

De Novo Diploid Genome Assembly and Haplotype Sequence Reconstruction

Jason Chin, Paul Peluso, David Rank / SFAF2015

FIND MEANING IN COMPLEXITY

Acknowledgement



This talk is about how to ~~make some interesting animation~~ reconstruct diploid genome with **continuous long reads**.

Arabidopsis Samples:

Joe Ecker
Chongyuan Luo
Ronan Omalley

String graph / daligner:

Gene Myers

All PacBio Colleagues

Open source tools :

Gephi
Graphviz
Python/NetworkX
Mummer3

Challenges Ahead

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Haplotype-resolved genome sequencing: experimental methods and applications

Matthew W. Snyder¹, Andrew Adey², Jacob O. Kitzman^{3,4} and Jay Shendure¹

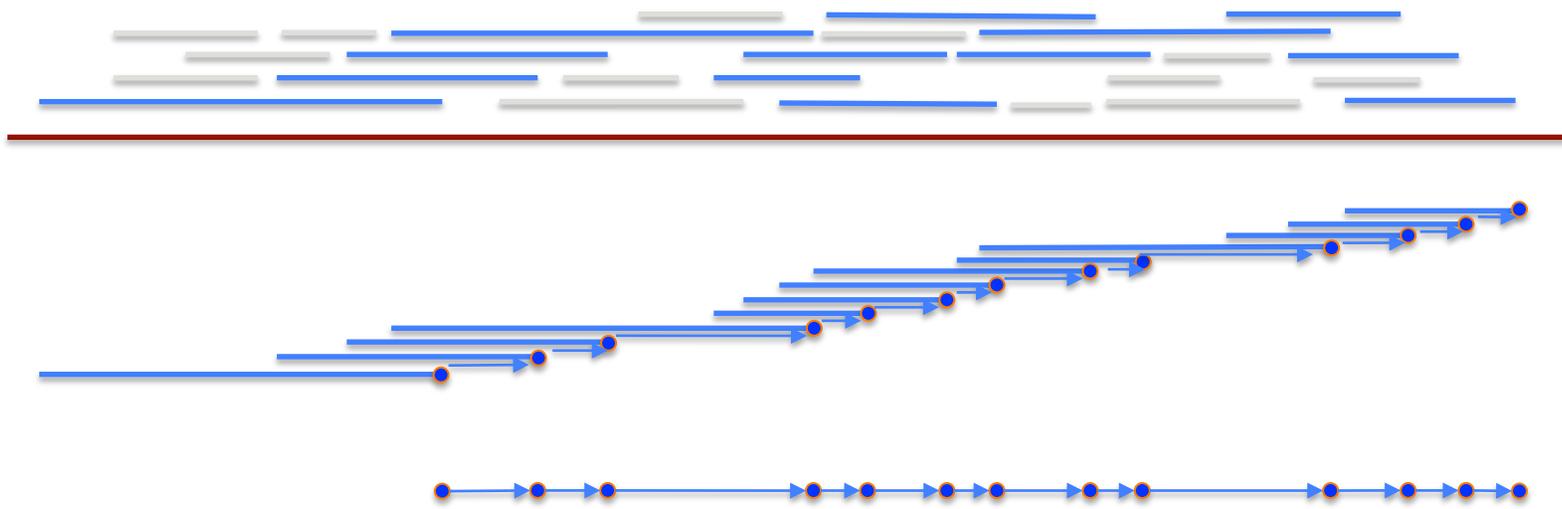
Nat Rev Genet. 2015 Jun;16(6):344-58

Can we just construct haplotype sequences *de novo* rather than calling and connecting just some sets of variant calls on top of references?

Comprehensiveness of variant types. Haplotypes consist of the full spectrum of genetic variation, including SNVs, short insertions and deletions (indels), structural rearrangements and copy-number polymorphisms. However, most methods for haplotype inference operate only at the level of SNVs and are restricted to unique single-copy sequences on autosomes. The development of algorithms that are capable of integrating multiple variant types into comprehensive assembled haplotypes represents an important challenge for the field.

As discussed above, many of the methods for experimental genome-wide haplotyping separate the genotyping step (data from which heterozygous variants are called) from the haplotyping step (data from which heterozygous variants are phased). Although in principle this could be extended to all forms of variation, challenges include the fact that calling of indels and structural variation from shotgun, short-read sequencing data remains challenging as well as the fact that complex structural variation may confound the algorithms used for calling haplotypes from dense or sparse experimental data. Progress towards this goal may be helped by Phase 3 of the 1000 Genomes Project, in which the phasing of all forms of genetic variation by inferential methods is an explicit goal.

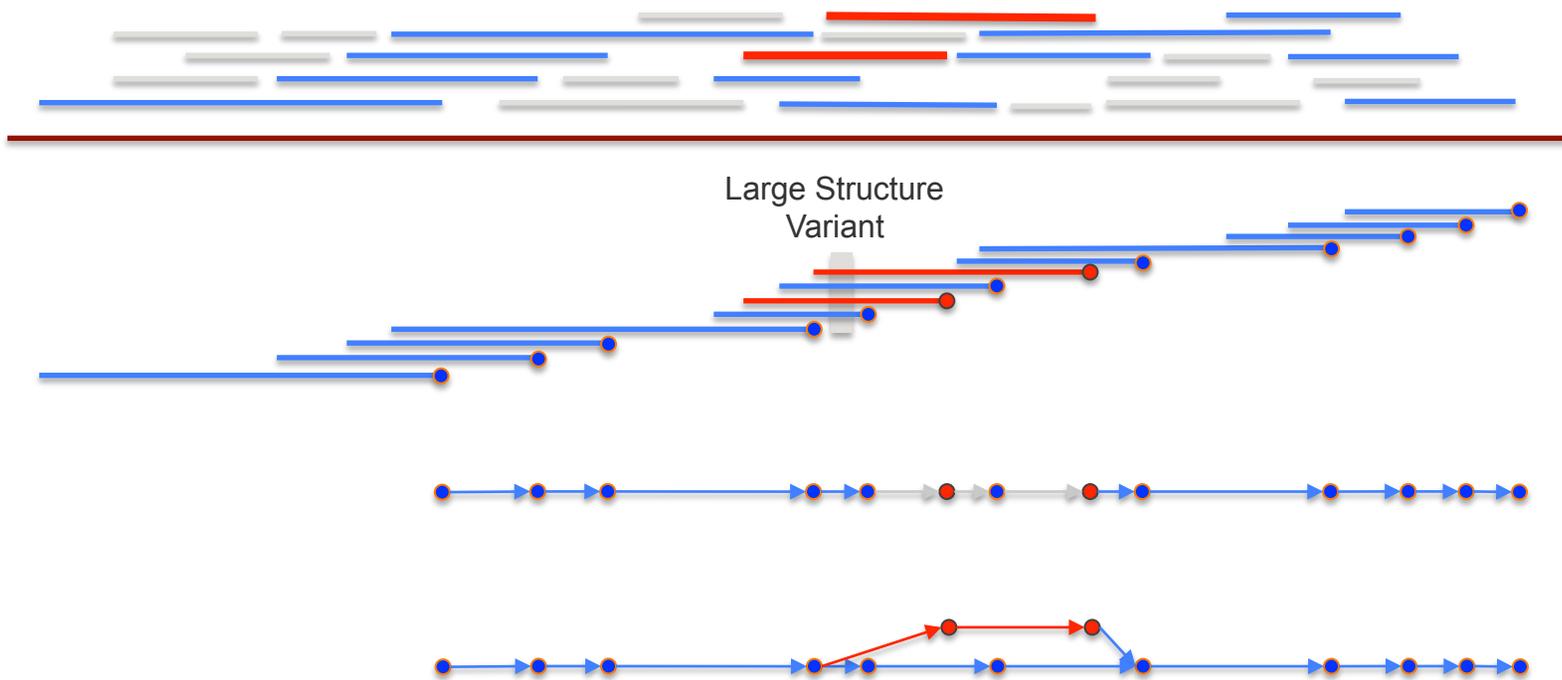
String Graph and *De Novo* Genome Assembly



String graph assembly for continuous long reads:

1. Remove contained reads (gray)
2. Overlaps to string graphs and tiling paths (blue)
3. The tiling path is corresponding to a path in the string graph.

