

Molecular Overview

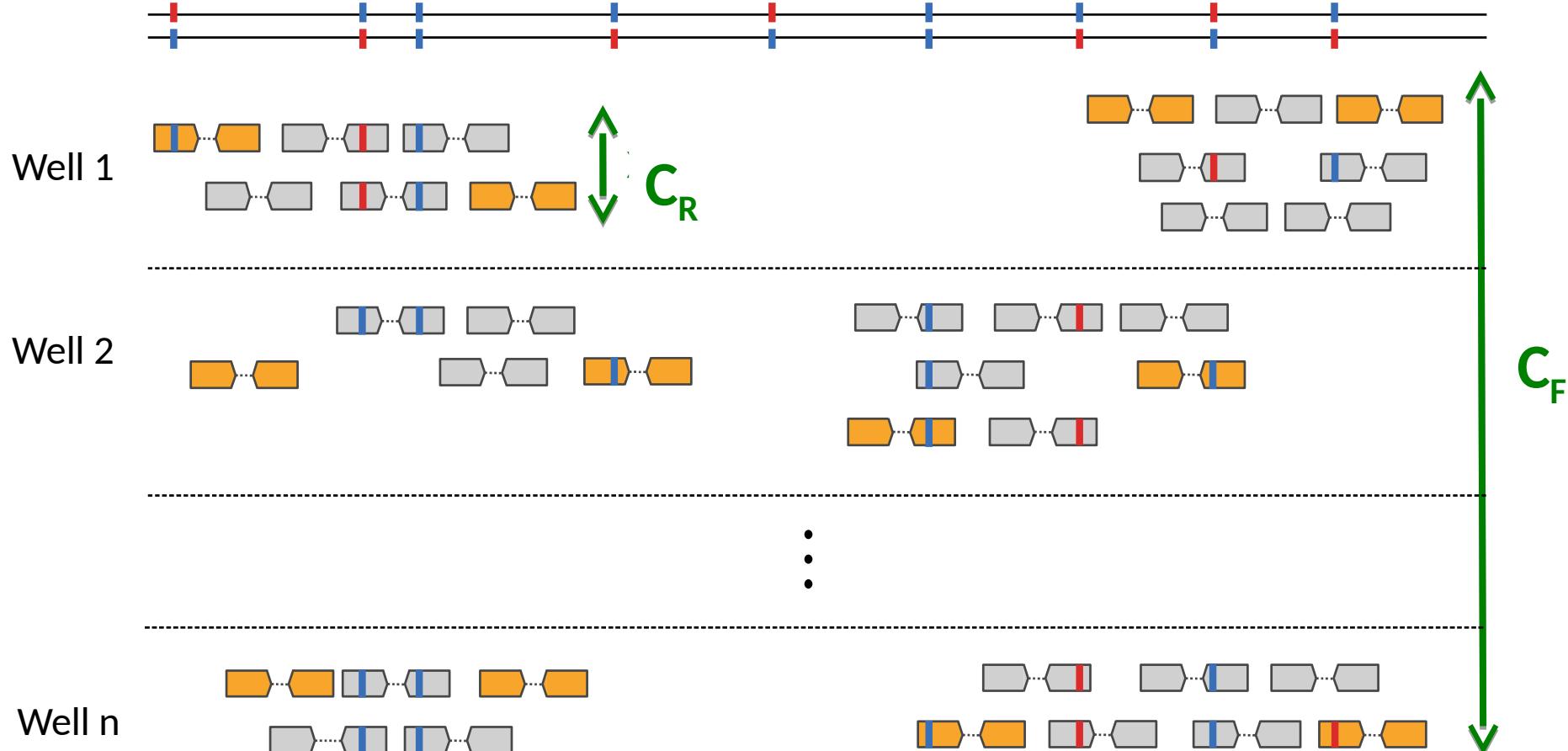
1. Sample DNA is sheared into fragments of about 10 kbp

2. Fragments are diluted and placed into 384 wells

3. Fragments are amplified through long-range PCR, cut into short fragments and barcoded

4. Short fragments are pooled together and sequenced

Read Clouds



$$\text{Coverage} = C_F C_R$$

- Normal (FFPE)
- Sequenced by:
 - Shotgun (40X)

- IDC (Fresh Frozen)
- Sequenced by:
 - Shotgun (40X)
 - Moleculo (78X)

$$C_F = 43, C_R = 1.8$$

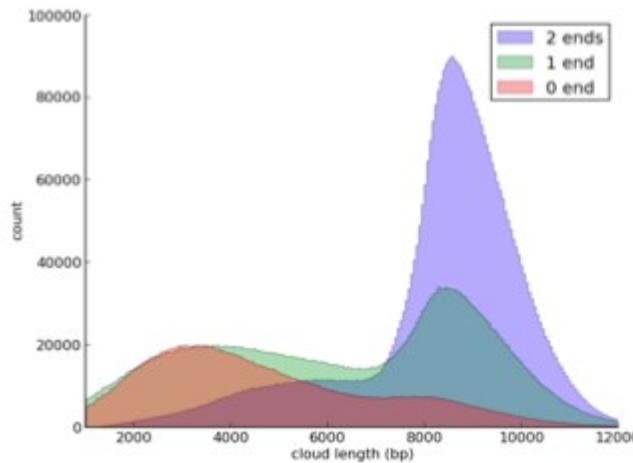
Align reads

Create clouds

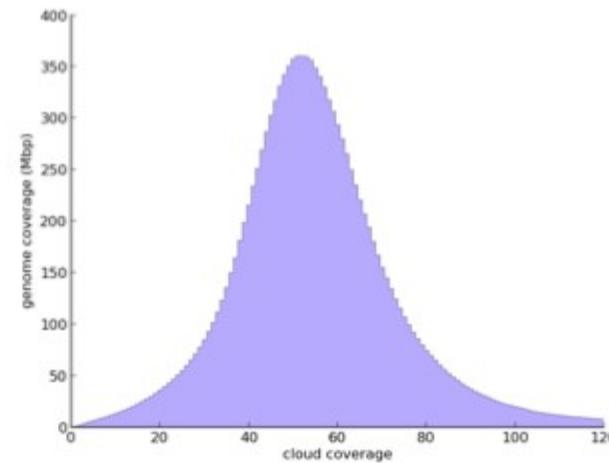
Realign reads

Call Variants (GATK)

Phase



(a) Read cloud sizes



(b) Read Cloud Genome Coverage

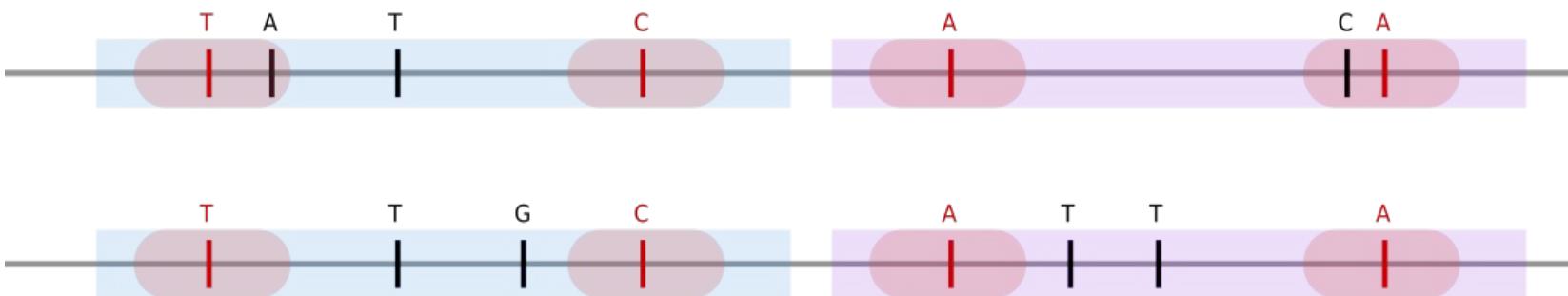
Identifying Variation in Segmental Duplications

~180 Mbp of human in almost exact repeats

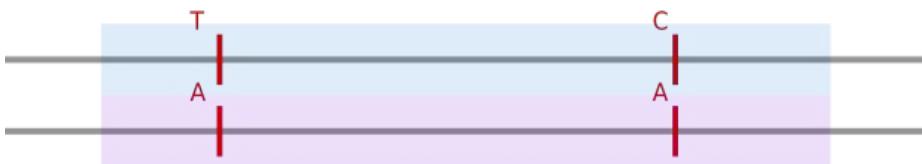
Novel variation impossible to detect with short reads alone

Single unique nucleotides (SUNs) in segdups can be leveraged by read clouds

Sequence Comparison

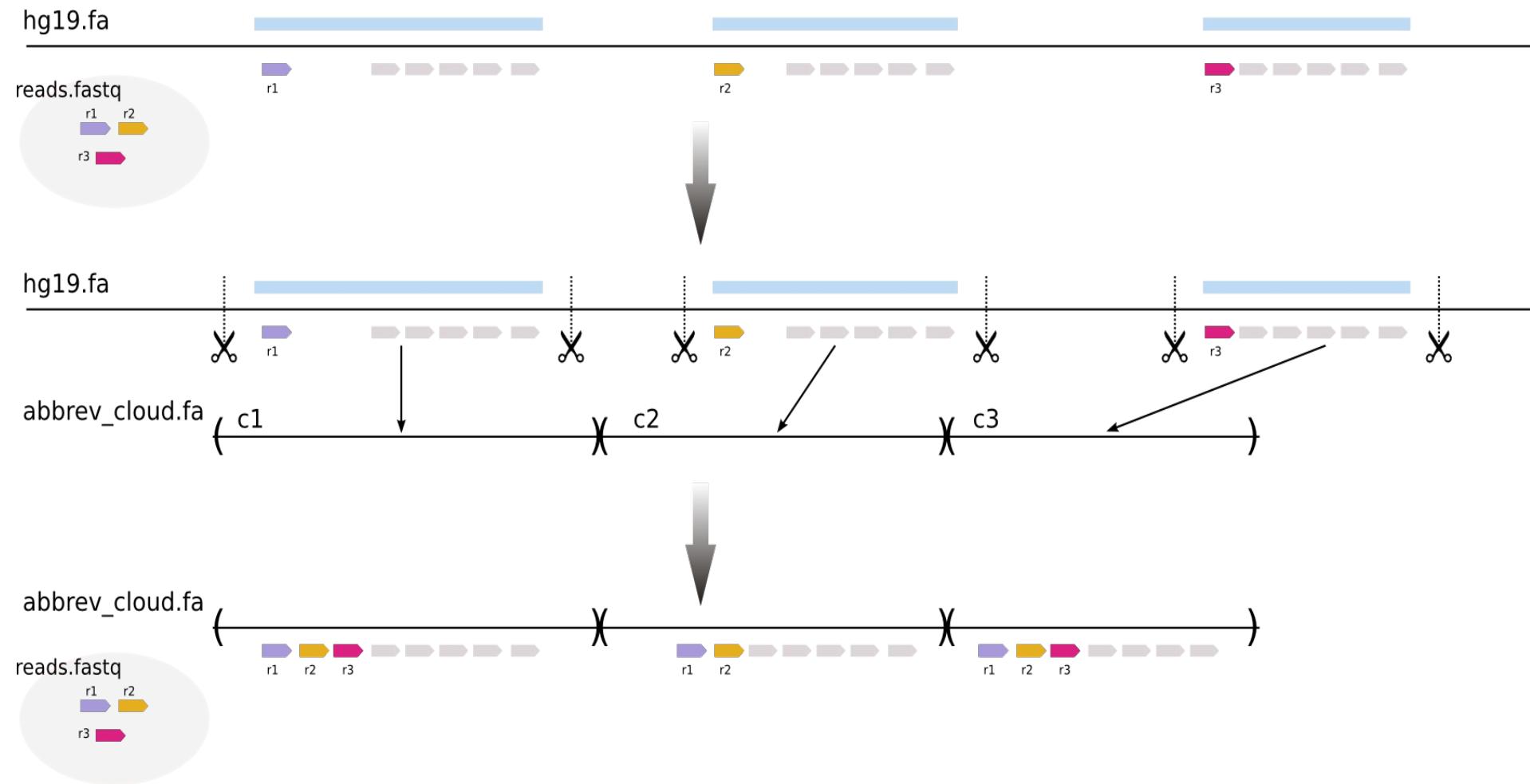


MSA reveals SUNs



Candidate Cloud and Alignment Generation

- Use Bowtie2 to align each well separately to the reference



MRF-based Realignment

Generative model of read cloud generation of set of reads R

$$P(R) = \sum_{\{\text{Underlying molecule set } M\}} P(M) P(R | M)$$

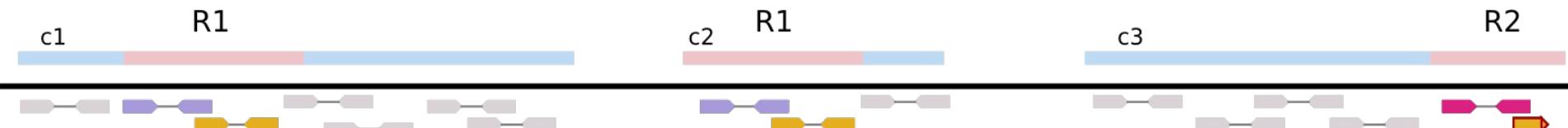
Markov Random Field (MRF)

