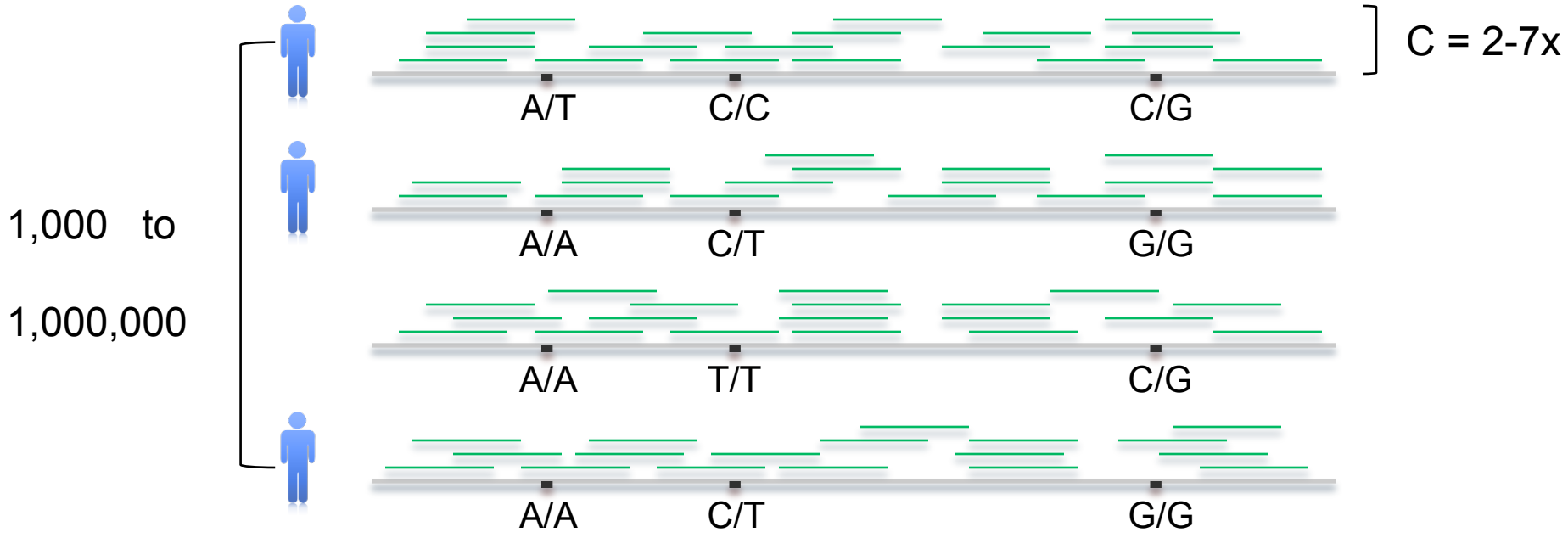
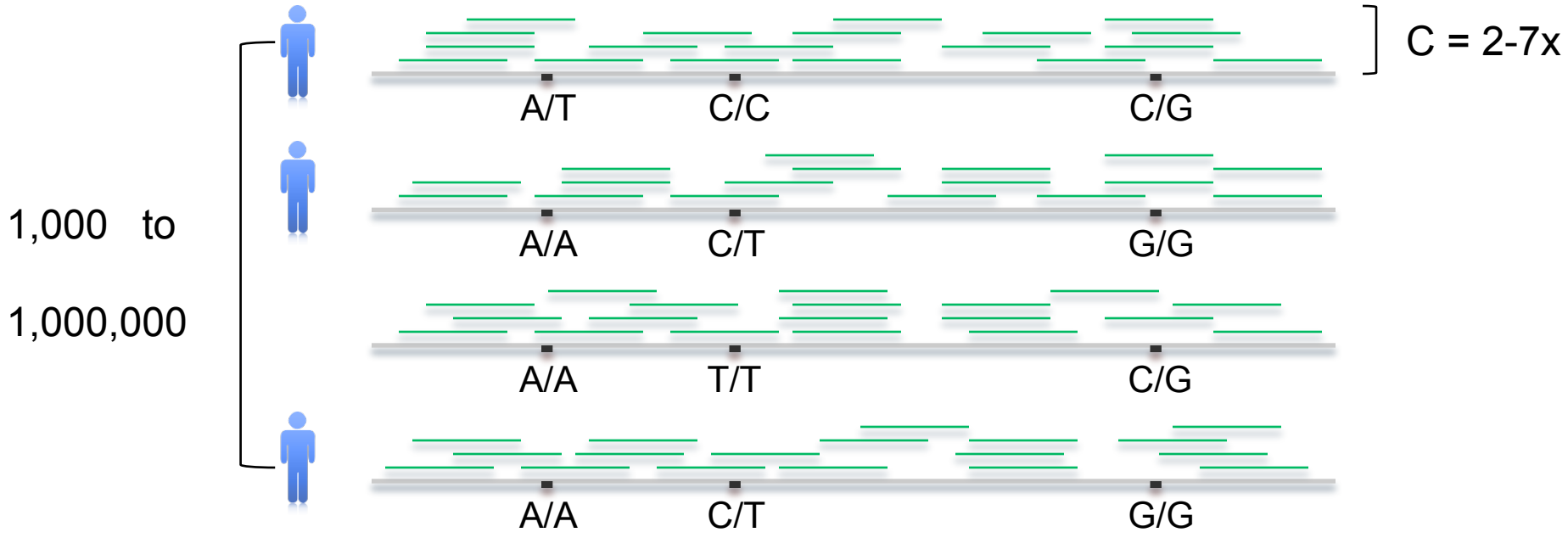
Population Sequencing



Population Sequencing



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [$p_{ijg} = \text{Prob}(g_{ij} = g \mid \text{data})$]

Population Sequencing

When C is high (>30x),

Prob(g_{ij} = *g* | data) ~

Prob(g_{ij} = *g* | reads mapping on (i, j))

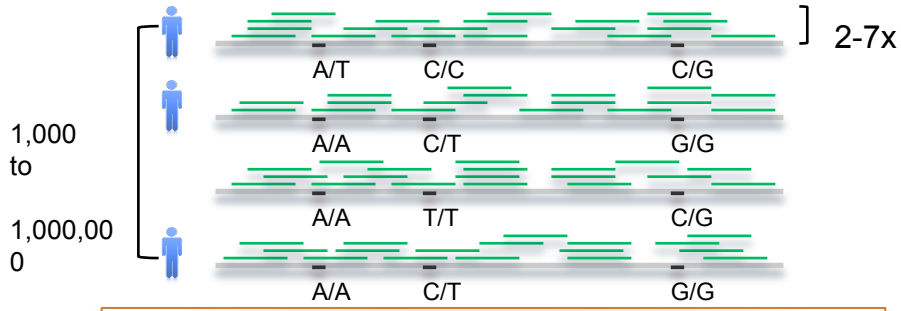
fast & easy

When C is low,

Prob(g_{ij} = *g* | data) needs to leverage LD:

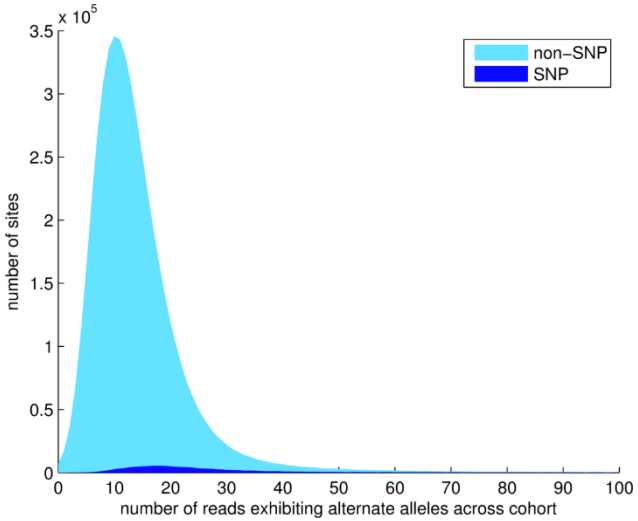
positions j' ≠ j in all individuals

in principle, intractable



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [p_{ijg} = Prob(g_{ij} = *g* | data)]

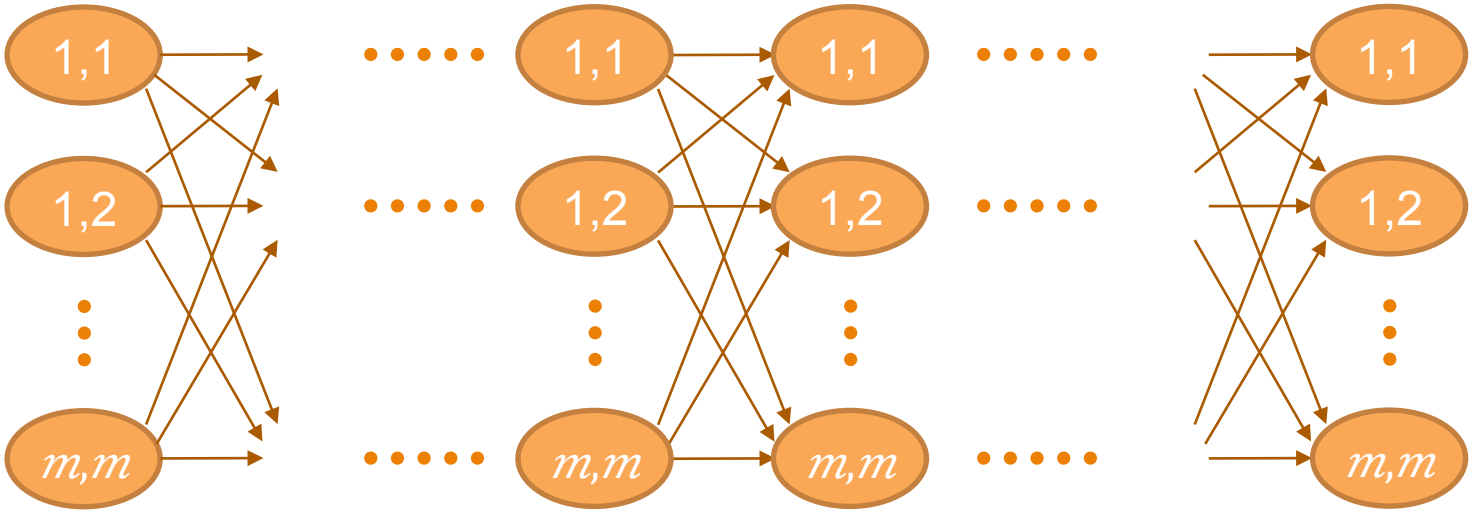


1000 Genomes Project, 2535 individuals, 7x sequencing

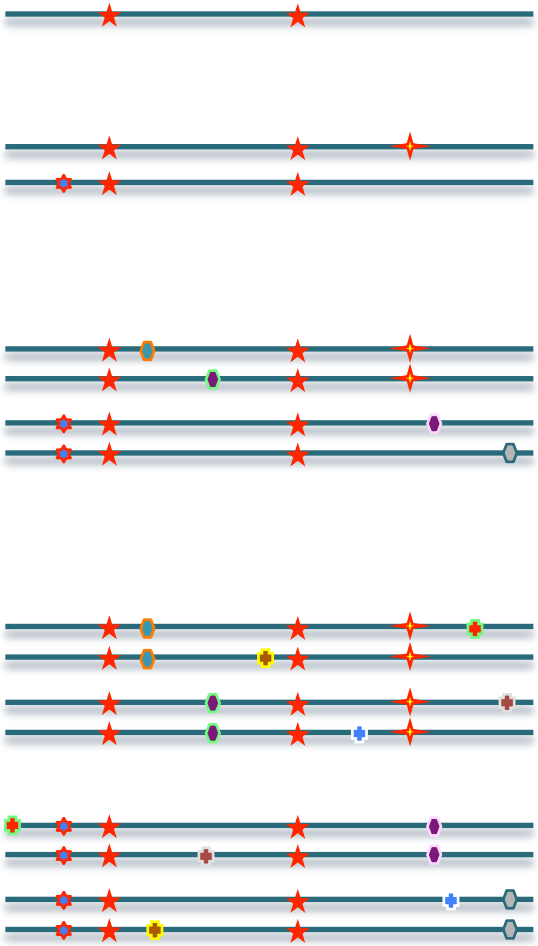
HMM-based models

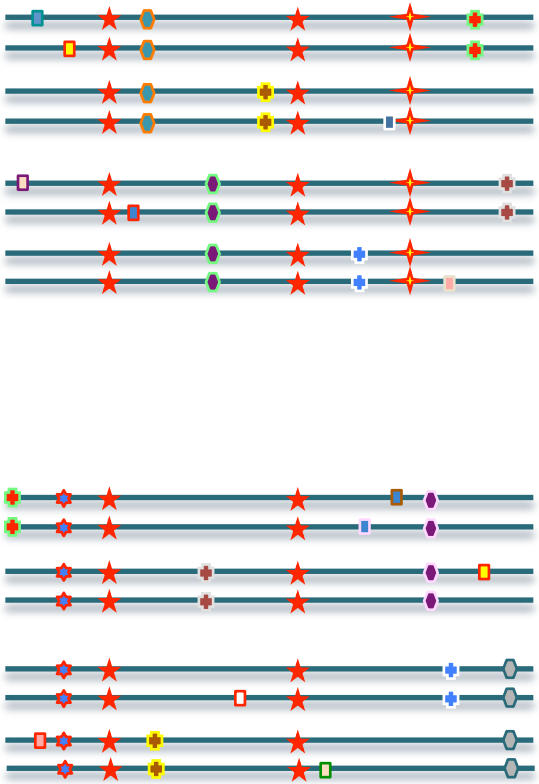
- Li and Stephens 2003

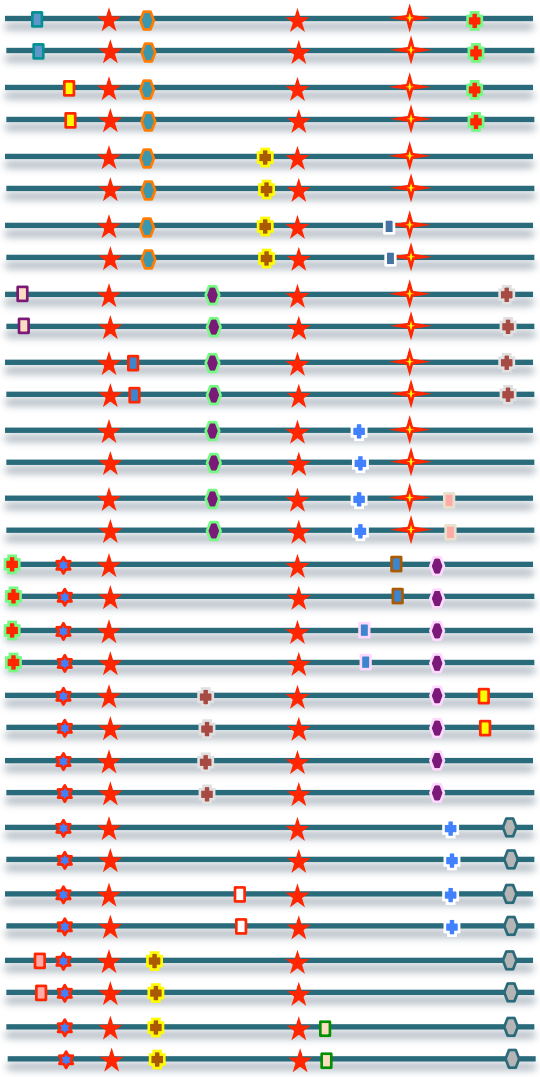
Given m reference haplotypes, and a target sample,
Find the most likely path of haplotype pairs
m^2 states, m^4 transitions per position