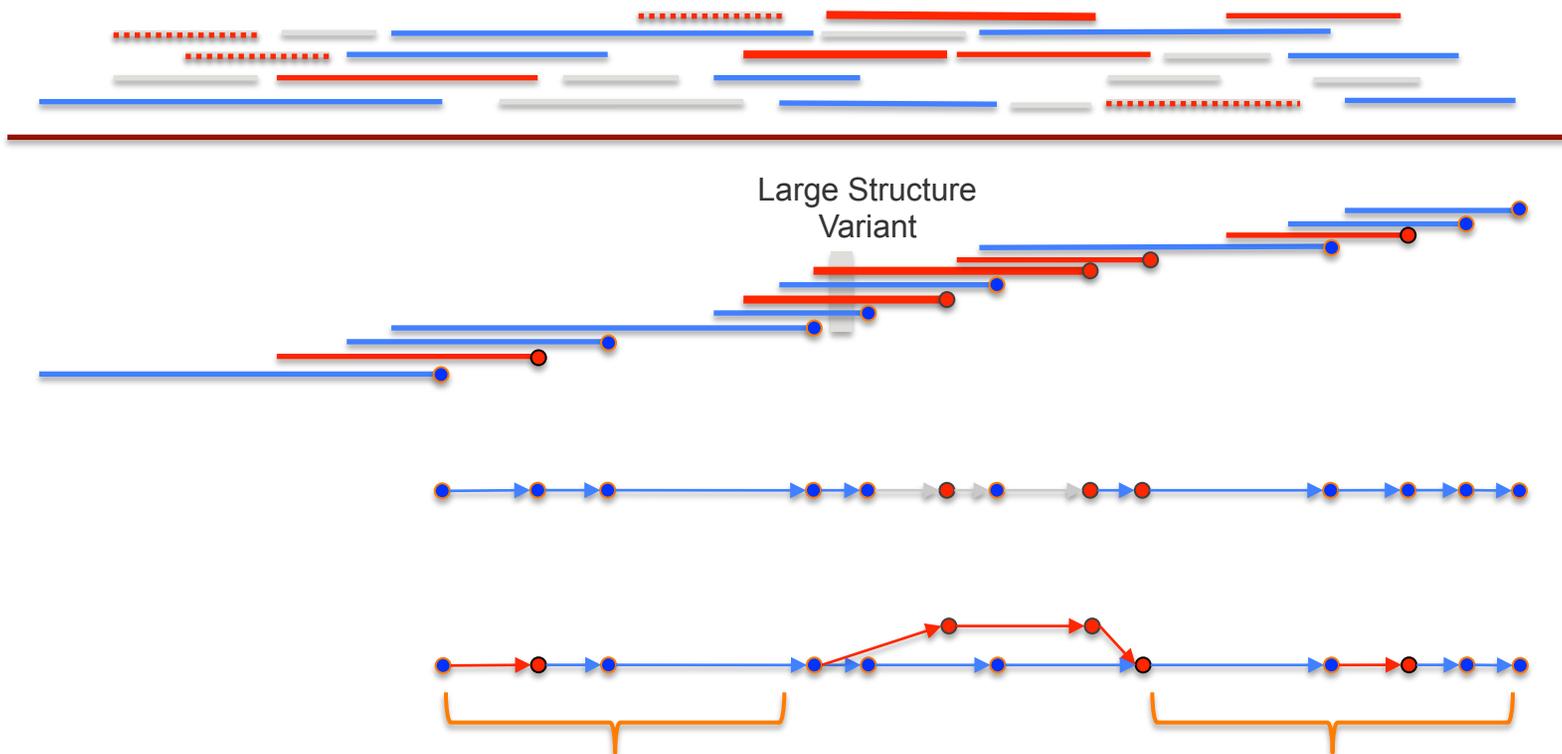
Polymorphism Causes “Bubbles” in the String Graph



Structure variants between haplotypes can create bubbles in the string graph

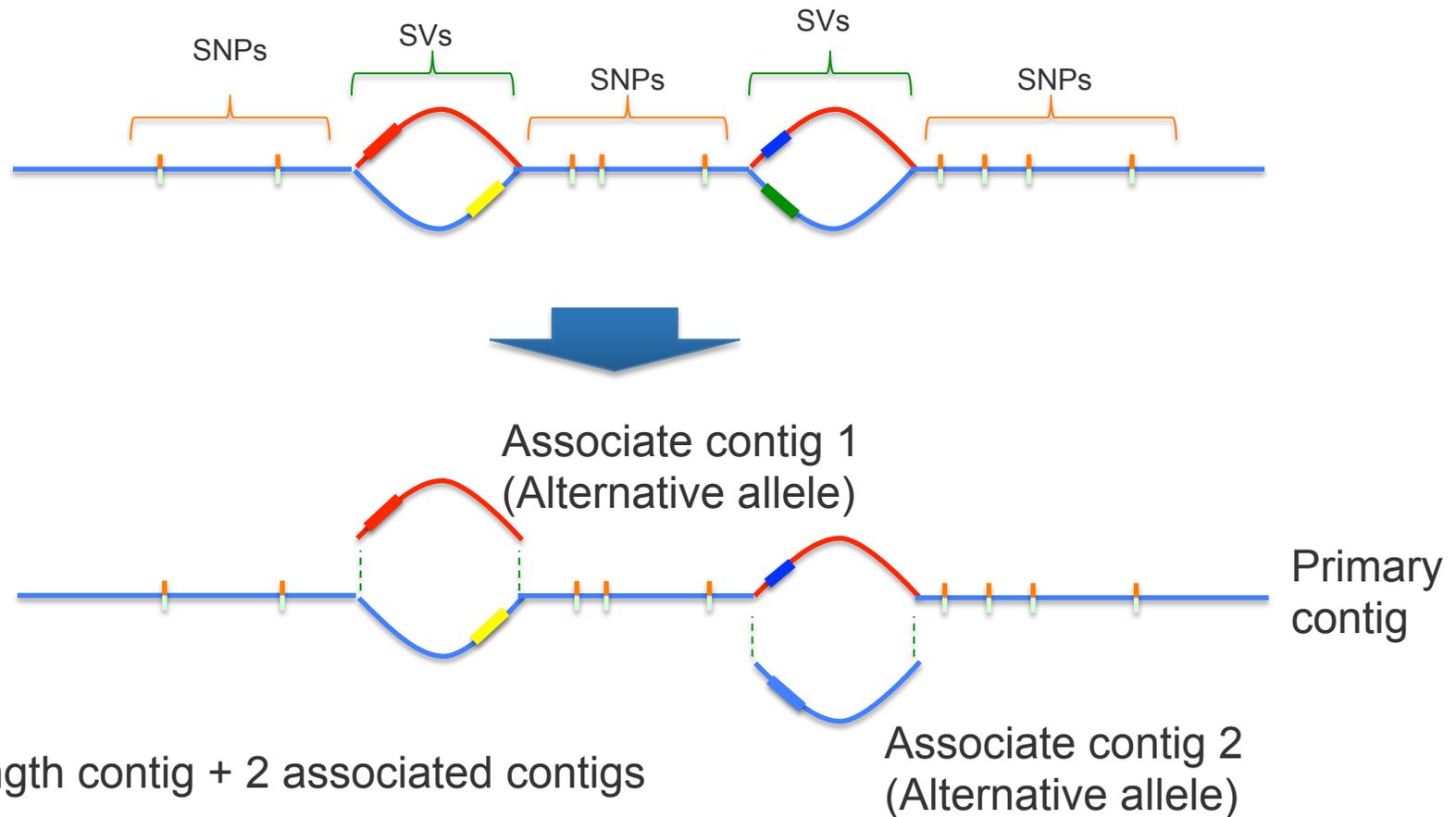
String Graph Structure and Subtle Base-Level Polymorphism

