Reference-quality diploid genomes without *de novo* assembly

Michael Schatz

January 16, 2018
PAG Bioinformatics Workshop

@mike_schatz / #PAGXXVI

Why NOT *de novo*?

De novo is necessary for the first genome of a species, but is it really necessary for genomes 2 through N?

1. ***De novo is slow:*** Large inputs and intermediate files. Algorithms are complex to compare all the reads to each other. Mammalian genomes need thousands of core hours and terabytes of space

2. ***De novo is demanding:*** Make the libraries just right, sequence the reads just right, set the parameters just right, launch and relaunch to optimize

3. ***De novo is unpredictable:*** Add a little more (or a little less) data, and your contig N50 drops in half. Errors creep in ranging from SNPs to SVs. Heuristics break when the data or genome structure are unexpected.

4. ***De novo is just the beginning.*** Annotation from scratch is really hard, and to use it (variant analysis/selection analysis/regulatory analysis) you will probably need to align to a reference anyways.

Why NOT *de novo*?

De novo is necessary for the first genome of a species, but is it really necessary for genomes 2 through N?

1. ***De novo is slow:*** Large inputs and intermediate files. Algorithms are com… …mes ne…

2. ***D…*** …eads ju… …ize

3. ***D…*** …nd y… …SVs. H… …d.

4. ***D…*** …ard, an… …you will probably need to align to a reference anyways.

Wouldn't it be great if we could assemble high quality genome sequences and capture all the genetic diversity in a species without this complexity?

Reference-Guided Assembly

1. ***High quality reference***
   - Contig N50 over 1Mbp
   - Scaffold N50 over 10Mbp
   - High Quality Gene Annotation (See VGP definition)
   - Your sample is sufficiently similar (~99% identity)

2. ***Sample specific data***
   - *SNPs and Indels*: Illumina-based (Illumina PE or 10X)
   - *Structural Variants*: Long Reads (PacBio or ONT)
   - *Phasing Data*: 10X and/or HiC; trios when available

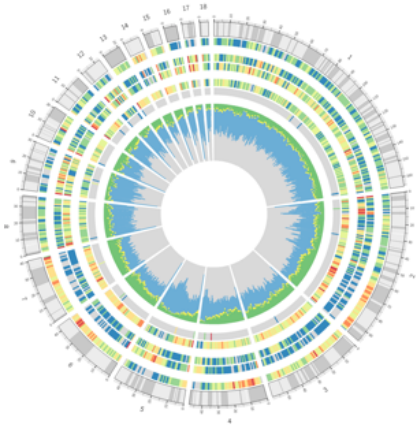
Data requirements similar to de novo, but less demanding, more accurate, and more predictable

Comparative Genome Assembly ("AMOScmp")

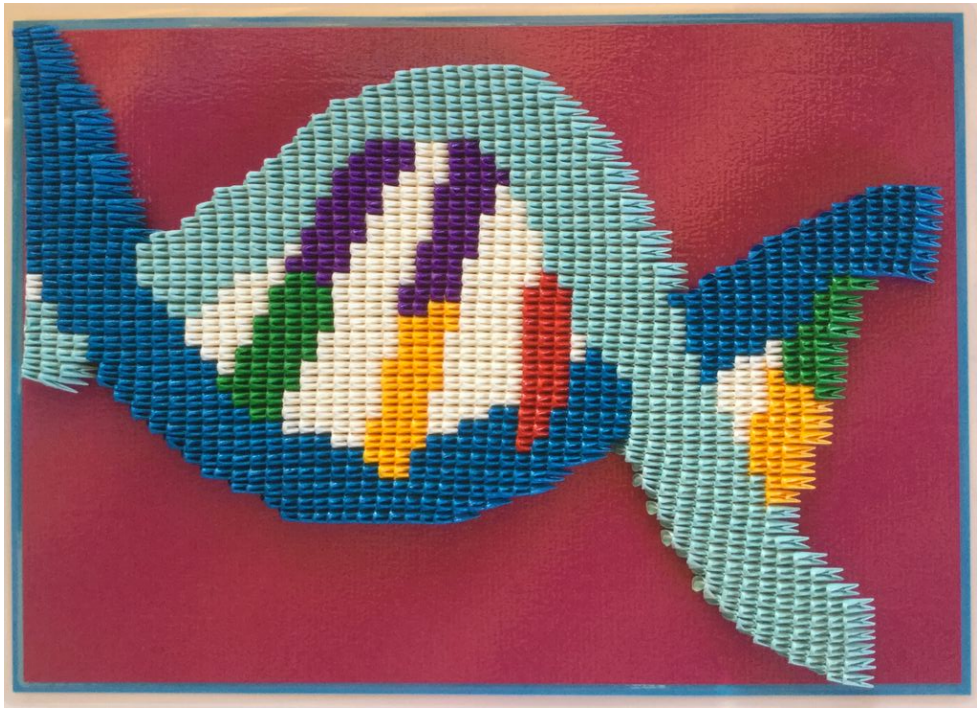
Pop *et al* (2004) *Briefings in Bioinformatics. Sep;5(3):237-48.*

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference



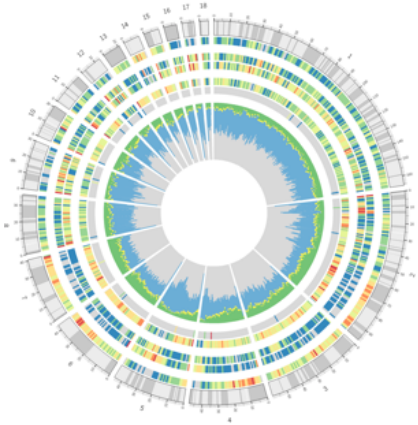
my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guido, Ring, & Gerstein labs

