In pursuit of perfect personal genomes

Michael Schatz

Feb 13, 2018
AGBT Informatics

@mike_schatz


"Without a doubt, this is the most important, most wondrous map ever produced by humankind."
June 26, 2000

- ***The "reference" doesn't represent *any* human***
- ***Your sample may contain unique genes, gene structures, and other sequences not in the reference***
- ***Mapping short reads to the reference can bias the results***
- ***The reference can limit analysis of how genome variant impact regulation and expression or allele-specific features***

- ***De novo assembly, while greatly improved, is still slow, demanding and unpredictable***

"Without a doubt, this is the most important, most wondrous map ever produced by humankind." *June 26, 2000*

Reference Guided Assembly

1. High quality reference

- Contig N50 over 1Mbp
- Scaffold N50 over 10Mbp
- High Quality Gene Annotation
- Your sample is sufficiently similar (~99%)

2. Sample specific data

- *SNPs and Indels:* Illumina-based (PE/10X)
- *Structural Variants*: Long PacBio/ONT
- *Phasing Data*: 10X and/or HiC; trios

Comparative Genome Assembly ("AMOScmp")

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

Reference Guided Assembly

1. High quality reference

- Contig N50 over 1Mbp
- Scaffold N50 over 10Mbp
- High Quality Gene Annotation
- Your sample is sufficiently similar (~99%)

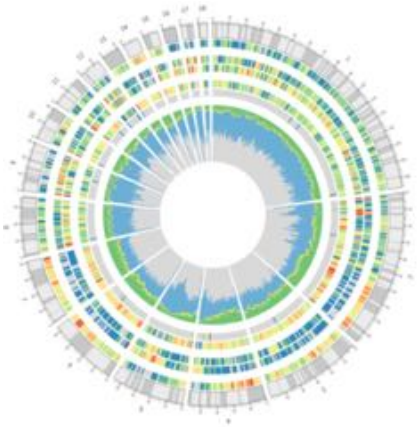
2. Sample specific data

- *SNPs and Indels*: Illumina-based (PE/10X)
- *Structural Variants*: Long PacBio/ONT
- *Phasing Data*: 10X and/or HiC; trios

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

CrossStitch

https://github.com/schatzlab/crossstitch



HQ Reference

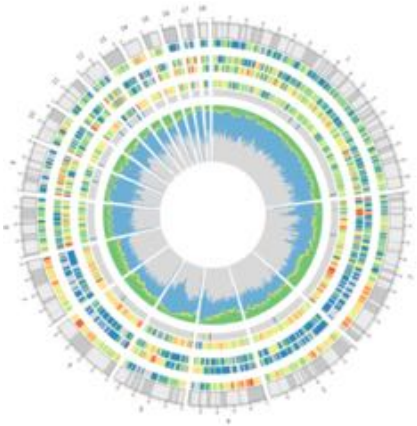
my.mat.fa

my.pat.fa

In collaboration with Sedlazeck, Gingeras, Guigo, 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