Reads (dashed red) contained in the other different haplotype might be missed in the graph



Fused paths where small base-level differences are collapsed

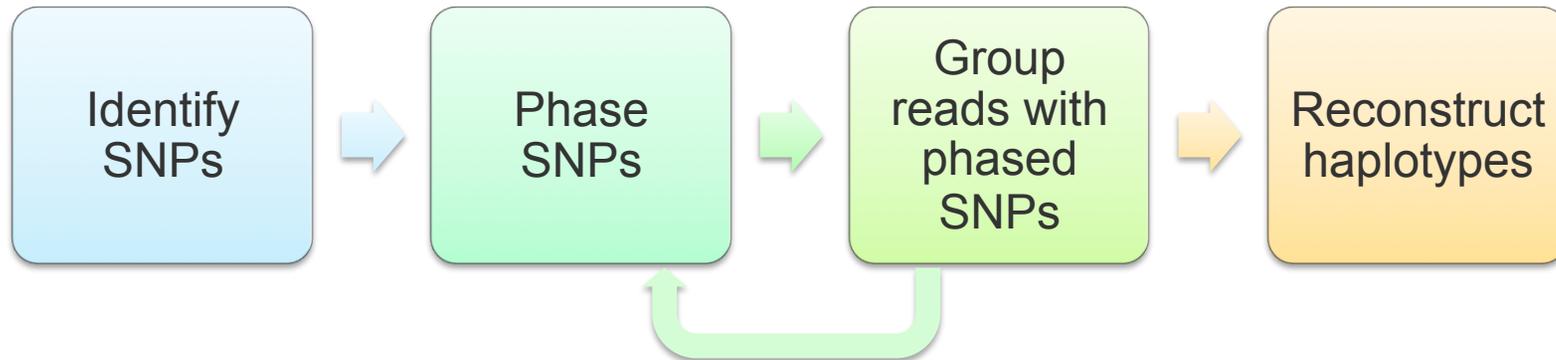
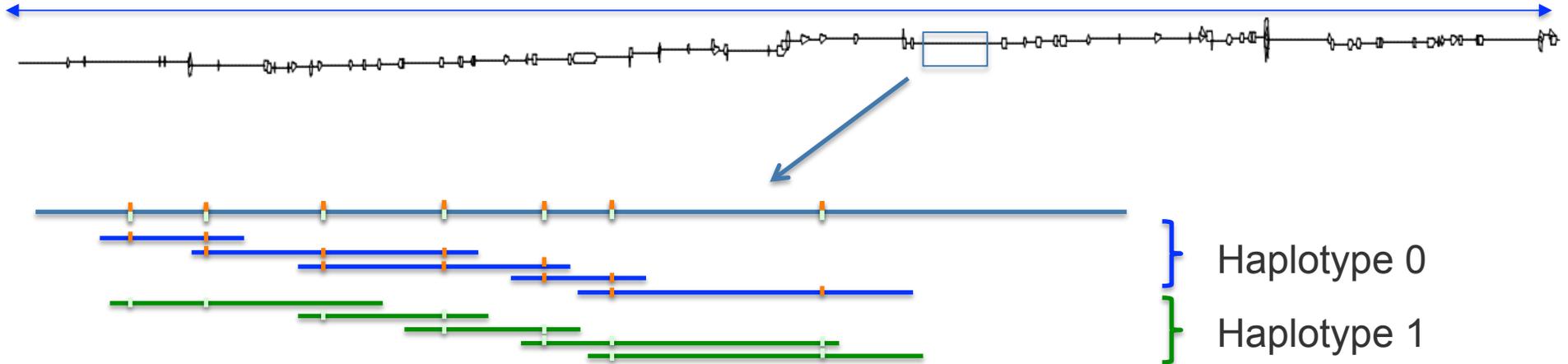
Stepping Stone Toward Haplotype Reconstruction to Catch All Variants



Keep the long-range information while maintaining the relations of the alternative alleles.

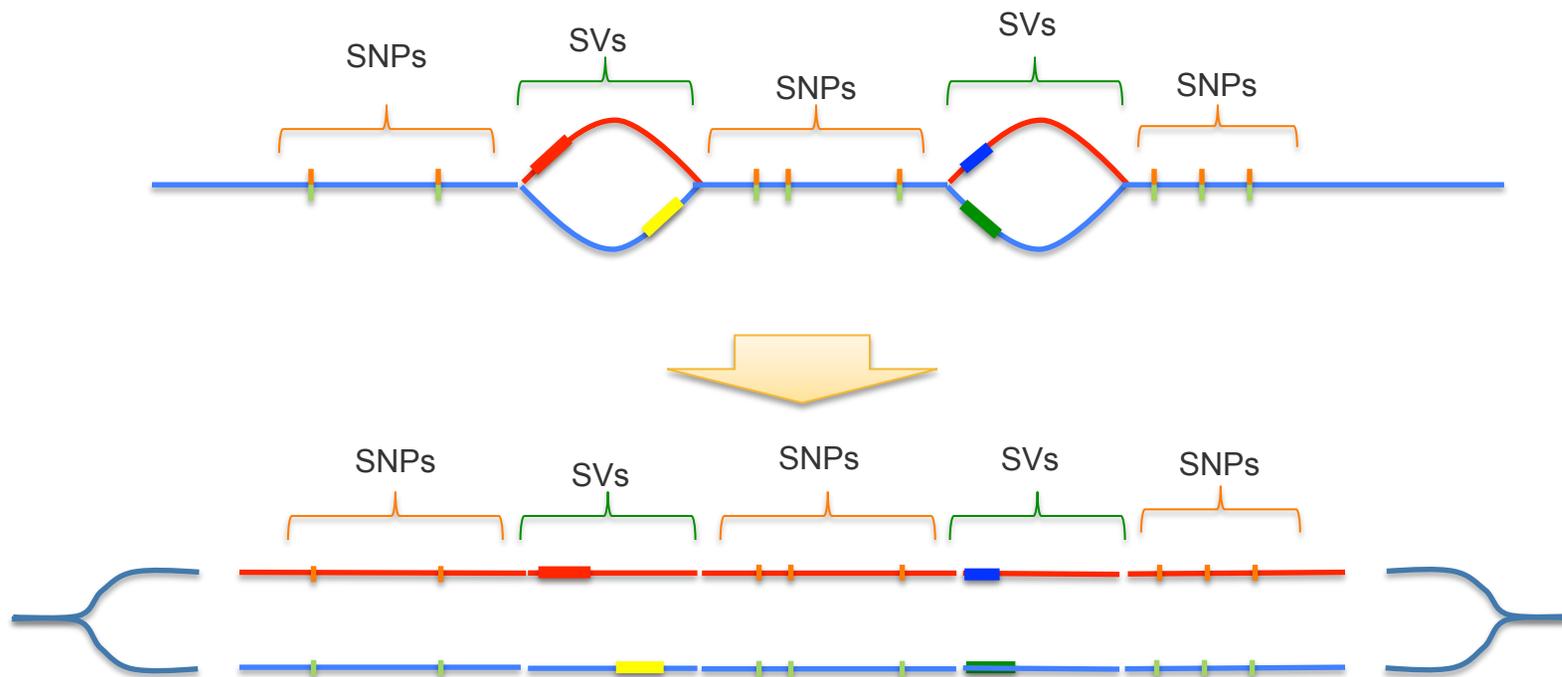
Phasing Variants Through Higher Identity Regions

