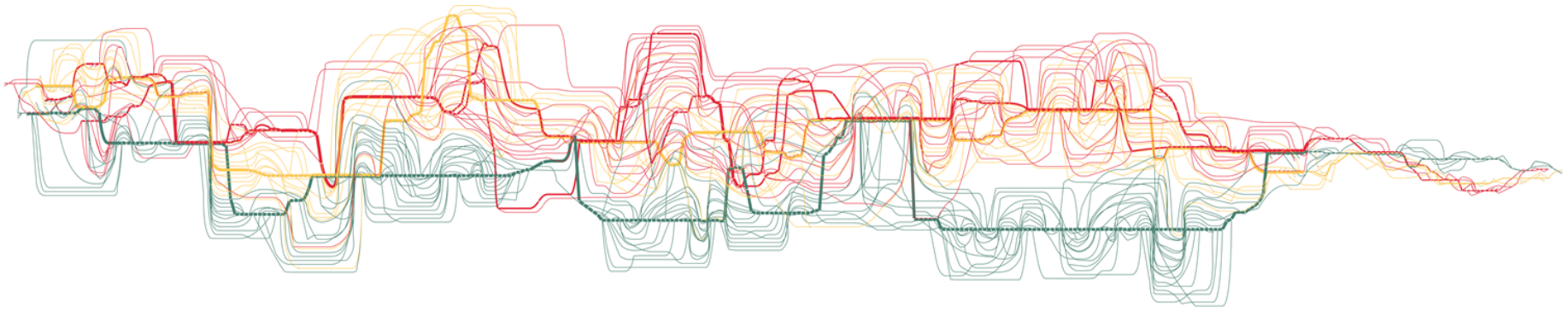
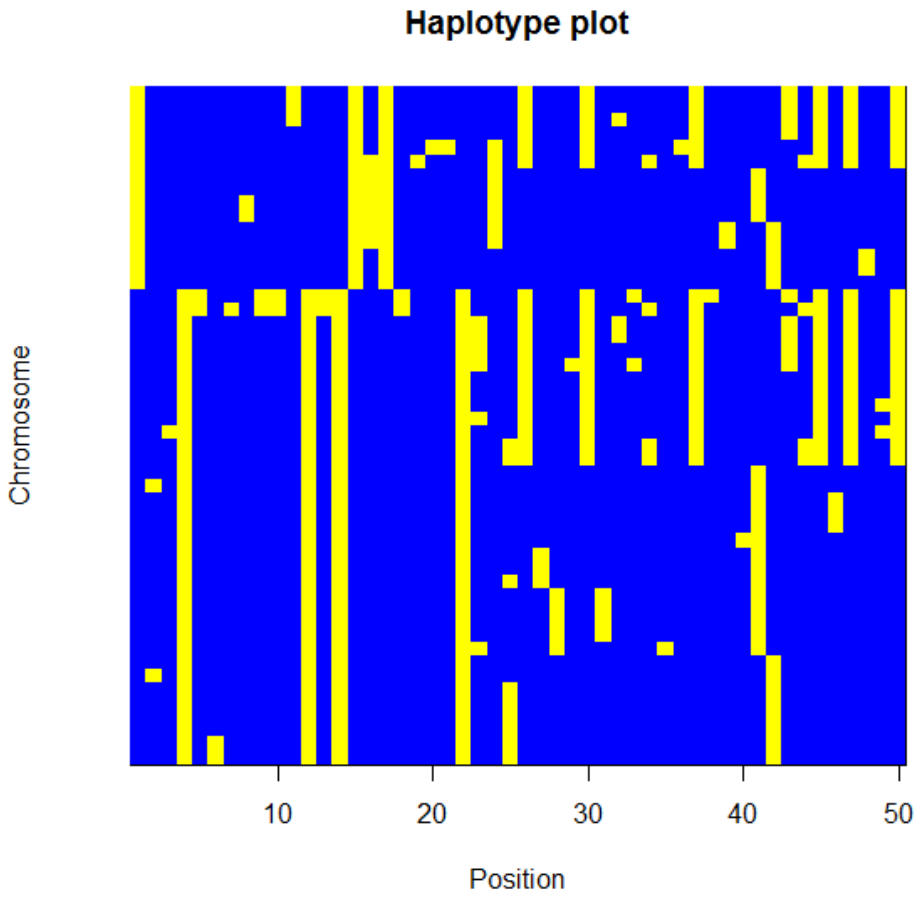
Graph structures for representing and analysing genetic variation

Gil McVean

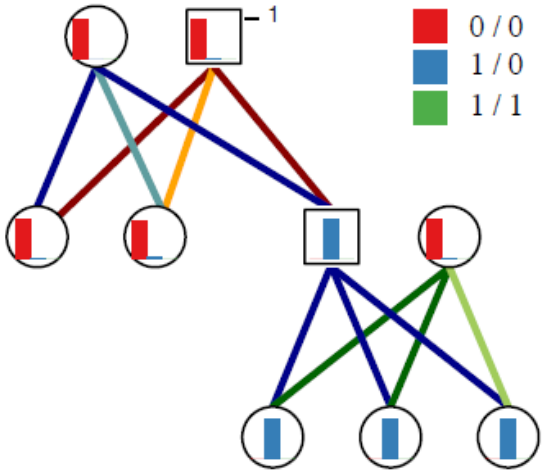


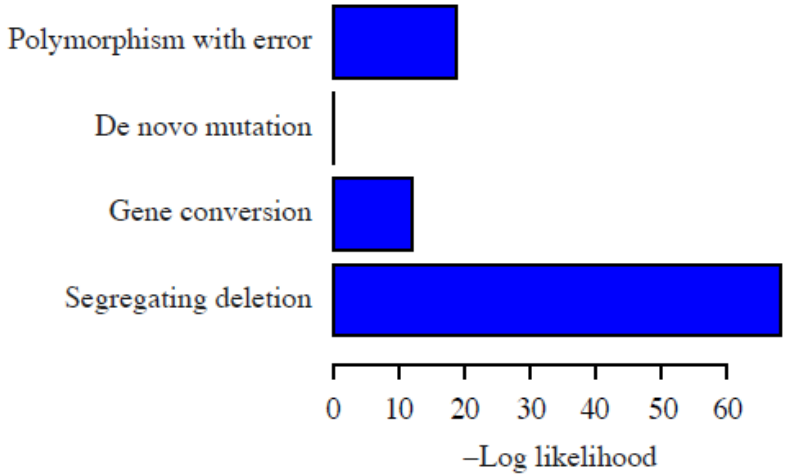
What is genetic variation data?



Binary incidence matrix

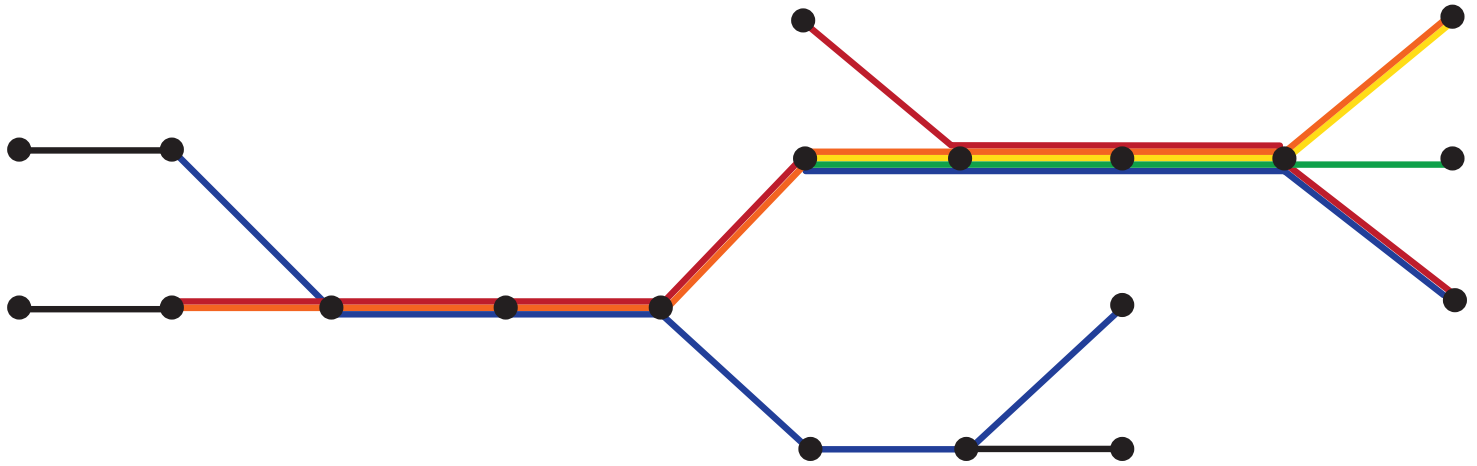
What is genetic variation data?





Genotype likelihoods

What is genetic variation data?

