CS681: Advanced Topics in Computational Biology

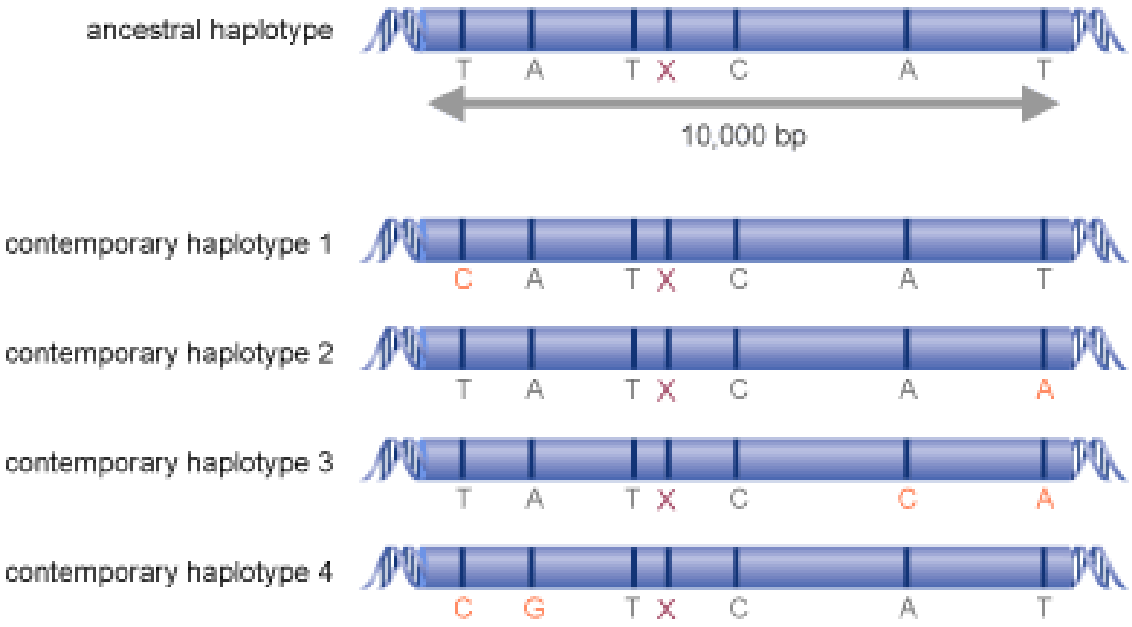
Can Alkan
EA509
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

HAPLOTYPE PHASING

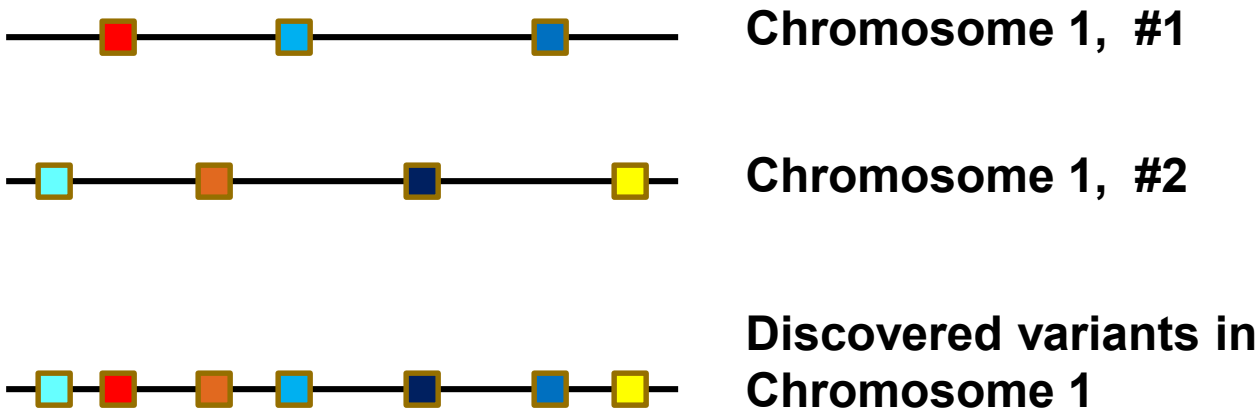
Haplotype

- "Haploid Genotype": *a combination of alleles at multiple loci that are transmitted together on the same chromosome*