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference

A -> C

Call SNPs
+ Indels

A/C -> A|C

Phase SNPs
+ Indels

GAT
-> XXX

Call SVs

GAT/XXX
-> GAT|XXX

Phase SVs
& Local Assembly

Phased
SNPs,
Indels &
SVs.vcf

Assemble
Diploid
Sequence

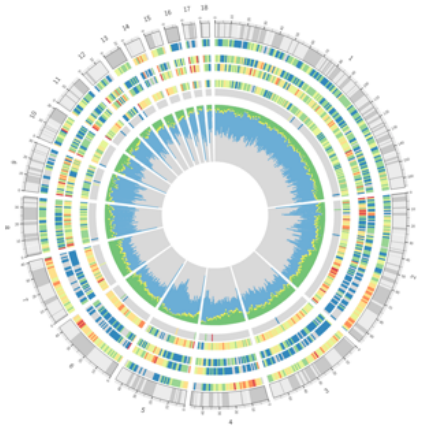
my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guido, Ring, & Gerstein labs

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference

A -> C

Call SNPs
+ Indels

A/C -> A|C

Phase SNPs
+ Indels

GAT
-> XXX

Call SVs

GAT/XXX
-> GAT|XXX

Phase SVs
& Local Assembly

Phased
SNPs,
Indels &
SVs.vcf

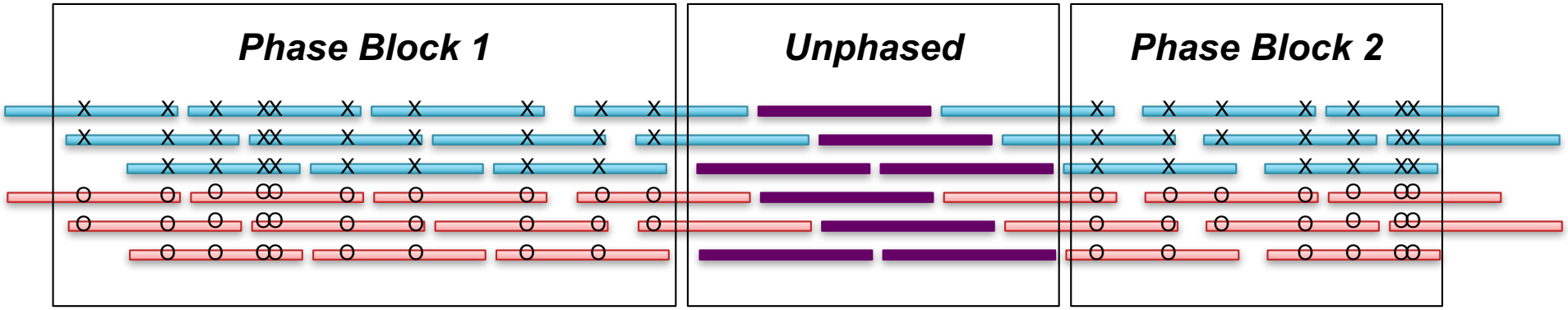
Assemble
Diploid
Sequence

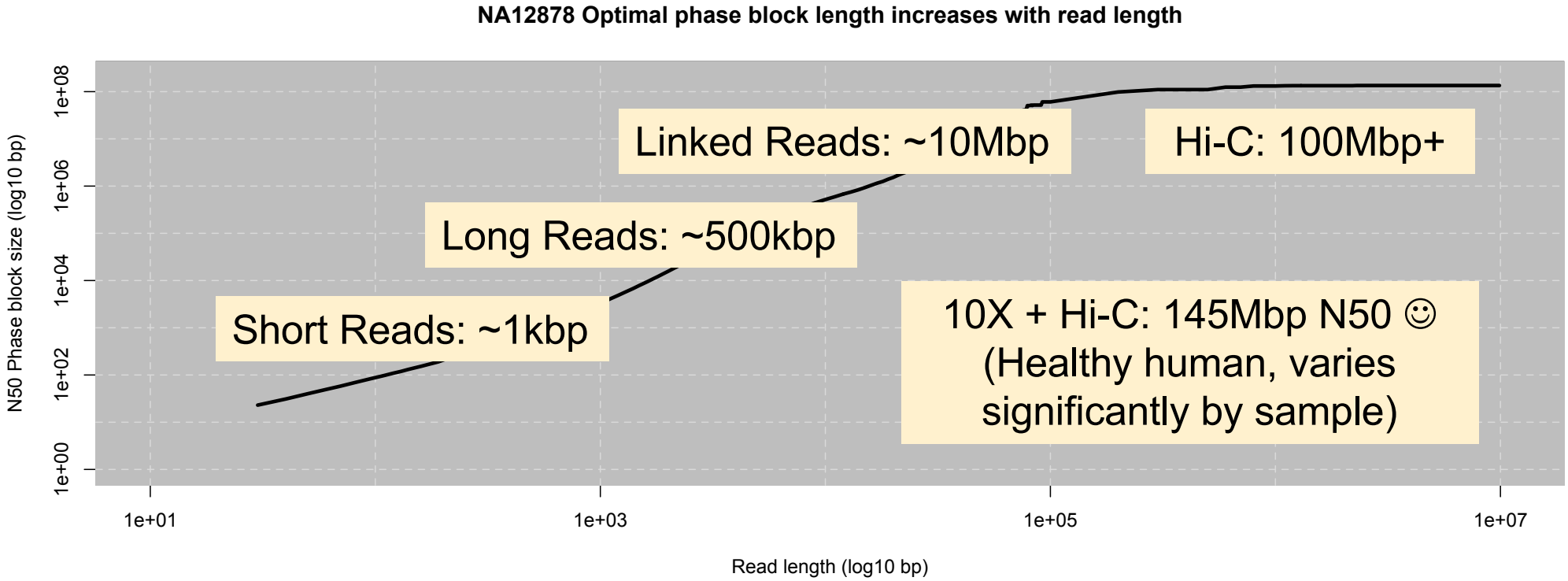
my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guido, Ring, & Gerstein labs