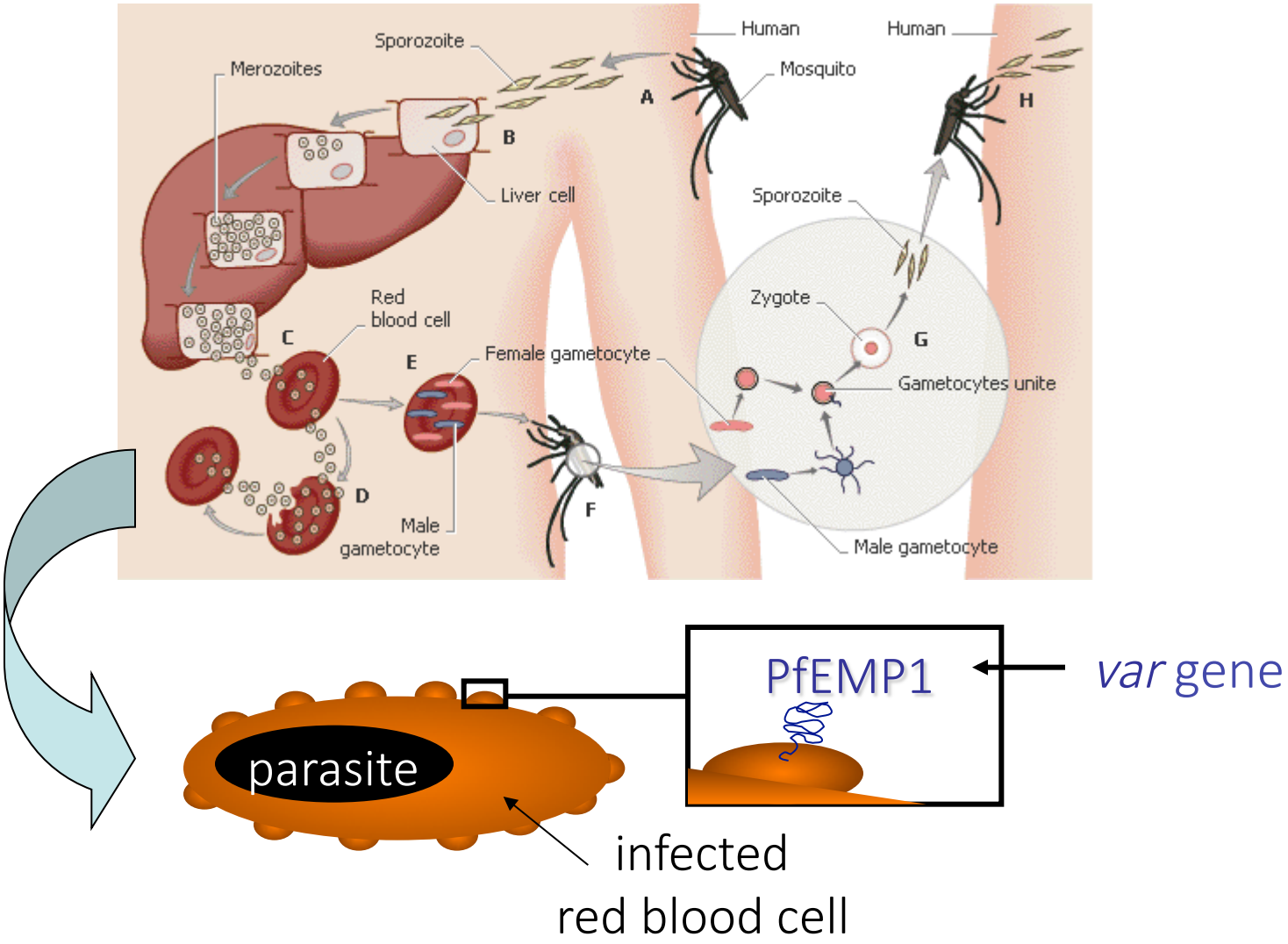


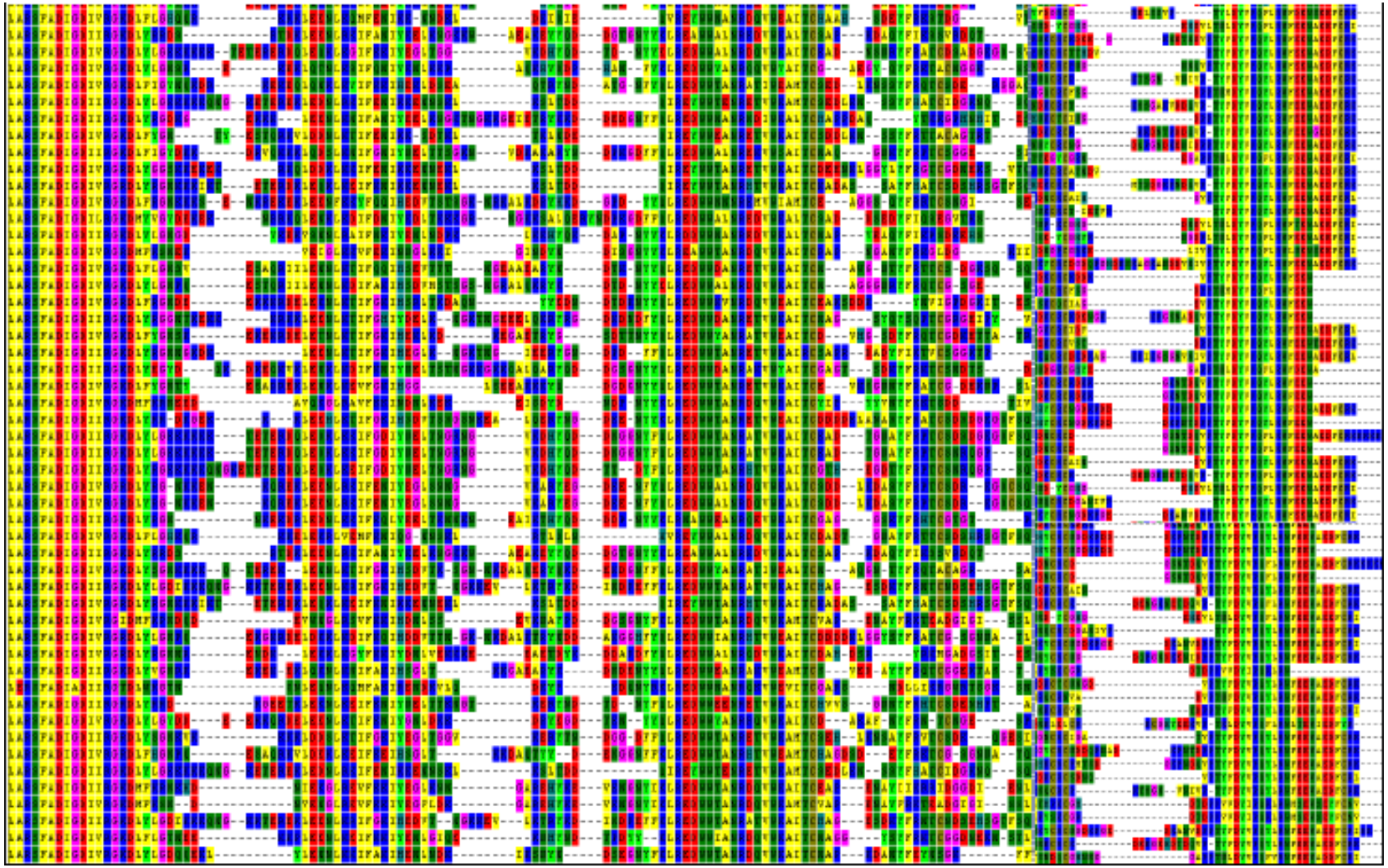
Graphs of primary sequence

What is this talk about?

- I want to convince you that there are types of variation that are not well represented by the binary incidence or genotype likelihood models.
- I want to convince you that this variation is interesting from an evolutionary and phenotypic perspective, hence the need for methods that can access and analyse such variation.
- I want to convince you that graph-based approaches are a powerful way to represent and analyse both known and novel sequence.
  - Reference graph for human variation.
  - Assembly of hypervariable genes.

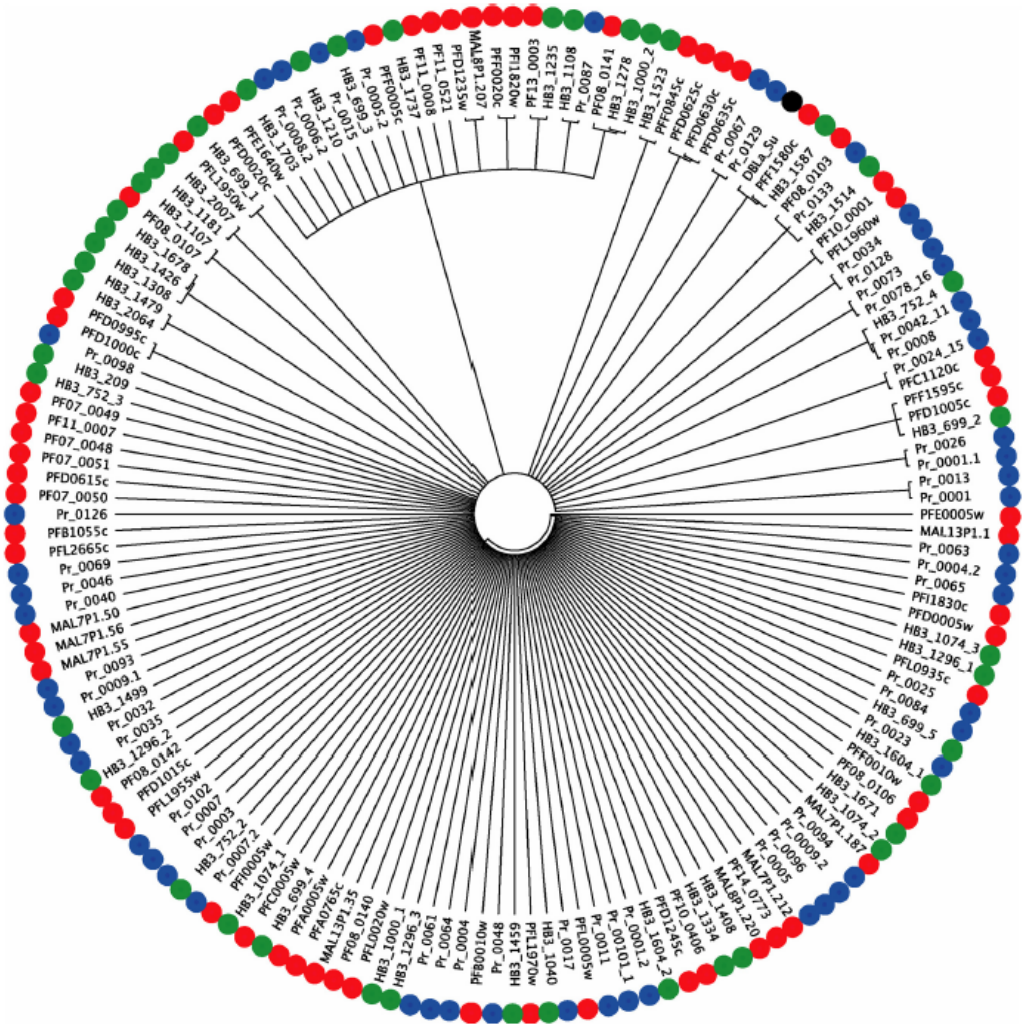
Example I: The *var* genes of *P. falciparum*