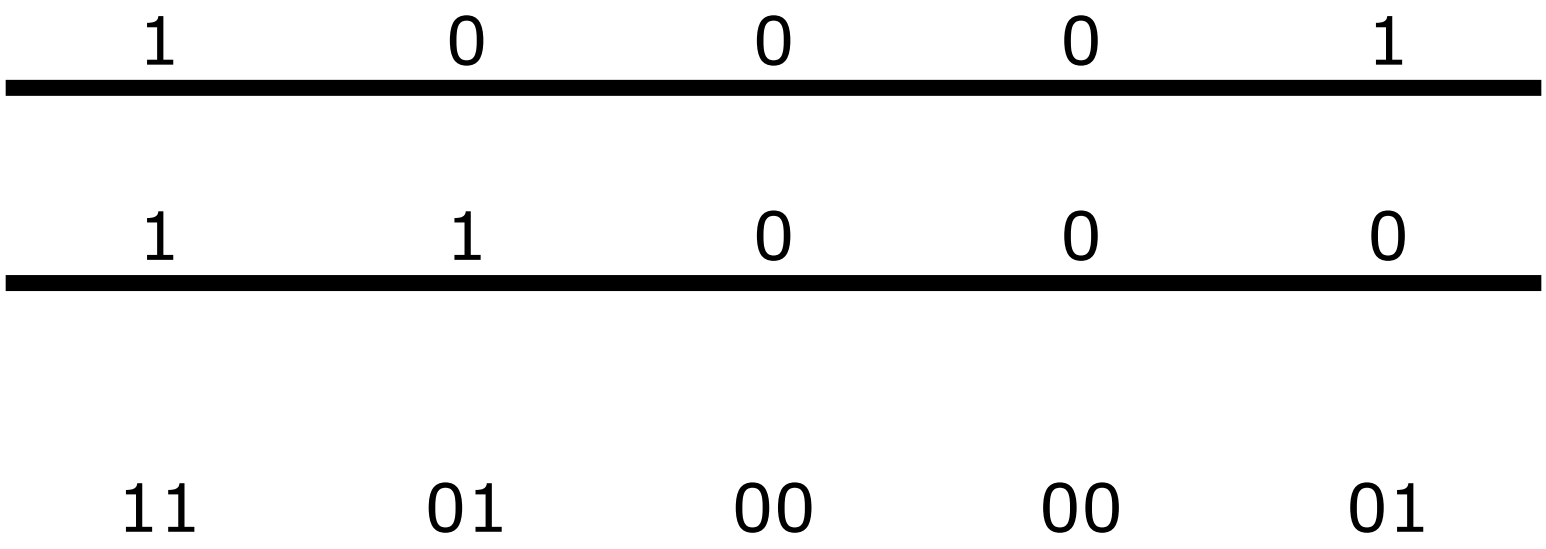
Haplotype resolution

- Variation discovery methods do not directly tell which copy of a chromosome a variant is located
- For heterozygous variants, it gets messy:



Haplotype resolution or haplotype phasing: finding which groups of variants "go together"

Haplotypes and genotypes (1)



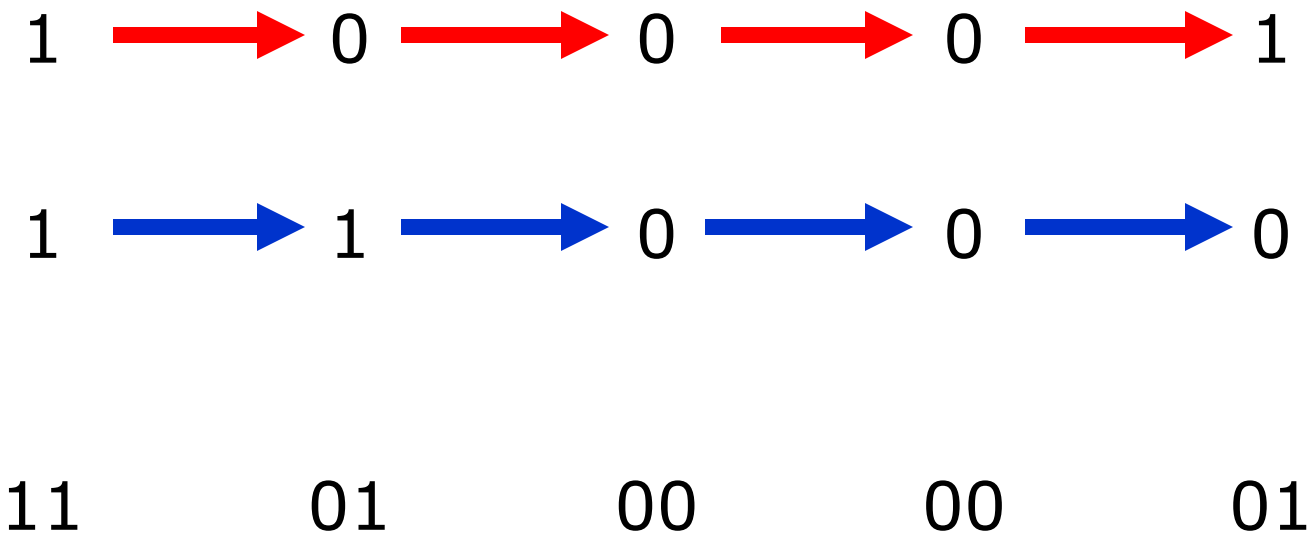
Slide from Andrew Morris

Haplotypes and genotypes (1)

