Phased diploid genomes using short, long, and linked reads

Michael Schatz

January 17, 2018
PAG G10k Workshop

@mike_schatz / #PAGXXVI

Selected Tools

1. Pre-assembly QC
2. SV Detection & Phasing
3. Post-assembly

Selected Tools

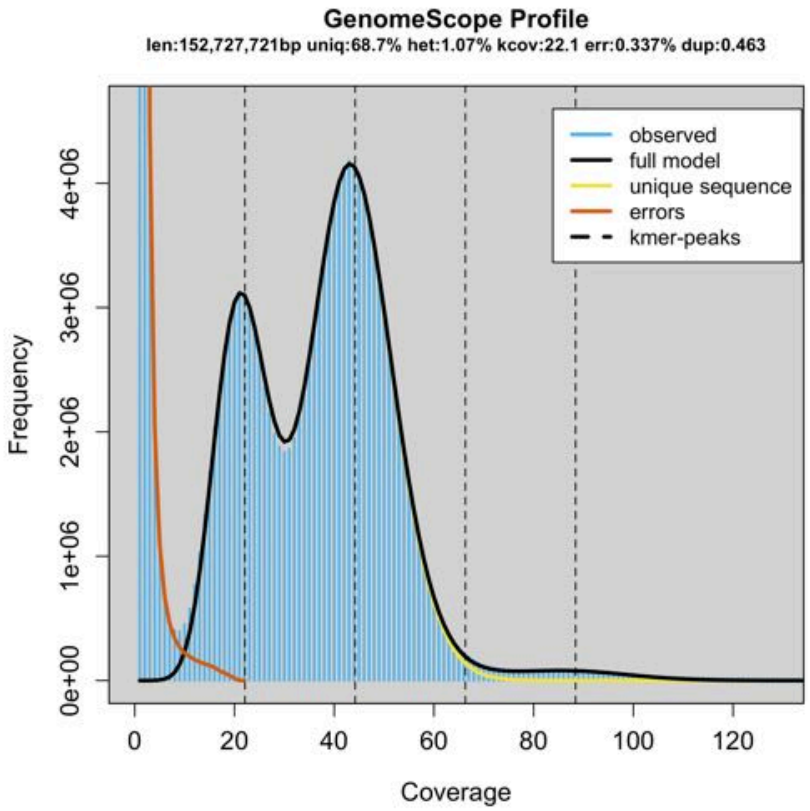
1. **Pre-assembly QC**

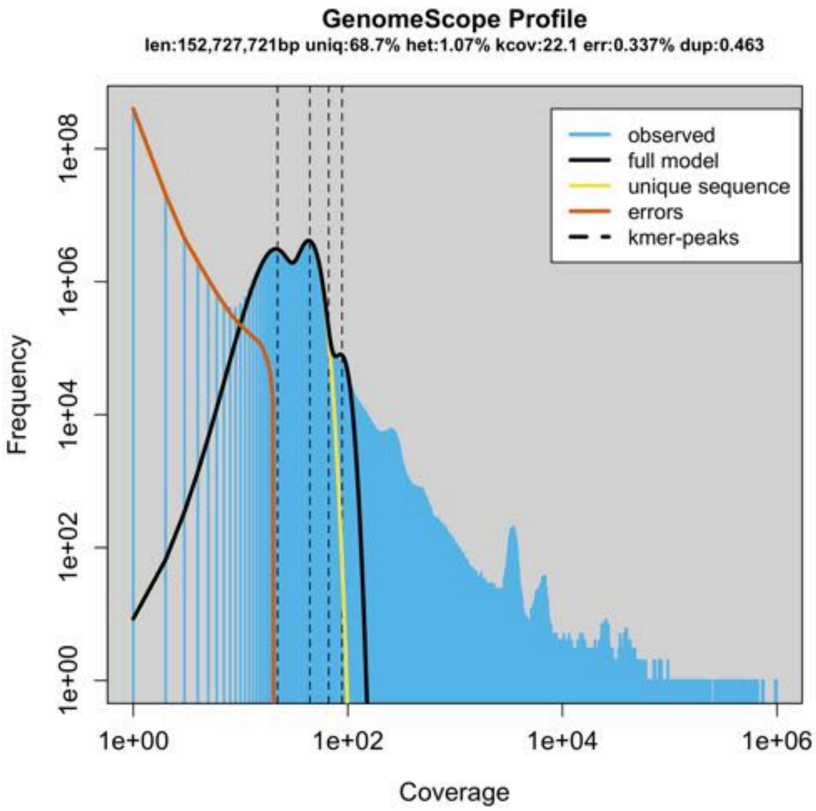
2. **SV Detection & Phasing**

3. **Post-assembly**

GenomeScope: Fast reference-free genome profiling

http://genomescope.org





Infer the properties of unassembled genomes from raw sequencing data:

- *Genome Size, Repeat Content, Rate of Heterozygosity*
- *Coverage, Read Error Rate, Rate of PCR Duplications*
- *Analysis of polypoid genomes in development*

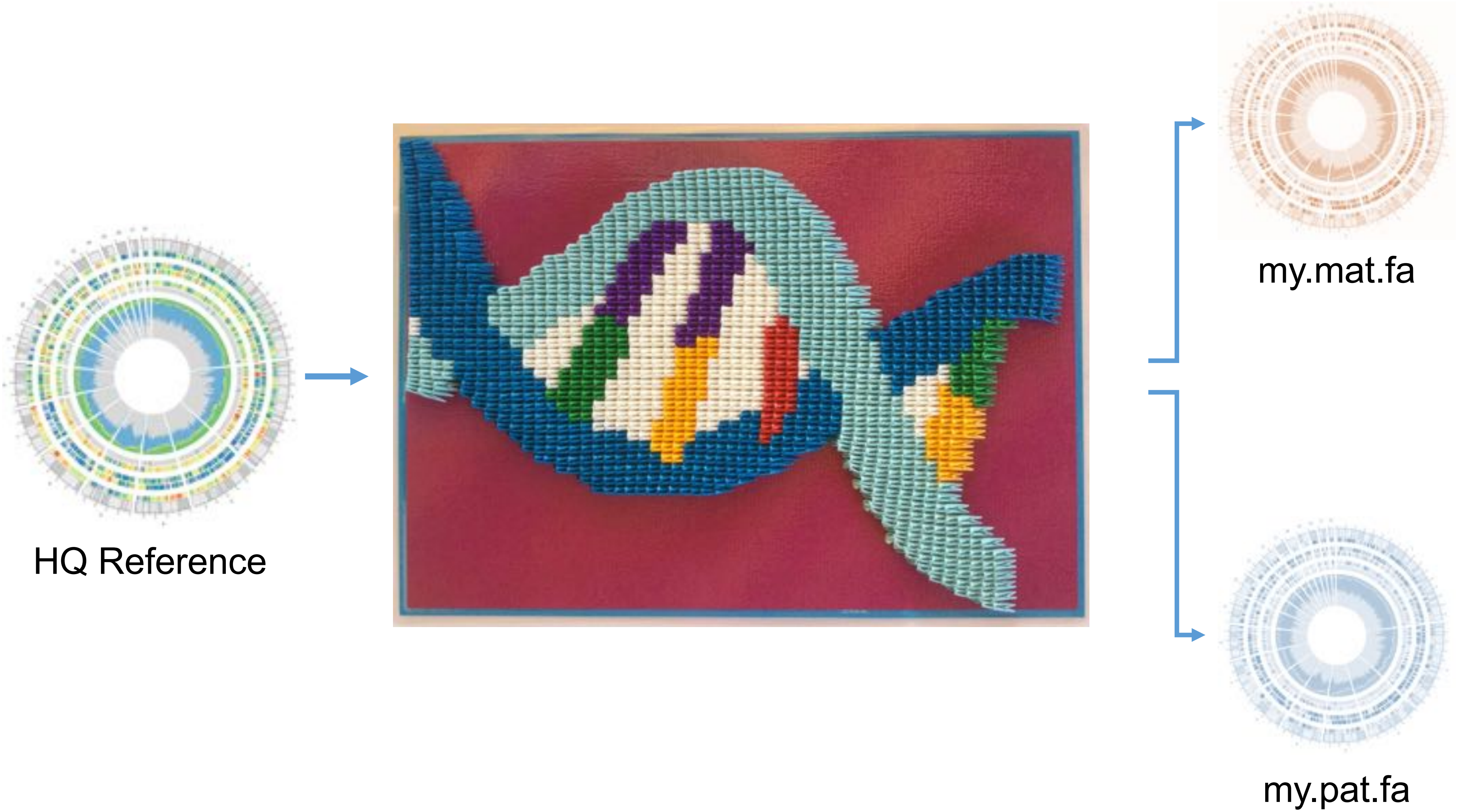
Vurture *et al.* (2017) *Bioinformatics. doi: https://doi.org/10.1093/bioinformatics/btx153*