- Read alignment quality
- Mate pairs biased to map together
- Reads biased to form clouds

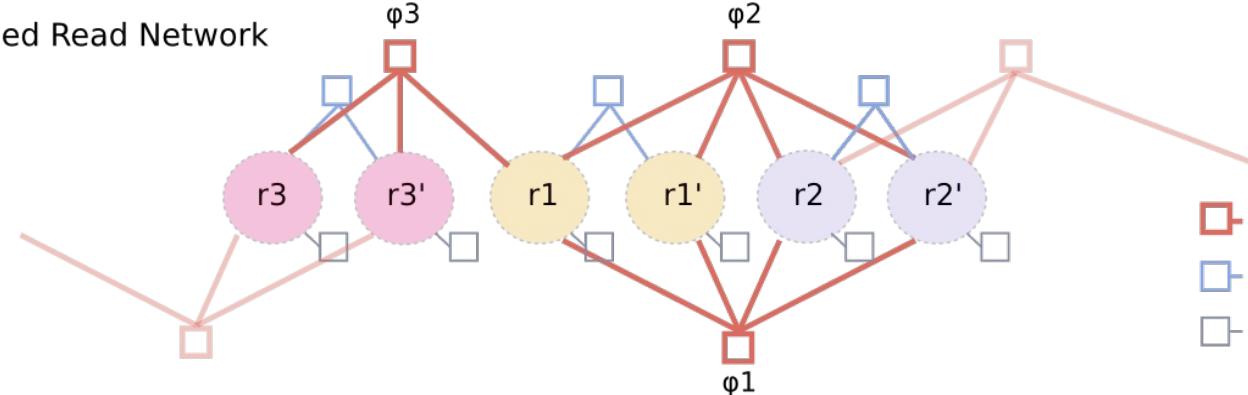
Optimize with Simulated Annealing

- Move a read
- Move a pair of reads
- Move a cloud of reads

Candidate Clouds

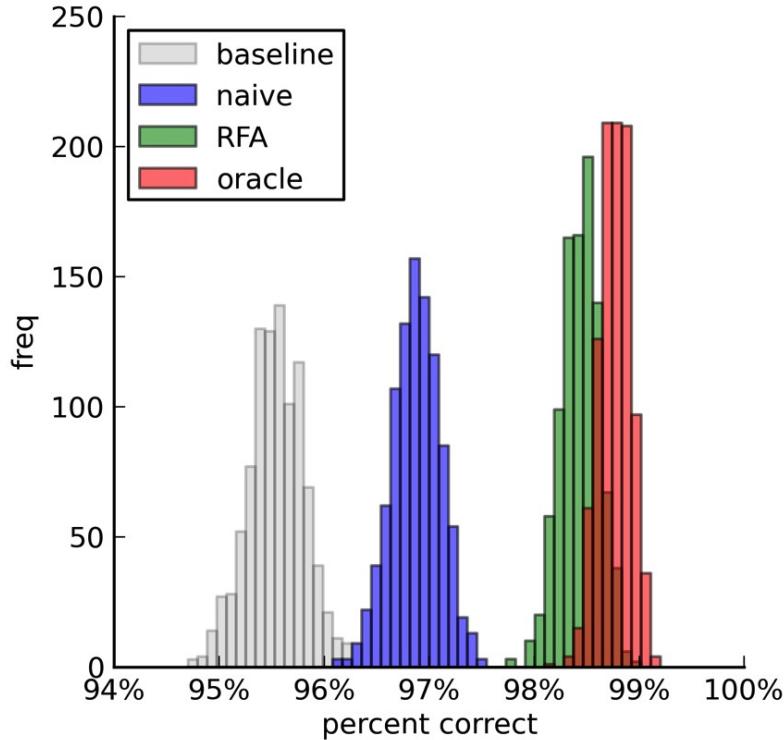


Induced Read Network



- = cloud potential
- = mate pair potential
- = read potential

Results of 1000 simulated read cloud wells



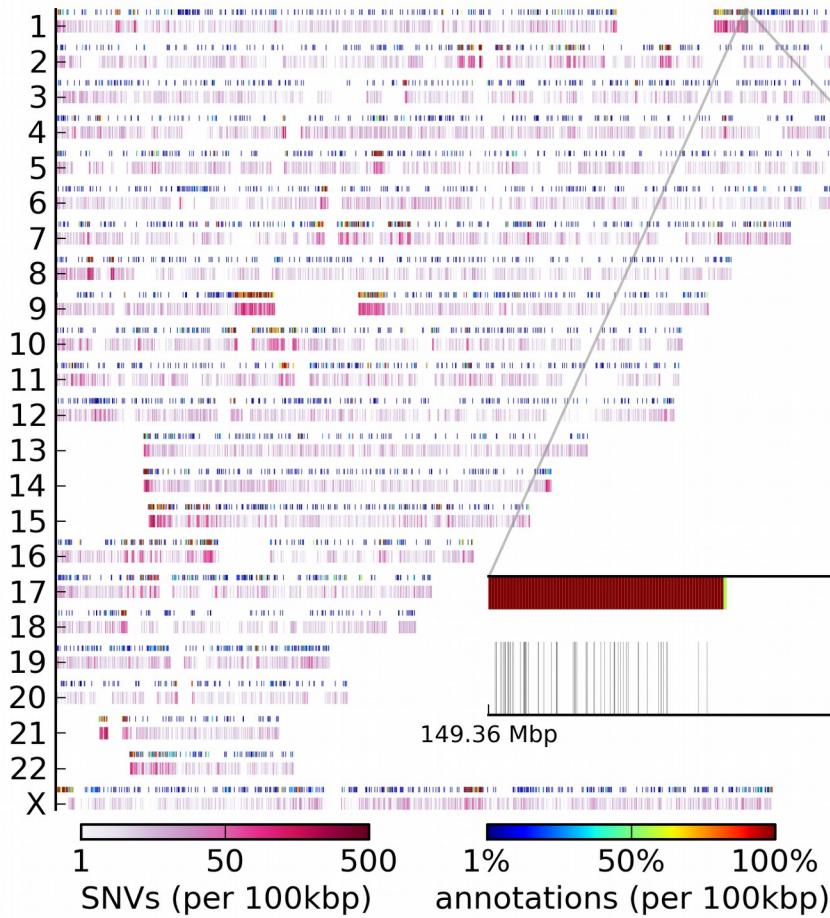
Element Class	frequency (%)	illuminated (%)
all	100.0	90.6
annotated	88.4	90.5
segdup	43.4	93.8
LINE	35.2	88.3
SINE	14.2	92.7
gene	7.0	95.5
LTR	6.3	92.3
Simple repeat	4.6	85.9
Satellite	2.3	88.1
Low complexity	1.6	89.0

Baseline: Bowtie2 baseline

Naive: Naive policy of using alignments to abbreviated reference

Oracle: Pick the true alignment among Bowtie2 alignments

Results on IDC Sample



Element Class	recovered SNVs (hetero)	unique SNVs (hetero)
all	85997 (44503)	19408 (12747)
annotated	82887 (42625)	19033 (12552)
segdup	28661 (16970)	16421 (11168)
LINE	35461 (17054)	2961 (1983)
SINE	20394 (9945)	3351 (2075)
gene	4972 (2915)	2592 (1869)
LTR	6693 (3582)	2018 (1289)
Simple repeat	2272 (1271)	1376 (755)
Satellite	2081 (1270)	388 (238)
Low complexity	387 (231)	187 (123)

