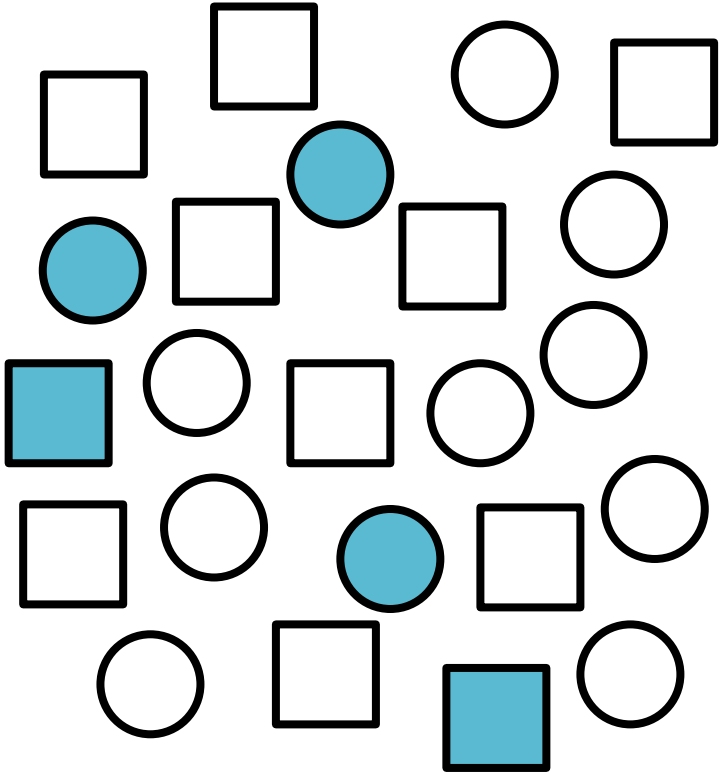
Family-based study



Family-based study



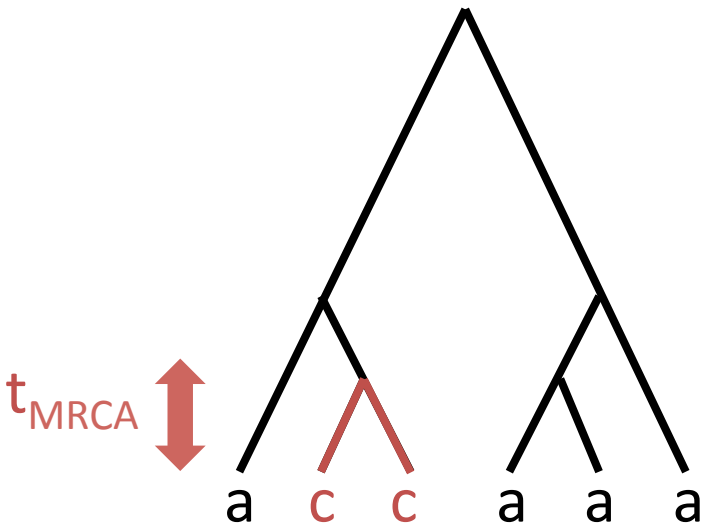
Population-based study



What features can distinguish recurrent and IBD alleles?

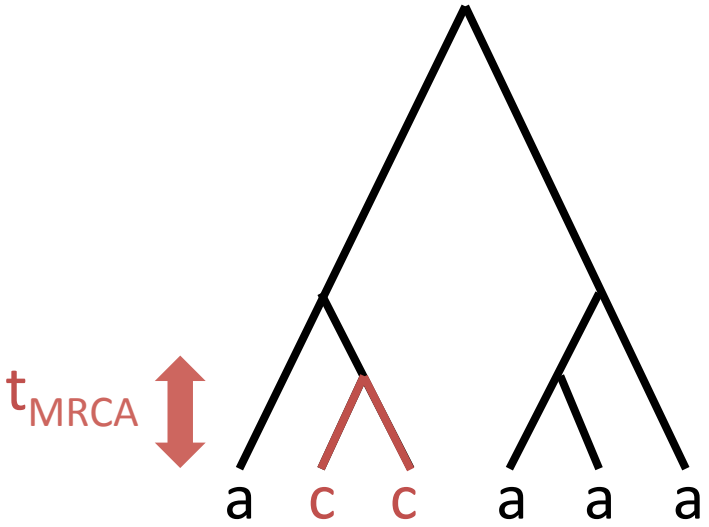
Differences in t_{MRCA} for IBD vs. recurrent alleles