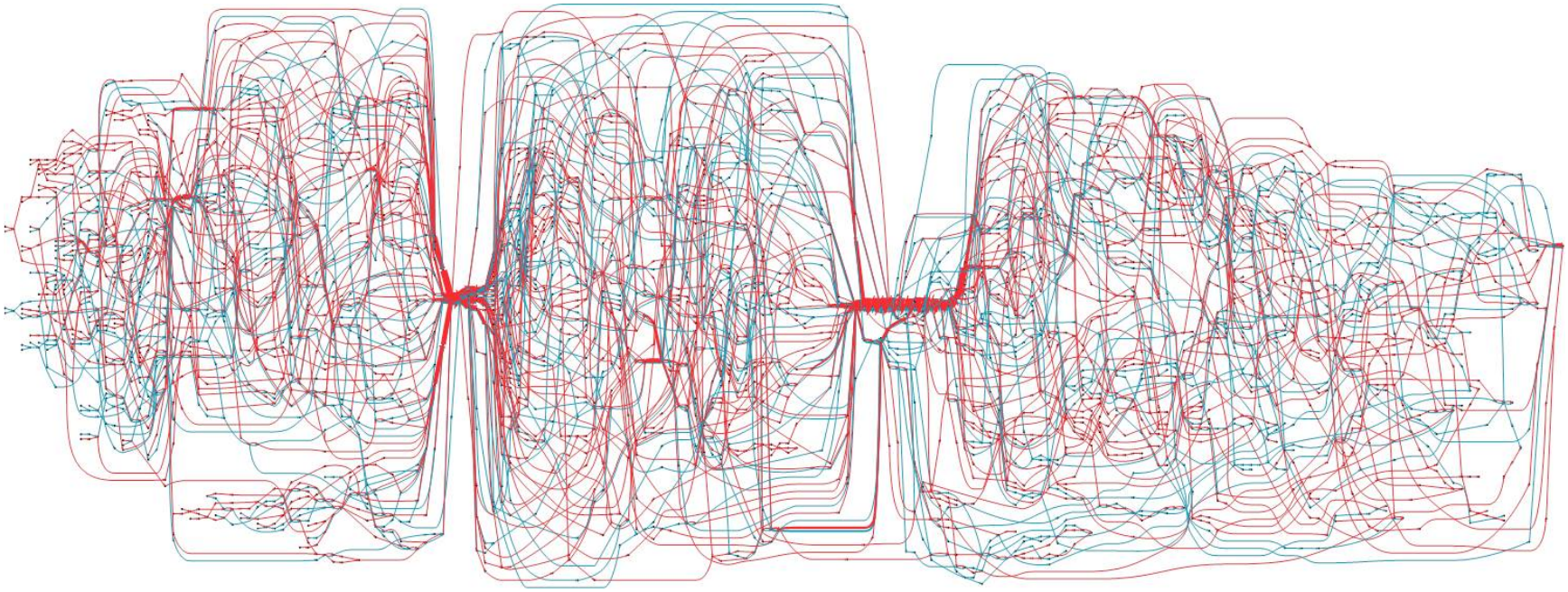


Alignment of *P. falciparum* DBLα domains

There is little structure in the basic alignment

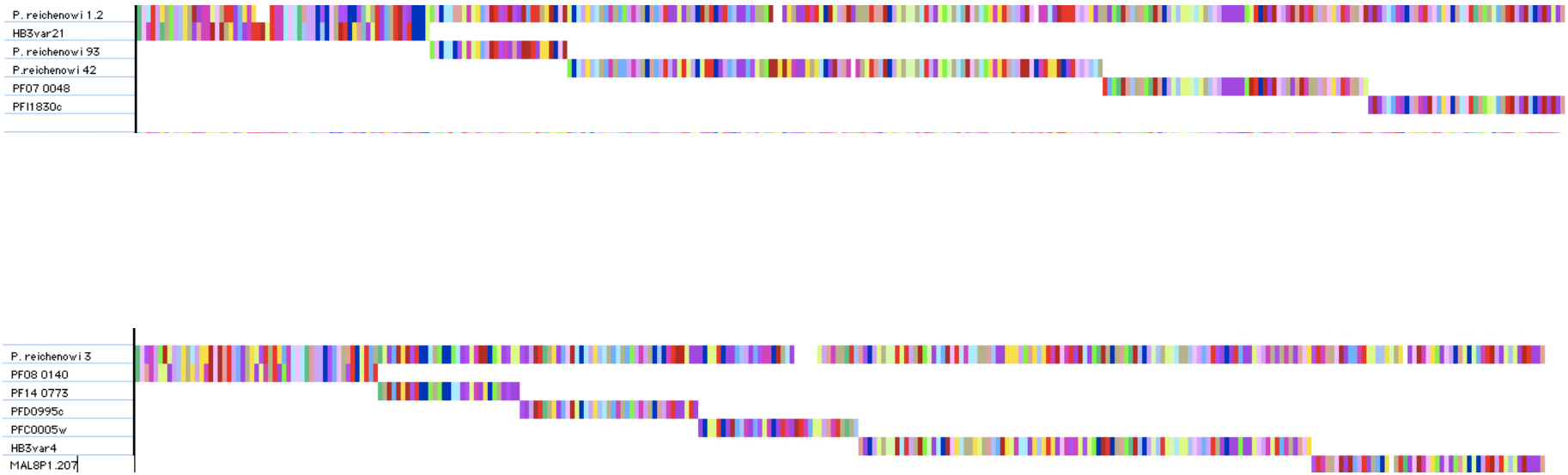


A graph of PfEMP1 DBLa sequence variation



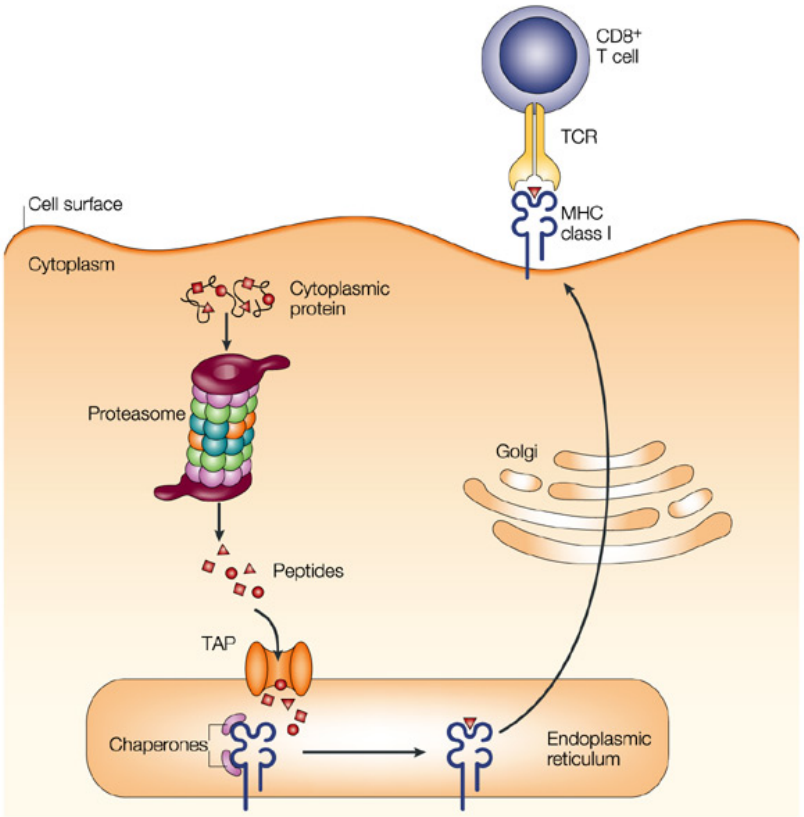
k = 9 (amino acid)