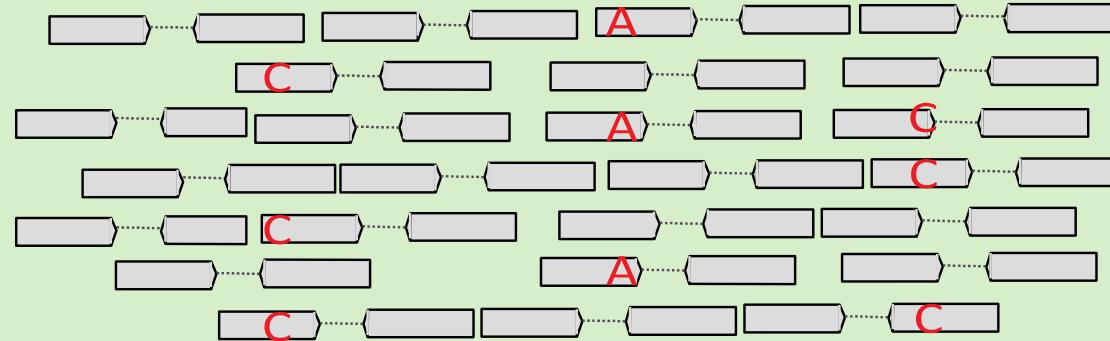
Multiplex PCR
346 SNVs, 94% validated

Phasing



Dorna Kashef

Sequence reads



Reference Genome

ATTACGAAAATTACGAACATACCATGGACACCTCG

Genotype

A/C

A/T

A/C

Haplotypes

A

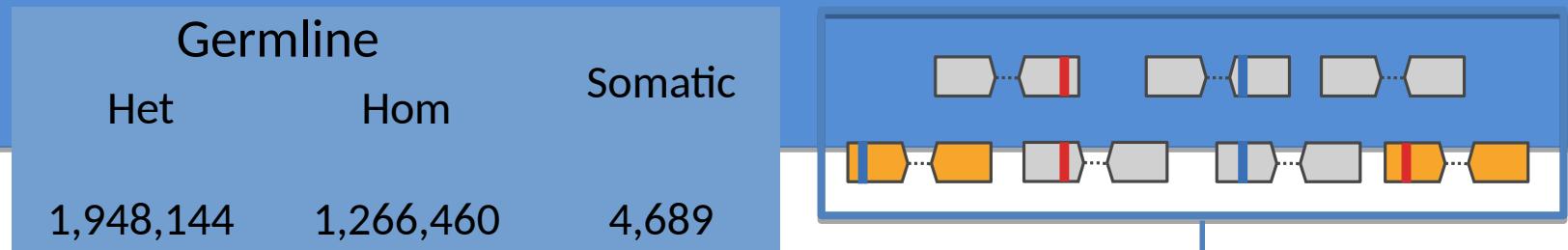
A

C

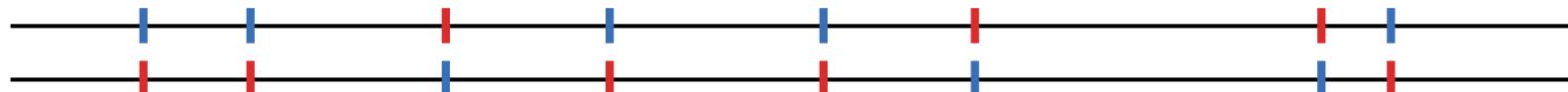
C

T

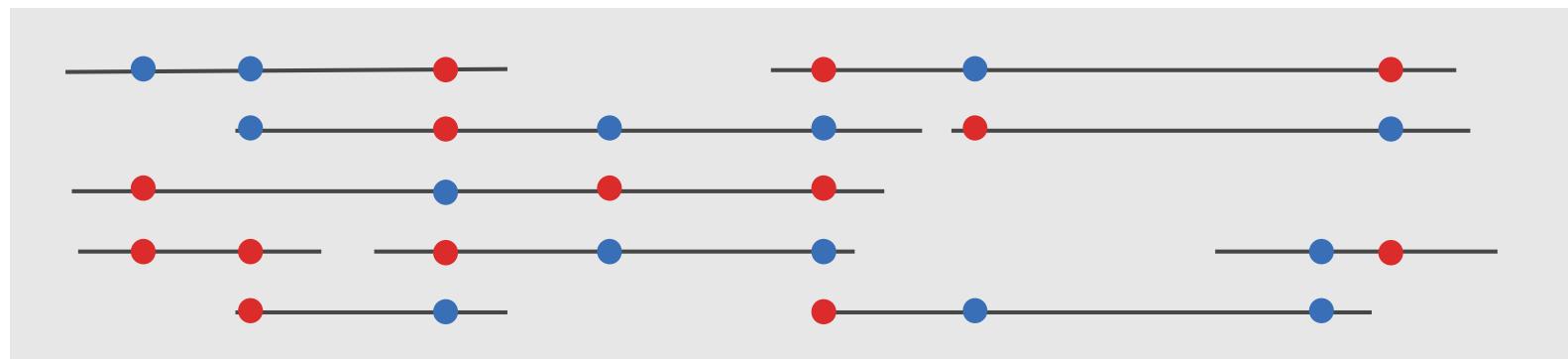
A



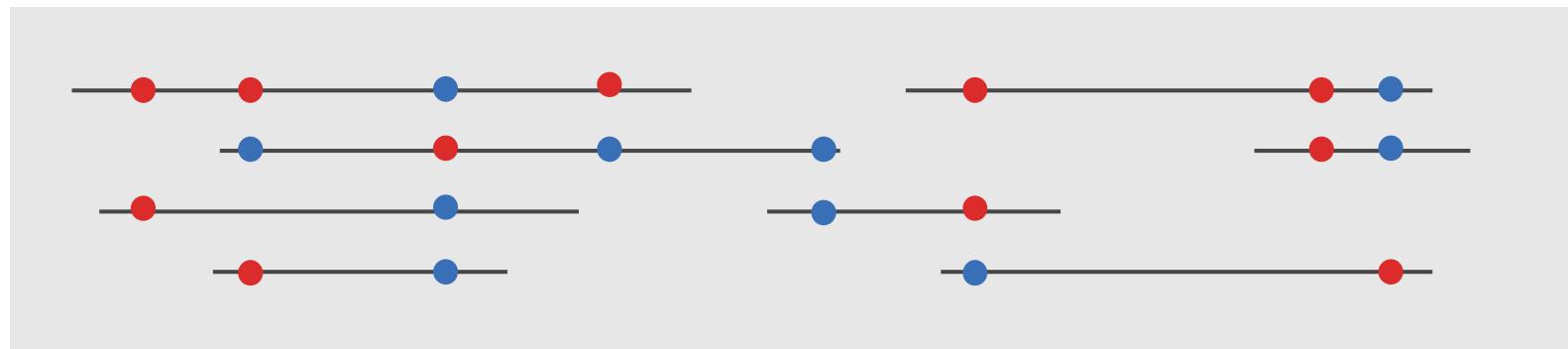
Diploid Genome



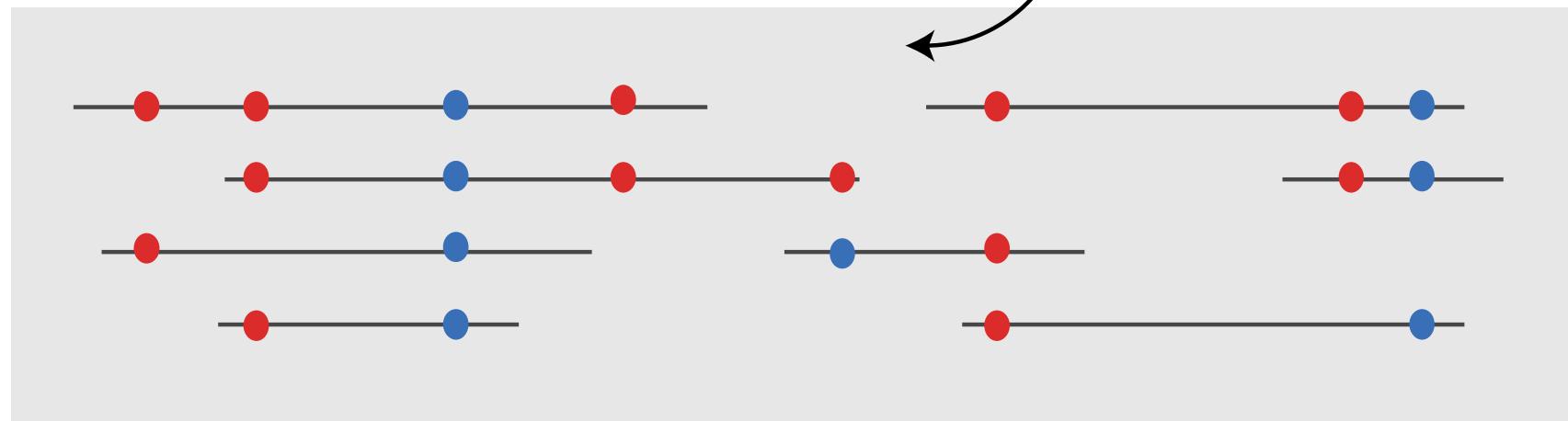
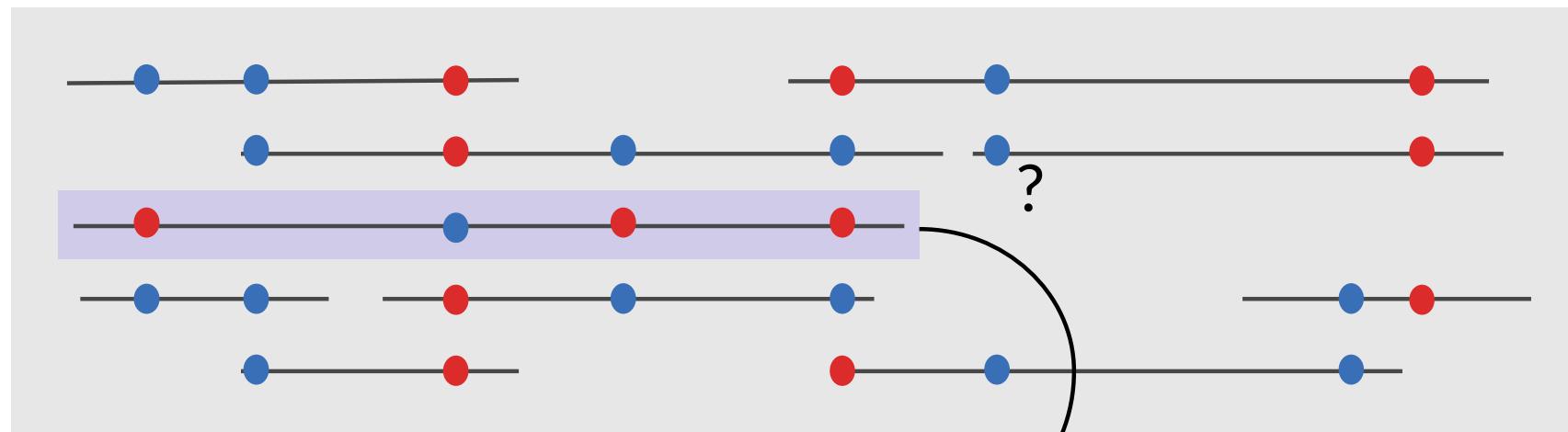
h_1



