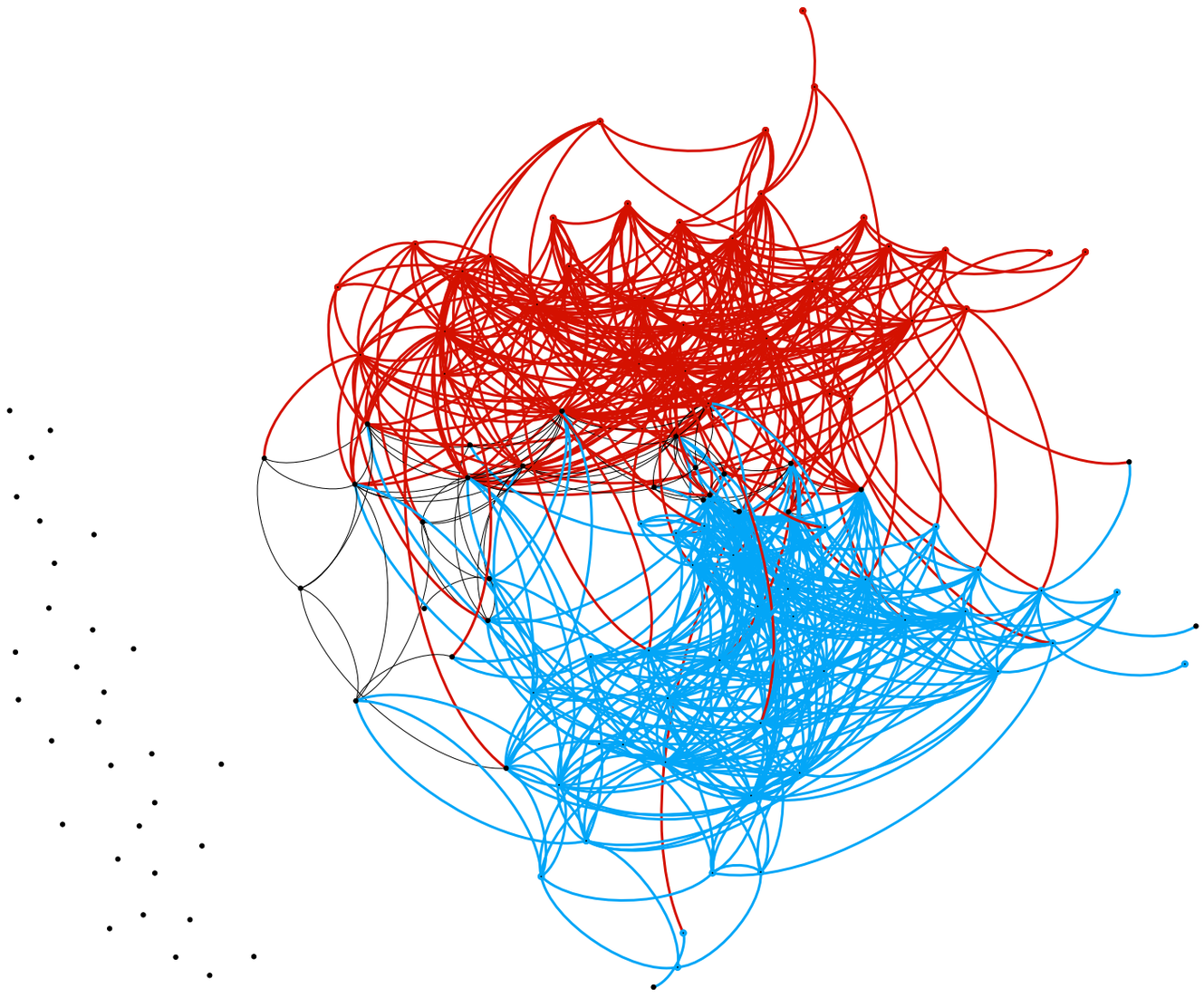
FALCON-Phase: Integrating PacBio and Hi-C for phased diploid assemblies



Friday December 7th

Zev Kronenberg

Talk outline

- **Diploid assembly and Hi-C**
- The FALCON-Phase method
- Evaluating the accuracy of FALCON-Phase

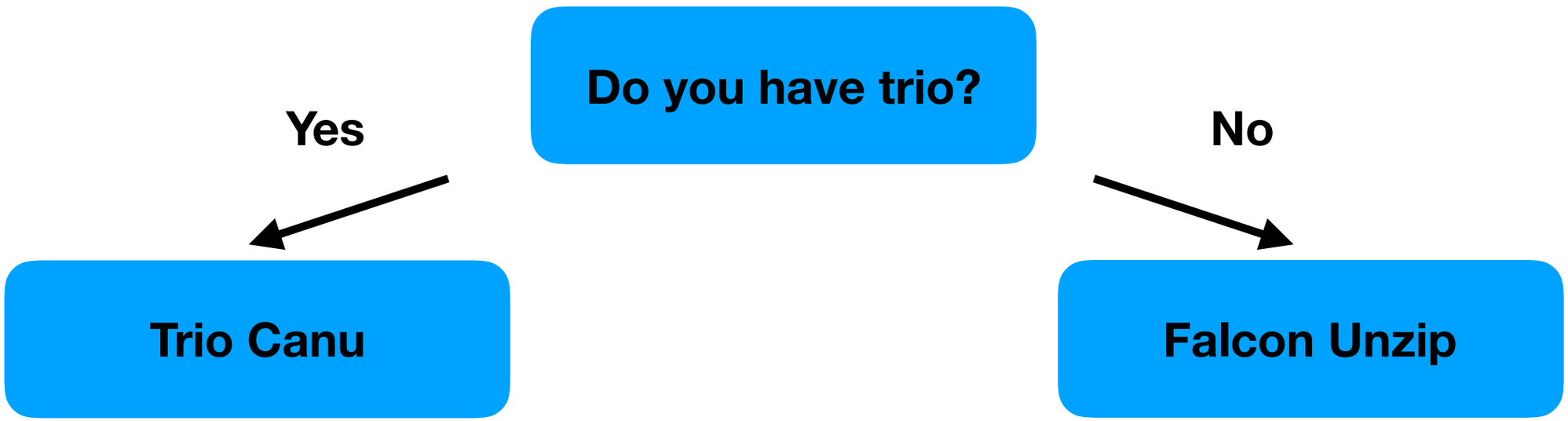
Diploid genome assembly

How big is the human genome?