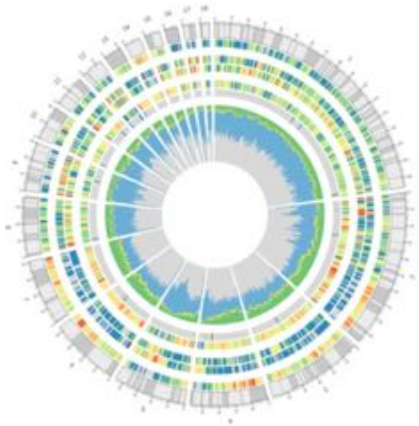
Assemble
Diploid
Sequence

my.mat.fa

my.pat.fa

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

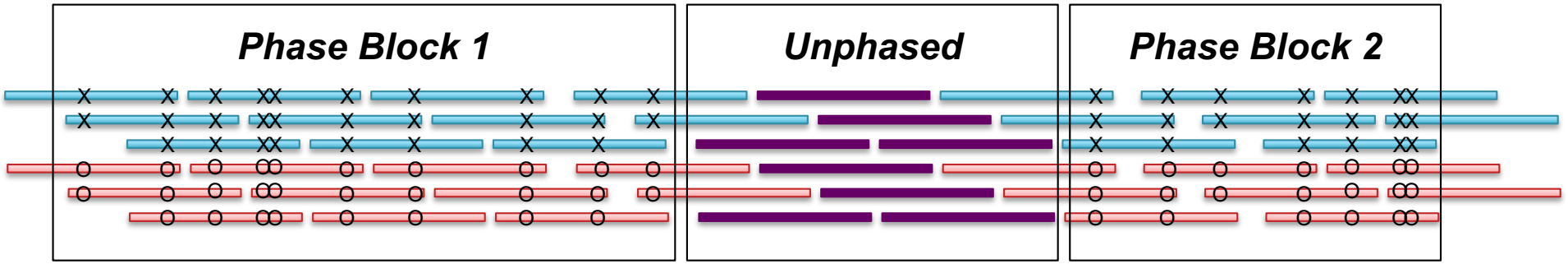
Phased
SNPs,
Indels &
SVs.vcf

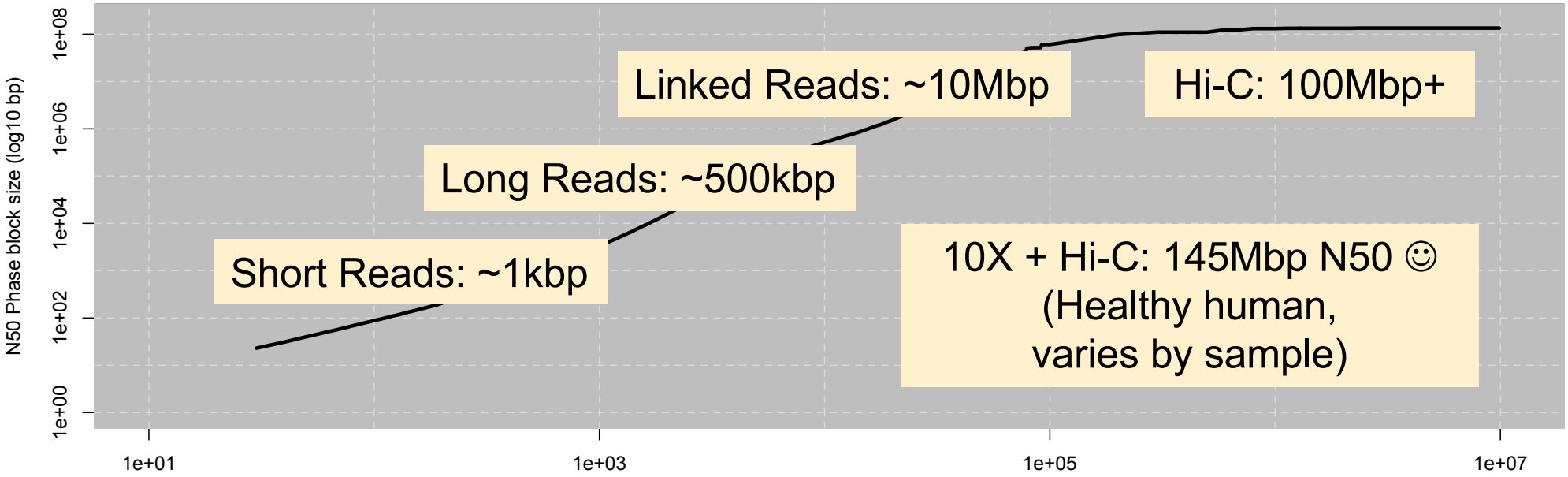
Assemble
Diploid
Sequence

my.mat.fa

my.pat.fa

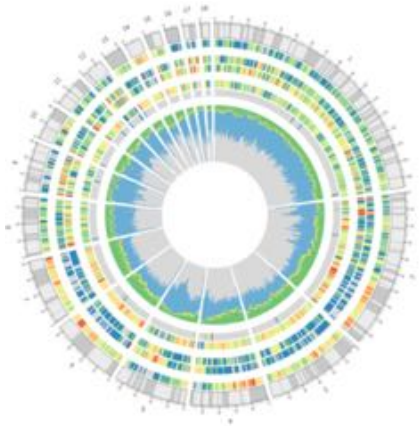
Phasing Results





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