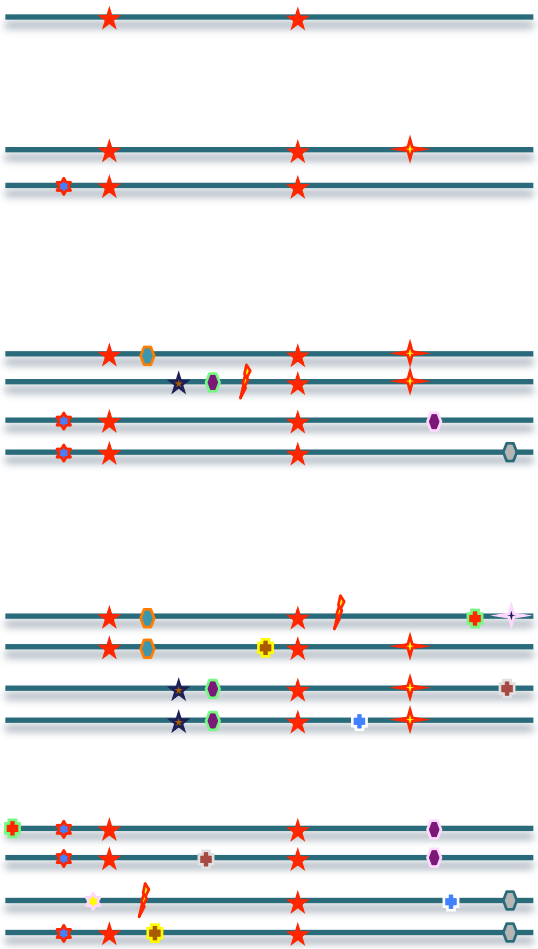
Evolution of a local haplotype

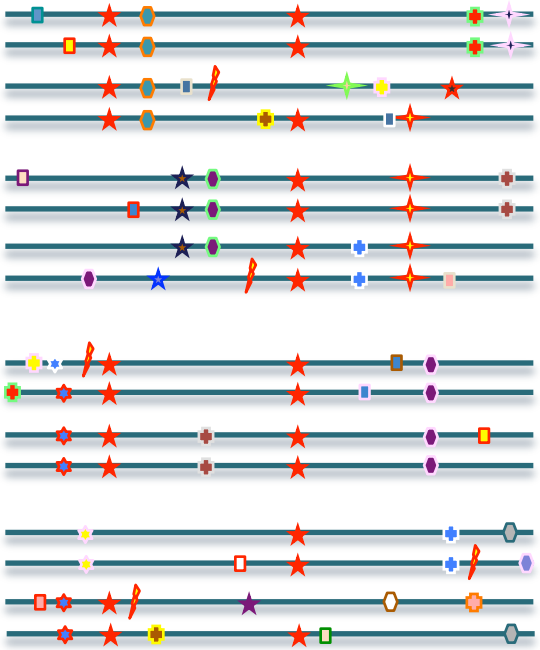


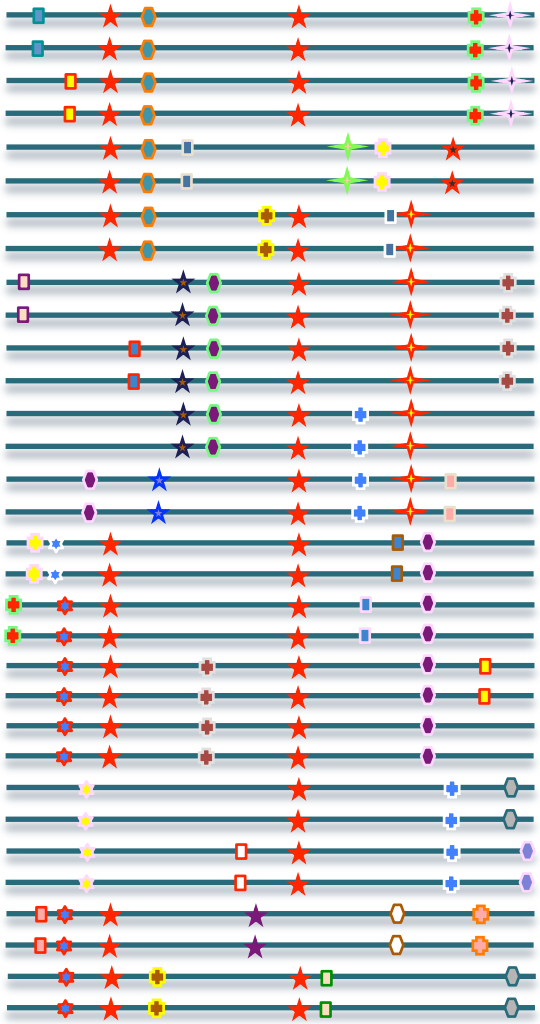




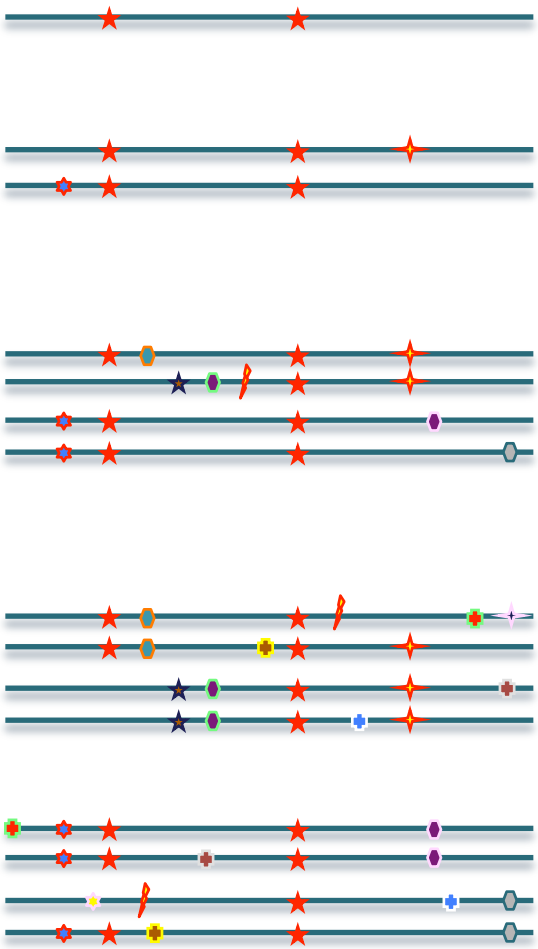
Evolution of haplotypes, recombination

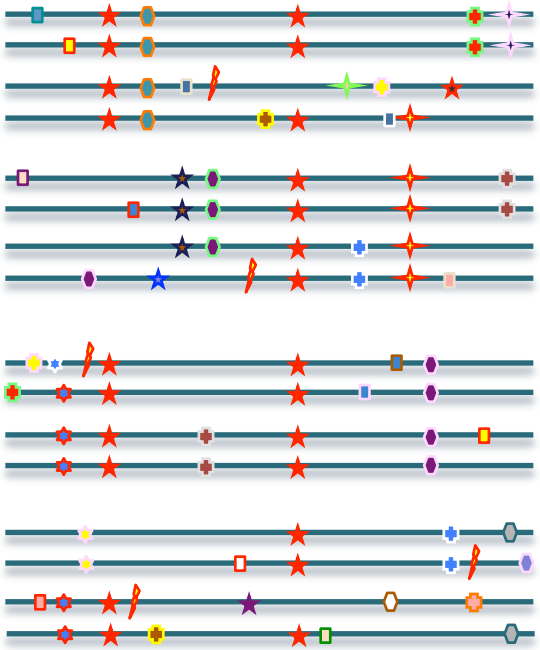


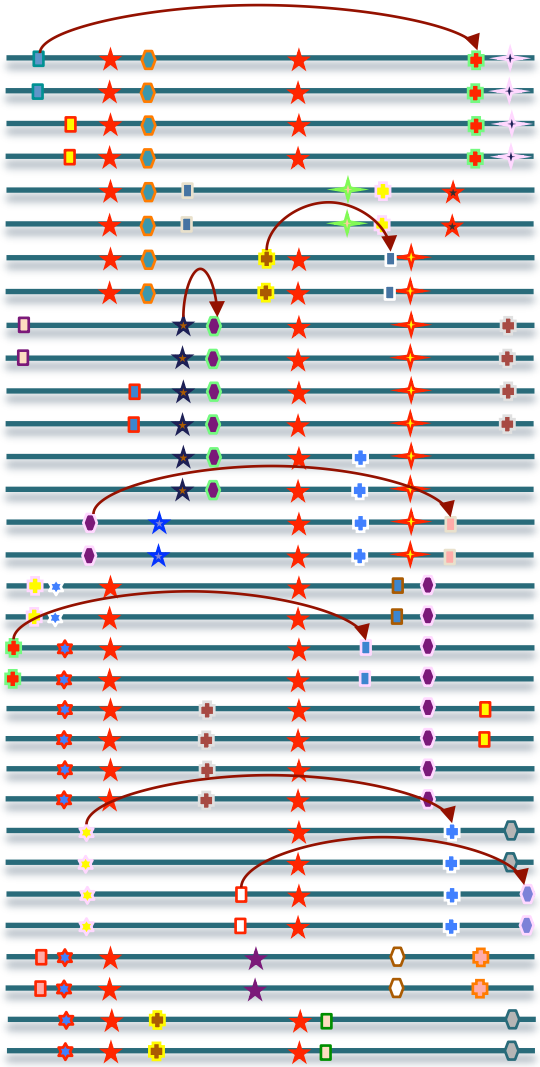




Evolution of haplotypes, recombination





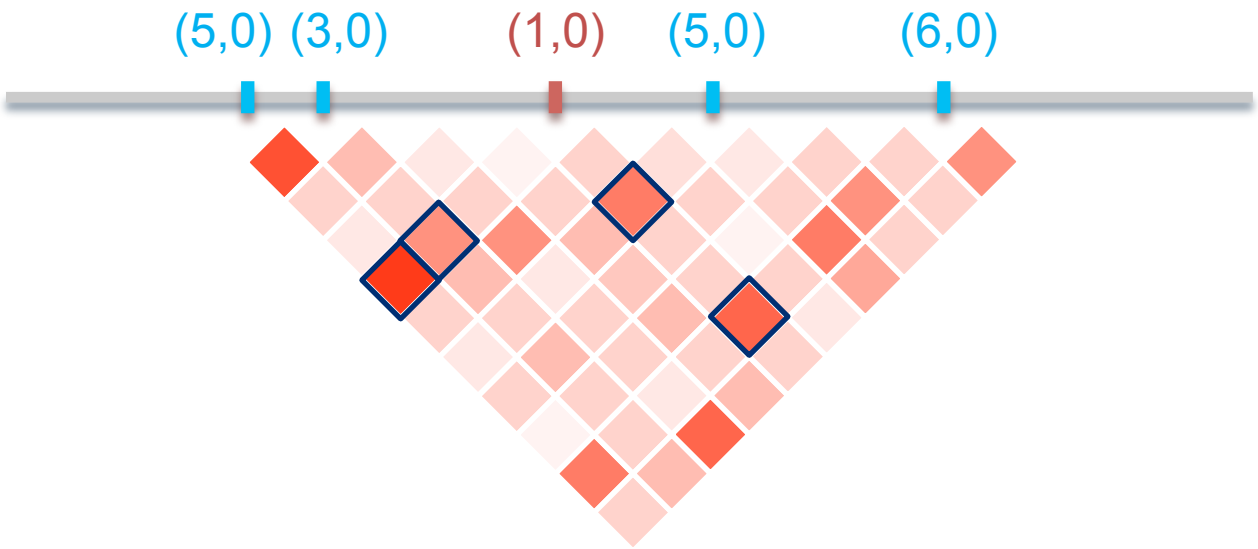


Informative Neighbors

target SNP

k-"nearest" neighbors

in terms of linkage disequilibrium



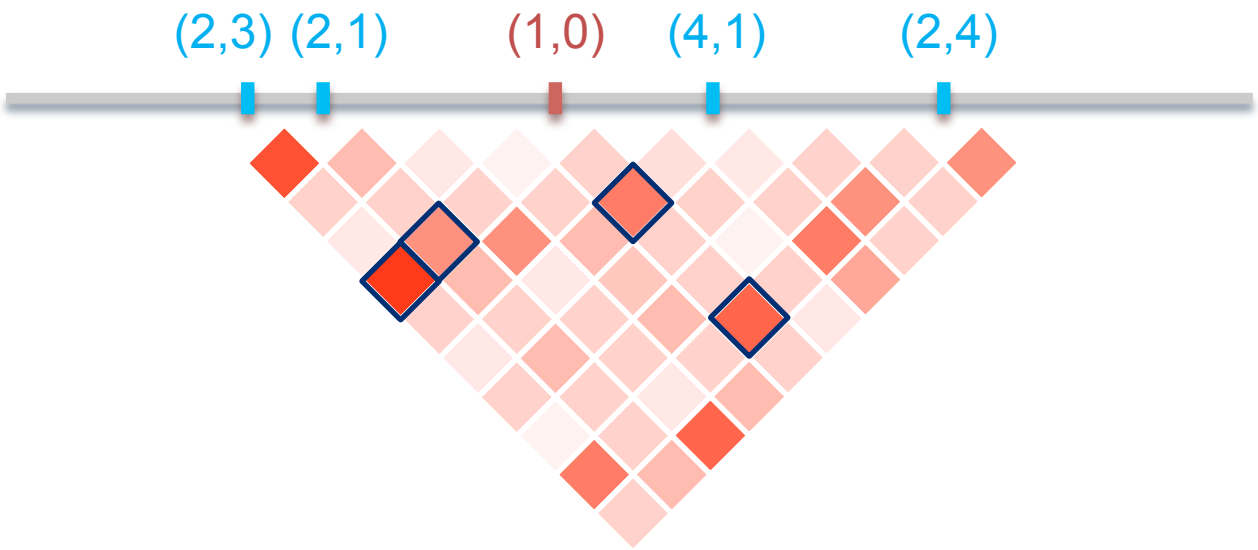
$(R_{ref}, R_{alt}) = \Sigma_{\{target,\ nbrs\}}(r_{ref}, r_{alt}) = (20, 0)$

Informative Neighbors

target SNP

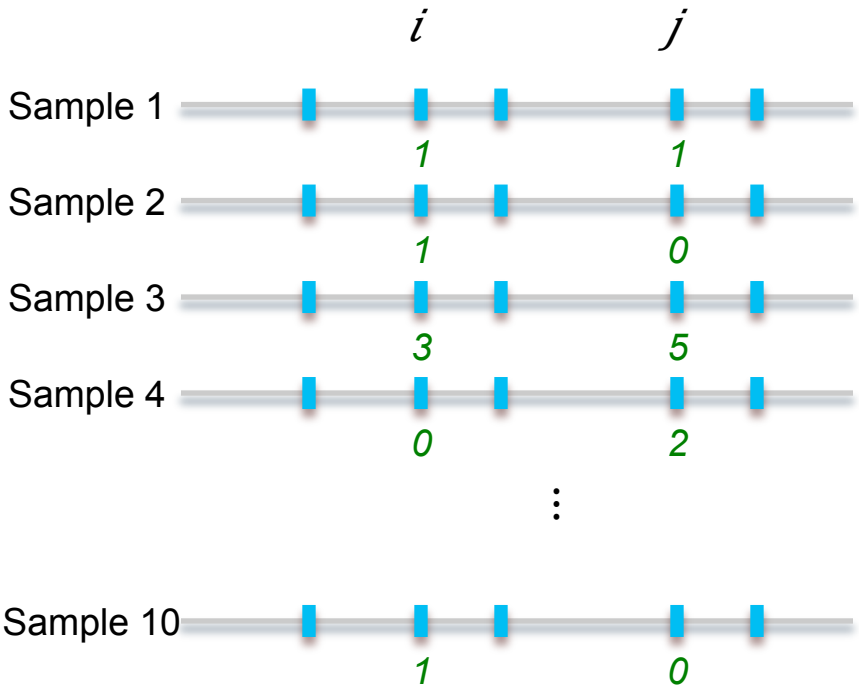
k-"nearest" neighbors

in terms of linkage disequilibrium



$(R_{ref}, R_{alt}) = \Sigma_{\{target,\ nbrs\}}(r_{ref}, r_{alt}) = (11, 9)$

How to pick k nearest neighbors fast



Correlation Coefficient:

$r^2 = (p_{AB} - p_A p_B)^2 / p_A p_B p_a p_b$

Caveat: need **genotyping**, **phasing**

Let

S_i = { samples covering minor allele }

S_i' = { read counts of minor allele }

$S_i = \{1, 2, 3, 10\}$

$S_j = \{1, 3, 4\}$

$S_i' = \{1, 2, 3, 3, 3, 10\}$

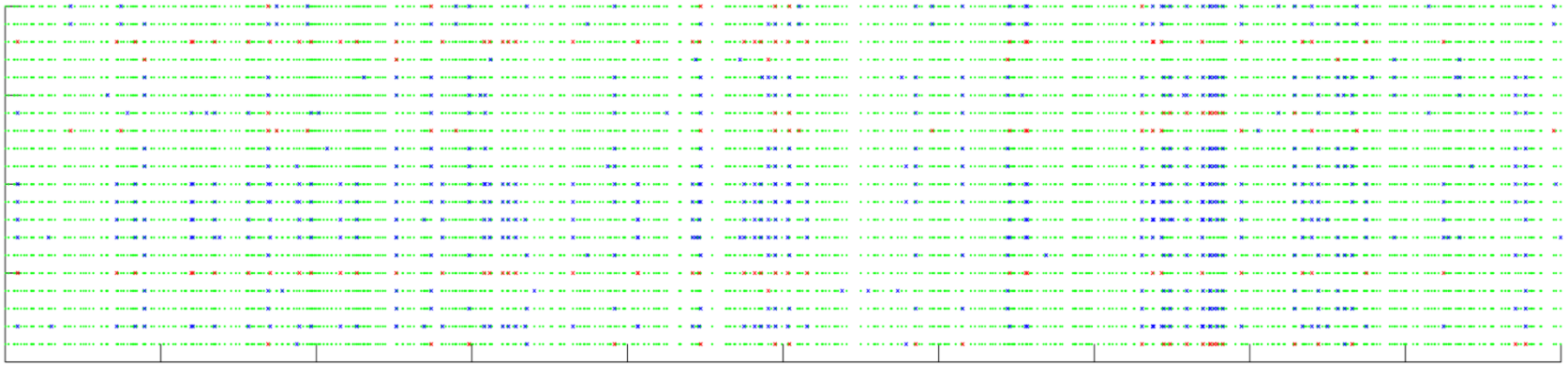
$S_j' = \{1, 3, 3, 3, 3, 3, 4, 4\}$

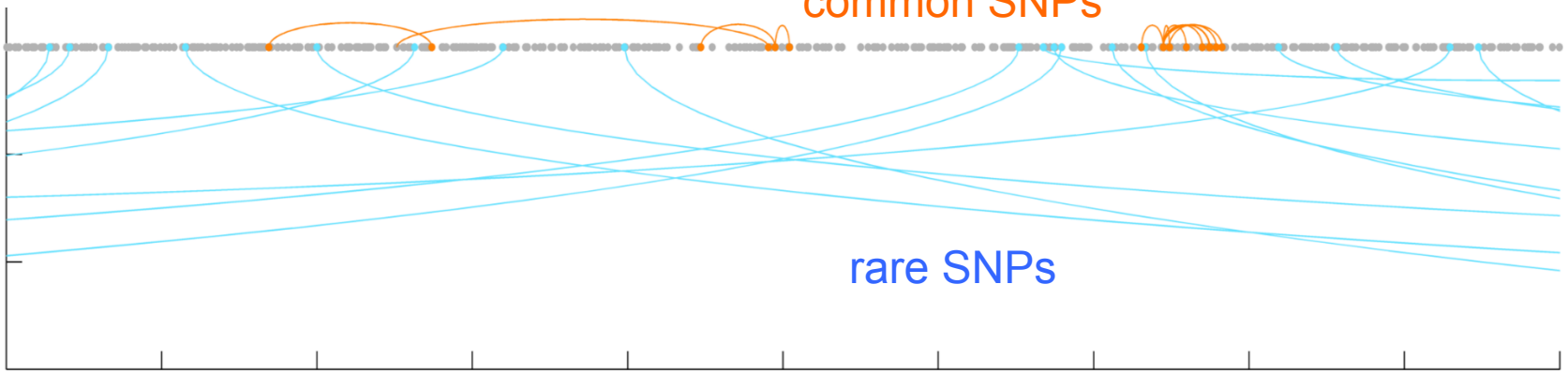
$Sim_1(i, j) = (S_i \cap S_j) / (S_i \cup S_j)$