Diploid genome assembly

~5.75Gb **~5.75Gb** **~6.0Gb**

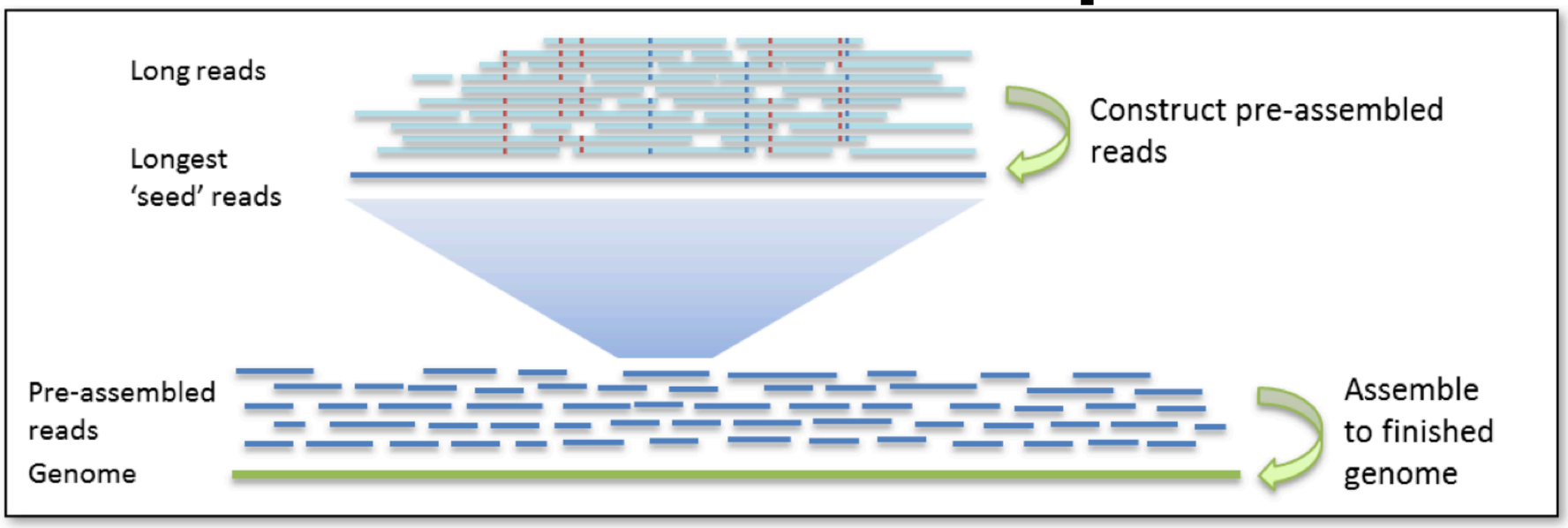
Methods for long-read diploid assembly



Sergey Koren, Adam Phillippy, Arang Rhie

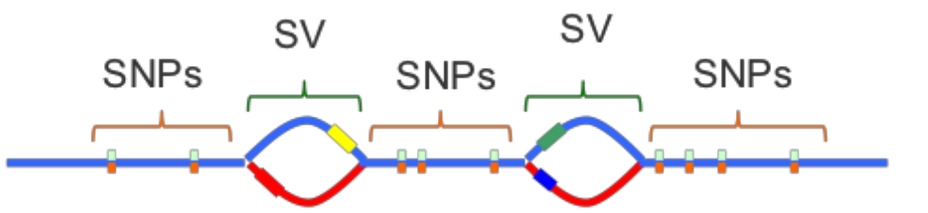
Jason Chin, Ivan Sovic, Sarah Kingan

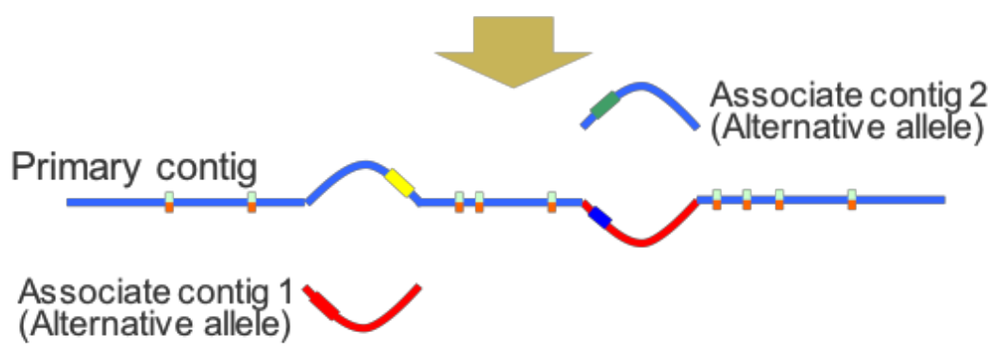
FALCON-Unzip



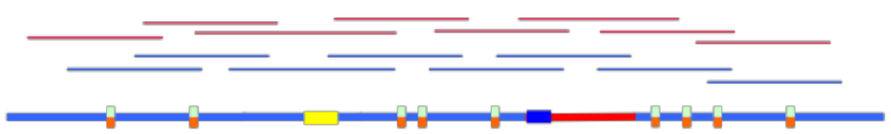
(a) FALCON

Initial assembly graph



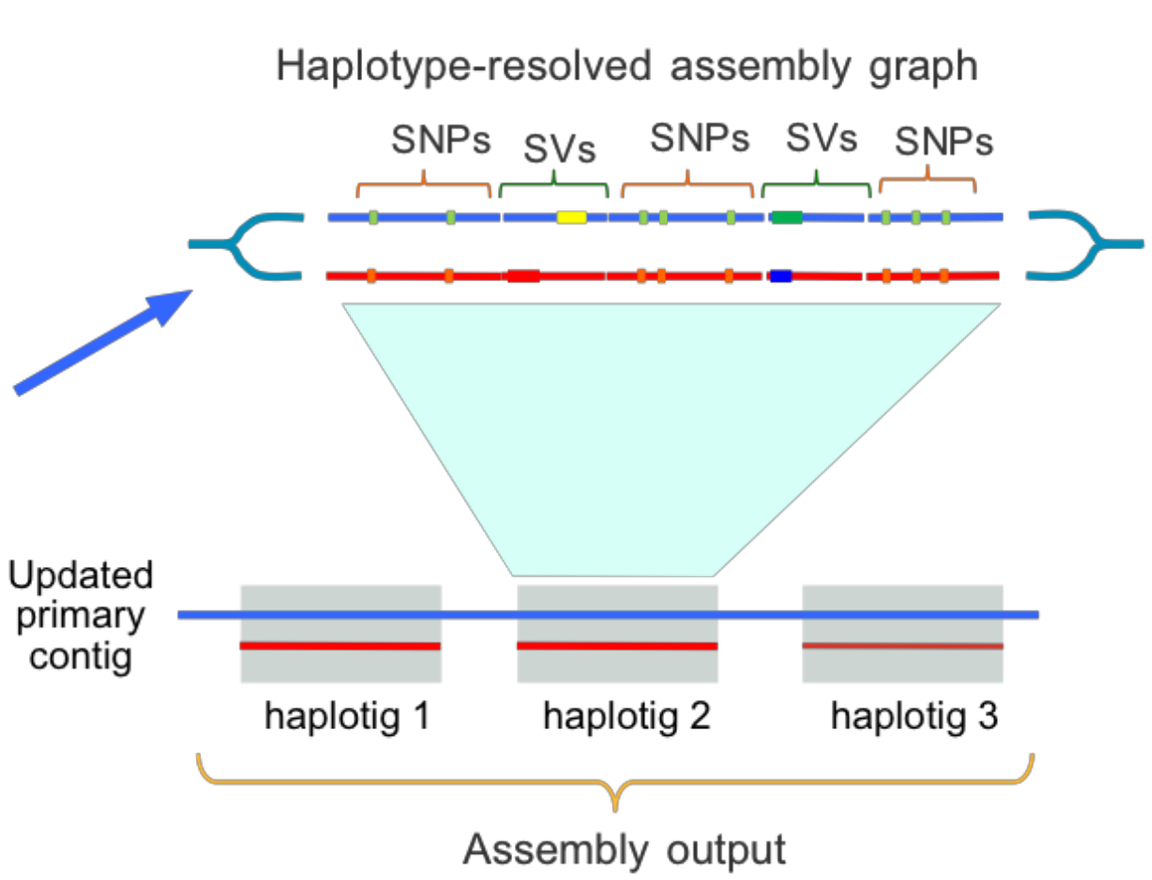


(b)

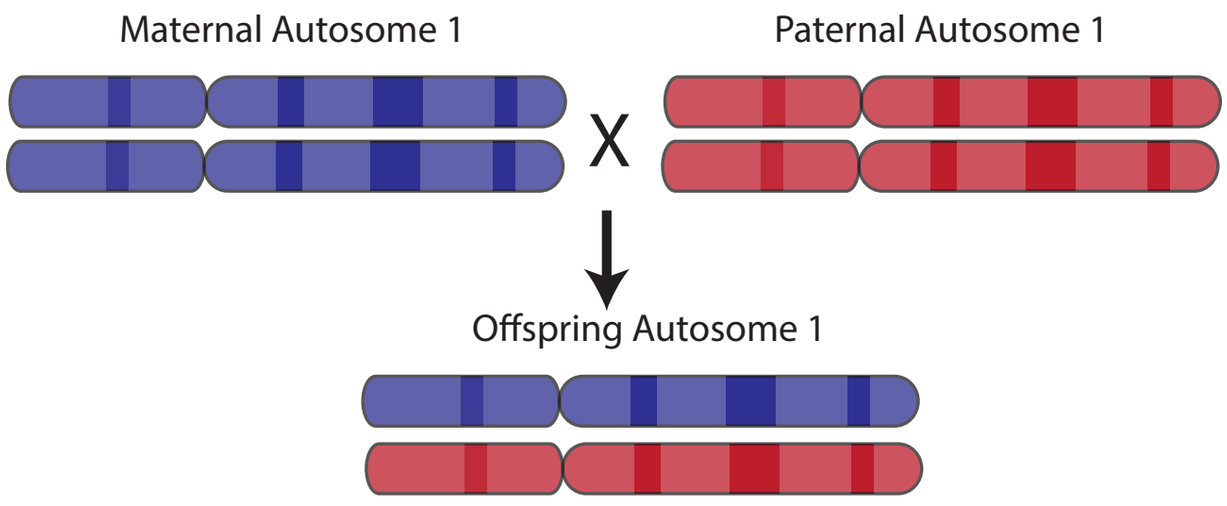


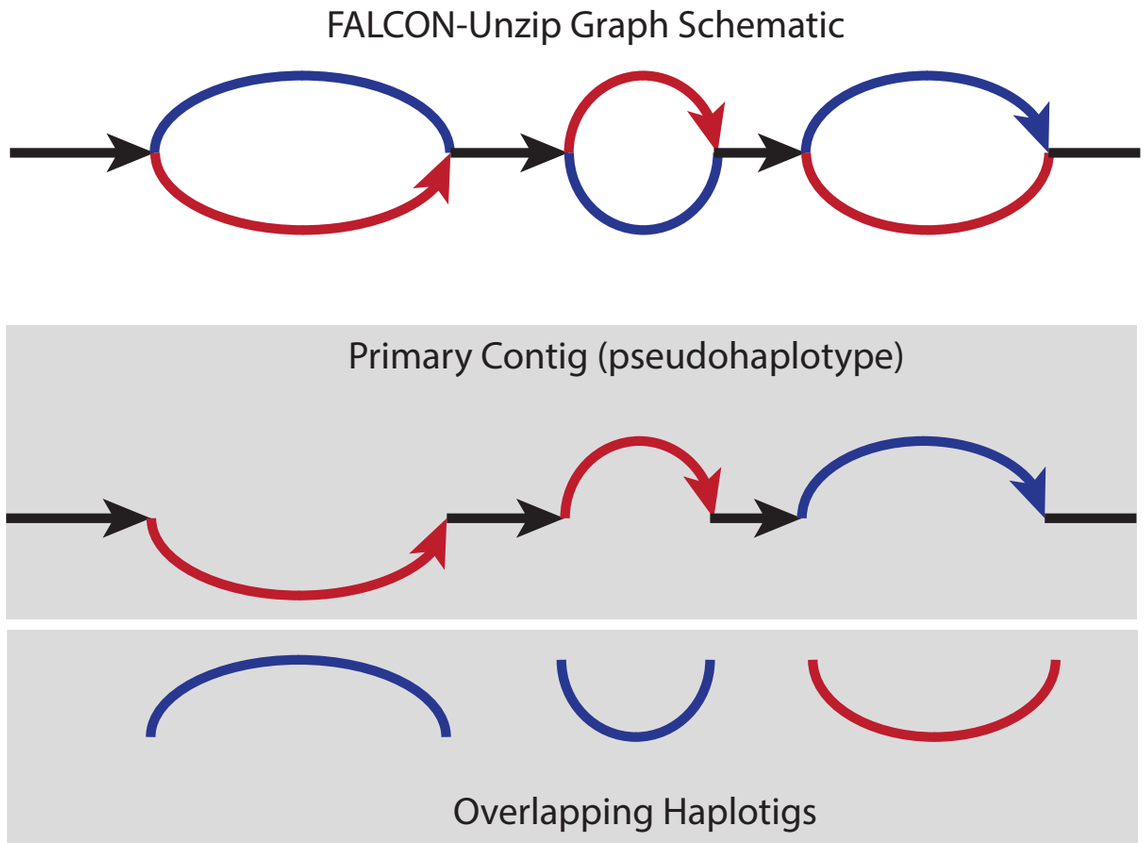
Phase heterozygous SNPs and
identify the haplotype of each read