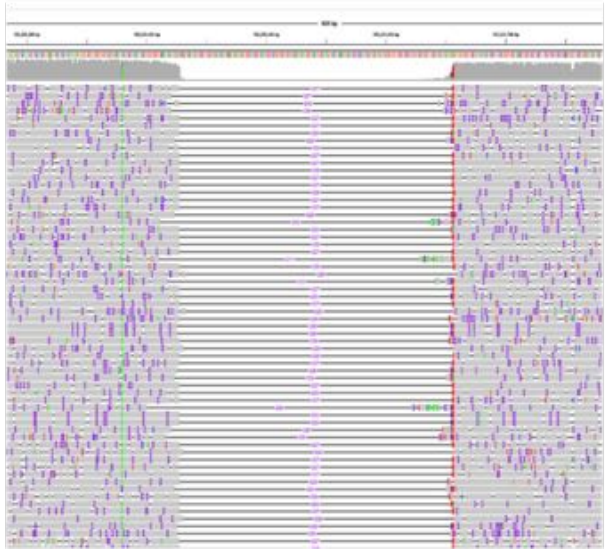
my.pat.fa

Fritz
Sedlazeck
Poster: 201

NGMLR + Sniffles

BWA-MEM

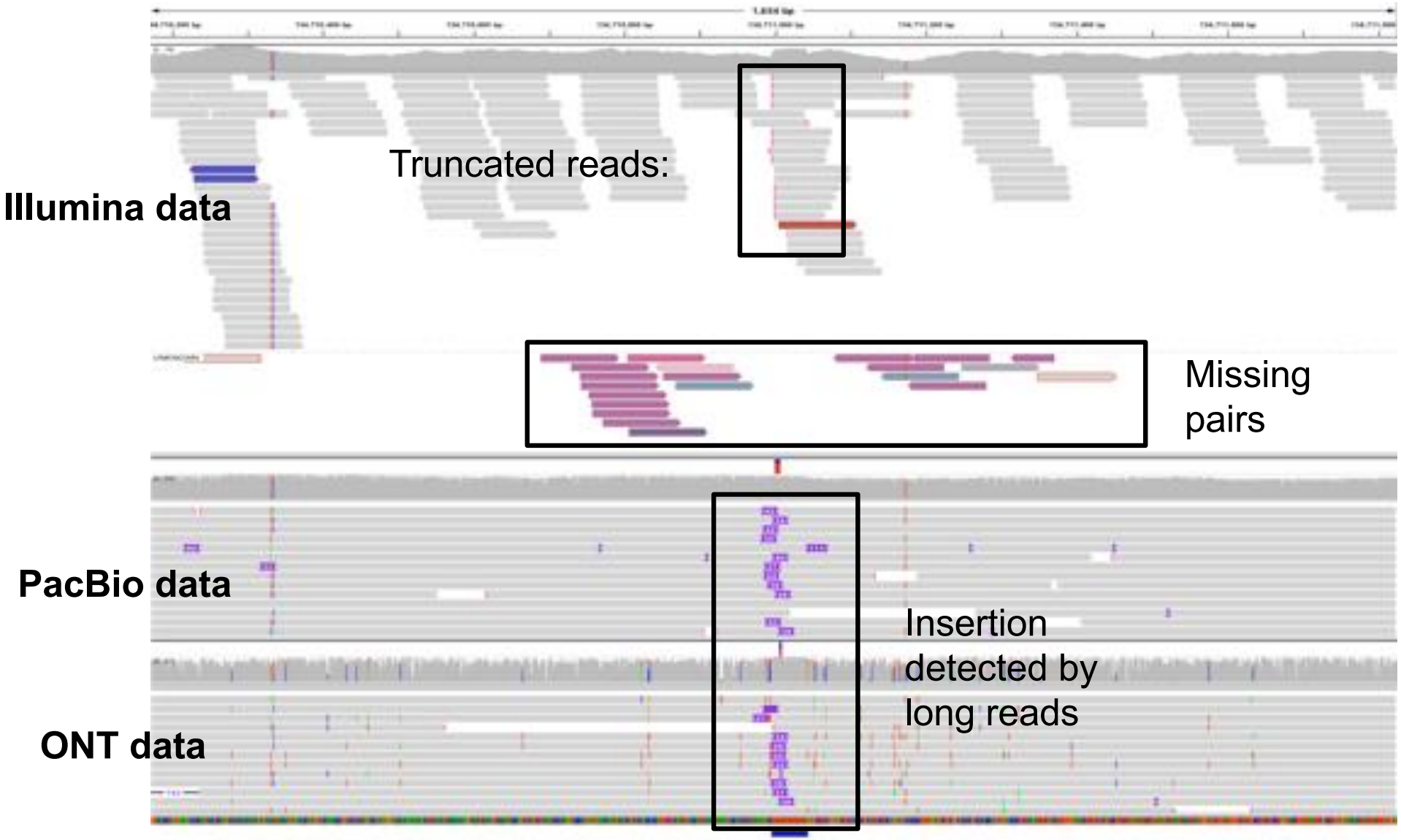
NGMLR



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*

Illumina data

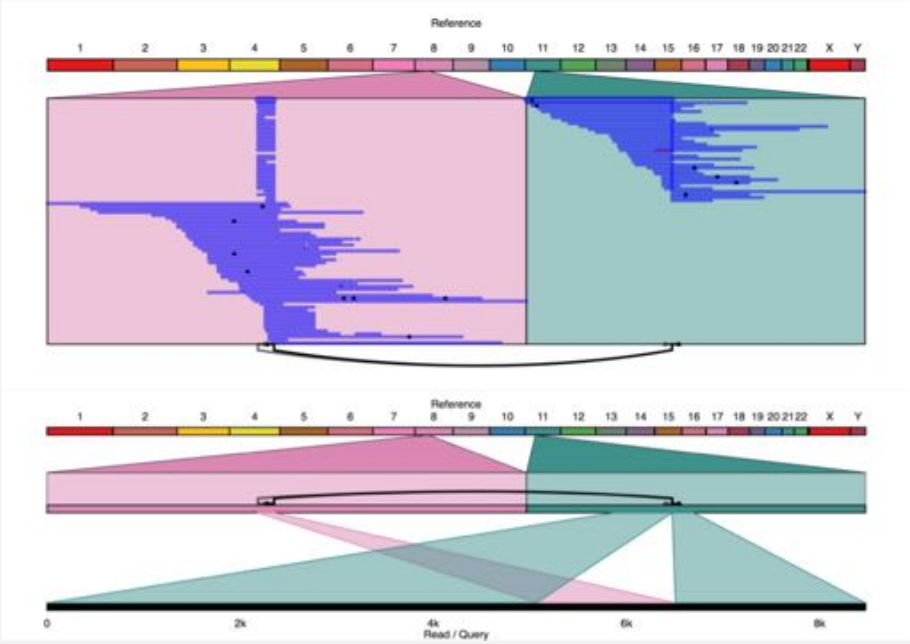

Illumina data
Truncated reads:
Missing pairs
PacBio data
Insertion detected by long reads
ONT data