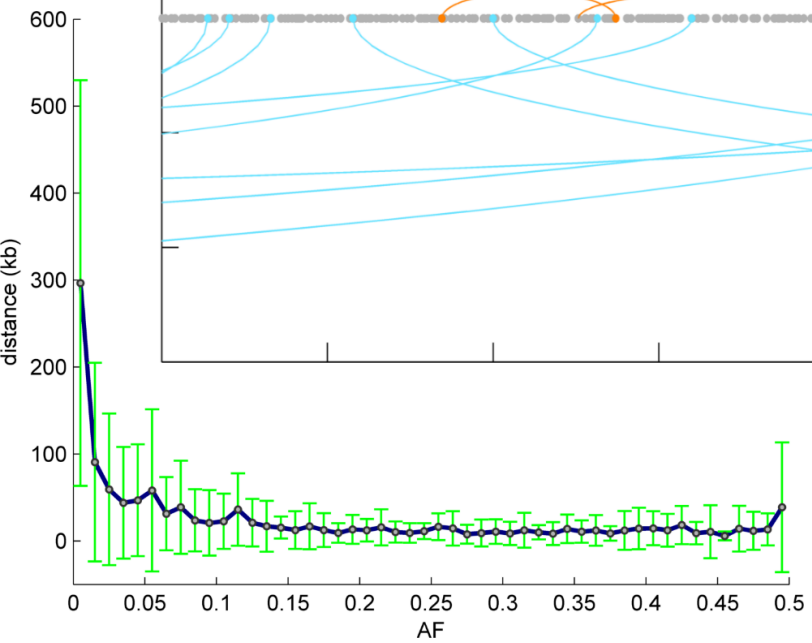
$Sim_2(i, j) = (S_i' \cap S_j') / (S_i' \cup S_j')$

$Sim_3(i, j) = ((S_i' \cap S_j') / (S_i' \cup S_j'))^2$

Genetic distance between NNs



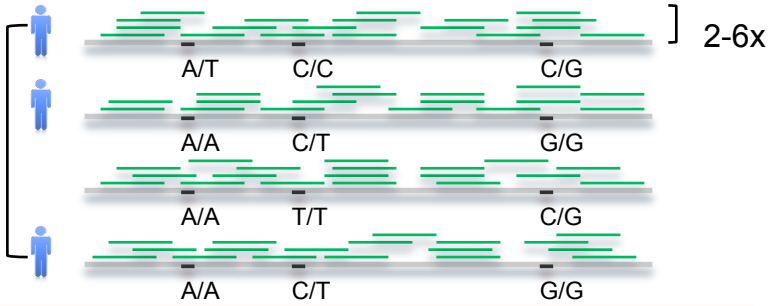




Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



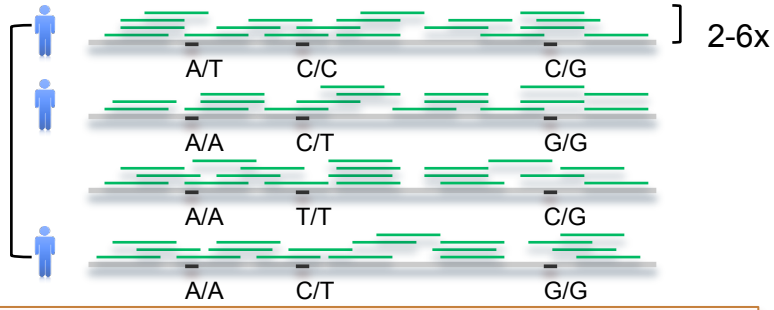
$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

Overview of Reveel

Reveel:

1. **Identify candidate polymorphic sites**
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

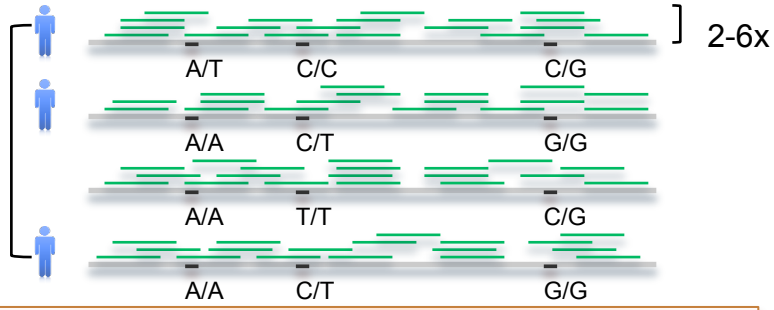
Candidate Polymorphic site

Essentially, pos'n j where some individuals have at least 2 reads with same minor allele

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. **Initialize $G^{(0)}$**
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

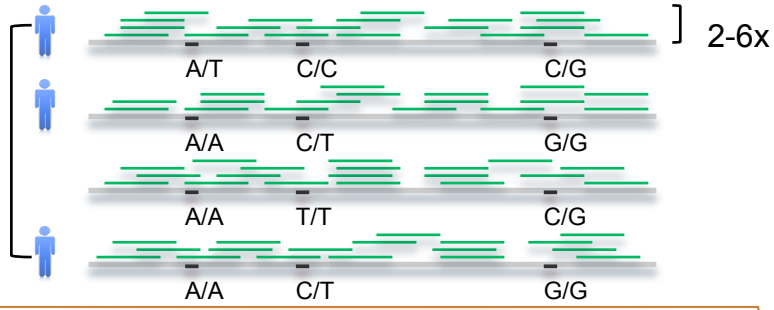
At each position j,

Use sum of read counts at j and its nearest neighbors

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. **Summarization/Maximization**
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

$p^{(n+1)}_{ijg} = P(g_{ij} = g \mid G^{(n)}, reads)$
$\sim P(g_{ij} = g \mid g_{kNN}, reads)$
$= P(reads \mid g_{ij} = g)\ P(g_{ij} = g \mid g_{kNN})$

$P(g_{ij} = g \mid g_{kNN}) =$

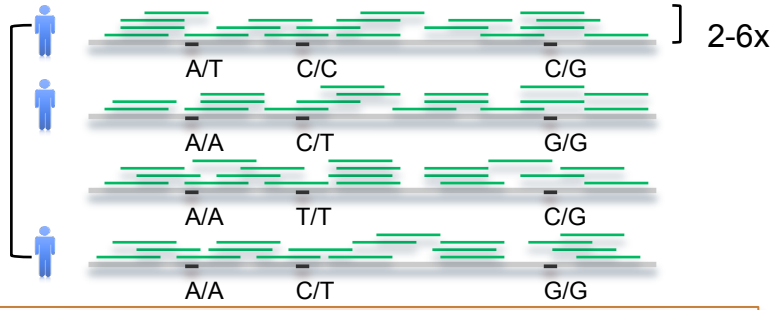
Let C_0, C_1, C_2 = # samples matching i in kNN, with j^{th} genotype pos'n = 0, 1, 2

$P(g_{ij} = g \mid g_{kNN}) = C_g / (C_0 + C_1 + C_2)$

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. **Recalculate k nearest neighbors**
   - **Approximate Correlation Coefficient (Schaid 2004)**
6. **Summarization/Maximization**
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

Correlation Coefficient:
$r^2 = (p_{AB} - p_A p_B)^2 / p_A p_B p_a p_b$

Caveat: need **genotyping**, **phasing**

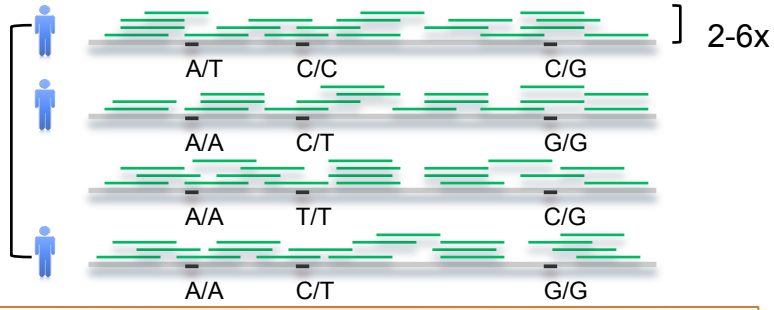
Schaid 2004:
$D = 1/N\ (2N_{AABB} + N_{AABb} + N_{AaBB} + ½N_{AaBb}) - 2p_A p_B$

A faster alternative:
$D = ½\ Sim_1(i, j) + p_A p_B (p_A + p_B - ½\ p_A p_B - 2)$

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. **Recalculate k nearest neighbors**
   - **Approximate CC, Entropy**
8. **Summarization/Maximization**



$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

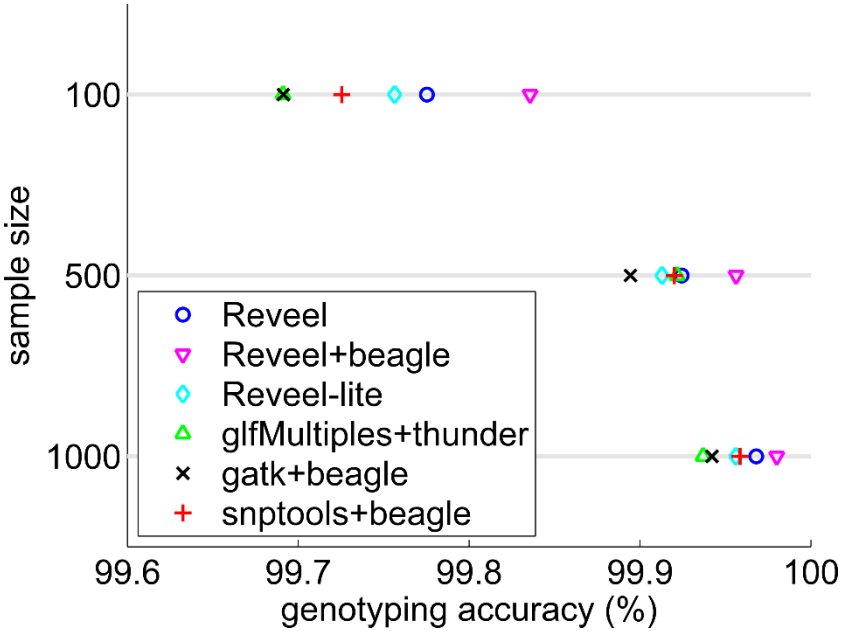
$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

Identify the sites at which decision trees formed by kNNs have high entropy

Reduce entropy by replacing one or more kNNs

Simulations

- Simulated data set: ~ low-coverage 1KGP
  - 2,535 samples, 1-Mbp region: Chr 20 (43-44 Mbp) of GRCh37
  - Using COSI to simulate variations
  - Reads same positions, lengths, qualities as reads in 1KGP
  - Injecting sequencing base errors, mapping with BWA



Sample size	100	500	1000	2535
Reveel	1.8	14.6	47.4	273
Reveel+Beagle	3.1	25.3	71.2	526
Reveel-lite	1.5	7.8	21.7	145
SNPTools+Beagle	8.2	217	1089	>5days
GATK+Beagle	13.4	388	1806	>5days
glfMultiples+Thunder	307	2736	6120	~15 days

Genotyping Accuracy

Computation Time (mins)

Performance on 1000 Genomes Data

population	# of samples in 1KGP	# of samples in HapMap3	population	# of samples in 1KGP	# of samples in HapMap3
ACB	96	0	**ASW**	66	**50**
BEB	86	0	CDX	99	0
CEU	99	**90**	**CHB**	103	**94**
CHS	108	0	CLM	94	0
ESN	99	0	FIN	99	0
GBR	92	0	**GIH**	106	**93**
GWD	113	0	IBS	107	0
ITU	103	0	**JPT**	104	**97**
KHV	101	0	**LWK**	101	**90**
MSL	85	0	**MXL**	67	**56**
PEL	86	0	PJL	96	0
PUR	105	0	STU	103	0
TSI	108	**96**	**YRI**	109	**103**

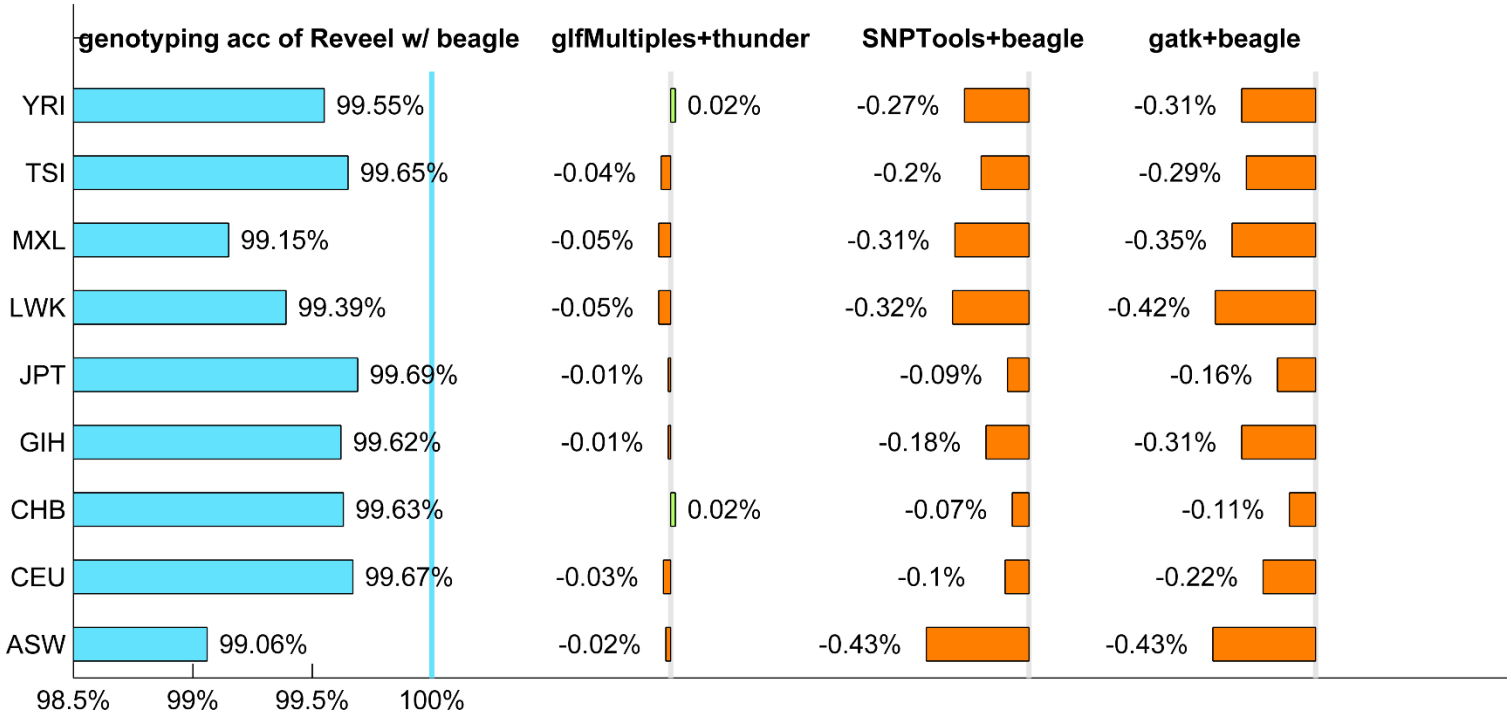
Compared on a 5-Mbp region on chromosome 20 (43-48 Mbp)

Performance on 1000 Genomes Data

HapMap 3 benchmark

769 invididuals from 9 populations

AF ≥ 5%,	2686 SNPs
5% > AF ≥ 1%,	368 SNPs
AF < 1%,	32 SNPs



Performance on 1000 Genomes Data

- SNP discovery
  - Discover 171,734 likely polymorphic sites in 26 populations
  - Benchmark: Complete Genomics data

