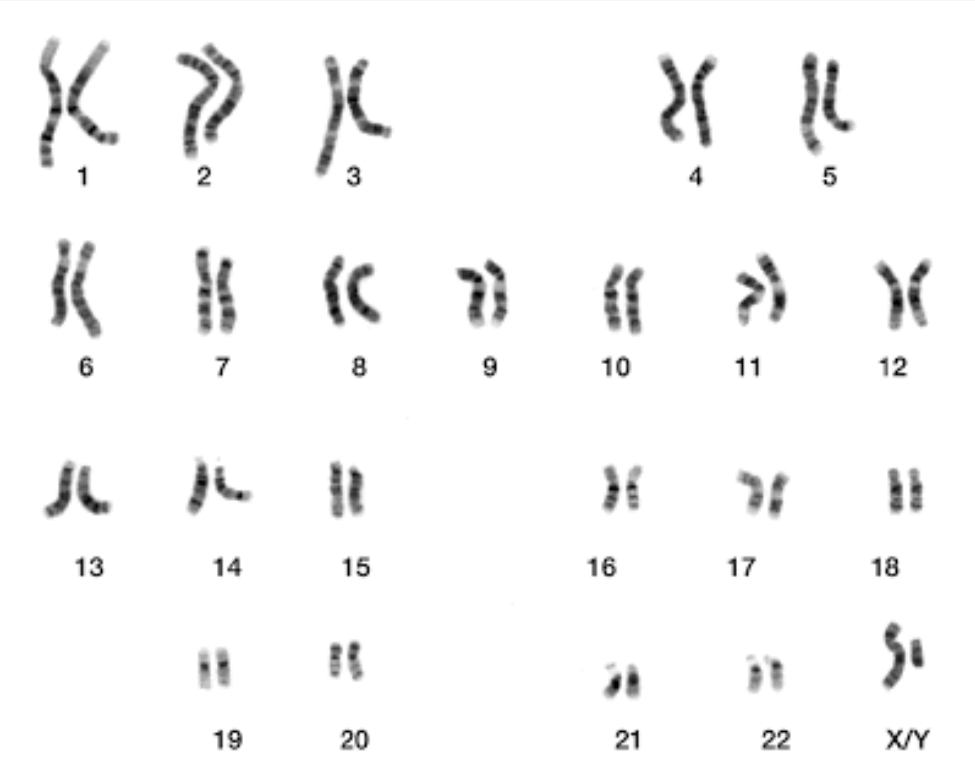
Haplotype assembly

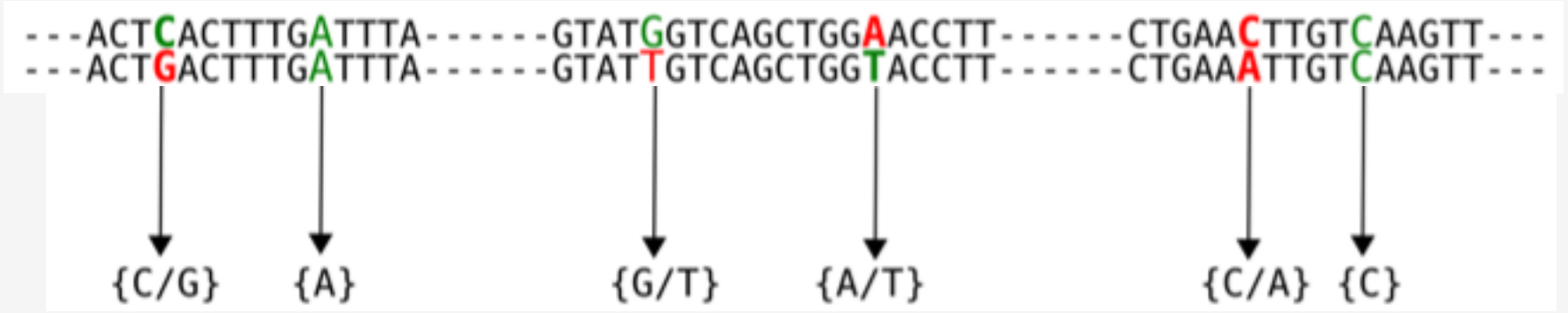
Vikas Bansal

Humans are diploid



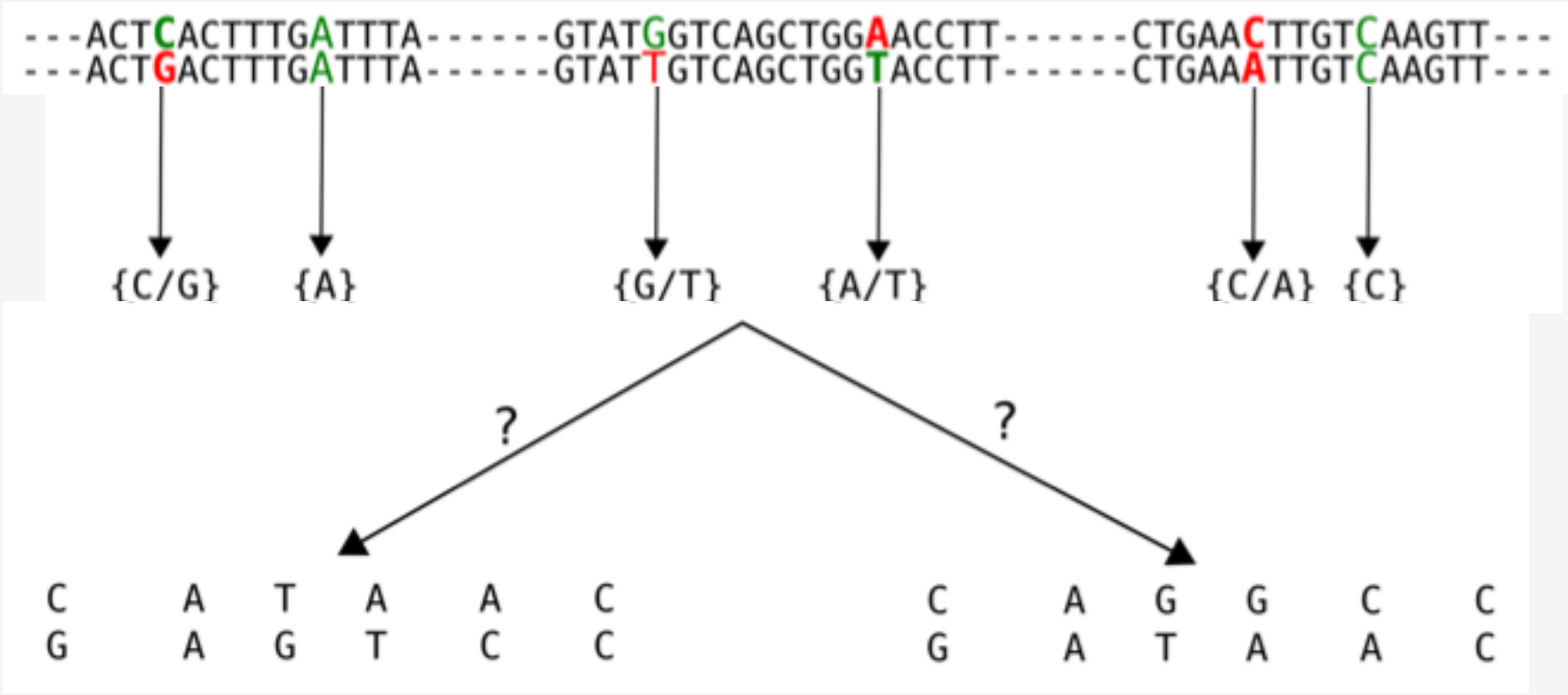
- Humans have two copies of each chromosome
  - Inherited from mother and father

Genotyping combines paternal and maternal information



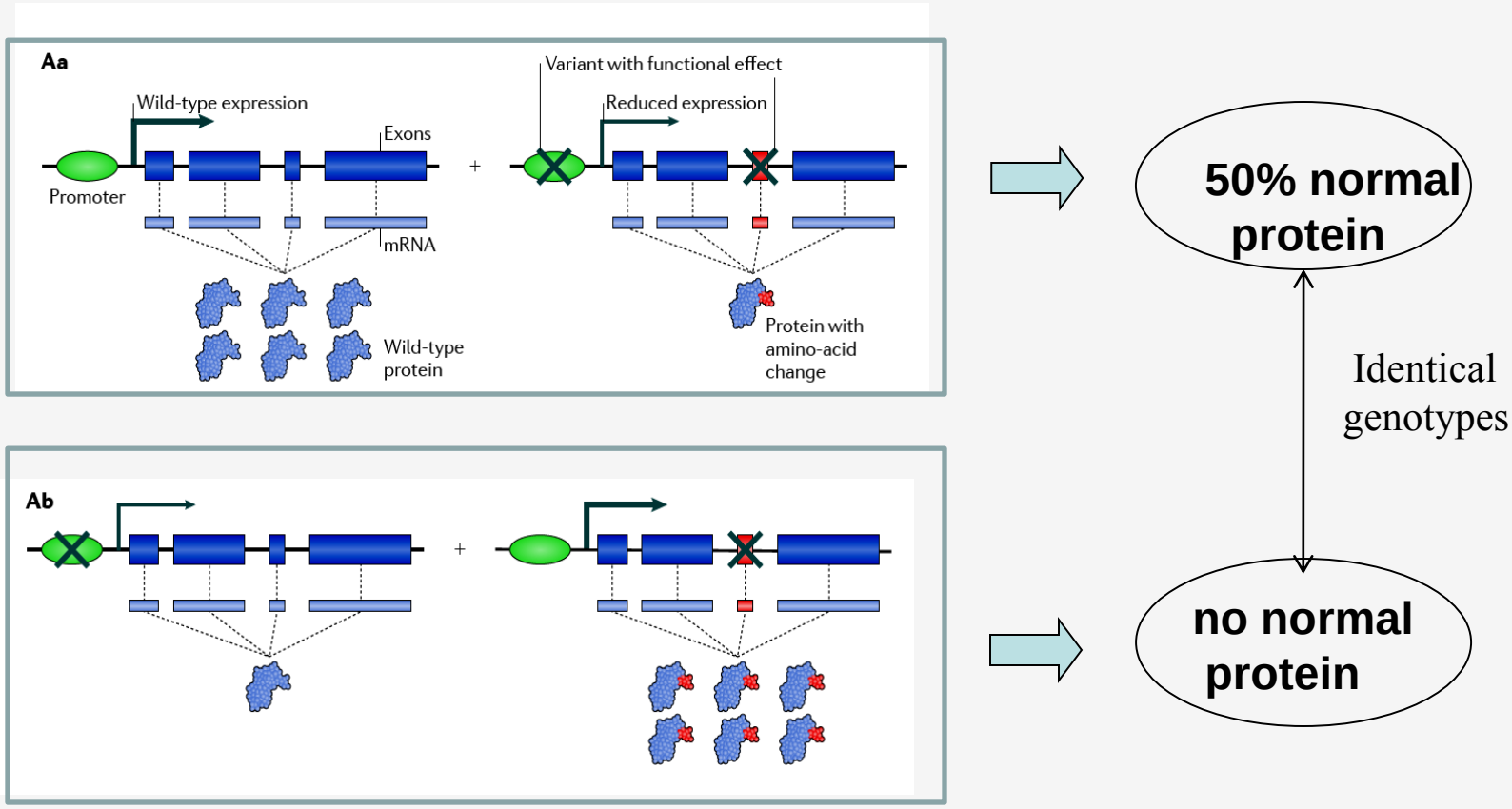
- haplotypes
  - C A G A C C
  - G A T T A C

Haplotype phasing problem



Why do we need haplotypes?

1. Finding disease genes



Why do we need haplotypes?