SVs in a typical healthy human

Sniffles calls

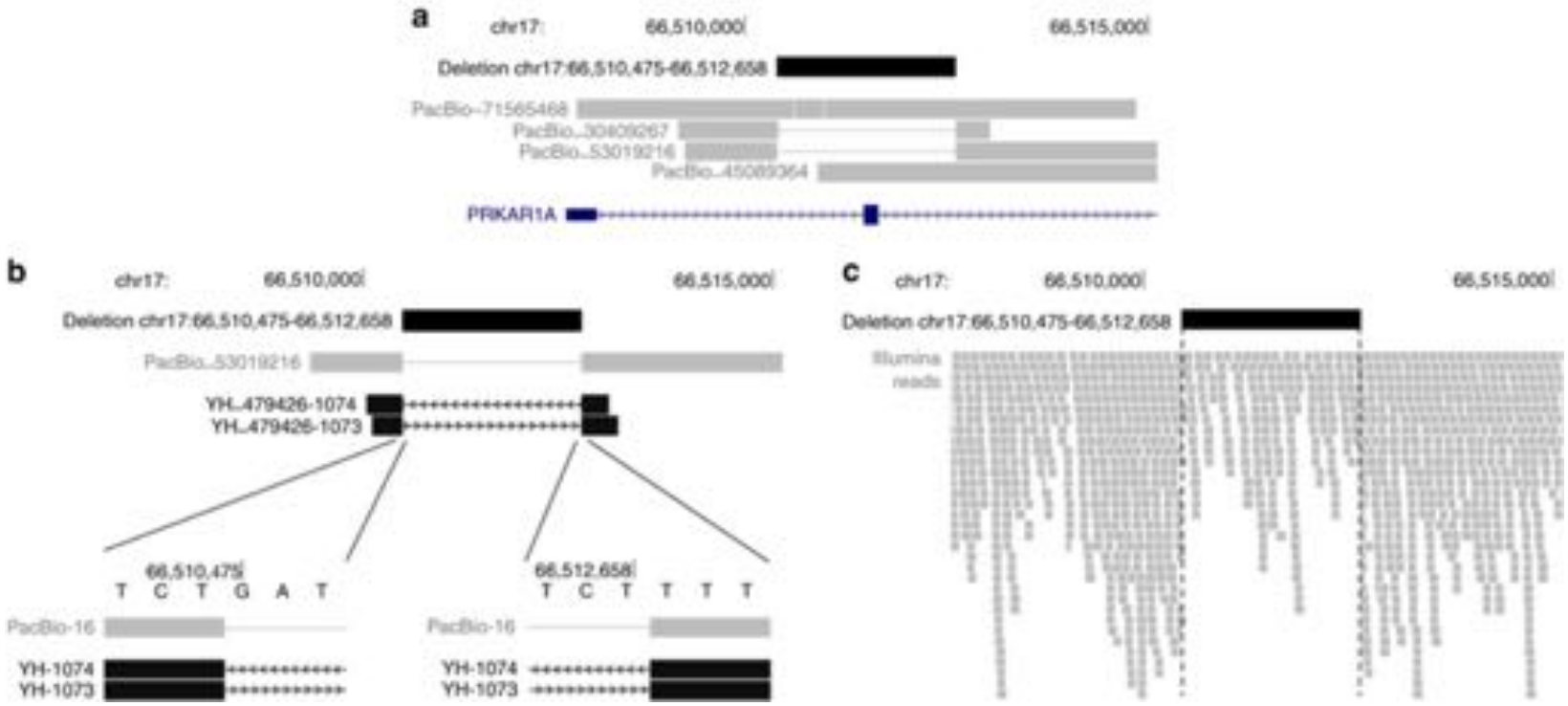
	All SVs (50bp+)	Large SVs (10kbp+)
Deletions	7,389	164
Duplications	1,284	139
Insertions	8,382	4
Inversions	229	116
Translocations	170	170
All	***17,454***	***593***

Translocation in Ribbon



Ribbon: Visualizing complex genome alignments and structural variation
Nattestad et al. (2016) *bioRxiv* doi: http://dx.doi.org/10.1101/082123

Structural Variations in Human Disease

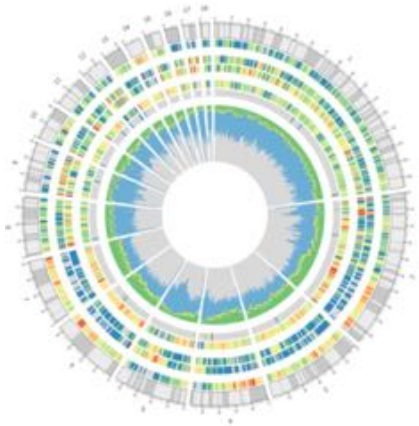


W. Richard McCombie
Thursday
@ 3:50p

Long-read genome sequencing identifies causal structural variation in a Mendelian disease
Merker et al (2017) *Genetics in Medicine. doi:10.1038/gim.2017.86*