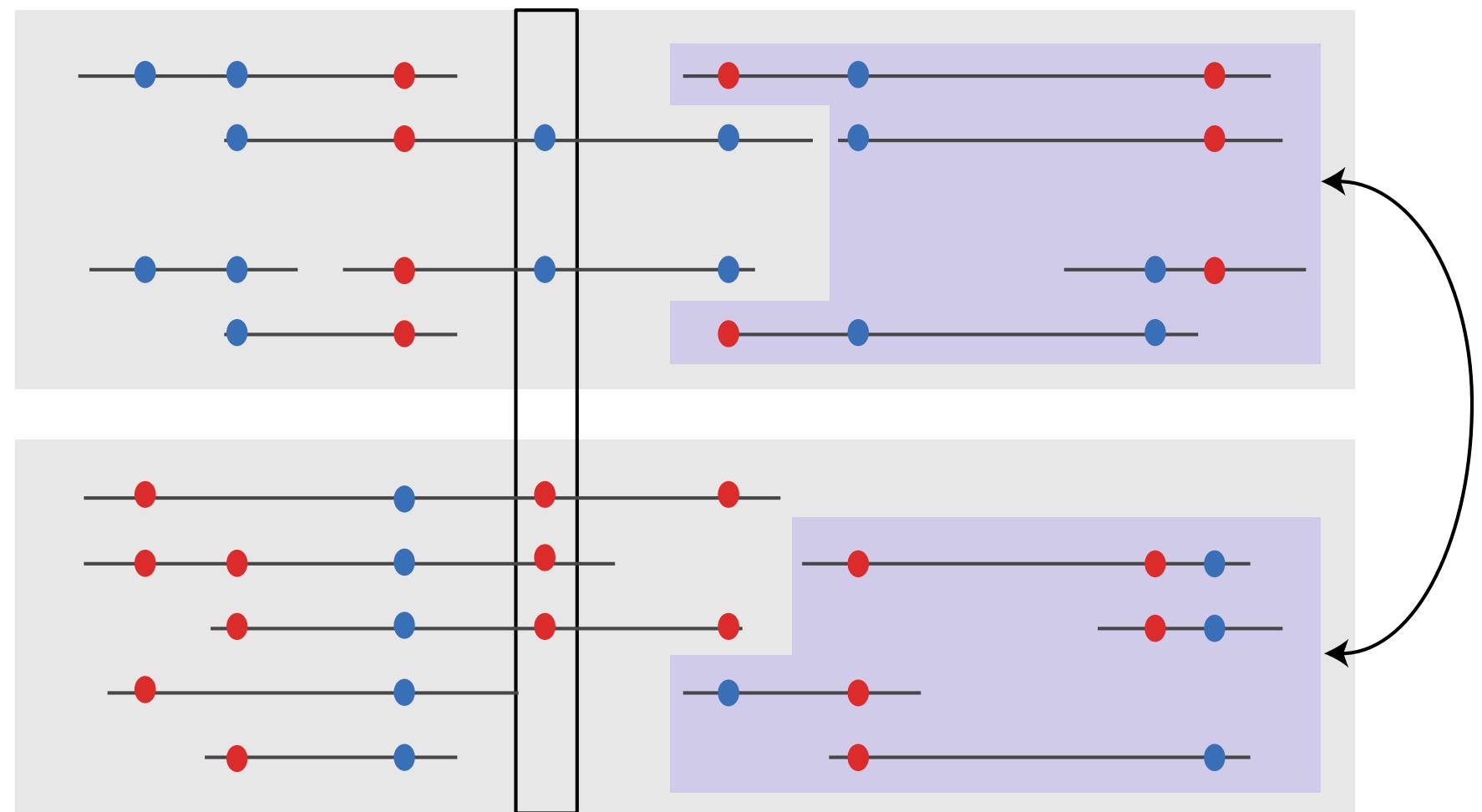
h_2



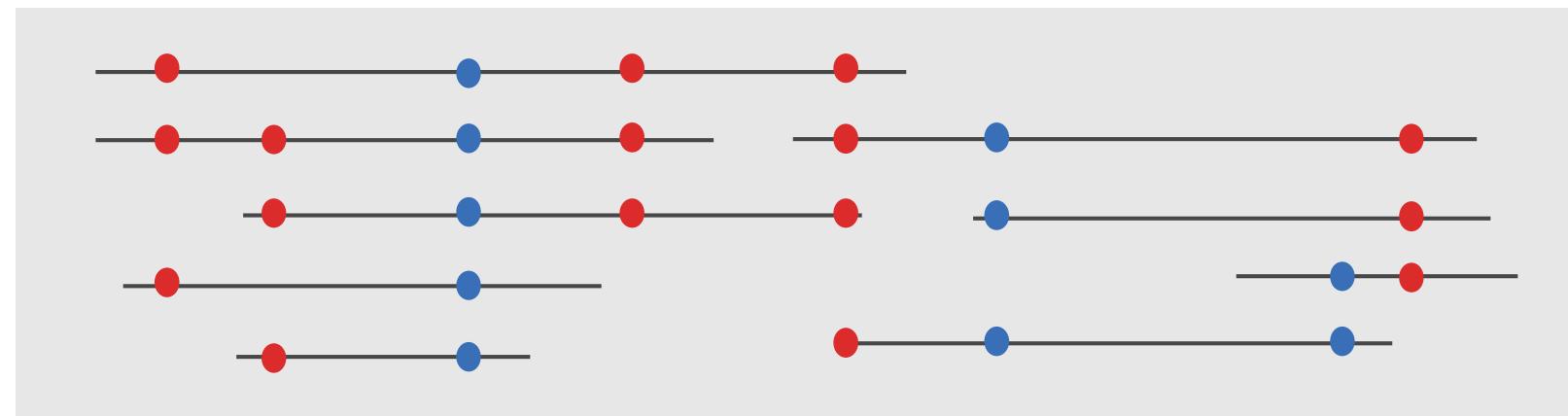
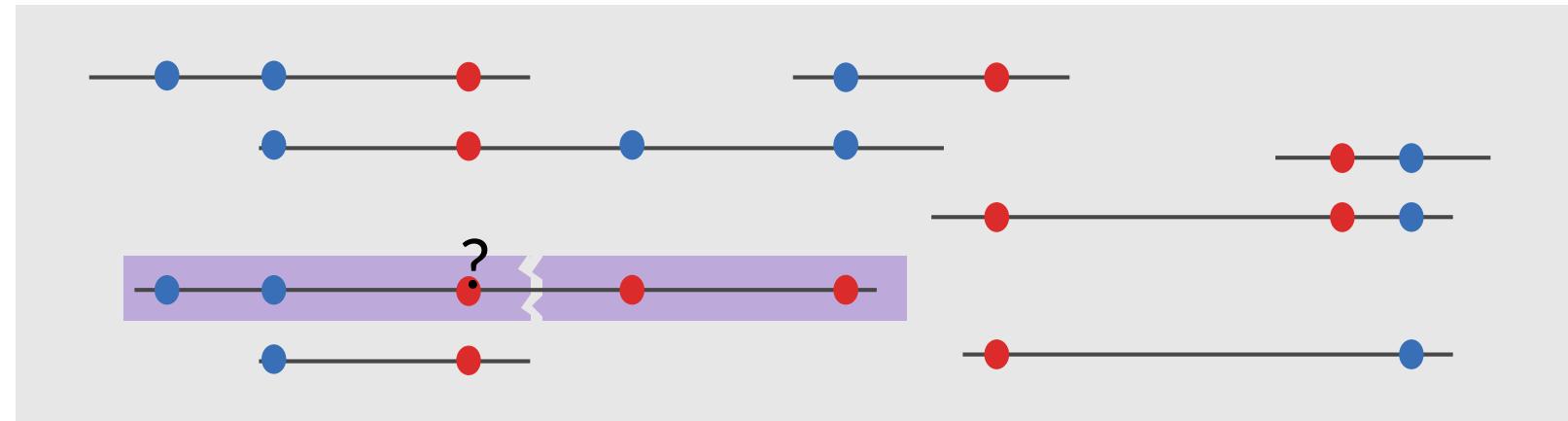
Move 1: assign a cloud to a haplotype