Selected Tools

1. Pre-assembly QC
2. **SV Detection & Phasing**
3. Post-assembly

CrossStitch

https://github.com/schatzlab/crossstitch



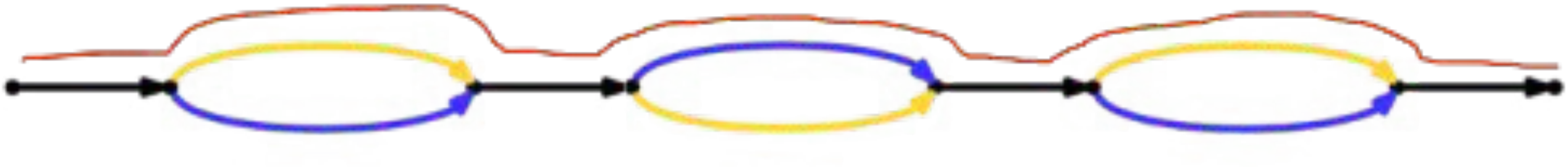
HQ Reference

my.mat.fa

my.pat.fa

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

CrossStitch for de novo assembly

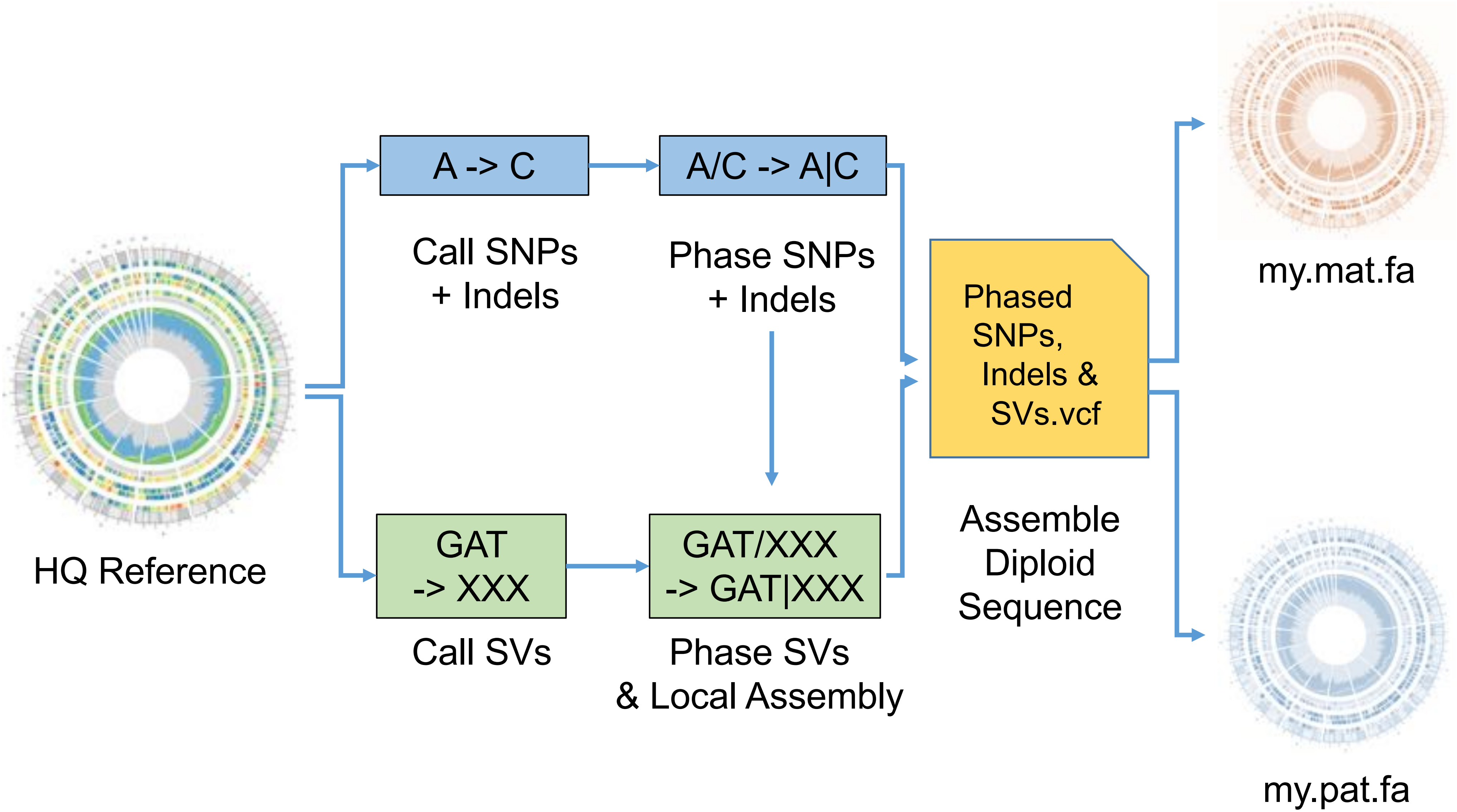


Rather than a reference genome, start from a "pseudohap" draft assembly

- Native output for FALCON and SuperNova (and regular Canu?)
- Will need to be careful to correctly recognize and traverse the bubbles

