Diploid Genome Assembly and Comprehensive Haplotype Sequence Reconstruction

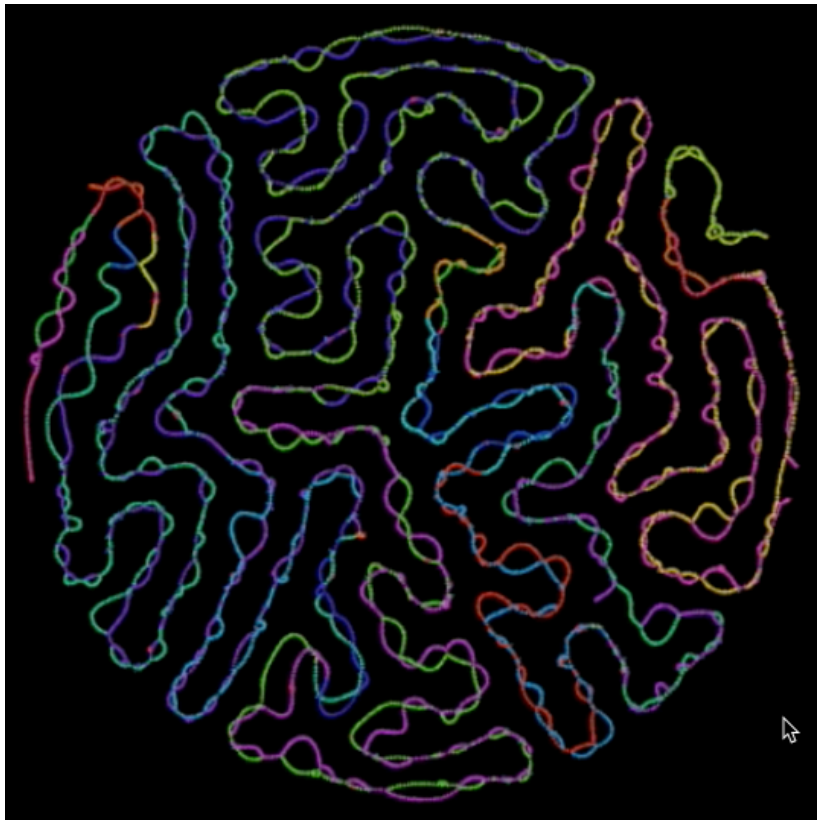
Jason Chin, Paul Peluso, David Rank, Fritz Sedlazeck, Maria Nattestad, Michael Schatz, Greg Concepcion, Alicia Clum, Kerrie Barry, Alex Copeland, Ronan O'Malley

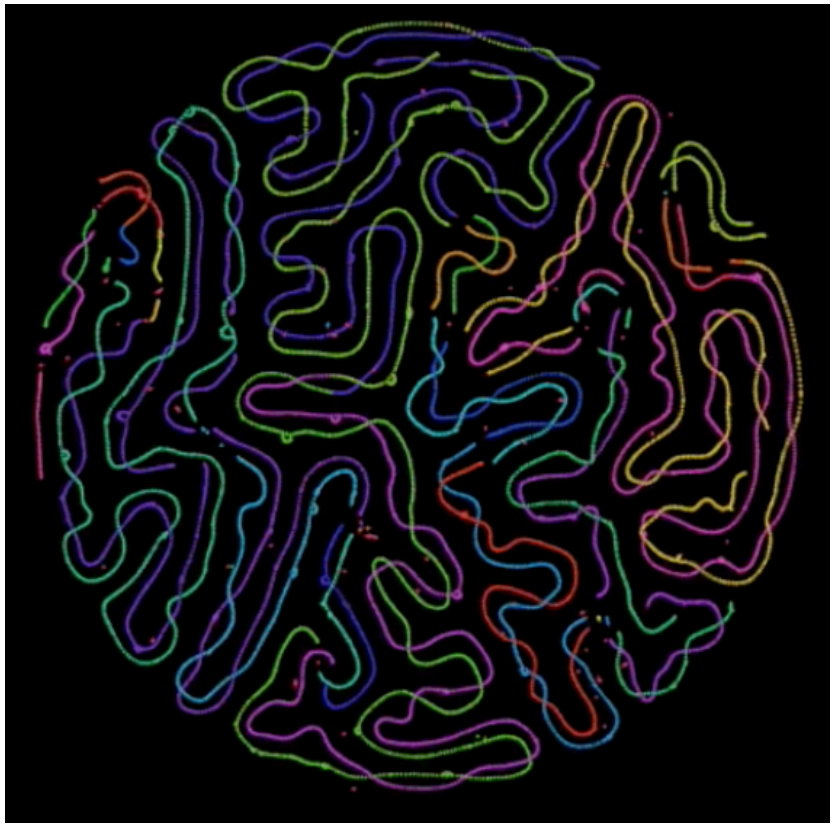
Acknowledgments

- All PacBio Colleagues
- Ronan O'Malley, Chongyuan Luo, Joseph Ecker (HHMI / The Salk Institute)
- Alicia Clum, Kerrie Barry, Alex Copeland (Joint Genome Institute)
- Maria Nattestad, Fritz Sedlazeck, Michael Schatz (CSHL)

- Open source toolsets
  - Daligner (https://dazzlerblog.wordpress.com), Gene Myers
  - BLASR (https://github.com/PacificBiosciences/blasr), Mark Chaisson
  - Python, NetworkX for rapid algorithm protyping
  - Gephi, Graphviz for graph visualization
  - FALCON (https://github.com/PacificBiosciences/falcon, https://github.com/PacificBiosciences/falcon)