2. Imputation

Genotype at subset of SNPs	C/G ? G/T ? C/A C
Infer phase	C ? G ? C C G ? T ? A C
Impute missing markers	C A G A C C G A T T A C

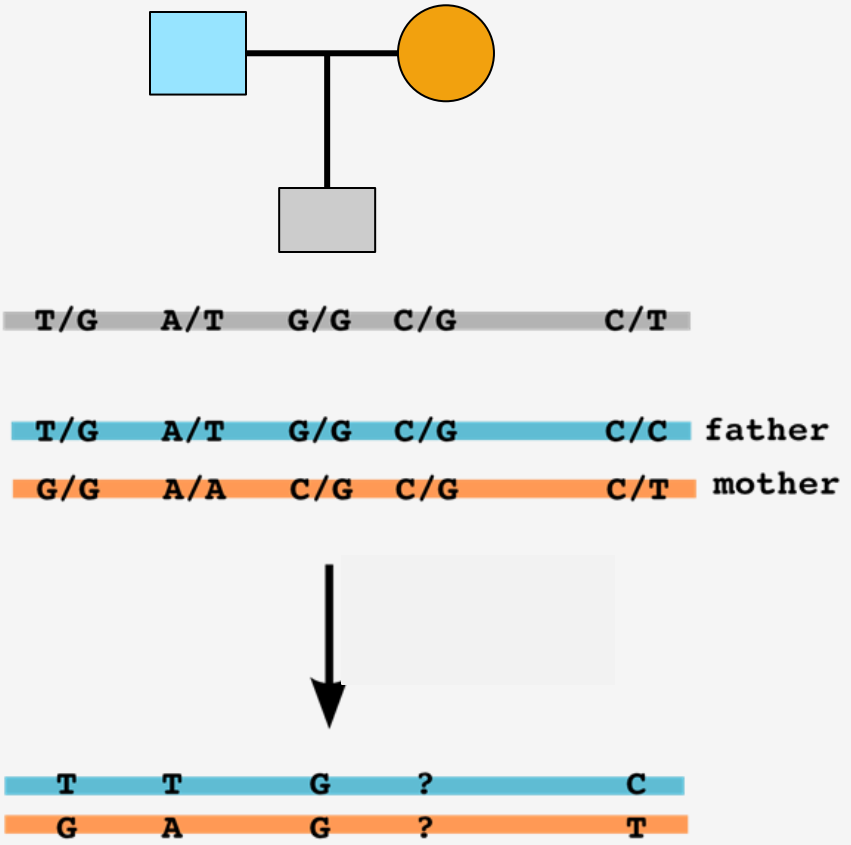
Haplotype reference panel

C	C	T	C	A	C
C	A	G	A	C	C
C	A	T	A	C	C
G	A	T	T	A	C
C	C	G	A	C	T
C	A	G	A	C	T
G	A	T	C	C	C
G	A	G	A	A	C

- several million SNPs can be imputed using 500K genotypes
- Widely used in genome-wide association studies

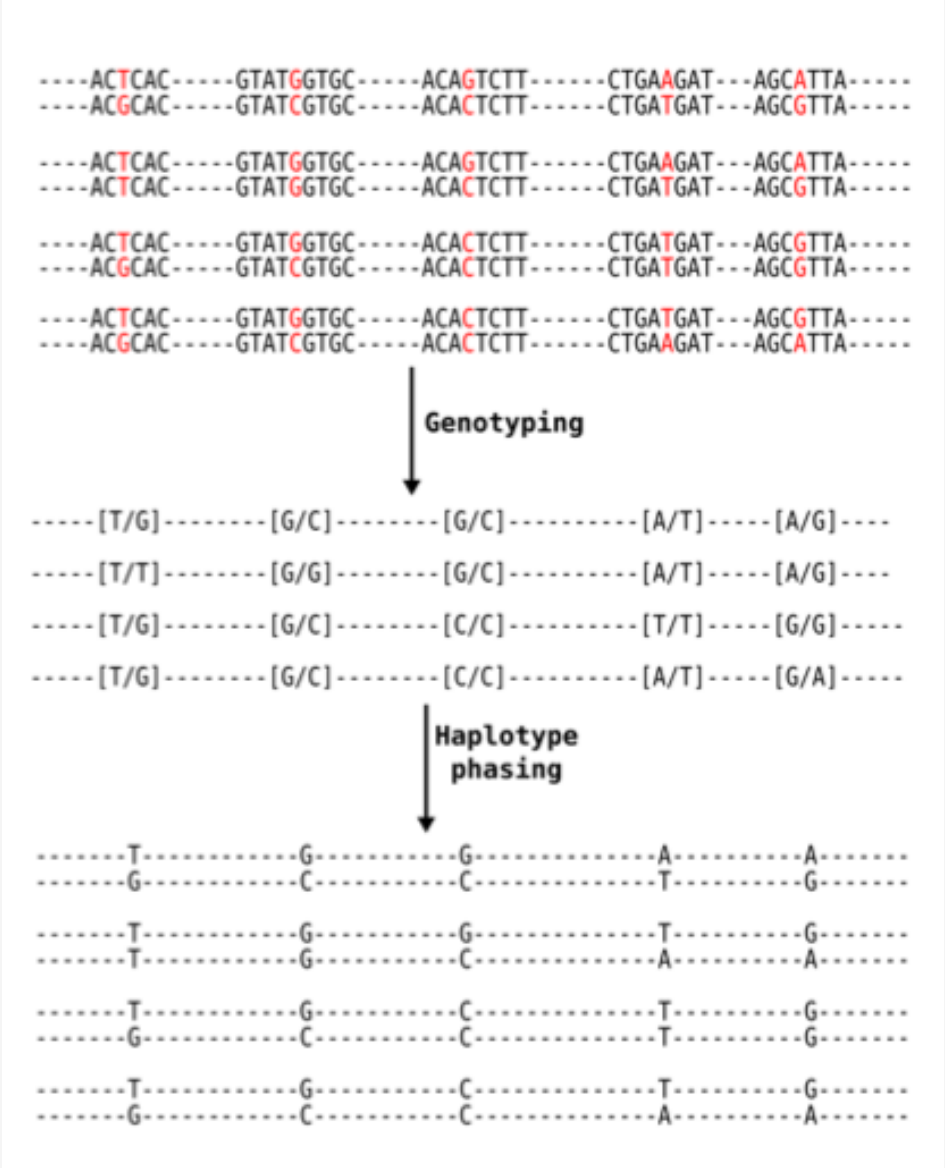
1. Family based phasing

- If child is heterozygous, and a parent is homozygous, we know which allele comes from which parent
- Can be generalized to pedigrees



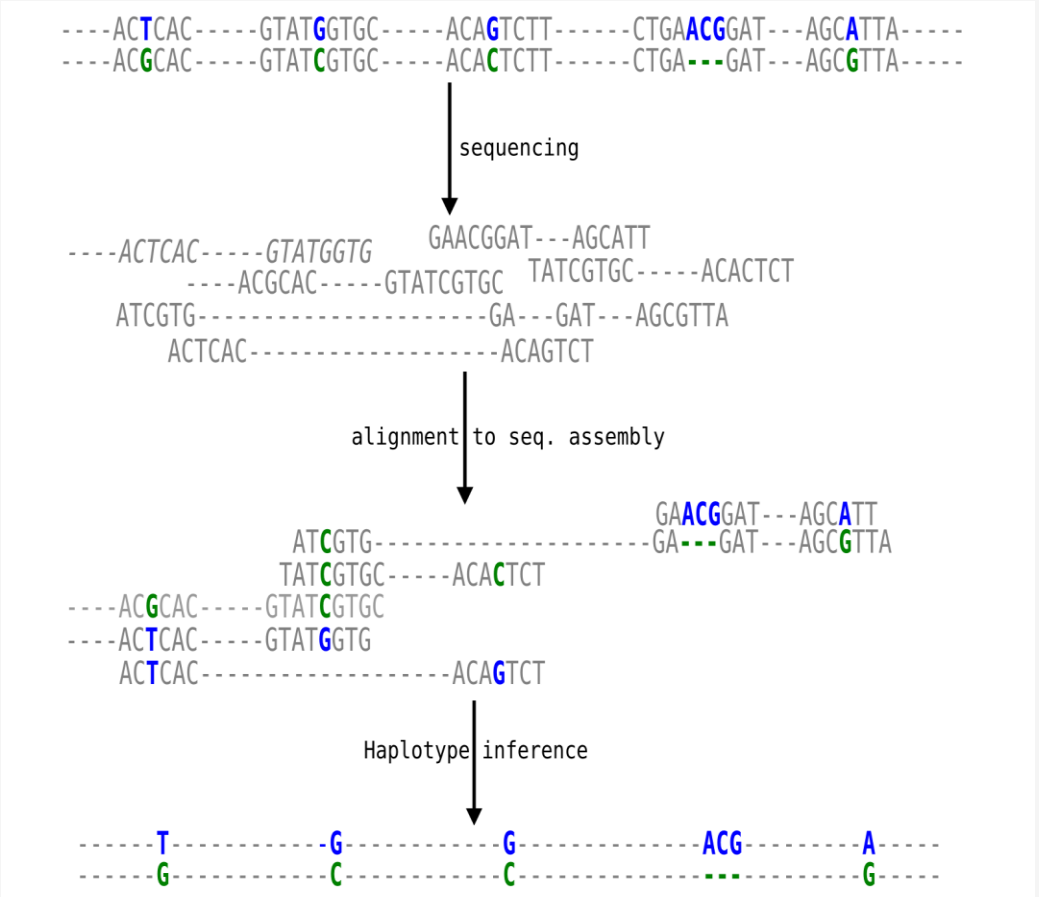
2. Haplotype phasing using populations

- Alleles at proximal SNPs are correlated (LD)
  - Few haplotypes are observed
- This phasing is most reliable for short regions



3. Sequence based haplotype phasing

- Reads that cover multiple heterozygous sites have phase information
- **Haplotype assembly**: use overlap between reads to infer two haplotypes for an individual



Personalized human genomes

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Hal
Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R.
Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vi
Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4]
Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig V

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Program in Genetics and G
Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 Department of Computer Scien
California, United States of America, 4 Genetics Department, Facultat de Biologia, Universitat de Barcel

**Presented here is a genome sequence of an individual human. It w
fragments, sequenced by Sanger dideoxy technology and assembled i
bases (Mb) of contiguous sequence with approximately 7.5-fold cove
modified version of the Celera assembler to facilitate the identification
individual diploid genome. Comparison of this genome and the Nationa
reference assembly revealed more than 4.1 million DNA variants, enc**

DNA pioneer Watson gets own genome map

By Nicholas Wade

Published: June 1, 2007

The full genome of James Watson, who jointly discovered the structure of DNA in 1953, has been deciphered, marking what some scientists believe is the gateway to an impending era of personalized genomic medicine.

5 diggs

digg it

Illumina unveils genome sequence of African male

nature.com — Illumina, a biotechnology company based in San Diego, California, announced on 6 February that it has sequenced the complete genome of an African man.

Haplotype-resolved genome sequencing of a Gujarati Indian individual

Nature Biotech 2011

Jacob O Kitzman[1], Alexandra P MacKenzie[1], Andrew Adey[1], Joseph B Hiatt[1], Rupali P Patwardhan[1], Peter H Sudmant[1], Sarah B Ng[1], Can Alkan[1,2], Ruolan Qiu[1], Evan E Eichler[1,2] & Jay Shendure[1]

Haplotype information is essential to the complete description and interpretation of genomes[1], genetic diversity[2] and genetic ancestry[3]. Although individual human genome sequencing is increasingly routine[4], nearly all such genomes are unresolved with respect to haplotype. Here we combine the throughput of massively parallel sequencing[5] with the contiguity information provided by large-insert cloning[6] to experimentally determine the haplotype-resolved genome of a South Asian individual. A single fosmid library was split into a modest number of pools, each providing ~3% physical coverage of the diploid genome. Sequencing of each pool yielded reads overwhelmingly derived from only one homologous chromosome at any given location. These data were combined with whole-genome shotgun sequence to directly phase 94% of ascertained heterozygous single nucleotide polymorphisms (SNPs) into long haplotype blocks (N50 of 386 kilobases (kbp)). This method also facilitates the analysis of structural variation, for example, to anchor novel insertions[7,8] to specific locations and haplotypes.

variants at which phased CHB/JPT HapMap data were available (CHB, Han Chinese from Beijing, China; JPT, Japanese from Tokyo, Japan)[12]. The genomes of a family of four have been sequenced and these relationships used to infer inheritance blocks[13]. Although they can be successful, inferential methods have limitations. Statistical phasing, whether based on genotyping[2] or sequencing[14], performs poorly when linkage disequilibrium is not high, and for rare variants. Phasing by pedigree analysis requires genome sequencing of many related individuals, increasing costs and limiting practical application.

We describe a cost-effective method for determining long-range haplotypes at a genome-wide scale by massively parallel sequencing of complex, haploid subsets of an individual genome (**Fig. 1**). We apply this method to the first reported whole-genome sequencing of a human of South Asian ancestry. The Indian subcontinent is home to myriad culturally and genetically diverse groups with distinct population histories[15]. We selected a female from the HapMap panel of 'Gujarati Indians in Houston' (GIH; NA20847) for sequencing. Notably, the imputation of genotypes for GIH was the least effective of all non-African populations in HapMap[2].

Haplotype assembly

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
A	G	A	G	-	-	-	-	-	-	-	-	-
C	T	T	-	-	-	-	-	-	-	-	-	-
-	-	A	G	T	-	-	-	-	-	-	-	-
-	-	A	-	-	T							
-	-	-	G	-	-	-	-	-	-	G	G	-
-	-	-	T	C	-	-	-	-	-	-	-	-
-	-	-	-	-	T	A	G	-	-	-	-	-
-	-	-	-	-	-	A	T	-	A	T	-	-
-	-	-	-	-	-	-	G	C	A	-	-	-
-	-	-	-	-	-	-	-	-	A	T	G	-
-	-	-	-	-	-	-	-	-	-	T	G	A
-	-	-	-	-	-	-	-	T	-	-	-	G

- The fragments are aligned to the reference genome
- Heterozygous sites known in advance
- Uninformative fragments and columns are removed
- Tri-allelic SNP columns are removed

Haplotype assembly formulation

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	–	–	–	–	–	–	–	–	–
1	1	1	–	–	–	–	–	–	–	–	–	–
–	–	0	0	1	–	–	–	–	–	–	–	–
–	–	0	–	–	0							
–	–	–	0	–	–	–	–	–	–	1	0	–
–	–	–	1	0	–	–	–	–	–	–	–	–
–	–	–	–	–	0	0	0	–	–	–	–	–
–	–	–	–	–	–	0	1	–	0	0	–	–
–	–	–	–	–	–	–	0	0	0	–	–	–
–	–	–	–	–	–	–	–	–	0	0	0	–
–	–	–	–	–	–	–	–	–	–	0	0	0
–	–	–	–	–	–	–	–	1	–	–	–	1

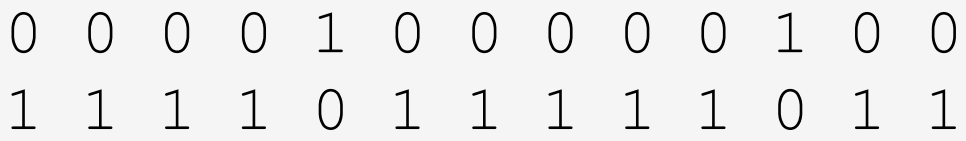
- The fragments are aligned to the reference genome
- Heterozygous sites known in advance
- Uninformative fragments and columns are removed
- Tri-allelic SNP columns are removed
- Relabel the two alleles using 0/1