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

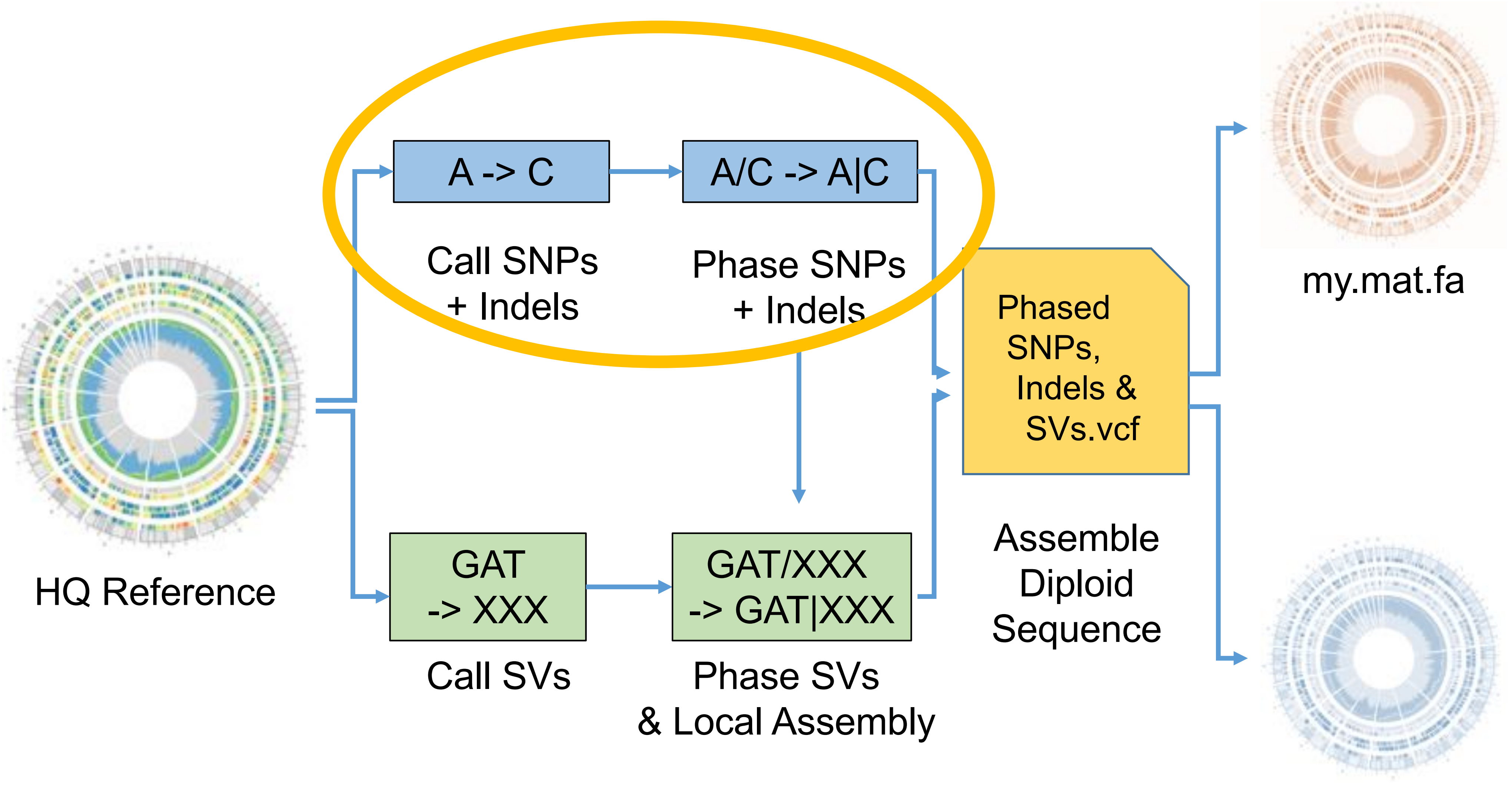
my.mat.fa

my.pat.fa

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

CrossStitch

https://github.com/schatzlab/crossstitch



A -> C

Call SNPs
+ Indels

A/C -> A|C

Phase SNPs
+ Indels

HQ Reference

GAT
-> XXX

Call SVs

GAT/XXX
-> GAT|XXX

Phase SVs
& Local Assembly

Phased
SNPs,
Indels &
SVs.vcf

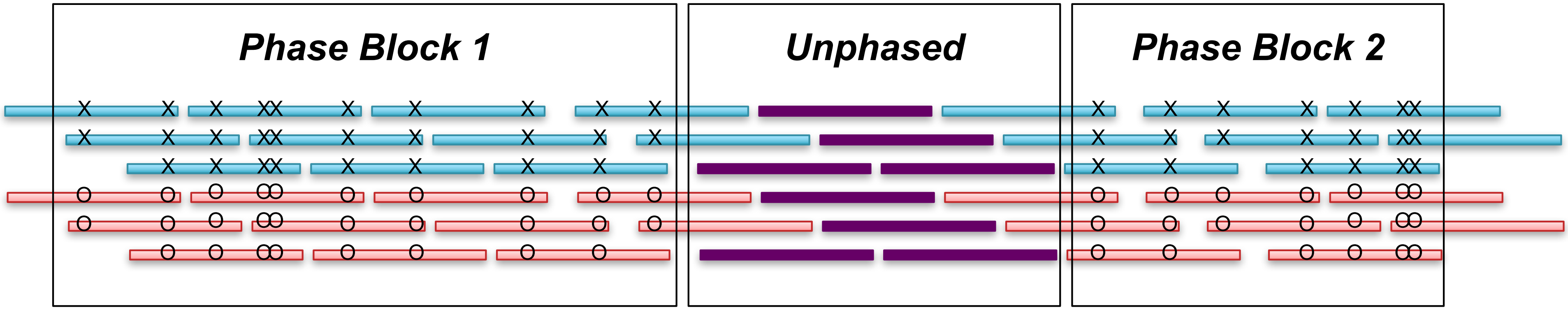
Assemble
Diploid
Sequence

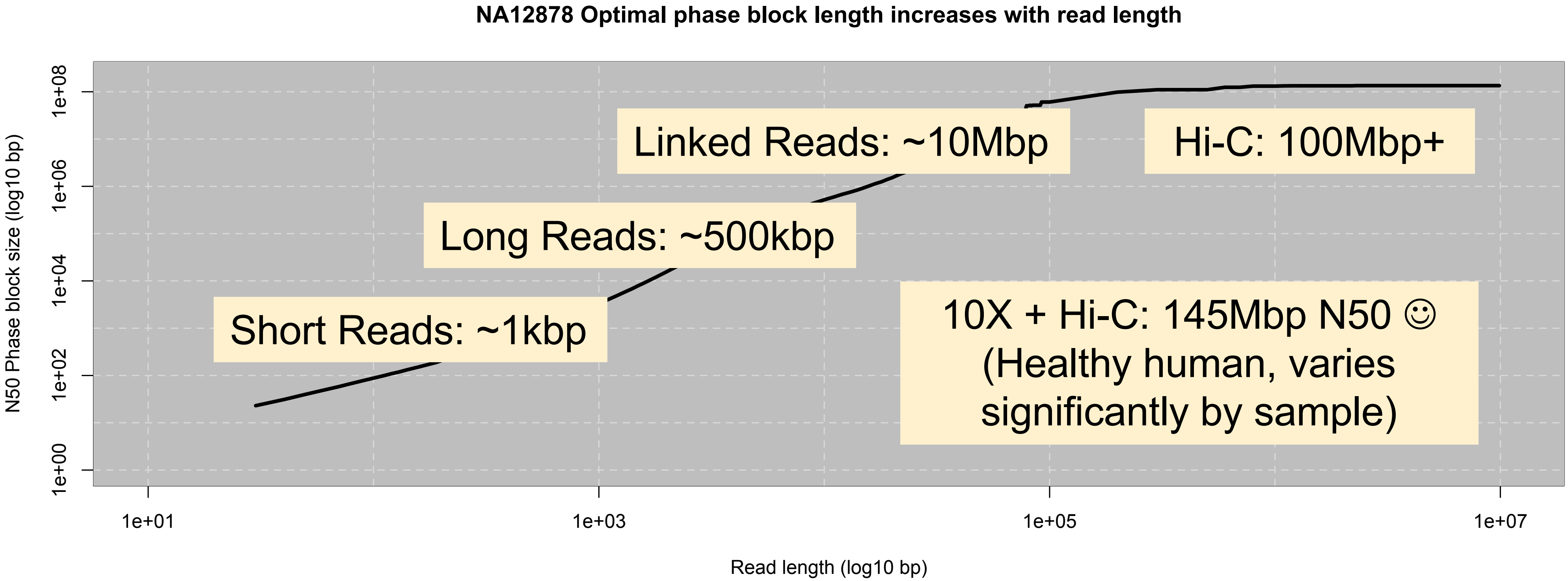
my.mat.fa

my.pat.fa

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

Phasing Results




NA12878 Optimal phase block length increases with read length

Short Reads: ~1kbp

Long Reads: ~500kbp

Linked Reads: ~10Mbp

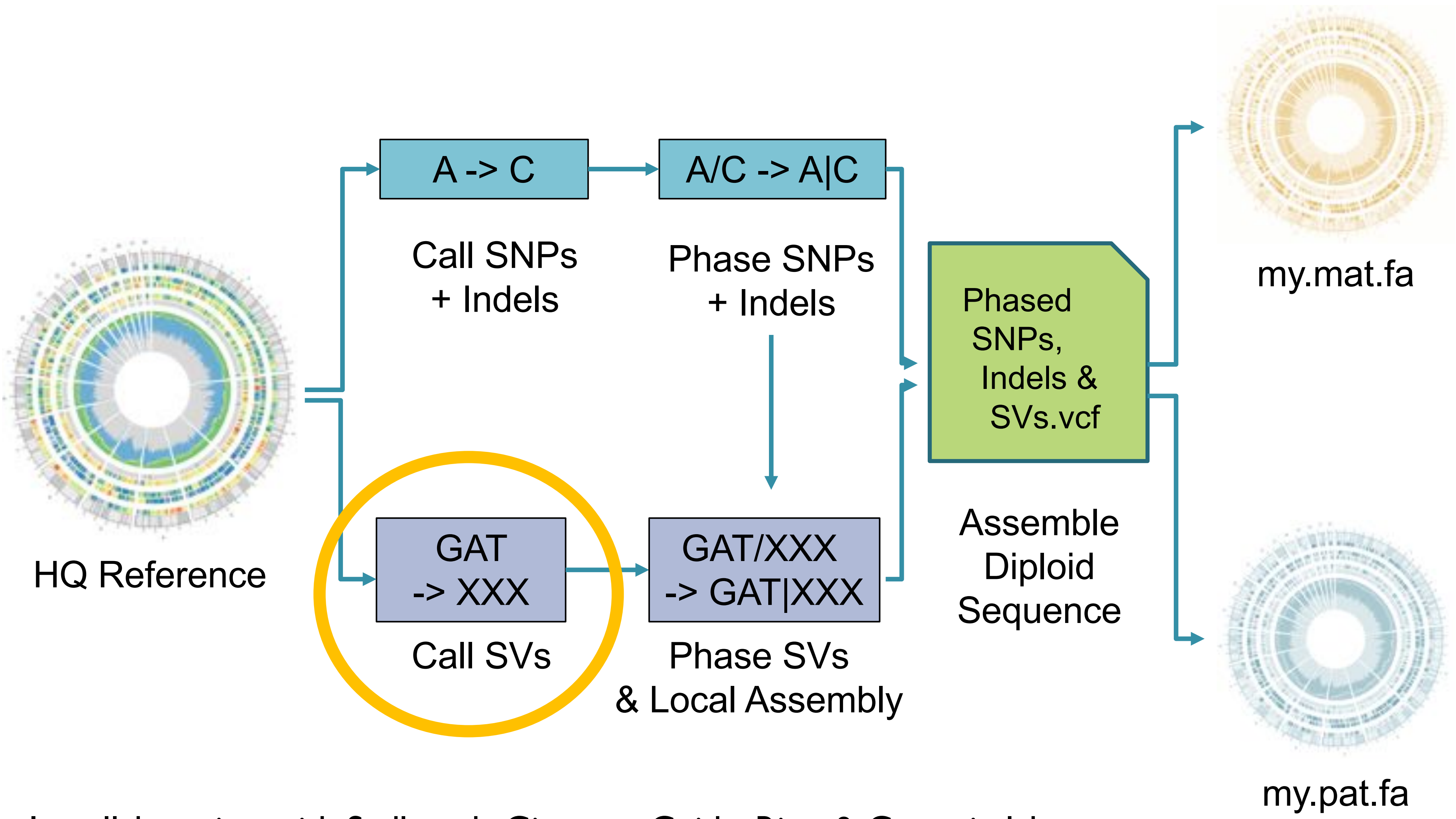
Hi-C: 100Mbp+

10X + Hi-C: 145Mbp N50 ☺
(Healthy human, varies significantly by sample)