SOLVING THE DIPLOID ASSEMBLY PROBLEM



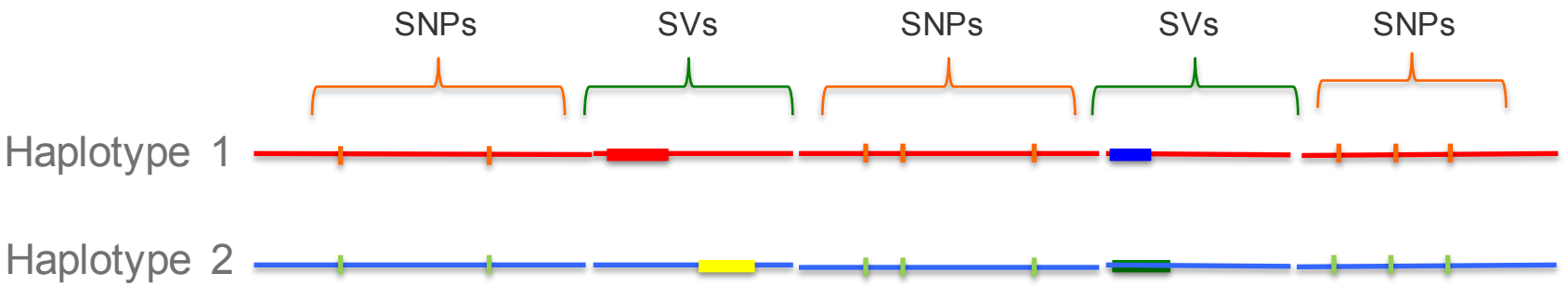


- Bubbles = big variants between the haplotypes
- Collapsed Path = smaller variants between the haplotypes

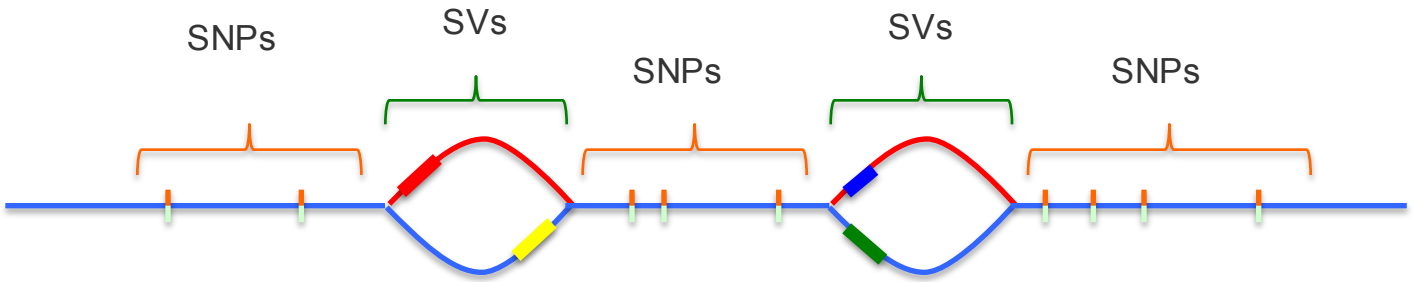
- Falcon (a polyploid-aware assembler) : generating the contigs through the bubbles
- Falcon Unzip: identifying smaller variants and using them to separate the haplotypes

WHY DO WE SEE BUBBLES?

Genome Sequences

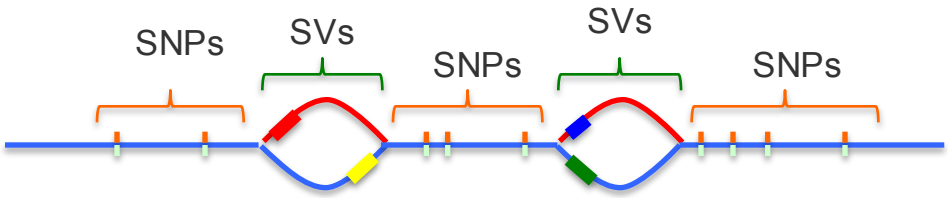


Assembly Graph



In most OLC assembler design, the overlapper does not catch differences at SNP level but structural variations are naturally segregated.

THE FALCON UNZIP PROCESS



FALCON

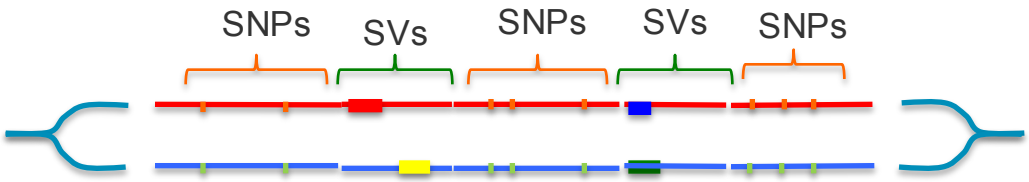
FALCON-Unzip

Associate contig 1
(Alternative allele)

Primary contig

Associate contig 2
(Alternative allele)

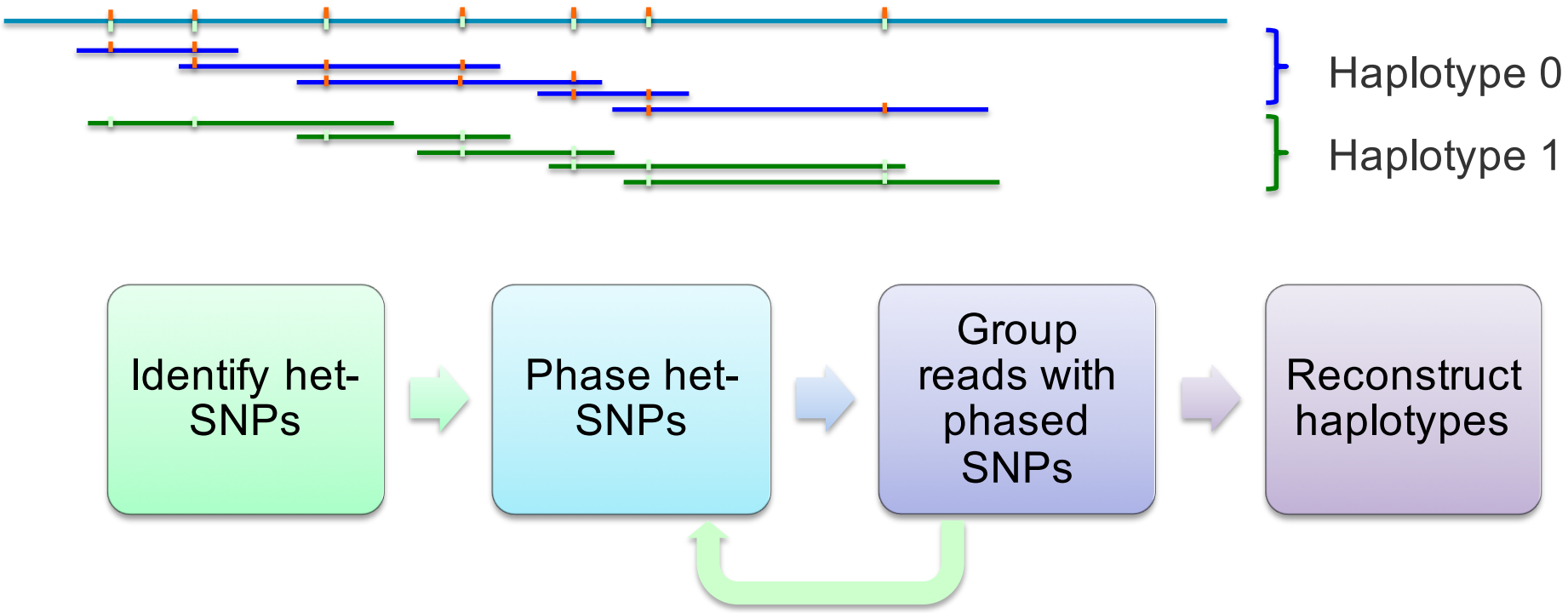
Augmented with haplotype information of each reads