Mosaic structures reveal ancient origin for hypervariable genes

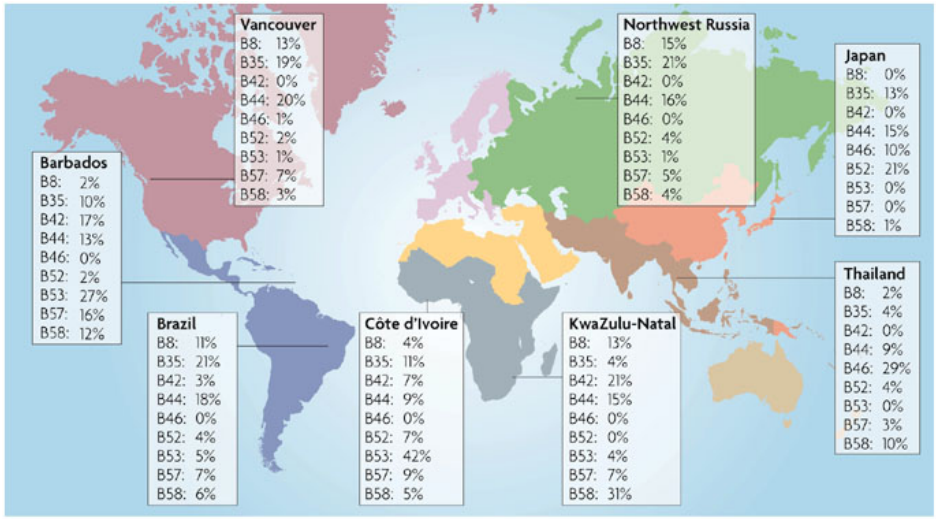


Zilversmit *et al.* 2013

Example II: Homology and paralogy in the Class I HLA genes

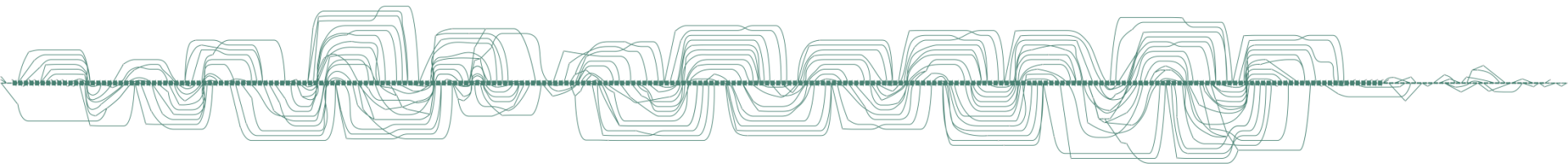


Nature Reviews | Immunology



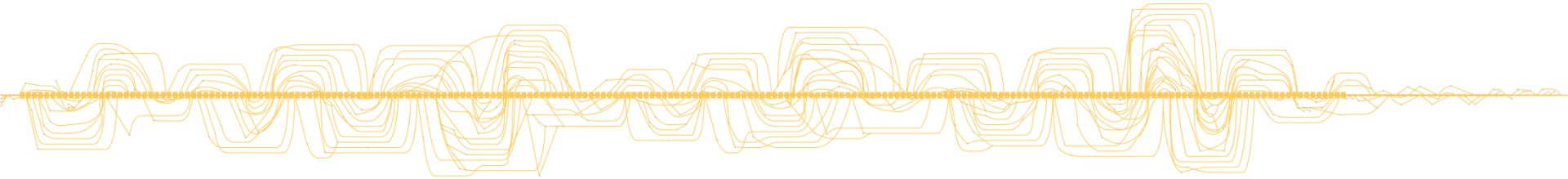
Nature Reviews | Immunology

Example II: Homology and paralogy in the Class I HLA genes



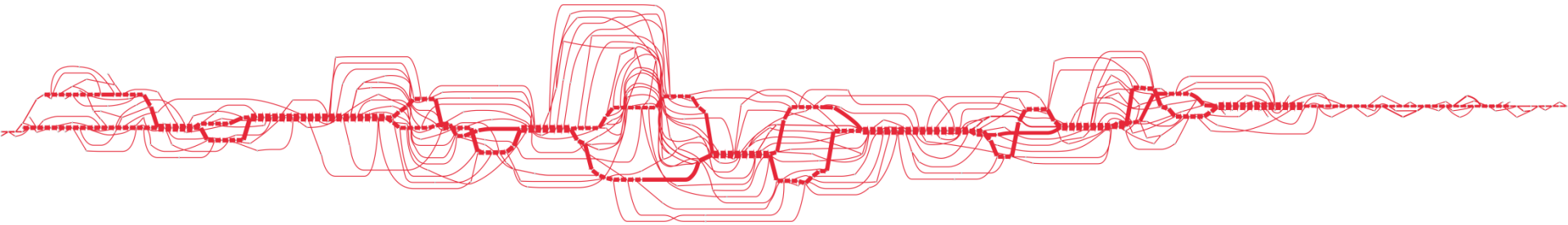
HLA-A
200 alleles, k = 31
Coding sequence

Example II: Homology and paralogy in the Class I HLA genes



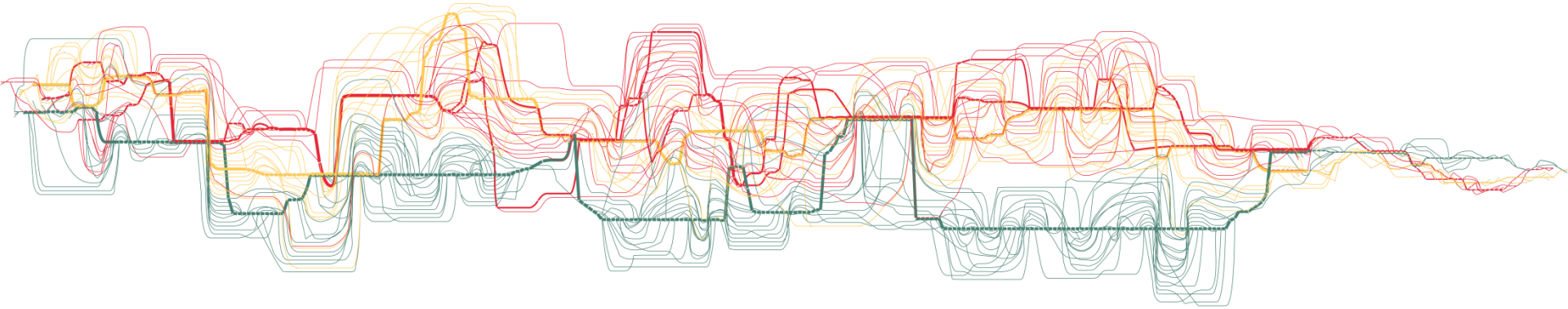
HLA-B
200 alleles, k = 31
Coding sequence

Example II: Homology and paralogy in the Class I HLA genes



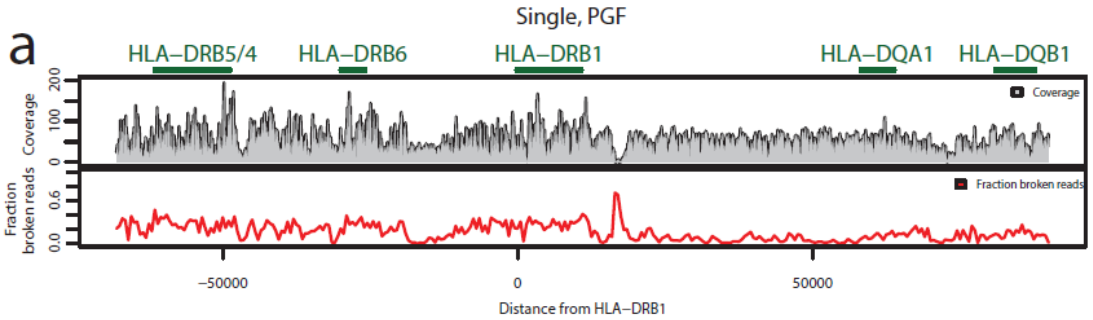
HLA-C
200 alleles, k = 31
Coding sequence

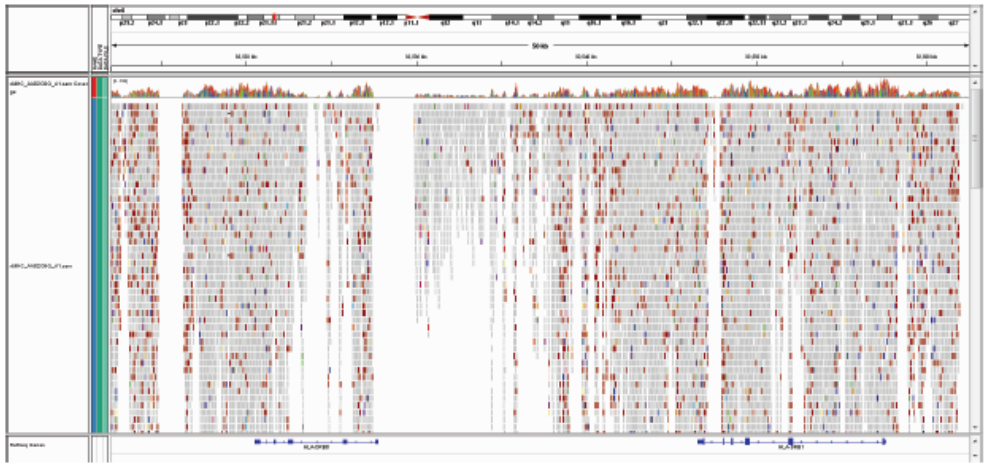
Example II: Homology and paralogy in the Class I HLA genes