Identical by descent (IBD):

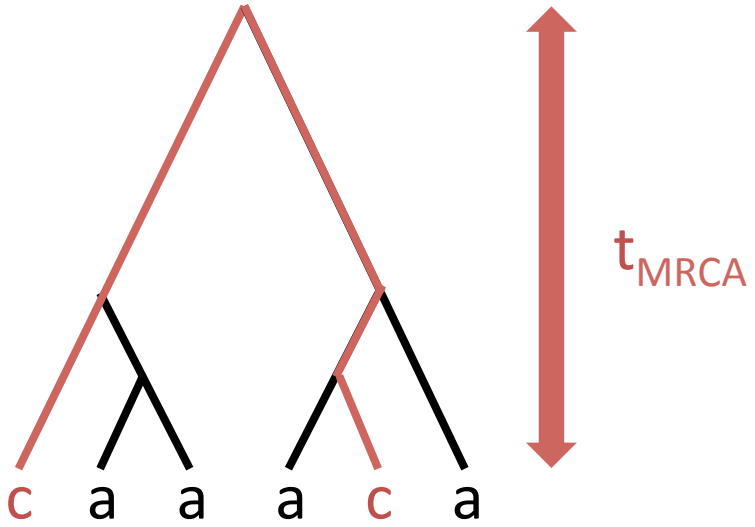


Differences in t_{MRCA} for IBD vs. recurrent alleles

Identical by descent (IBD):



Recurrent:



Population-level sequencing data with diploid genotypes

0 0 0 0 1 0 0 1 0 0 1 0 0 2 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1

0 1 0 0 2 0 0 0 0 0 0 0 0 0 1 0 **1** 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 0 1

0 0 0 0 1 0 0 0 0 0 1 0 0 2 0 0 0 0 0 1 0 0 2 0 0 1 0 0 1 0 2

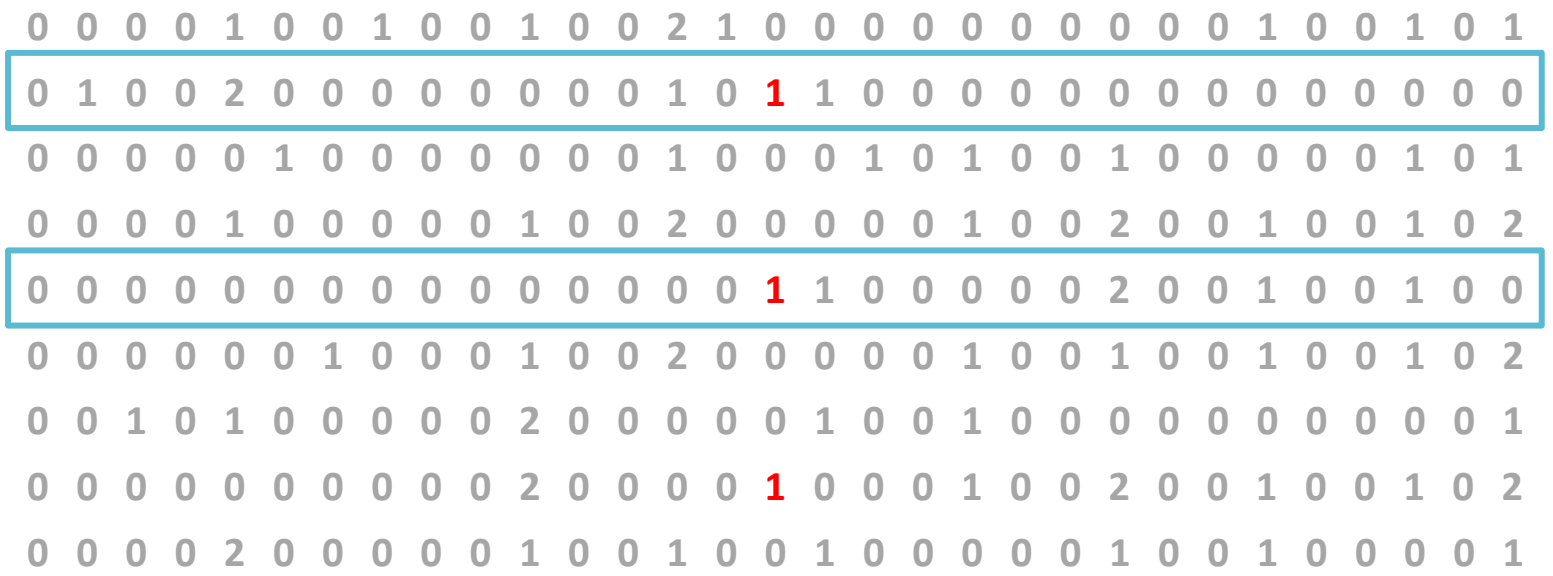
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 1 0 0 0 0 0 2 0 0 1 0 0 1 0 0