1	0	0	0	1
1	1	0	0	0
11	01	00	00	01

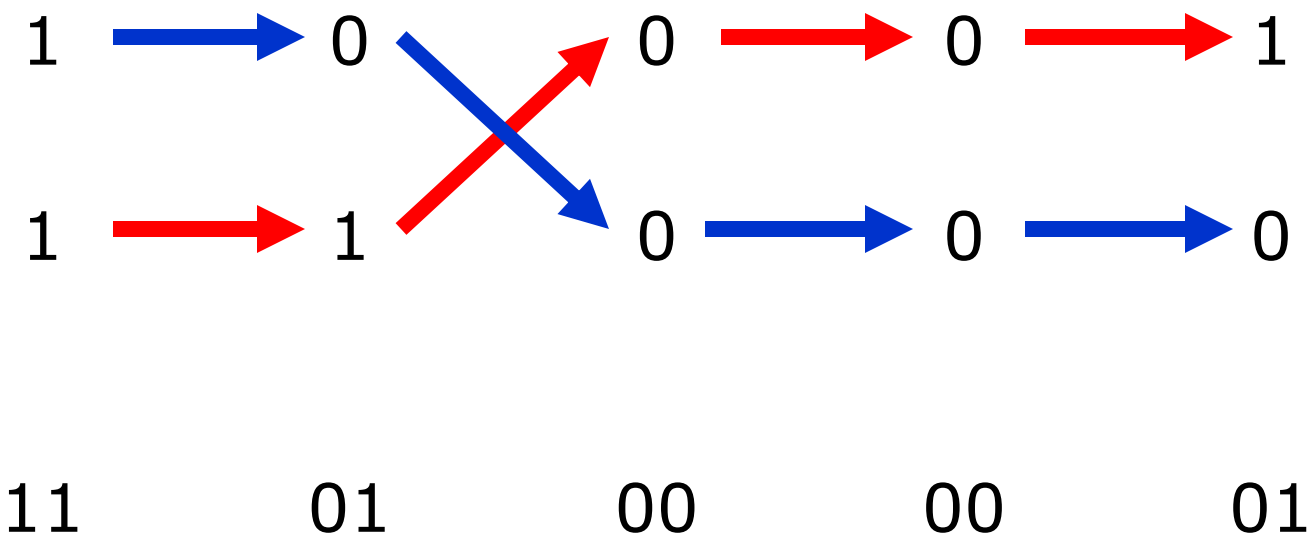
Slide from Andrew Morris

Haplotypes and genotypes (1)



Slide from Andrew Morris

Haplotypes and genotypes (1)



Slide from Andrew Morris

Haplotypes and genotypes (2)

- Individuals that are homozygous at every locus, or heterozygous at just one locus can be trivially **resolved**.
- Individuals that are heterozygous at k loci are consistent with 2^{k-1} configurations of haplotypes.

Slide from Andrew Morris

Why do we need haplotypes?

- Correlation between alleles at closely linked locations
- Fine-scale mapping studies.
- Association studies with multiple markers in candidate genes.
- Investigating patterns of linkage disequilibrium (LD) across genomic regions.
- Inferring population histories.

Slide from Andrew Morris

Simplex family data (1)

00 01 00 11 (M) x 01 11 01 01 (F)

00 01 01 01

Slide from Andrew Morris

Simplex family data (1)

00 01 00 11 x 01 11 01 01

(M) (F)

00 01 01 01

Slide from Andrew Morris

Simplex family data (1)

00 01 00 11 x 01 11 01 01

(M) (F)

00 01 01 01

Inferred haplotypes: 0001 / 0110

Slide from Andrew Morris

Simplex family data (2)

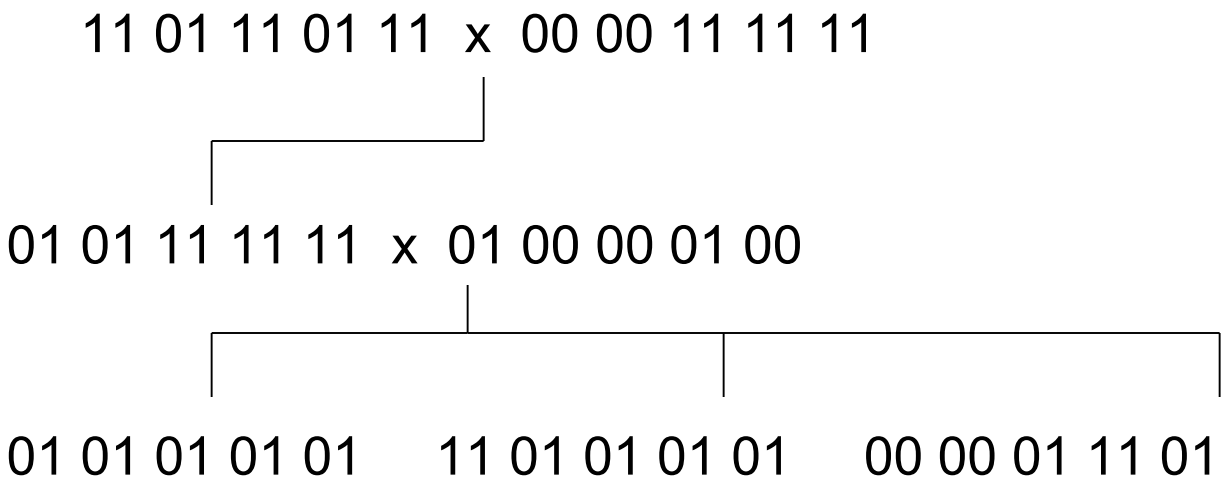
00 01 00 01 (M) x 01 01 00 01 (F)