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

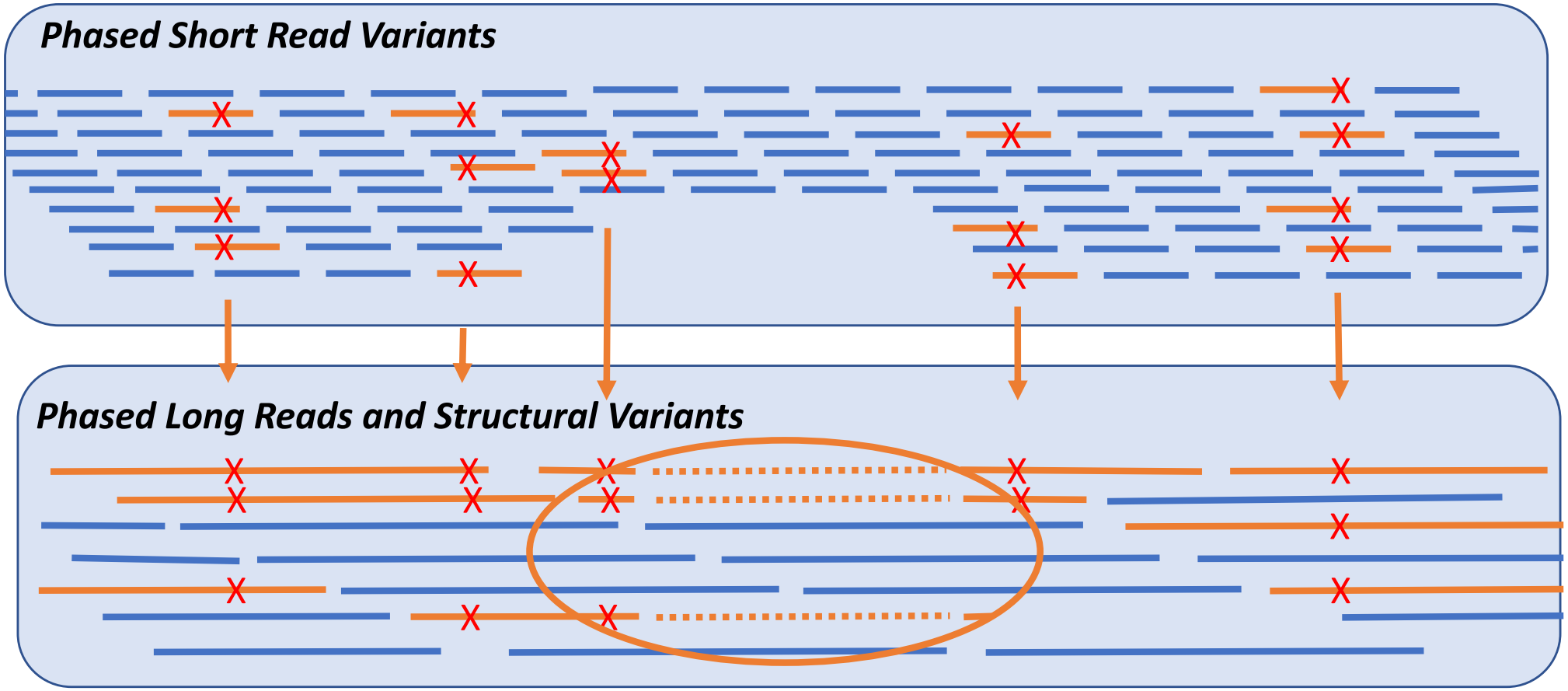
Phased
SNPs,
Indels &
SVs.vcf

Assemble
Diploid
Sequence

my.mat.fa

my.pat.fa

Hybrid Phasing and Local Assembly

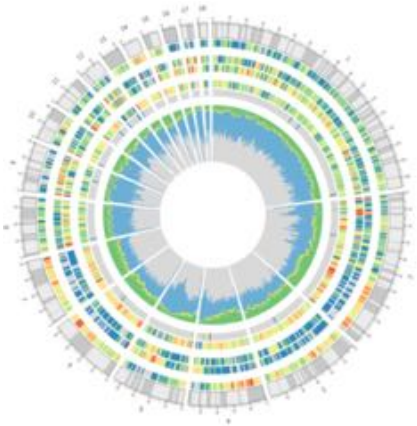


Phase SVs: Determine the haplotype of each read and each SV

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

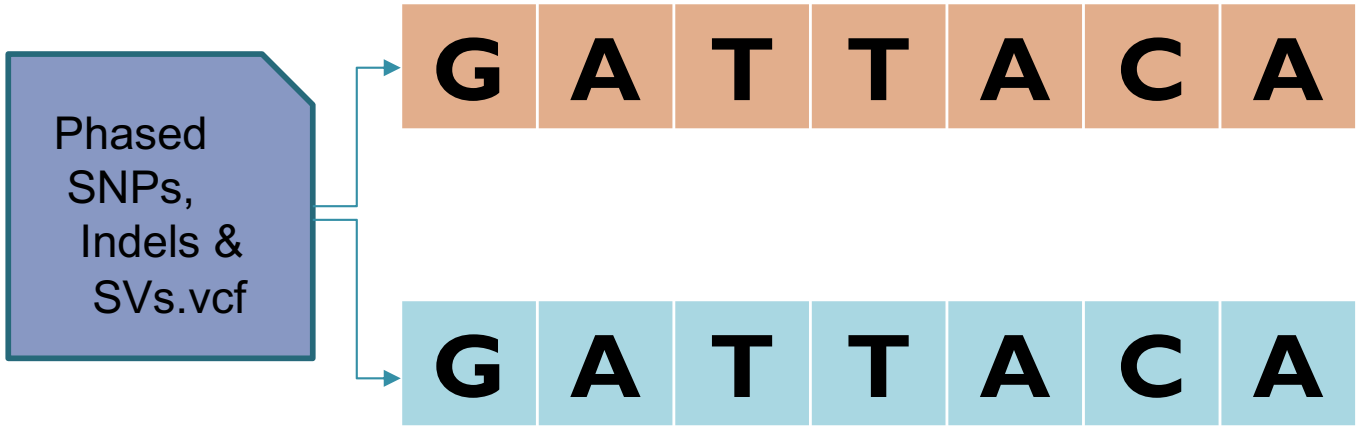
Phased
SNPs,