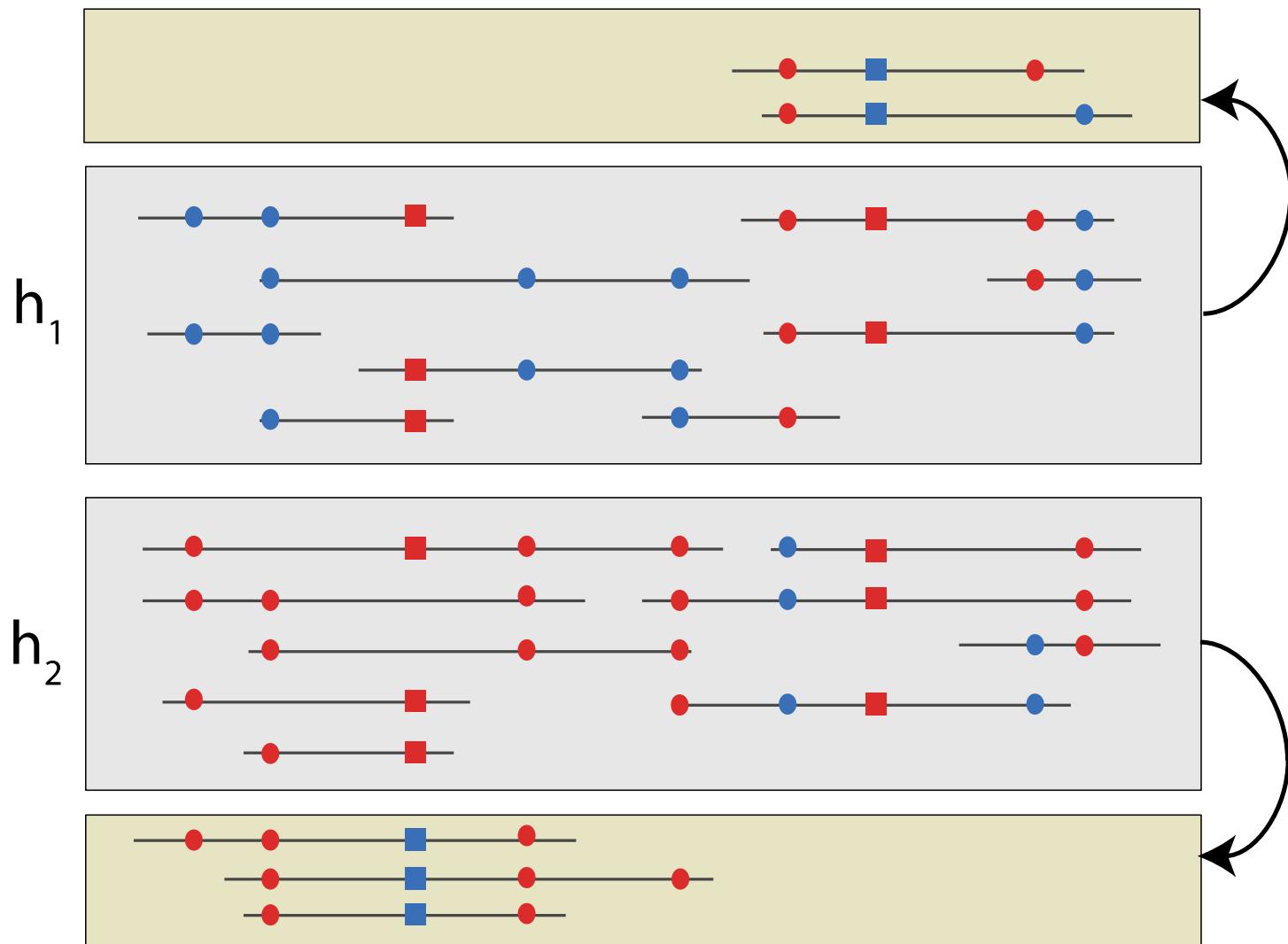
Move 2: unwind a switch error



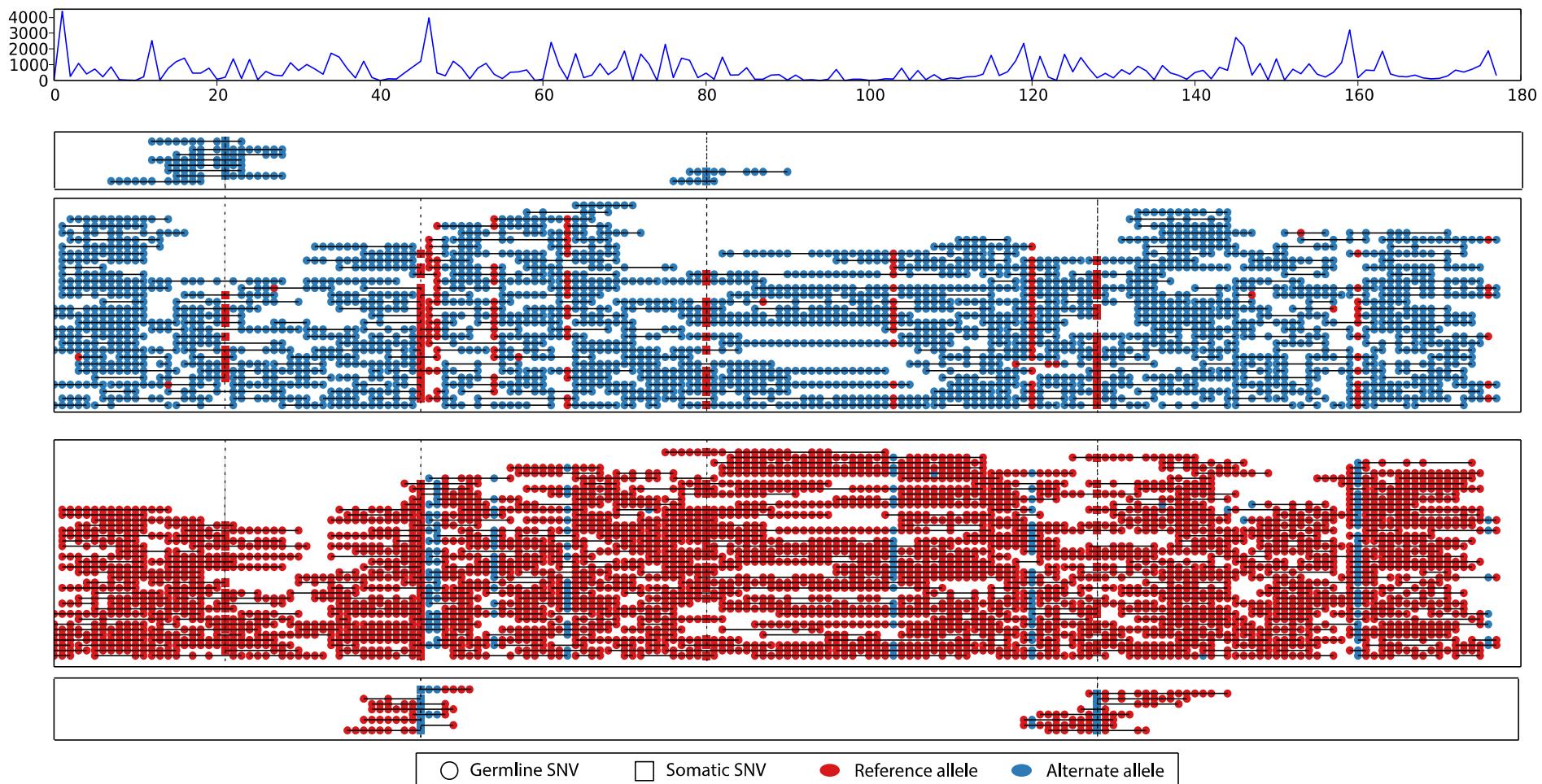
Move 3: flag/unflag clouds as mixed



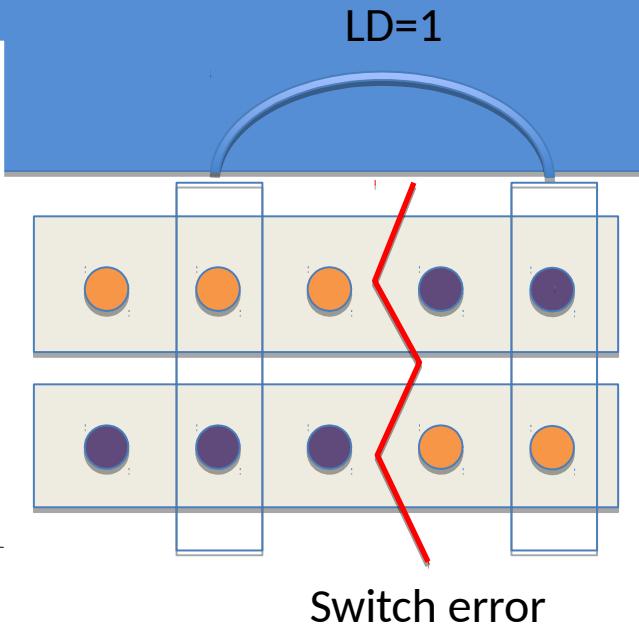
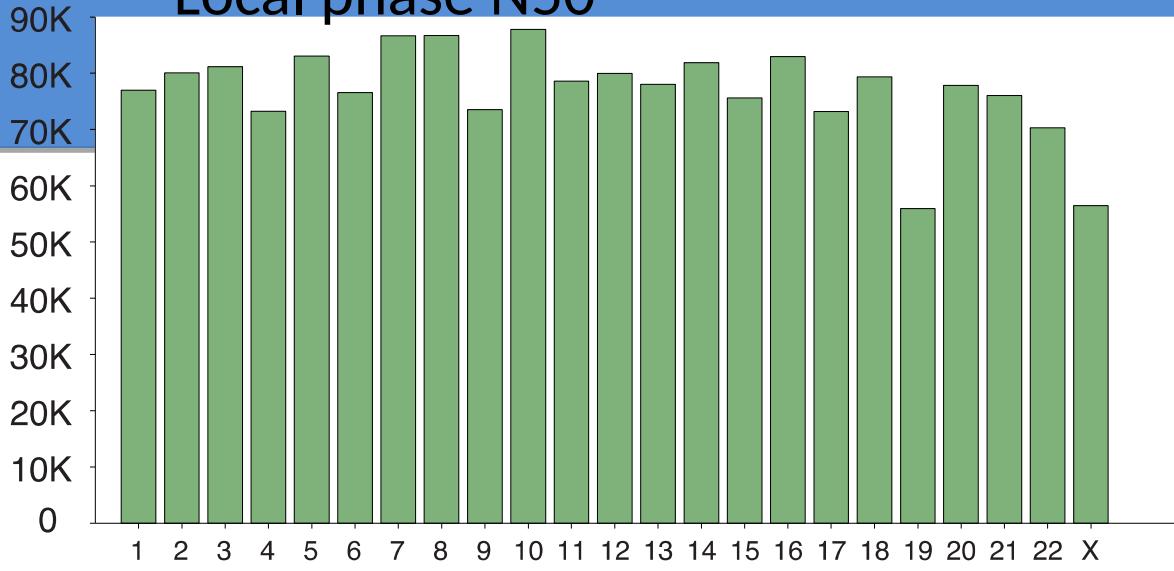
Separate somatic haplotypes



Sample Output



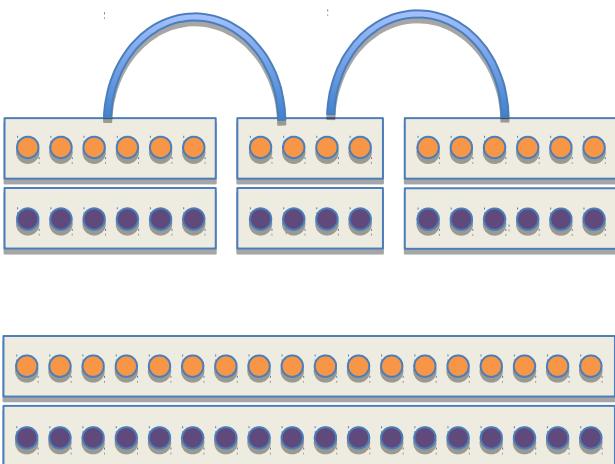
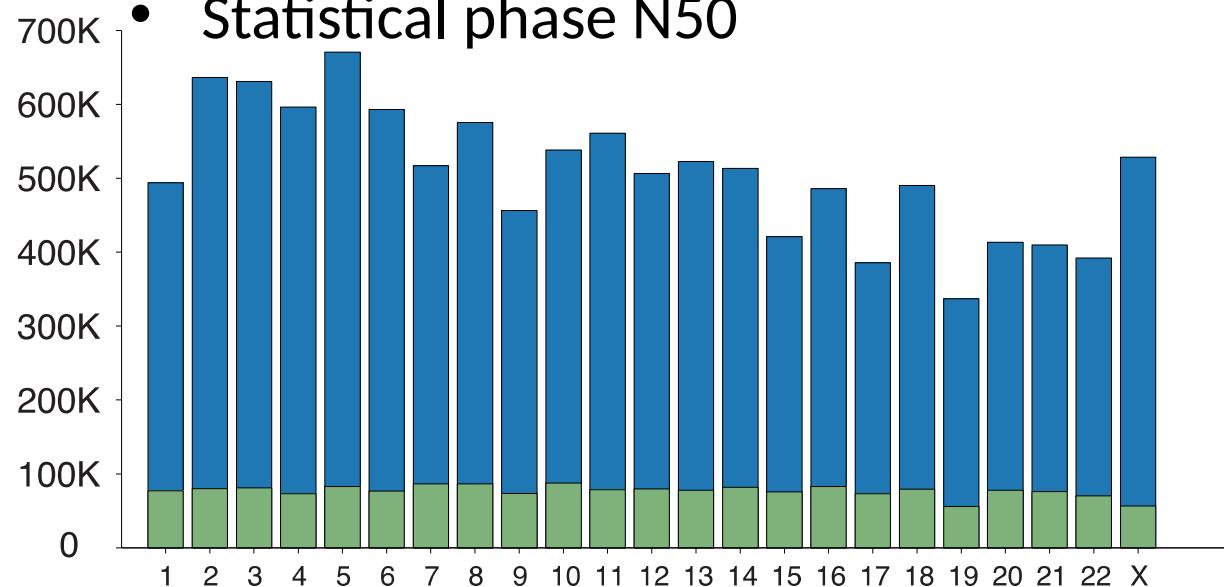
- Local phase N50



- Switch Errors

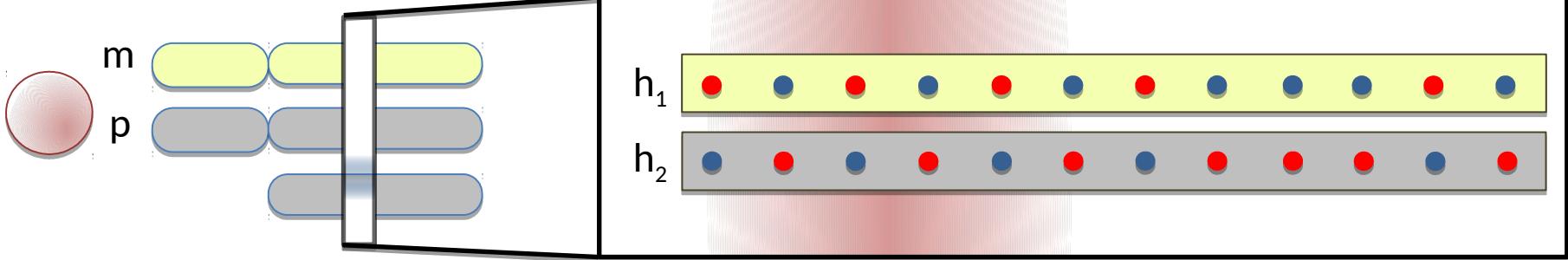


- Statistical phase N50



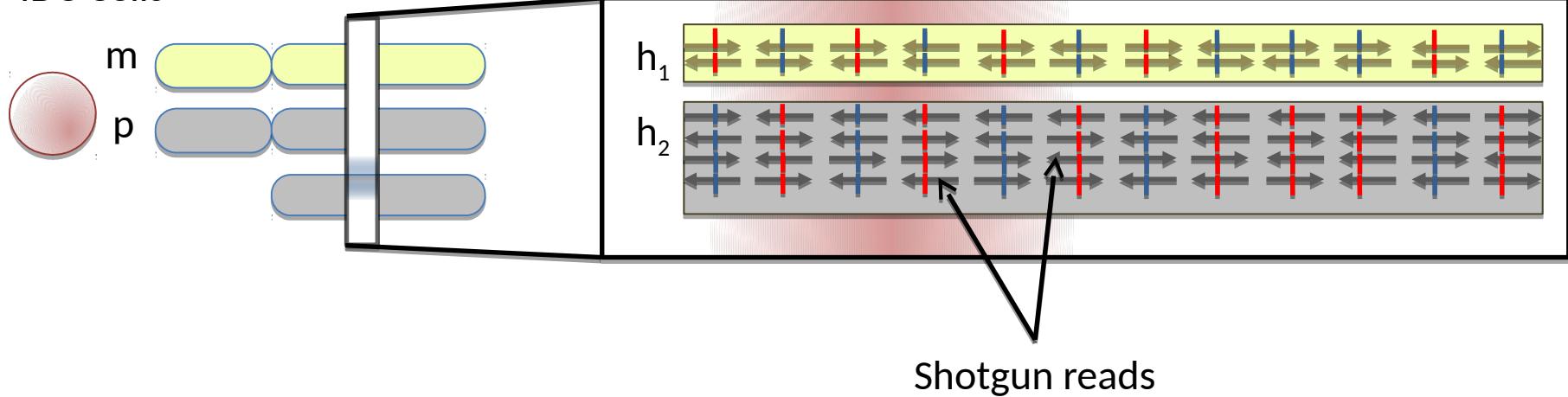
- We can use aneuploidy information to detect switch errors

IDC Cells



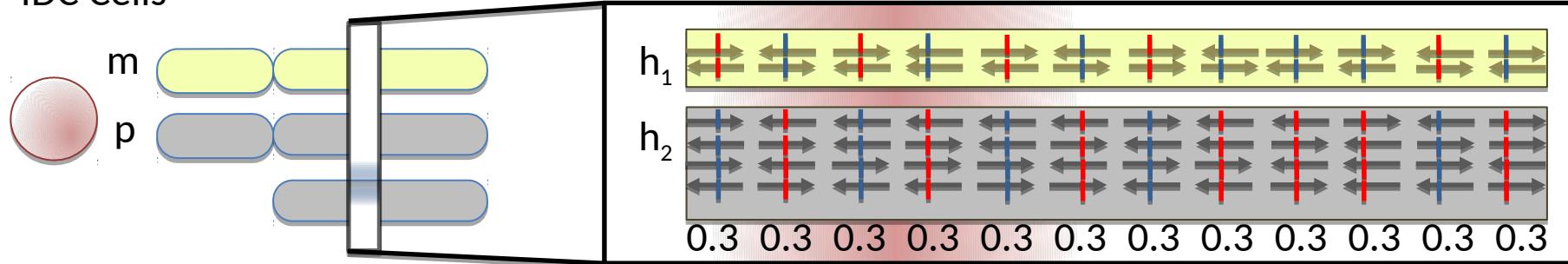
- We can use aneuploidy information to detect switch errors

IDC Cells

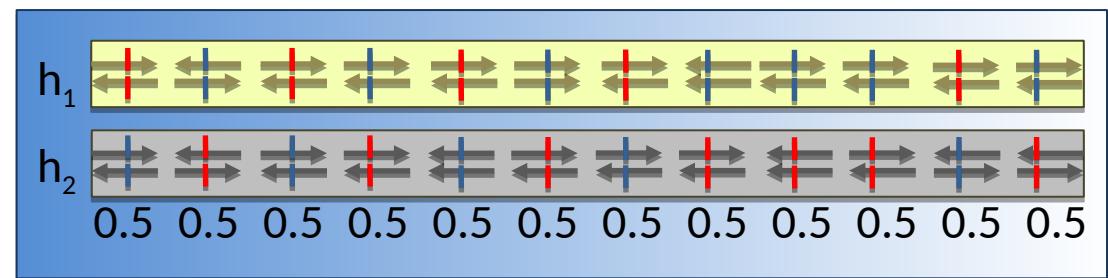


- We can use aneuploidy information to detect switch errors

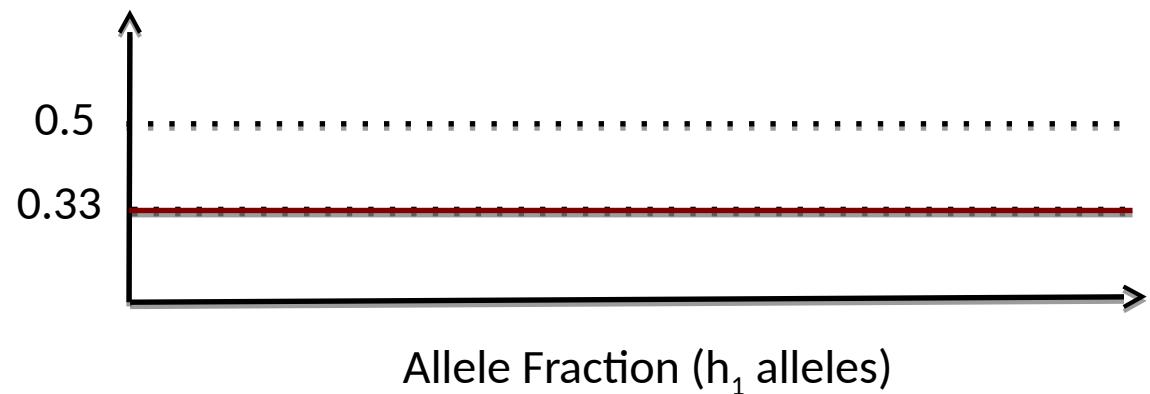
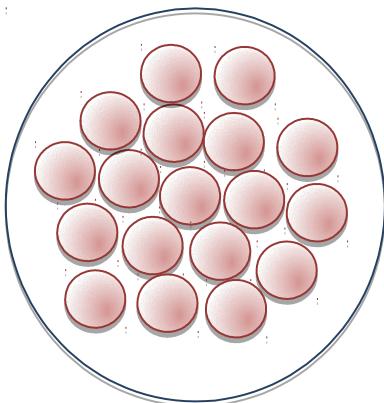
IDC Cells