Haplotype assembly

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	–	–	–	–	–	–	–	–	–
1	1	1	–	–	–	–	–	–	–	–	–	–
–	–	0	0	1	–	–	–	–	–	–	–	–
–	–	0	–	–	0							
–	–	–	0	–	–	–	–	–	–	1	0	–
–	–	–	1	0	–	–	–	–	–	–	–	–
–	–	–	–	–	0	0	0	–	–	–	–	–
–	–	–	–	–	–	0	1	–	0	0	–	–
–	–	–	–	–	–	–	0	0	0	–	–	–
–	–	–	–	–	–	–	–	–	0	0	0	–
–	–	–	–	–	–	–	–	–	–	0	0	0
–	–	–	–	–	–	–	–	1	–	–	–	1

- The fragments are aligned to the reference genome
- Heterozygous sites known in advance
- Uninformative fragments and columns are removed
- Tri-allelic SNP columns are removed.
- Relabel the two alleles using 0/1

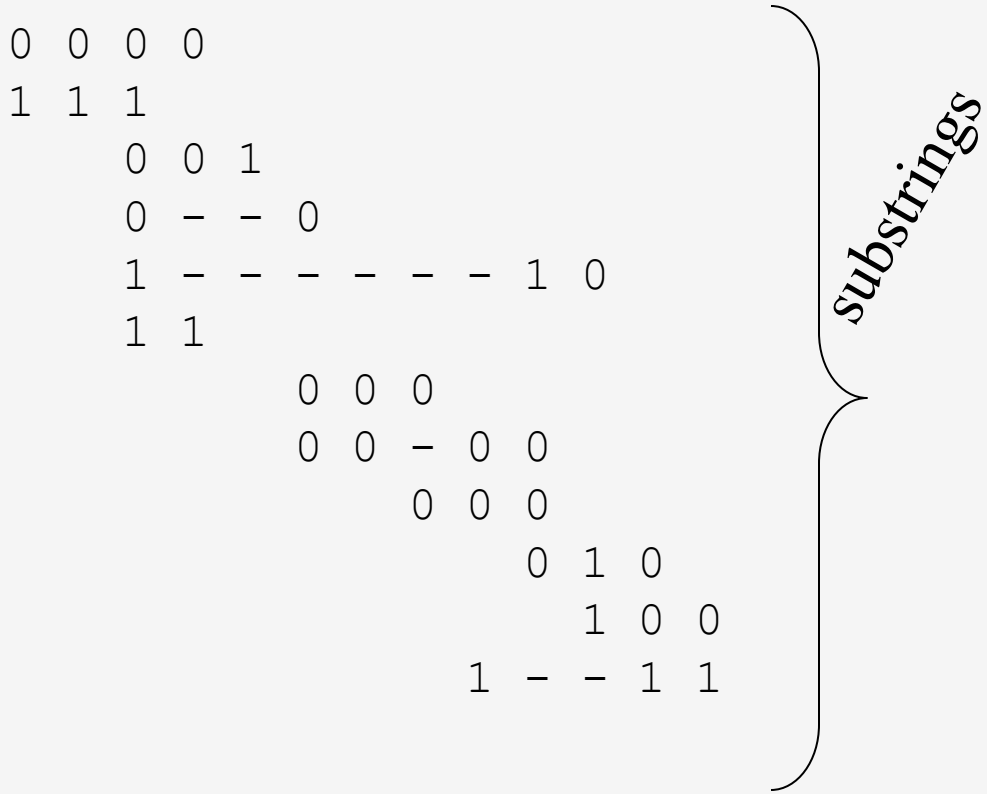
A combinatorial problem



- Consider a binary string (and its complement)

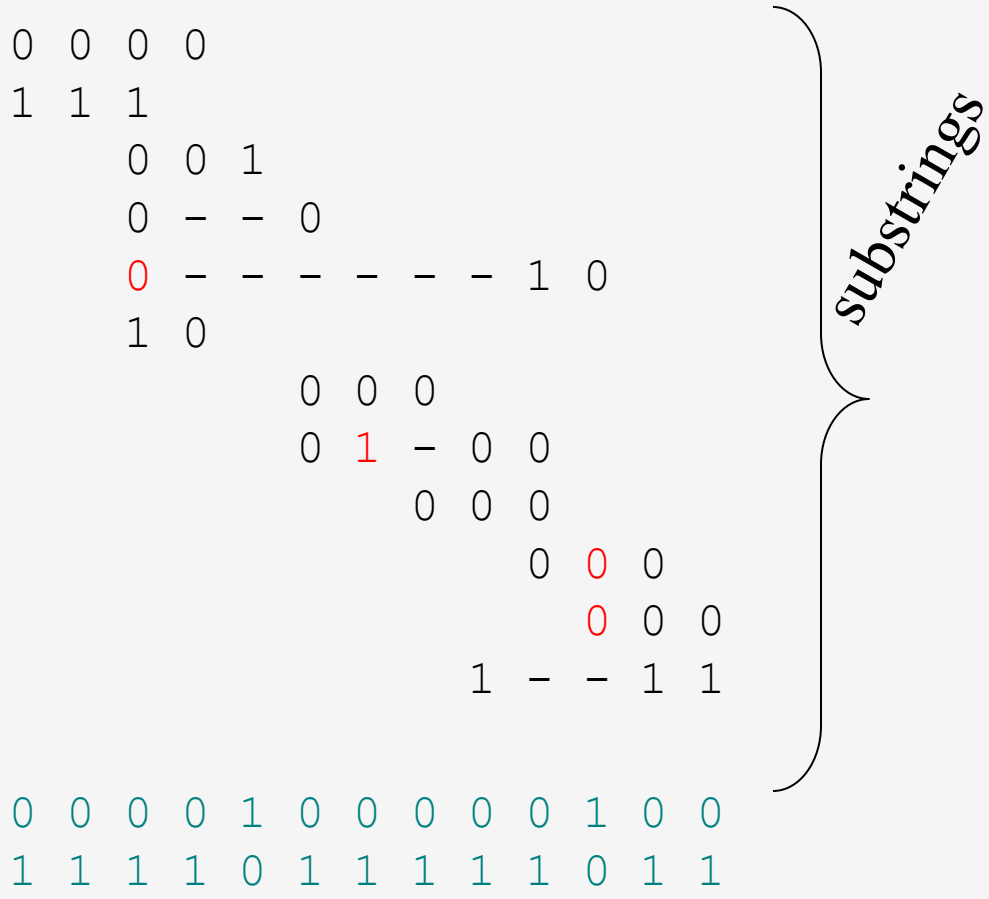
Sampling from the string, and its complement

- The string is revealed to us only through a collection of substrings of the string, and its complement.
- Given the substrings, can the string be reconstructed?



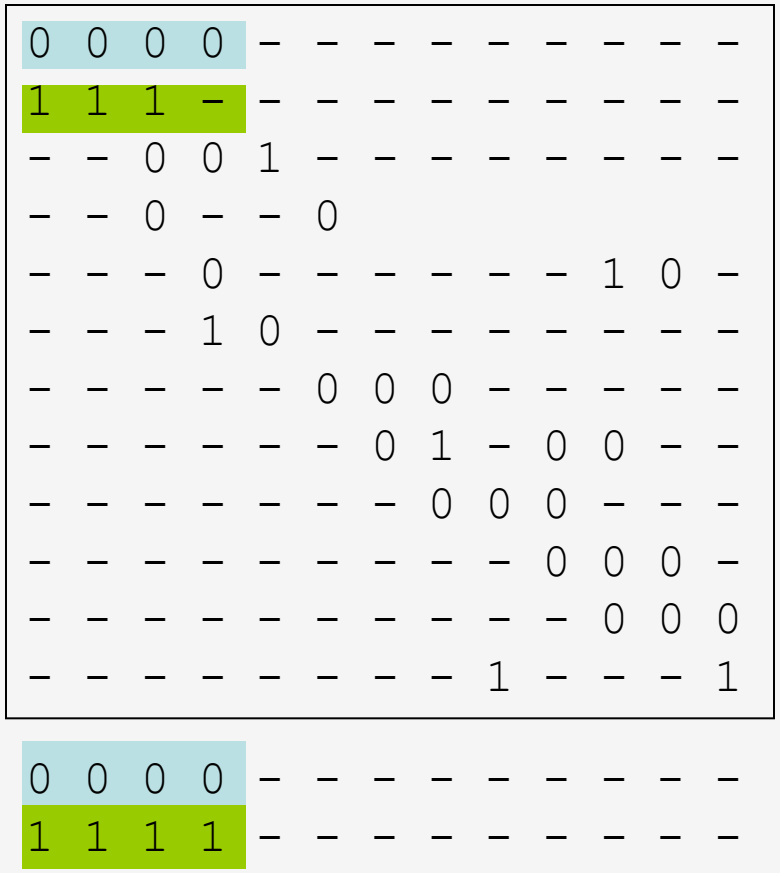
Inference in the presence of errors

- The error free reconstruction is unique.
- The problem becomes much harder if some of the substrings have errors (do not match the consensus)
- Define MEC: Minimum # calls that need to be changed for each fragment to be matched perfectly
- MEC reconstruction: Find the string that minimizes the MEC error.
- MEC reconstruction is NP-hard even when all fragments have length 2!