(c) FALCON-Unzip



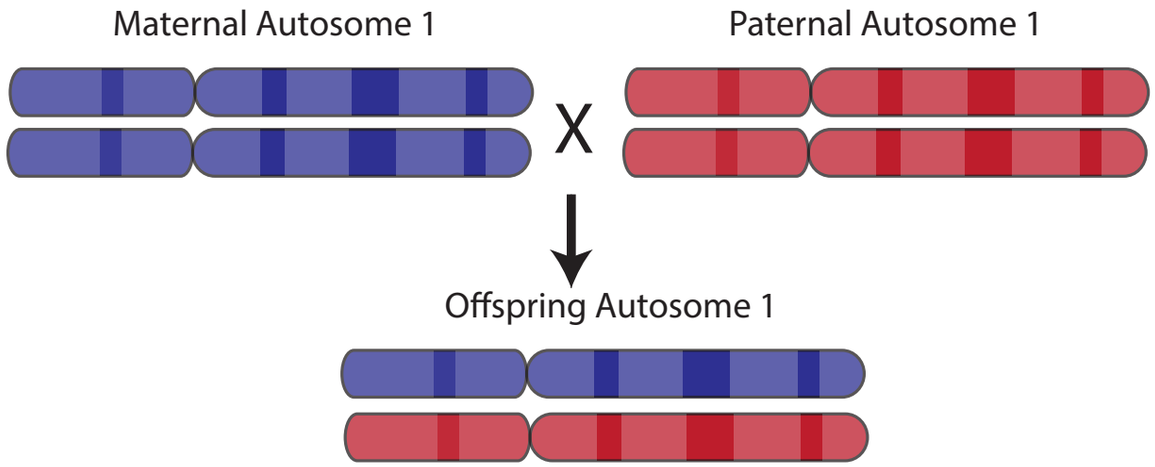
FALCON-Unzip Summarized

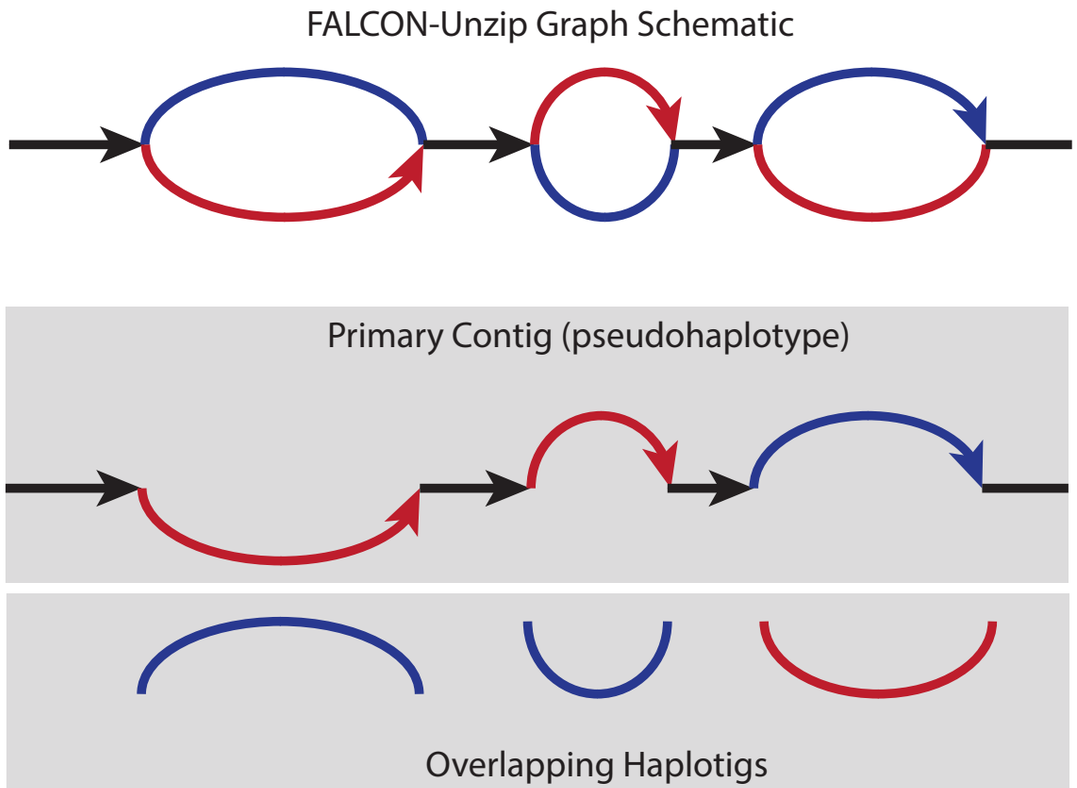




Unsolved challenges in FALCON-Unzip

- The position of haplotigs relative to the primary contigs is unknown.
- High divergence between haplotypes results in two primary contigs.
- Phase switching within haplotigs
- **Phase switching between haplotigs**



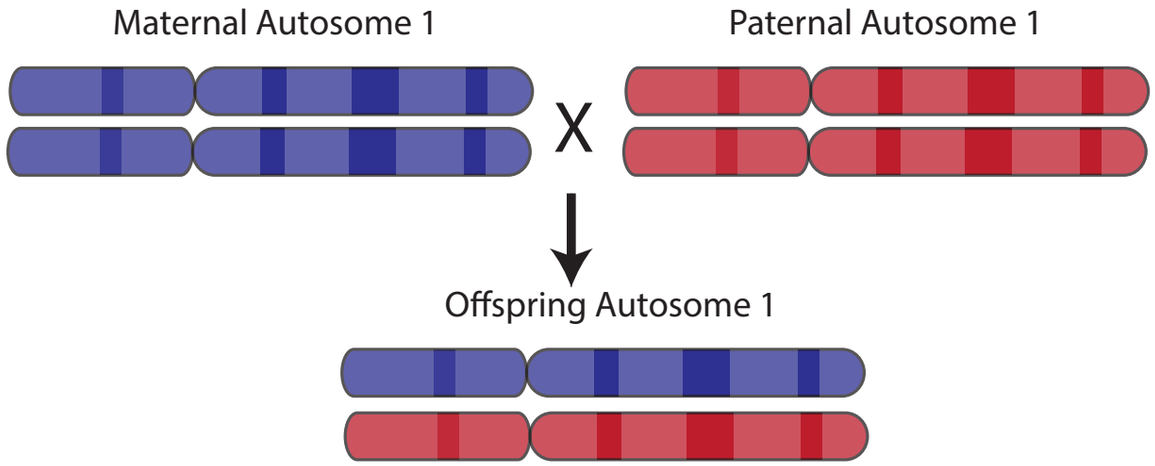


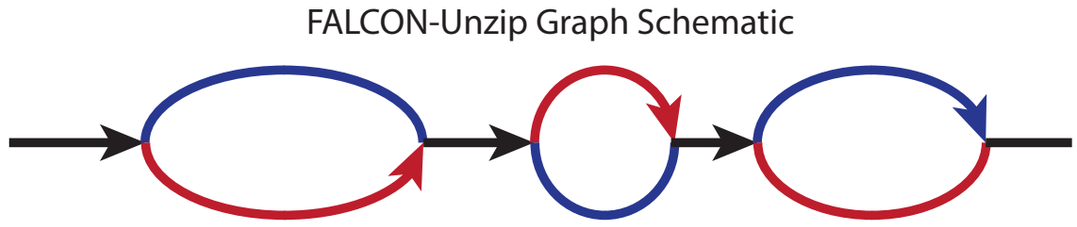
A proposed idea

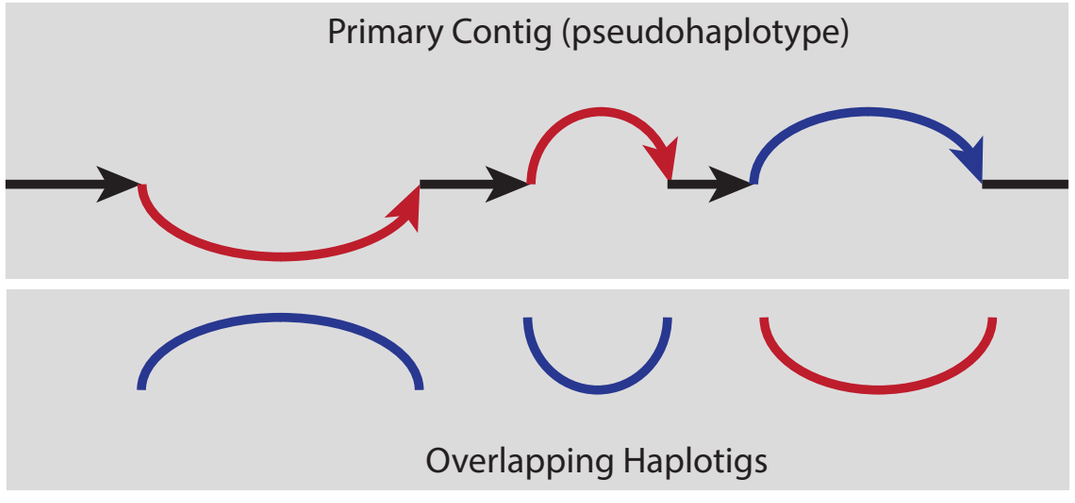
- **Phase switching between haplotigs**
- **PAG 2017**

Can Hi-C be used to phase across haplotigs?

Sarah Kingan (PacBio)







A proposed idea

- **Phase switching between haplotigs**
- **PAG 2017**

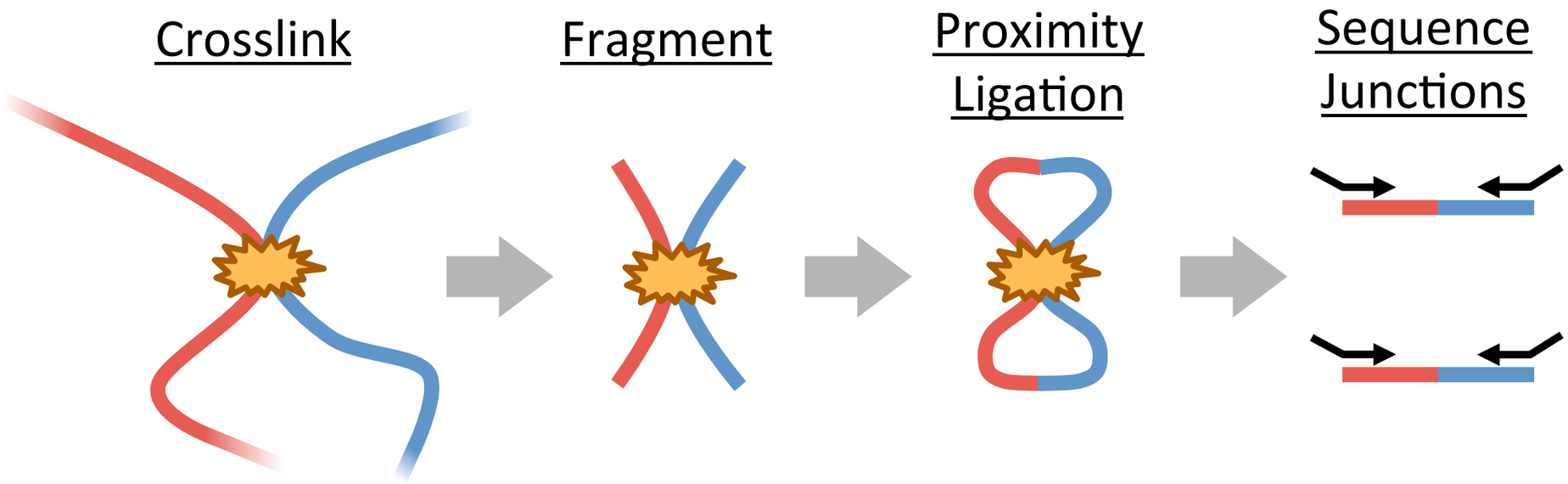
Sounds like fun!

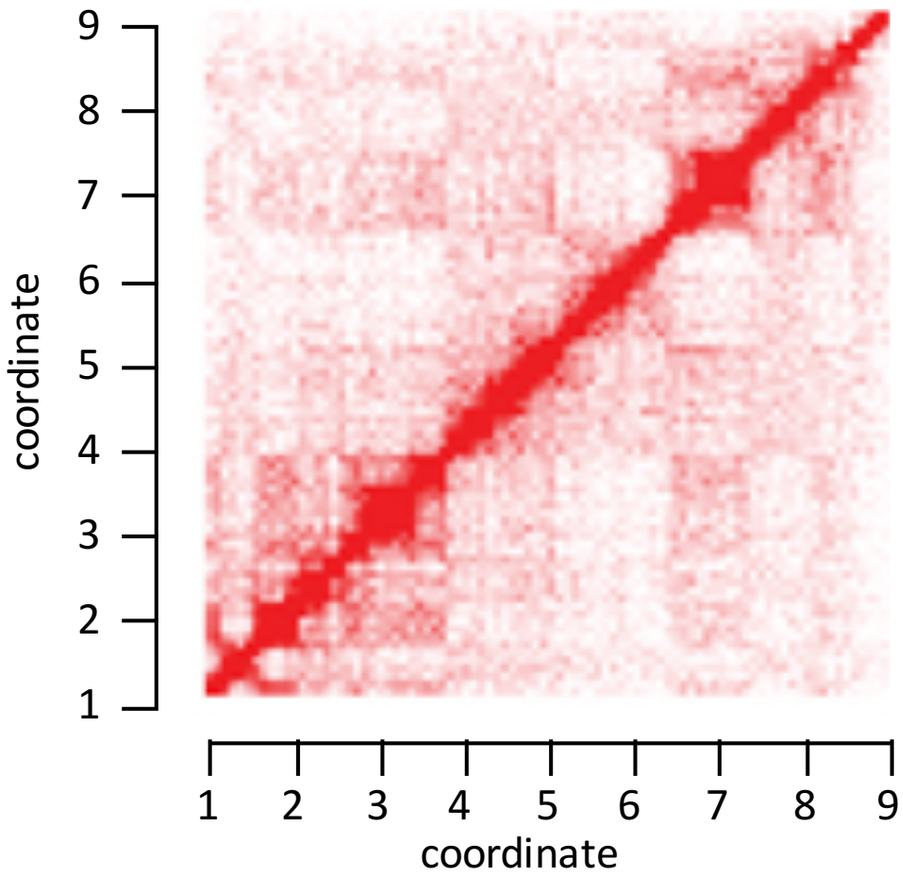
Zev (Phase Genomics)

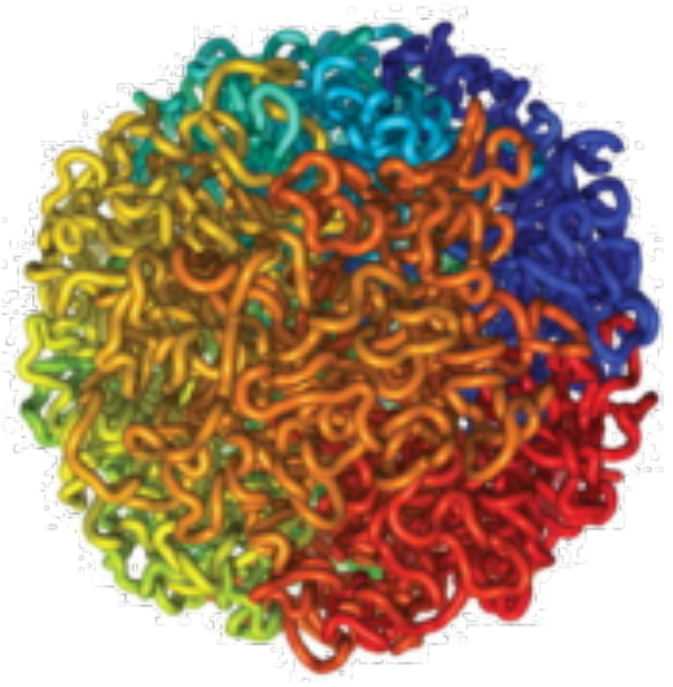
Can Hi-C be used to phase across haplotigs?