A 9 Mbp contig spanning through the MHC region of a diploid human genome



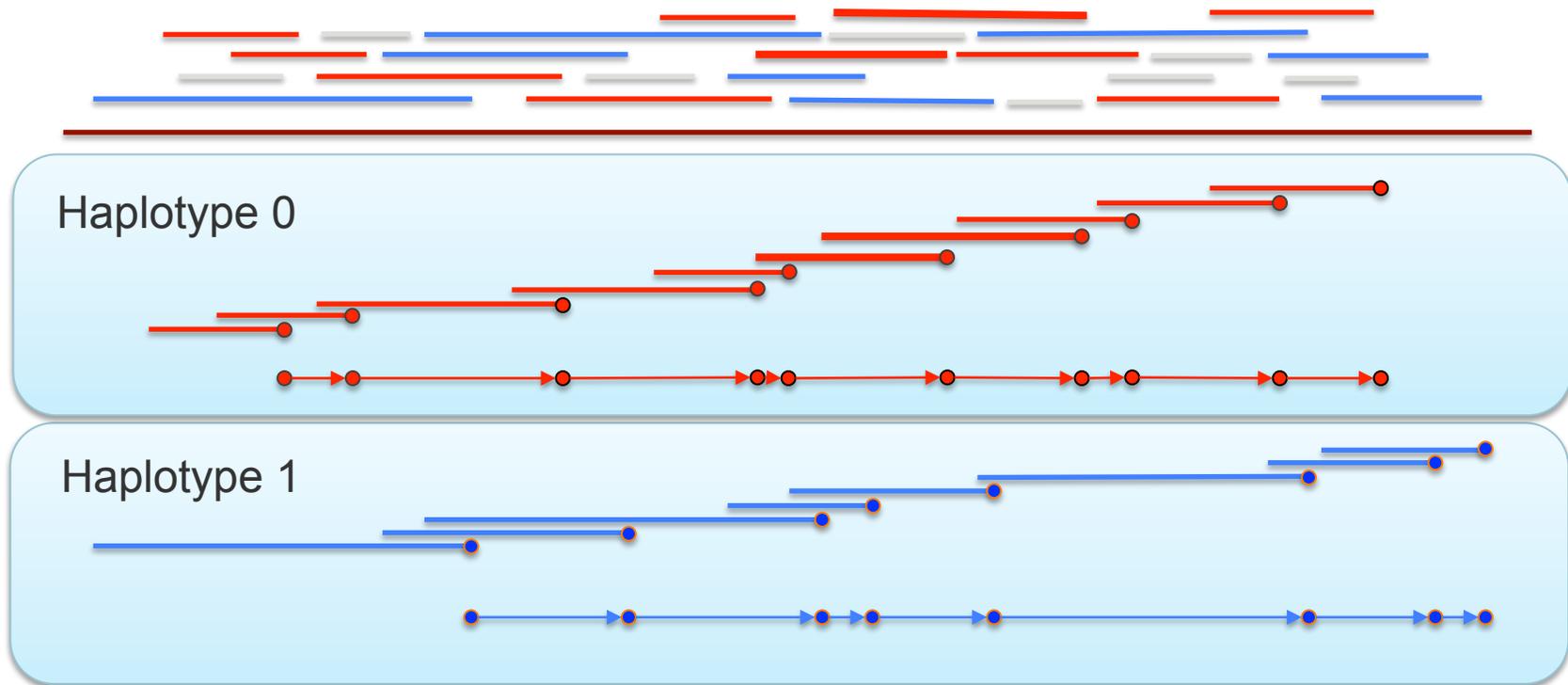
Group the SNPs and reads simultaneously for reconstructing haplotypes different only by small variations.

“Unzipping Collapsed Paths” with SNP Information



Chaining together all kinds of variants to assemble haplotigs for a diploid genome

When Everything is Perfect

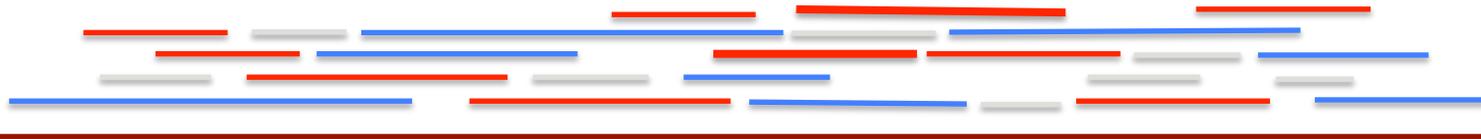


Haplotype 0

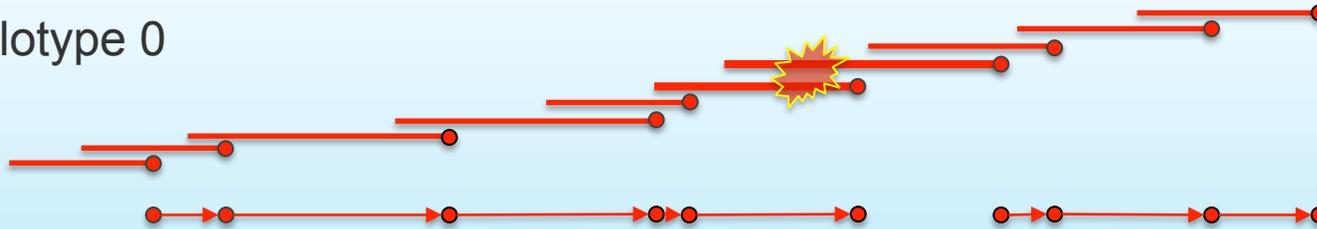
Haplotype 1

Problem Solved!!
(Only with perfect data and perhaps perfect “boring” genomes)

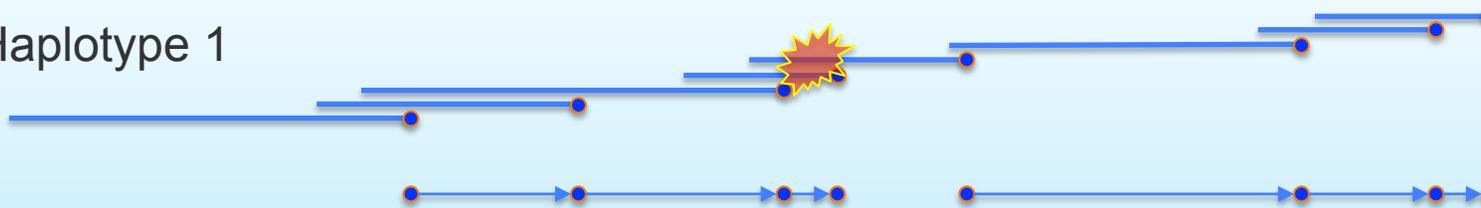
Structure Variations Can Fragment Haplotype Blocks