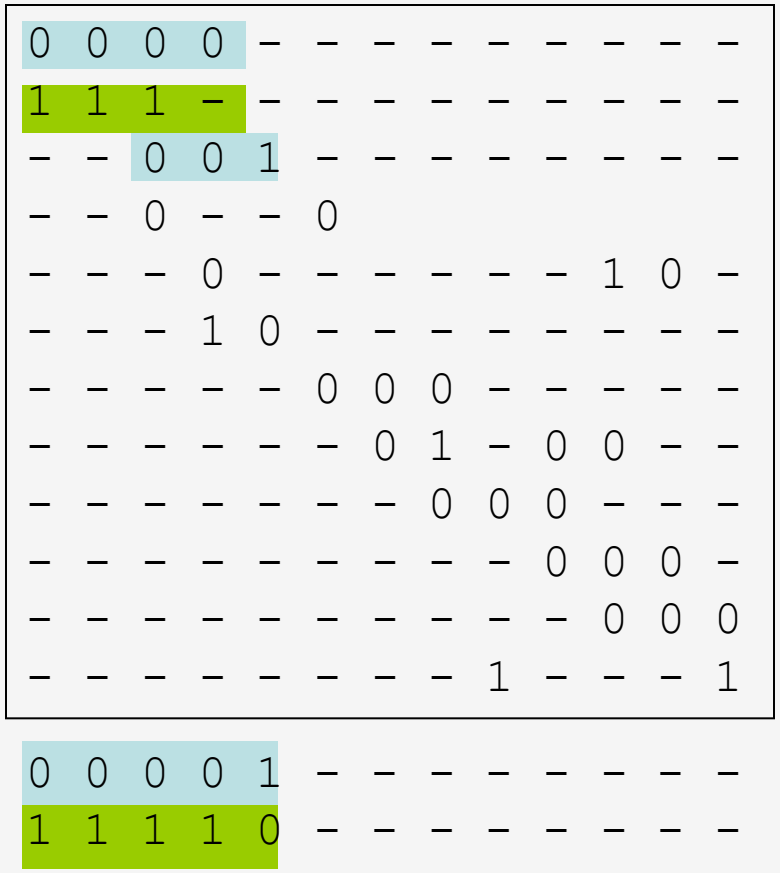
A simple greedy approach

- Greedily select a fragment that extends the current haplotype



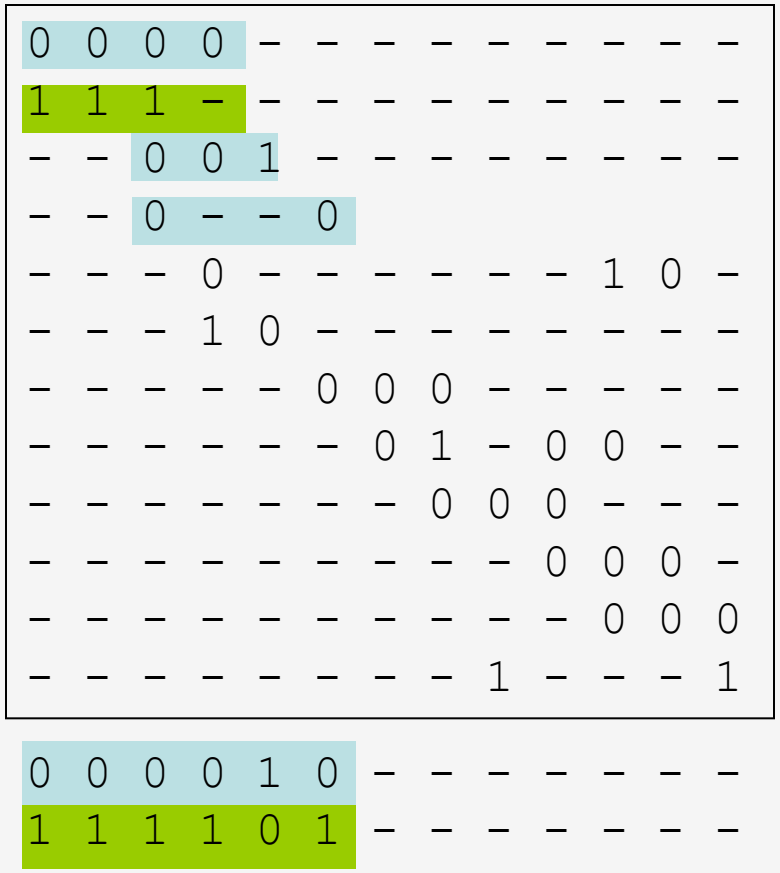
A simple greedy approach

- Greedily select a fragment that extends the current haplotype



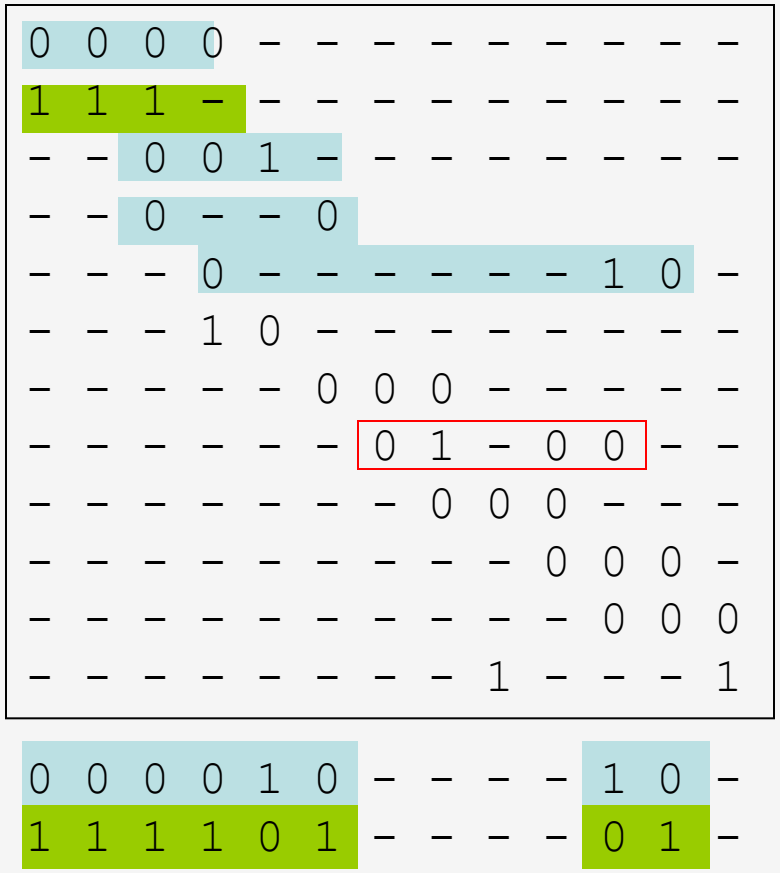
A simple greedy approach

- Greedily select a fragment that extends the current haplotype



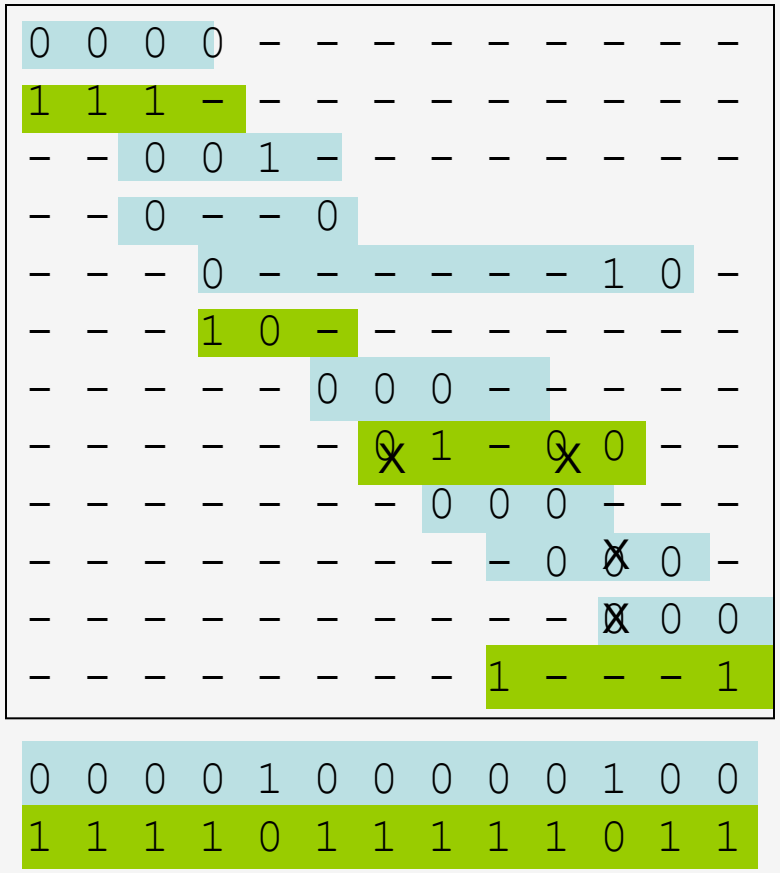
A simple greedy approach

- Some fragments will not match without error.
- These are ‘assigned & corrected’ greedily



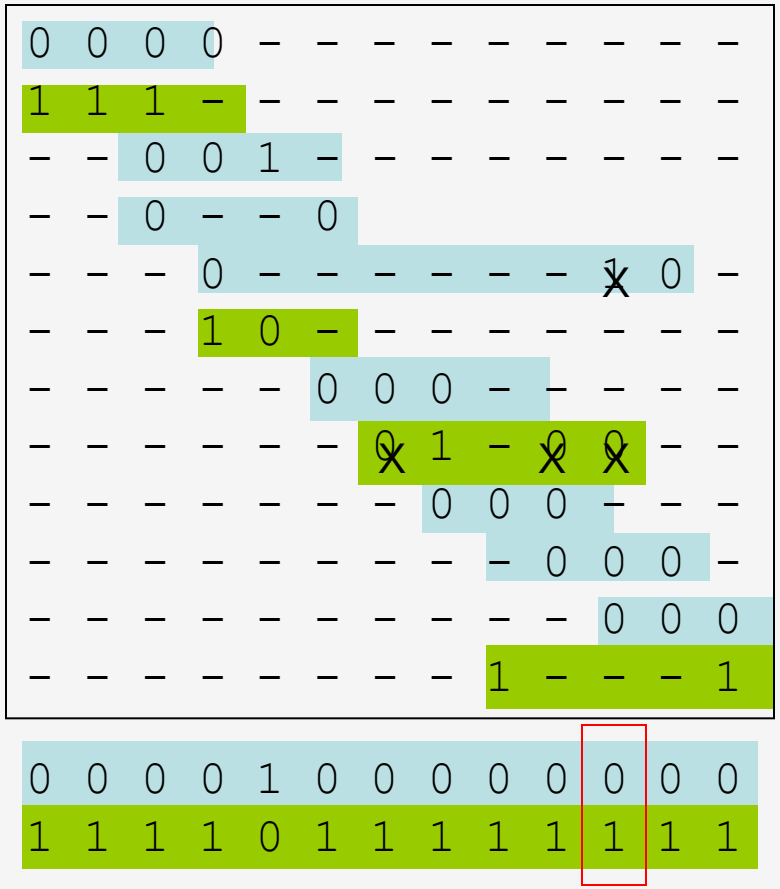
Greedy haplotype assembly

- MEC: The minimum number of base-calls that need to be flipped for an error free assignment
- MEC score =



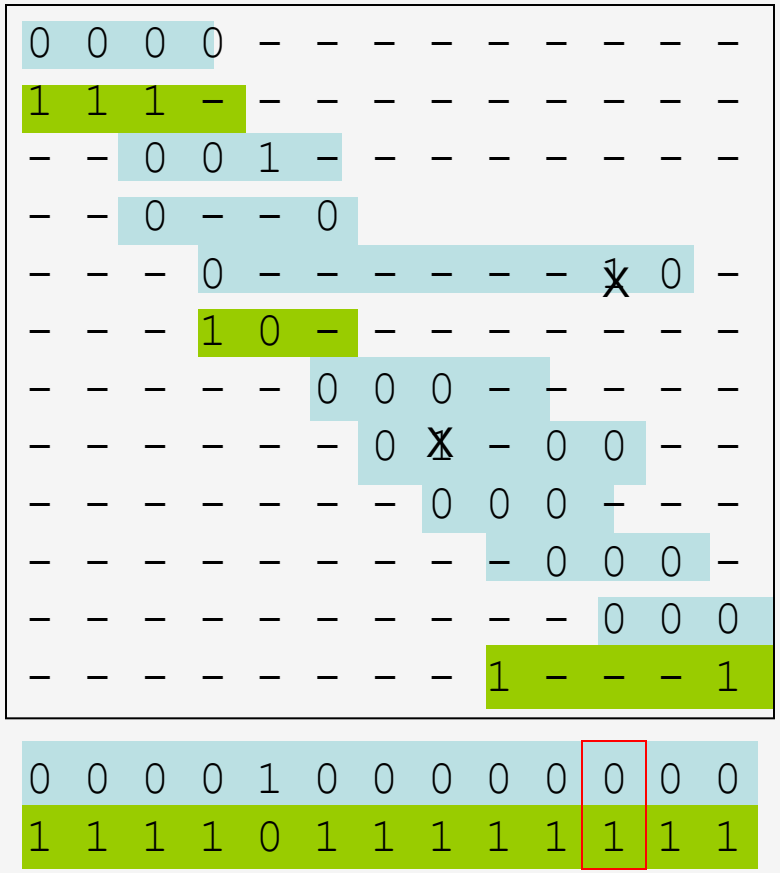
Modifying the haplotypes

- The Greedy approach often leads to suboptimal solutions
- A local flipping of the current haplotype might improve the MEC



Haplotype to Haplotype

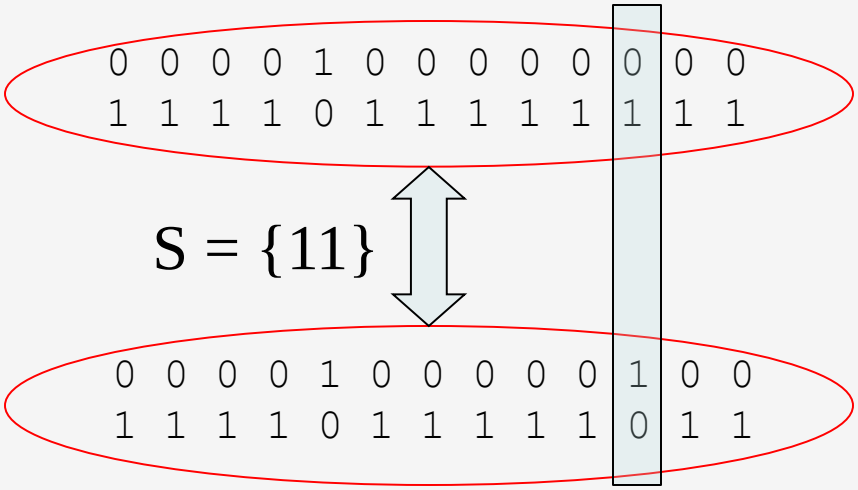
- The haplotype change also involves a reassignment of fragments.
- Here, the MEC error reduced to 2.
- This suggests a generic strategy
- Start with a haplotype, and move to a new one if it can improve the MEC



Haplotype to Haplotype

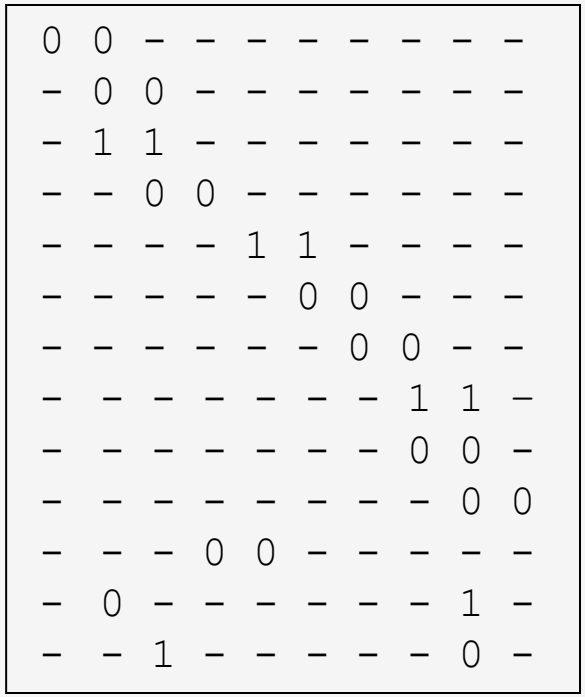
- A simple neighborhood is defined by flipping one column at a time (Ex: col. 11)

- difficult to get out of local minima using single flips
- "right" move cannot be chosen independently of the fragment matrix and the current solution



A difficult example

- $MEC(H_1) = 2$
- $MEC(H_2) = 1$
- Moving from H_1 to H_2 requires changing multiple columns
- Cannot be done using local greedy moves



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

H_1

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

H_2

Haplotype assembly: recap

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	–	–	–	–	–	–	–	–	–
1	1	1	–	–	–	–	–	–	–	–	–	–
–	–	0	0	1	–	–	–	–	–	–	–	–
–	–	0	–	–	0							
–	–	–	0	–	–	–	–	–	–	1	1	–
–	–	–	1	0	–	–	–	–	–	–	–	–
–	–	–	–	–	0	0	0	–	–	–	–	–
–	–	–	–	–	–	0	1	–	0	0	–	–
–	–	–	–	–	–	–	0	0	0	–	–	–
–	–	–	–	–	–	–	–	–	0	0	0	–
–	–	–	–	–	–	–	–	–	–	0	0	0
–	–	–	–	–	–	–	–	1	–	–	–	1