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

	Sniffles	Falcon	LongRanger	SuperNova	SURVIVOR2	MegaHit
Sniffles	17,139					
Falcon	7,857	12,241				
LongRanger	2,823	1,946	3,785			
SuperNova	3,394	2,837	1,486	18,862		
SURVIVOR2	3,291	2,163	2,274	1,646	6,631	
MegaHit	1,858	1,529	569	1,378	687	3,855

PacBio

10X Genomics

Illumina

Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

Inner triplets

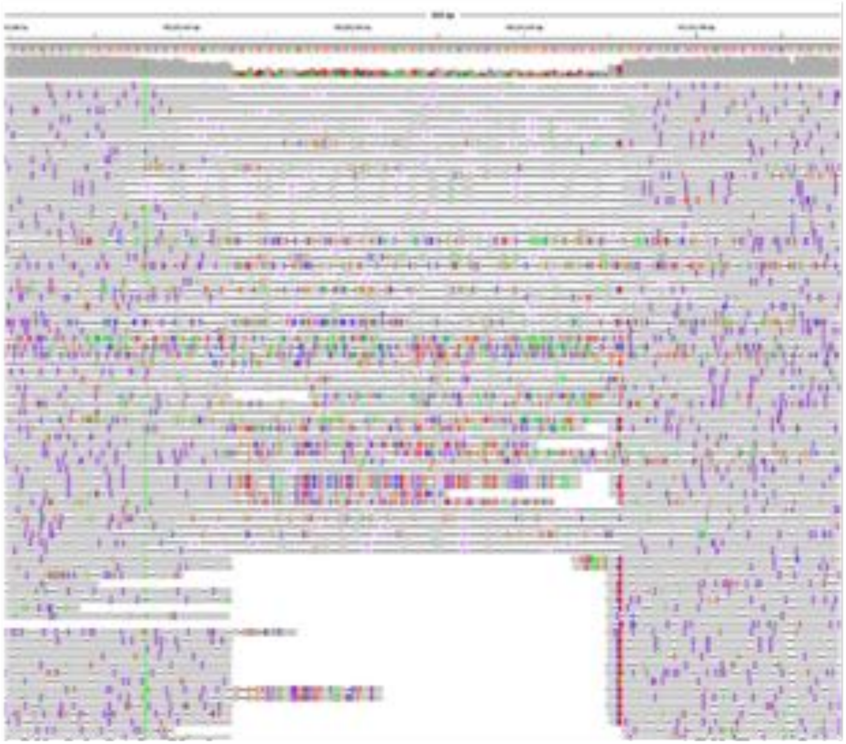
- Concordance by Assembly
- Concordance by Mappers

Overall:

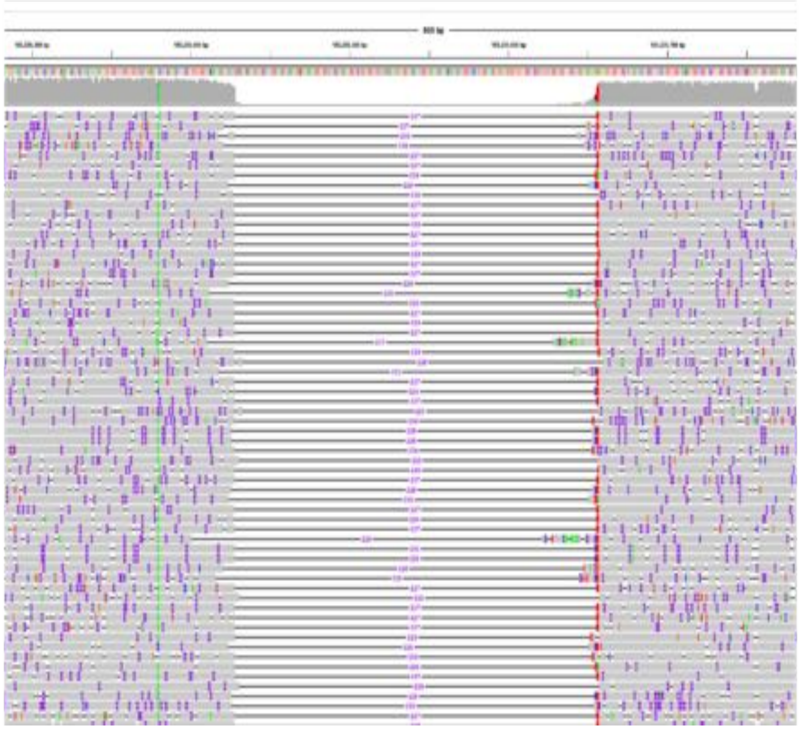
- Lonnnnnnng reads give the most variants with the best concordance ☺

NGMLR + Sniffles

BWA-MEM:



NGMLR:

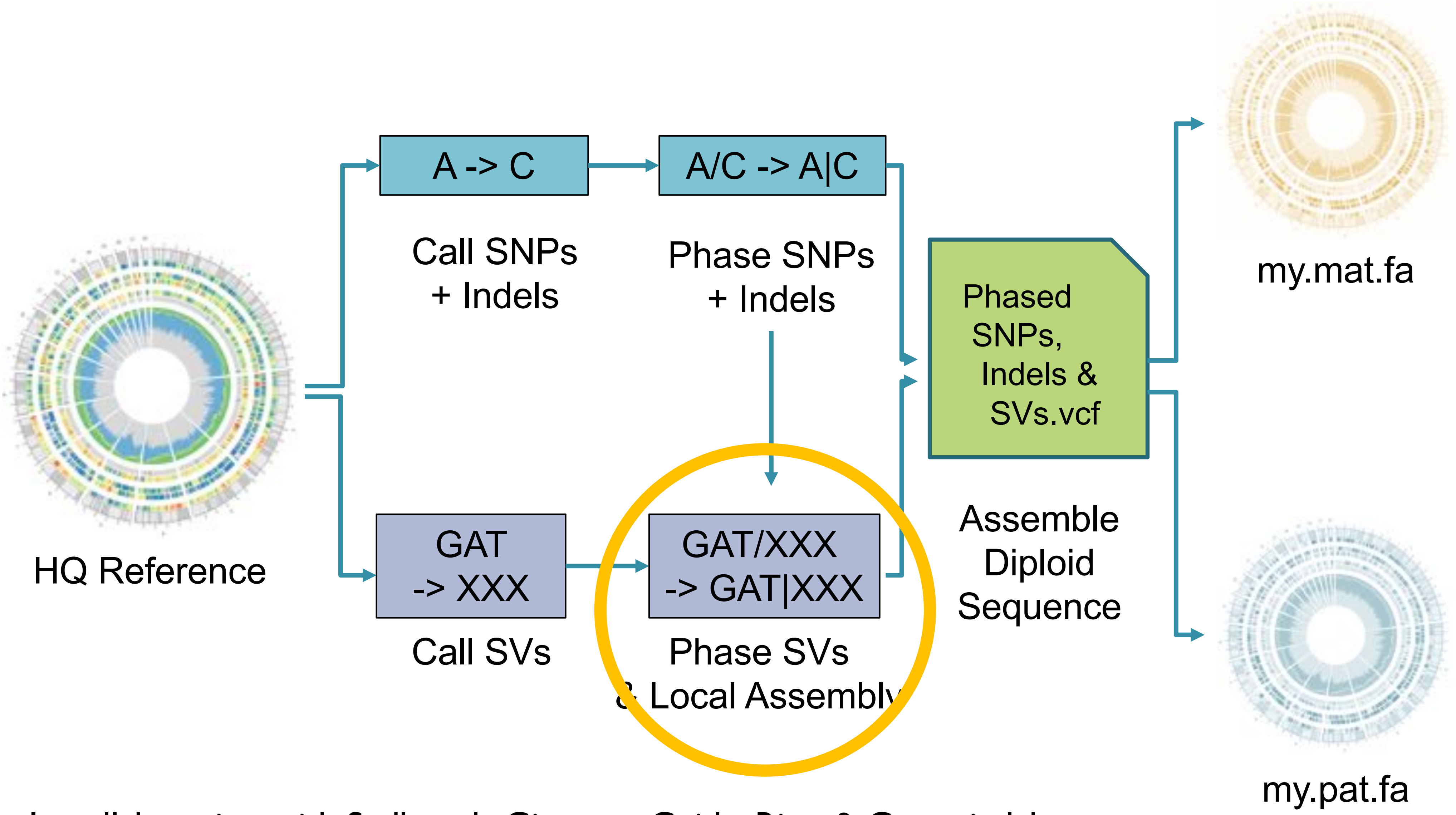


NGMLR: Convex gap penalty to balance frequent small sequencing errors with larger SVs
Sniffles: Scan within and between split reads to accurately find SVs (Ins, Del, Dup, Inv, Trans)
Mendelian concordance >95%, experimental validation also very high