Sarah Kingan (PacBio)

What is Hi-C?

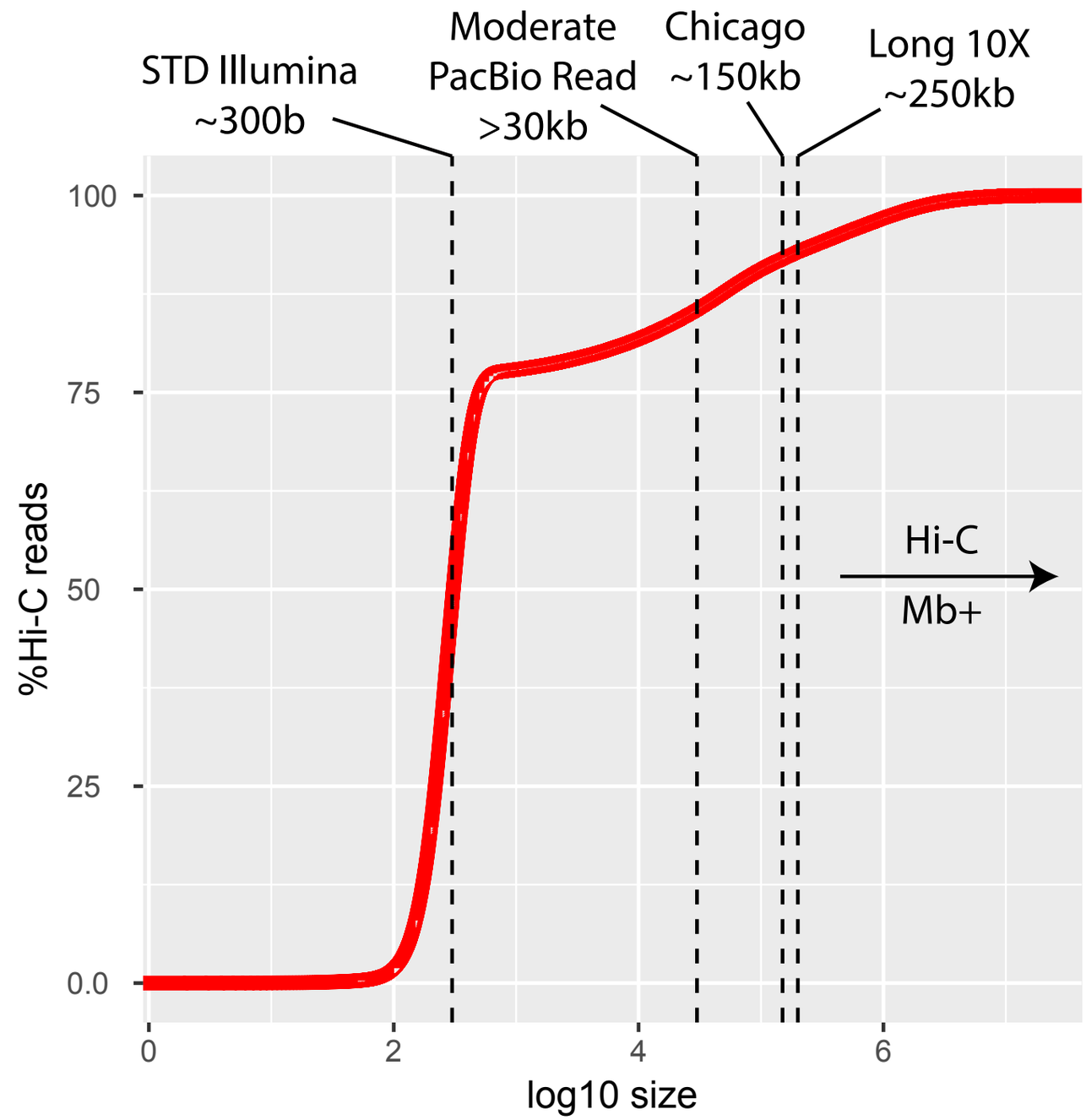






Lieberman-Aiden, *et. al*. Science, 2009

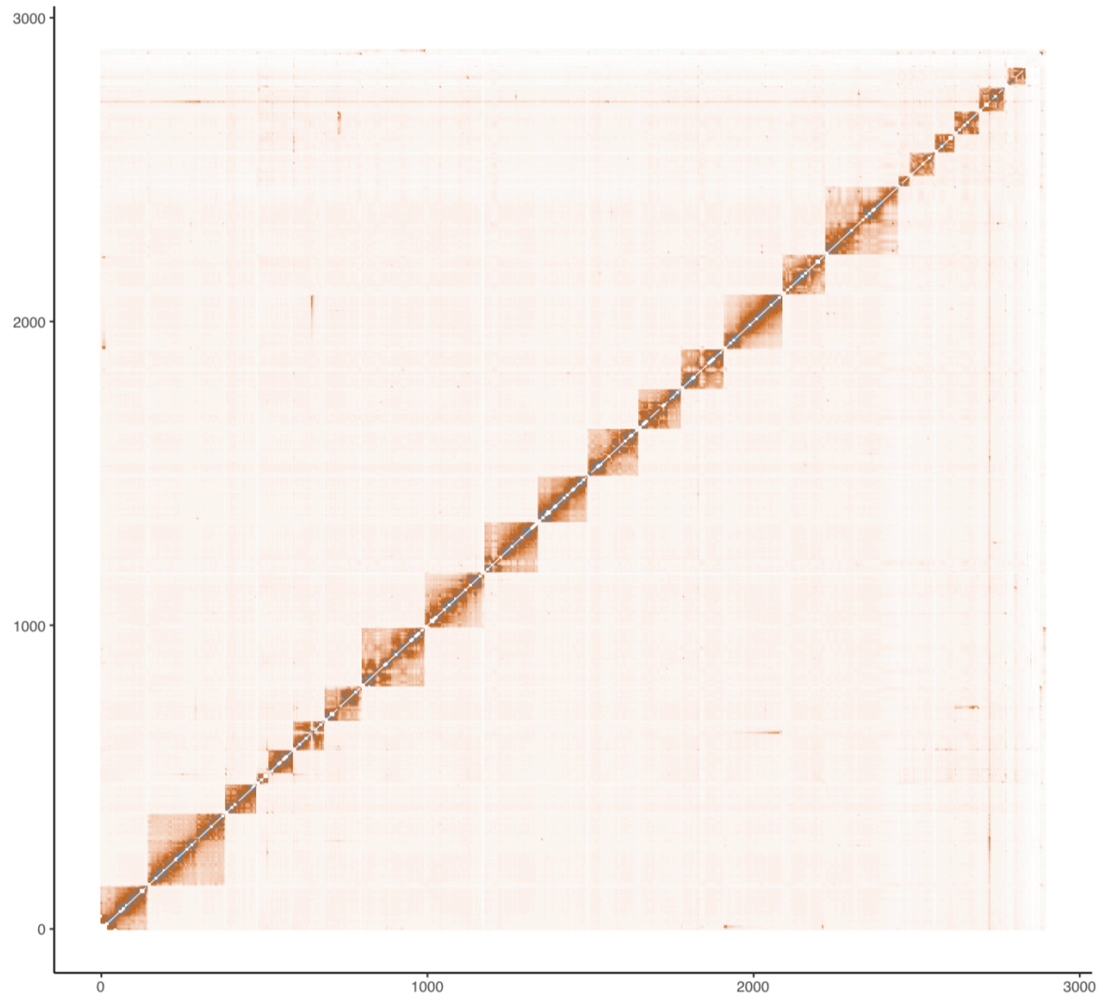
What is Hi-C?


STD Illumina
~300b
Moderate
PacBio Read
>30kb
Chicago
~150kb
Long 10X
~250kb
100
75
50
25
0.0
%Hi-C reads
0
2
4
6
log10 size
Hi-C
Mb+

What is Hi-C good for?

- **Scaffolding genomes**
- **Metagenomic deconvolution**
- **Characterizing large scale structural variation**
- **Phasing**
- **The MAJORITY of Hi-C molecules come from the same chromosome**

The most contiguous human assembly:

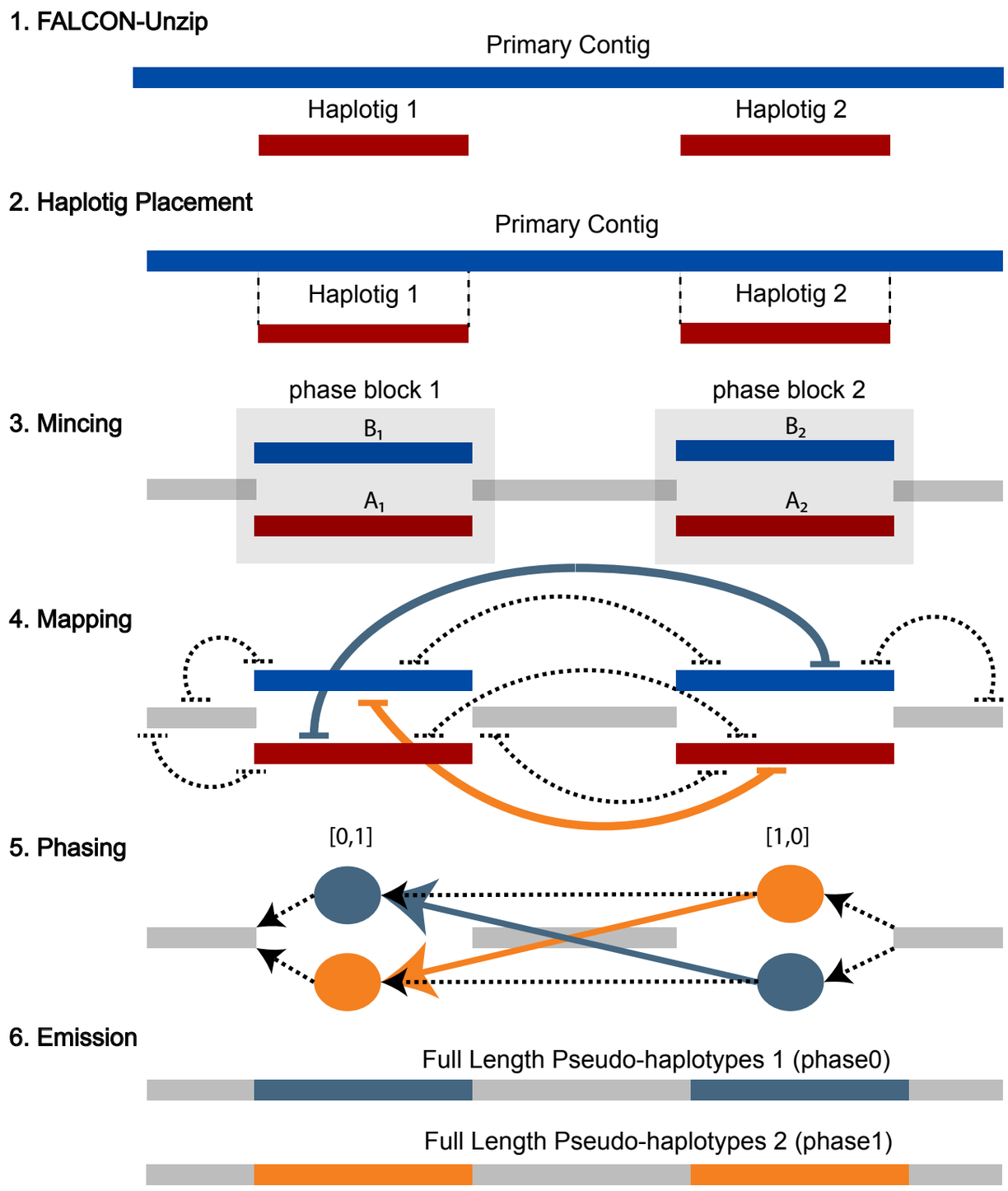


Talk outline

- Diploid assembly and Hi-C
- **The FALCON-Phase method**
- Evaluating the accuracy of FALCON-Phase

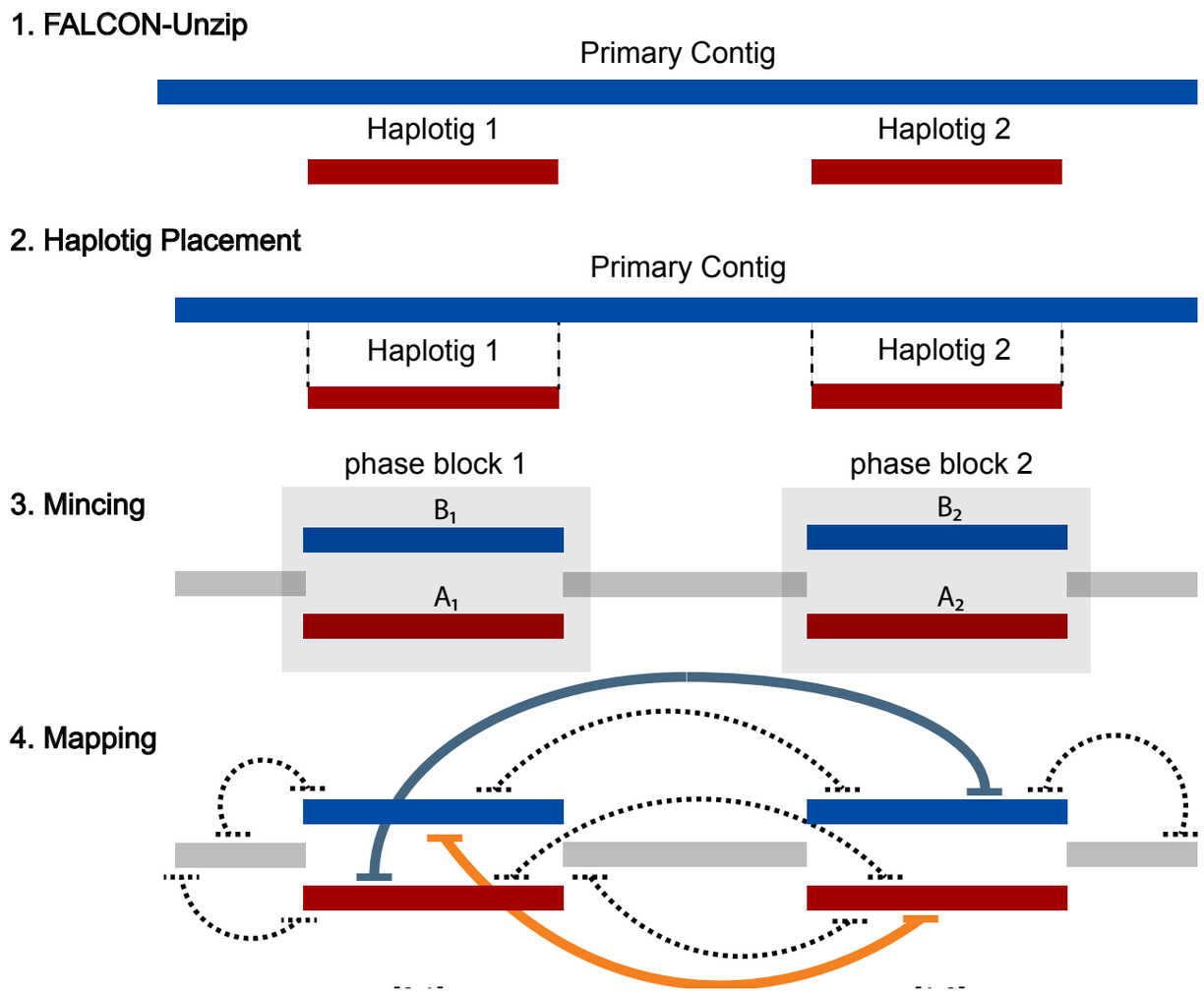
The FALCON-Phase pipeline

- **Input 1: FALCON-Unzip contigs and haplotigs**
- **Input 2: Hi-C fastq**
- **Output: pseudo haplotypes**



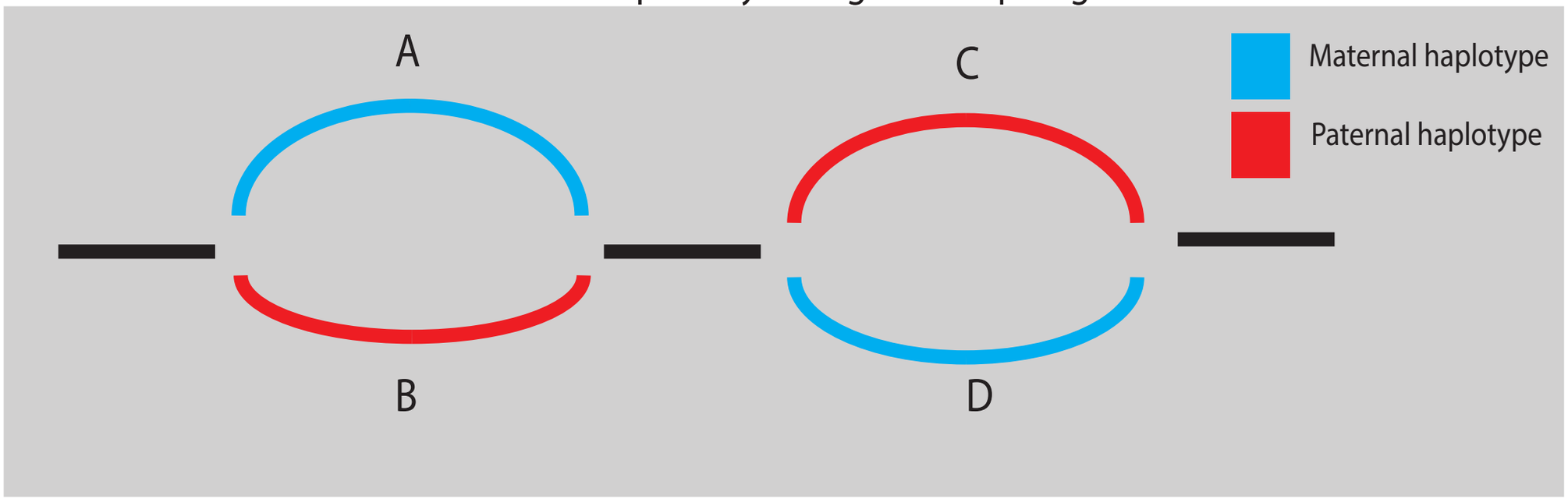
The FALCON-Phase pipeline

- **Preprocessing steps:**
  - **Place haplotigs on their parent contigs (mummer+filtering)**
  - **Create A minced diploid FASTA**
  - **Hi-C data are mapped to minced FASTA**
  - **Hi-C data are filtered on MQ > 10**



Step 5. Phasing

Minced primary contigs and haplotigs:



Mapping count:

	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	.
D	7	0	3	12

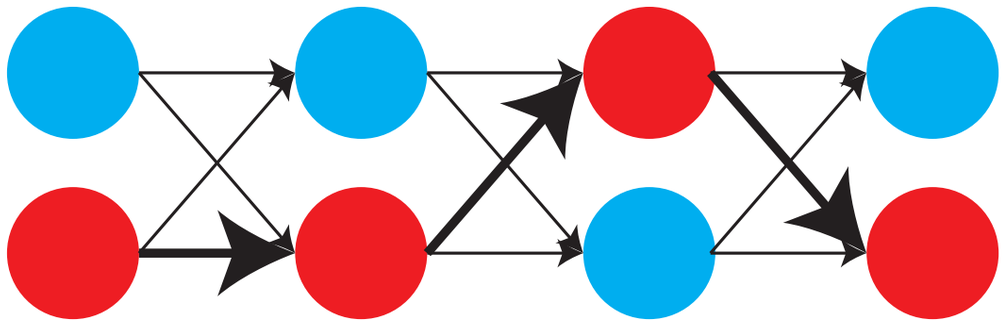
Goal:

To maximize Hi-C link counts within Phase0 and Phase1 haplotigs

Attempt 1: Greedy solution

Mapping count:

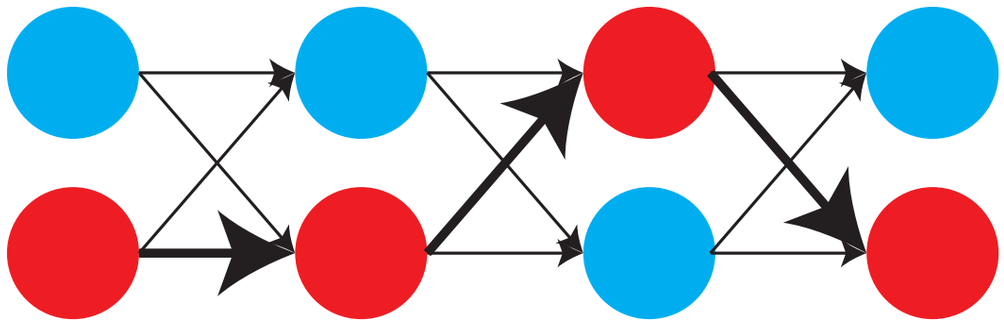
	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	.
D	7	0	3	12