Phasing Results

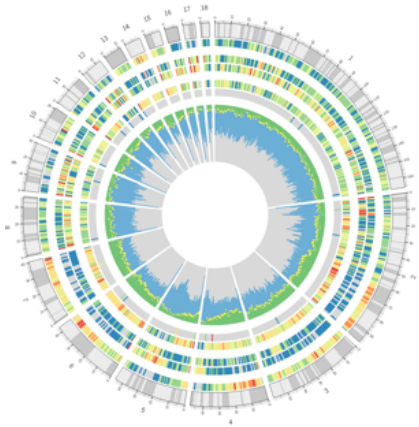




HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies
Edge, P, Bafna, V, Bansal, V (2016) *Genome Research*. doi: 10.1101/gr.213462.116

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference

A -> C

Call SNPs
+ Indels

A/C -> A|C

Phase SNPs
+ Indels

GAT
-> XXX

Call SVs

GAT/XXX
-> GAT|XXX

Phase SVs
& Local Assembly

Phased
SNPs,
Indels &
SVs.vcf

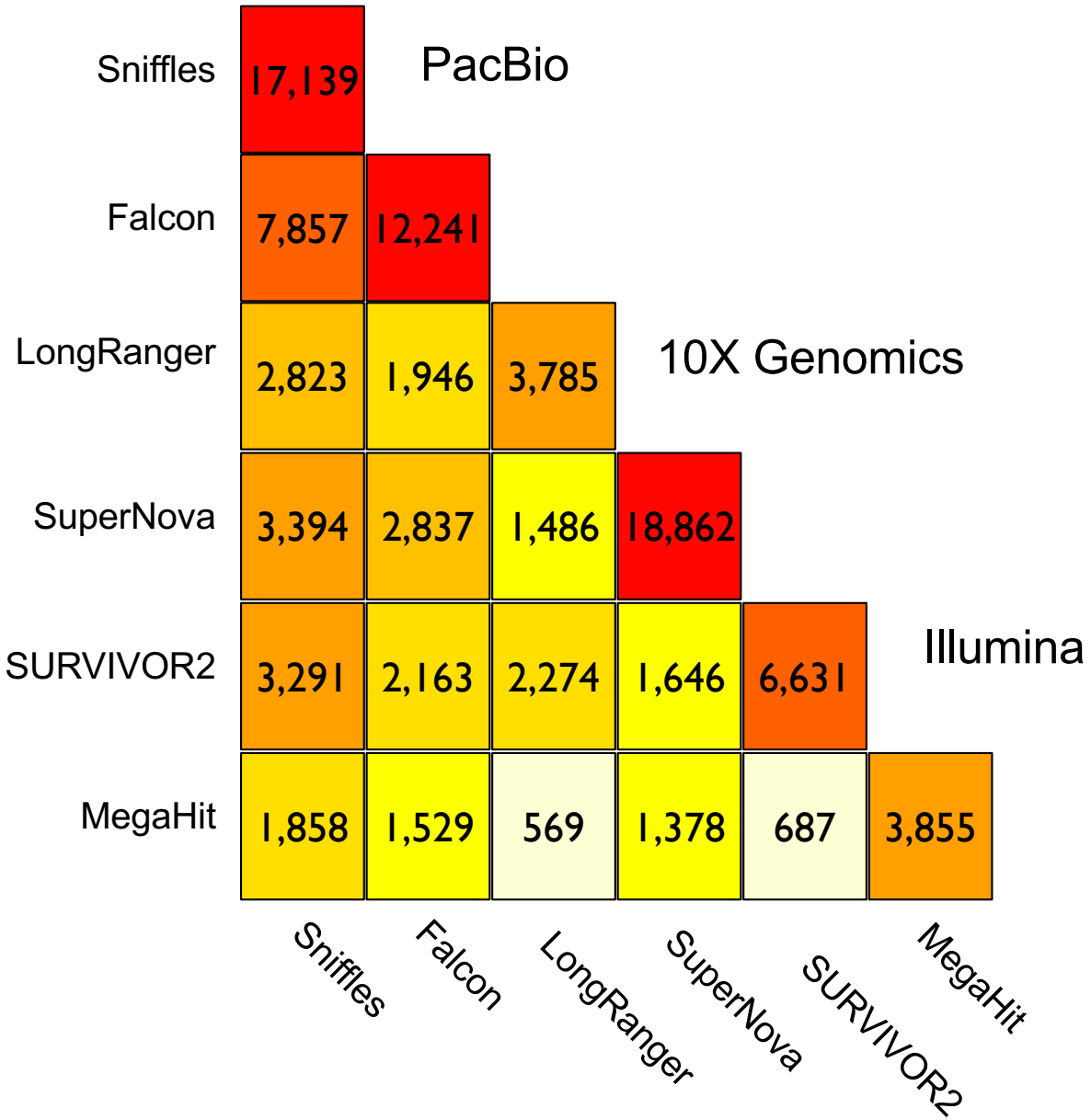
Assemble
Diploid
Sequence

my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guido, Ring, & Gerstein labs