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

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

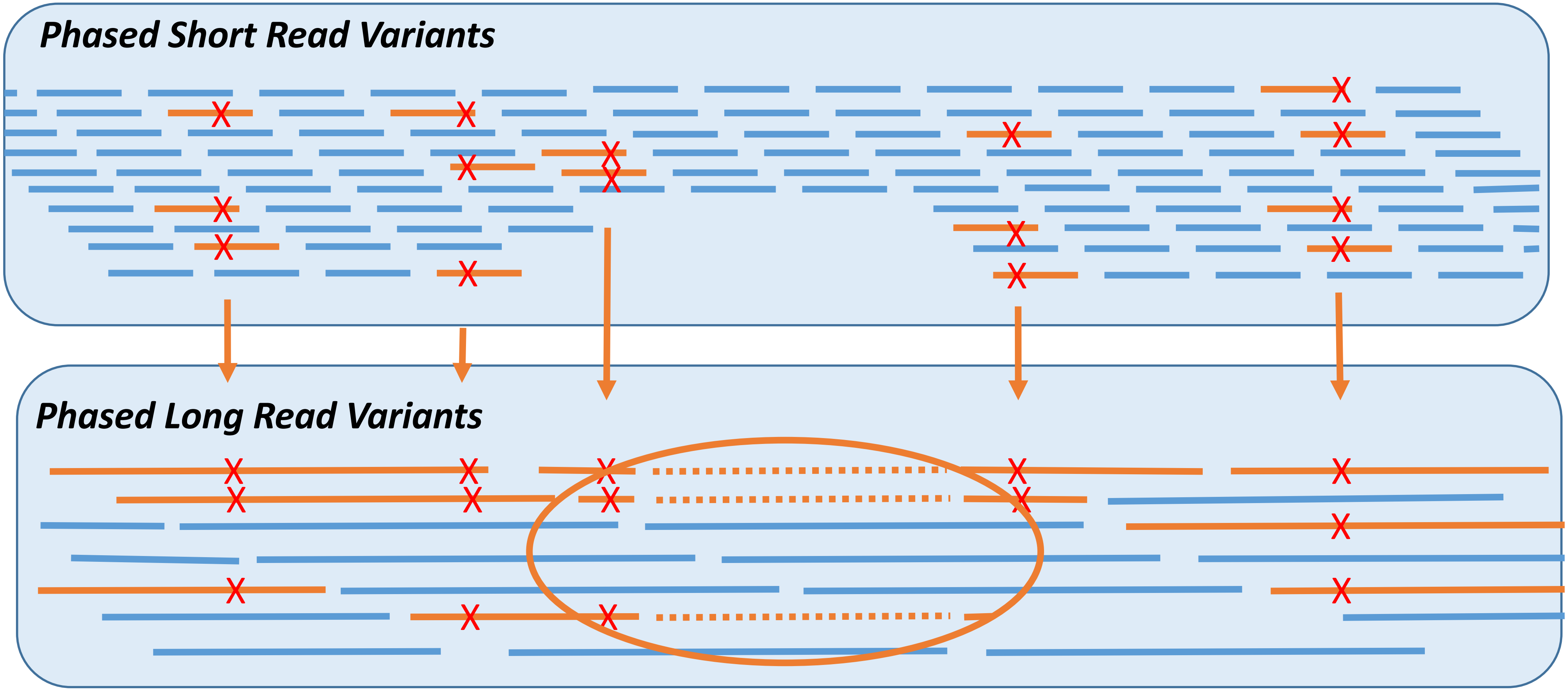
my.mat.fa

my.pat.fa

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

Local Assembly and SV Phasing

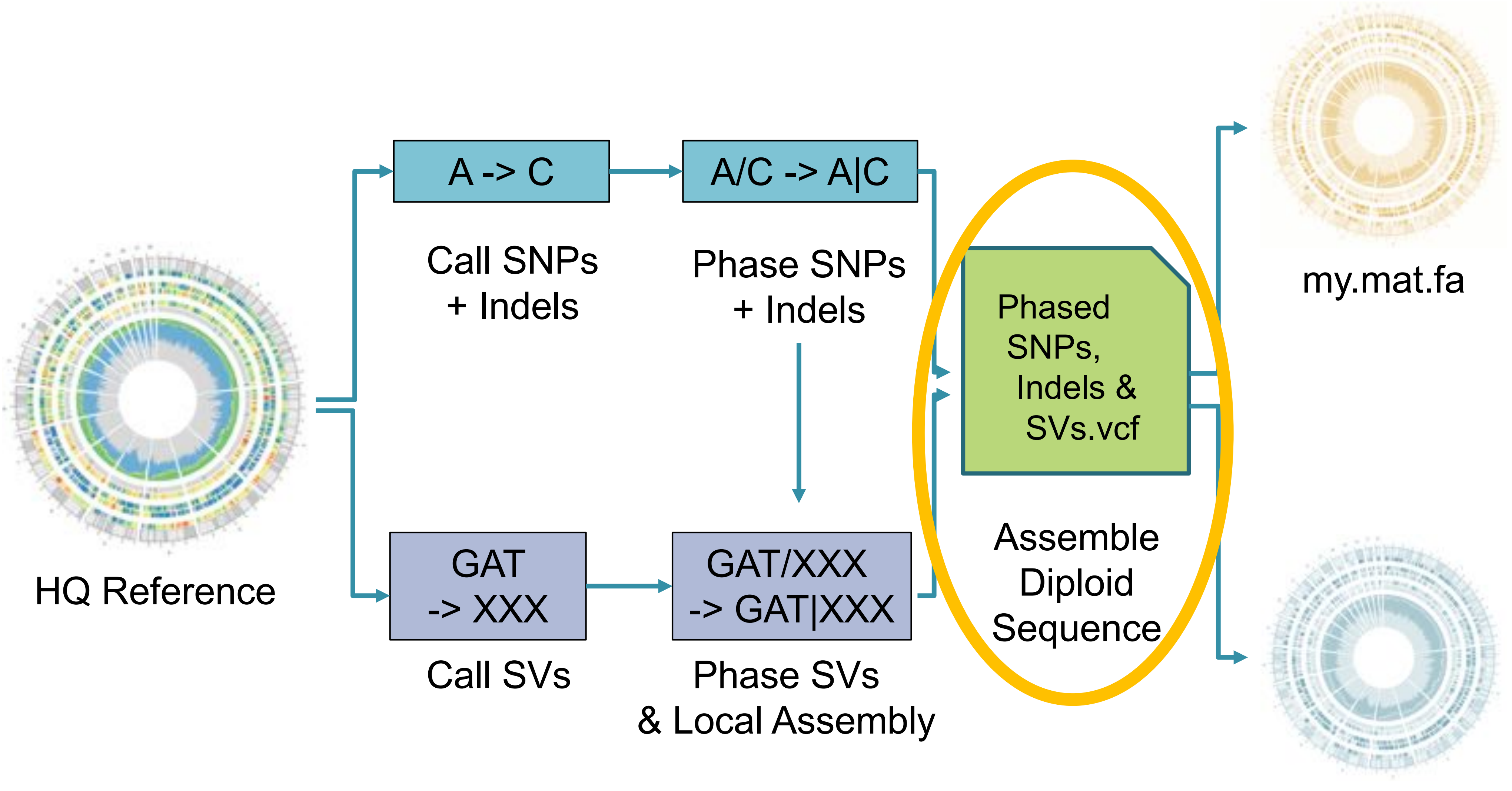
Transfer the phasing of the short read variants to the long reads
The phased long reads allow the SVs to be phased