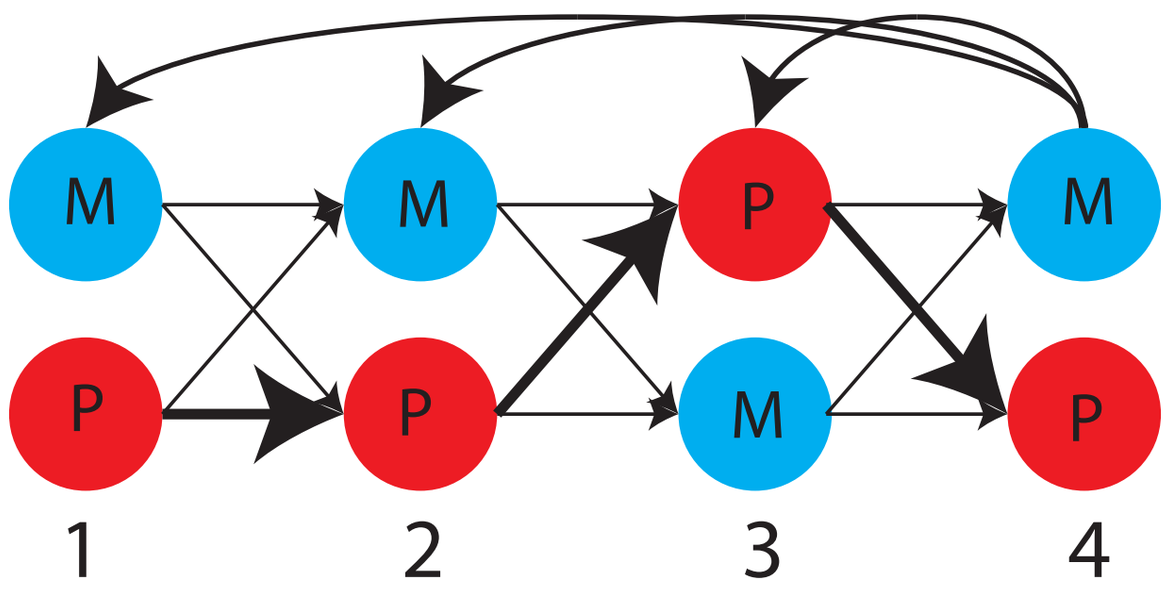
Attempt 1: Greedy solution

Mapping count:

	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	.
D	7	0	3	12



Why did the solution fail?



- **Phase switching errors persist**
- **Some phase blocks have no Hi-C data**
- **Better to consider all past phase blocks**

Stochastic phasing algorithm

Normalized Hi-C matrix: *M*

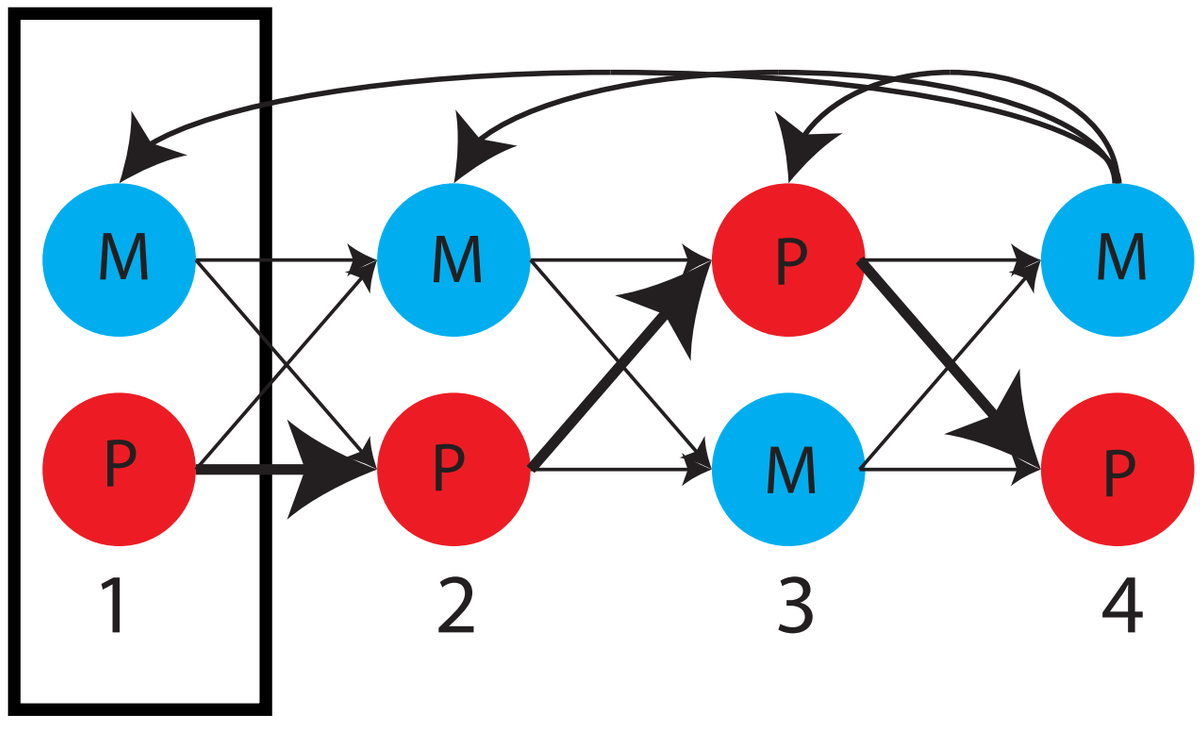
$$\hat{\mathrm{M}}_{\mathrm{i,j}} := \frac{M_{i,j}}{z_i + z_j}$$

	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	·
D	7	0	3	12

Phase block index (overlapping pairs): *C*
e.g: 0:1,2:3,5:6,8:9

Array of temporary phase assignment: *T*
e.g. 0,1,0,0,1

Phase block one



Stochastic phasing algorithm

Normalized Hi-C matrix: *M*

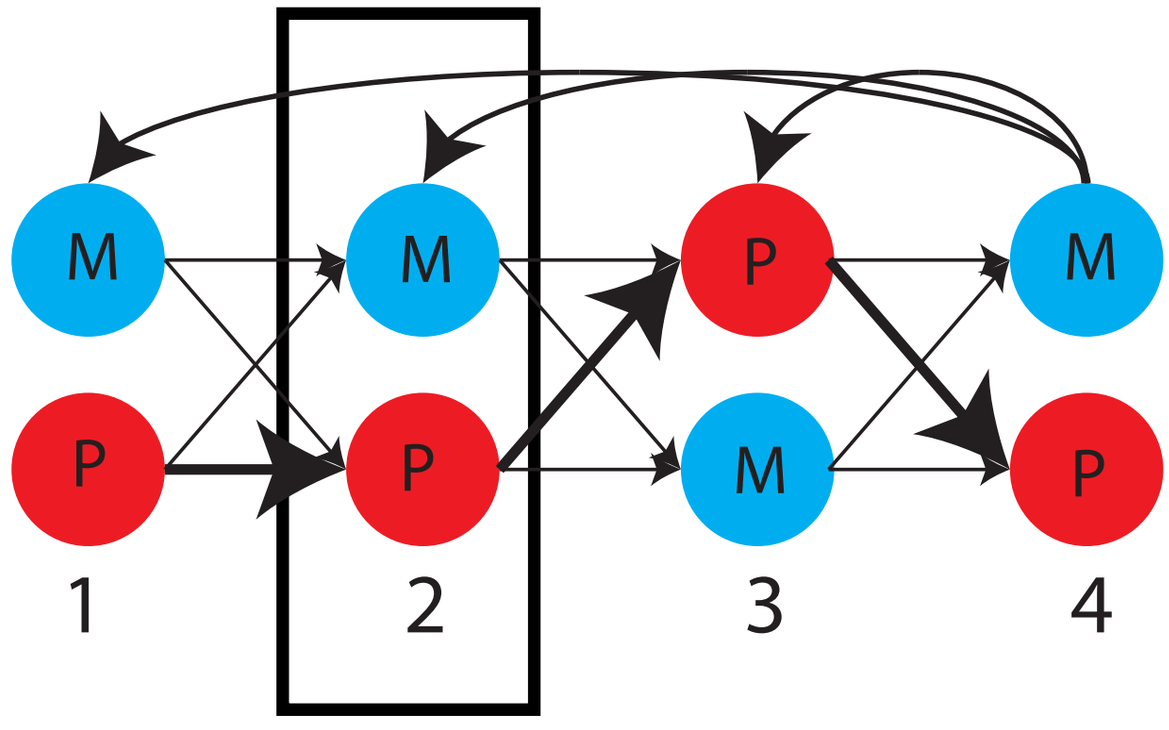
$$\hat{\mathrm{M}}_{\mathrm{i,j}} := \frac{M_{i,j}}{z_i + z_j}$$