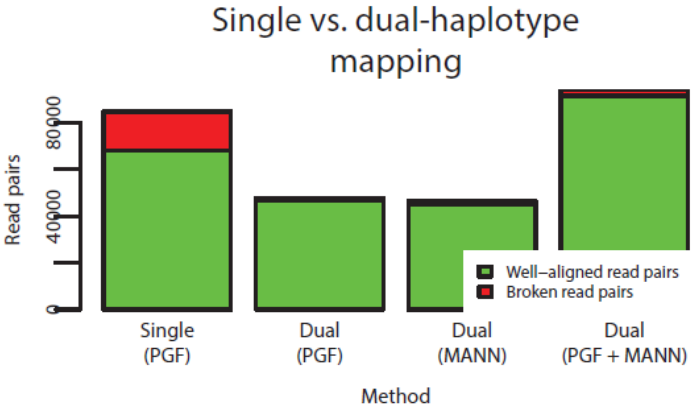


HLA-A, HLA-B, HLA-C
600 alleles, k = 31
Coding sequence

Example III: Structural variation in the HLA Class II region







Motivation and questions

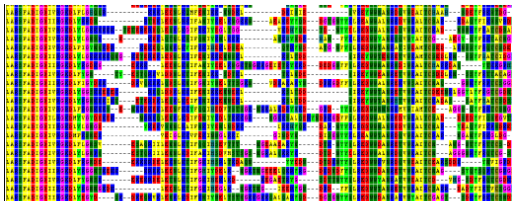
- One of the most powerful insights from population genetics is that novel sequences tend to look like those we've already seen, though with mutation and recombination.

- Moreover, relatively few sequences are often needed to capture the vast majority of sequence space.

- The big question is how to formalise this relationship so that we can best assemble and interpret the genome of the next sample.

Graph structures for representing sequence and variation

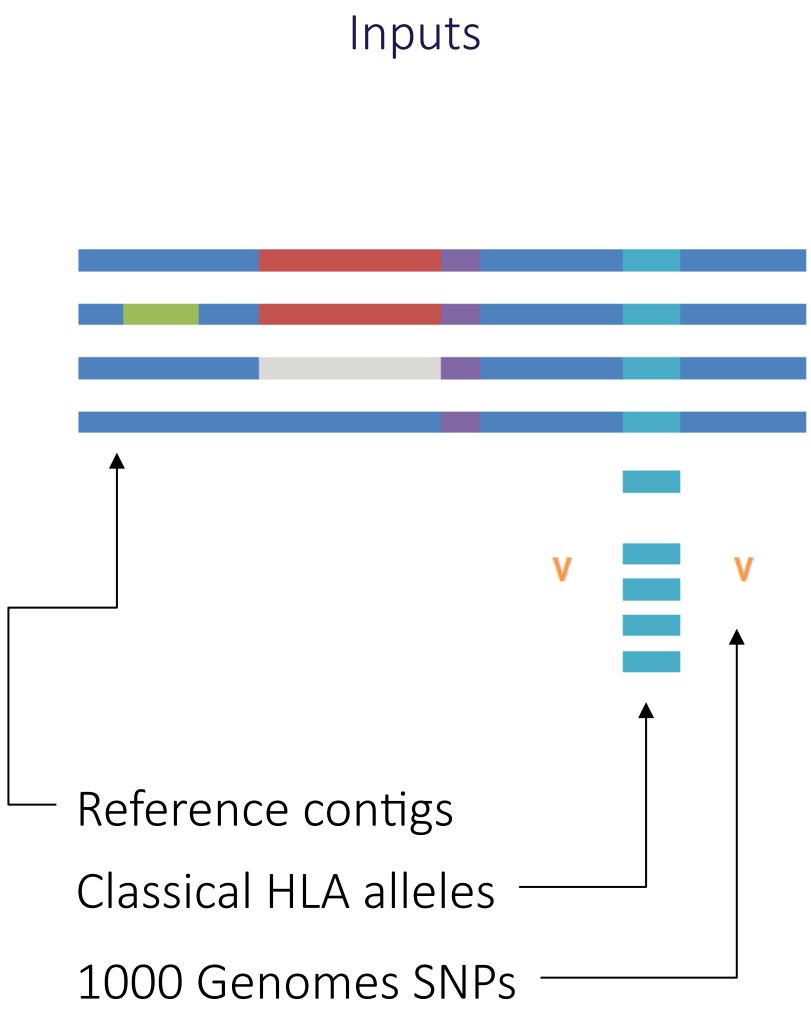
Multiple sequence alignment



Statement of homology

No statement of homology

A population reference graph (PRG) for the HLA



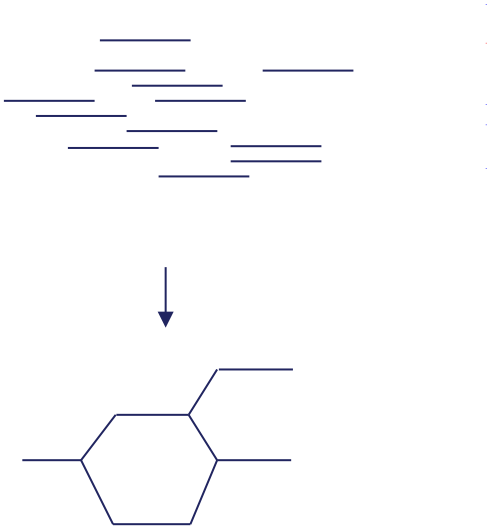
Features of the PRG

- It is a compression of the input data
  - Long-range information can be retained if necessary as coloured paths
- It is a generative model
  - New genomes can be simulated by choosing paths through the PRG
- Its structure suggests an efficient method for genome inference in a novel sample
  - Use an HMM where emissions are the reads or a summary of them (diagnostic kmers associated with each string)
  - Current implementation is not optimal, but goal was to re-use as much of current tool chain as possible.

Implementation

Stage 1

Reads converted to cleaned de Bruijn graph

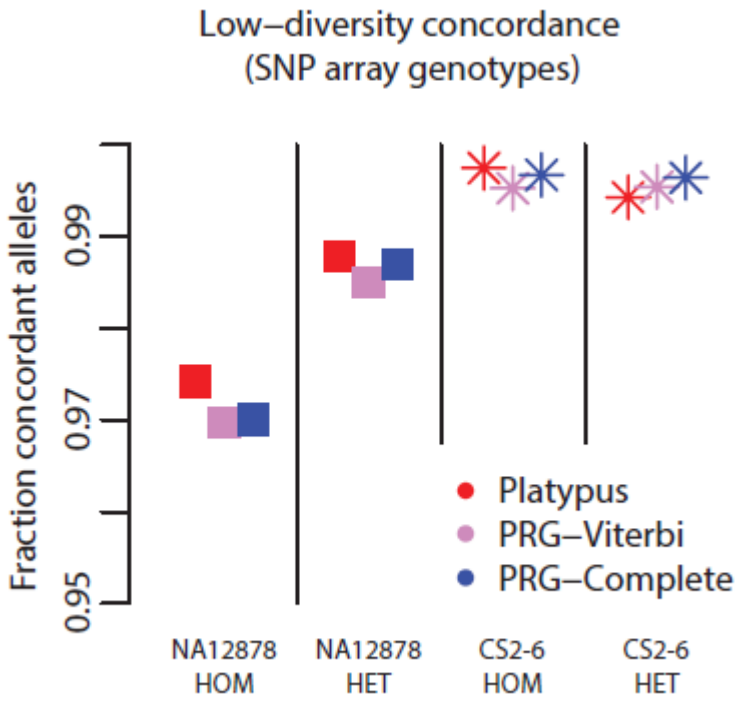


De Bruijn Graph (dBG)

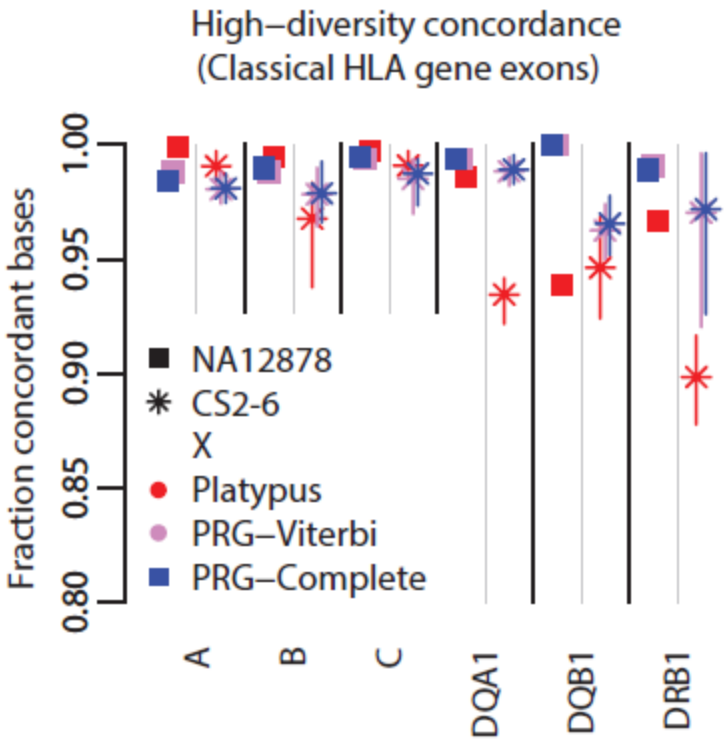
Evaluation

- Compare to Stampy/Platypus as ‘best-practice’ mapping-based approach
- Evaluate on four data types
  - SNP array data
  - Sequence based typing (Sanger) of classical HLA alleles
  - Kmer recovery from high throughput sequencing data
  - Long-read (10kb) Moleculo data
- Two sets of samples
  - NA12878
  - Five cohort samples from a GSK drug-safety study (CS2-6) [Not Moleculo]