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

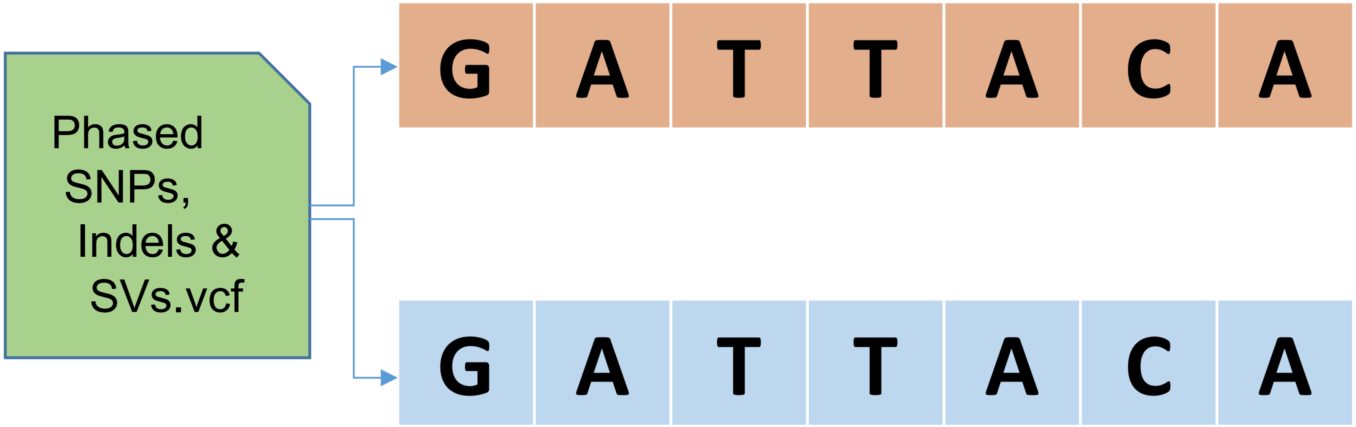
my.mat.fa

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



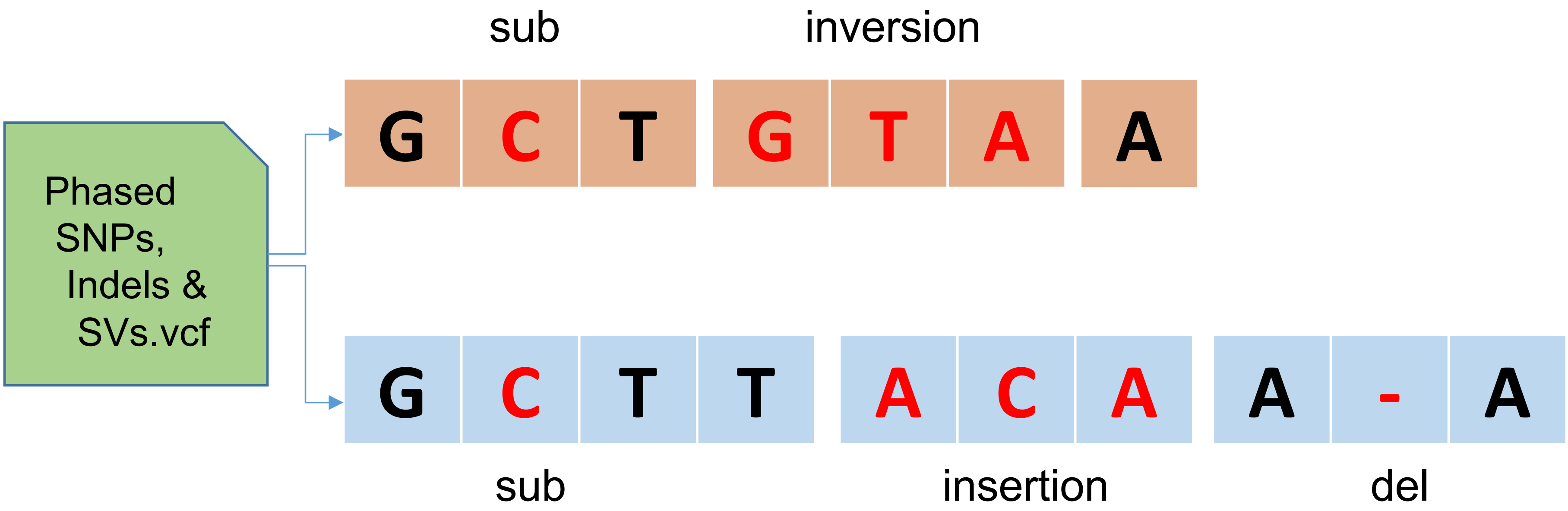
Phased SNPs, Indels & SVs.vcf

G	A	T	T	A	C	A

G	A	T	T	A	C	A

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



Phased SNPs, Indels & SVs.vcf

sub inversion

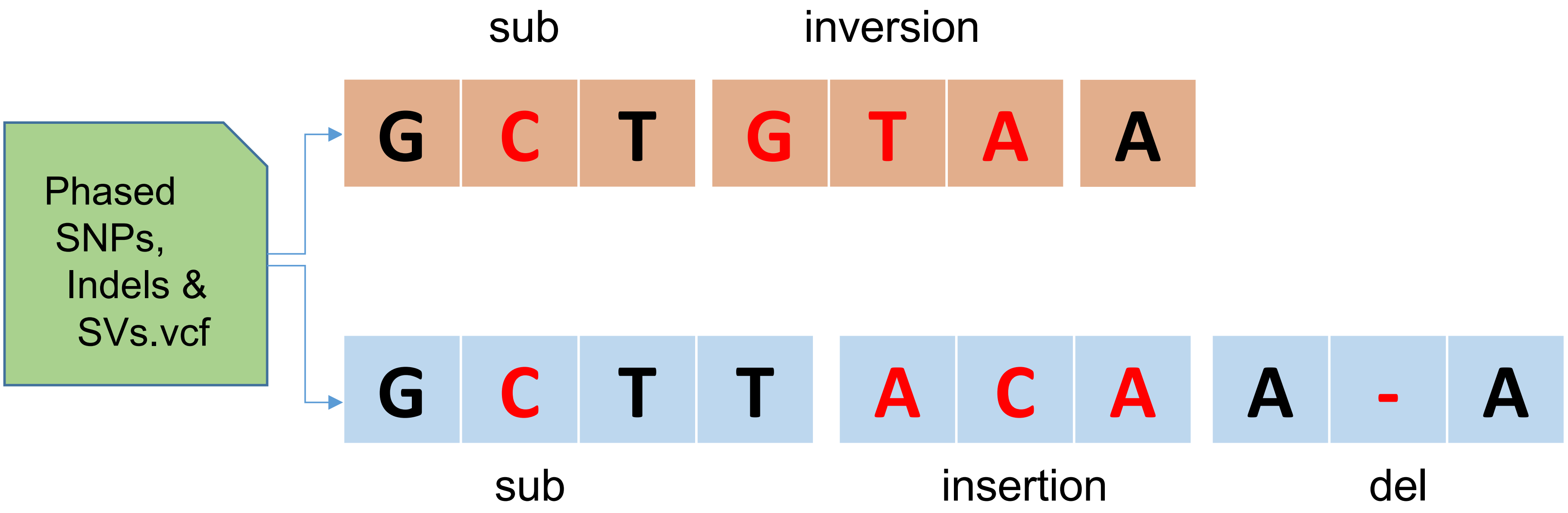
G C T G T A A

G C T T A C A A - A

sub insertion del

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

ENCODE

MaizeCODE

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)

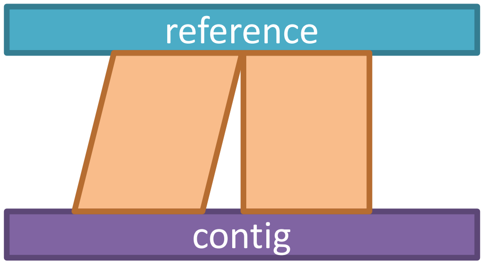
Selected Tools

1. Pre-assembly QC
2. SV Detection & Phasing
3. **Post-assembly**

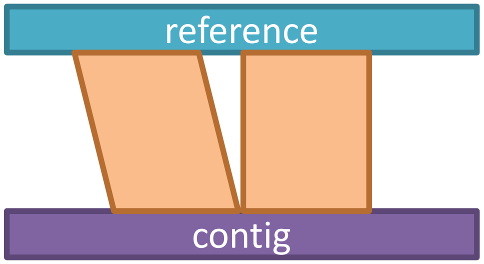
Assemblytics: Assembly-Based Variant-Caller