00 01 00 01

- Cannot be fully resolved…

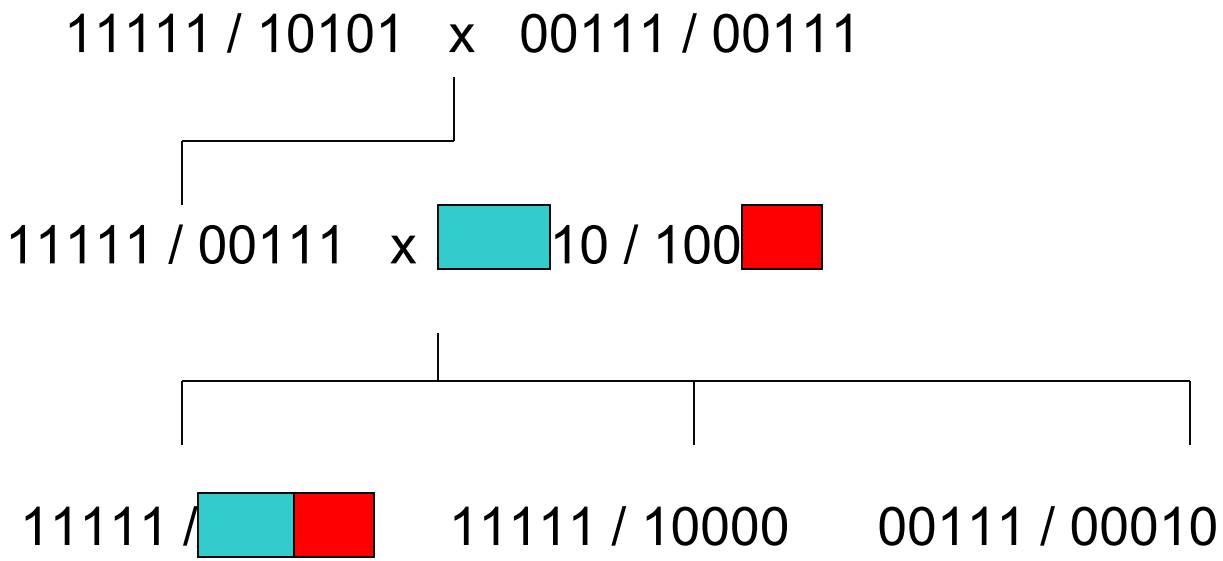
Slide from Andrew Morris

Pedigree data (1)



Slide from Andrew Morris

Pedigree data (1)



Slide from Andrew Morris

Pedigree data (2)

- Many combinations of haplotypes may be consistent with pedigree genotype data.
- Complex computational problem.
- Need to make assumptions about recombination.
- SIMWALK and MERLIN.

Slide from Andrew Morris

Statistical approaches to reconstruct haplotypes in unrelated individuals

- **Parsimony** methods: Clark's algorithm.
- **Likelihood** methods: E-M algorithm.
- **Bayesian** methods: PHASE algorithm.

- **Aims**: **reconstruct** haplotypes and/or **estimate population frequencies**.

Slide from Andrew Morris

Clark's algorithm (1)

- Reconstruct haplotypes in unresolved individuals via parsimony.
- Minimise number of haplotypes observed in sample.
- Microsatellite or SNP genotypes.

Slide from Andrew Morris

Clark's algorithm (2)

1. Search for **resolved** individuals, and record all recovered haplotypes.
2. Compare each **unresolved** individual with list of recovered haplotypes.
3. If a recovered haplotype is identified, individual is resolved.
4. Complimentary haplotype added to list of recovered haplotypes.
5. Repeat 2-4 until all individuals are resolved or no more haplotypes can be recovered.

Slide from Andrew Morris

Example

(A)	00 01 01 00
(B)	00 00 00 00
(C)	00 01 00 00
(D)	01 11 01 11
(E)	00 11 01 01
(F)	01 11 11 00
(G)	00 01 11 01
(H)	00 01 01 11
(I)	00 00 00 00
(J)	00 00 00 11

Slide from Andrew Morris

Example

(A)	00 01 01 00
(B)	00 00 00 00
(C)	00 01 00 00
(D)	01 11 01 11
(E)	00 11 01 01
(F)	01 11 11 00
(G)	00 01 11 01
(H)	00 01 01 11
(I)	00 00 00 00
(J)	00 00 00 11

Slide from Andrew Morris

Example

(A)	00 01 01 00
(B)	**0000 / 0000**
(C)	**0000 / 0100**
(D)	01 11 01 11
(E)	00 11 01 01
(F)	**0110 / 1110**
(G)	00 01 11 01
(H)	00 01 01 11
(I)	**0000 / 0000**
(J)	**0001 / 0001**

- Recovered haplotypes:

0000

0100

0110

1110

0001

Slide from Andrew Morris

Example

(A) 00 01 01 00
(B) **0000 / 0000**
(C) **0000 / 0100**
(D) 01 11 01 11
(E) 00 11 01 01
(F) **0110 / 1110**
(G) 00 01 11 01
(H) 00 01 01 11
(I) **0000 / 0000**
(J) **0001 / 0001**