Updated primary contig + “associate haplotigs”

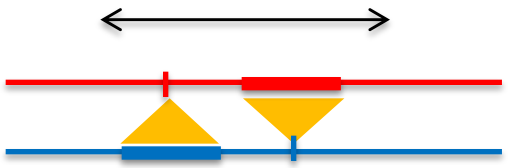
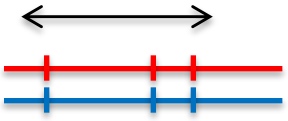
PHASING READ INTO HAPLOTYPE GROUPS

Align SMRT reads to the initial primary contig

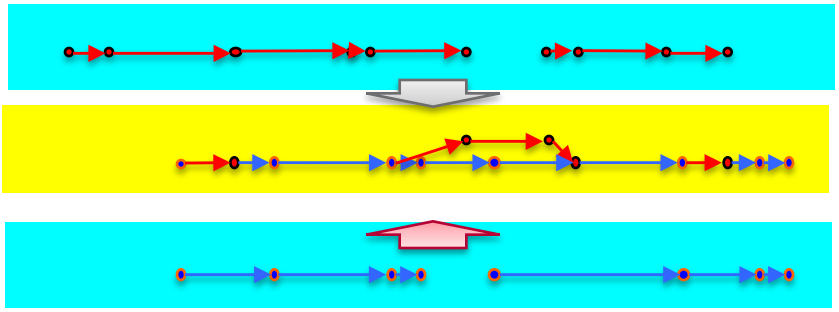


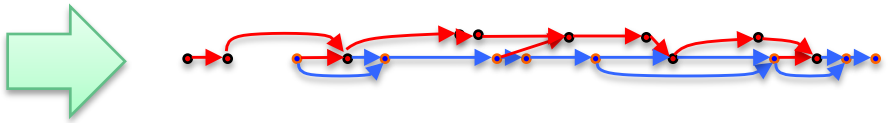
More het-SNPs in longer reads: 8% to 15% sequence error rate is not an issues given enough long read coverage for phasing.

QUESTION: HOW TO RESOLVE STRUCTURAL VARIATIONS & HET-SNPS PHASING AT ONCE

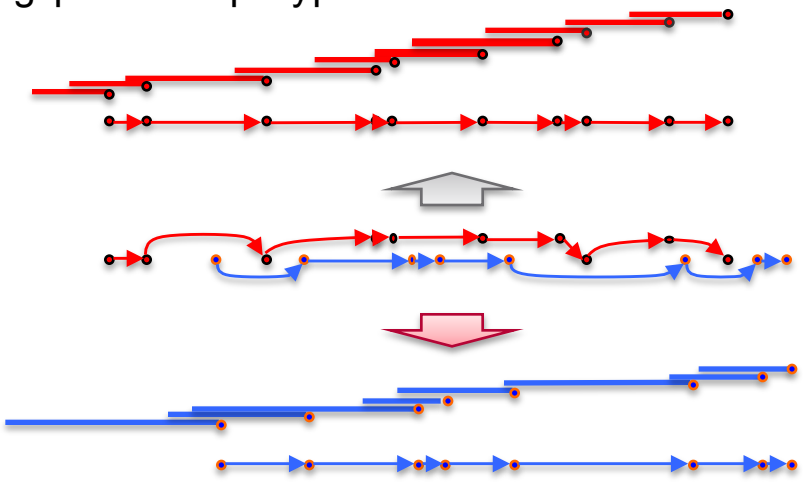
	Information Sources	Pros & Cons	Assembly graph features
Structural Variations	3 kb – 100 kb 	✓ Overlap-layout process catches SV haplotypes ✗ Collapsed paths when there is no SV	✓ Nearby SVs may be phased automatically ✗ Haplotype-fused paths
het-SNP	300 b – 10 kb 	✓ Easy to group SNPs/reads into different haplotypes ✗ No phasing information associated with SVs	✓ Haplotype-specific paths ✗ More fragmented contigs

MERGE HAPLOTYPE INFORMATION AND "UNZIP"