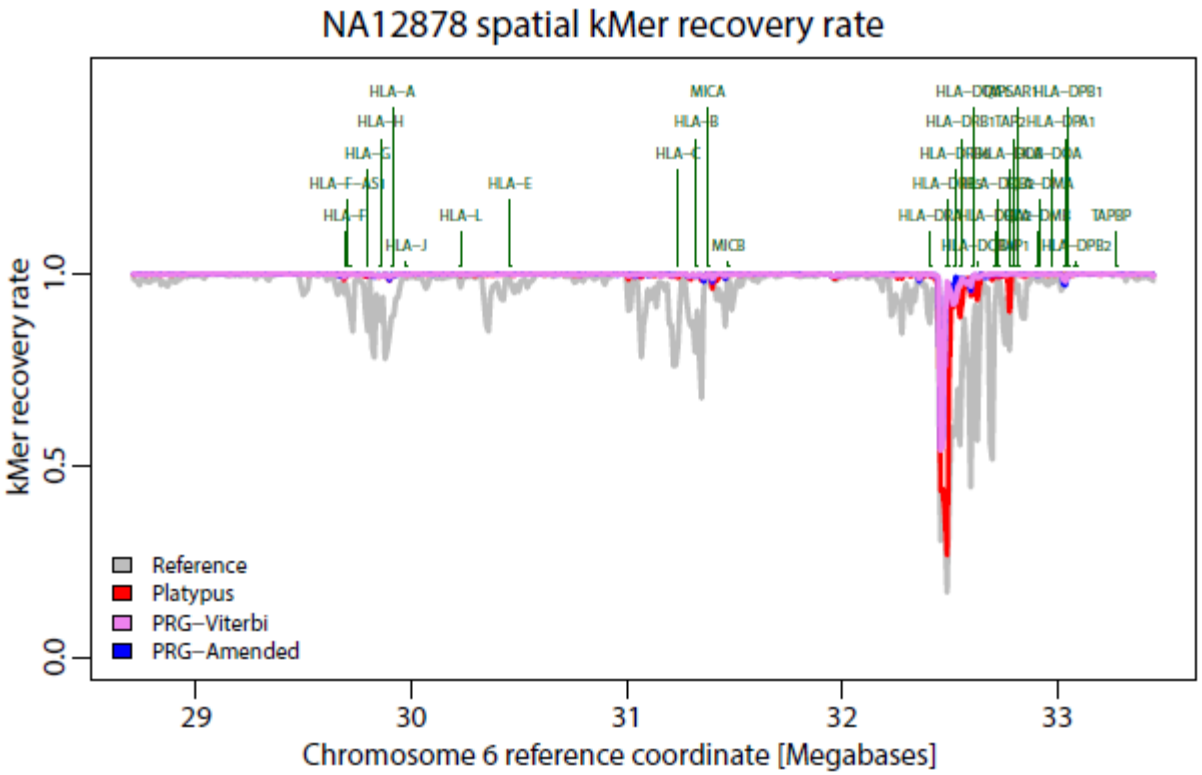
Comparison to SNP array data

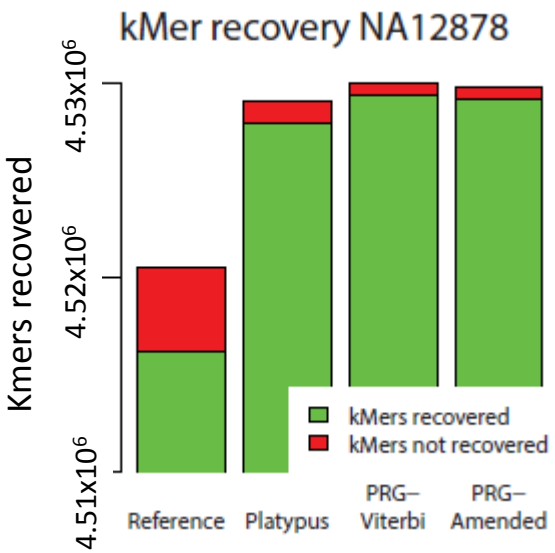


Comparison to Sanger sequence at classical HLA alleles

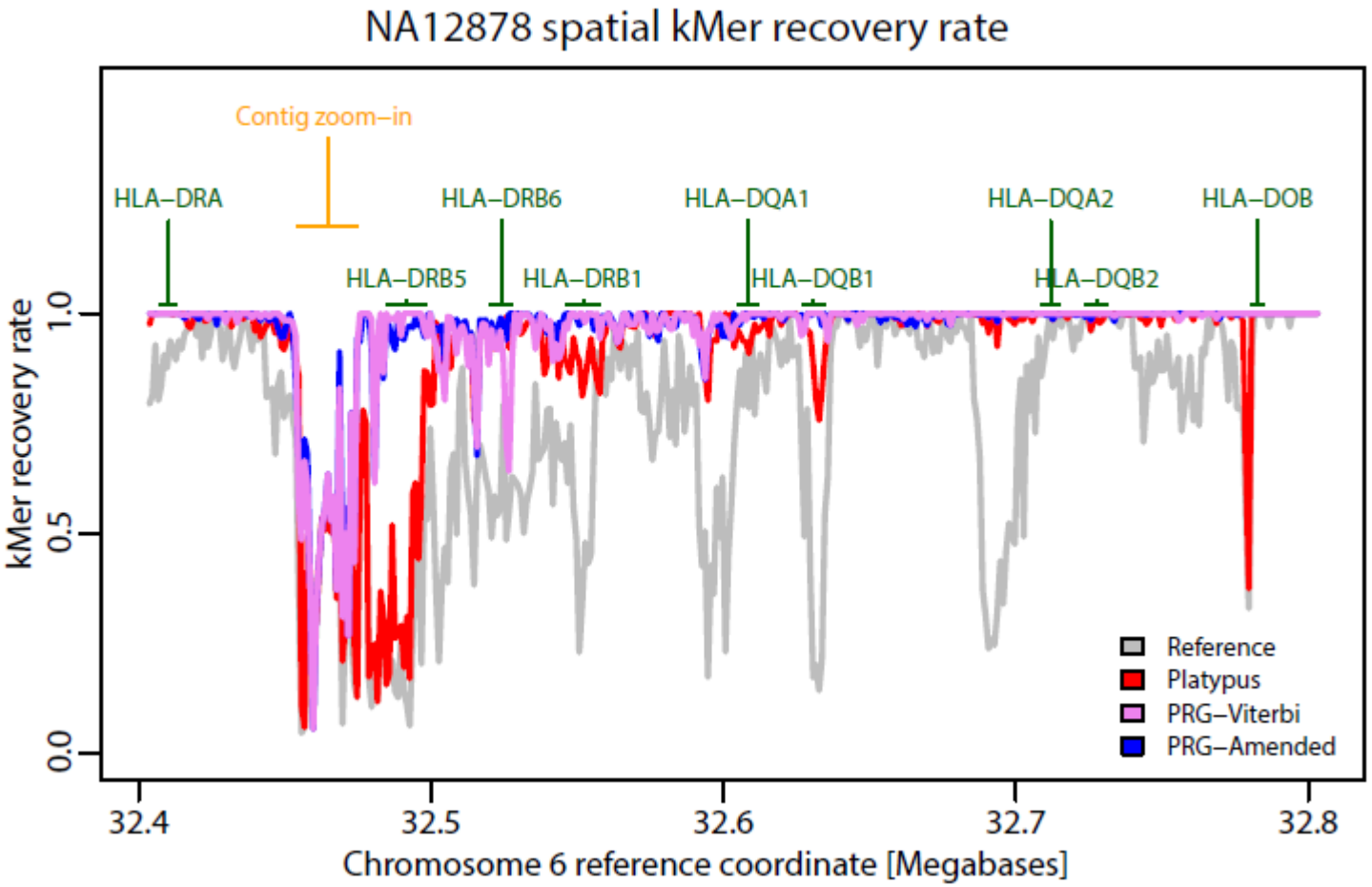


Recovery of kmers across HLA (NA12878)