	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	·
D	7	0	3	12

Phase block index (overlapping pairs): *C*
e.g: 0:1,2:3,5:6,8:9

Array of temporary phase assignment: *T*
e.g. 0,1,0,0,1

Phase block two



Stochastic phasing algorithm

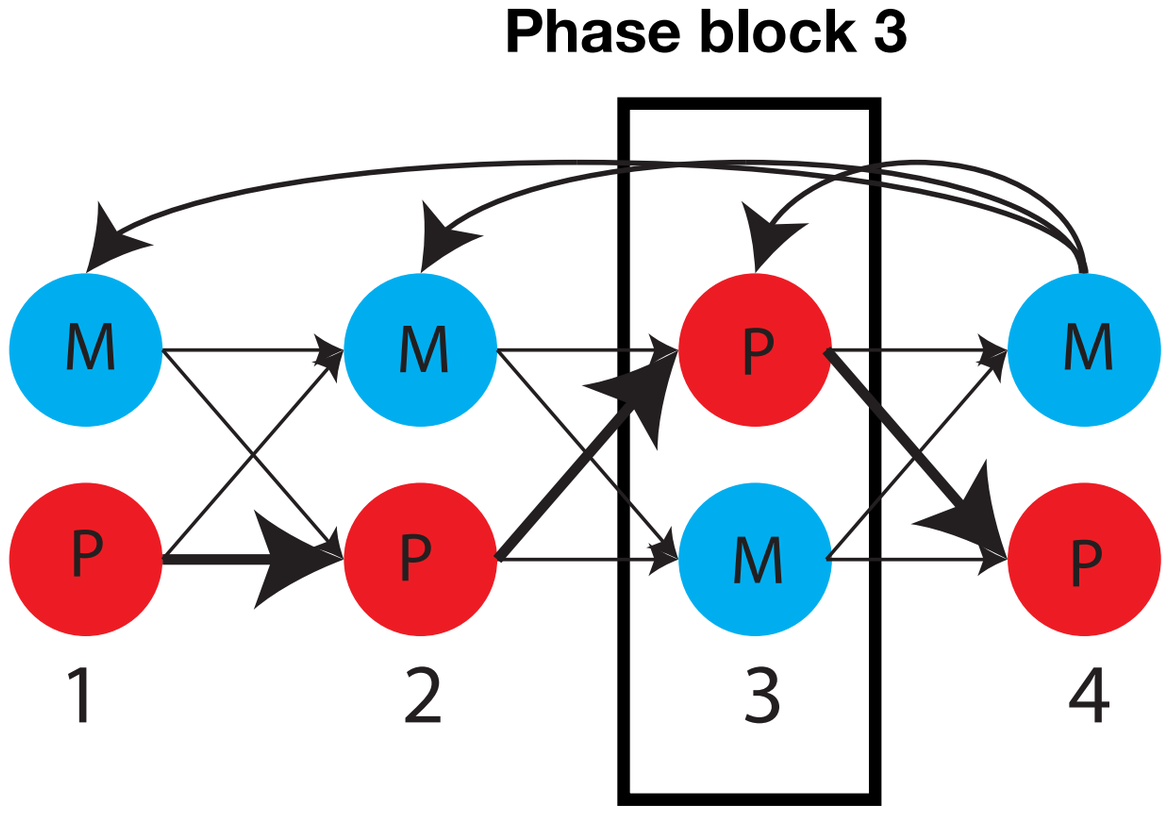
Normalized Hi-C matrix: *M*

$$\hat{\mathrm{M}}_{\mathrm{i,j}} := \frac{M_{i,j}}{z_i + z_j}$$

	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	·
D	7	0	3	12

Phase block index (overlapping pairs): *C*
e.g: 0:1,2:3,5:6,8:9

Array of temporary phase assignment: *T*
e.g. 0,1,0,0,1



Stochastic phasing algorithm

Normalized Hi-C matrix: *M*

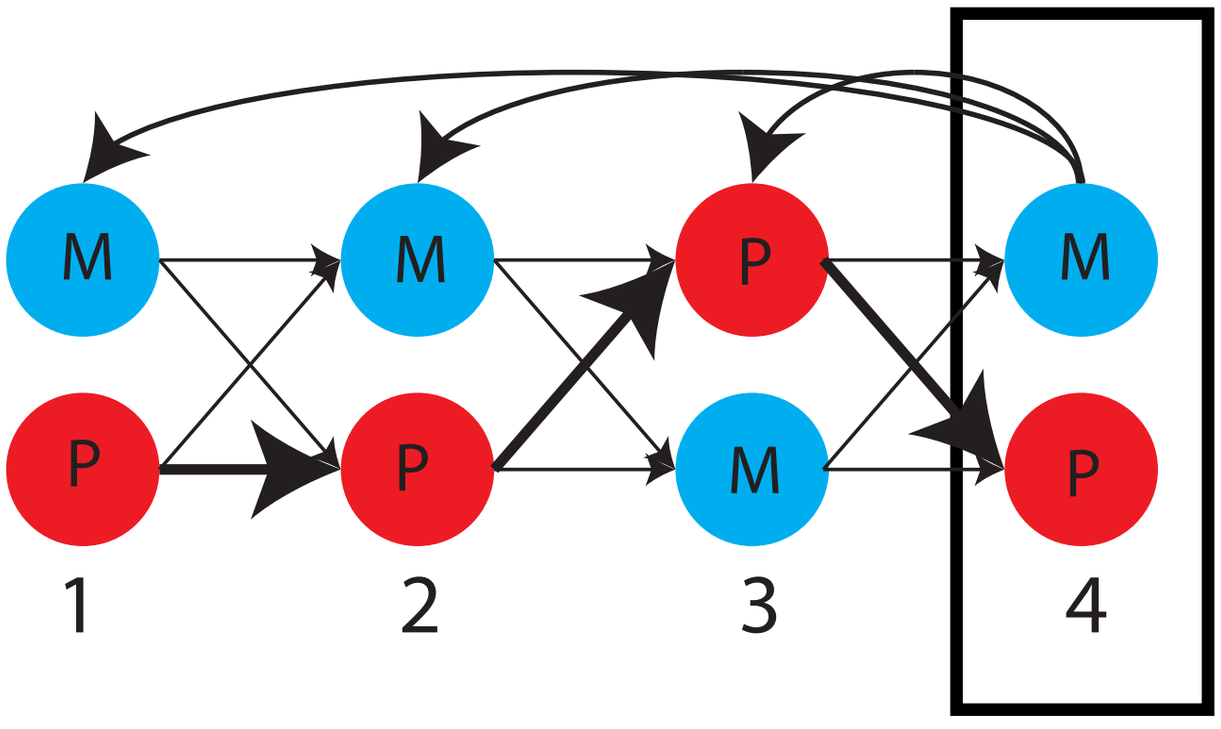
$$\hat{\mathrm{M}}_{\mathrm{i,j}} := \frac{M_{i,j}}{z_i + z_j}$$

	A	B	C	D
A	18	.	.	.
B	1	15	.	.
C	2	9	15	·
D	7	0	3	12

Phase block index (overlapping pairs): *C*
e.g: 0:1,2:3,5:6,8:9

Array of temporary phase assignment: *T*
e.g. 0,1,0,0,1

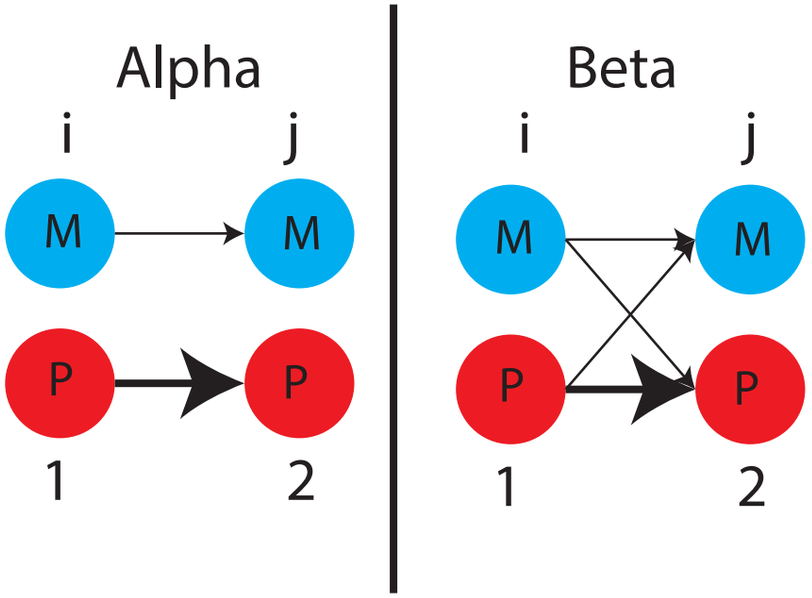
Phase block 4

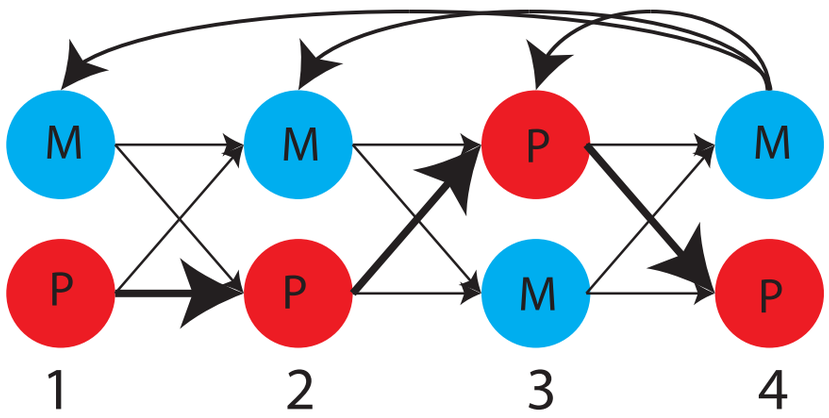


Stochastic phasing algorithm

$$phaseFreq(i, T, \hat{M}, C) = \frac{\sum_{j=0}^{j<i} \gamma(i,j) * \alpha(i,j)}{\sum_{j=0}^{j<i} \beta(i,j)}$$

$$\gamma(\mathrm{i,j}) = \begin{cases} 1, & T[i] = T[j] \\ 0, & T[i] \neq T[j] \end{cases}$$





Phase block update

for $i \leftarrow 0$ **to** n **do**
  for $j \leftarrow 1$ **to** m **do**
    $T[j] \leftarrow 1$;
    if $phaseFreq(j,\ T,\ \hat{M},\ C) < runif()$ **then**
      $T[j] \leftarrow 0$;
    end
    if $i > b$ **and** $T[j] = 1$ **then**
      $P[j] \leftarrow P[j] + 1$;
    end
  end
end