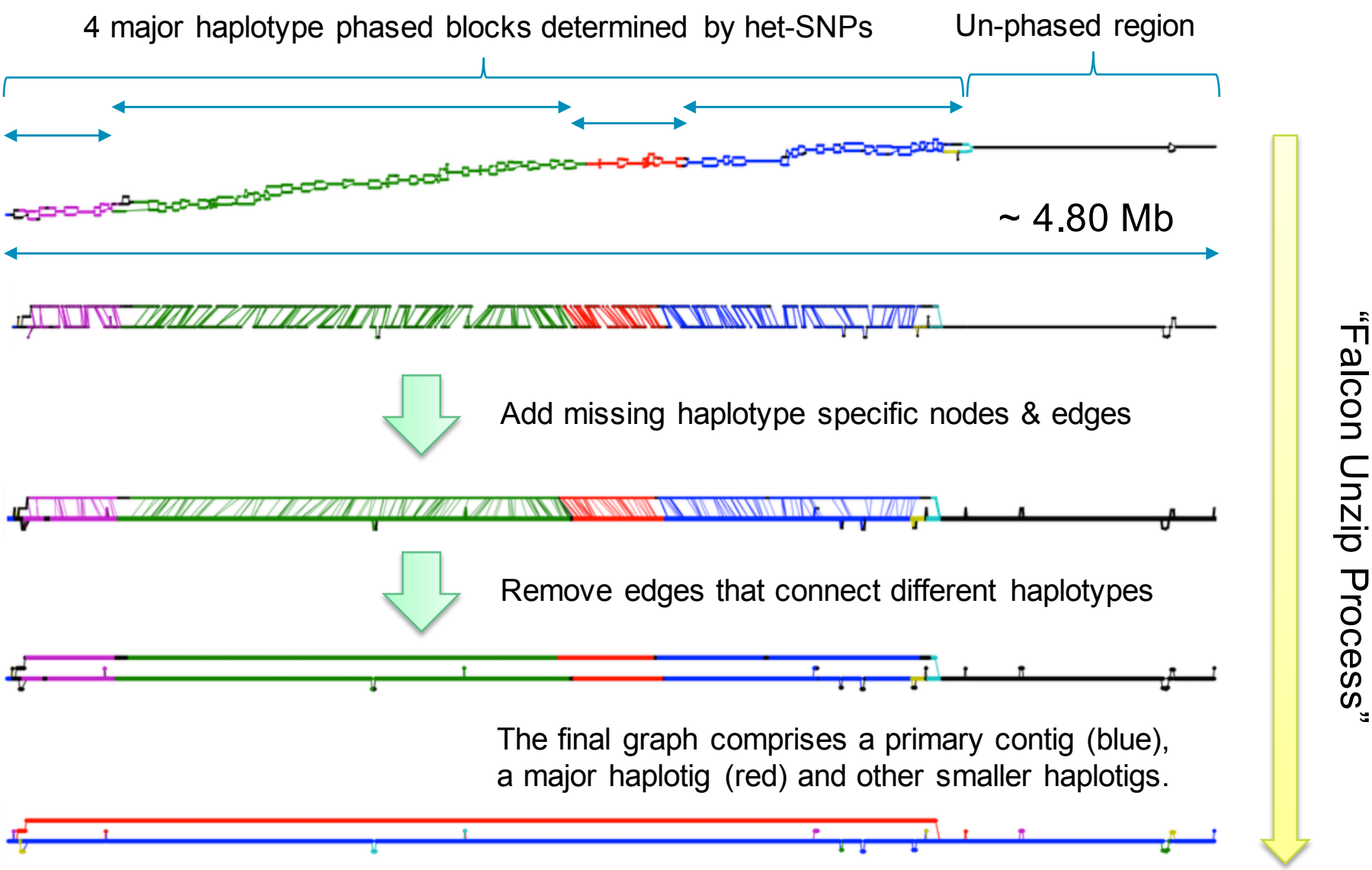
Tiling path of haplotype 0



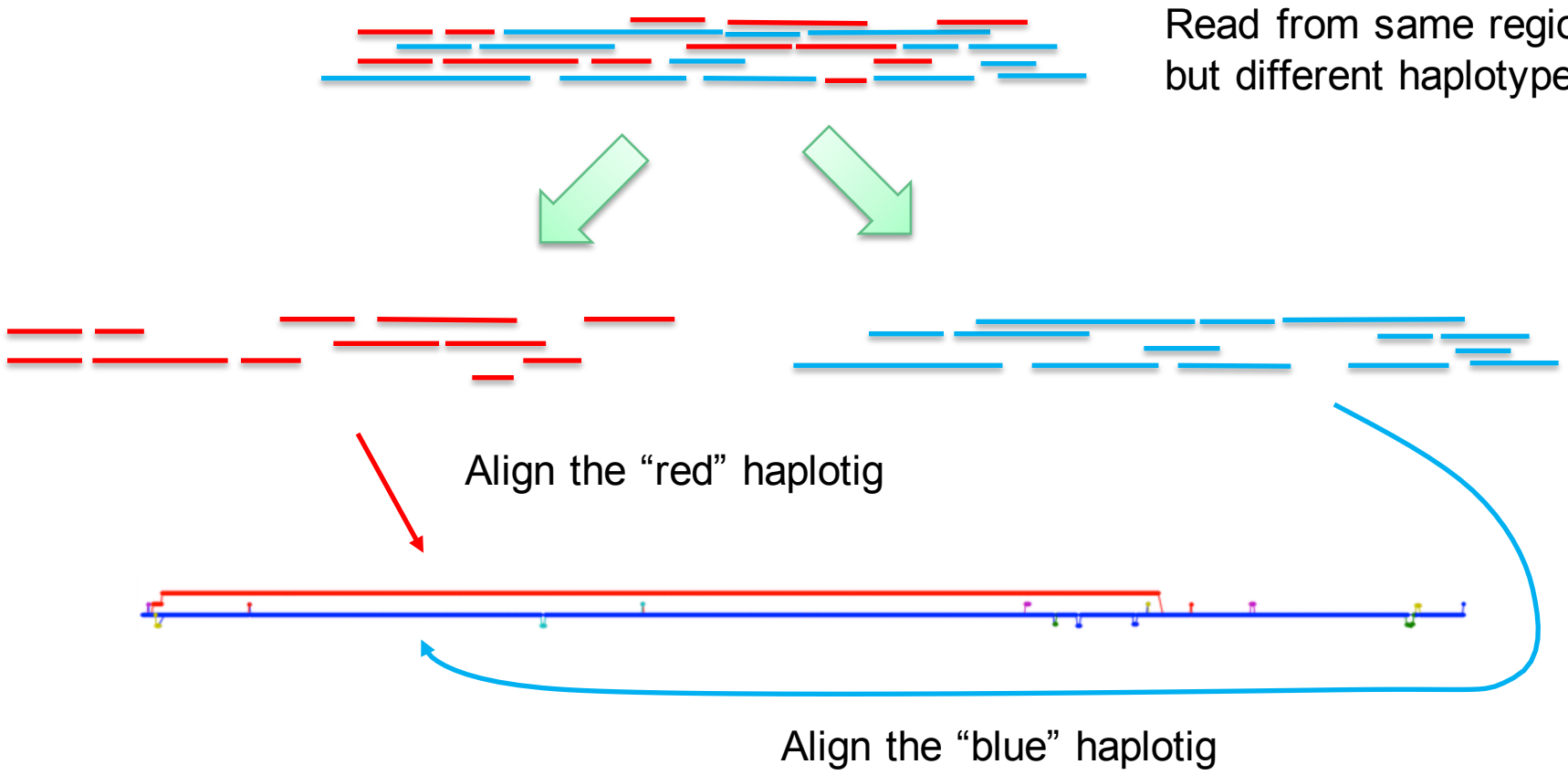
Remove edges connecting different haplotypes