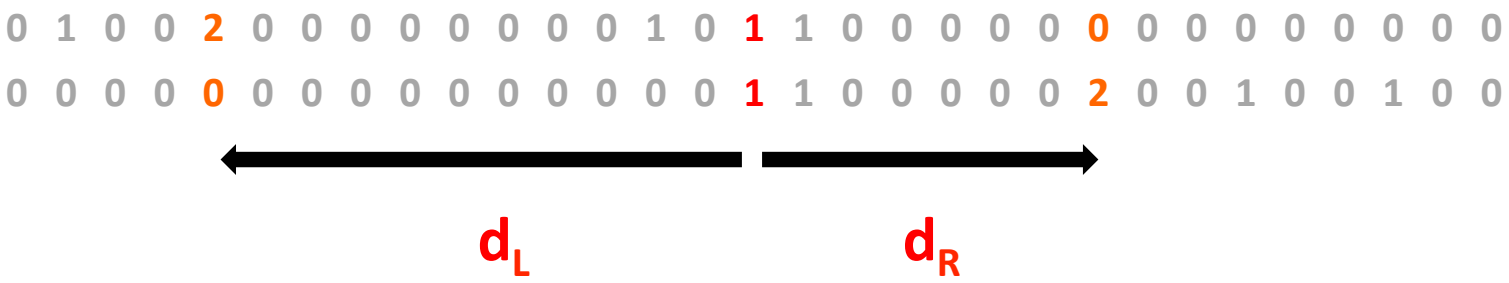
0 0 0 0 0 0 1 0 0 0 1 0 0 2 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 2

0 0 1 0 1 0 0 0 0 0 2 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1

0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 **1** 0 0 0 1 0 0 2 0 0 1 0 0 1 0 2

0 0 0 0 2 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1


0 0 0 0 1 0 0 1 0 0 1 0 0 2 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1
0 1 0 0 2 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 0 1
0 0 0 0 1 0 0 0 0 0 1 0 0 2 0 0 0 0 0 1 0 0 2 0 0 1 0 0 1 0 2
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 2 0 0 1 0 0 1 0 0
0 0 0 0 0 0 1 0 0 0 1 0 0 2 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 2
0 0 1 0 1 0 0 0 0 0 2 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 1 0 0 0 1 0 0 2 0 0 1 0 0 1 0 2
0 0 0 0 2 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1



Mathieson & McVean, 2014

If the t_{MRCA} of two alleles is known, the conditional probability distribution of the recombination distance is:

$$f(d_L \mid t_{MRCA})$$

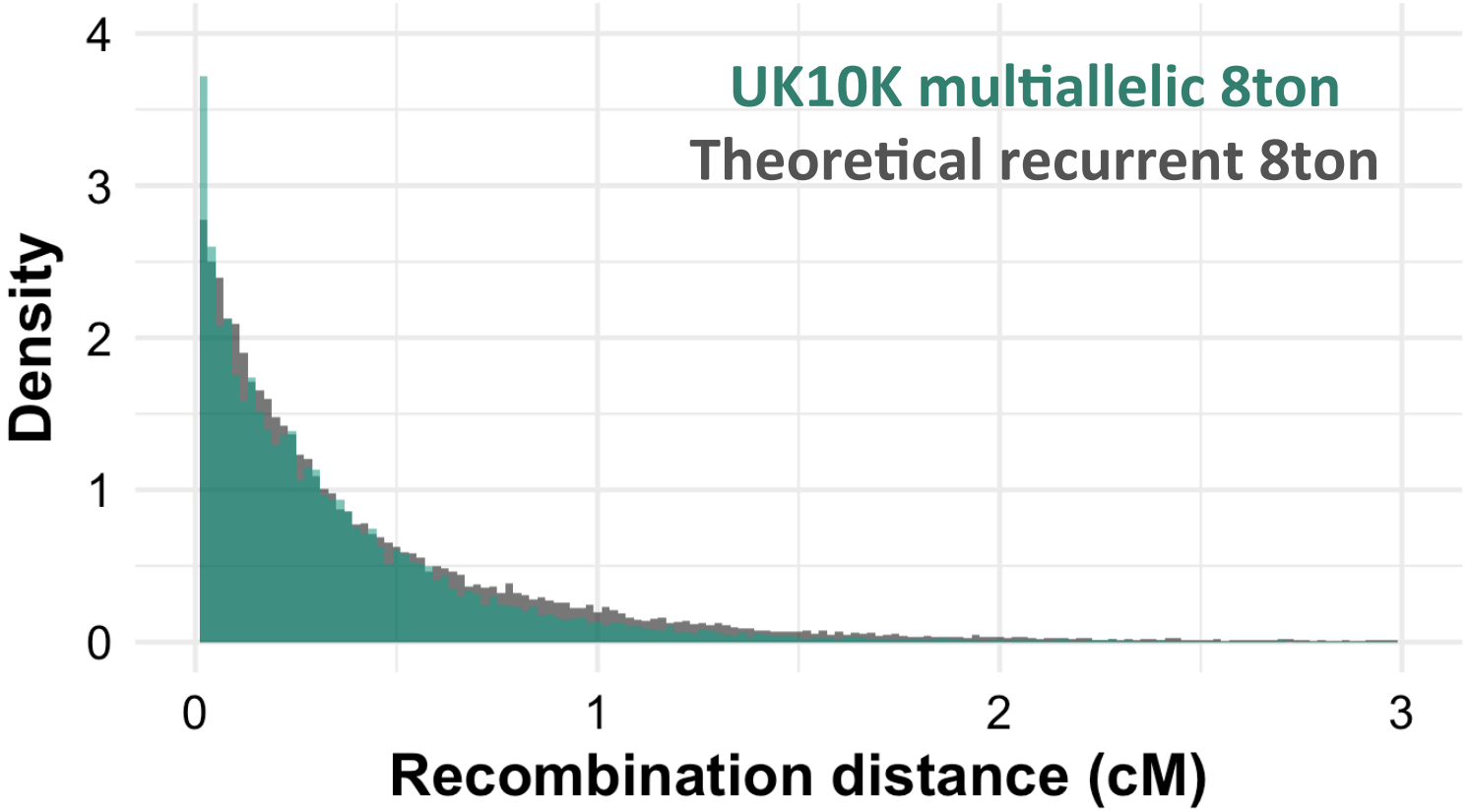
If the t_{MRCA} of two alleles is known, the conditional probability distribution of the recombination distance is:

$$f(d_L \mid t_{MRCA})$$

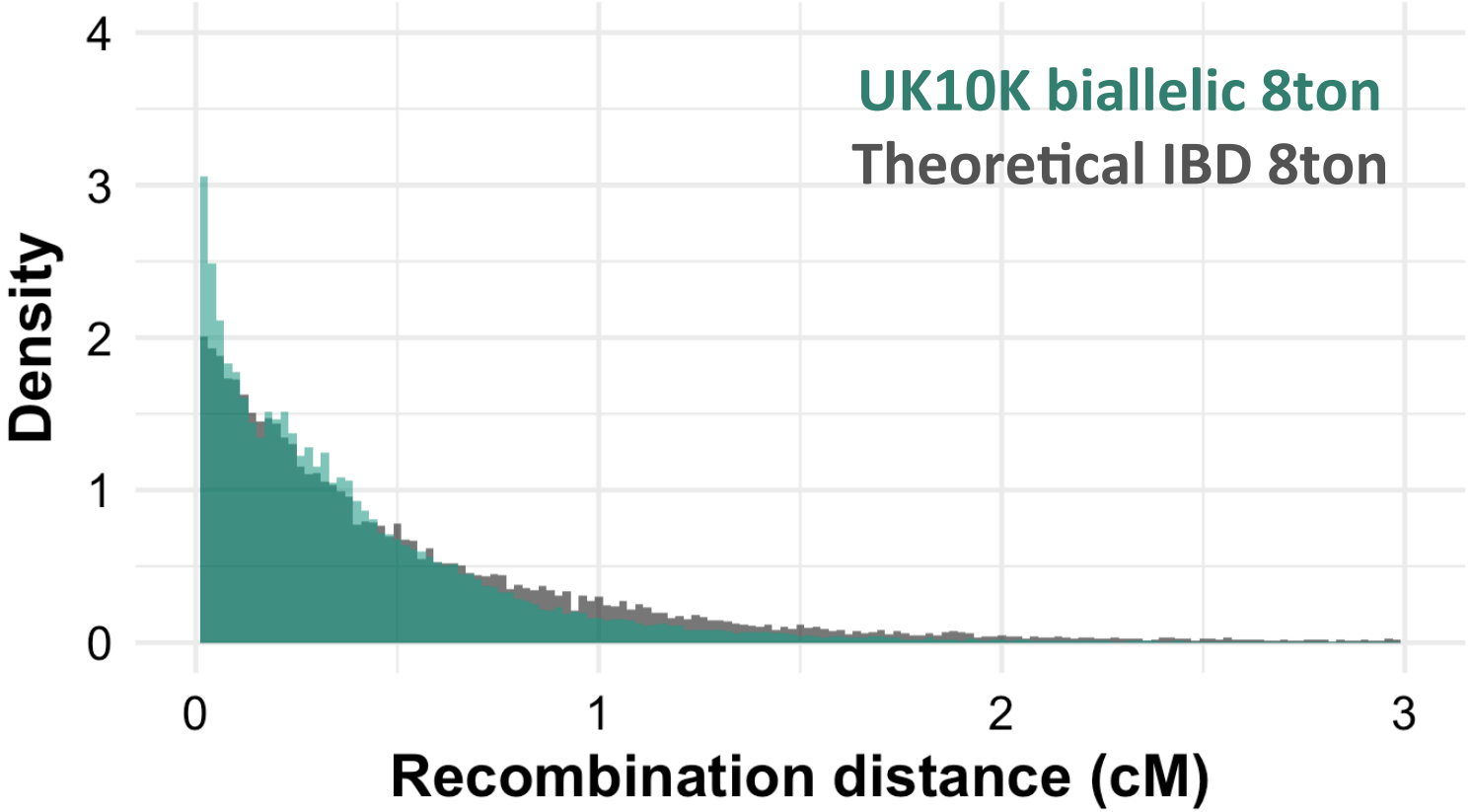
With the probability distribution of the t_{MRCA} for recurrent and IBD alleles, we can calculated the probability of d_L:

$$f(d_L) = \int_{t_{MRCA}} f(d_L \mid t_{MRCA})\, f(t_{MRCA})\, dt_{MRCA}$$

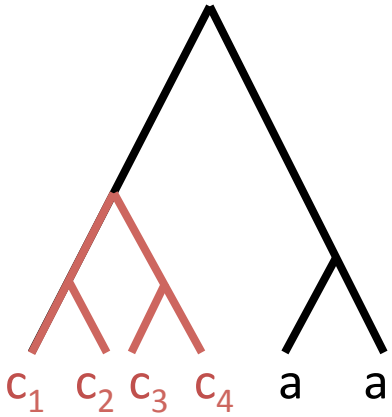
Theory vs. data: recurrent mutations

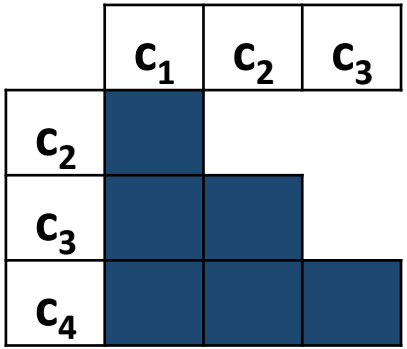


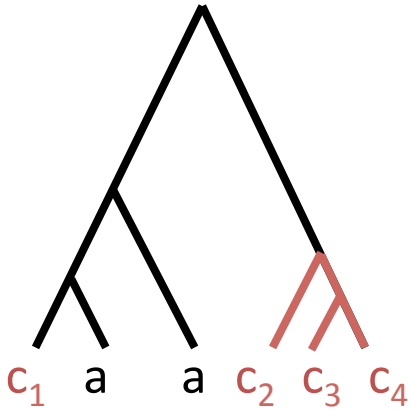
Theory vs. data: IBD mutations

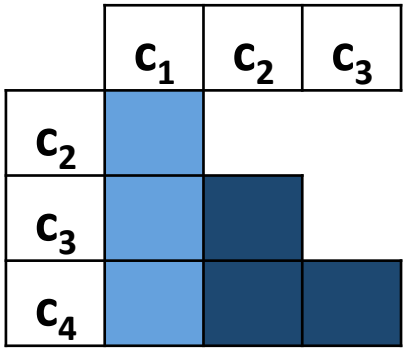


Recombination distances follow a predictable pattern



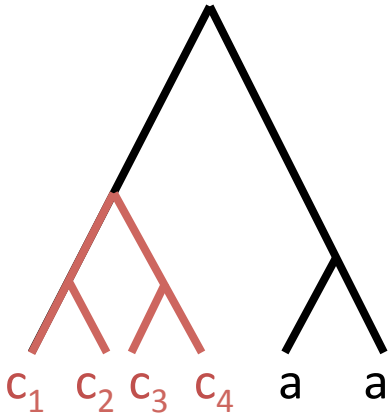


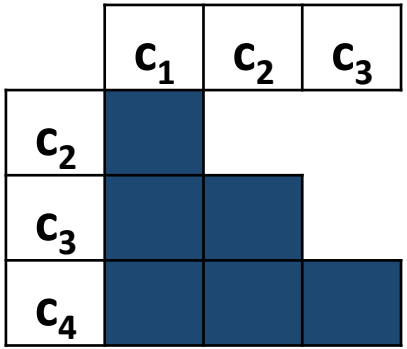


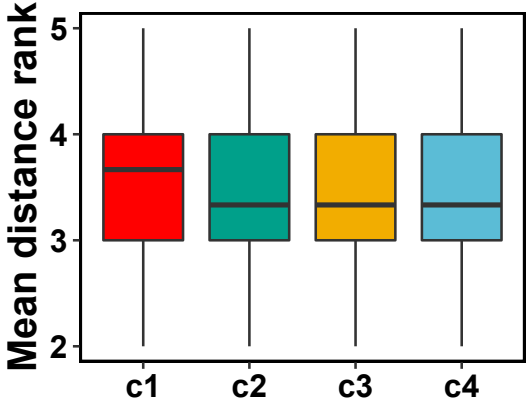


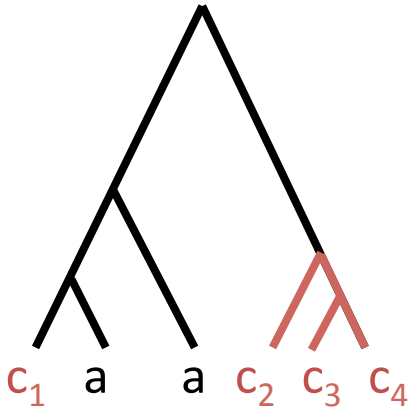
short t_{MRCA}, long rec. dist. **long t_{MRCA}, short rec. dist.**

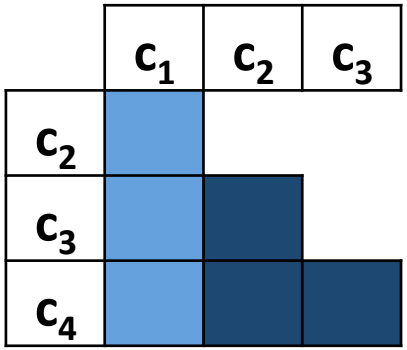
Recombination distances follow a predictable pattern