- Recovered haplotypes:

0000
0100
0110
1110
0001

Slide from Andrew Morris

Example

(A) **0000 / 0110**
(B) **0000 / 0000**
(C) **0000 / 0100**
(D) 01 11 01 11
(E) 00 11 01 01
(F) **0110 / 1110**
(G) 00 01 11 01
(H) 00 01 01 11
(I) **0000 / 0000**
(J) **0001 / 0001**

- Recovered haplotypes:

0000 0111
0100
0110
1110
0001

Slide from Andrew Morris

Example

(A) **0000 / 0110**
(B) **0000 / 0000**
(C) **0000 / 0100**
(D) 01 11 01 11
(E) **0100 / 0111**
(F) **0110 / 1110**
(G) 00 01 11 01
(H) 00 01 01 11
(I) **0000 / 0000**
(J) **0001 / 0001**

- Recovered haplotypes:

0000 0111
0100 0011
0110
1110
0001

Slide from Andrew Morris

Example

(A) **0000 / 0110**
(B) **0000 / 0000**
(C) **0000 / 0100**
(D) **0111 / 1101**
(E) **0100 / 0111**
(F) **0110 / 1110**
(G) **0110 / 0011**
(H) **0001 / 0111**
(I) **0000 / 0000**
(J) **0001 / 0001**

- Recovered haplotypes:

0000	0111
0100	0011
0110	1101
1110	
0001	

Slide from Andrew Morris

Example: problem…

(A) **0000 / 0110**
(B) **0000 / 0000**
(C) **0000 / 0100**
(D) 01 11 01 11
(E) **0100 / 0111**
(F) **0110 / 1110**
(G) 00 01 11 01
(H) 00 01 01 11
(I) **0000 / 0000**
(J) **0001 / 0001**

- Recovered haplotypes:

0000 0111
0100 0011
0110
1110
0001

Slide from Andrew Morris

Example: problem…

(A) **0000 / 0110**
(B) **0000 / 0000**
(C) **0000 / 0100**
(D) 01 11 01 11
(E) **0100 / 0111**
(F) **0110 / 1110**
(G) 00 01 11 01
(H) 00 01 01 11
(I) **0000 / 0000**
(J) **0001 / 0001**

- Recovered haplotypes:

0000 0111
0100 0010
0110
1110
0001

Slide from Andrew Morris

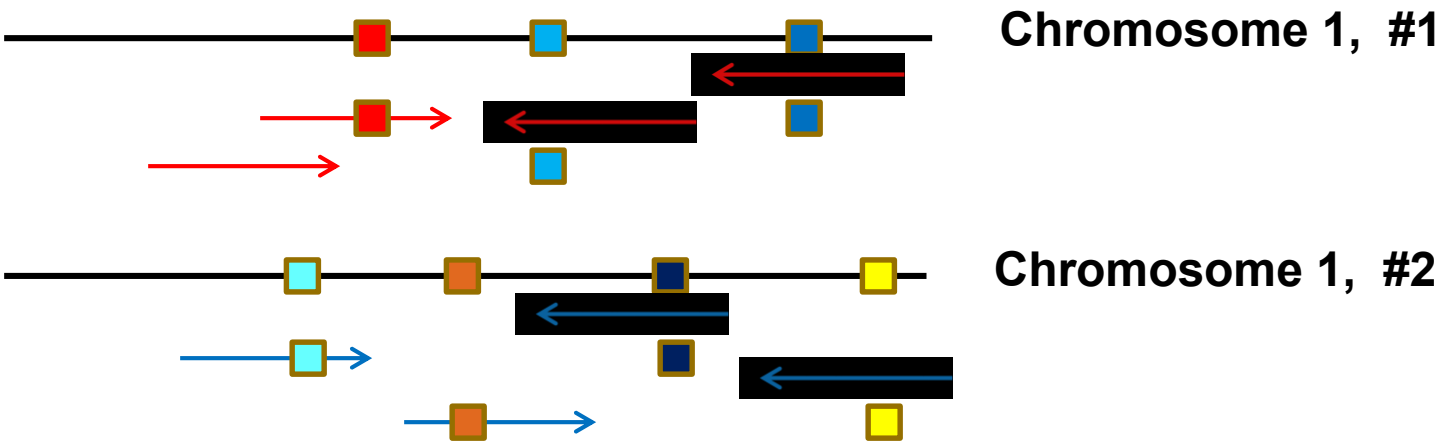
Clark's algorithm: problems

- Multiple solutions: try many different orderings of individuals.
- No starting point for algorithm.
- Algorithm may leave many unresolved individuals.
- How to deal with missing data?