SVs using Short, Long and Linked Reads



Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

Inner triplets

- Concordance by Assembly
- Concordance by Mappers

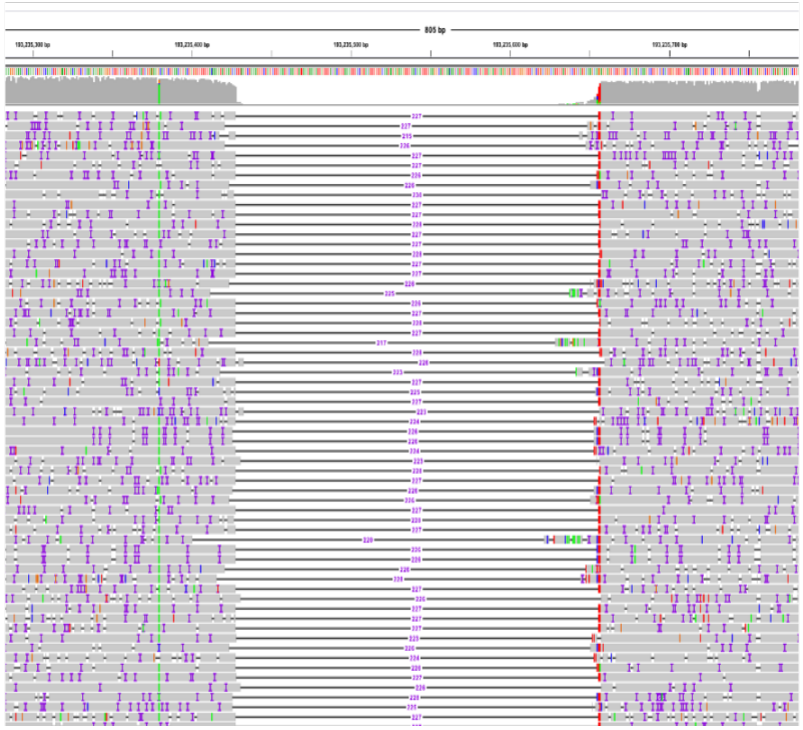
Overall:

- Lonnnnnnng reads give the most variants with the best concordance ☺

NGMLR + Sniffles

BWA-MEM:

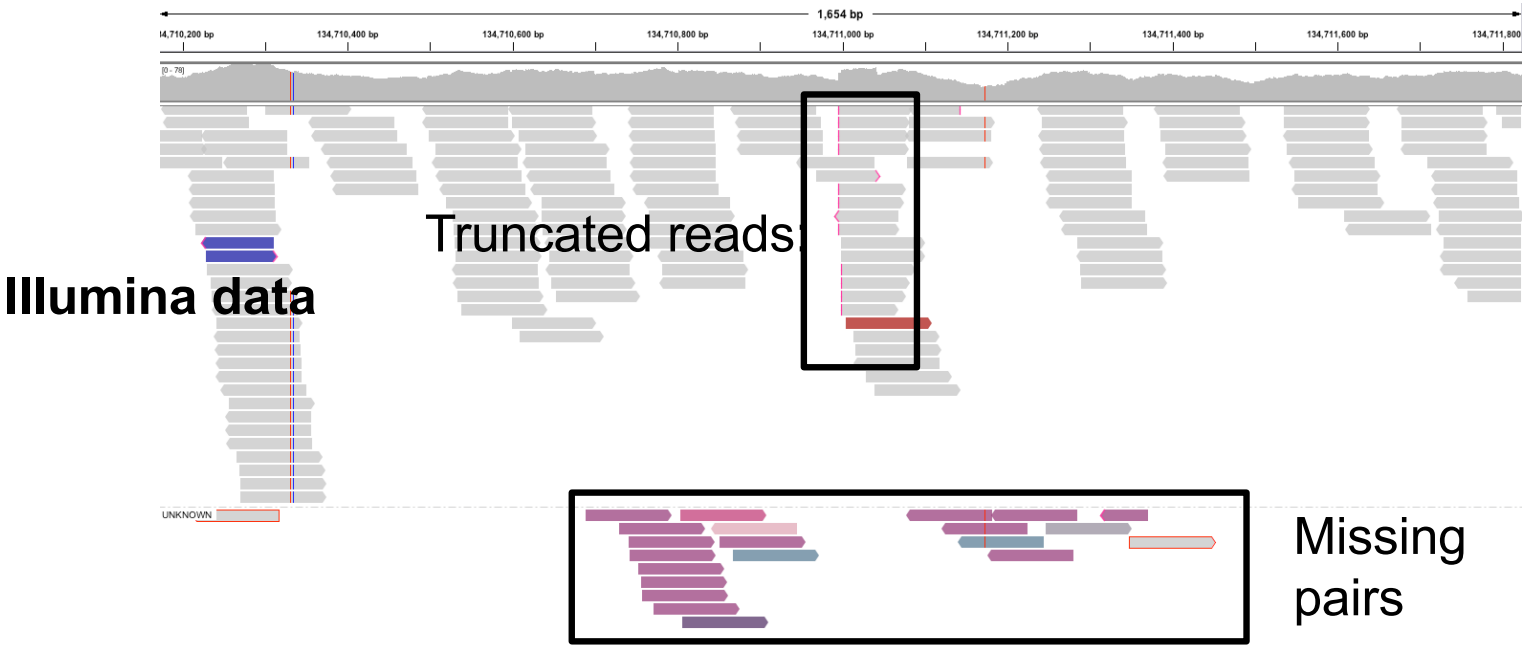
NGMLR:



NGMLR: Convex gap penalty to balance frequent small sequencing errors with larger SVs
Sniffles: Scan within and between split reads to accurately find SVs (Ins, Del, Dup, Inv, Trans)
Mendelian concordance >95%, experimental validation also very high

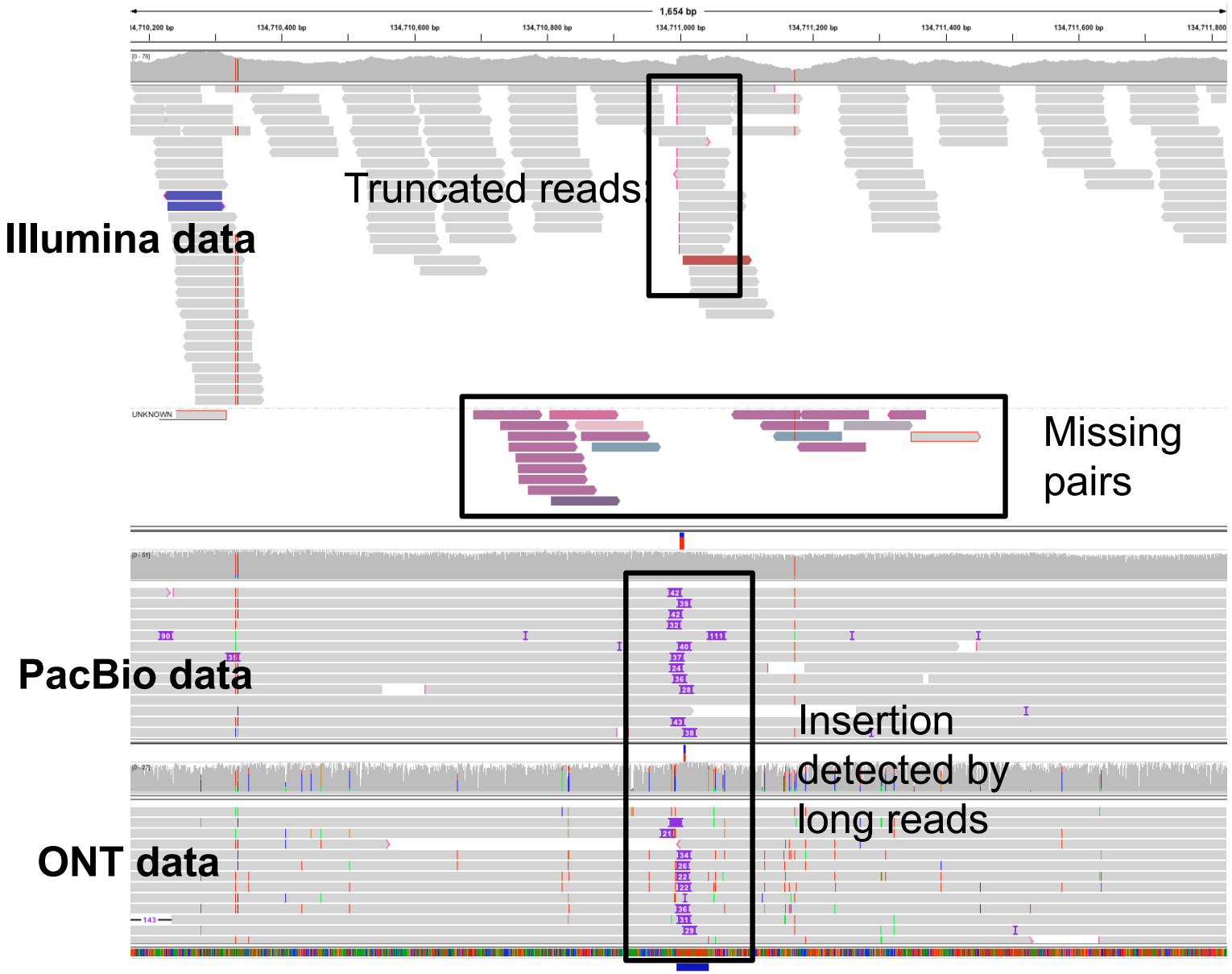
Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

No more false positives!



Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

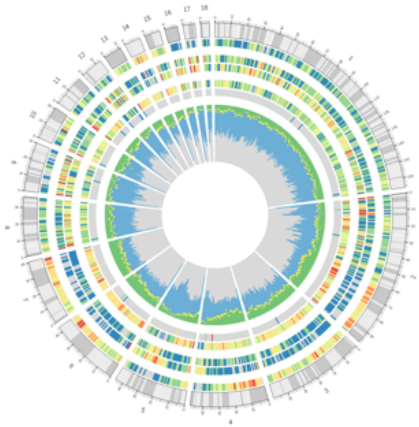
No more false positives!



Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference

A -> C

Call SNPs
+ Indels

A/C -> A|C

Phase SNPs
+ Indels

GAT
-> XXX

Call SVs

GAT/XXX
-> GAT|XXX

Phase SVs
& Local Assembly

Phased
SNPs,
Indels &
SVs.vcf

Assemble
Diploid
Sequence

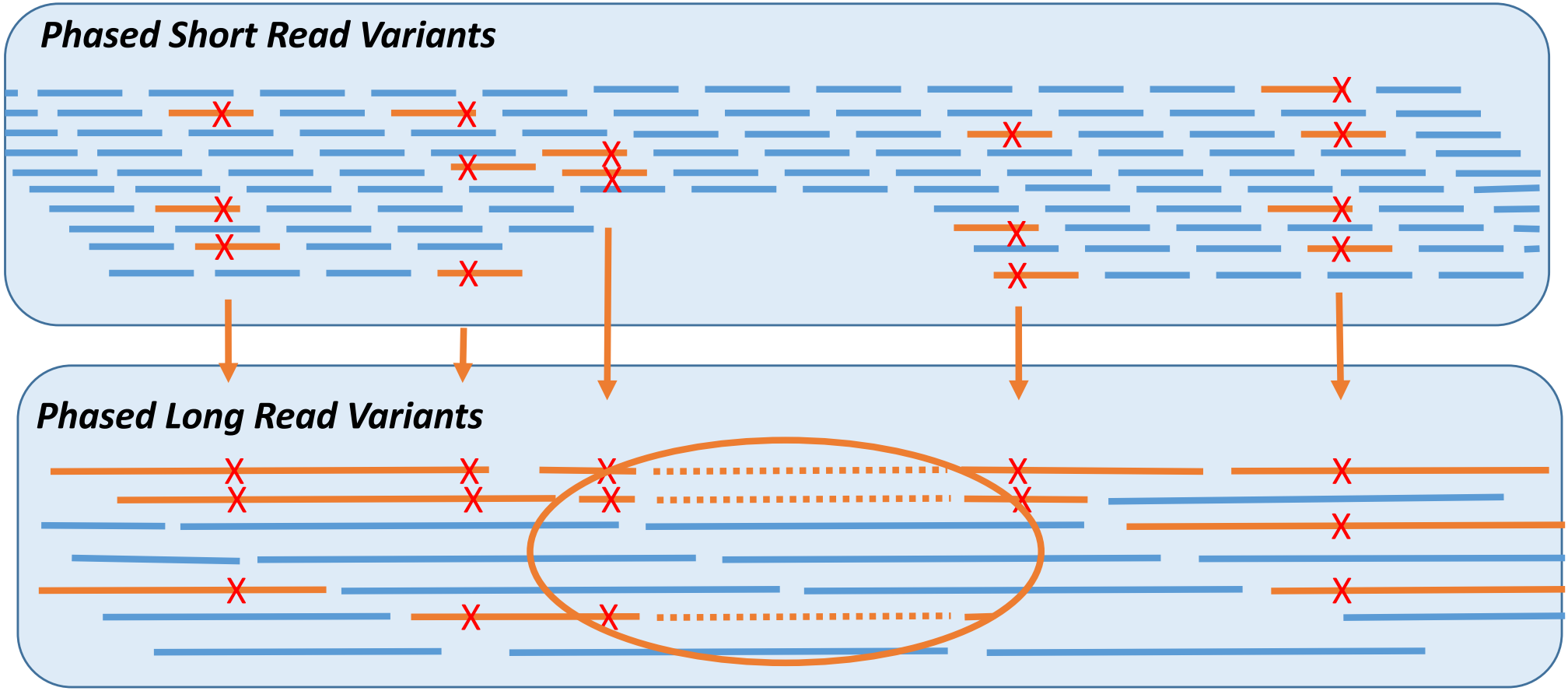
my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guido, Ring, & Gerstein labs

Local Assembly and SV Phasing

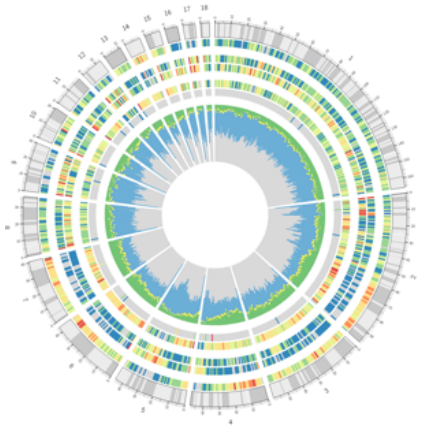
Transfer the phasing of the short read variants to the long reads
The phased long reads allow the SVs to be phased



Phase SVs: Make sure SVs are associated with the correct haplotype
Local Assembly: Refine sequence of insertions, resolve complex nested variants