Tiling path of haplotype 1

PUT EVERYTHING TOGETHER

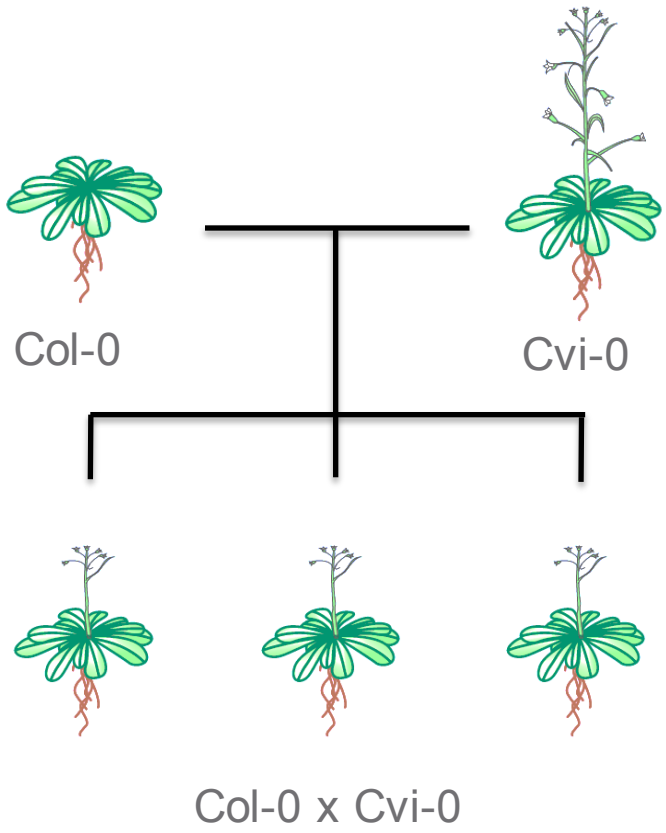


POLISHING: ALLELE-SPECIFIC ALIGNMENT FOR FINAL CONSENSUS



“Augmented alignment”: Each read has extra attribute (e.g., contig identifier, phasing block, haplotype phase), an aligner uses those information to place the read to specific reference sequence or regions.

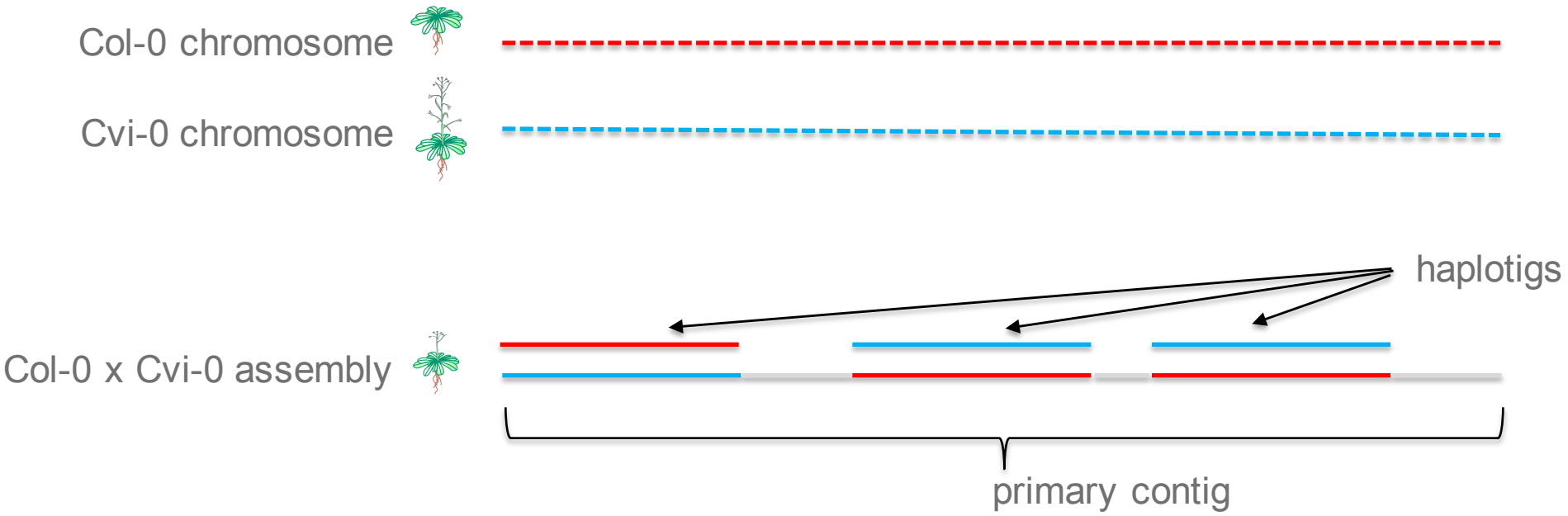
CONSTRUCT *ARABIDOPSIS THALIANA* COL-0 X CVI-0 DIPLOID F1 LINE



- Two inbred lines sequenced in 2013 (P4 chemistry), assembled as haploid genomes
- F1 line constructed and sequenced in 2015 (P6 chemistry), assembled with FALCON and FALCON-Unzip

Image credits: Pajoro, et al, Trends in plant science 21.1 (2016): 6-8.