Normal Cells

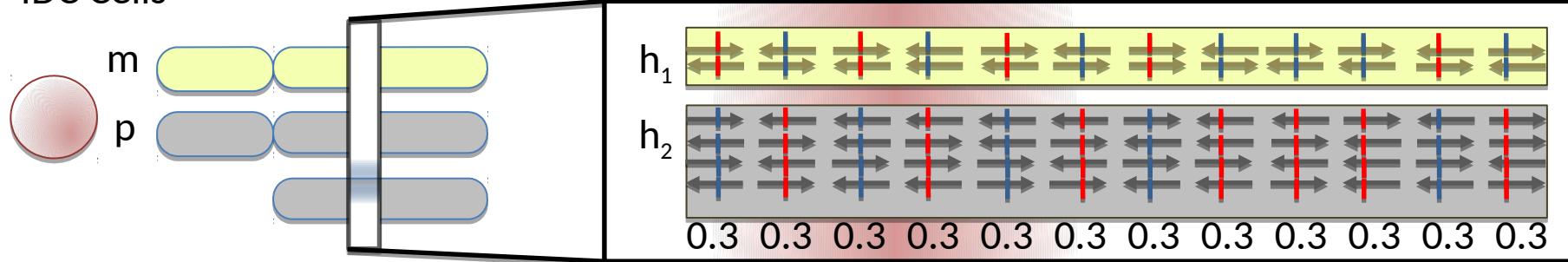


IDC Sample

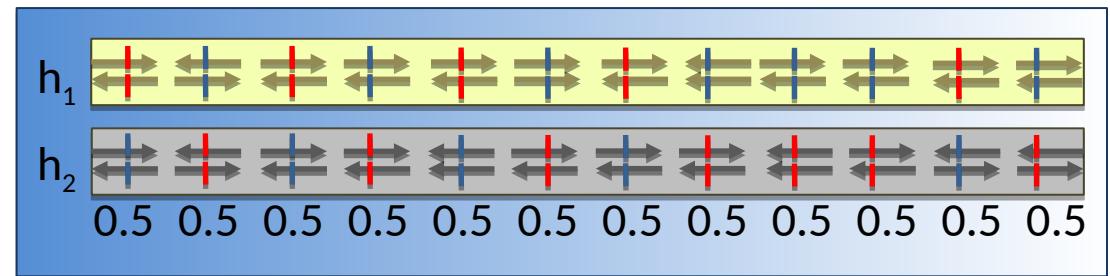


- We can use aneuploidy information to detect switch errors

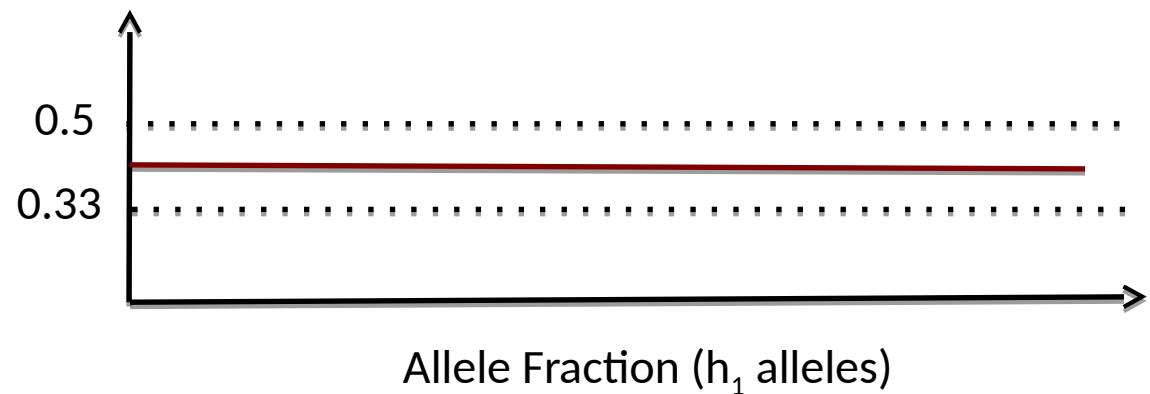
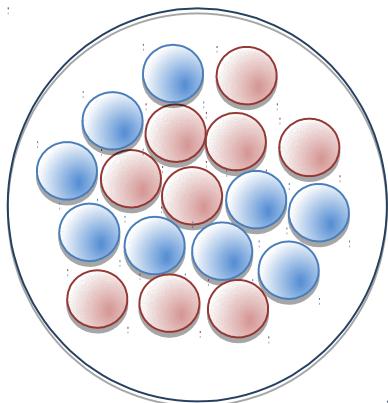
IDC Cells



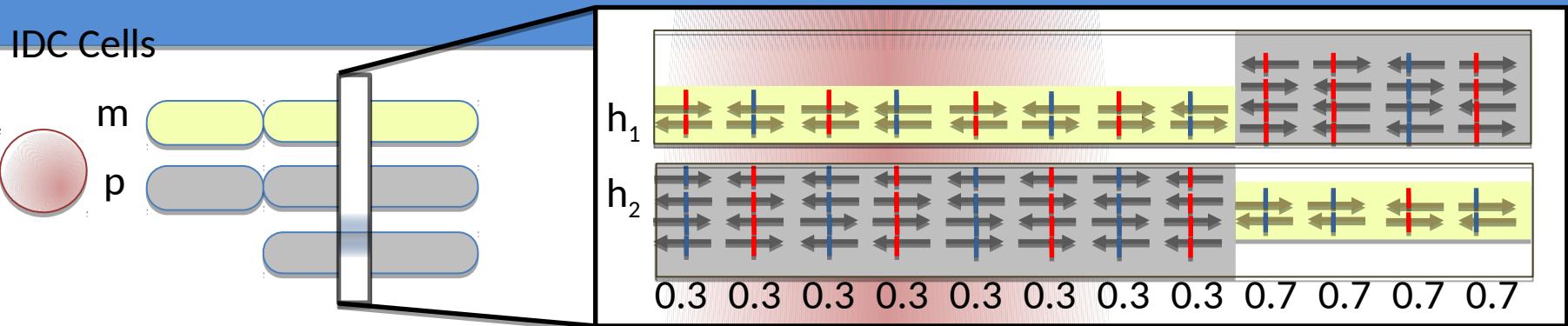
Normal Cells



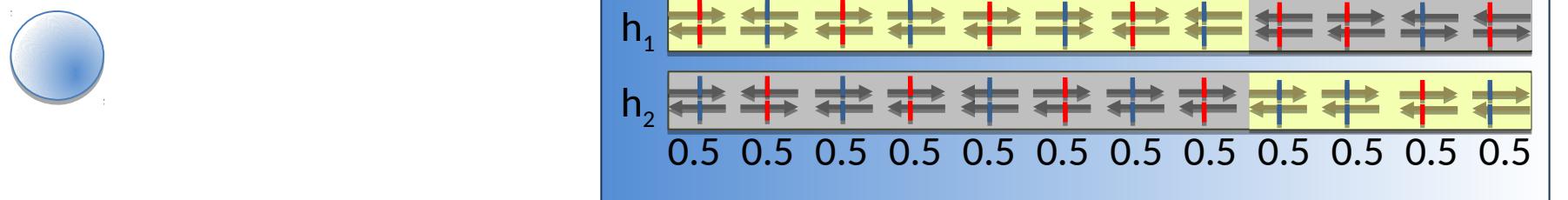
IDC Sample



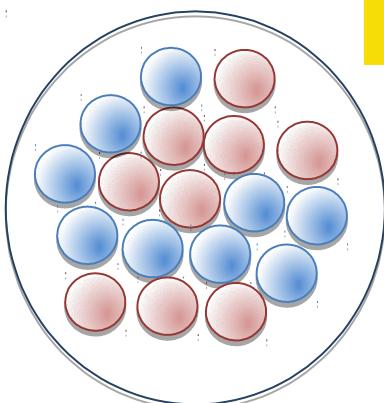
- We can use aneuploidy information to detect switch errors