Slide from Andrew Morris

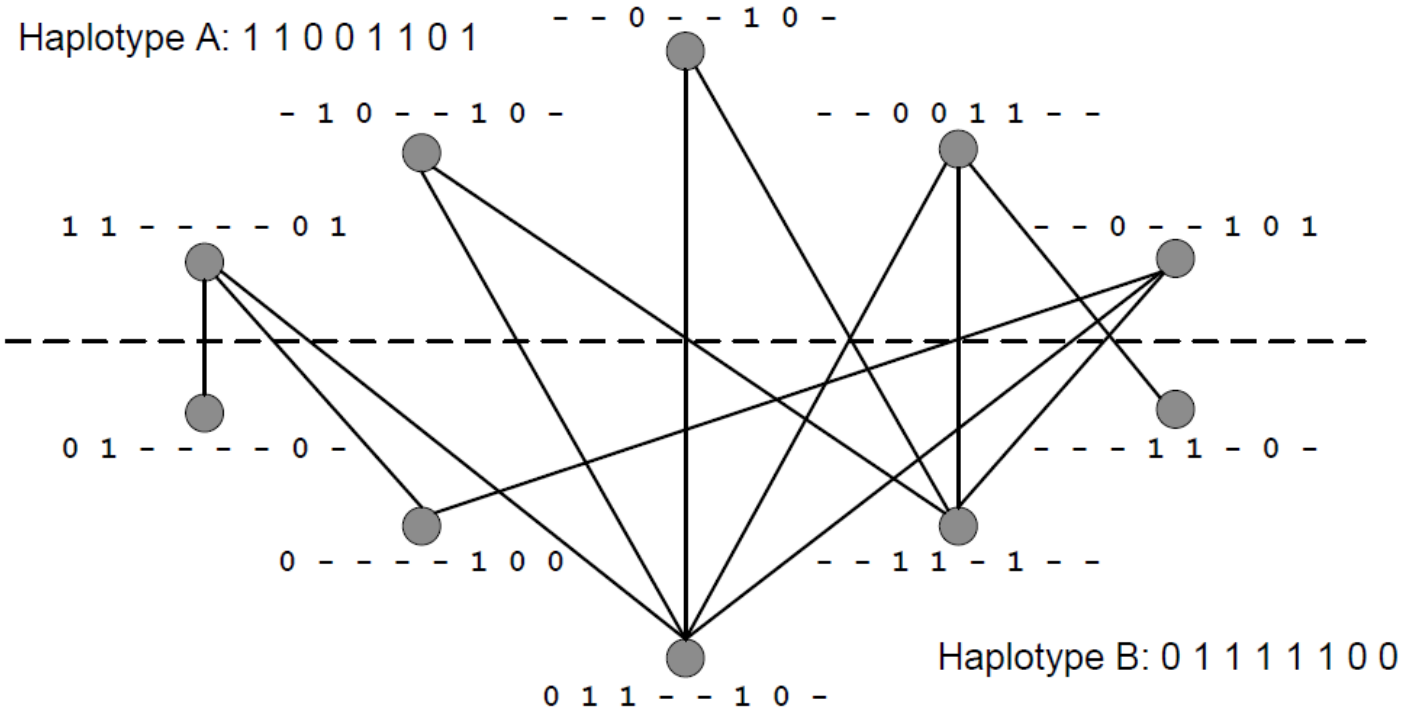
Haplotype phasing with PE sequences

PE sequences are from the same molecule, thus same haplotype



- Build initial shared haplotypes from PE reads
- Assemble shared haplotypes to get larger phased blocks