Defined point

Insertion

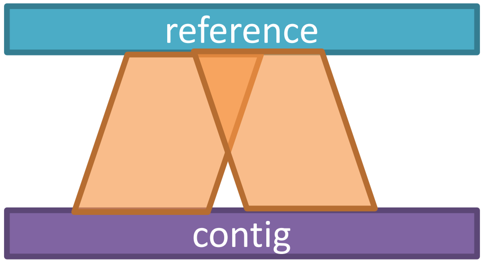


Deletion

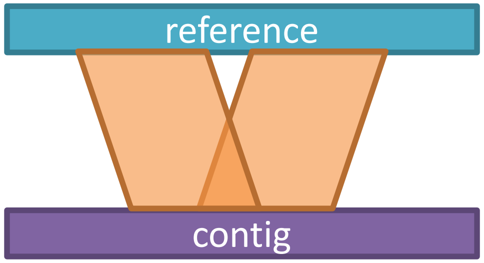


Overlapping alignments suggest tandem repeat

Tandem Expansions

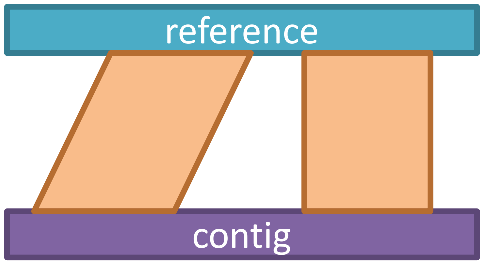


Tandem Contractions

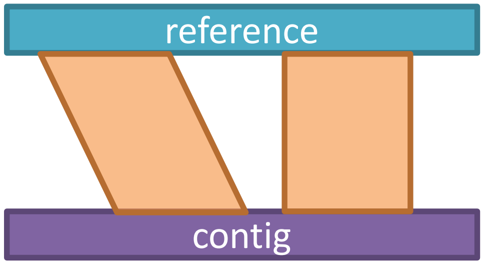


Gap where sequences do not align uniquely suggests a repeat

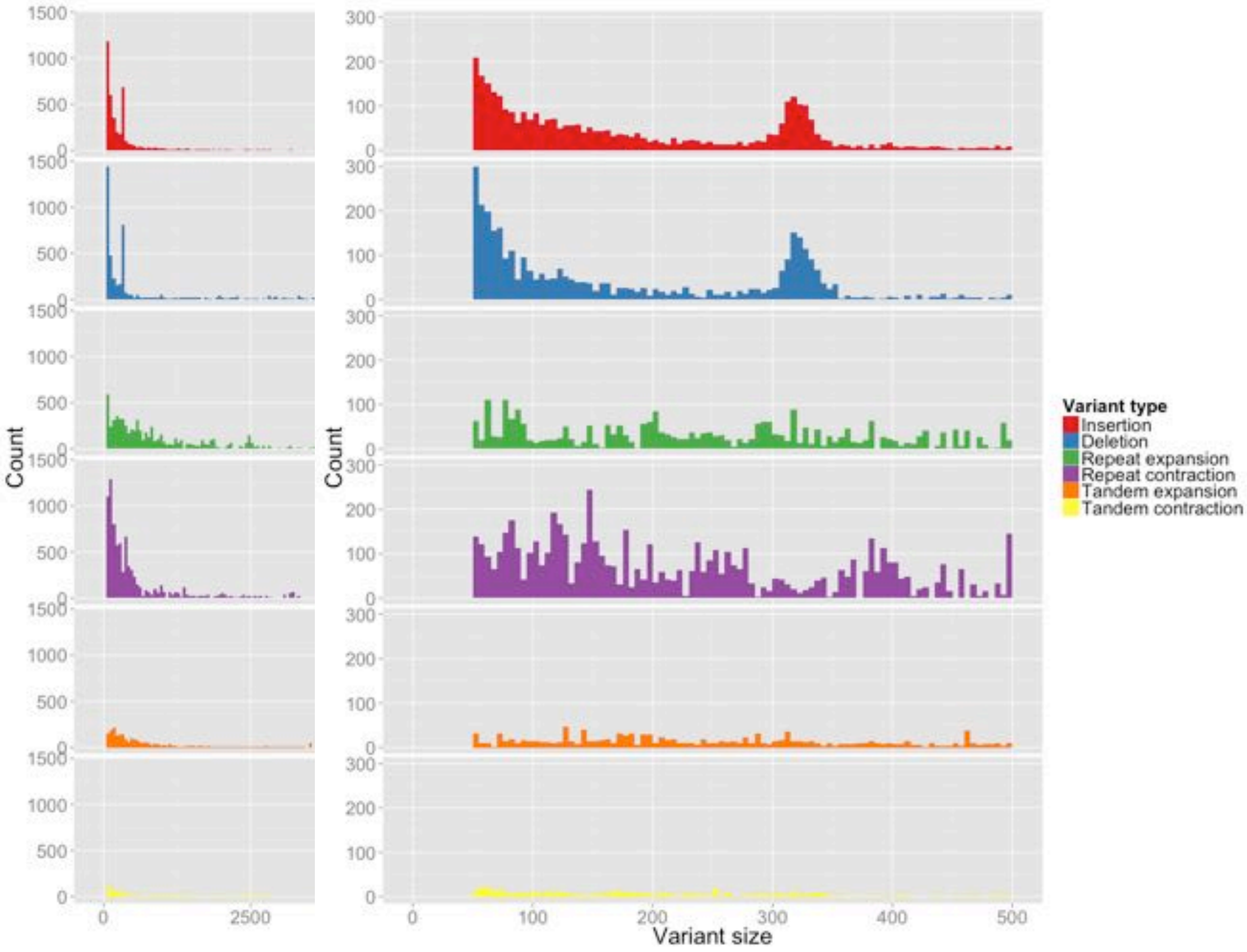
Repeat Expansions

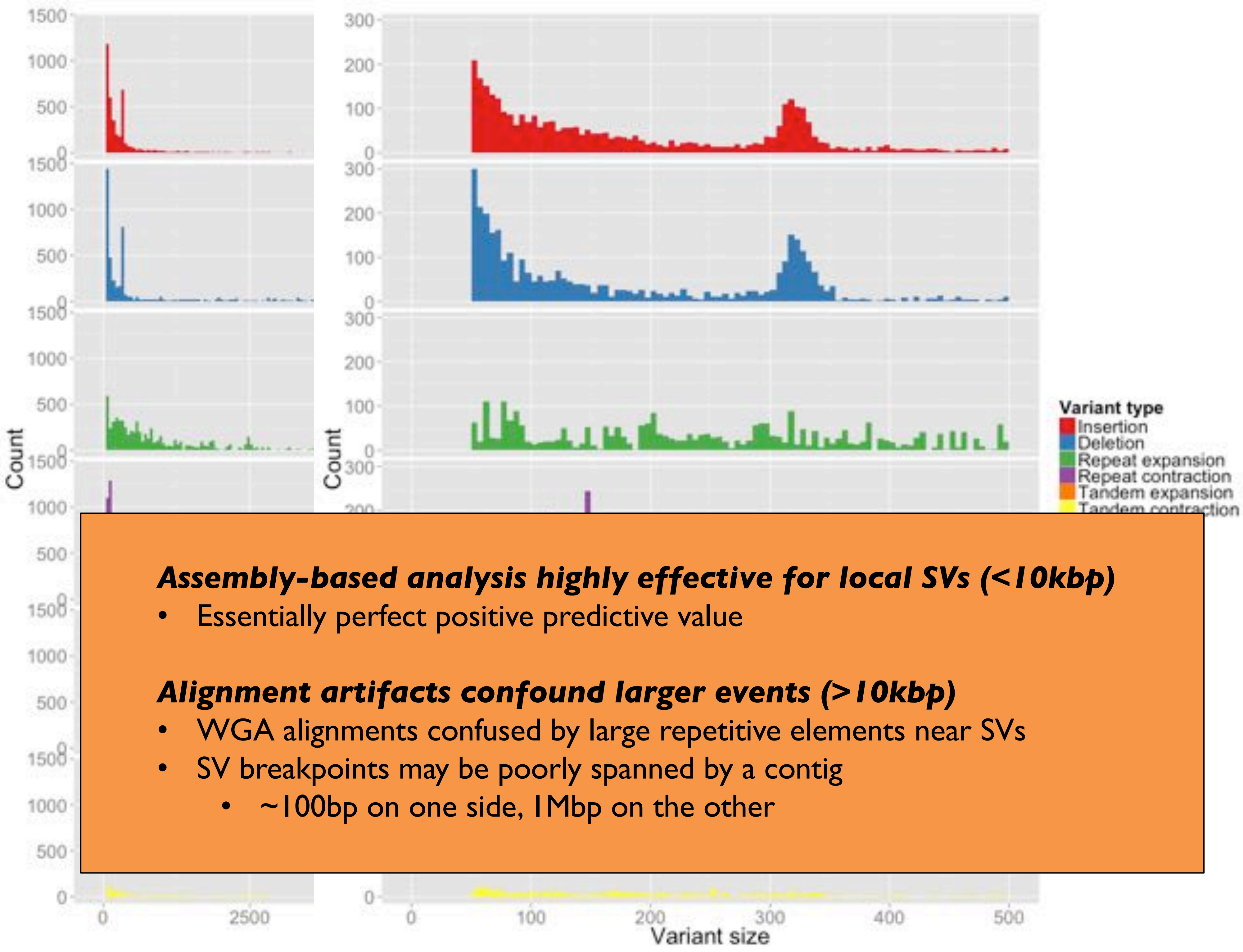


Repeat Contractions



Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad, M, Schatz, MC (2016) Bioinformatics doi: 10.1093/bioinformatics/btw369


Count
1500
1000
500
0
0
2500
Count
300
200
100
0
0
100
200
300
400
500
Variant size
Variant type
🟥 Insertion
🟦 Deletion
🟩 Repeat expansion
🟪 Repeat contraction
🟧 Tandem expansion
🟨 Tandem contraction



Assembly-based analysis highly effective for local SVs (<10kbp)