Method	false positive rate	sensitivity
Reveel	0.021%	95.80%
Reveel+Beagle	0.020%	95.92%
Reveel-lite	0.021%	95.80%
GATK+Beagle	0.035%	95.62%
glfMultiples+Thunder	0.037%	95.80%
SNPTools+Beagle	0.035%	95.66%
GotCloud (w/ filters)	0.007%	91.29%
Integrated (w/ filters)	0.011%	95.89%

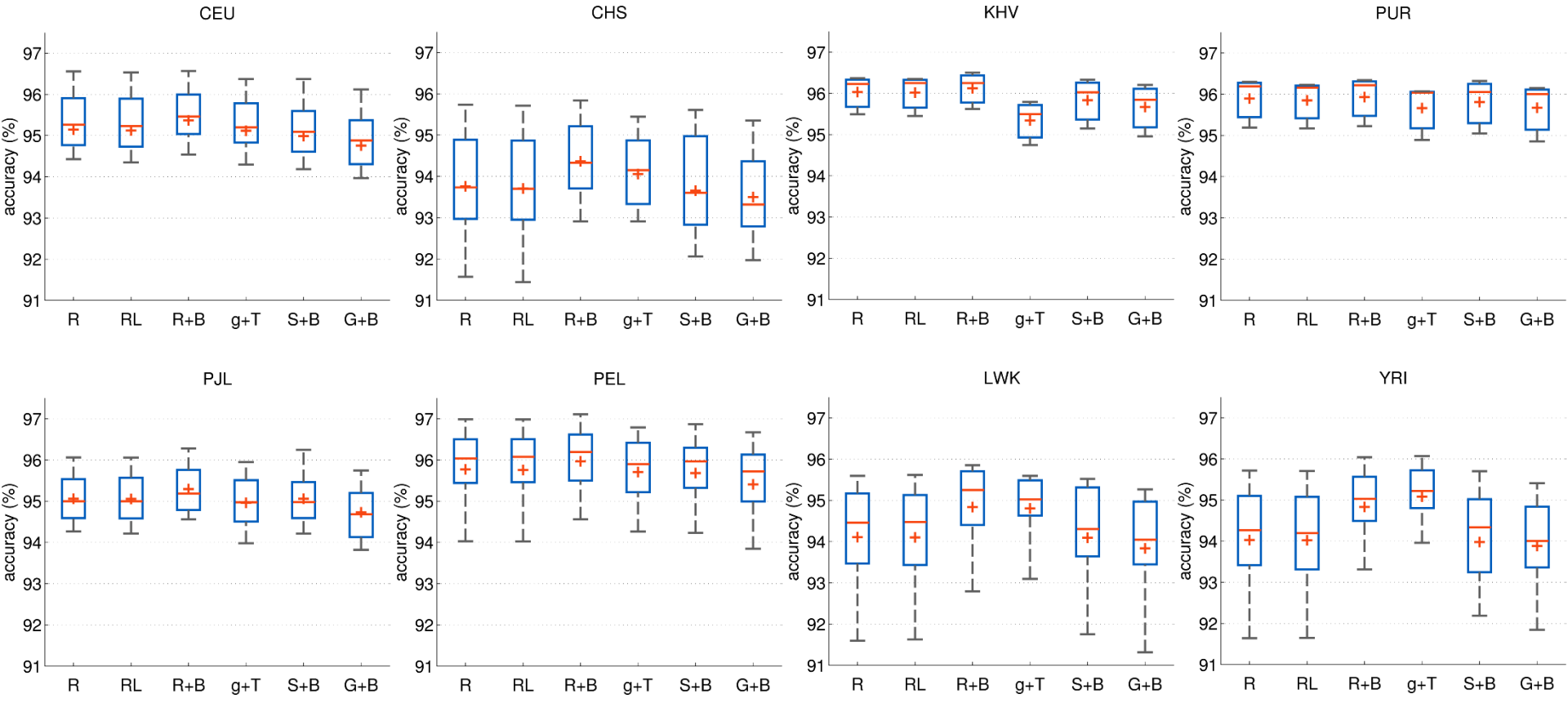
Performance on 1KGP Trios

- SNP discovery
  - Discover 171,734 likely polymorphic sites in 26 populations
  - Benchmark: 1KGP Pilot2 Trios

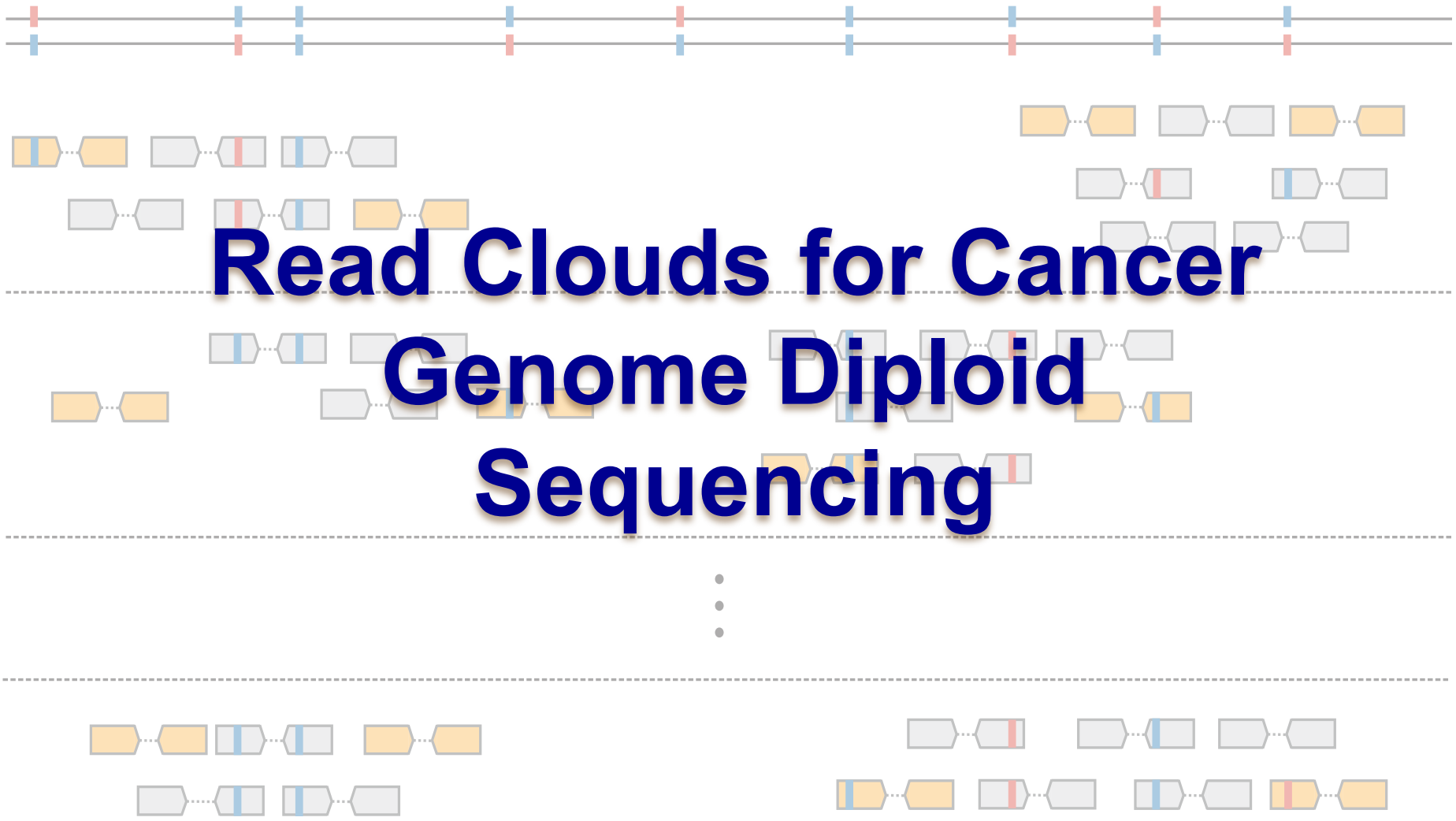
Method	false positive rate	sensitivity
Reveel	0.031%	97.06%
Reveel+Beagle	0.031%	97.53%
Reveel-lite	0.031%	97.06%
GATK+Beagle	0.040%	97.38%
glfMultiples+Thunder	0.048%	98.17%
SNPTools+Beagle	0.044%	96.81%
GotCloud (w/ filters)	0.011%	91.46%
Integrated (w/ filters)	0.023%	98.32%

Performance on 1000 Genomes Data

- Genotyping
  - Benchmark: Complete Genomics samples

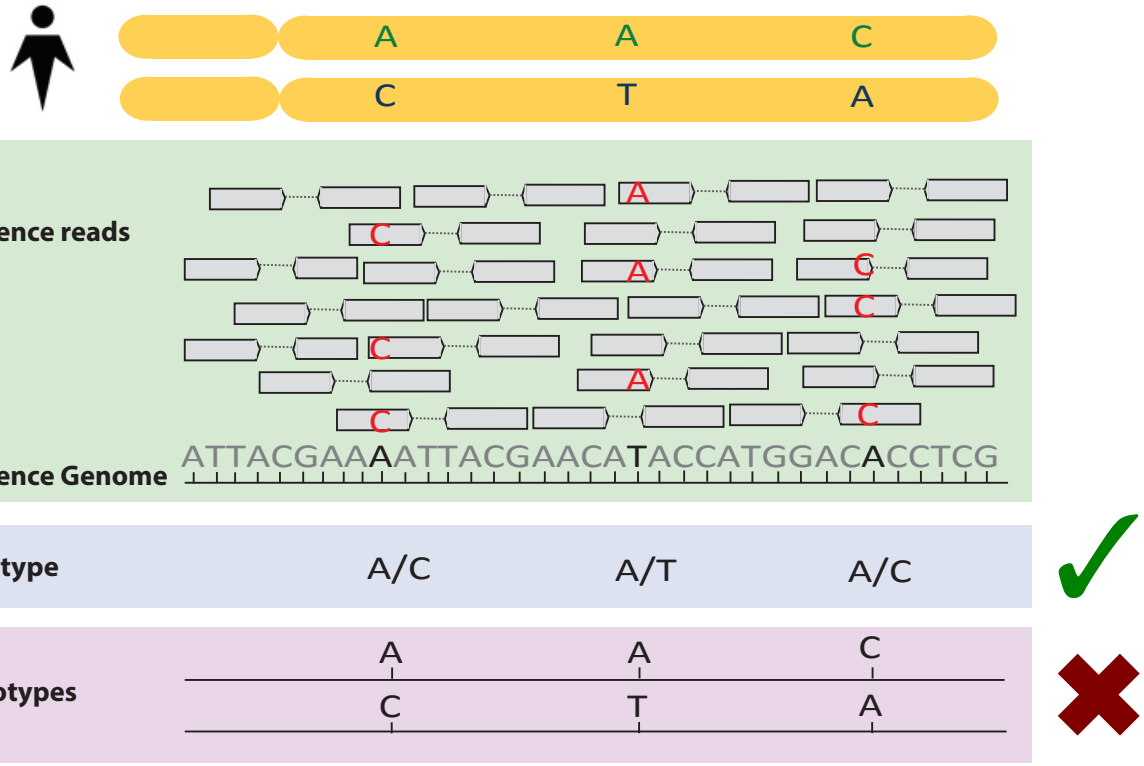


R: Reveel; RL: Reveel-lite; R+B: Reveel+Beagle; g+T: glfMultiples+Thunder; S+B: SNPTools+Beagle; G+B: GATK+Beagle

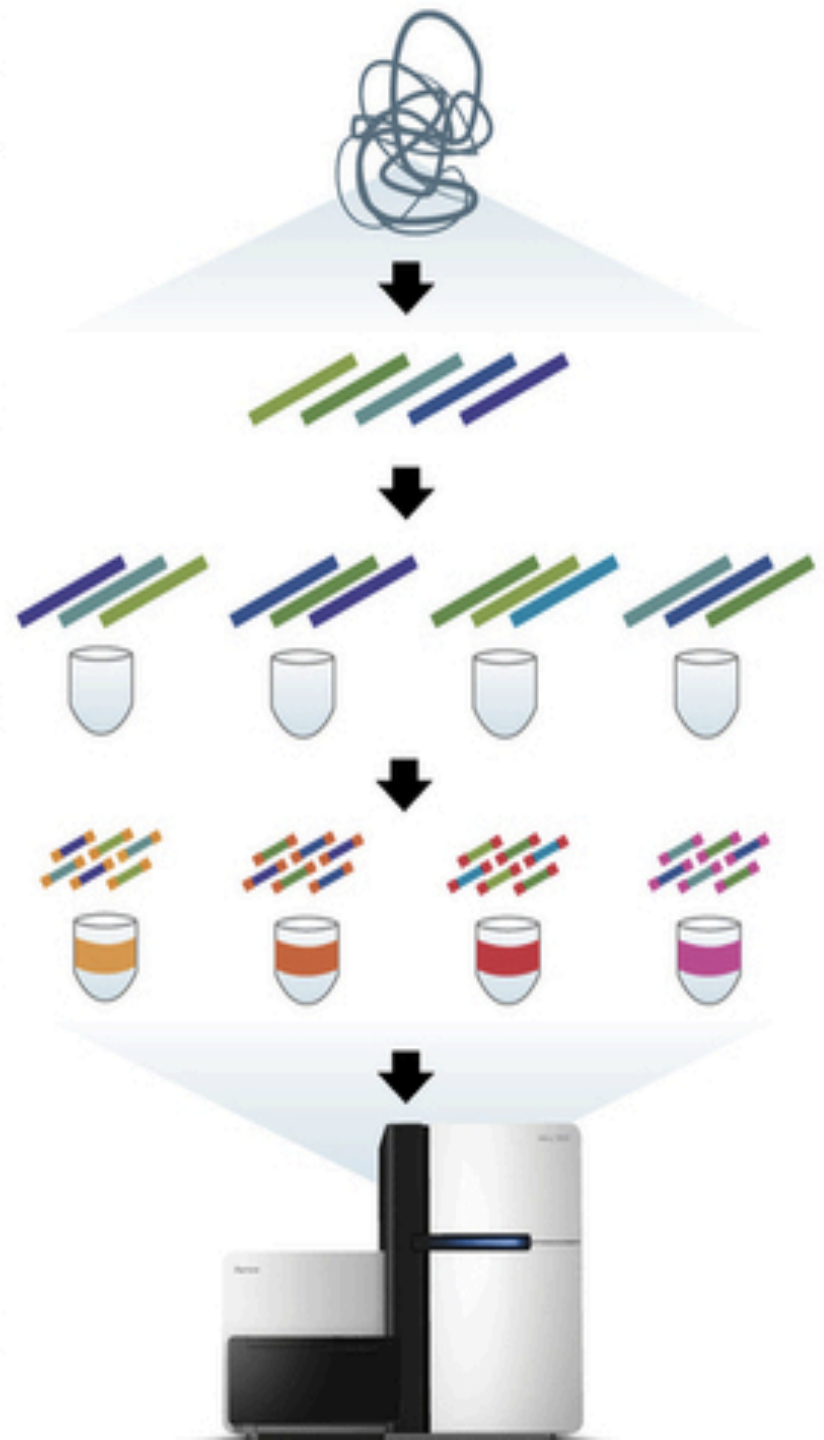


Read Clouds for Cancer Genome Diploid Sequencing

Shortcomings of Short Reads



- Unphased genotypes
- No variants detected in high-fidelity repeats (6% of genome)
- Low-accuracy structural variants
- Challenging regions such as HLA



Moleculo Overview

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

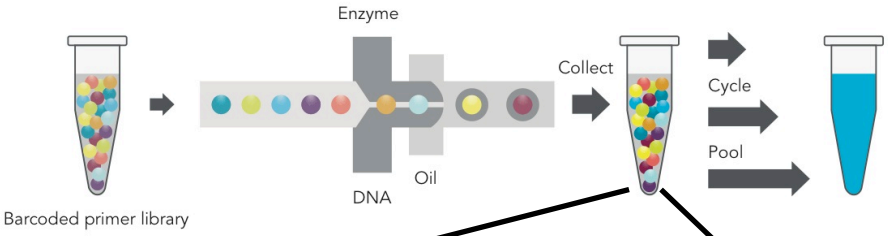
2. Fragments are diluted and placed into 384 wells