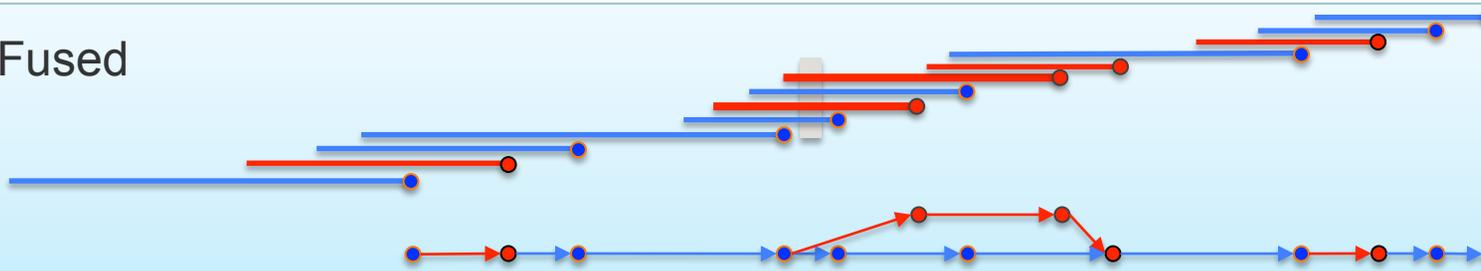
Haplotype 0



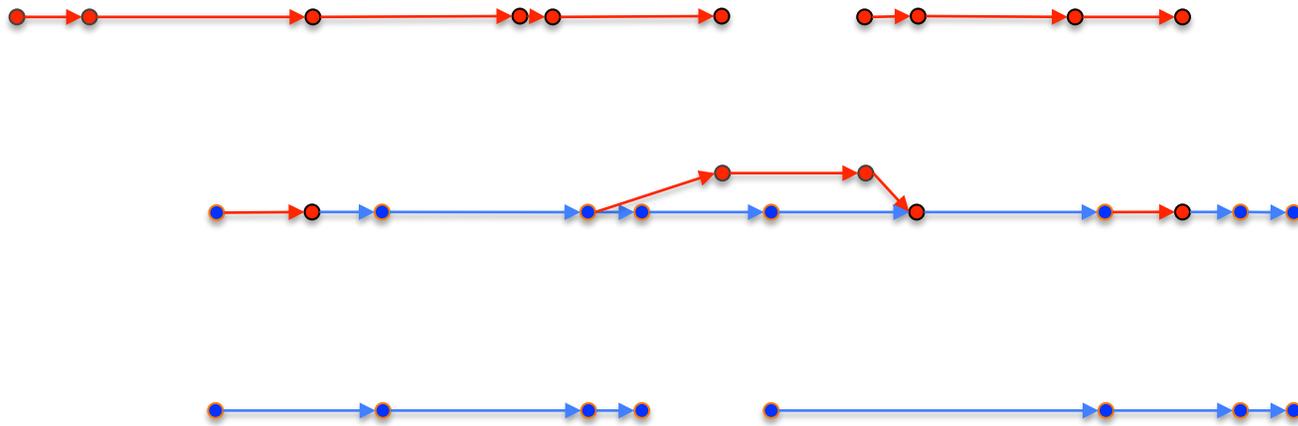
Haplotype 1



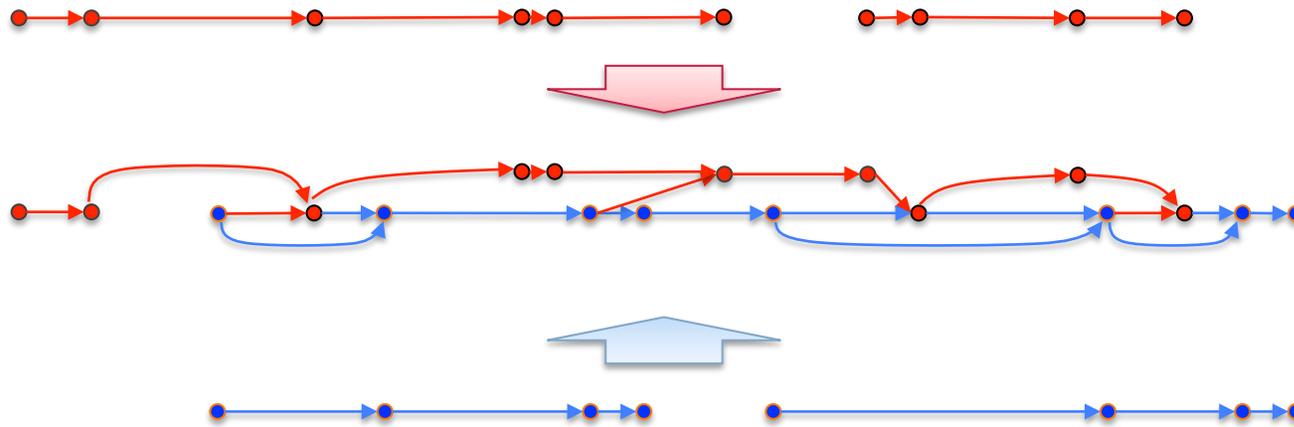
Fused



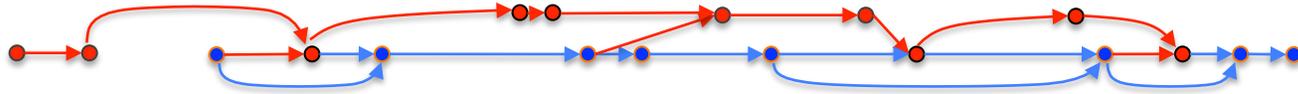
We Need to Combine the Graphs for Full Resolution



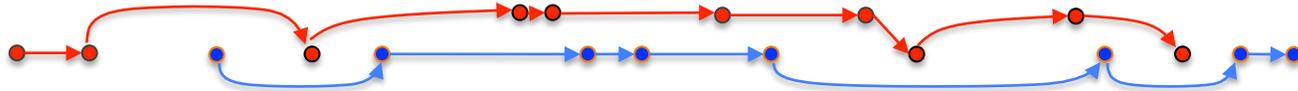
We Need to Combine the Graphs for Full Resolution



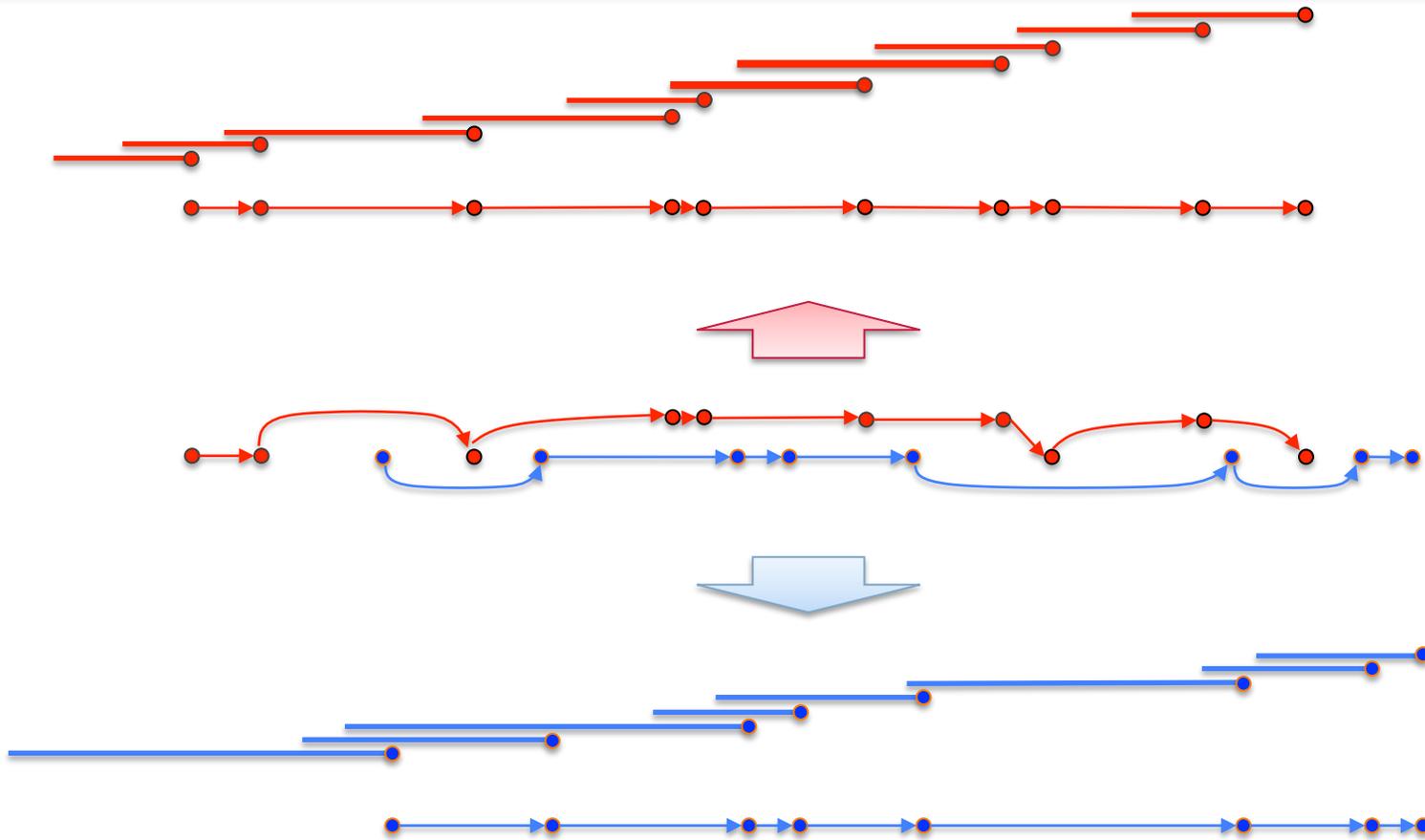
We Need to Combine the Graphs for Full Resolution