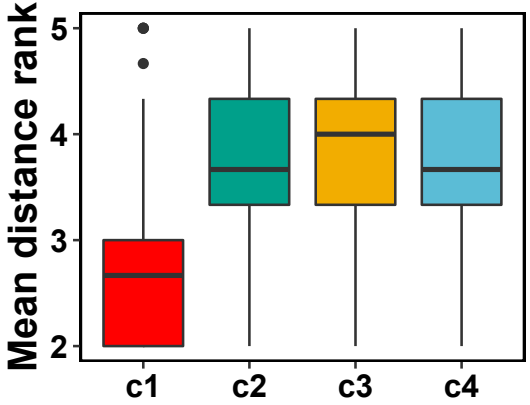












short t_{MRCA}, long rec. dist. **long t_{MRCA}, short rec. dist.**

Statistical approach

- Calculate likelihood of observed data under 2 scenarios (IBD or recurrent):
  - Recombination distances on right & left hand sides

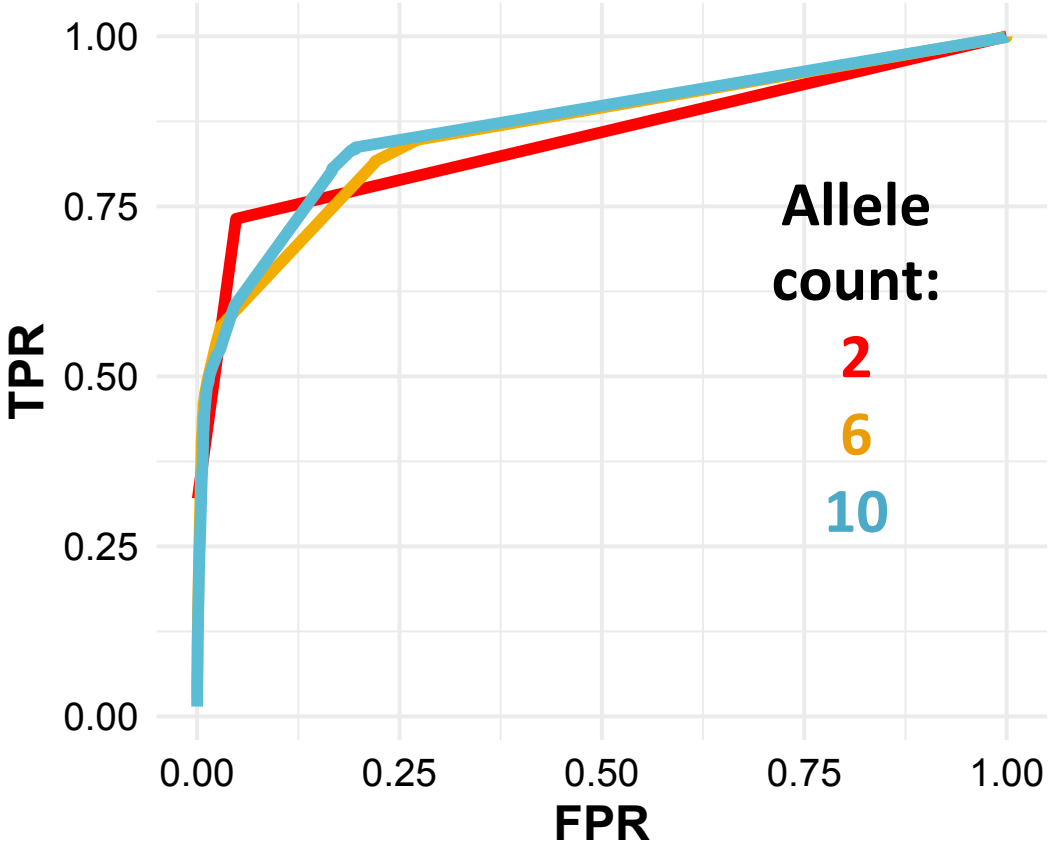
Statistical approach

- Calculate likelihood of observed data under 2 scenarios (IBD or recurrent):
  - Recombination distances on right & left hand sides
  - Distance ranks on right & left hand sides