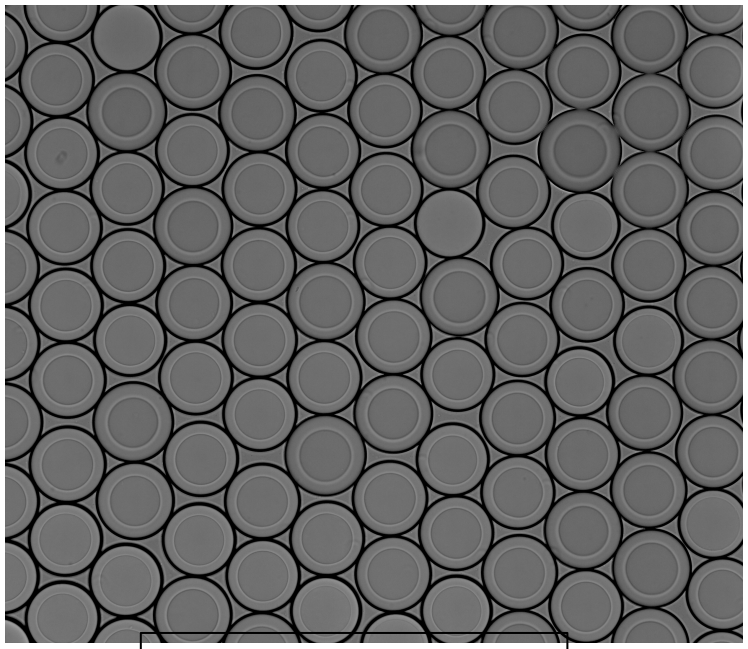
3. Fragments are amplified through long-range PCR, cut into short fragments and barcoded

4. Short fragments are pooled together and sequenced

10x System

Massively Parallel Partitioning

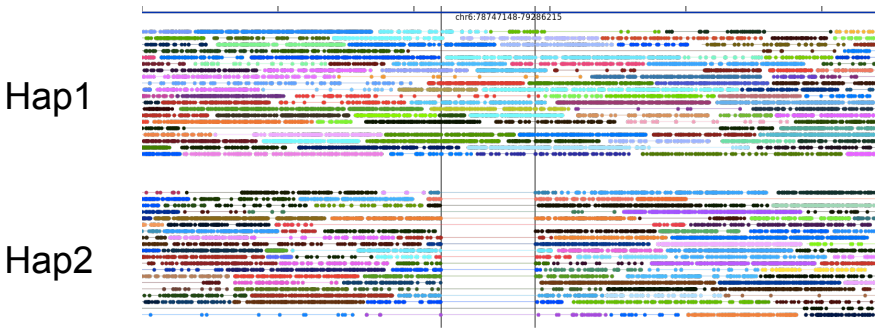




X 200,000+

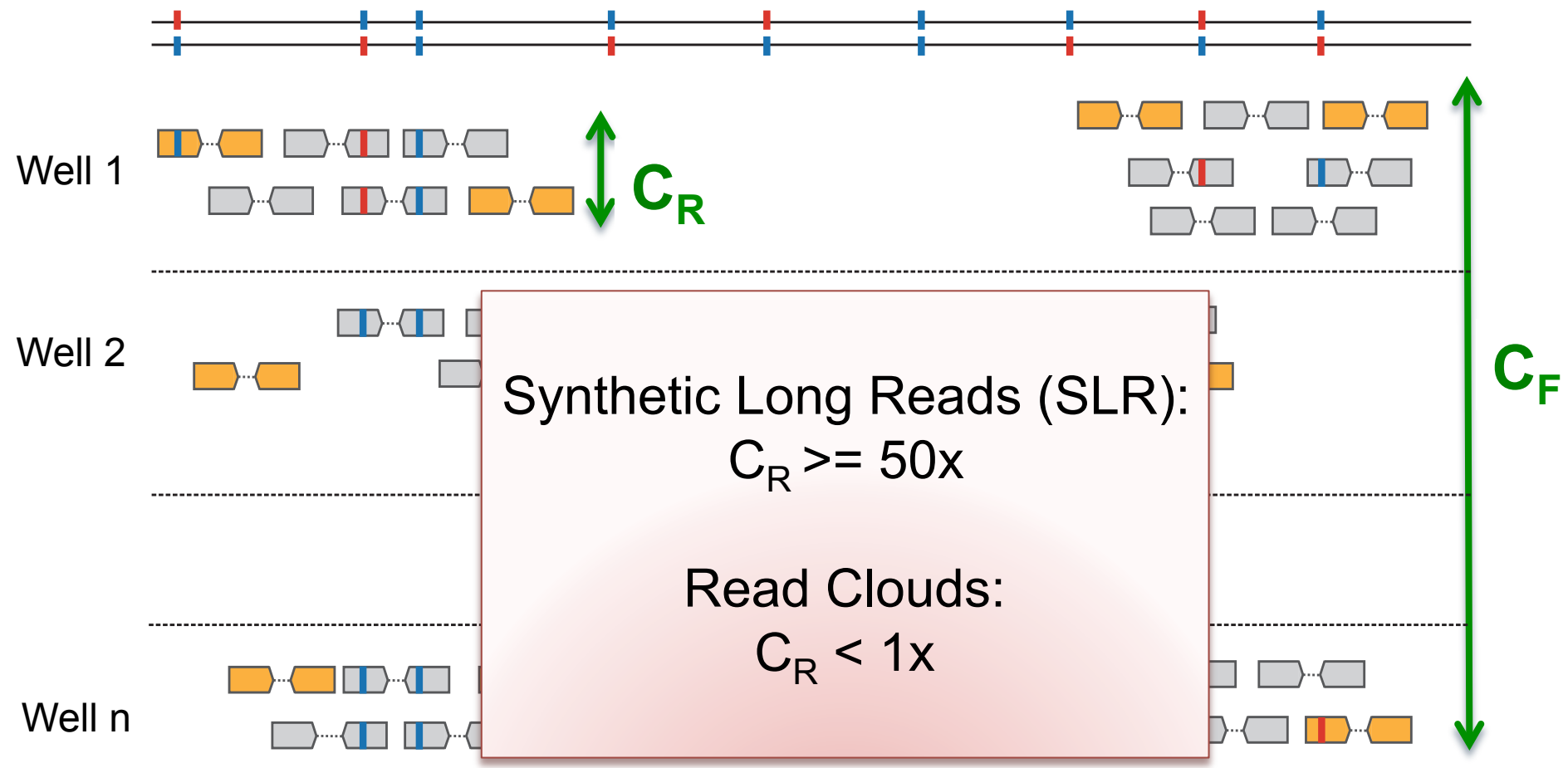
10X Instrument & Reagents

Read Clouds (“linked reads”)



Phased 60Kb deletion

Read Clouds



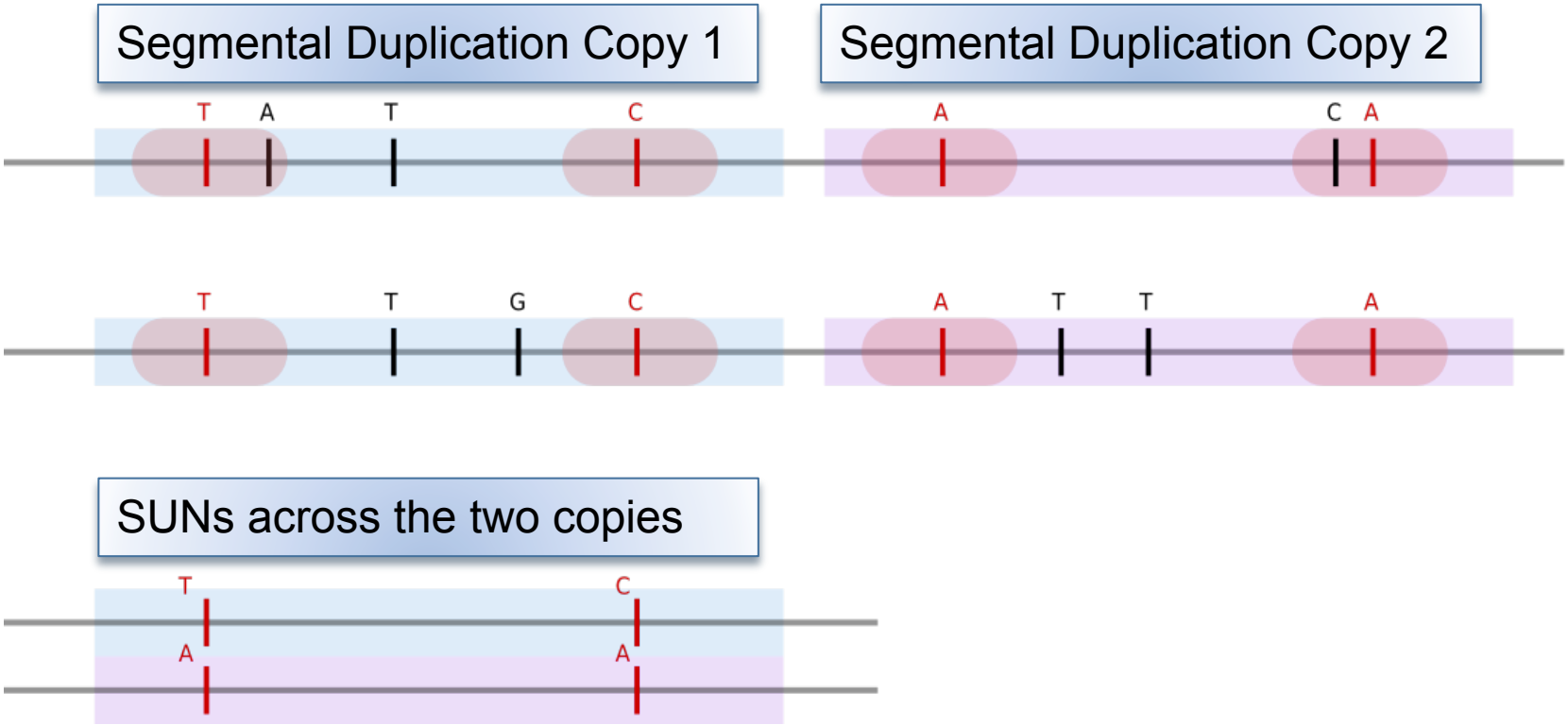
Coverage = $C_F C_R$

Identifying Variation in Segmental Duplications

~180 Mbp of human in almost exact repeats

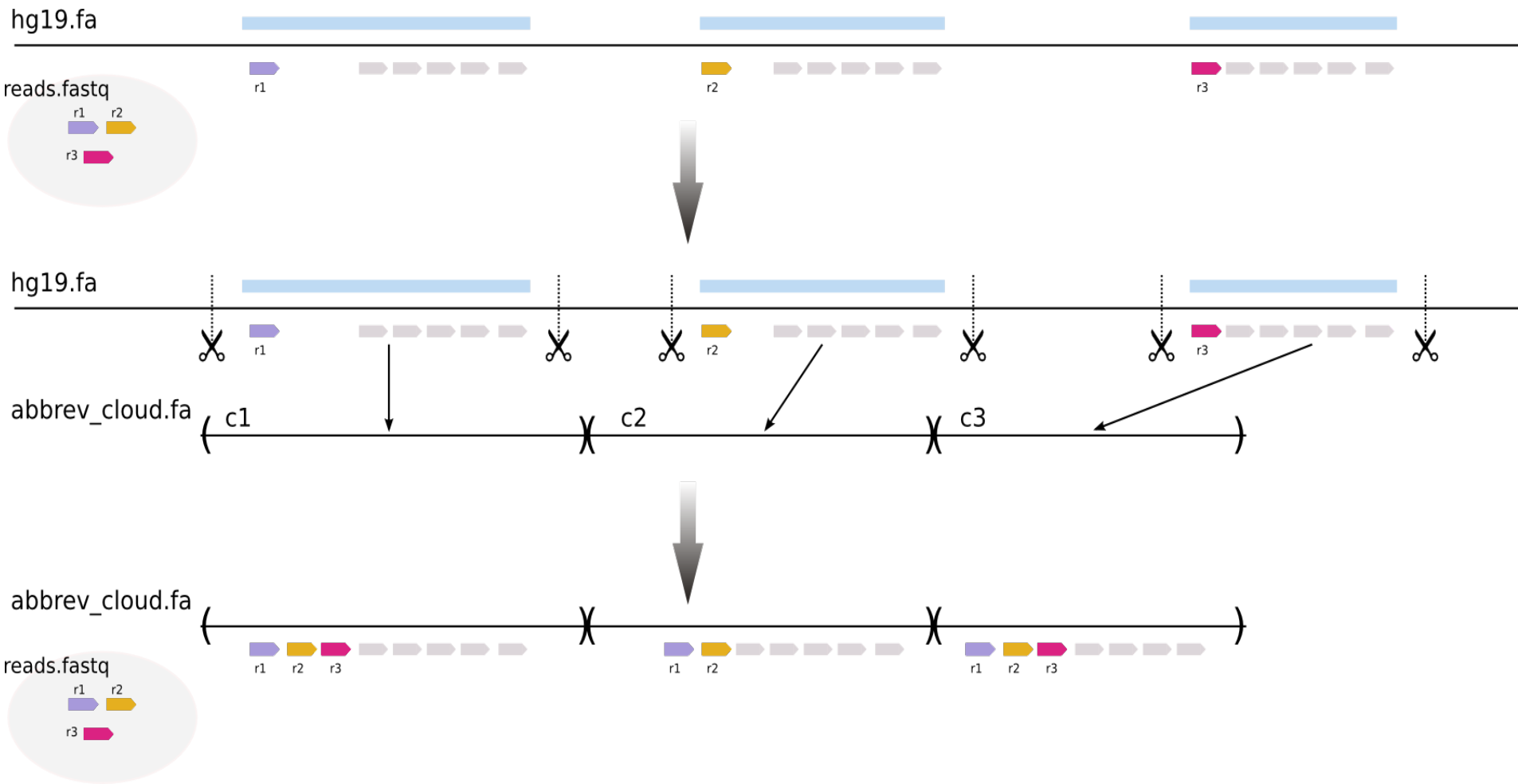
Novel variation impossible to detect with short reads alone

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



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

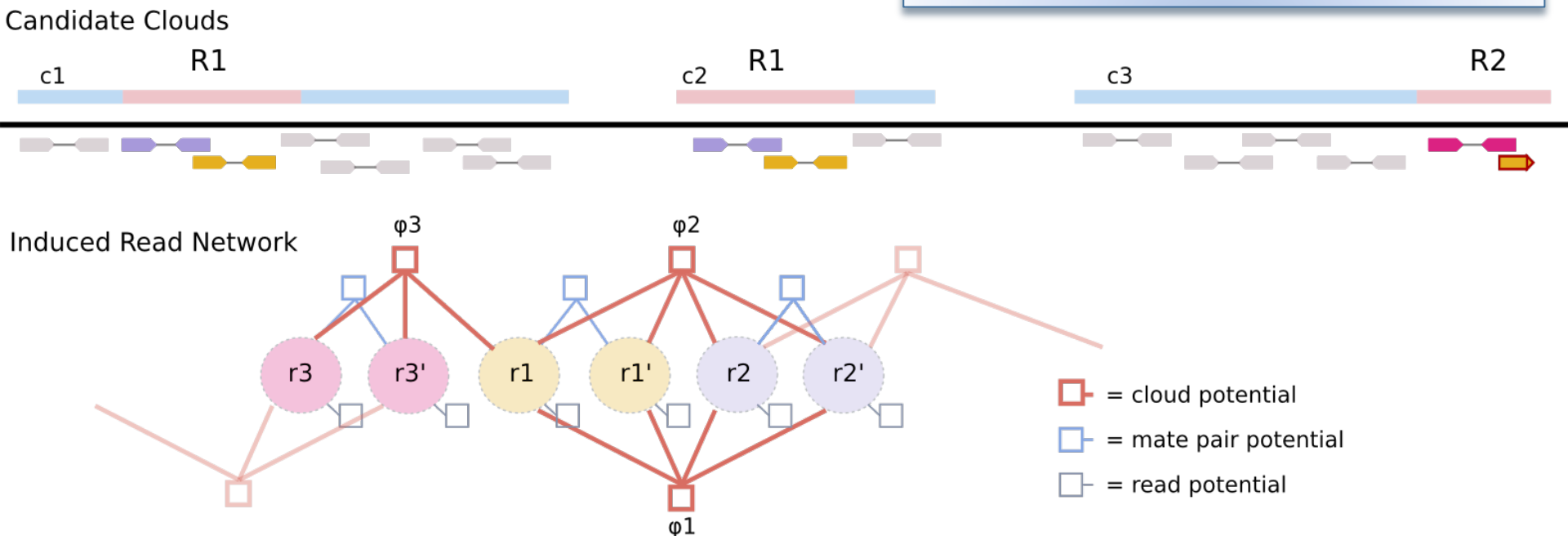
$$P(R) = \sum_{\{\text{Underlying molecule set M}\}} P(M)\, P(R \mid M)$$

Markov Random Field (MRF)

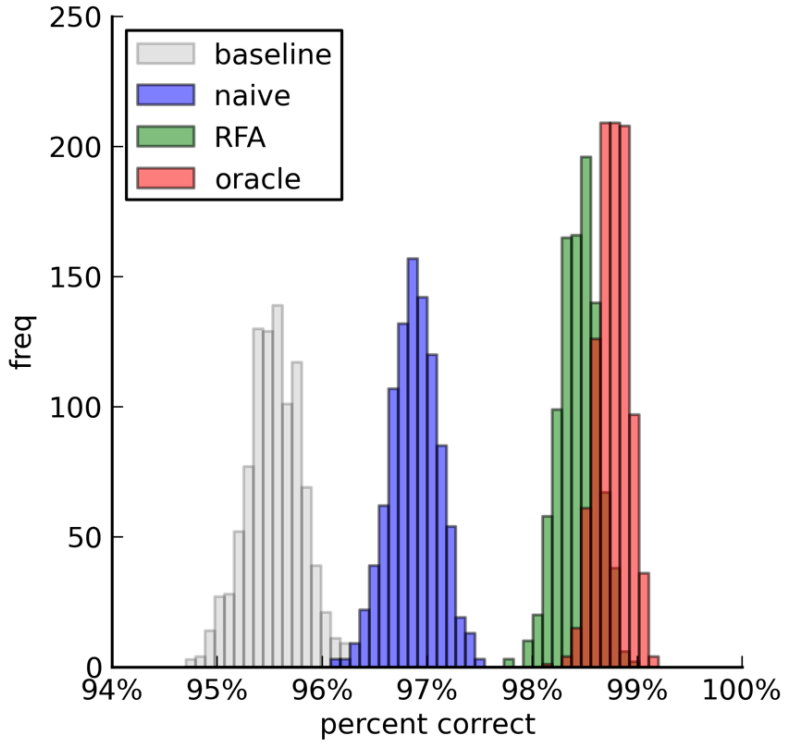
- Read alignment quality
- Mate pairs biased to map together
- Reads biased to form clouds