Remove “Crossing-Phase” Edges



Reconstructing Haplotigs with SV and SNPs



Arabidopsis Synthetic Diploid Genome For Algorithm Development

- Two inbred lines, CVI and Col-0, were sequenced separately about 1.5 years ago with P5C3 chemistry
- *In silico* mixture of the two datasets to emulate a diploid genome at about 80x coverage.
- Falcon assembly result:
 - N50 = 2.50Mb, Total : 130 Mb (primary)
 - Largest “fused contig” = 9.49 Mb
- High SV density: big SV every 80 kb
- High SNP density: SNP every 100 to 300 bp



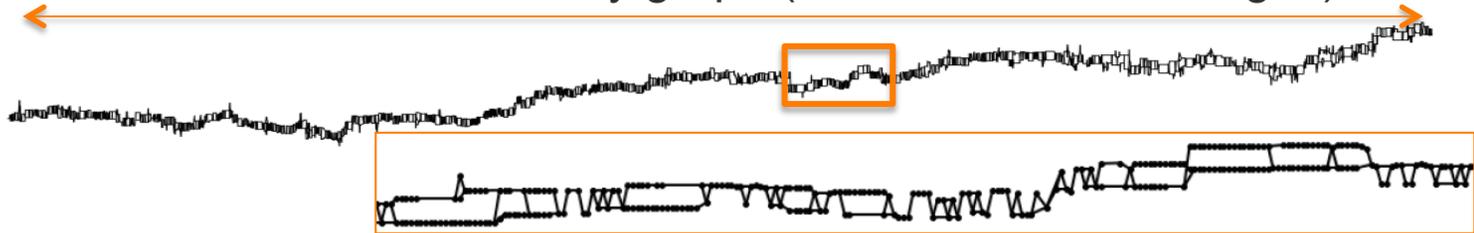
9.49 Mb Fused assembly graph



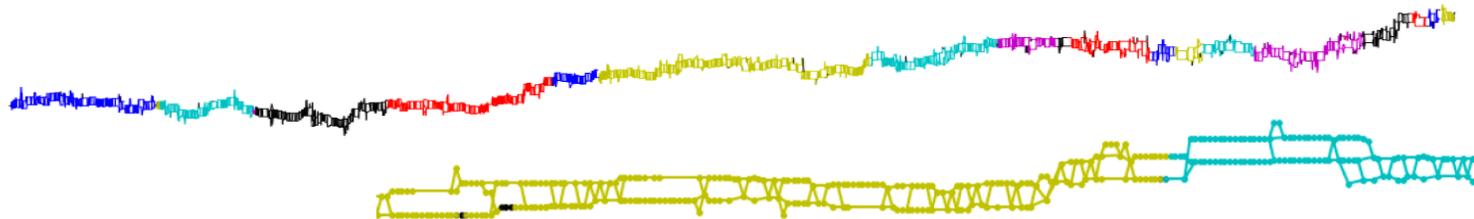
Falcon Unzip Results

9.49 Mb Fused assembly graph (7344 nodes, 8859 edges)

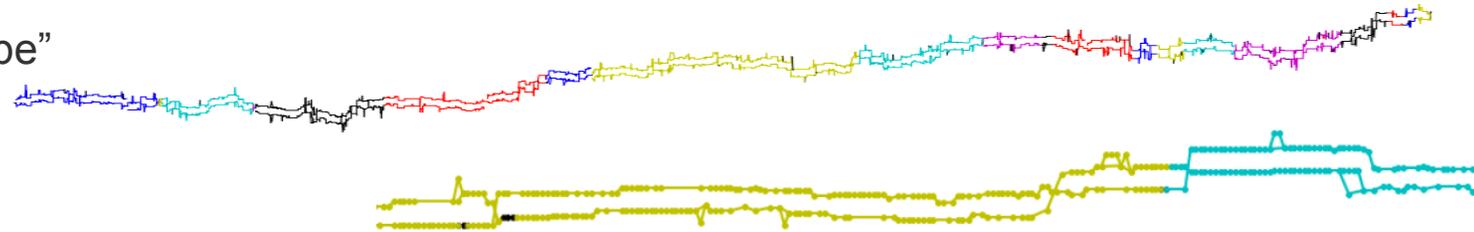
Fused
Assembly
Graph



Add phased
read
information

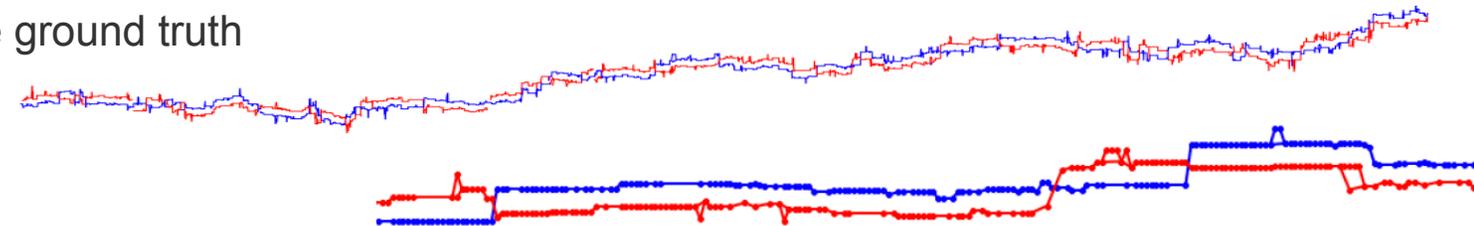


Remove
“cross-haplotype”
edges



Check with the ground truth

Blue: CVI
Red: COL-0



44 phased haplotigs:

N50 = 831.9 kb, Total: 18.4 Mb (~ 2 x 9.49 Mb), Max 1.39 Mb

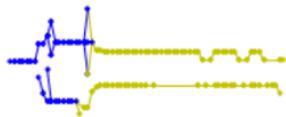
No switch error observed

HuRef MHC Region

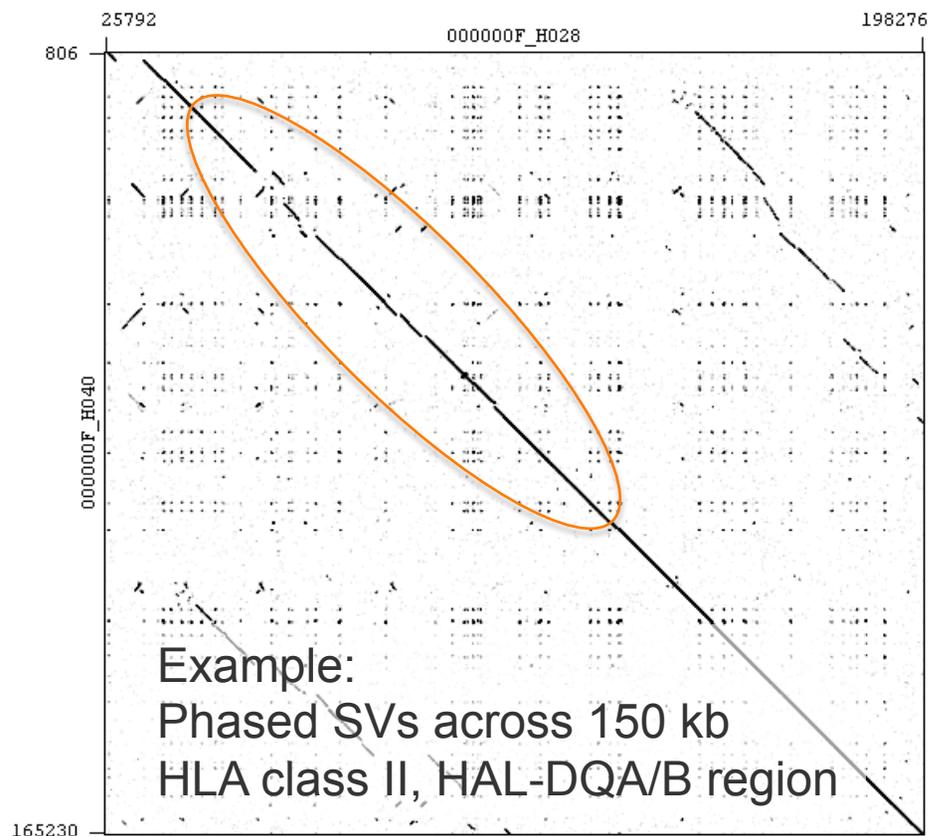
9.21 Mb Haplotig Assembly Graph (3079 nodes, 3997 edges)



Total 70 haplotigs
Total size 14,918,026 bp
N50 size 483,236 bp

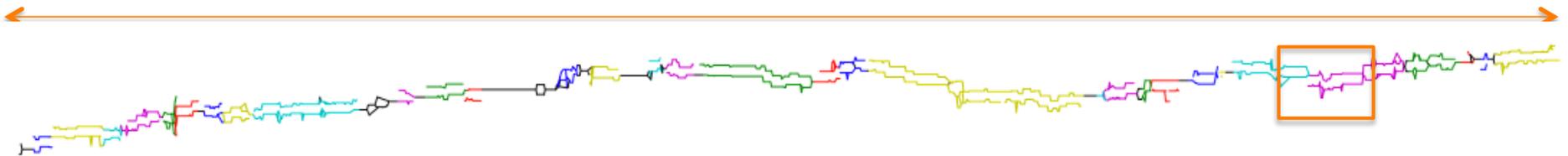


Haplotigs extended with SV information



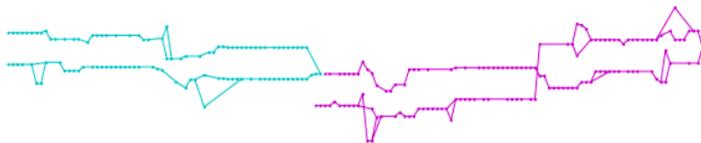
HuRef KIR Region

The HuRef KIR gene cluster is within one contig (5.56 Mb, 2041 nodes, 2495 edges)

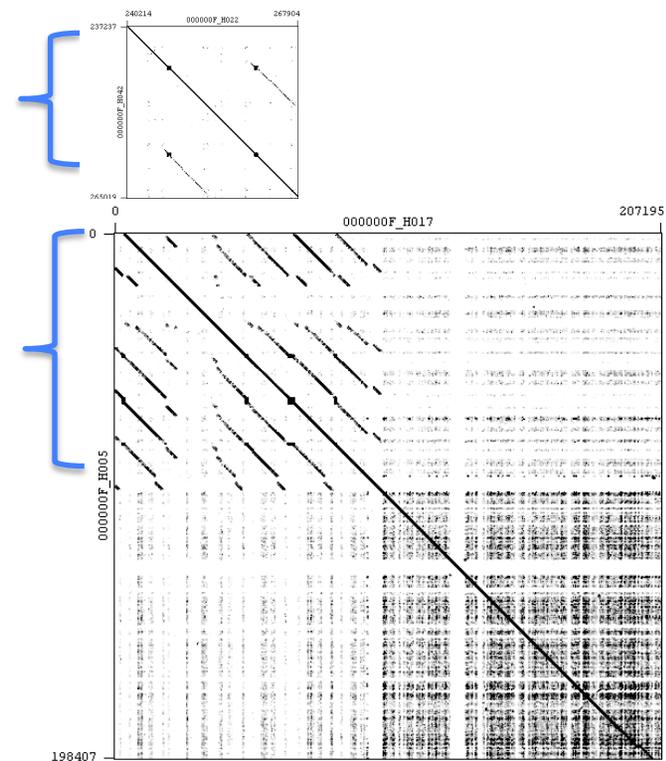


Total 74 haplotigs
Total size 9,405,867 bp
N50 size 254,502 bp

Phased
haplotigs
(span through
KIR2DL1 -
KIR3DL2)



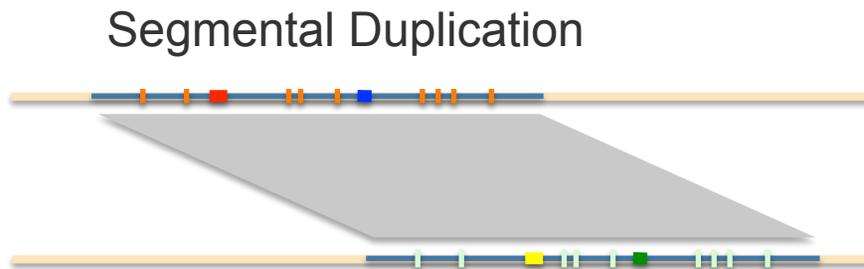
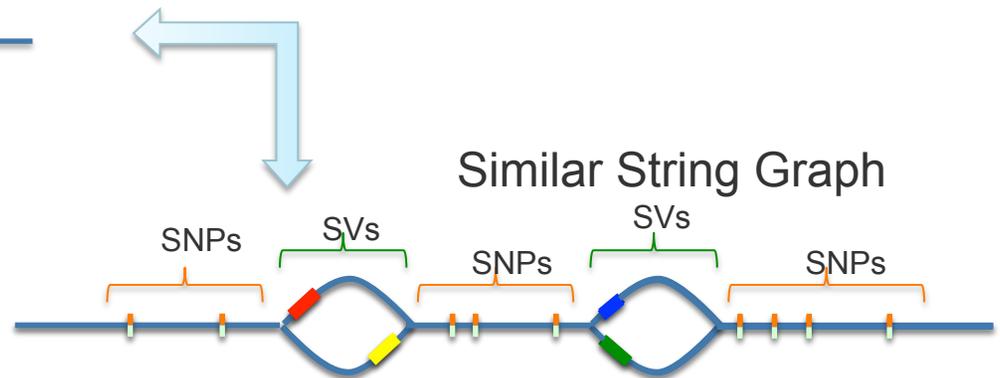
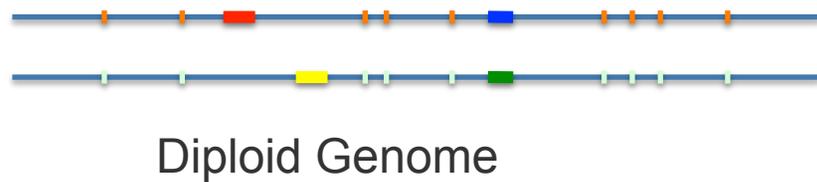
Haplotigs discontinuity caused by local repeats. (Need improved algorithm)