Statistical approach

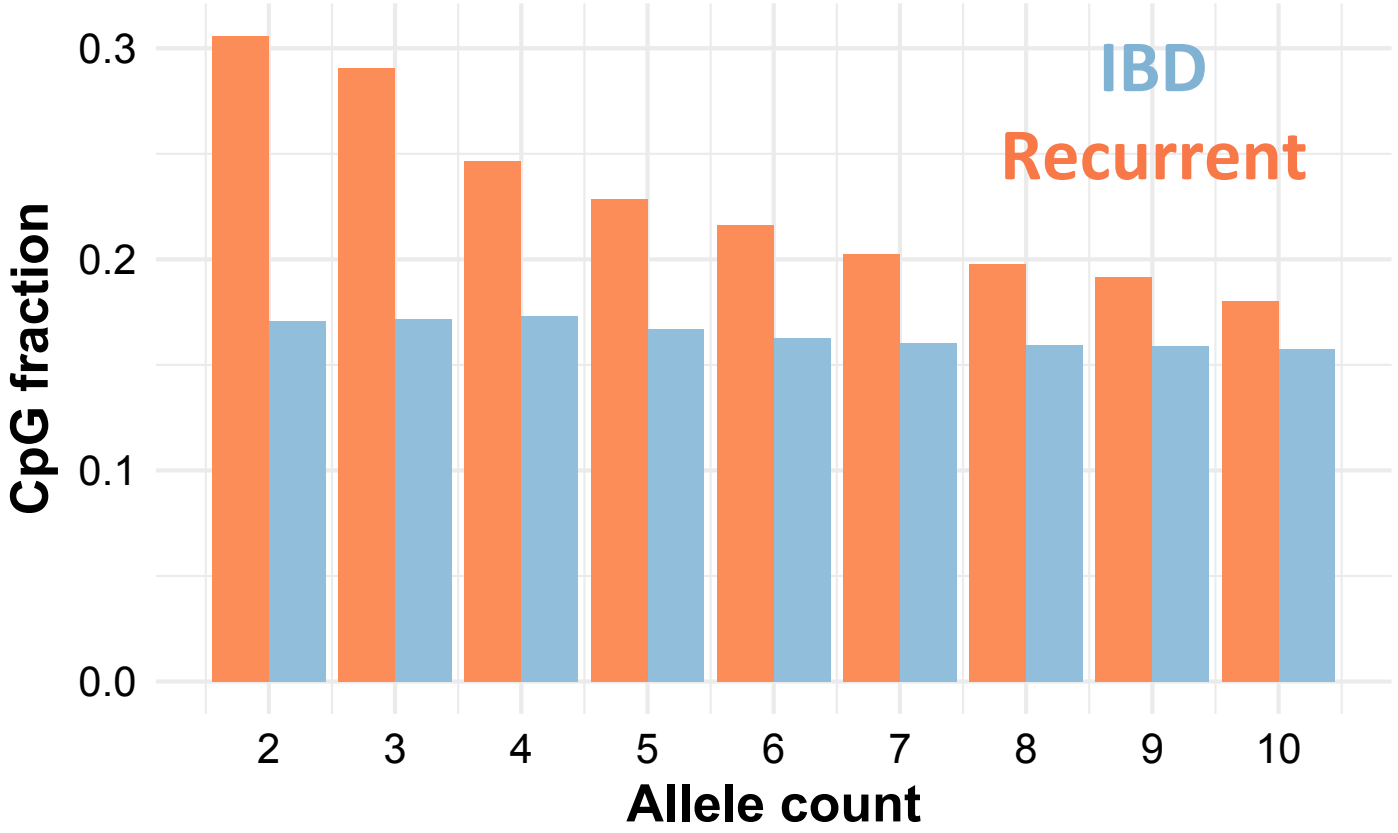
- Calculate likelihood of observed data under 2 scenarios (IBD or recurrent):
  - Recombination distances on right & left hand sides
  - Distance ranks on right & left hand sides
- Compute test statistic of composite likelihood ratio

Statistic performance depends on allele count



Application to UK10K: CpG enrichment

Application to UK10K: CpG enrichment



What's next?

- Application to empirical datasets (e.g. UK10K)
  - Updated measurement of SFS

What's next?

- Application to empirical datasets (e.g. UK10K)
  - Updated measurement of SFS
  - Mutation rate variation

What's next?

- Application to empirical datasets (e.g. UK10K)
  - Updated measurement of SFS
  - Mutation rate variation

- Rare variant burden tests

Thank you!

Voight Lab

Ben Voight
Paul Babb
Diana Cousminer
Kat Gawronski
Kim Lorenz
Katie Siewert
Chris Thom

Thesis Committee

Casey Brown
Maja Bucan
Struan Grant
Sarah Tishkoff

Funding

Genetics Training Grant
T32GM008216