X

0	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**
1	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**0**

H

- Input = heterozygous sites and fragments (reads with alleles at sites)

$$\mathrm{MEC}(X_i, H) = \min\{d(X_i, h), d(X_i, \bar{h})\}$$

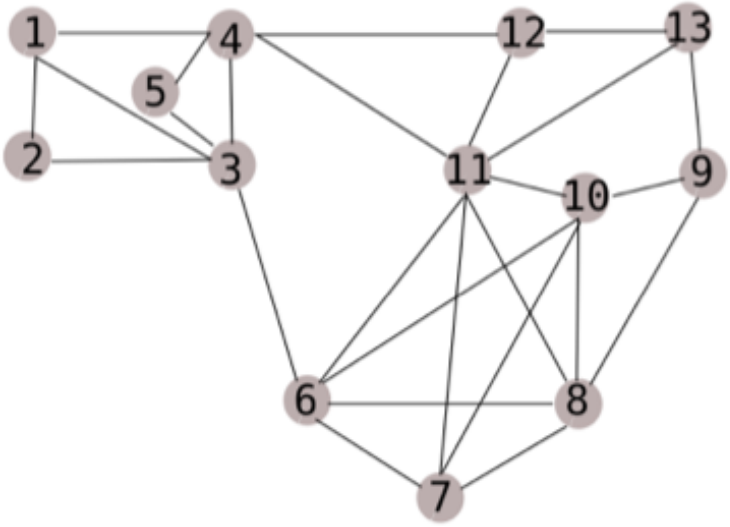
- $\mathrm{MEC}(X,H) = \Sigma\ \mathrm{MEC}(X_i,H)$
- Goal: reconstruct $H = (h,\bar{h})$ that minimizes MEC score
- NP-hard problem with exponential number of potential solutions

Haplotype assembly: approaches

- Greedy approach
  - Stuck in local optima
- Exact (Dynamic programming)
  - complexity is exponential in # of variants per read
- HapCUT
  - Greedy updates
  - use the graph structure of the fragment matrix to determine the updates

fragment-haplotype consistency graph

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1



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

H

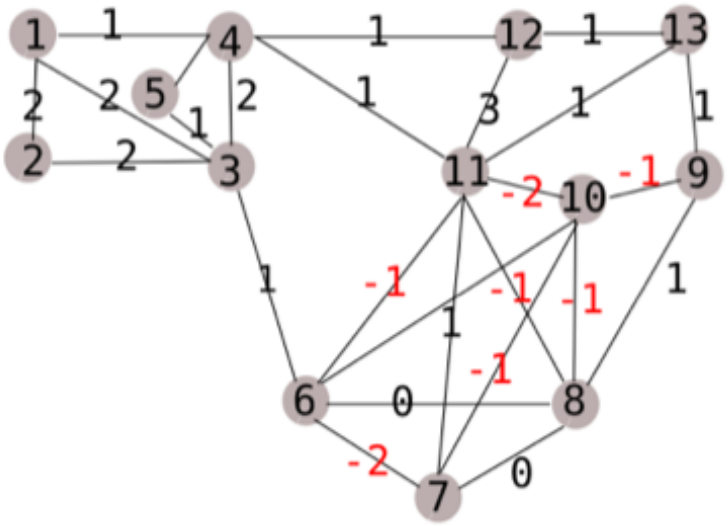
- Each column/variant is a node in the graph
- (x,y) is an edge if there is a fragment ‘touching’ columns x and y

Weighting graph edges

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1

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

H



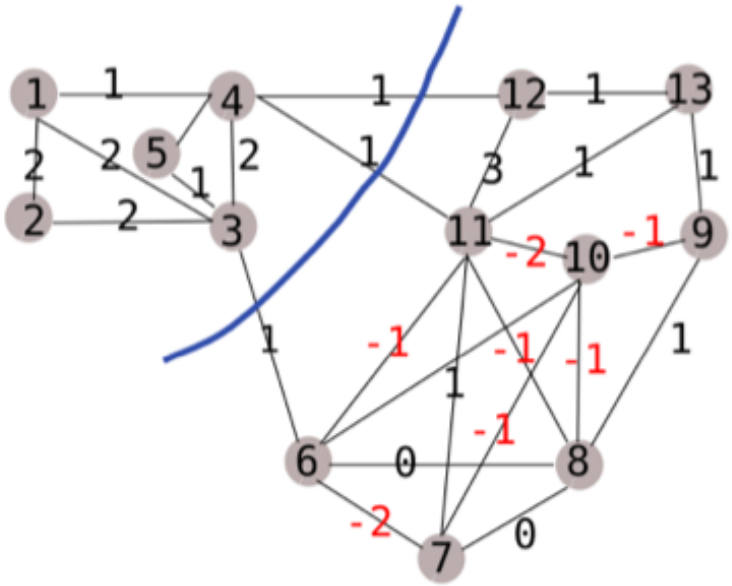
- w(x,y) = # fragments matching H – # fragments mismatching H

Cuts

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1

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

H



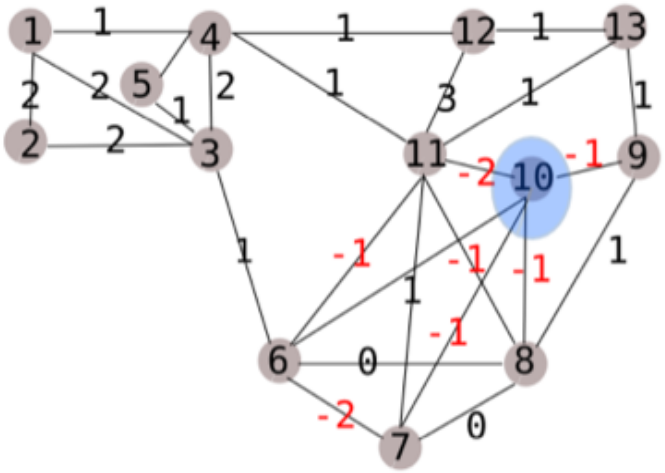
$S = \{ 1,2,3,4,5\}$

$W(S,\overline{S}) = 1 + 1 + 1 = 3$

- Cut ‘S’ is a bipartition of the vertices of the graph

Negative cuts can reduce MEC score

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1



0	0	0	0	1	0	1	1	1	0	1	1	1	H	
1	1	1	1	0	1	0	0	0	1	0	0	0		
0	0	0	0	1	0	1	1	1	1	1	1	1	H_S	
1	1	1	1	0	1	0	0	0	0	0	0	0		

$W(S,S) = -2 -1 -1 -1 = -5$

$MEC(H_s) - MEC(H) = -3$

if $W(S, \overline{S}) < 0$, $\quad MEC(H_S) < MEC(H)$

HapCUT algorithm

Initialization: Choose an initial haplotype configuration H^1 randomly.

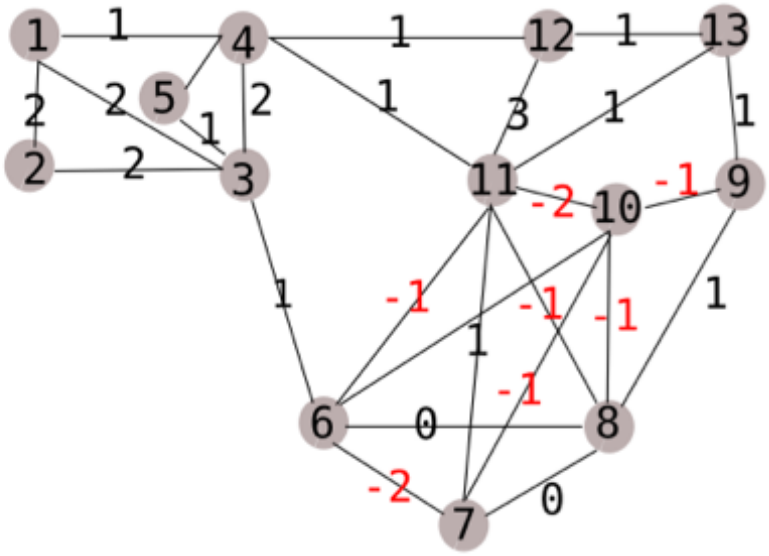
Iteration: For $t = 1, 2, \ldots$

1. Construct the read-haplotype consistency graph $G(H^t)$
2. Compute a cut $(S, \overline{S})$ in $G(H^t)$ such that $w(S, \overline{S}) < 0$
3. If $\text{MEC}(H^t_S) \leq \text{MEC}(H^t)$, $H^{t+1} = H^t_S$
4. Else $H^{t+1} = H^t$

Example

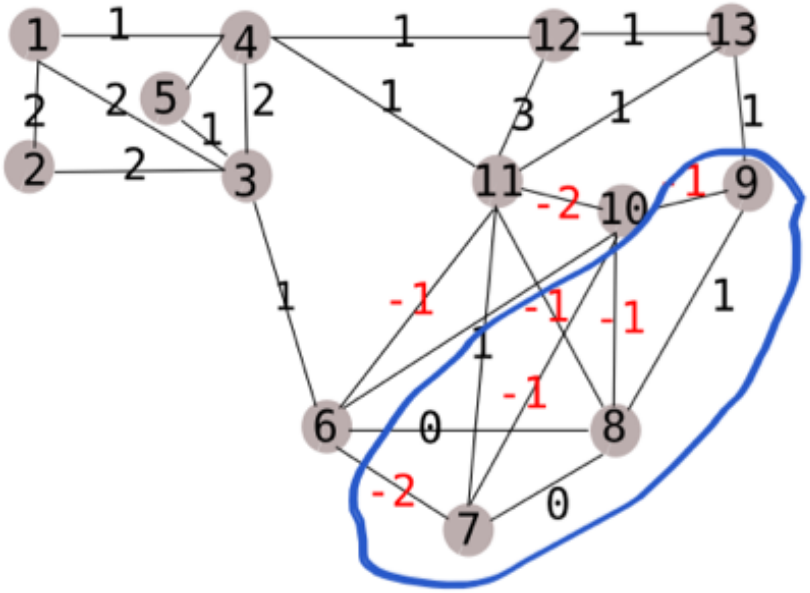
1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1

0	0	0	0	1	0	1	1	1	0	1	1	1	H
1	1	1	1	0	1	0	0	0	1	0	0	0	



First Cut

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1



0	0	0	0	1	0	1	1	1	0	1	1	1	H
1	1	1	1	0	1	0	0	0	1	0	0	0	

$W(S,\overline{S}) = -2 + (-3) + 0 = -5$

0	0	0	0	1	0	0	0	0	0	1	1	1	H_S
1	1	1	1	0	1	1	1	1	1	0	0	0	

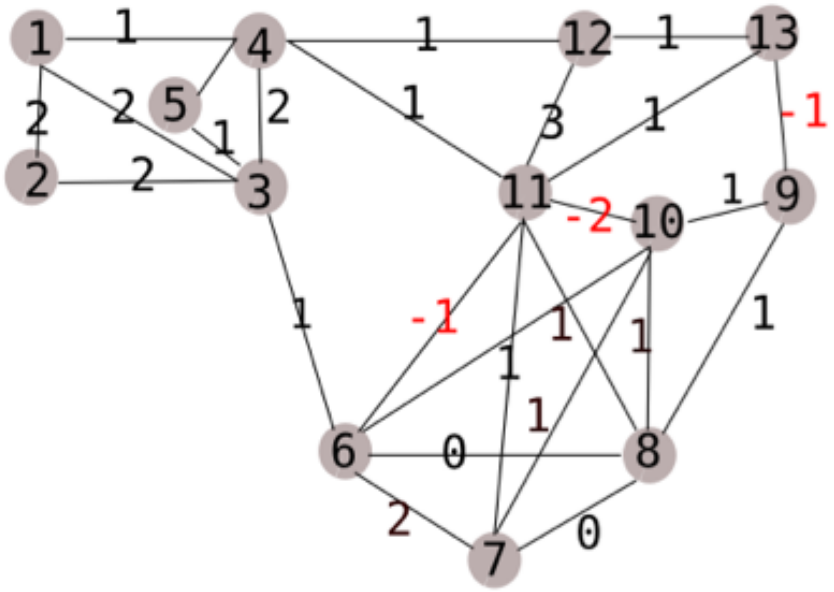
$MEC(H_S) - MEC(H) = -3$

Update edge weights

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1

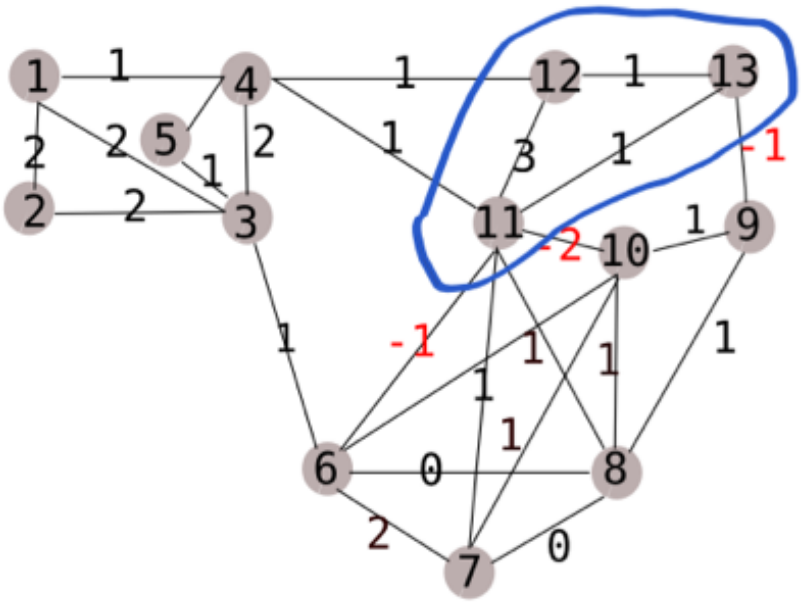
0	0	0	0	1	0	0	0	0	0	1	1	1
1	1	1	1	0	1	1	1	1	1	0	0	0