Fragment conflict graph



Two fragments conflict if they cover a common SNP with different alleles

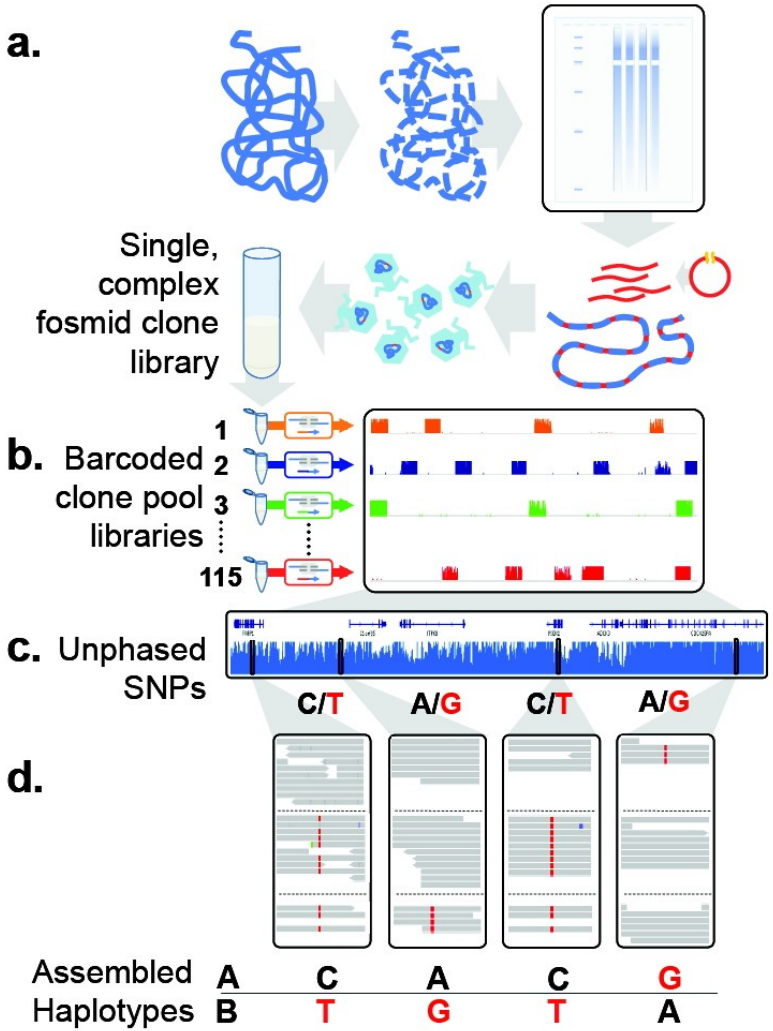
Halldorsson et al., PSB 2011

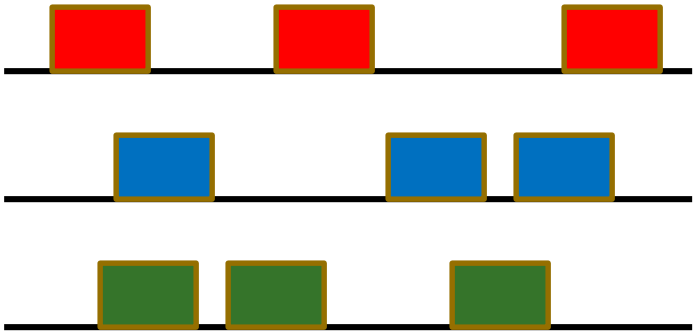
Pooled clone sequencing

- Instead of short paired-ends, use fosmids (40 kb)
  - Build fosmid library
  - Dilute the concentration of the library to cover the genome ~5X
  - Merge ~5000 fosmids in a pool
    - Total 114 pools
  - Sequence pools & separate fosmids *in silico*