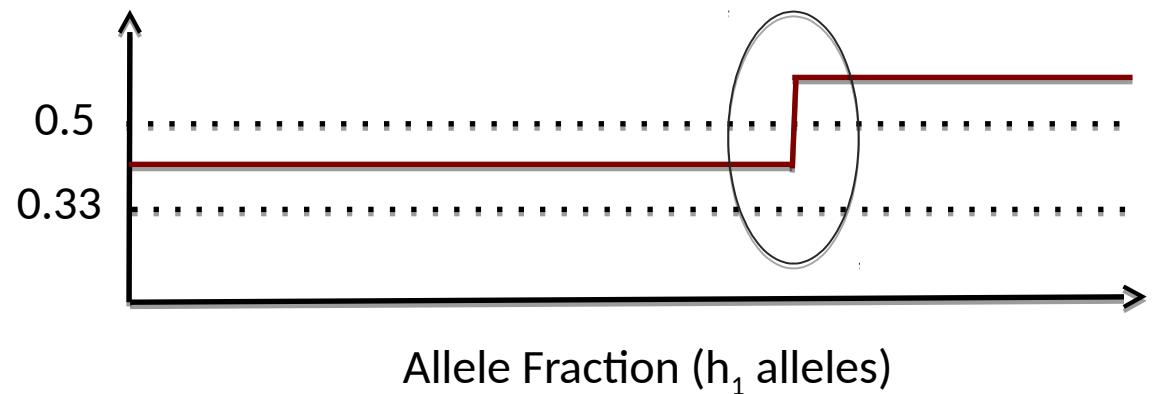
Normal Cells



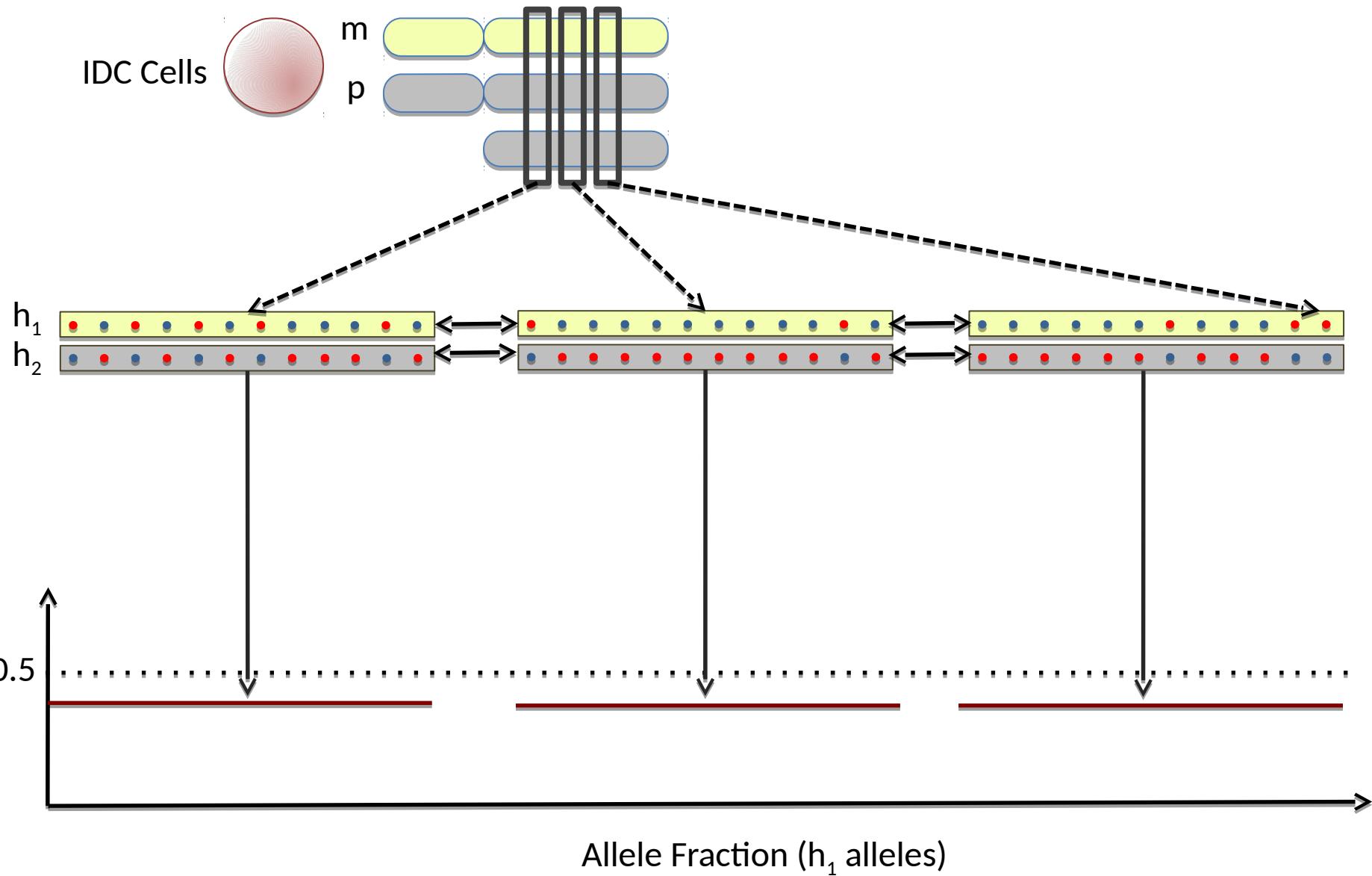
IDC Sample

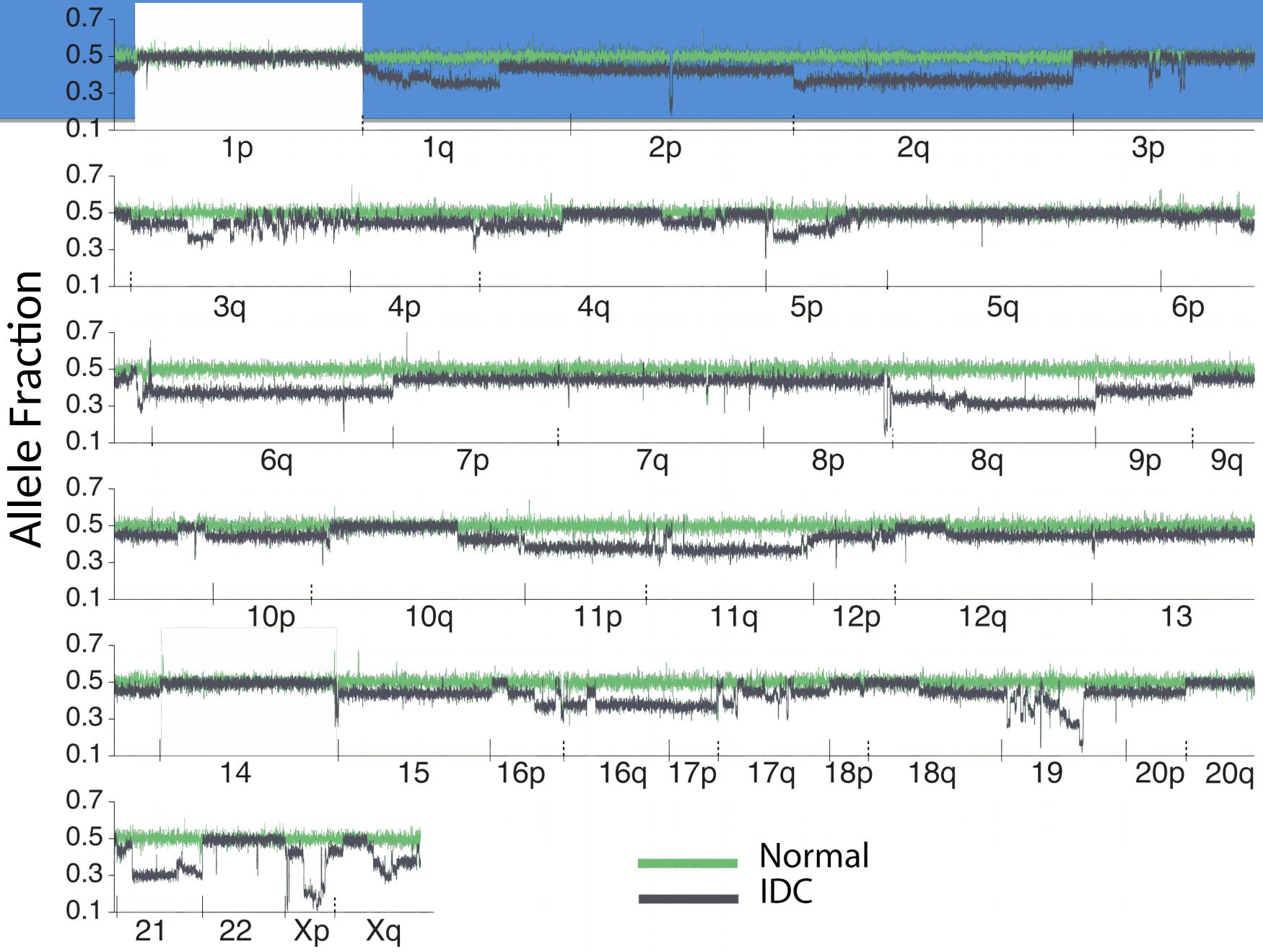


Switch error rate: 0.7/Mbp

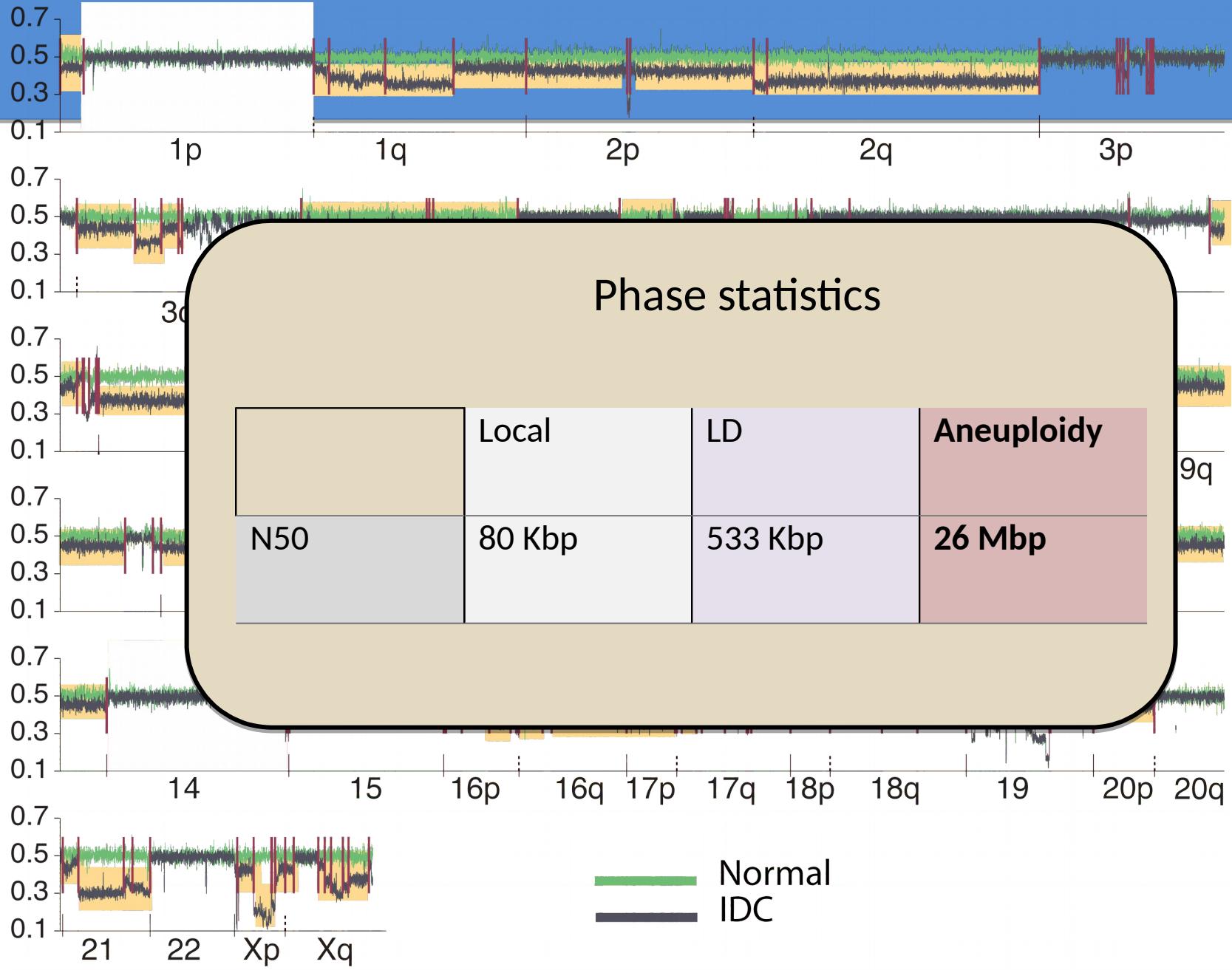


- We can connect haplotype blocks in regions of aneuploidy.