Indels &
SVs.vcf

Assemble
Diploid
Sequence

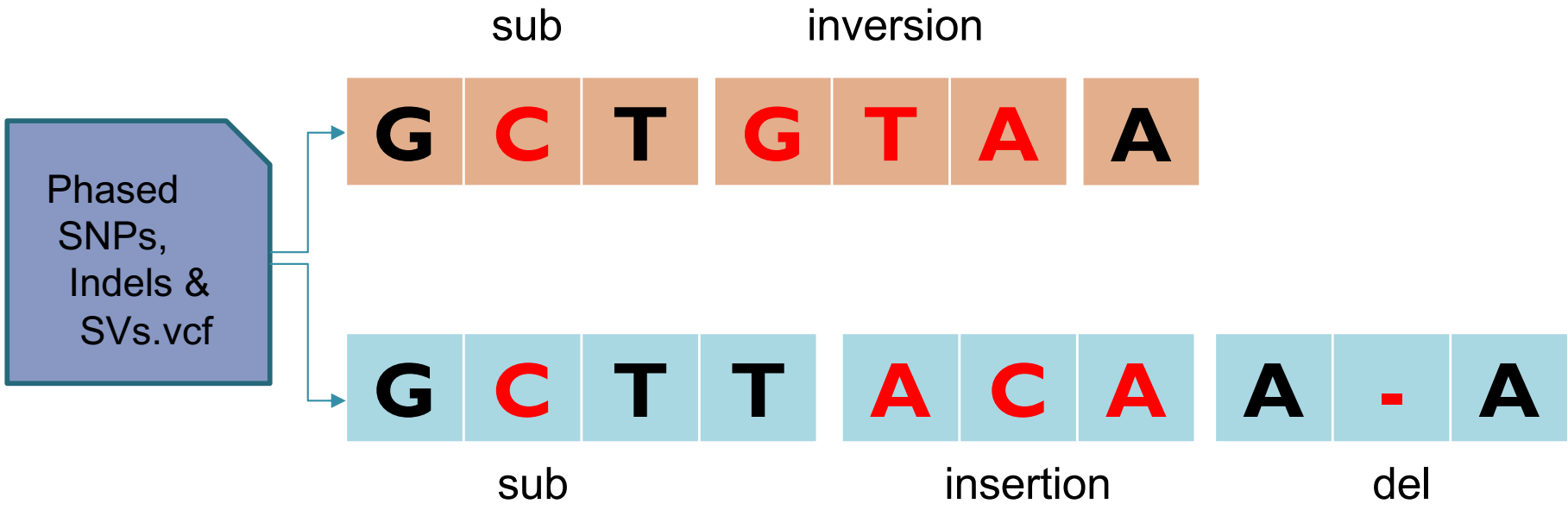
my.mat.fa

my.pat.fa

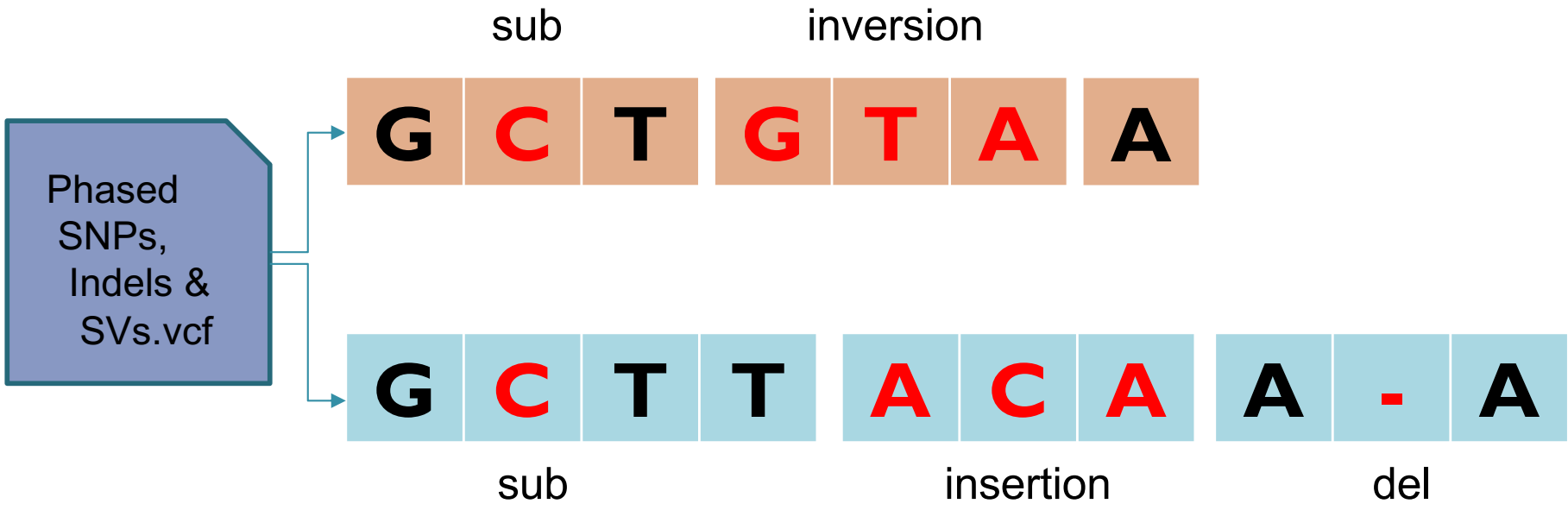
Assembling a "Perfect" Personalized Diploid Genome



Assembling a "Perfect" Personalized Diploid Genome



Assembling a "Perfect" Personalized Diploid Genome



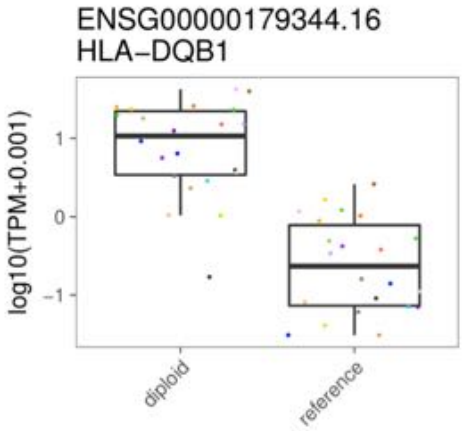
Stitching based on AlleleSeq pipeline enhanced for SVs (Rozowsky et al, 2011)