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference

A -> C

Call SNPs
+ Indels

A/C -> A|C

Phase SNPs
+ Indels

GAT
-> XXX

Call SVs

GAT/XXX
-> GAT|XXX

Phase SVs
& Local Assembly

Phased
SNPs,
Indels &
SVs.vcf

Assemble
Diploid
Sequence

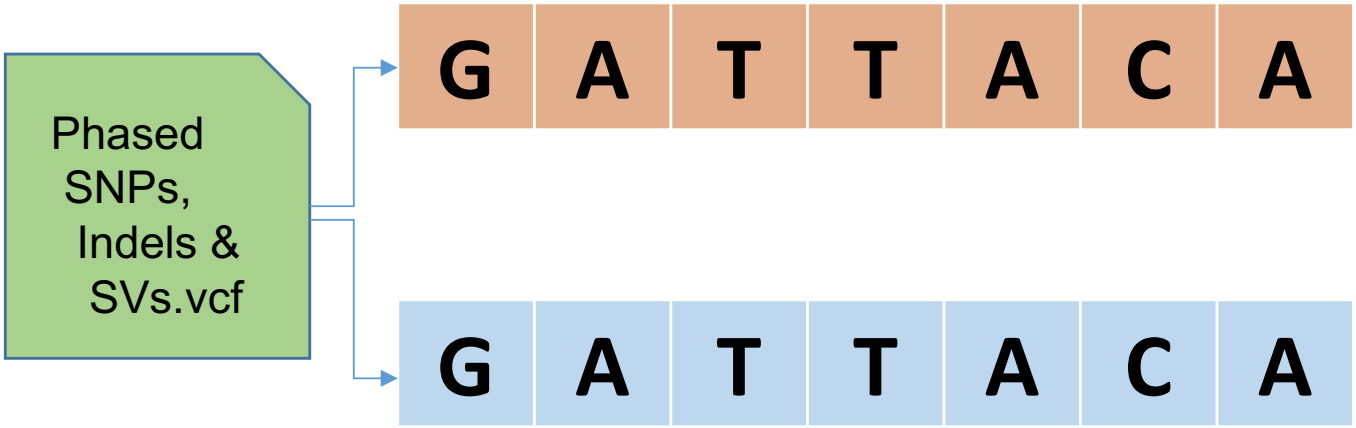
my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guido, Ring, & Gerstein labs

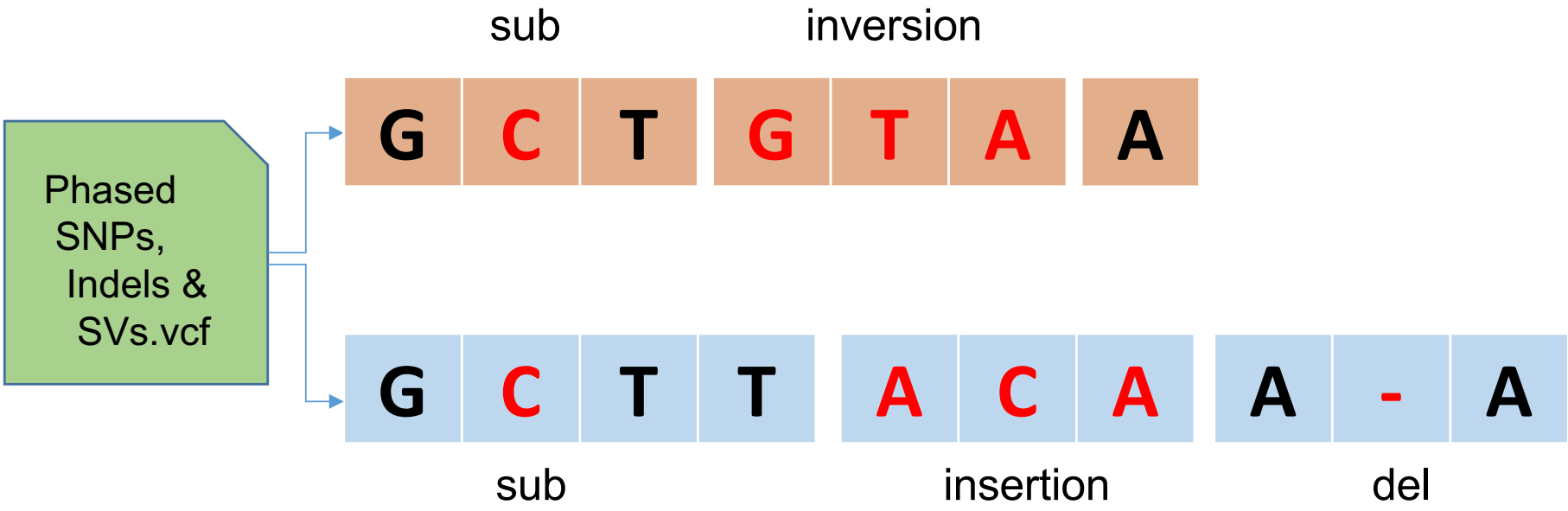
Assembling a "Perfect" Personalized Diploid Genome

Carefully "stitch" the phased variants into the reference genome at the right position to create a pair of phased chromosome fasta files