H



Second Cut

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1



$S = \{11,12,13\}$

0	0	0	0	1	0	0	0	0	0	1	1	1	H
1	1	1	1	0	1	1	1	1	1	0	0	0	

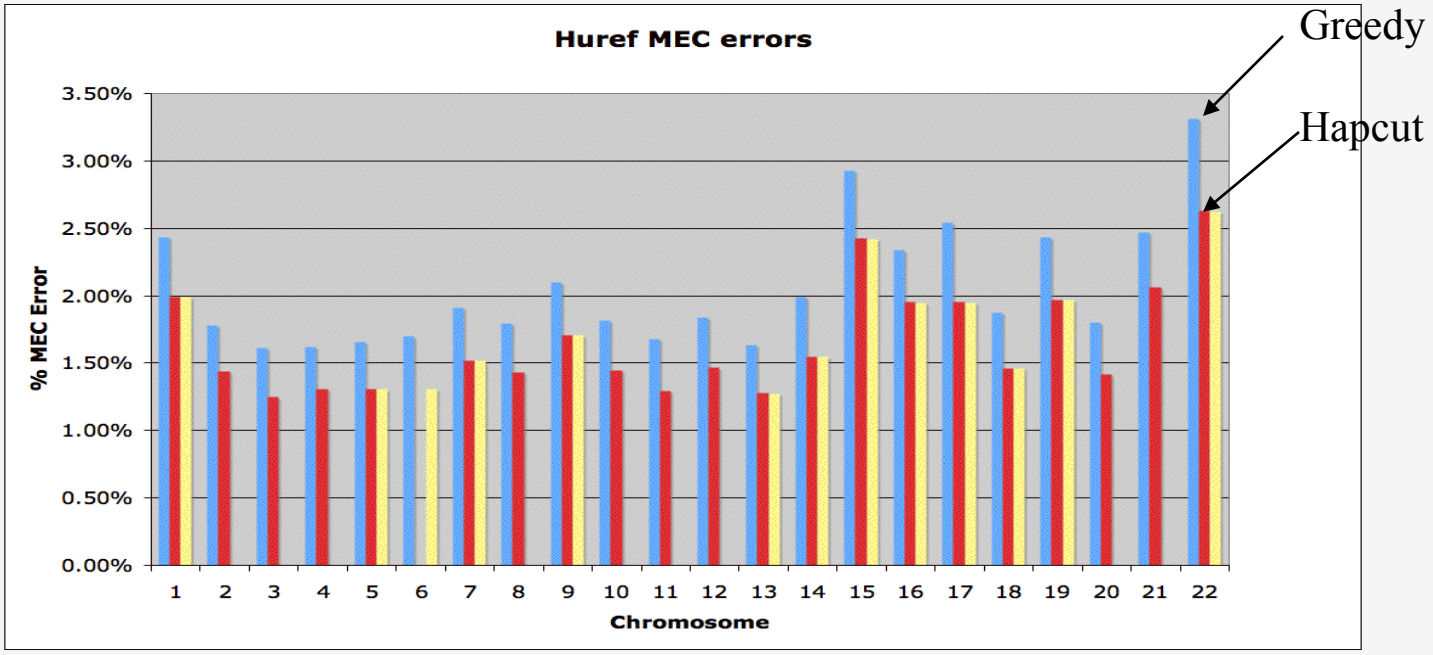
$W(S,\overline{S}) = -2 + 1 + (-1) = -2$

0	0	0	0	1	0	0	0	0	0	0	0	0	H_S
1	1	1	1	0	1	1	1	1	1	1	1	1	

$MEC(H_s) - MEC(H) = -2$

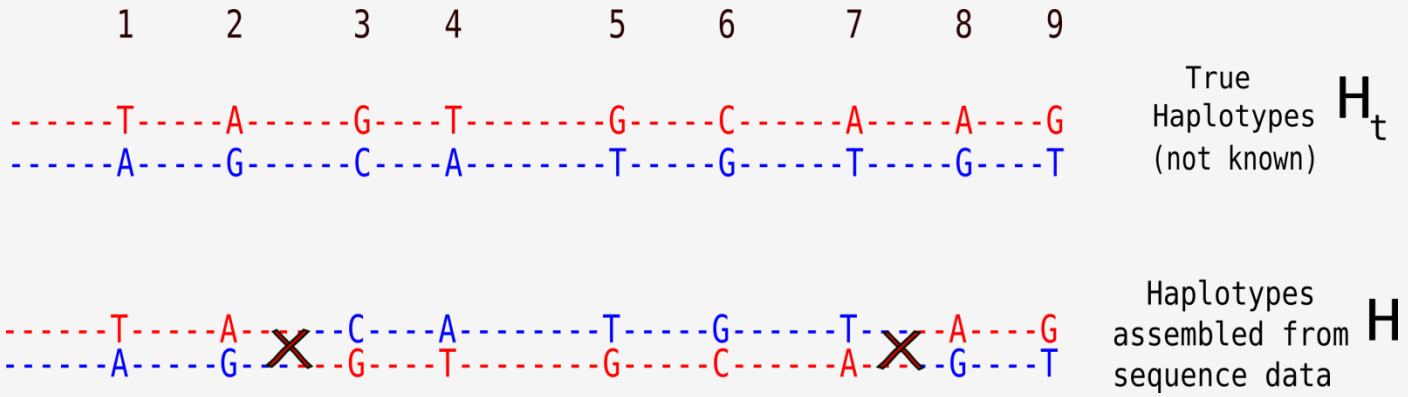
Haplotype assembly for Huref

- Craig Venter's genome sequenced using Sanger sequencing (PloS Biology 2007)
- Chromosome 22 statistics
  - Fragment matrix = 25K (sites) x 53K ('useful' fragments)
  - N50 haplotype length=350Kb



Switch error rate of haplotypes

- MEC error rate measures consistency of haplotypes with the sequenced fragments/reads
- Switch error rate measures absolute accuracy of haplotypes



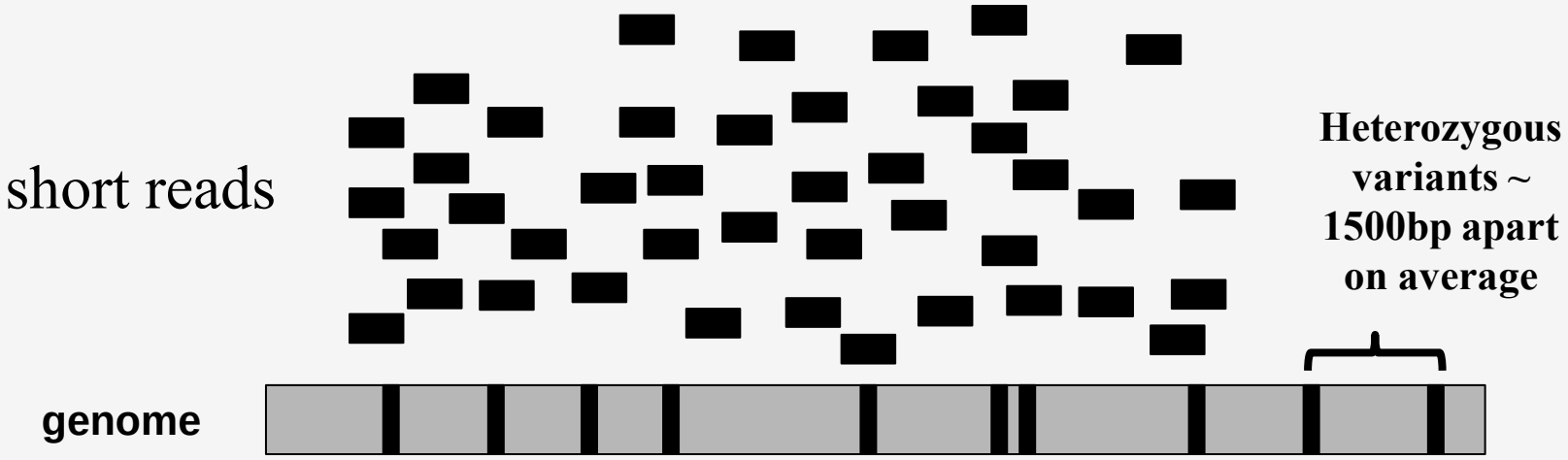
Switch error rate = 2/8 = 0.25

Three components of haplotype assembly

- Accuracy
  - solved using computational methods (HapCUT)
- Length of assembled haplotypes
  - Determined by sequencing technology and rate of heterozygosity
- Variant calls
  - Accurate set of heterozygous variants

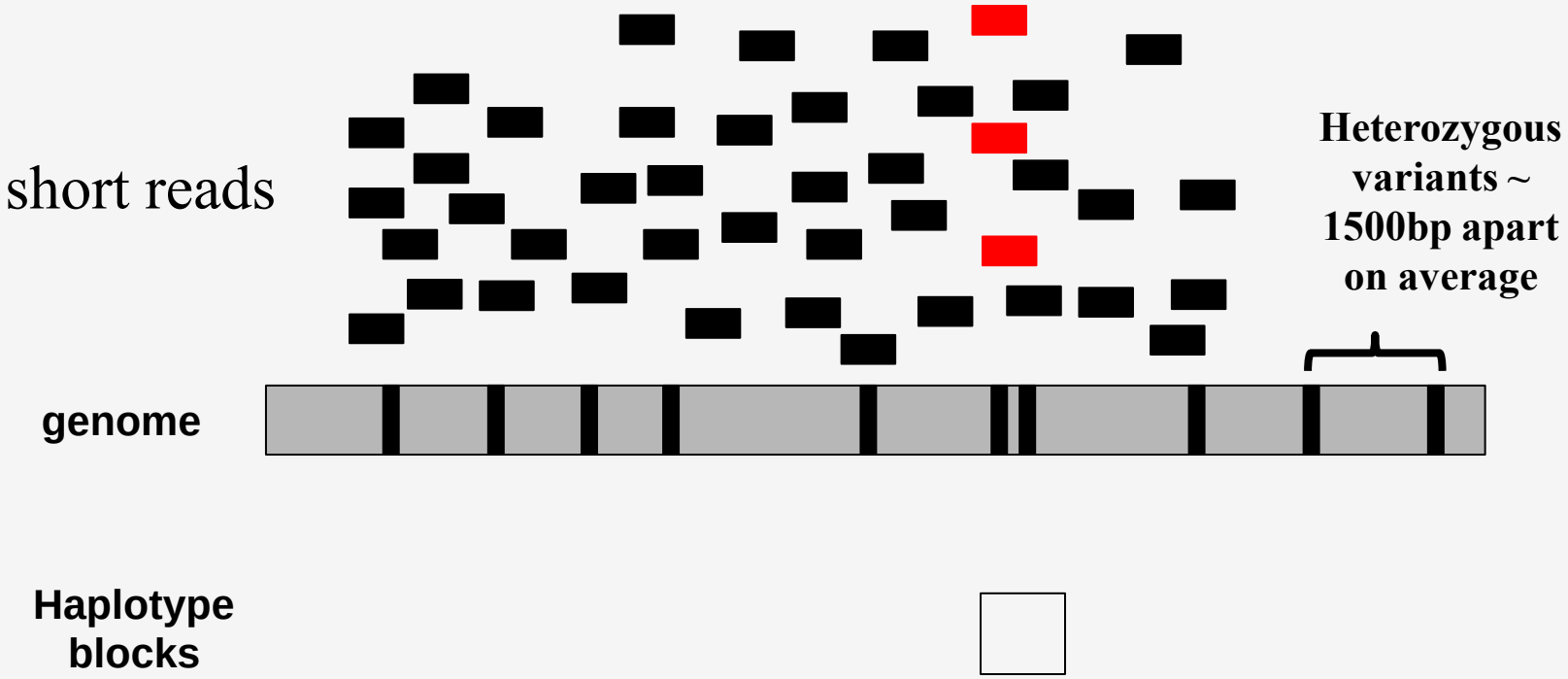
Can we do haplotype assembly with Illumina sequencing?