Array of temporary phase assignment: *T*

Number of iterations: *N*

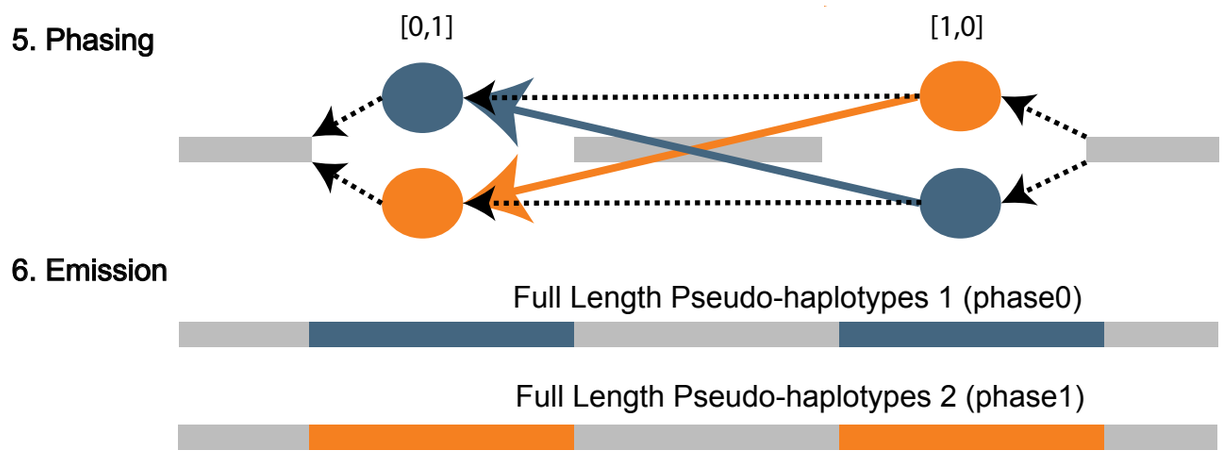
Number of Phase blocks: *M*

Array of phase counts: *P*

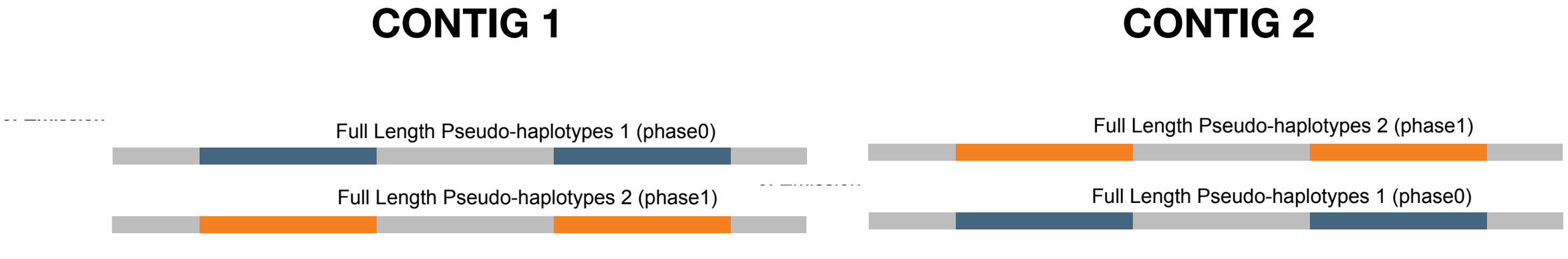
Burn-in: B

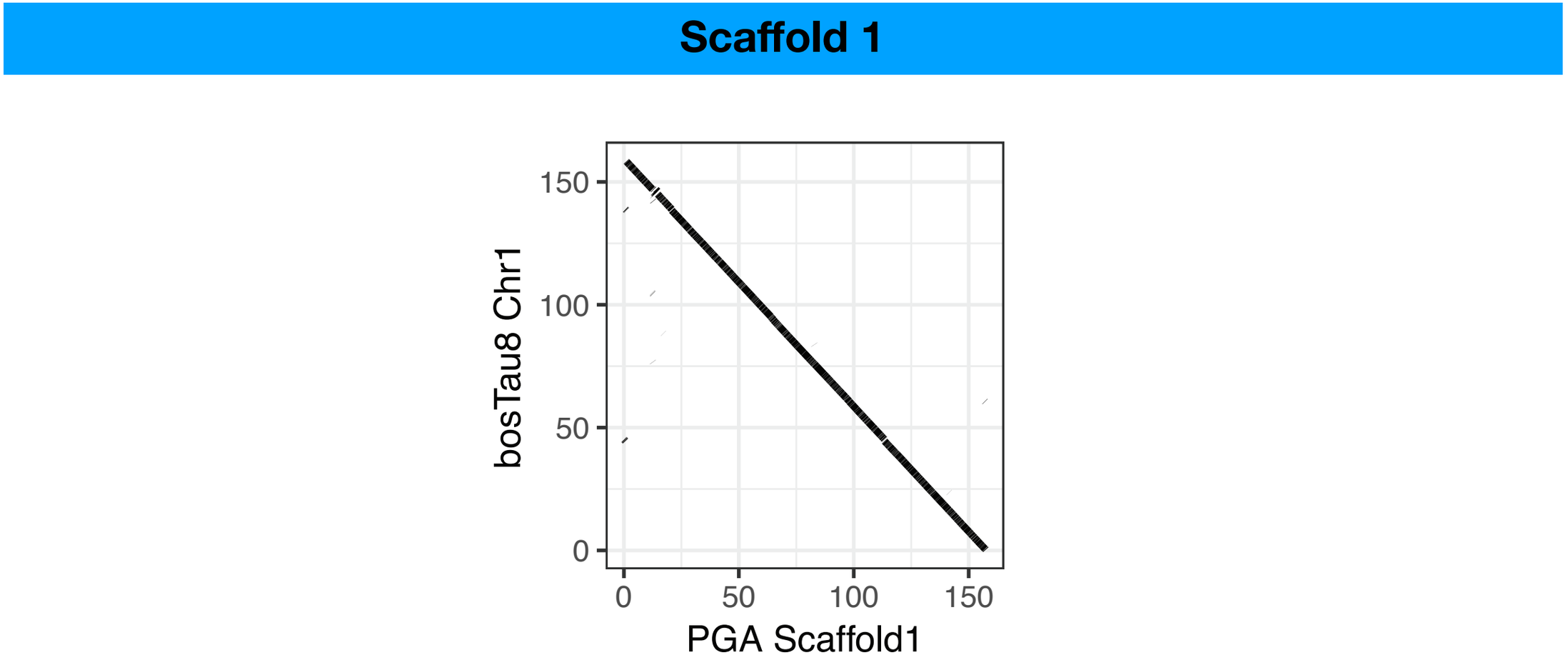
The FALCON-Phase pipeline

- **Emission step:**
  - **Parse phasing results**
  - **Reconstitute two pseudo haplotypes**



The FALCON-Phase method can be repeated at chromosome scale!





Talk outline

- Diploid assembly and Hi-C
- The FALCON-Phase method
- **Evaluating the accuracy of FALCON-Phase**

Two evaluation methods - three datasets

- Align FALCON-Phase output against Trio Canu (truth)
- Trio based SNP calling

Brahman x Angus

Heterozygosity 0.92%
Unzipped 88%

Zebra Finch

Heterozygosity 1.57%
Unzipped 74%

HG00733

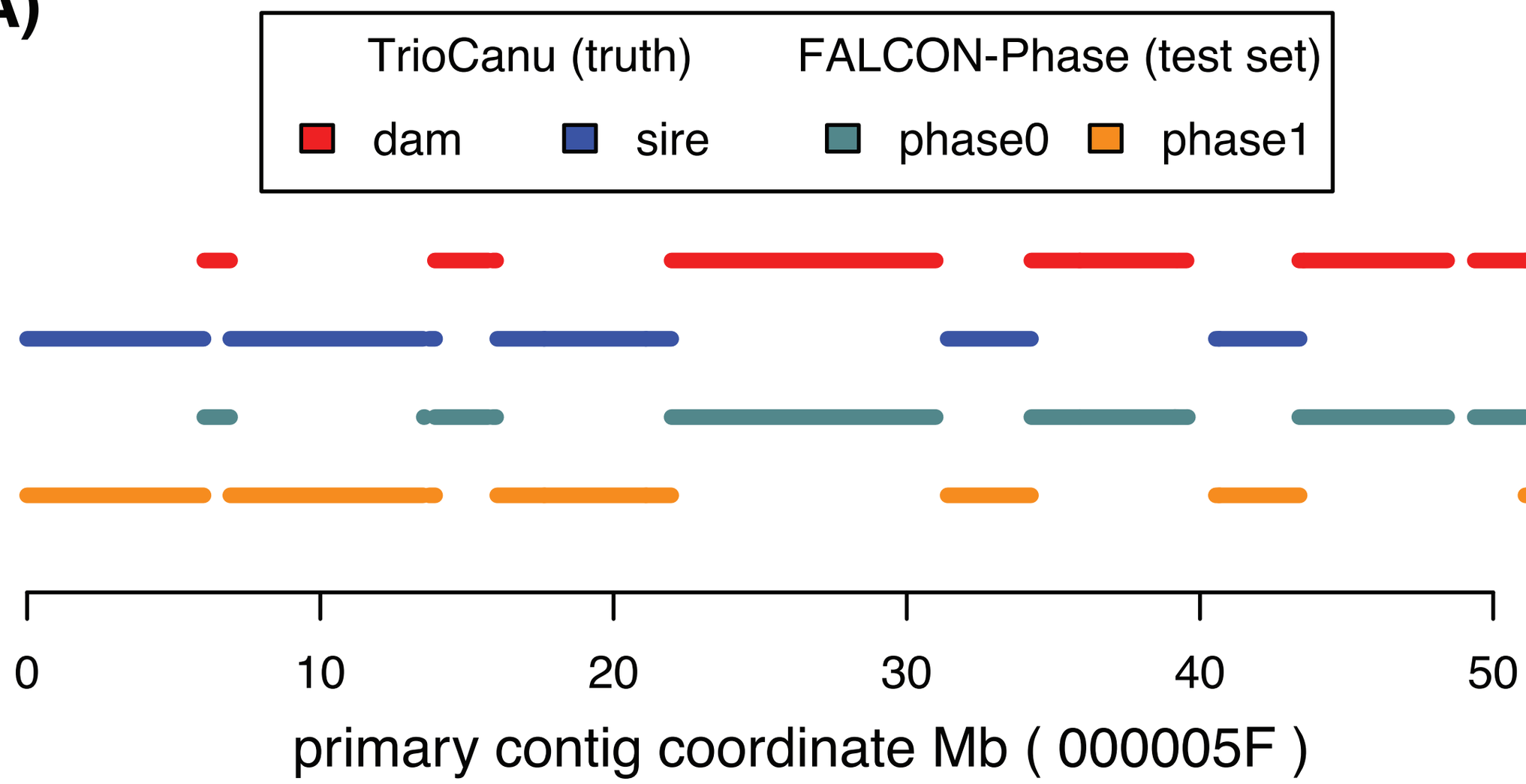
Heterozygosity 0.171%
Unzipped 84%

FALCON-Phase is accurate*

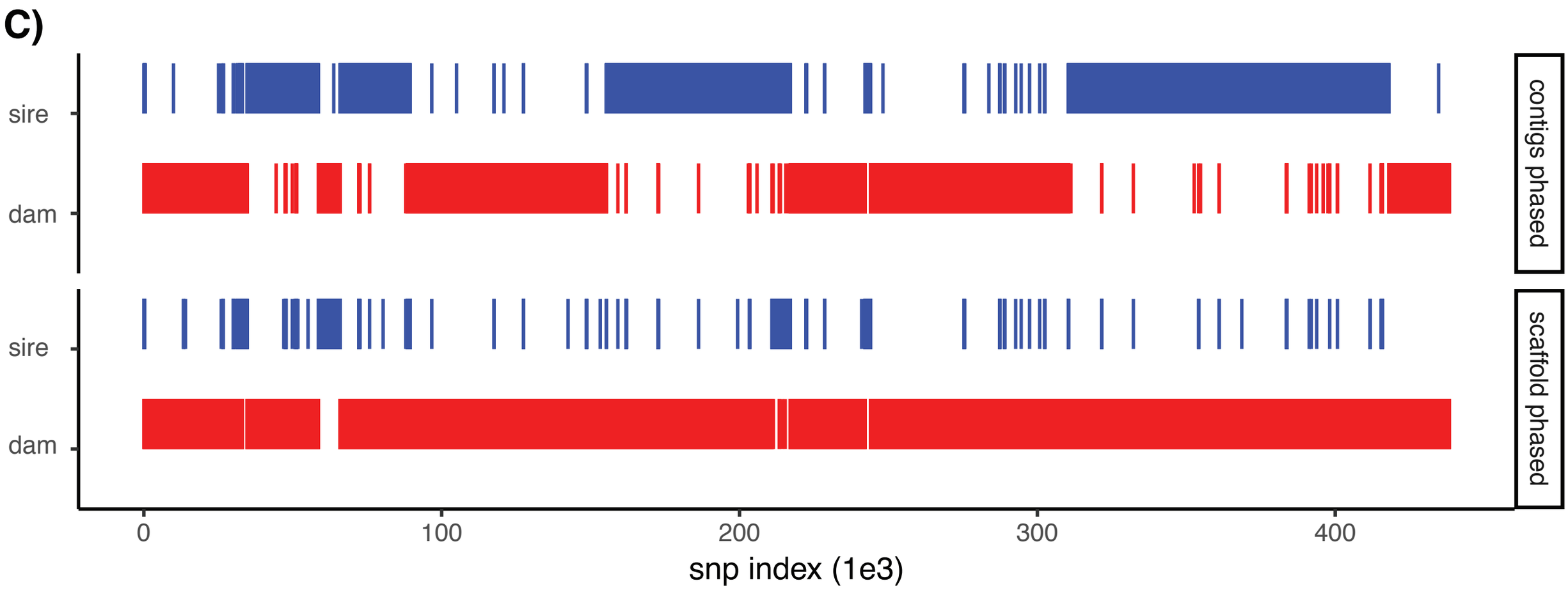
	Zebra Finch	F1 Bull	Puerto Rican
TrioCanu assay	99.4%	96.7%	NA
SNP assay	91.6%	96.7%	80%

FALCON-Phase accuracy compared against Trio Canu

A)



FALCON-Phase accuracy based on Trio SNP calls is > 96%



Conclusions

- **Hi-C is a powerful datatype for scaffolding and phasing genomes**
- **The FALCON-Phase algorithm is effective at phasing diploid contigs and scaffolds**
- **The future: Directly considering Hi-C in PacBio assembly graphs**

Many acknowledgements

PHASE GENOMICS

Shawn Sullivan & Ivan Liachko

Evan Eichler
Katy Munson

USDA

Tim Smith

PACBIO®

Sarah Kingan, Richard Hall

THE UNIVERSITY of ADELAIDE

John Williams
Stephan Hiendleder

National Human Genome Research Institute

Sergey Koren, Arang Rhie, Adam Phillipy

VGP VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

Questions?