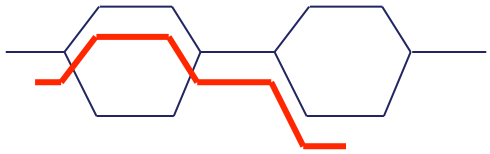




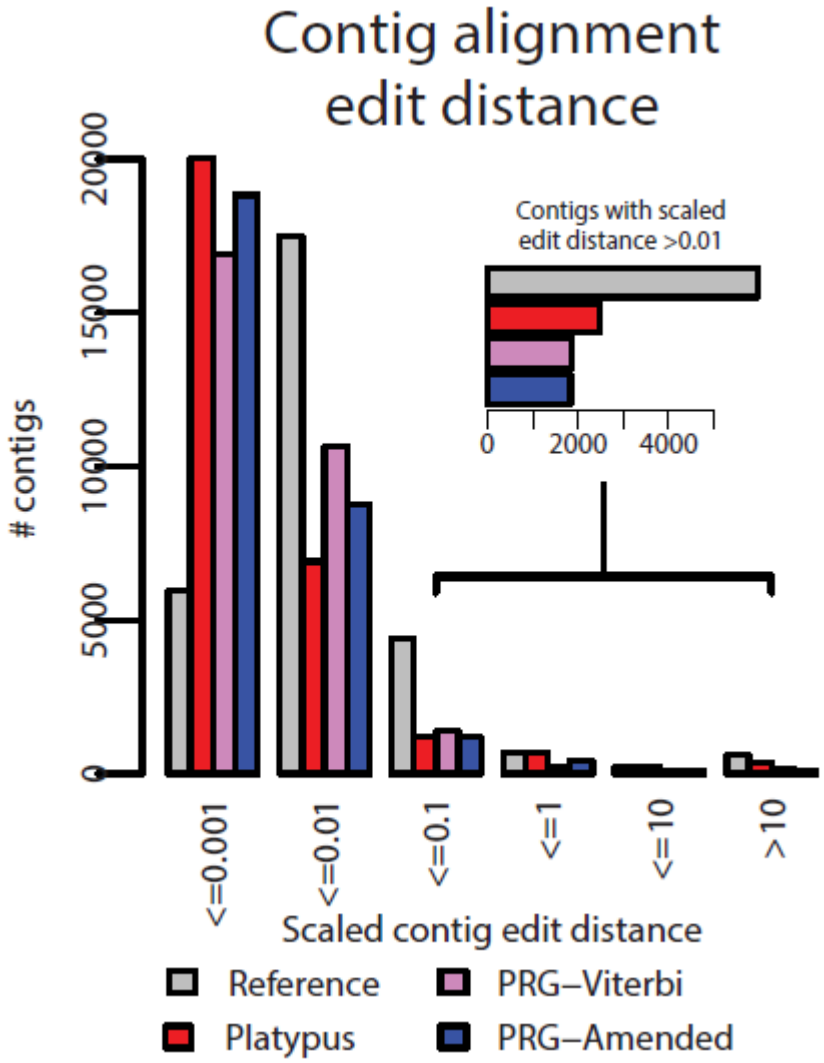
Zoom-in of kmer recovery



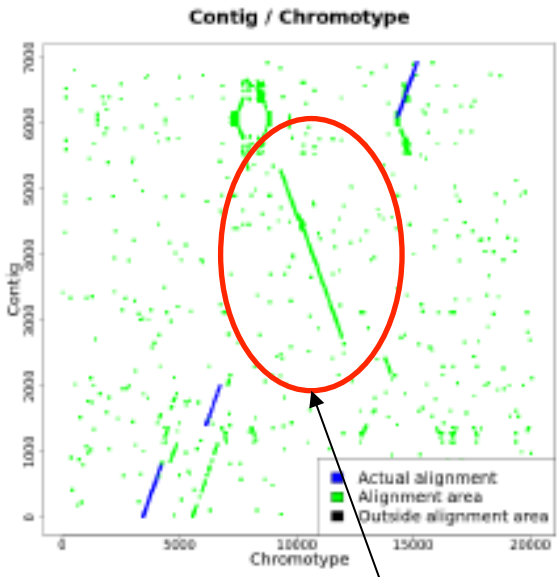
Comparison of Moleculo contigs (NA12878)

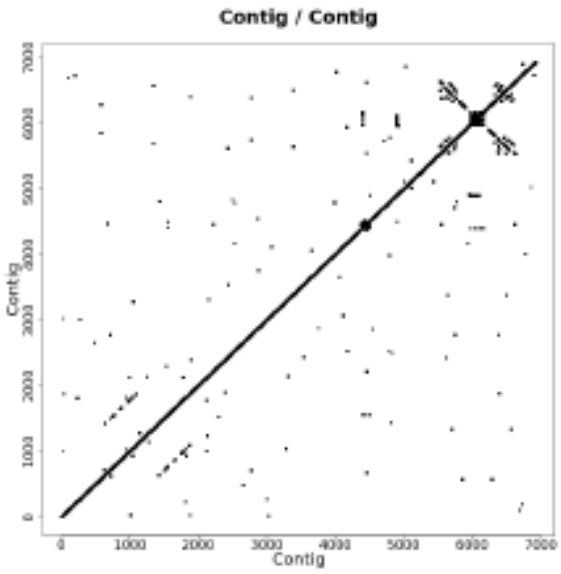


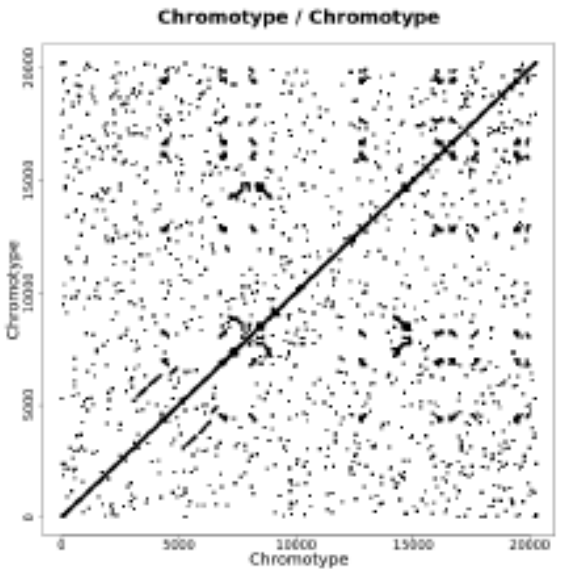
Contig aligned to chromotype



Evidence for ‘missing’ variation in Class II region







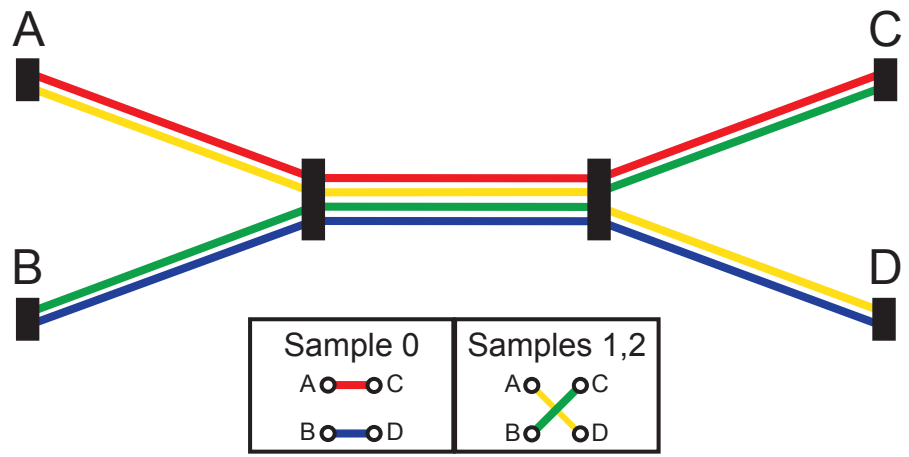
Looks like inversion within region

Extending the method – A new data structure

- We have end-to-end prototype for a population reference graph and its use to assemble variation within the HLA region.

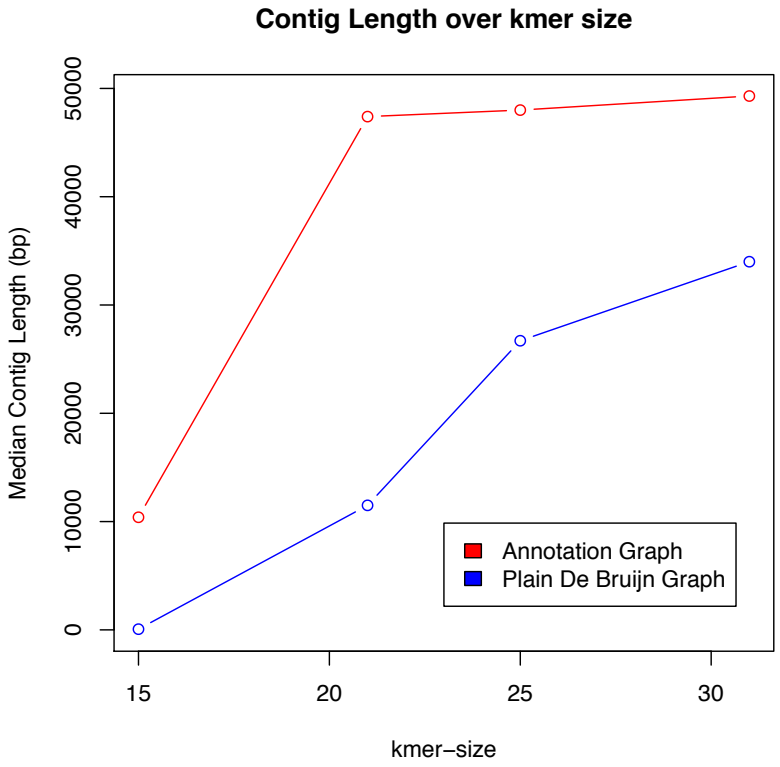
- The current implementation is not optimal in a few regards:
  - Use of de Bruijn graph throws away longer-range read data
  - Two step chromotype -> re-mapping is inefficient and doesn't add much

- Both issues can be solved with a novel data structure: annotated de Bruijn graph
  - Related to idea of Conway and Bromage (2011).