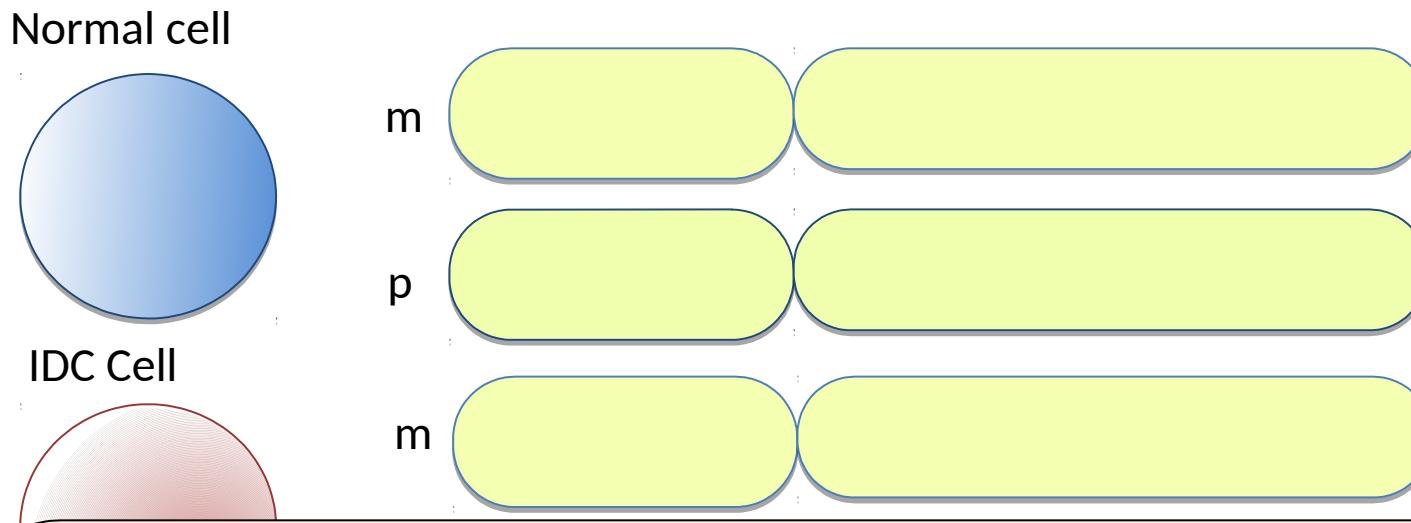
Allele Fraction



Normal
IDC

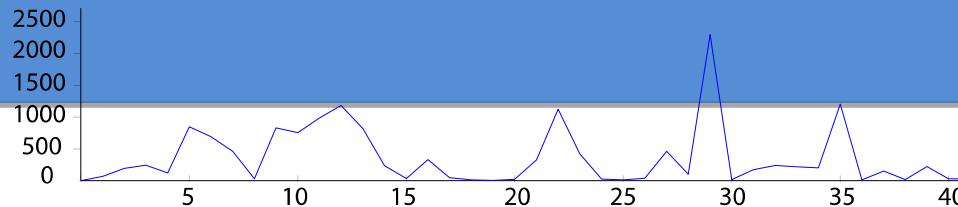
Phase statistics

Validation: Somatic SNVs in LOH

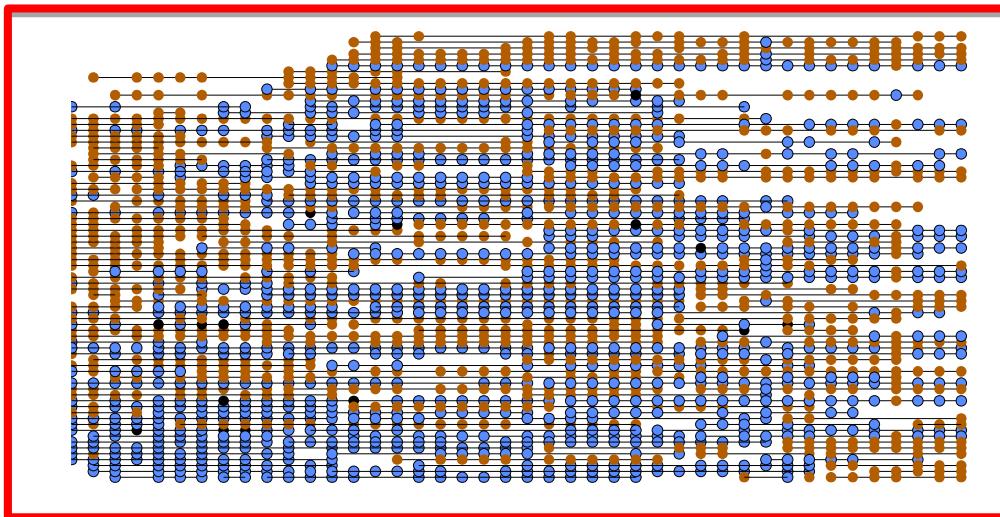


- 628 Mbp (20% of genome) exhibit LOH
- 308 somatic SNVs
 - 304 are assigned correctly

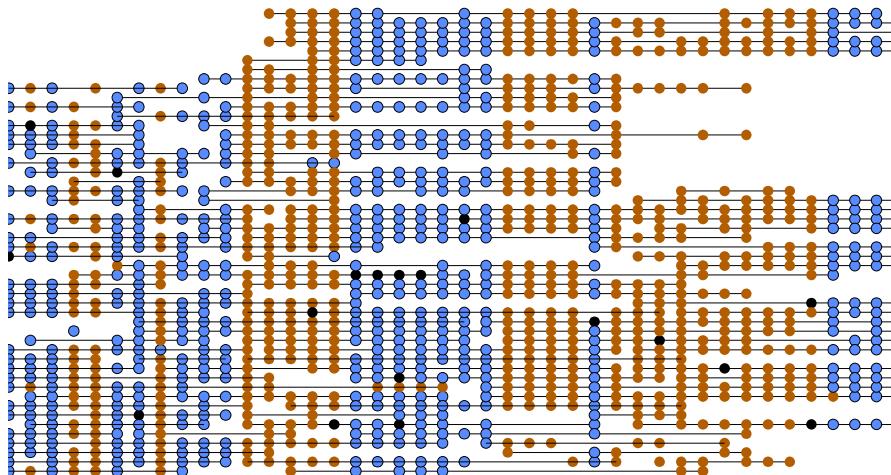
Chr17: 21305362-21320556



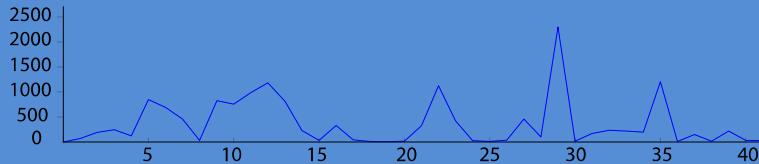
h_1



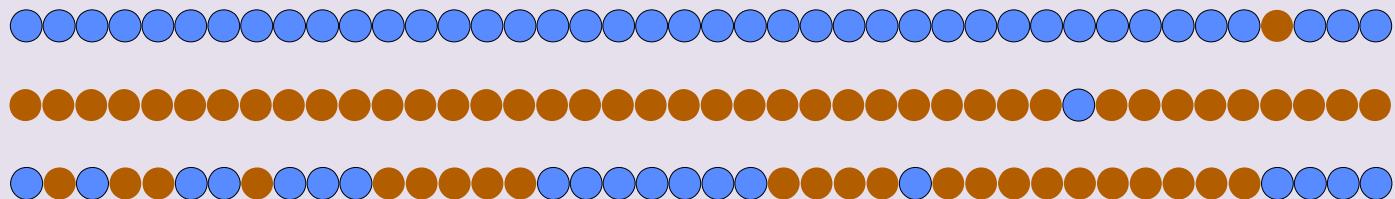
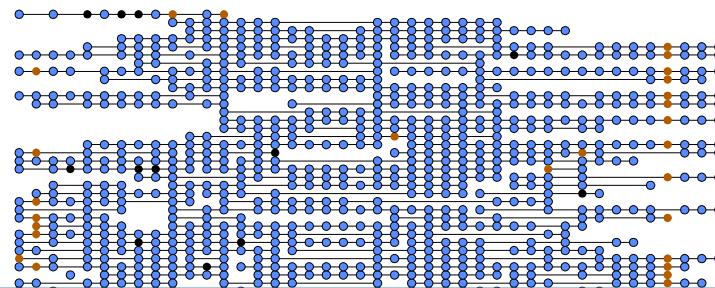
h_2



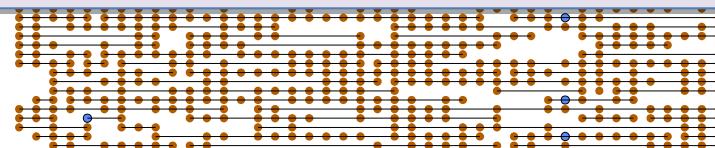
Chr17: 21305362-21320556



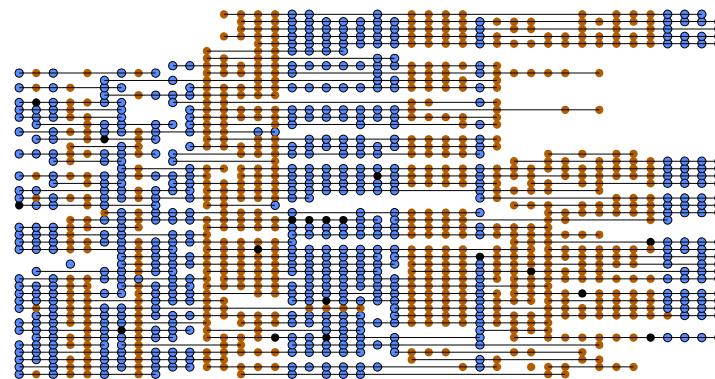
h_1



1

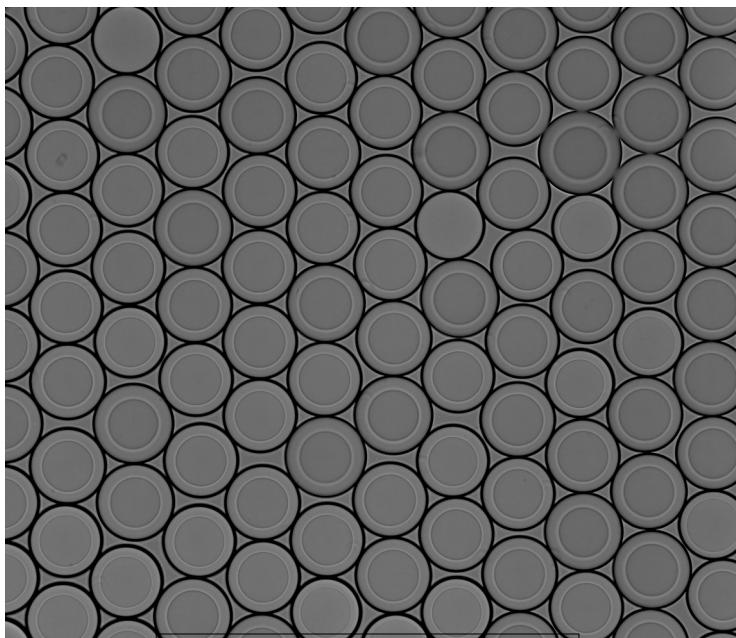
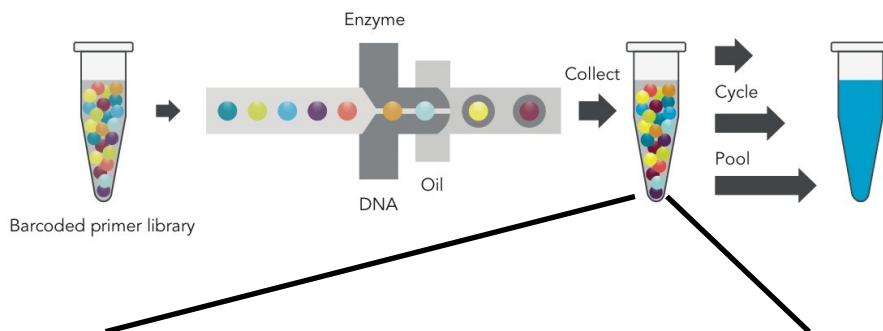


h_2



10x System

Massively Parallel Partitioning



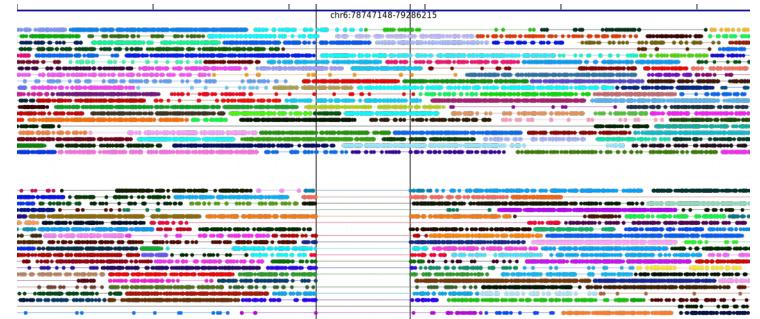
X 100,000+

10X Instrument & Reagents



Read Clouds ("linked reads")

Hap1



Further Readings for the Curious

- **Fantastic Cancer Reviews**

- Hanahan and Weinberg. 2000. The hallmarks of cancer. *Cell* 100: 57-70.
- Hanahan and Weinberg. 2011. Hallmarks of cancer: the next generation. *Cell* 144, 646-74.

- **Reviews of Cancer Genomics**

- Meyerson, Matthew, Stacey Gabriel, and Gad Getz. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* 11, no. 10 (October): 685-696. doi:10.1038/nrg2841.
<http://www.nature.com/doifinder/10.1038/nrg2841>.
- Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* 13, 795-806 (2012).

- **Variant Calling**

- Dalca, Adrian V, and Michael Brudno. 2010. Genome variation discovery with high-throughput sequencing data. *Briefings in bioinformatics* 11, no. 1 (January):
<http://www.ncbi.nlm.nih.gov/pubmed/20053733>.
- Medvedev, Paul, Monica Stanciu, and Michael Brudno. 2009. Computational methods for discovering structural variation with next-generation sequencing. *nature methods* 6, no. 11 <http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1374.html>.