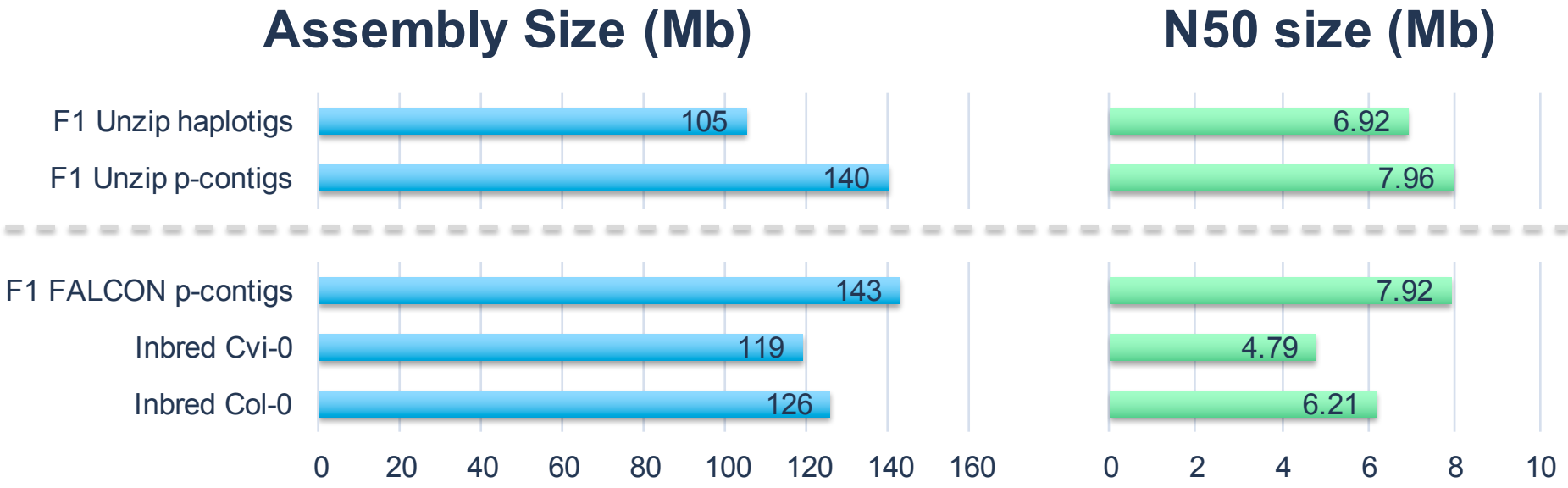
DIPLOID ASSEMBLY PRIMARY CONTIGS AND HAPLOTIGS



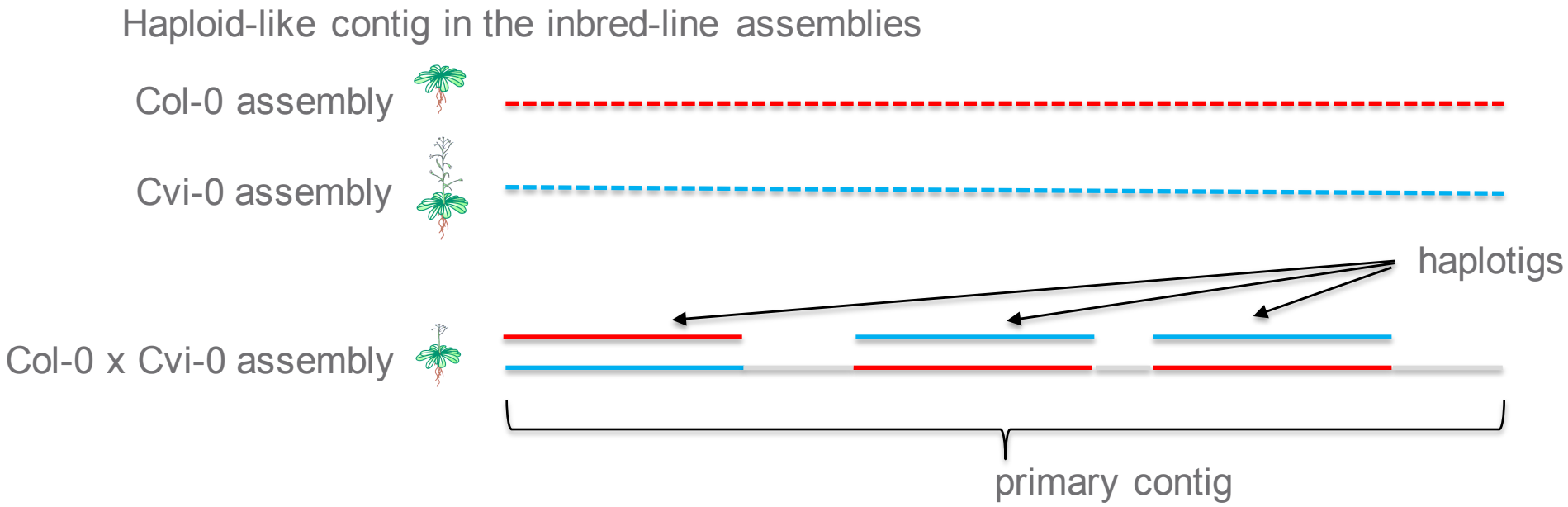
- Primary contigs ~ 1n representation of the genome
- Haplotigs ~ phased sequences from where the homologuous chromosomes are distinguishable