Structure enables error correction and use of paired-end information

k-agnostic data structure – approximating a string graph



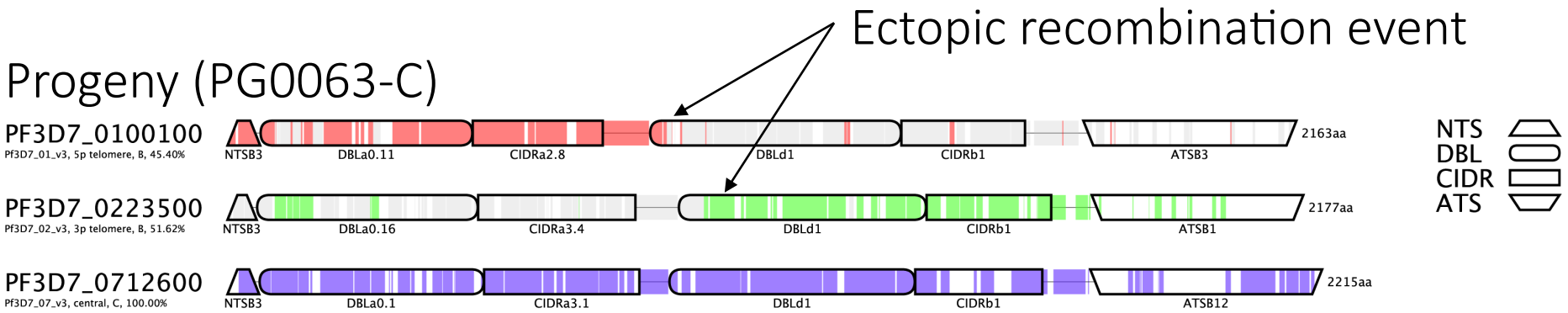
Simulation: Staph Genome, 100bp singled-ended error-free reads, 50x coverage

- Basic de Bruijn graph – c. 50Gb for one human.
- One additional human c. 1Gb.
- Should scale roughly as log(n).
- Easy to operate – “drop-in” model.

Example: Using novel structure to assemble *var* genes

Sample	Algorithm	Contigs	Max length	N50	Junctions resolved
PG0051_C (3D7)	supernode	3544	3286	194	0
PG0051_C (3D7)	single-end	1874	6437	591	5175 (59%)
PG0051_C (3D7)	paired-end (one-way)	1564	7412	982	7498 (72%)
PG0051_C (3D7)	paired-end (two-way)	1517	7352	993	7552 (73%)
PG0052_C (HB3)	supernode	1710	4512	291	0
PG0052_C (HB3)	single-end	966	5007	903	2559 (59%)
PG0052_C (HB3)	paired-end (one-way)	802	5062	1503	3748 (72%)
PG0052_C (HB3)	paired-end (two-way)	786	5062	1526	3998 (74%)
PG0063_C (progeny)	supernode	2802	3114	191	0
PG0063_C (progeny)	single-end	1430	5807	697	4665 (63%)
PG0063_C (progeny)	paired-end (one-way)	1180	6429	1185	6413 (74%)
PG0063_C (progeny)	paired-end (two-way)	1146	7300	1229	7044 (77%)

Contigs identify recombinant sequences among progreny



With thanks to

- Funders:
  - The Wellcome Trust, GSK, EPSRC, The Royal Society

- People:

Alexander Dilthey

Isaac Turner

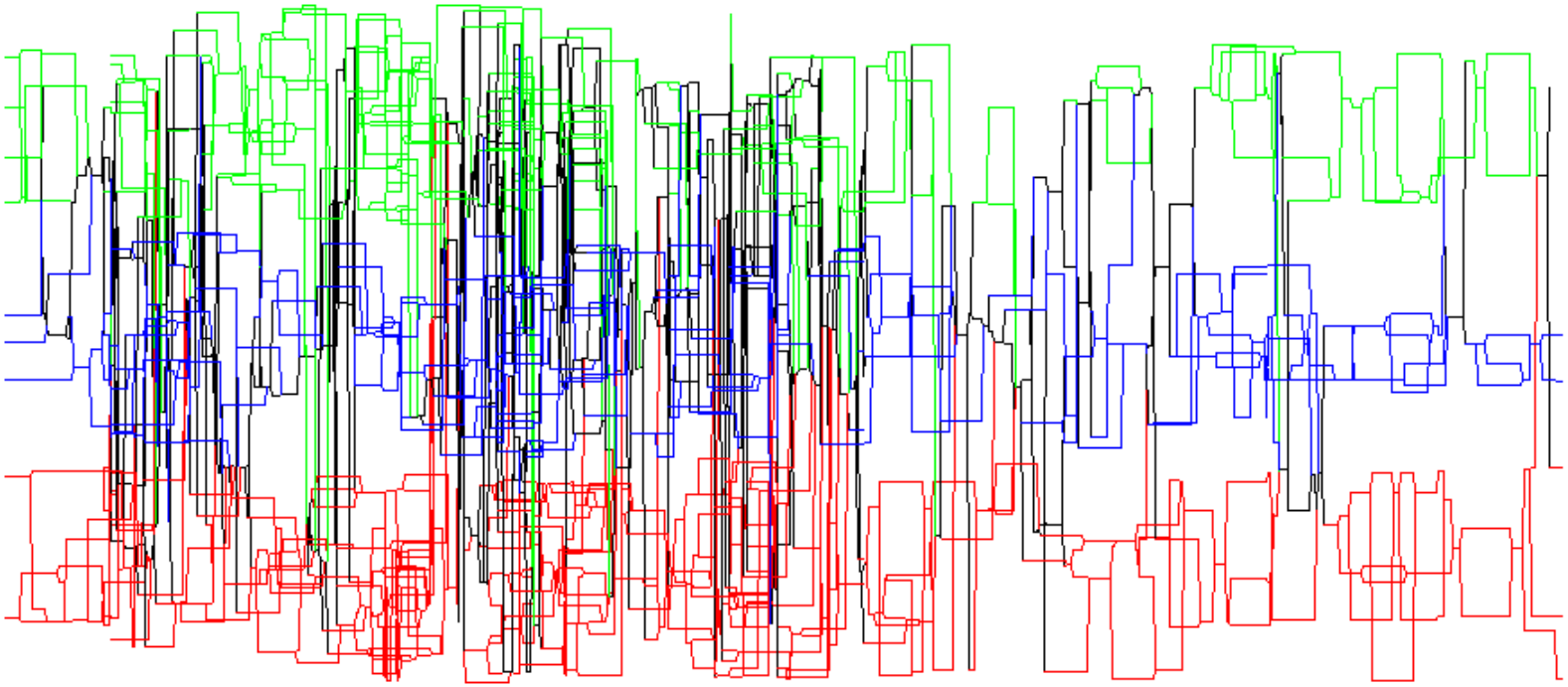
Zam Iqbal

Martine Zilversmit

Kiran Garimella

Kmer sharing demonstrates complexity of classical HLA allele sequence

- Structure of graph along CDS of 100 A, 100 B and 100 C alleles

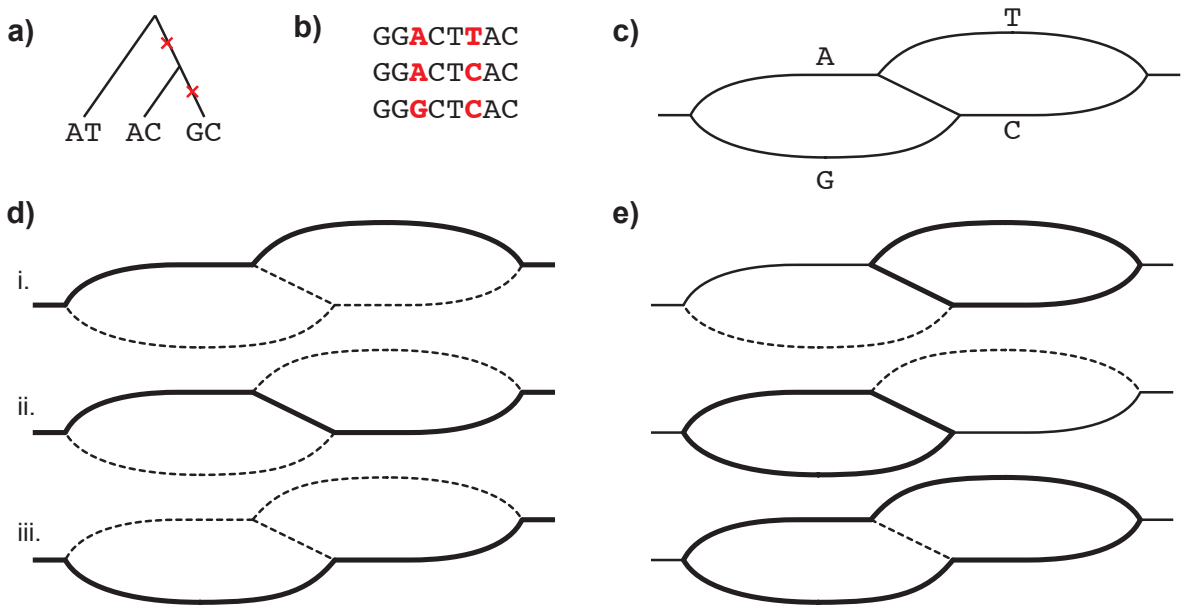


HLA-A only
HLA-B only
HLA-C only
Shared

k=50

Annotated De Bruijn Graph: Variant calling

1. Identify forks in the graph
2. Follow each path in each sample
3. Find where contigs join to find bubbles



a) b) polymorphisms in the population; c) variant induced graph structure
d) Contigs assembled; e) contigs combined to reconstruct bubbles