- Maintains a mapping from reference to personal genome coordinates for liftover

Using 10X + HiC + PacBio, assemble nearly perfect diploid human genomes

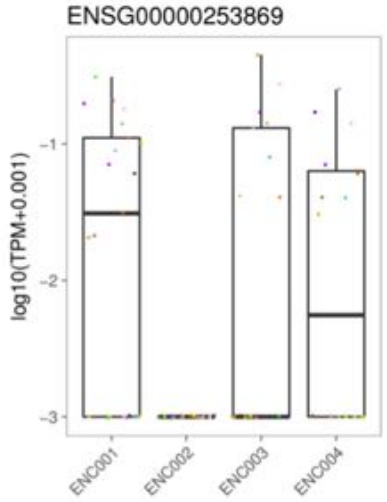
- Phased diploid genome can be aligned or aligned against

Improved mapping of functional data



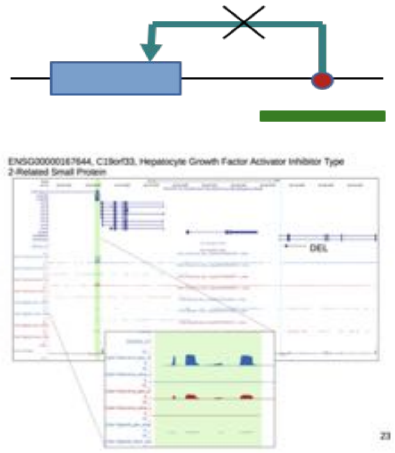
- Typically 10k – 100k additional mapped RNA-seq reads per sample; mappability more complicated

Expression of deleted genes and promoters



- Heterozygous or homozygous deletions of genes and promoters often show reduced expression

SVs intersecting eQTLS

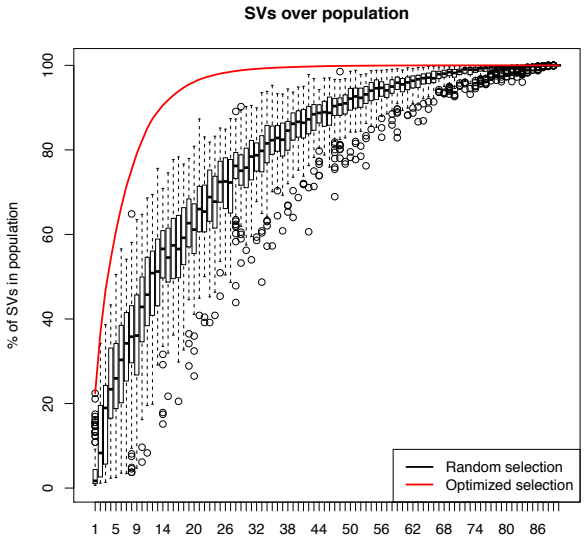


- Deletions overlapping a SNP eQTL affects the expression of the target gene; further analysis in progress

Reference-quality Genomes without *de novo* assembly

Why should we assemble perfect personal genomes?

- Pathogenic and other important variants might be missed
- Improved mapping, fixes "differential" expression, allele-specific
- Explore interplay between variation, regulation, and expression



Multiple sequencing technologies & approaches needed

- *>20x coverage PacBio/ONT*: Best Resolution of SVs
- *>20x coverage 10X/HIC:* Best Phasing
- Trio or Population-based phasing also possible to reduce costs

We have just begun to explore the universe of variants present

- Also need to push these ideas into single cell and population scale analysis

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

CSHL
Gingeras Lab
McCombie Lab

GRC
Roderic Guido
Alessandra Breschi
Anna Vlasova

University of Vienna
Arndt von Haeseler
Philipp Rescheneder

DNAnexus
Maria Nattestad

PacBio
Greg Concepcion

ENCODE Partners
Berstein Lab
Gerstein Lab
Myers Lab
Ren Lab
Snyder Lab
Stam Lab
Wold Lab