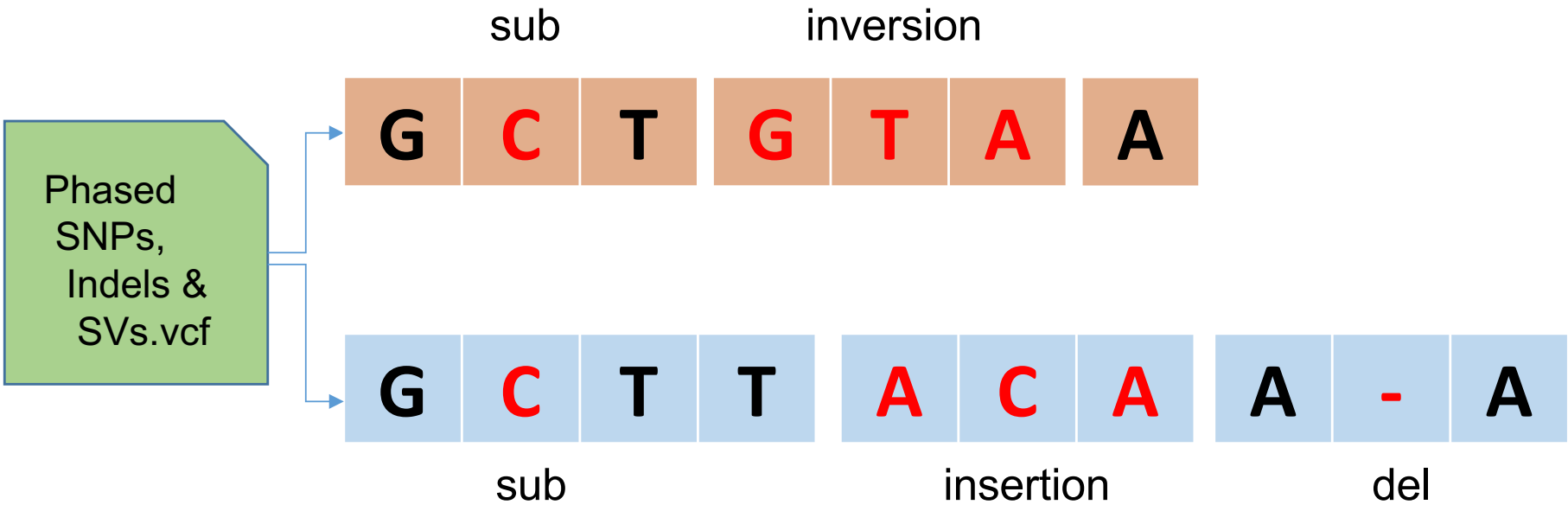
Assembling a "Perfect" Personalized Diploid Genome

Carefully "stitch" the phased variants into the reference genome at the right position to create a pair of phased chromosome fasta files



Assembling a "Perfect" Personalized Diploid Genome

Carefully "stitch" the phased variants into the reference genome at the right position to create a pair of phased chromosome fasta files



Stitching based on AlleleSeq pipeline enhanced for SVs (Rozowsky et al, 2011)

- Maintains a mapping from reference to personal genome coordinates to make lift over of annotation straightforward to compute

Using 10X + HiC + PacBio, assemble essentially perfect diploid human genomes with haplotypes spanning entire chromosomes

- Phased diploid genome can be aligned or aligned against just like a de novo genome assembly

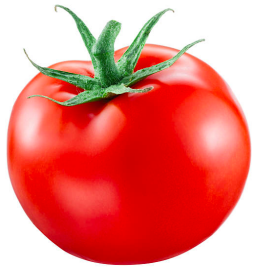
Applications

Expression & Regulation

Foundation for mapping functional data

- Discover novel genes and gene fusions
- Analyze differential expression in CNVs
- Discover new regulatory regions
- Analyze allele-specific expression

Population Genetics



Framework for GWAS of Structural Variations

- Identified SVs in >900 accessions using short reads
- Assembling the top 50 lines using long & linked reads
- Perform GWAS of breeding traits

Polyploidy

Studying heterozygosity in sugarcane

- Have a high quality PacBio-based assembly of POJ2878 using FALCON (140kbp N50)
- Developing new methods for phasing (9-14 copies of each chromosome)

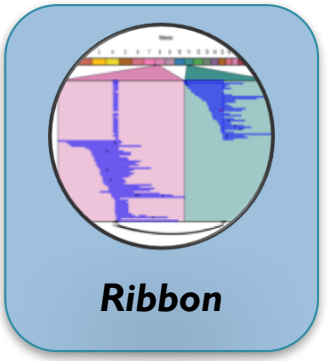
Reference-quality Genomes without *de novo* assembly

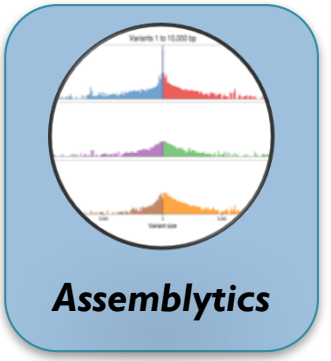
De novo assembly is essential for exploring new species

- Reference-free or mapping to a distant reference is difficult to impossible

But once the first genome of a species has been assembled, shouldn't the second genome be a little easier?

- Use the right combination of data to capture and phase all types of variants
- Overnight analysis to create a high quality personalized genome that are more accurate, more predictable, and easier to use than a de novo assembly
- The personalized diploid genome will be a platform for functional and evolutionary analysis in many species





Acknowledgements

Schatz Lab
Mike Alonge
Amelia Bateman
Charlotte Darby
Han Fang
Michael Kirsche
Sam Kovaka
Laurent Luo
Srividya Ramakrishnan
T. Rhyker Ranallo-Benavide
Your Name Here

Baylor Medicine
Fritz Sedlazeck

University of Vienna
Arndt von Haeseler
Philipp Rescheneder

DNAnexus
Maria Nattestad

CSHL
Gingeras Lab
Jackson Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

SBU
Skiena Lab
Patro Lab

GRC
Roderic Guido
Alessandra Breschi
Anna Vlasova

Yale
Gerstein Lab

JHU
Battle Lab
Langmead Lab
Leek Lab
Salzberg Lab
Taylor Lab
Timp Lab
Wheelan Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

PacBio
Greg Concepcion

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION

VERTEBRATE
GENOMES
PROJECT

DIGITAL NOAH'S ARK GENOME LIBRARY

Sunset Room, 9am-5pm, tomorrow

Thank you!

@mike_schatz