A read is only useful for phasing (**haplotype informative)** if it covers 2 or more heterozygous variants



Illumina sequencing results in very short haplotype blocks

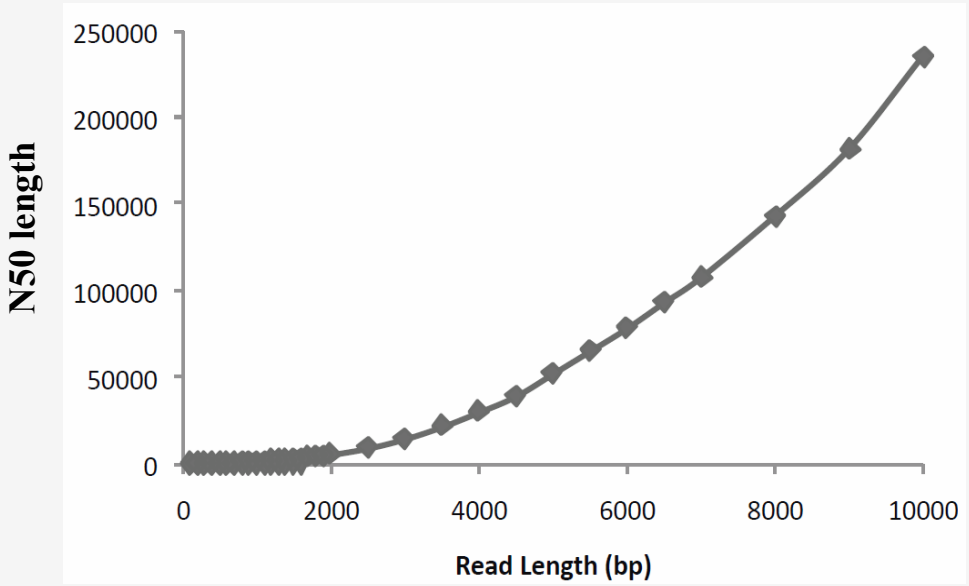
A read is only useful for phasing (**haplotype informative)** if it covers 2 or more heterozygous variants

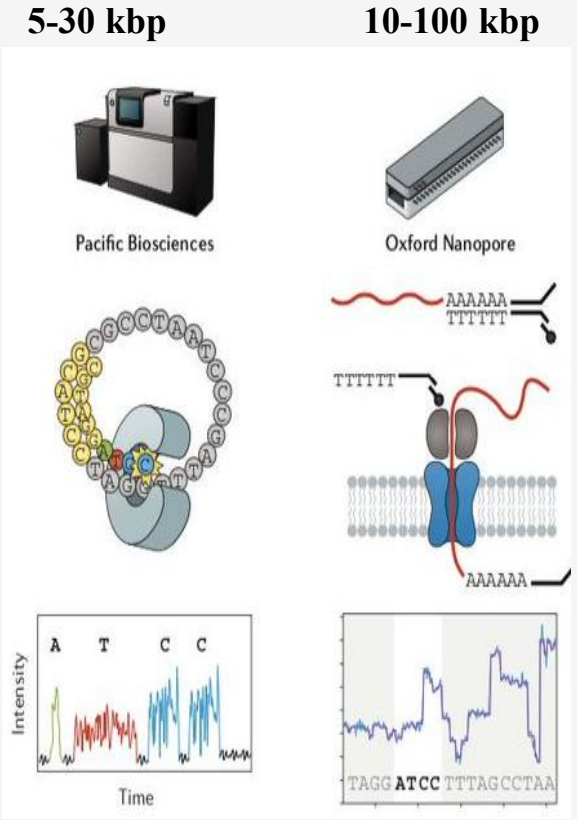


Long reads needed for haplotype assembly

A read is only useful for phasing (**haplotype informative)** if it covers 2 or more heterozygous variants

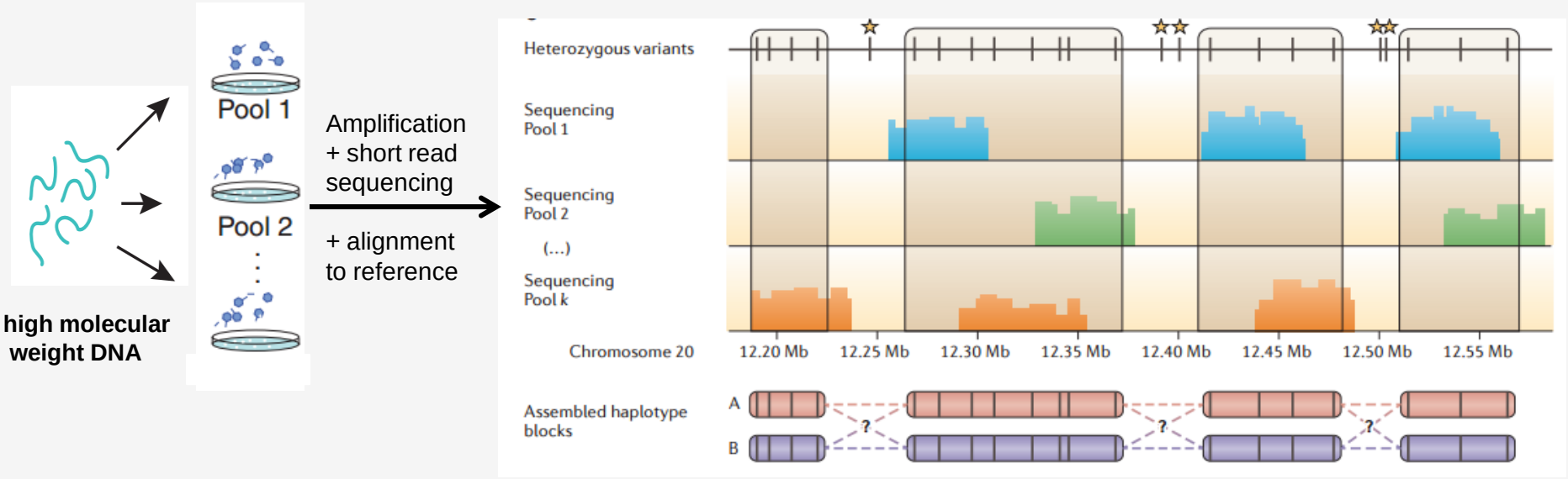
Long reads needed for haplotype assembly





- Pacific Biosciences sequencing costs $5000-10,000 per human genome
- Nanopore cost is lower but high error rates (8-15%)

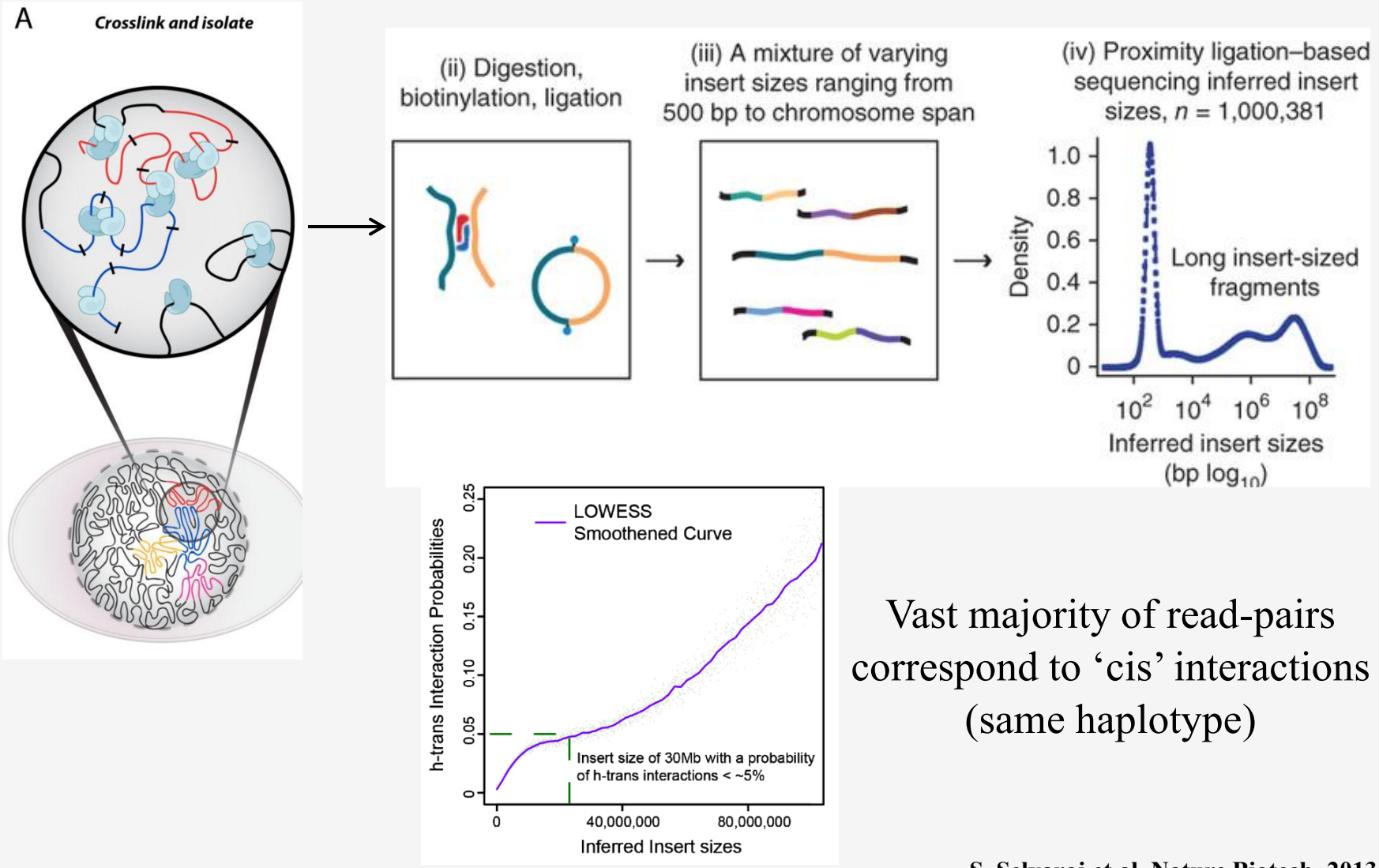
Virtual long reads from short Illumina reads



- long-range haplotype information maintained in short reads
- haplotypes assembled using algorithms such as HapCUT

Snyder et al. Nat. Gen. 2015

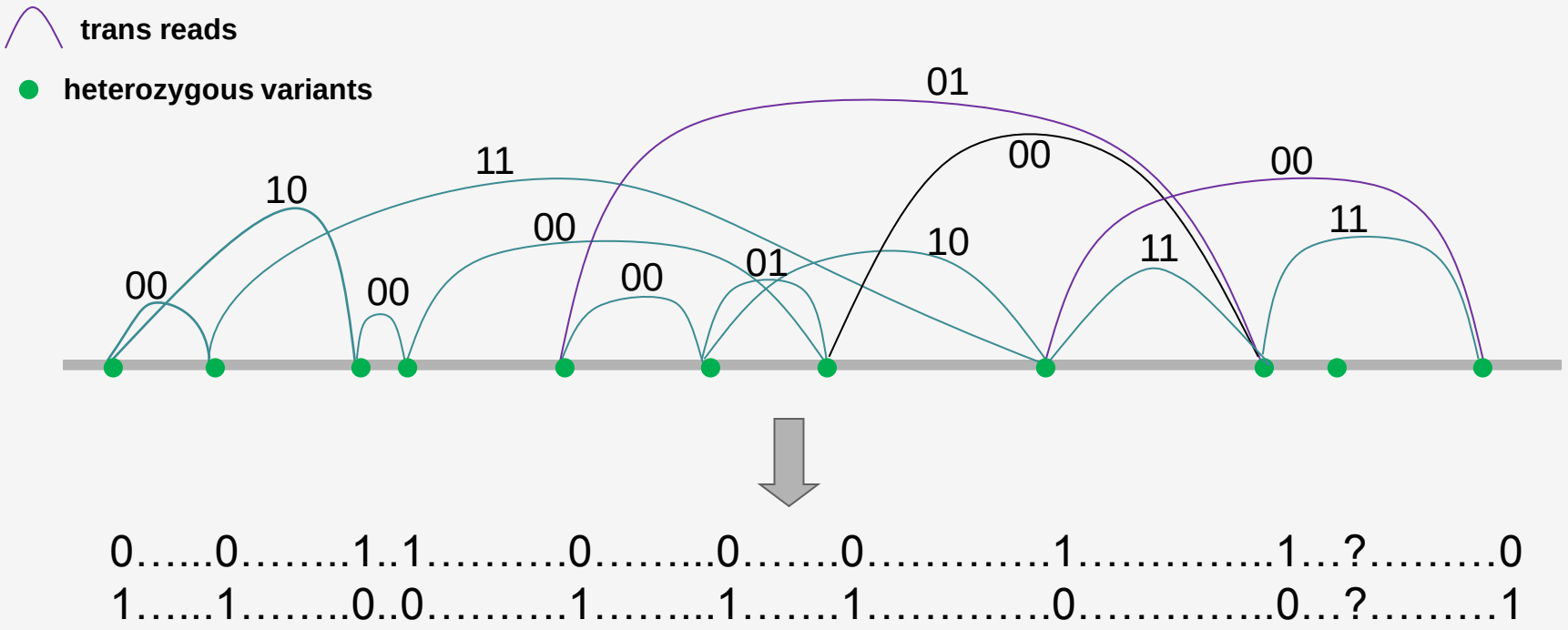
Hi-C sequencing generates read-pairs that link distal DNA fragments



Vast majority of read-pairs correspond to ‘cis’ interactions (same haplotype)