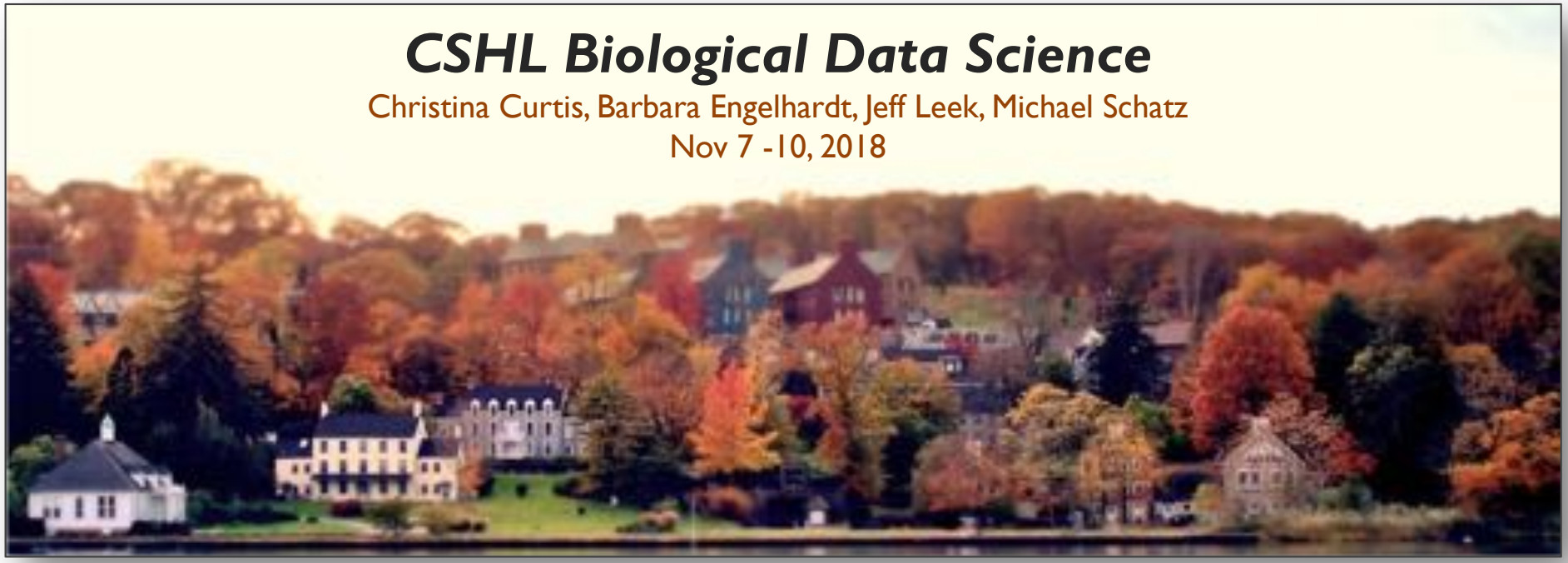
+ All ENCODE Members

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

ALFRED P. SLOAN FOUNDATION

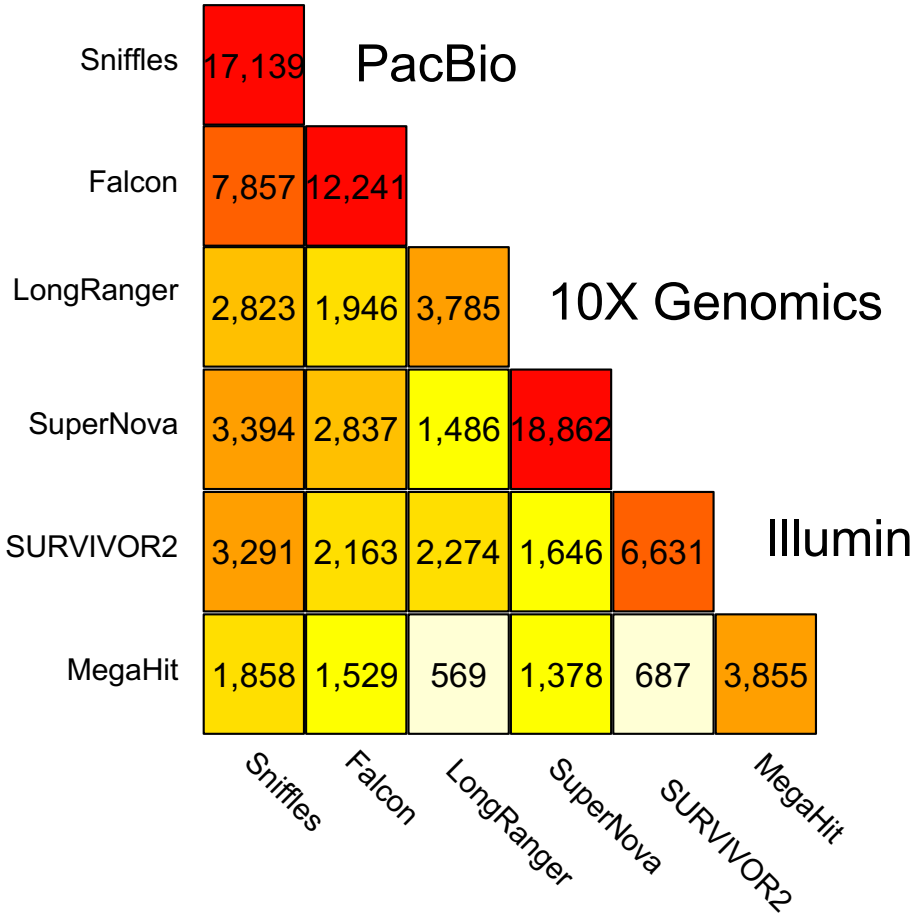


Thank you

http://schatz-lab.org

@mike_schatz

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 ☺

Expression & Regulation

Foundation for mapping functional data

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

Population Genetics

Framework for GWAS of Structural Variations

- Many GWAS SNPs appear to be in linkage with SVs that are the likely functional variant
- Resequencing key individuals with phenotype data available

Tumor Progression

Chromosome instability in breast cancer

- 10X, PacBio and Oxford Nanopore sequecning of breast cancer samples from Northwell Health
- Cell lines, patient tissues, and patient-derived organoids

Analysis in progress…

- Construct personal genome and personal annotation for all individuals
- Expression changes due to SVs overlapping functional elements, i.e. enhancers, eQTLs SNPs and short indel analysis
- Novel transcription elements in insertions
- Chimeric transcripts in reference and personal genomes
- Allele specific expression and binding
- Integrate other functional assays to perform tissue specific analysis, i.e. smallRNAs, RAMPAGE, ChiP-seq
- … and many more …