Kitzman et al., Nature Biotechnology, 2011

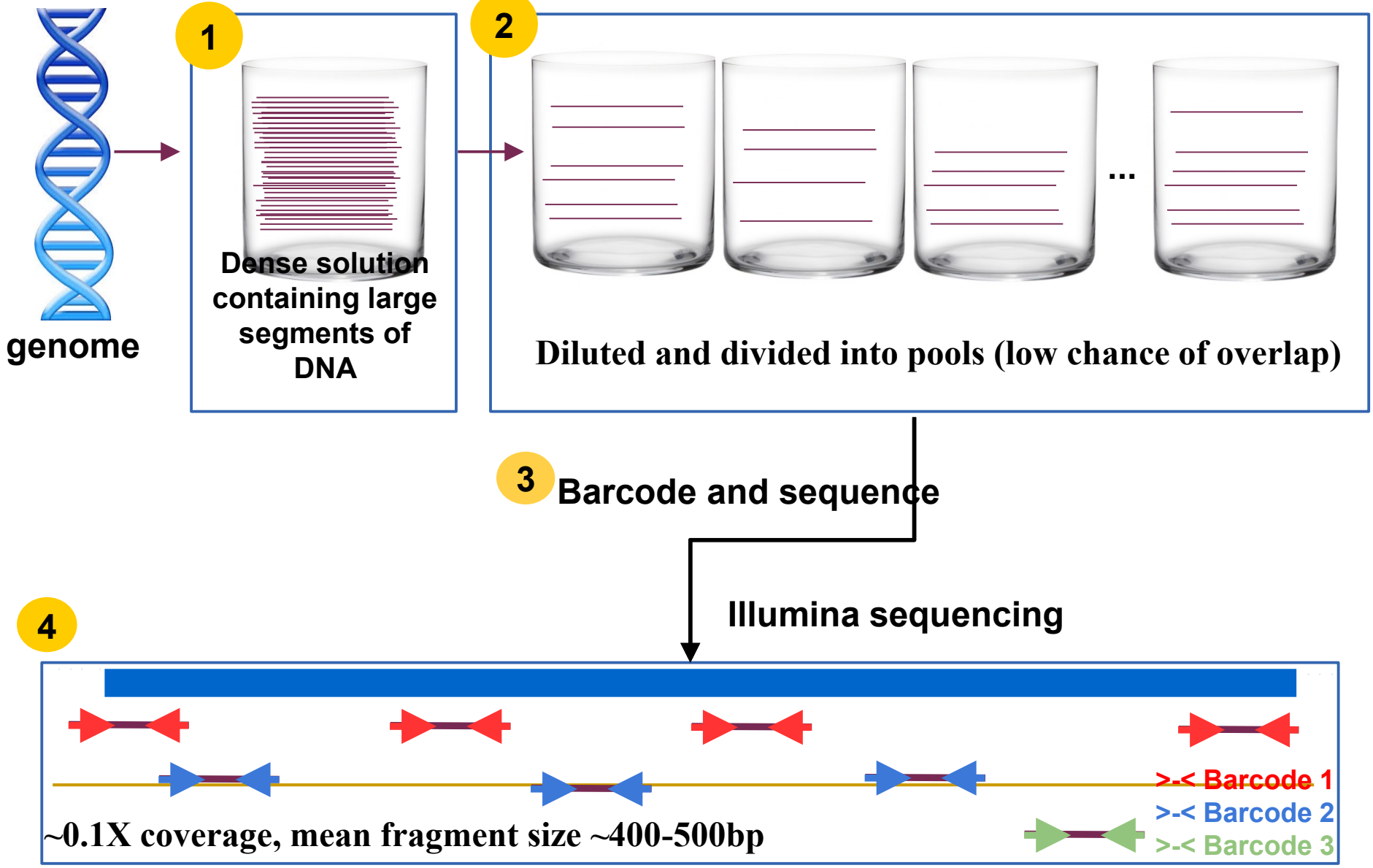
Pooled clone sequencing





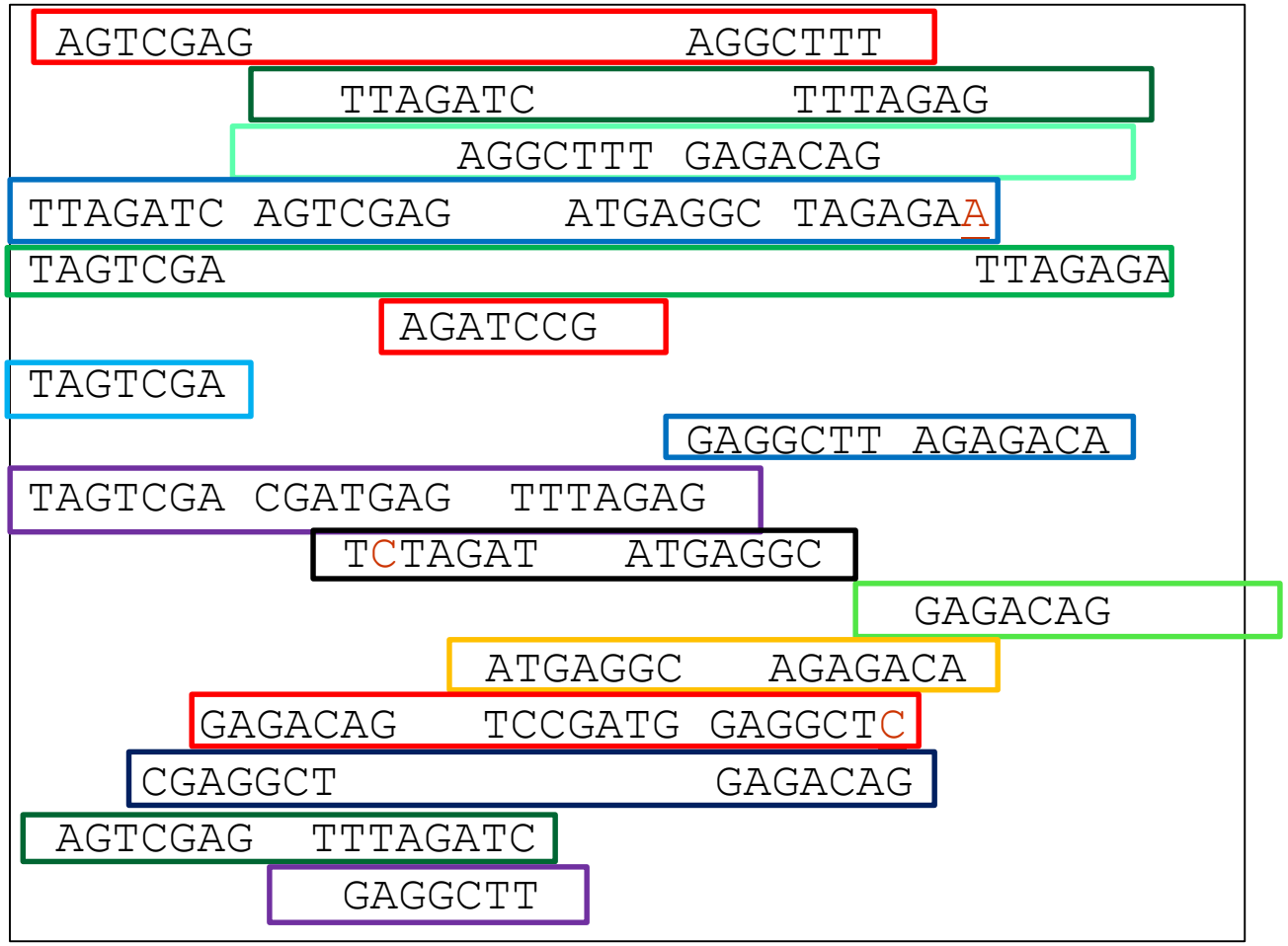
- Each fosmid represents one haplotype
- Resolve in ~40 kb blocks
- Extend blocks by overlapping fosmids in different pools

Long Range Information: Linked-Reads



A quick example – Linked-Reads

sample TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG



reference TACCGTCGAGCCTTTAGATCCGATGAG--TTTAGAGACAG

10x Genomics Linked-Reads

- ~45 Kb (average) molecules
- Automated process
  - No cloning bias, but size distribution problematic
- ~0.1x coverage per molecule
- Up to 4M barcodes
  - ~2-3 molecules per barcode