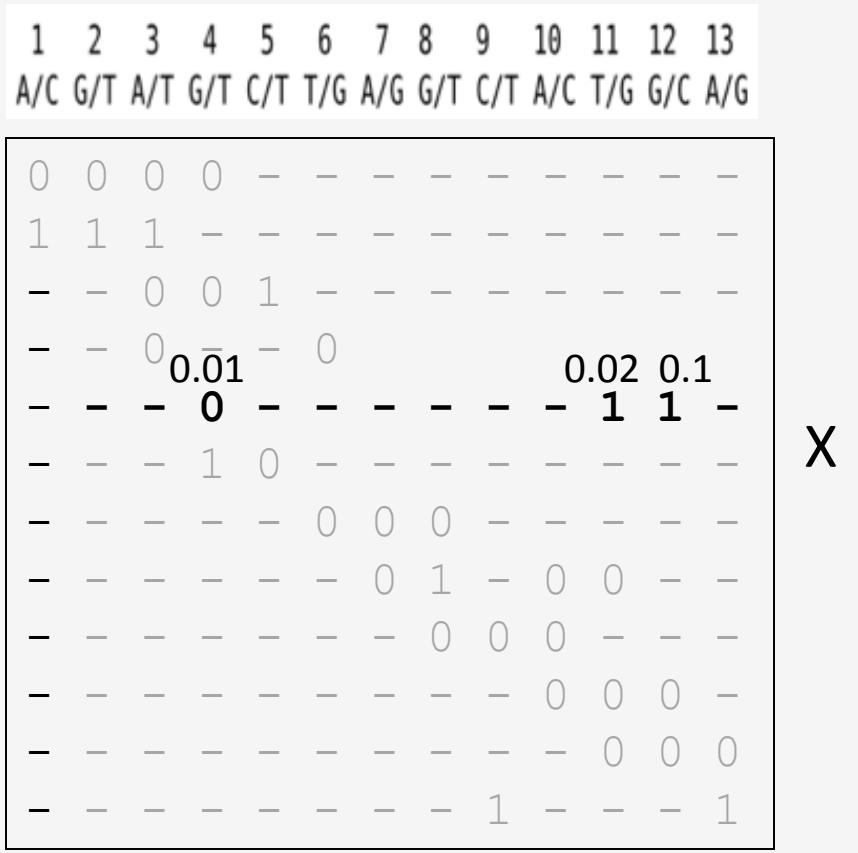
S. Selvaraj et al. Nature Biotech. 2013

Haplotype assembly using Hi-C



- Average error rate of 2-3% due to trans-interactions
- HapCUT is well-suited for this data
- Chromosomal-spanning haplotypes can be assembled from 18x whole-genome Hi-C data (NA12878) with 98% accuracy

Haplotype assembly: likelihood formulation



X

0 0 0 0 1 0 1 1 1 0 1 1 1 h_1

1 1 1 1 0 1 0 0 0 1 0 0 0 h_2

- Probability of error (q_{ij}) estimated from sequencing quality values
- Unknown variable H = (h_1, h_2)
- $P(X_5 \mid h_1, q) = (1-.01)\,(1-.02)(1-.1) = 0.89$
- $P(X_5 \mid h_2, q) = (.01)\,(.02)(.1) = 0.00002$
- $P(X_5 \mid H=(h_1, h_2), q) = (0.89+0.00002)/2 = 0.445$

Haplotype assembly: likelihood formulation

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	–	–	–	–	–	–	–	–	–
1	1	1	–	–	–	–	–	–	–	–	–	–
–	–	0	0	1	–	–	–	–	–	–	–	–
–	–	0	–	–	0							
–	–	–	0	–	–	–	–	–	–	1	1	–
–	–	–	1	0	–	–	–	–	–	–	–	–
–	–	–	–	–	0	0	0	–	–	–	–	–
–	–	–	–	–	–	0	1	–	0	0	–	–
–	–	–	–	–	–	–	0	0	0	–	–	–
–	–	–	–	–	–	–	–	–	0	0	0	–
–	–	–	–	–	–	–	–	–	–	0	0	0
–	–	–	–	–	–	–	–	1	–	–	–	1

X

0	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	h_1
1	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	h_2

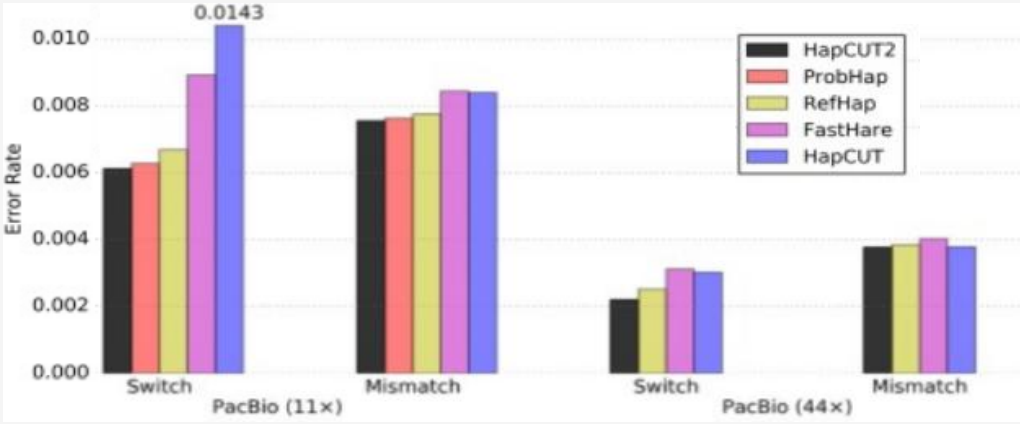
- We can compute the likelihood of X (fragment matrix) given H, q

$$\Pr(X \mid H, q) = \tilde{\prod}_i \Pr(X_i \mid H, q)$$

- The goal is to find H that maximizes likelihood (instead of minimizing MEC score)

Hapcut2: accurate haplotypes using noisy long reads

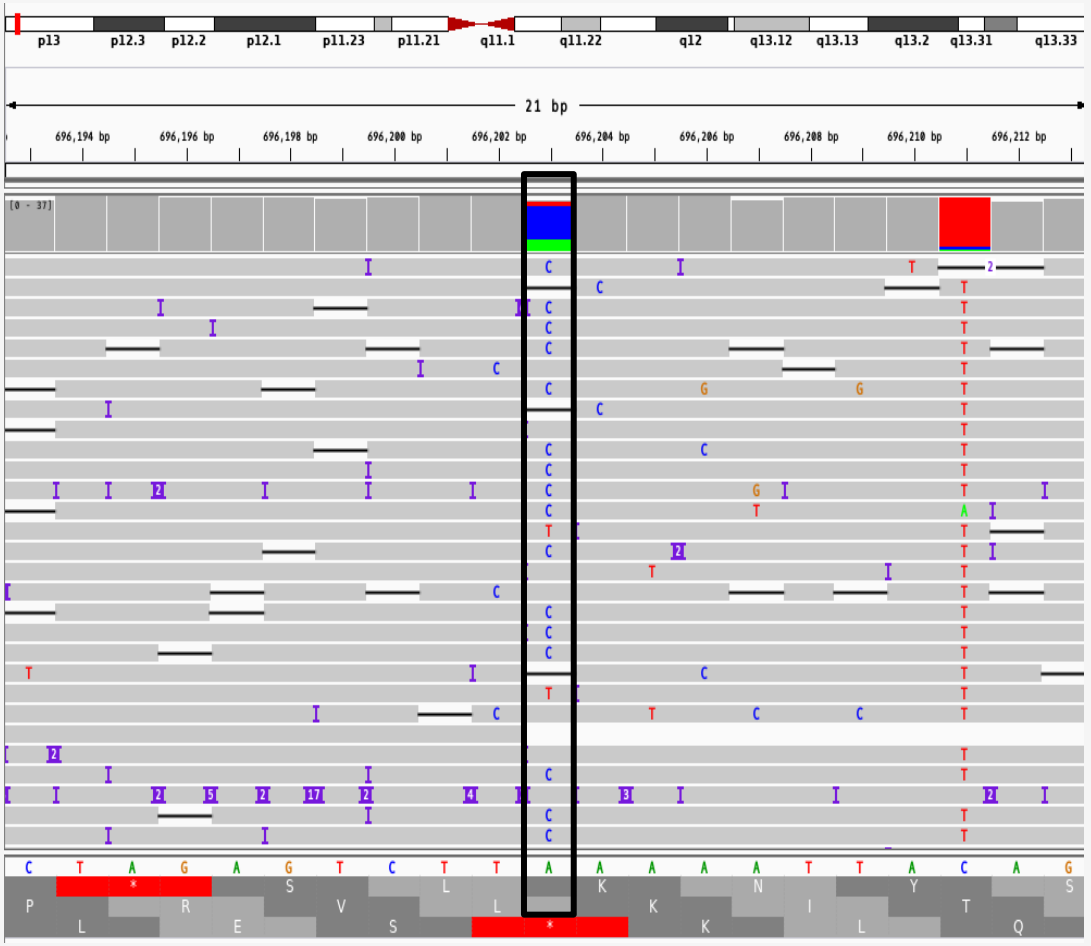
1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0							
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	-	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	-
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1



- Low switch error rate (< 0.1%) for human genomes using PacBio reads with 8-11% error rate
- Need variants called using Illumina sequencing as input for phasing

Variant calling is difficult with PacBio reads

- For PacBio, local alignments are highly unreliable, and quality values are not meaningful
- Variant calling methods developed for Illumina sequencing have low accuracy



Haplotype information can inform variant calling

A/T	G/C	T/C
–	–	0 0 0 – –
–	–	0 1 0 – –
–	–	0 0 0 – –
–	–	0 0 0 – –
–	–	1 0 1 – –
–	–	1 1 1 – –
–	–	1 0 1 – –

– – 0 0 0 – – H_1
– – 1 1 1 – –

– – 0 1 0 – – H_2
– – 1 0 1 – –

- Three heterozygous sites covered by 7 reads
- Likelihood calculation
  - $p(X| H_1, q=0.1) = 2.6 \times 10^{-5}$
  - $p(X| H_2, q=0.1) = 2.6 \times 10^{-5}$

Haplotype information can inform variant calling

A/T	G/C	T/C				
–	–	0	0	0	–	–
–	–	0	1	0	–	–
–	–	0	0	0	–	–
–	–	0	0	0	–	–
–	–	1	0	1	–	–
–	–	1	1	1	–	–
–	–	1	0	1	–	–

–	–	0	0	0	–	–	H_1
–	–	1	1	1	–	–	
–	–	0	1	0	–	–	H_2
–	–	1	0	1	–	–	
–	–	0	0	0	–	–	H_3
–	–	1	0	1	–	–	

- Three heterozygous sites covered by 7 reads
- Likelihood calculation
  - $p(X| H_1, q=0.1) = 2.6 \times 10^{-5}$
  - $p(X| H_2, q=0.1) = 2.6 \times 10^{-5}$
  - $p(X| H_3, q=0.1) = 1.7 \times 10^{-3}$

Generalized haplotype assembly problem

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	–	–	–	–	–	–	–	–	–
1	1	1	–	–	–	–	–	–	–	–	–	–
–	–	0	0	1	–	–	–	–	–	–	–	–
–	–	0	–	–	0							
–	–	–	0	–	–	–	–	–	–	1	1	–
–	–	–	1	0	–	–	–	–	–	–	–	–
–	–	–	–	–	0	0	0	–	–	–	–	–
–	–	–	–	–	–	0	1	–	0	0	–	–
–	–	–	–	–	–	–	0	0	0	–	–	–
–	–	–	–	–	–	–	–	–	0	0	0	–
–	–	–	–	–	–	–	–	–	–	0	0	0
–	–	–	–	–	–	–	–	1	–	–	–	1

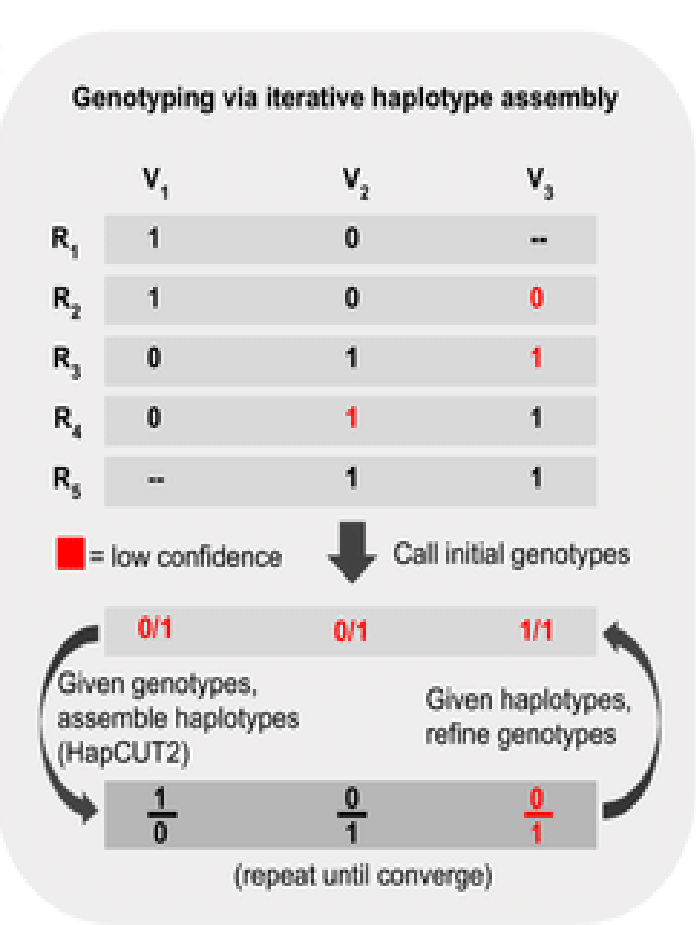
0 0 0 0 1 0 1 1 0 0 1 1 1 h_1

1 1 1 0 0 1 0 0 1 0 0 0 0 h_2

- constrained version of haplotype assembly
  - $h_1[i] + h_2[i] = 1$
- Generalized version: sites can be homozygous (0 or 1 for both alleles)
- More difficult to solve (space of haplotypes increased)

Haplotype-informed variant calling using long reads

1. Identify candidate set of variants (V) using local alignment information
2. Generate fragment matrix for V using aligned reads and assign genotypes
3. Repeat
   - I. Assemble haplotypes using HapCUT2 for heterozygous variants
   - II. Update genotype for each variant conditional on genotypes for other variants



- Achieves high accuracy (precision = 0.997 and recall=0.97) for SNV calling using Pacific Biosciences reads (Nat. Comm. 2019)