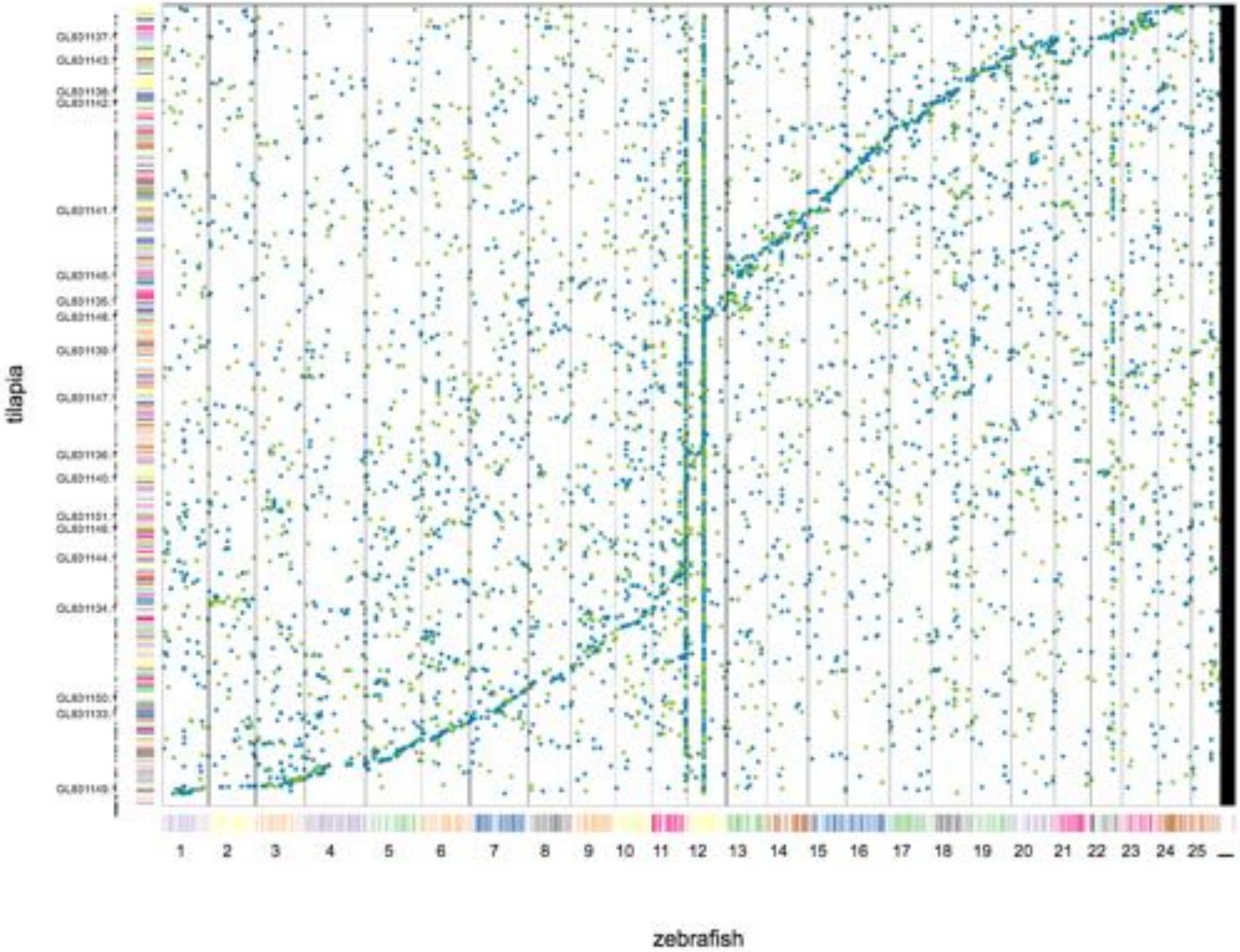
- Essentially perfect positive predictive value

Alignment artifacts confound larger events (>10kbp)

- WGA alignments confused by large repetitive elements near SVs
- SV breakpoints may be poorly spanned by a contig
  - ~100bp on one side, 1Mbp on the other

Dot: Interactive Dot plots for Comparative Genomics

https://github.com/dnanexus/dot



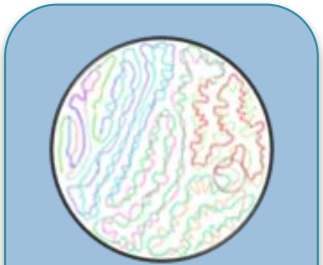
In pursuit of perfect genome sequencing

- ***Strive for Perfection: 100% Correct and 100% Complete***
  - The key for perfect genomes is lonnnnnnnnng reads ☺
  - Expect new insights on the causes of diseases, forces of evolution

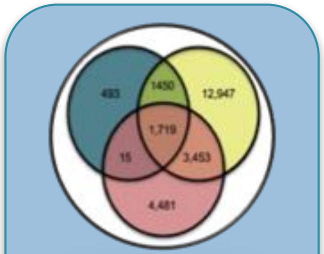
- ***Multiple sequencing technologies & approaches needed***
  - *PacBio*: Best Resolution of SVs
  - *10X/HIC*: Best Phasing
  - *De novo*: Best Resolution of small SVs
  - *Mapping*: Best resolution of large SVs

- ***We have just begun to explore the universe of variants present***
  - Tens of thousands of SVs per person, many megabases of variation
  - Also need to push these ideas into single cell and population scale analysis

CrossStitch

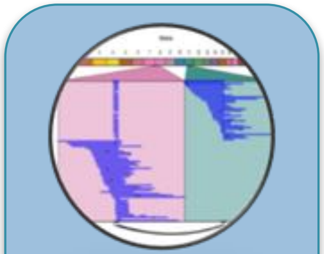


FALCON

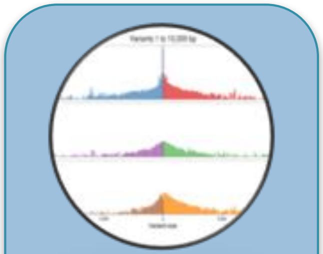


SURVIVOR

NGM+Sniffles



Ribbon



Assemblytics

http://schatz-lab.org

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

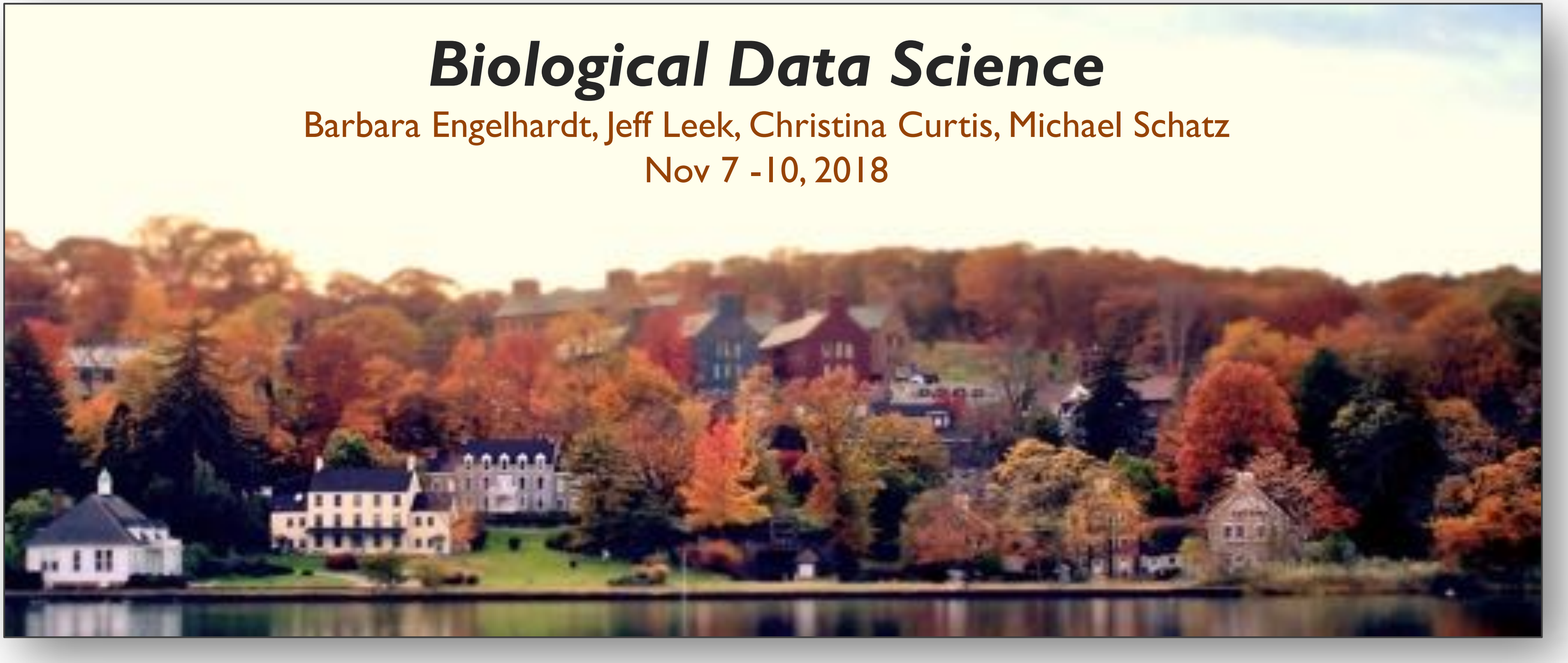
NSF

National Human
Genome Research
Institute

U.S. DEPARTMENT OF
ENERGY

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN
FOUNDATION



Biological Data Science

Barbara Engelhardt, Jeff Leek, Christina Curtis, Michael Schatz

Nov 7 -10, 2018

Thank you!

@mike_schatz