198407

20

Assembly Graph vs. Diploid Genome vs. Segmental Duplication



Resolve Segmental Duplication In Human Genome

A 360-kb interchromosomal duplication of the human HYDIN locus

Genomics 88 (2006) 762–771

Norman A. Doggett ^{a,*}, Gary Xie ^a, Linda J. Meincke ^a, Robert D. Sutherland ^a, Mark O. Mundt ^a,
Nicolas S. Berbari ^b, Brian E. Davy ^b, Michael L. Robinson ^{b,1}, M. Katharine Rudd ^c,
James L. Weber ^d, Raymond L. Stallings ^e, Cliff Han ^a

^a DOE Joint Genome Institute and Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

^b Division of Molecular and Human Genetics, Children's Research Institute, Ohio State University, 700 Children's Drive, Columbus, OH 43205, USA

^c Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C3-168, Seattle, WA 98109, USA

^d Center for Medical Genetics, Marshfield Medical Research Foundation, 1000 North Oak Avenue, Marshfield, WI 54449, USA

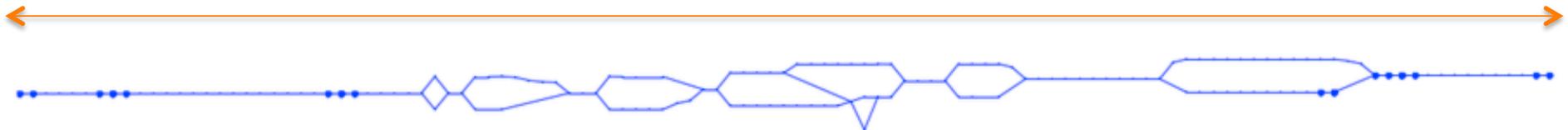
^e Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA

Missing in NCBI35/NCBI36, Unlocalized in GRCh36, Finished in GRCh38

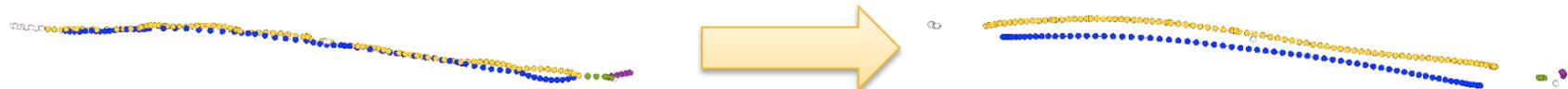
A CHM13 Contig Assembly Graph

(Mapped to GRCh38 chr16:70,811,384-71,168,671 and chr1:146,477,550-146,946,987)

421 kb



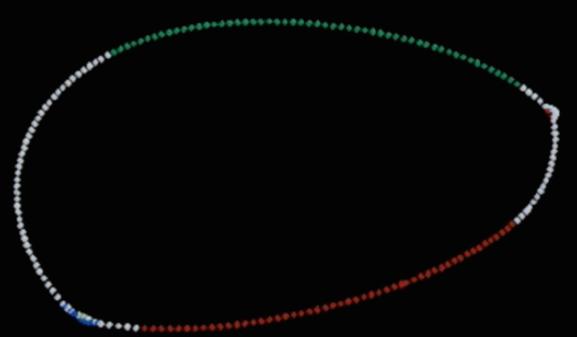
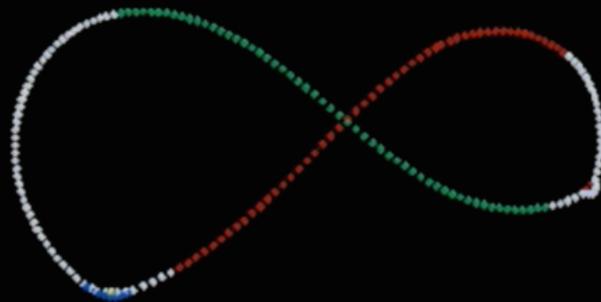
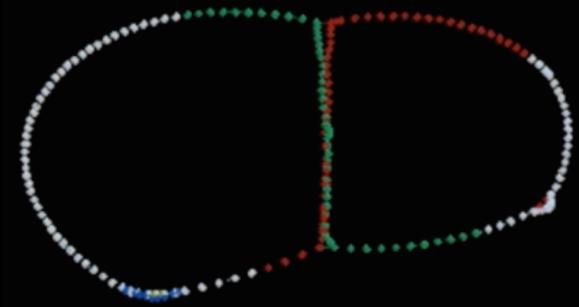
(In Discover NA12878 Assembly, this region has 13 contigs and 12 gaps.)



Falcon unzip

22

BAC Assembly With Internal Repeats

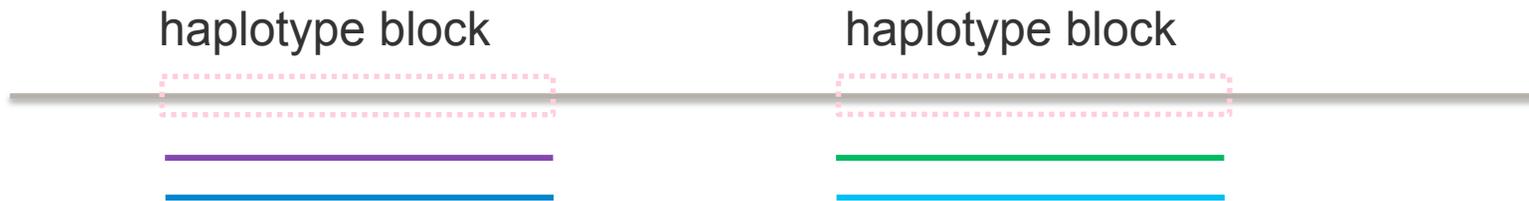


Representation of Haplotype Information

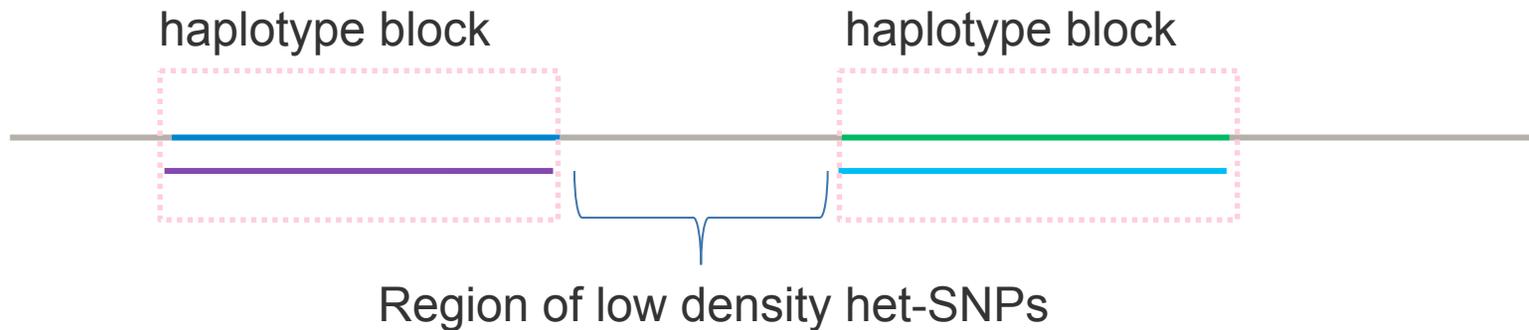
Unphased contig + phased variant calls



Phase-fused primary contig + ordered haplotigs



Primary contig with phased sequence + alternative haplotigs



Future Outlook

We just see the tip of the iceberg....

Re-sequencing with short reads:

Need a reference genome
Mostly SNP information

High contiguity assembly
with continuous long reads:

Resolve haplotype information *de novo*
Detect all structural variations
Better annotation
Build graph genome model
Enable comparative genomics
at chromosome scale and more



PACIFIC
BIOSCIENCES®

For Research Use Only. Not for use in diagnostic procedures. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, and Iso-Seq are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. All other trademarks are the sole property of their respective owners.