ARABIDOPSIS THALIANA F1 DIPLOID ASSEMBLY STATISTICS

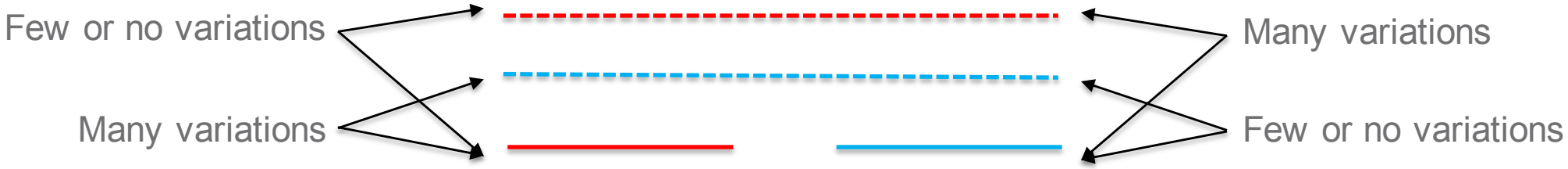
Strain	Inbred Col-0	Inbred Cvi-0	Col-0 x Cvi-0 F1		
Assembler	CA/HGAP	CA/HGAP	FALCON primary contigs	FALCON-Unzip primary contigs	FALCON-Unzip haplotigs
Assembly Size (Mb)	126	119	143	140	105
# contigs	1325	194	426	172	248
N50 size (Mb)	6.210	4.79	7.92	**7.96**	**6.92**
Max Contig size (Mb)	10.25	11.25	13.39	13.32	11.65



EVALUATE THE DIPLOID ASSEMBLY RESULT

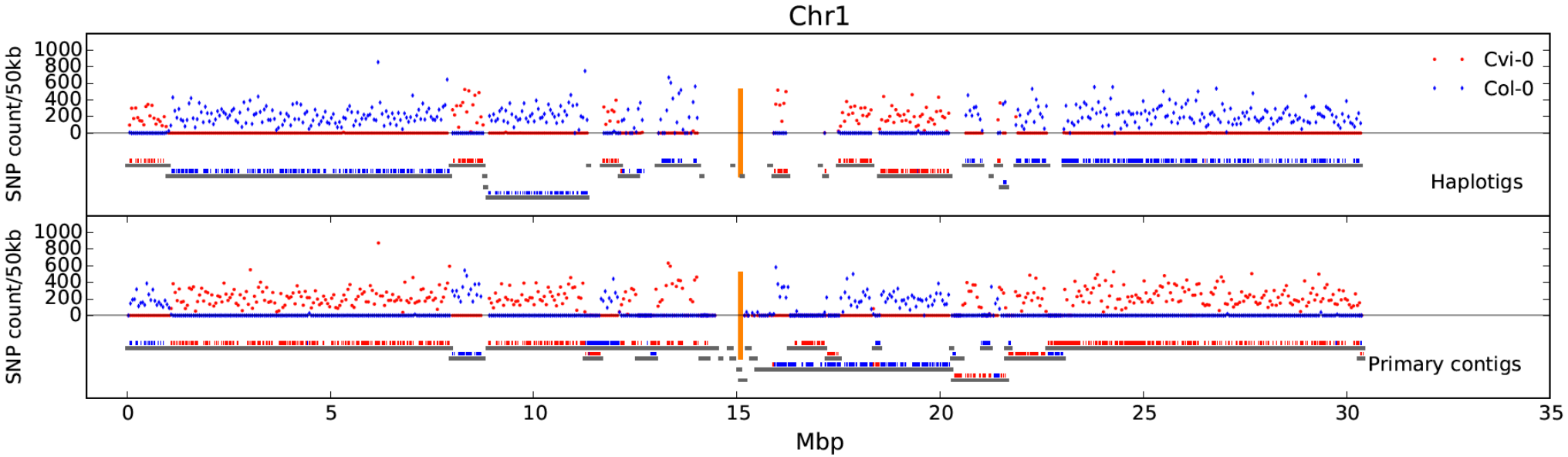


By aligning the haplotigs to the parental genome assemblies, we can evaluate the haplotigs' quality, e.g. haplotyping accuracy and CDS prediction consistency.