Optimize with Simulated Annealing

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



Validation on Simulations



Previously Dark Regions

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

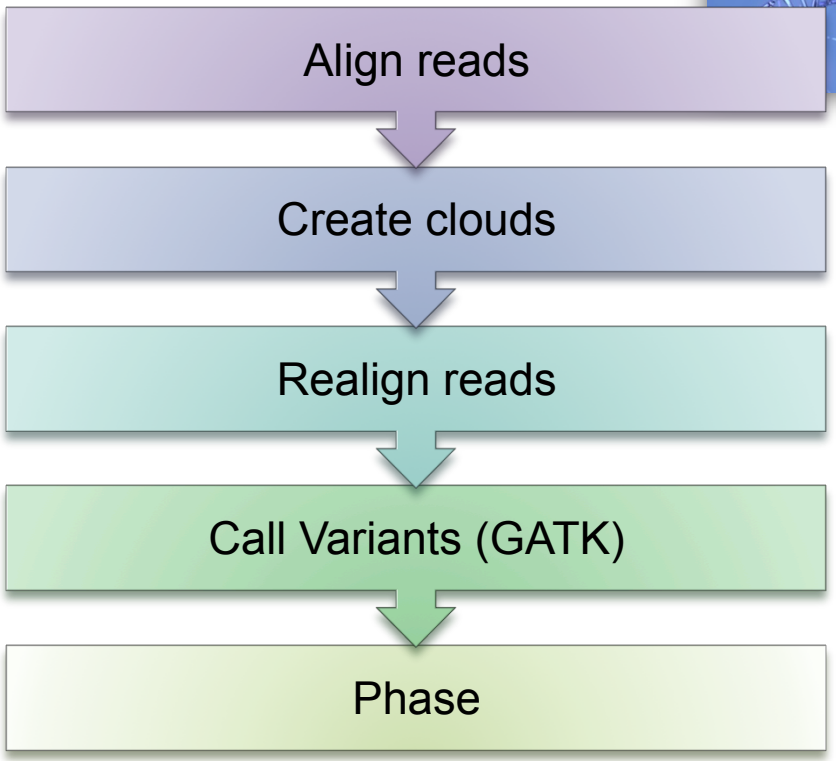
Naive: Naive policy of using alignments to abbreviated reference

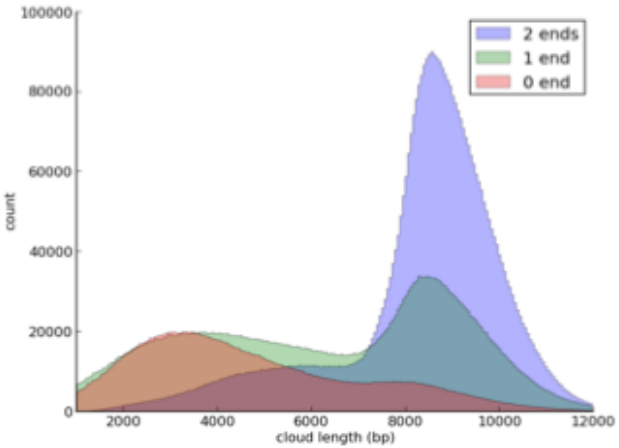
Oracle: Pick the true alignment among Bowtie2 alignments

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

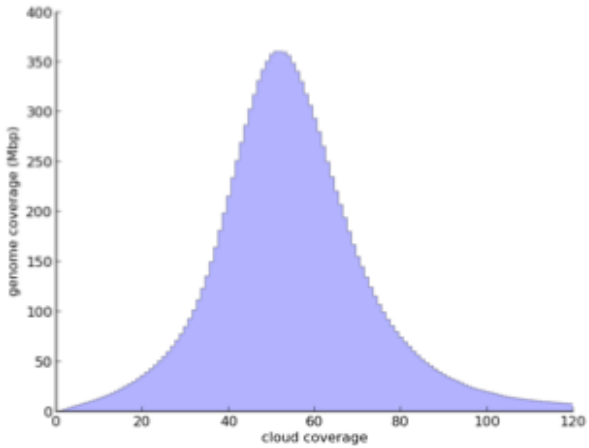
- IDC (Fresh Frozen)
- Sequenced by:
  - Shotgun (40X)
  - Moleculo (78X)

C_F=43, C_R=1.8





(a) Read cloud sizes

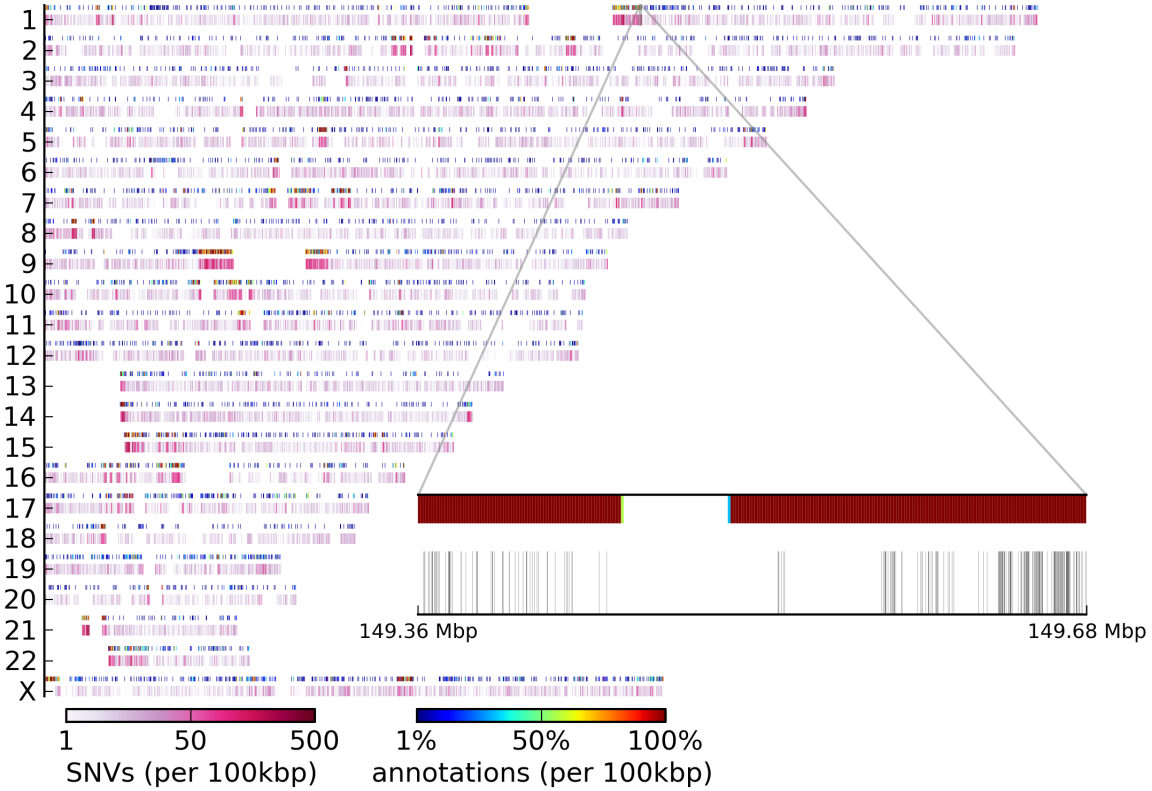


(b) Read Cloud Genome Coverage

Results on IDC Sample - SNVs

Novel SNVs within 6% dark genome:

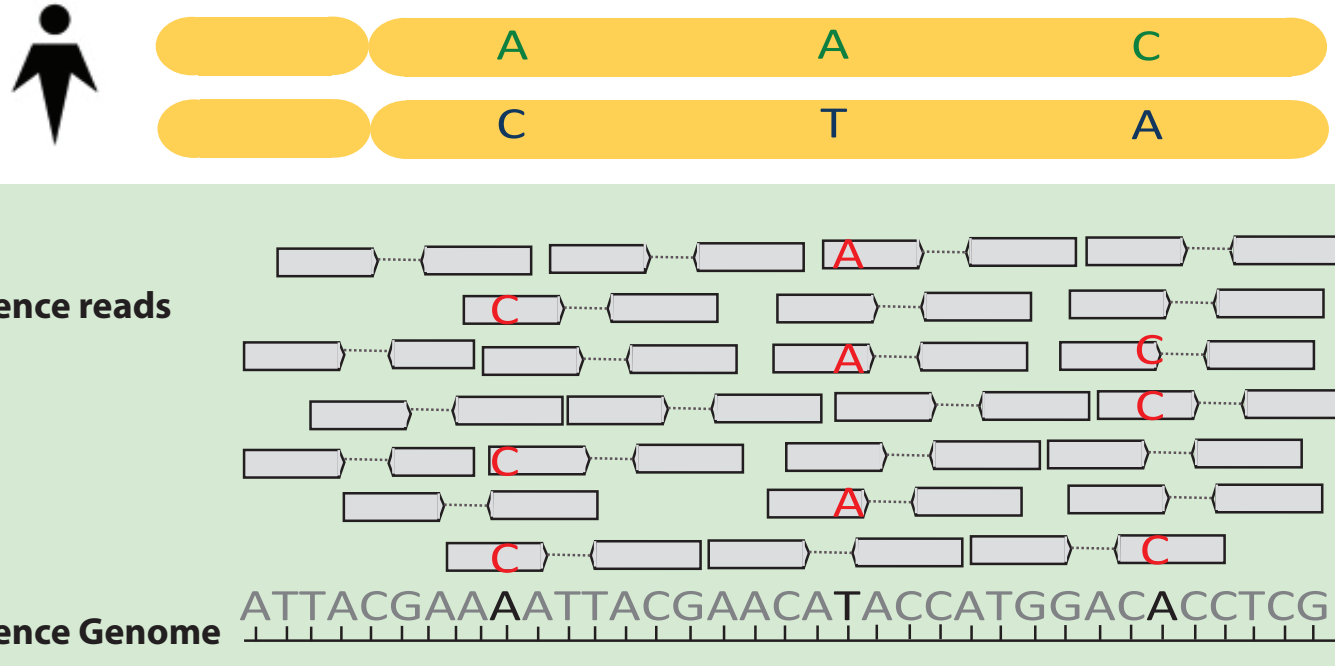
131,957 Heterozygous, 65,529 Homozygous



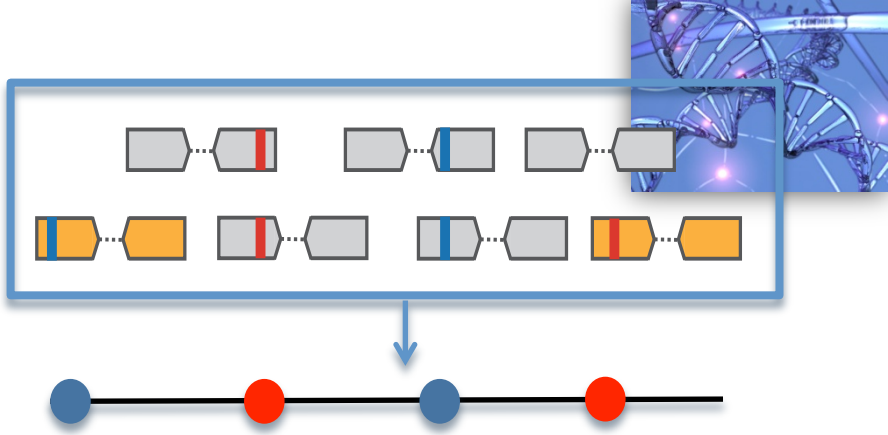
97% of segdups mappable

Multiplex PCR 346 SNVs, **94% validated**

Phasing – an MCMC approach

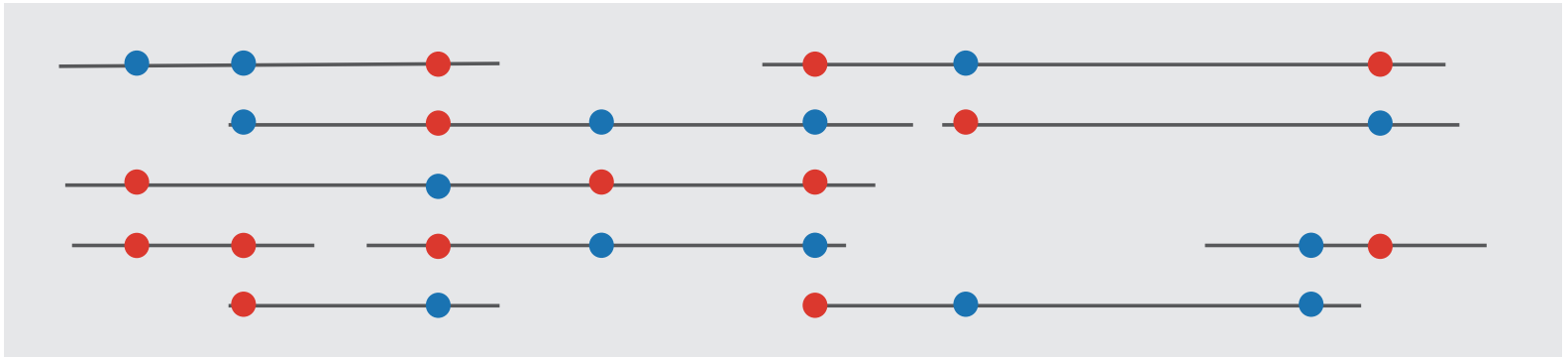


Germline		Somatic
Het	Hom	
1,948,144	1,266,460	4,689

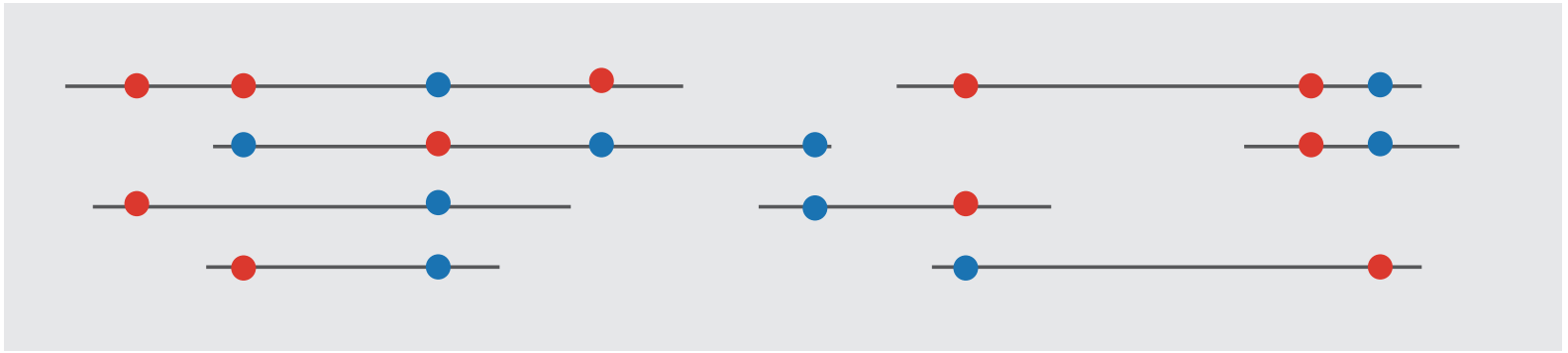


Diploid Genome

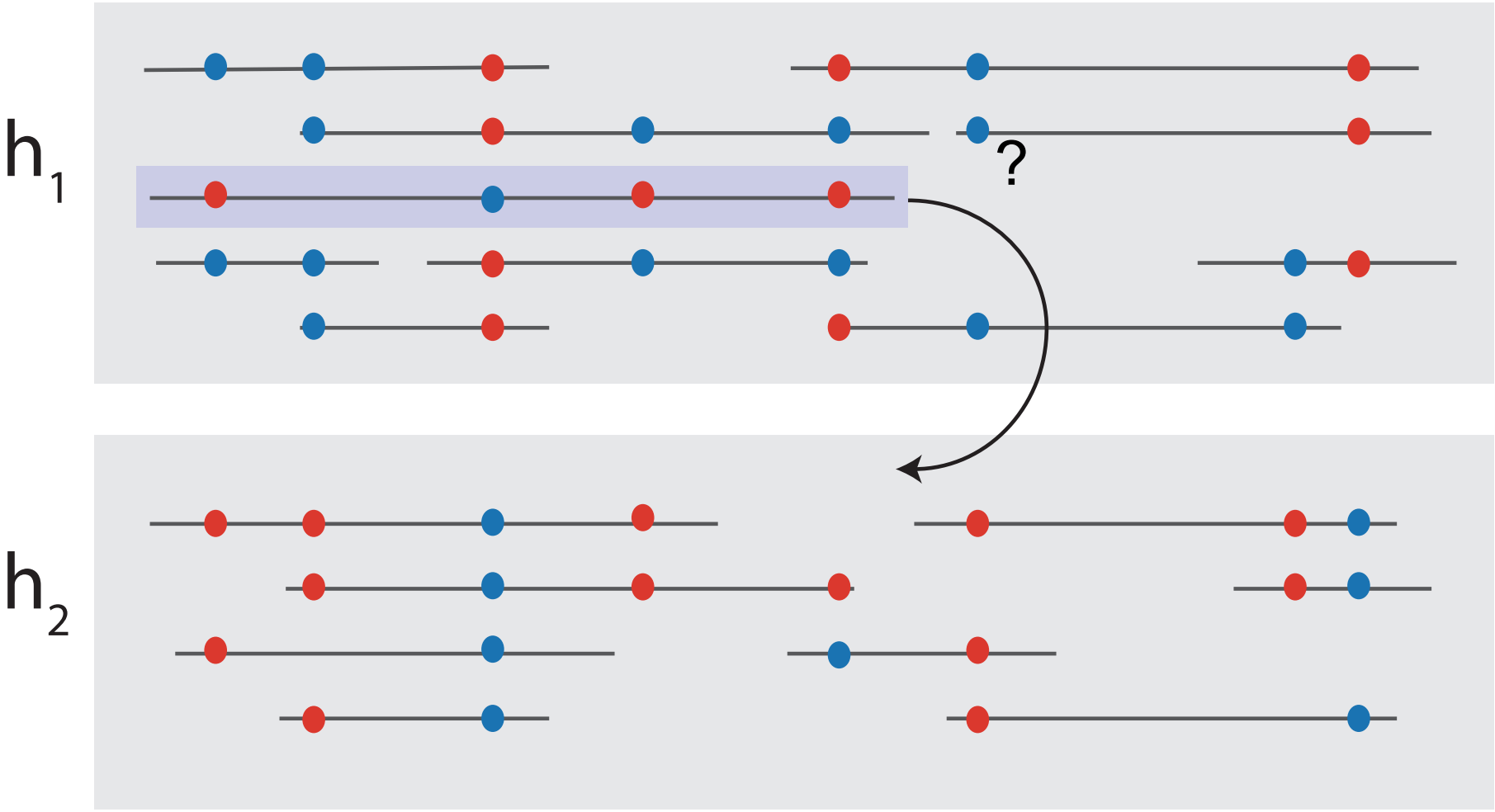
h_1



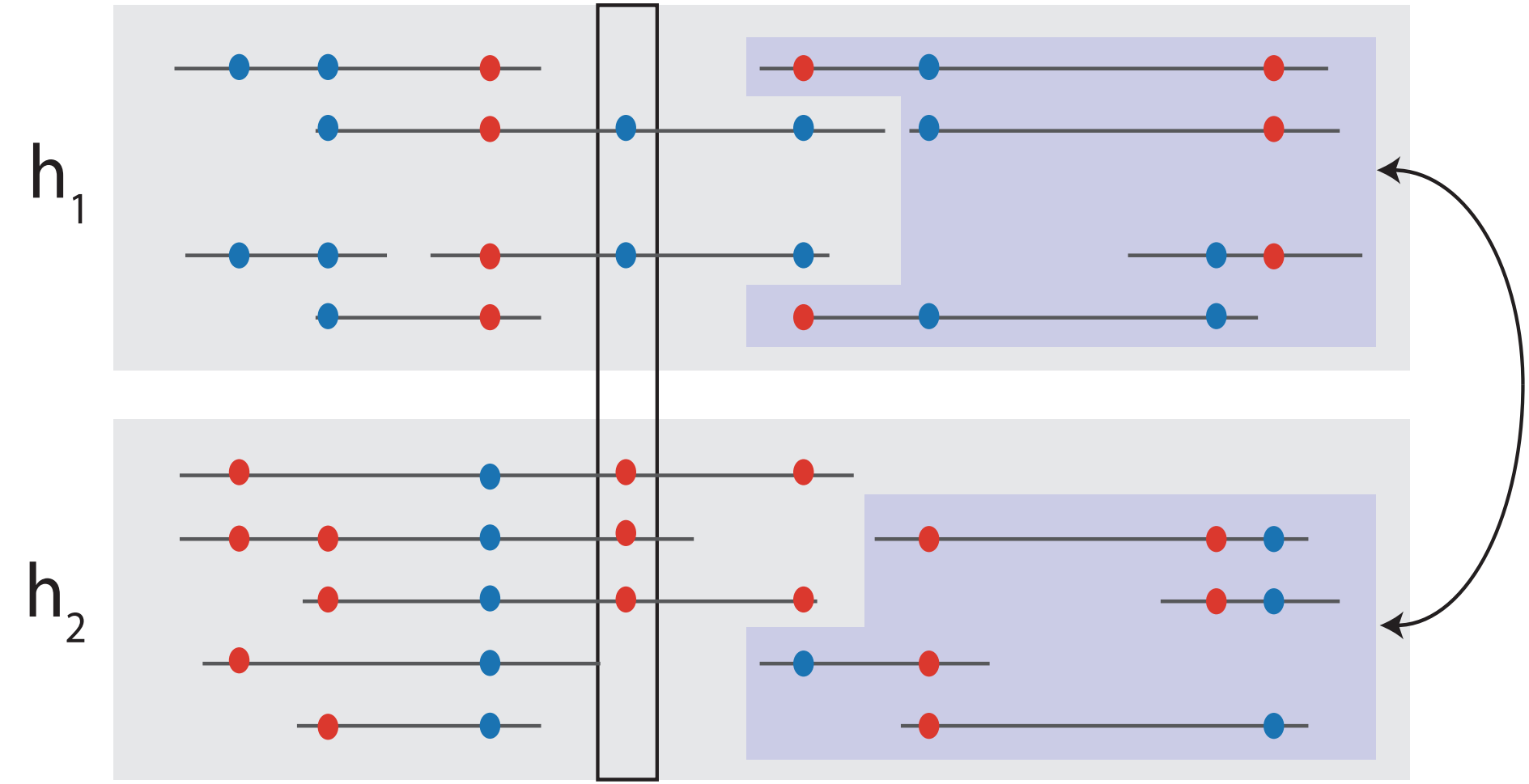
h_2



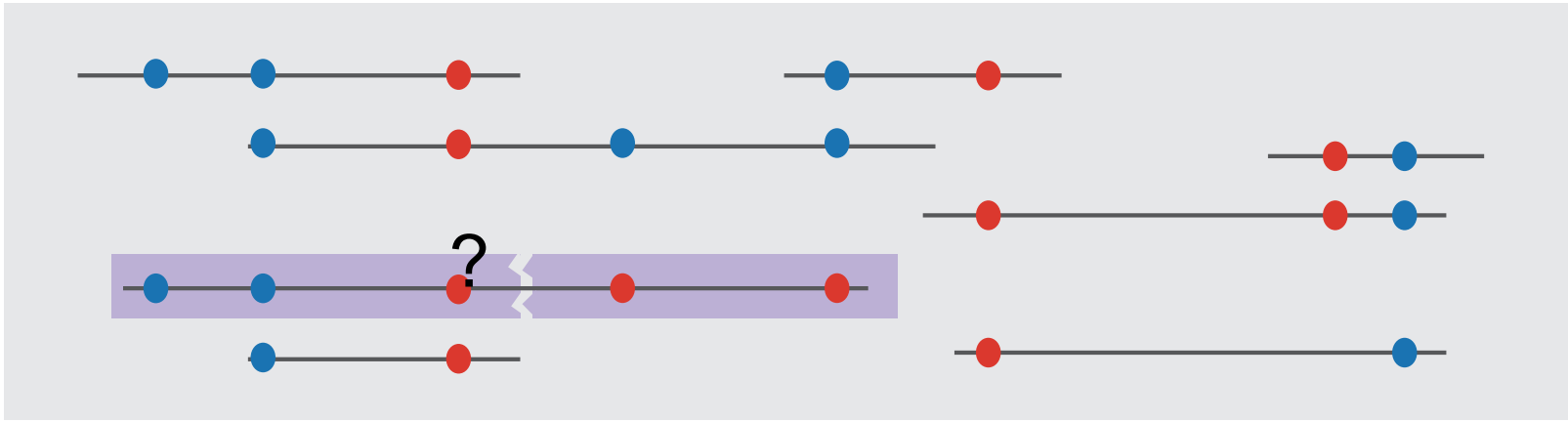
Move 1: assign a cloud to a haplotype

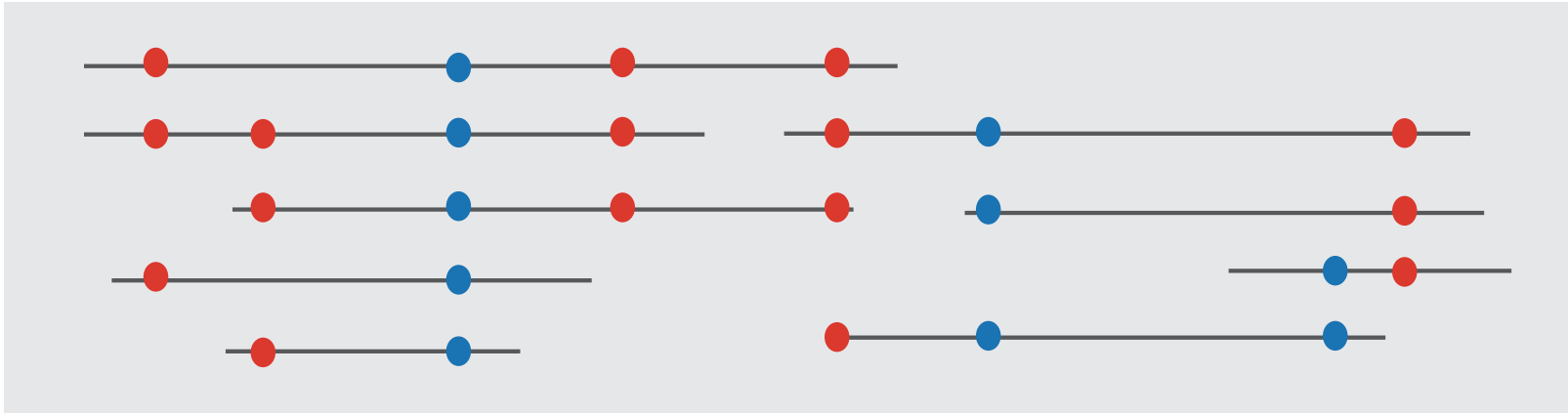


Move 2: unwind a switch error

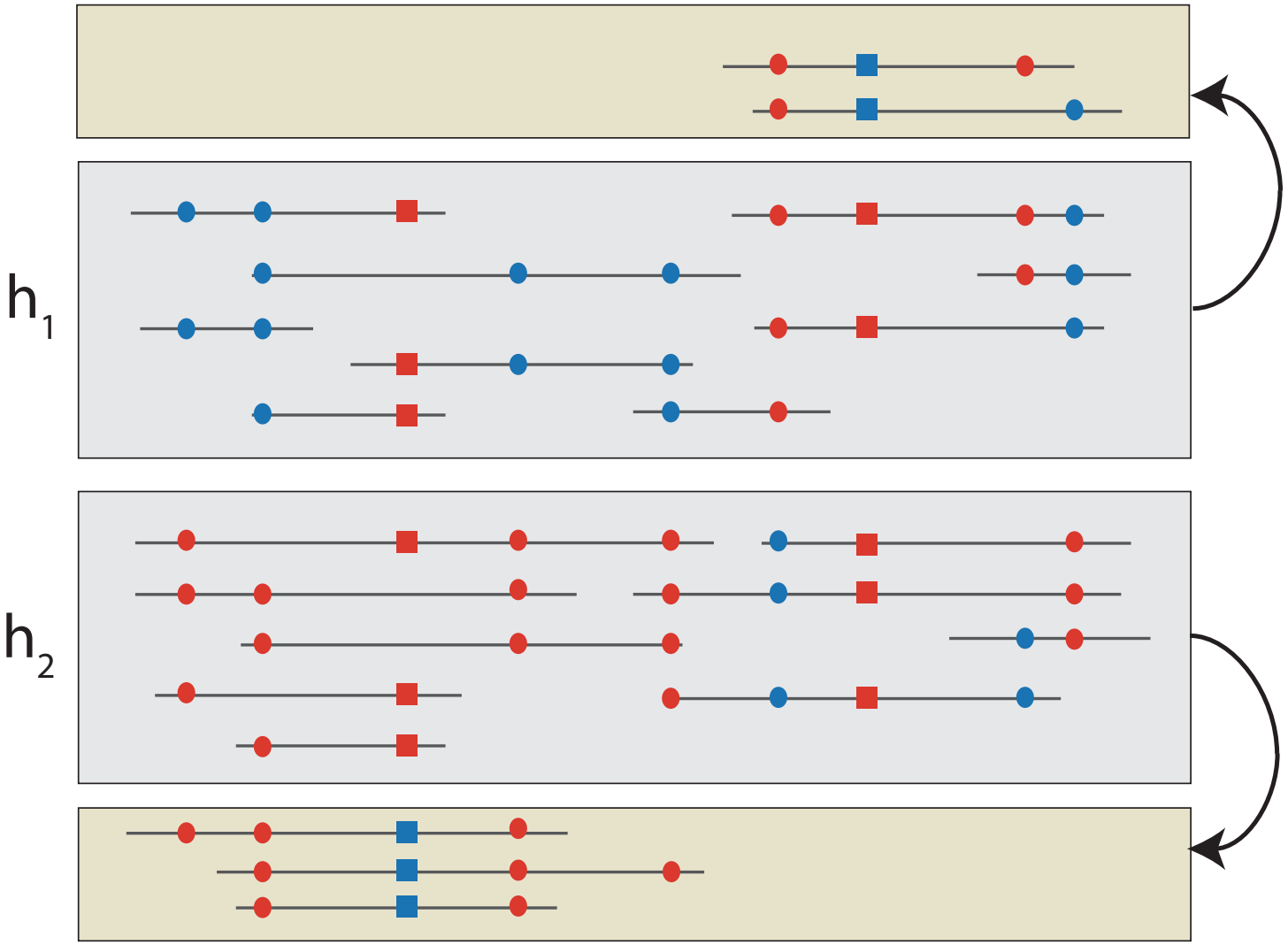


Move 3: flag/unflag clouds as mixed

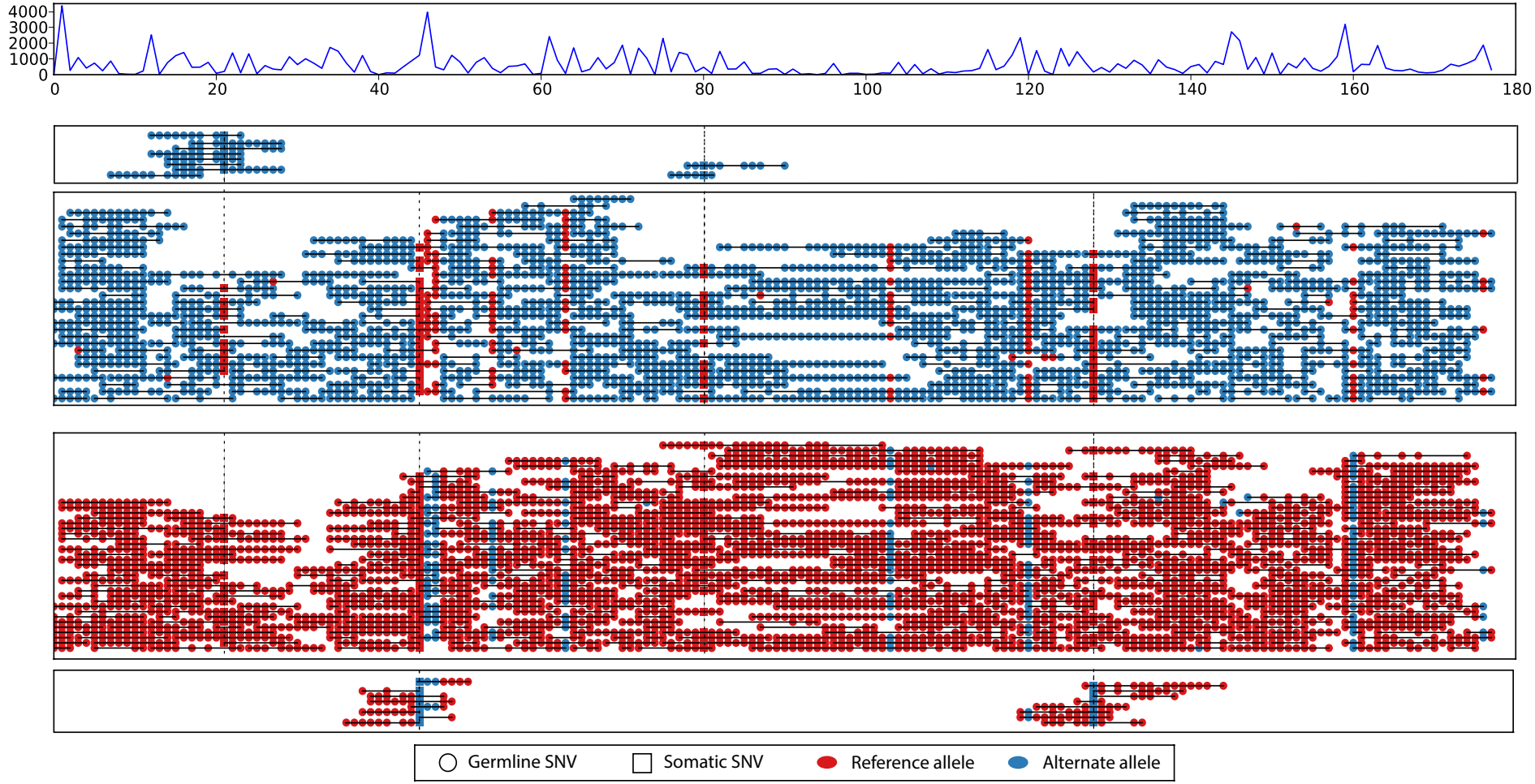




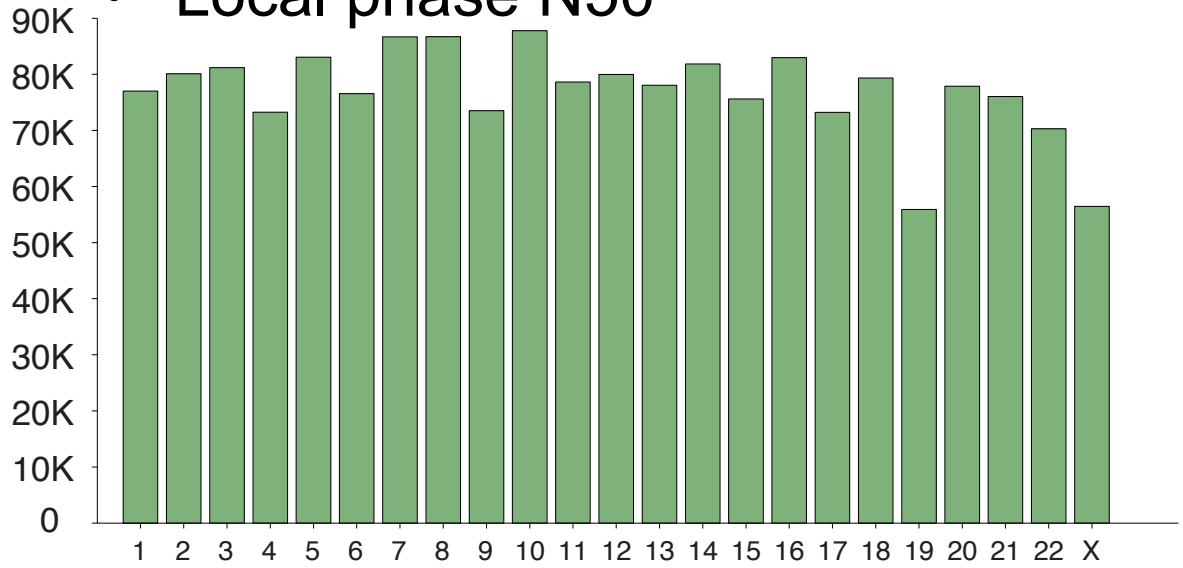
Separate somatic haplotypes

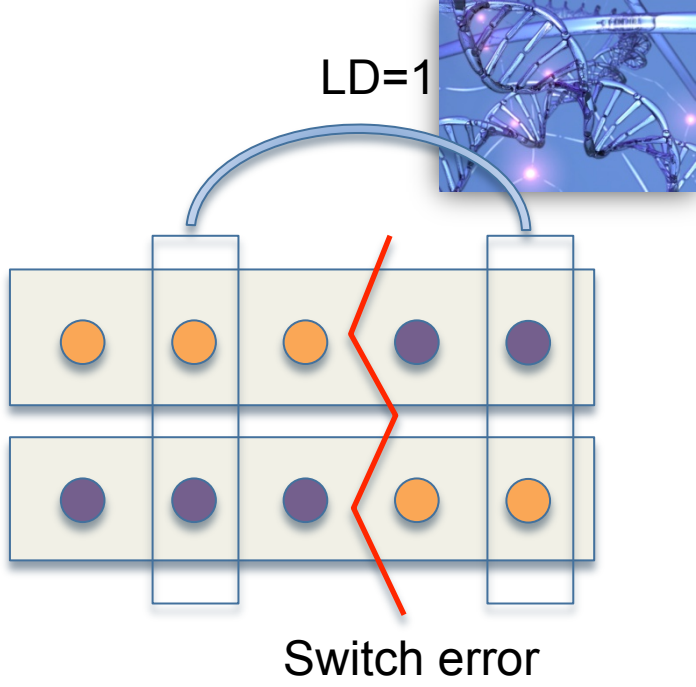


Sample Output



- Local phase N50

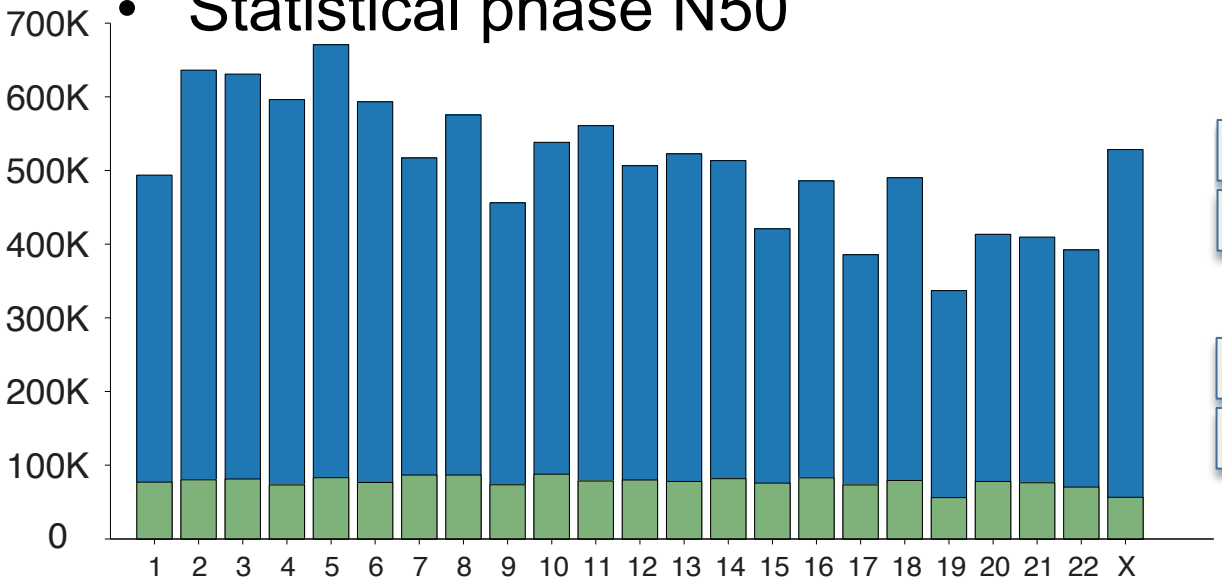


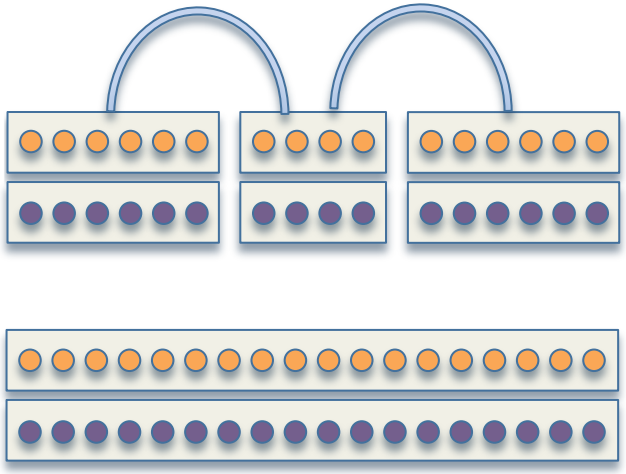


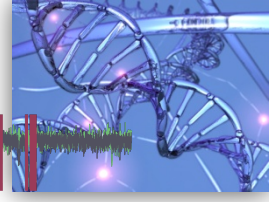
- Switch Errors

2	-	-	-	-	1	1	-	-	-	1	2	1	-	-	-	2	1	-	-	-	-	-

- Statistical phase N50







Allele Fraction

0.7
0.5
0.3
0.1

1p 1q 2p 2q 3p

9q

14 15 16p 16q 17p 17q 18p 18q 19 20p 20q

21 22 Xp Xq

Normal
IDC

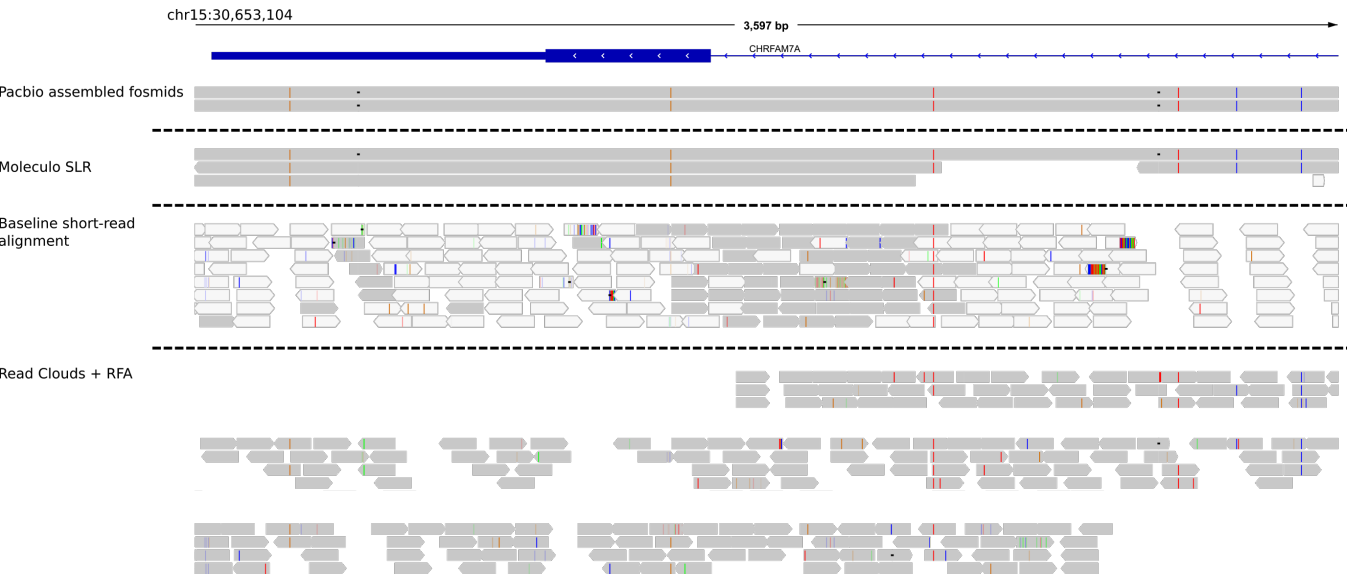
Phase statistics

	Local	LD	**Aneuploidy**
N50	80 Kbp	533 Kbp	**26 Mbp**

Moleculo on NA12878 - SNVs

- 3 lanes of Moleculo sequencing by Illumina
  - C_F = 21x; C_R = 0.8x on average
- Within 6% repeat DNA, where coverage is sufficient:
  - **50,314** novel mutations (35,092 heterozygous)

  Of which, within Moleculo SLR assembled regions:
  - 9,651 homozygous, 99.5% validated (>=2 SLR reads)
  - 24,333 heterozygous, 92.0% validated



Antonnaci, Eichler et al. 986kbp BACs within SDs of high structural variation:

RFA calls 301 novel SNVs, of which:

126 homozygous
97.6% in BACs

175 heterozygous
53% in BACs

10X on NA12878 – Phasing

Mean Coverage	63X
C_F , C_R	200x, 0.3x
N50 Phase Block	20.6Mb
SNP Short Switch	0.3%
SNP Long Switch	0.001%
% SNPs Phased	97.0%
% Genes Fully Phased (<100kb)	94.4%

10X on NA12878 – SNV detection

RFA-10X

- 10X team, with help from Alex Bishara
  - Fast version of RFA, part of 10X suite

SNVs: 4,474 k (1,666 k hom)

Recovered: 238 k (39 k hom)

Baseline: Illumina Platinum Genomes, 200x

	Moleculo SLR	Eichler BACs	
Region Overlap	130,870	749	
Homozygous Overlap	99,6%	100%	100%
Heterozygous Overlap	88.2%	52.9%	66.9%

	% of SNVs
Segmental Duplications	54%
LINE	22.3%
SINE	9.5%
Genes	7.8%
LTR	7.5%
Satellite	5.5%
Simple repeats	1.3%
Low complexity	0.4%

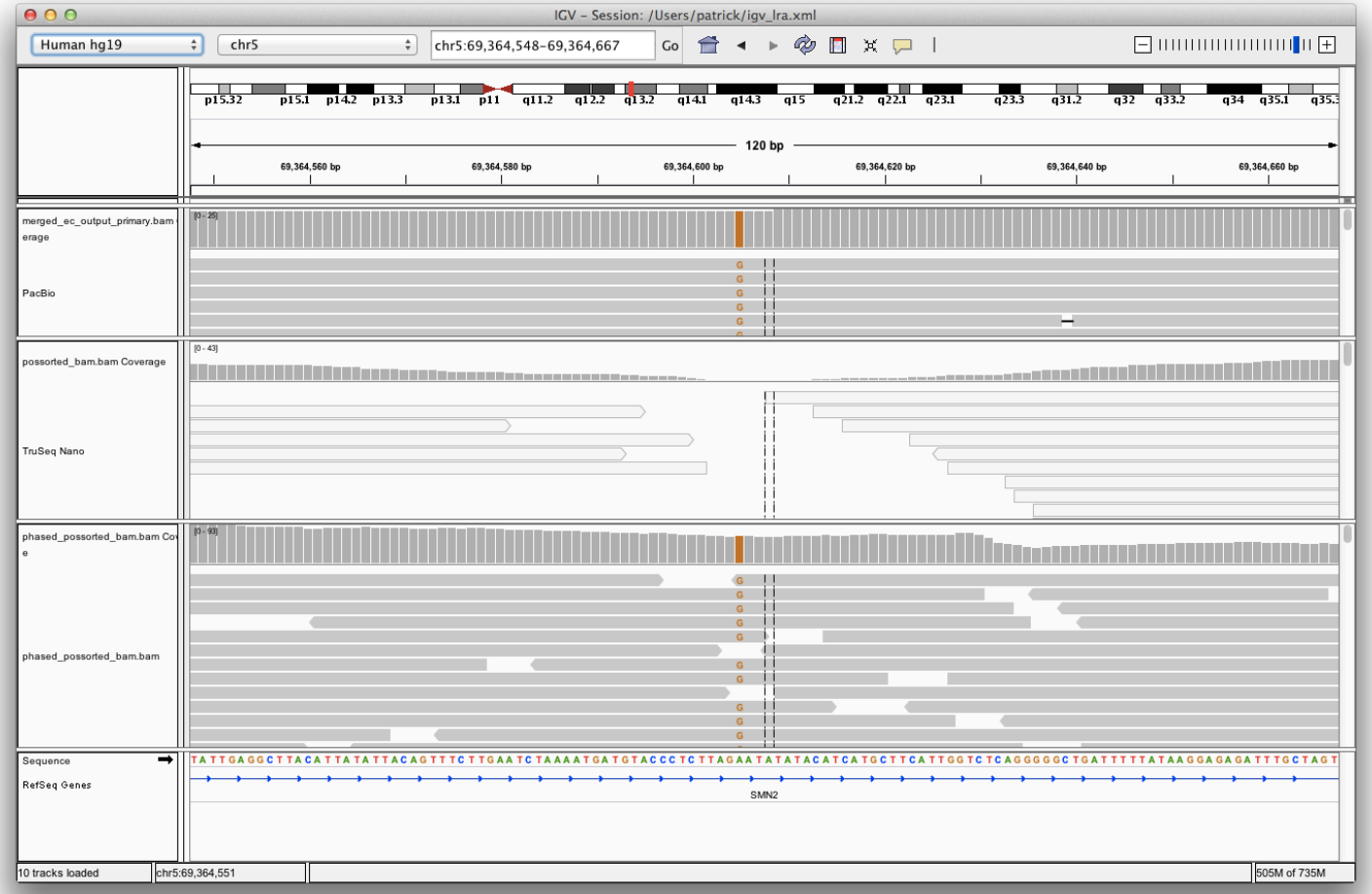
Variants in Clinically Relevant Genes

SMN1 Gene: Associated w/ Spinal Muscular Atrophy
SMN1 & SMN2: 500K w/ >99% identity

PacBio

Standard Illumina

Linked-Reads



Variants in Clinically Relevant Genes

PMS2: DNA repair (associated w/ Lynch Syndrome)

PacBio

Standard Illumina

Linked-Reads

Variants in Clinically Relevant Genes

PMS2: DNA repair (associated w/ Lynch Syndrome)

PacBio

Standard Illumina

Linked-Reads

Questions