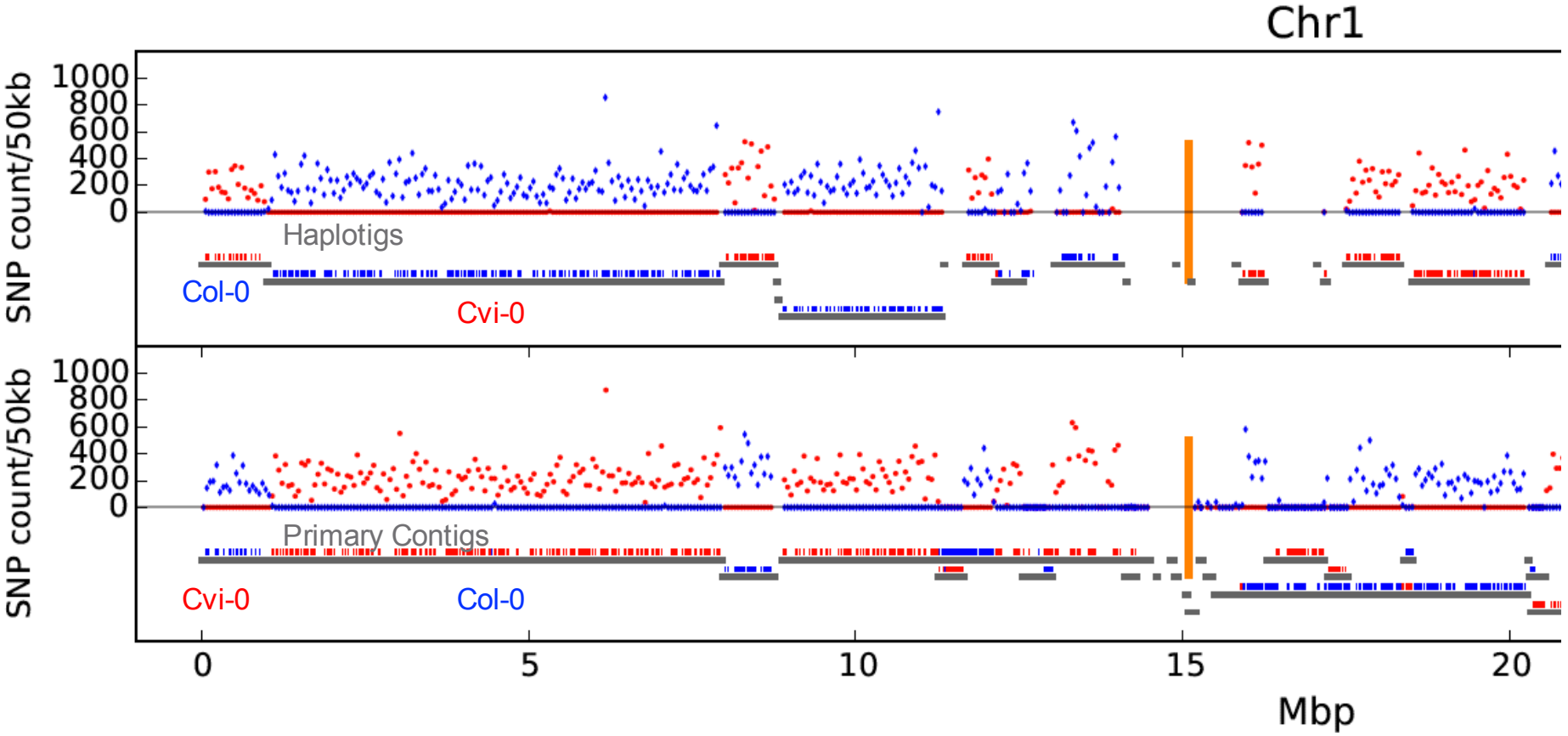
COMPARE F1 ASSEMBLY TO THE INBRED ASSEMBLIES

- We call the SNP and SVs against the parental inbred assemblies for all primary contigs and haplotigs.
- Most haplotigs can be fully assigned to one of the parental haplotypes.



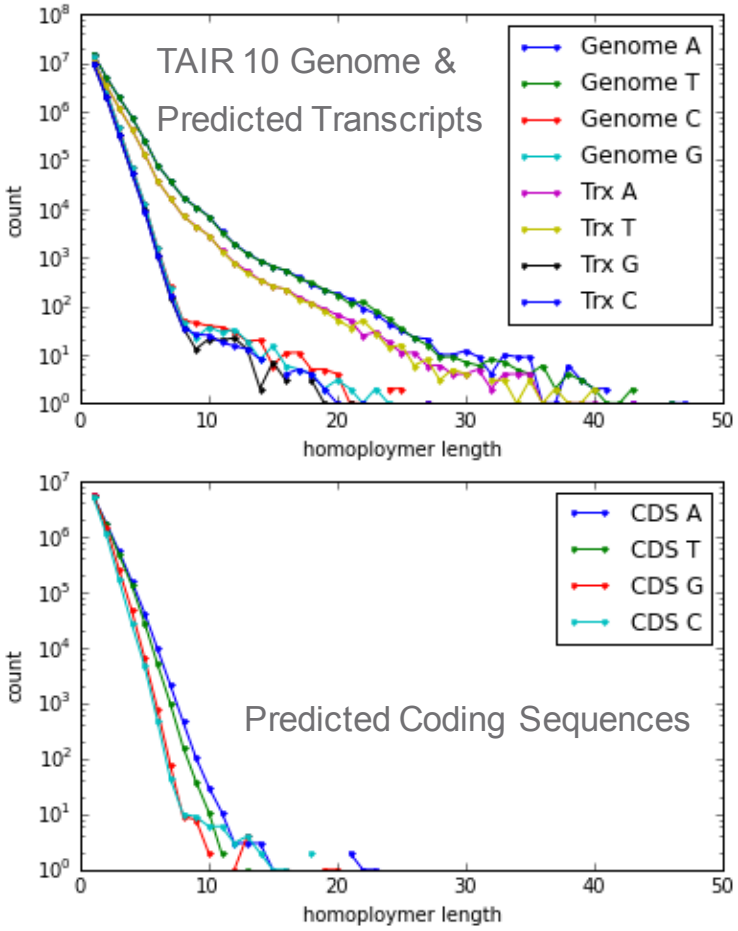
COMPARE F1 ASSEMBLY TO THE INBRED ASSEMBLIES

- We call the SNP and SVs against the parental inbred assemblies for all primary contigs and haplotigs.
- Most haplotigs can be fully assigned to one of the parental haplotypes.



ANNOTATION COMPARISION

Homopolymer Length Distributions



Compare *de novo* gene prediction (with AUGUSTUS (Stanke 2003)) between different assemblies

Assemblies		TAIR 10	Col-0	Cvi-0
	Number of predicted CDS	27,946	30,006	27,393
		100% indel-free full length overlaps		
Col-0	30006	25,966 (92.9%)		
Col-0 x Cvi-0	56775	25,865 (92.5%)	26,537 (88.4%)	27,370 (99.9%)

OTHER SMALLER AND LARGER DIPLOID GENOMES

	Clavicorona pyxidata (Coral Fungus)	Cabernet Sauvignon+*	Human*
Haploid Genome Size:	~ 44 Mb	~ 500 Mb	~ 3 Gb
FALCON-Unzip Results: Primary contig size	41.9 Mb	591.0 Mb	2.76 Gb
Primary contig N50	1.5 Mb	2.2 Mb	22.9 Mb
Haplotig size	25.5 Mb	372.2 Mb	2.0 Gb
Haplotig N50	872 kb	767 kb	330 kb

+Led by Cantu lab, UC Davis and Cramer lab, UN Reno

*Preliminary results. Fast file system and efficient computational infrastructure are currently needed for large genomes.

SUMMARY

- Single data type for routine diploid assembly
- Large genomes are more computationally challenging but it is mostly an engineering problem now:
  - Haplotype phasing improvement, incorporate 3rd party phasing code
  - Develop a sequence aligner for "augmented alignment" for faster Quiver consensus process
- FALCON-Unzip code: (No code, No truth!!) if you like to hack it for now, email me (jchin@pacb.com)
- Want to attack the algorithm problem for polyploid assembly? Let us help you!

Thanks for your attention!

www.pacb.com