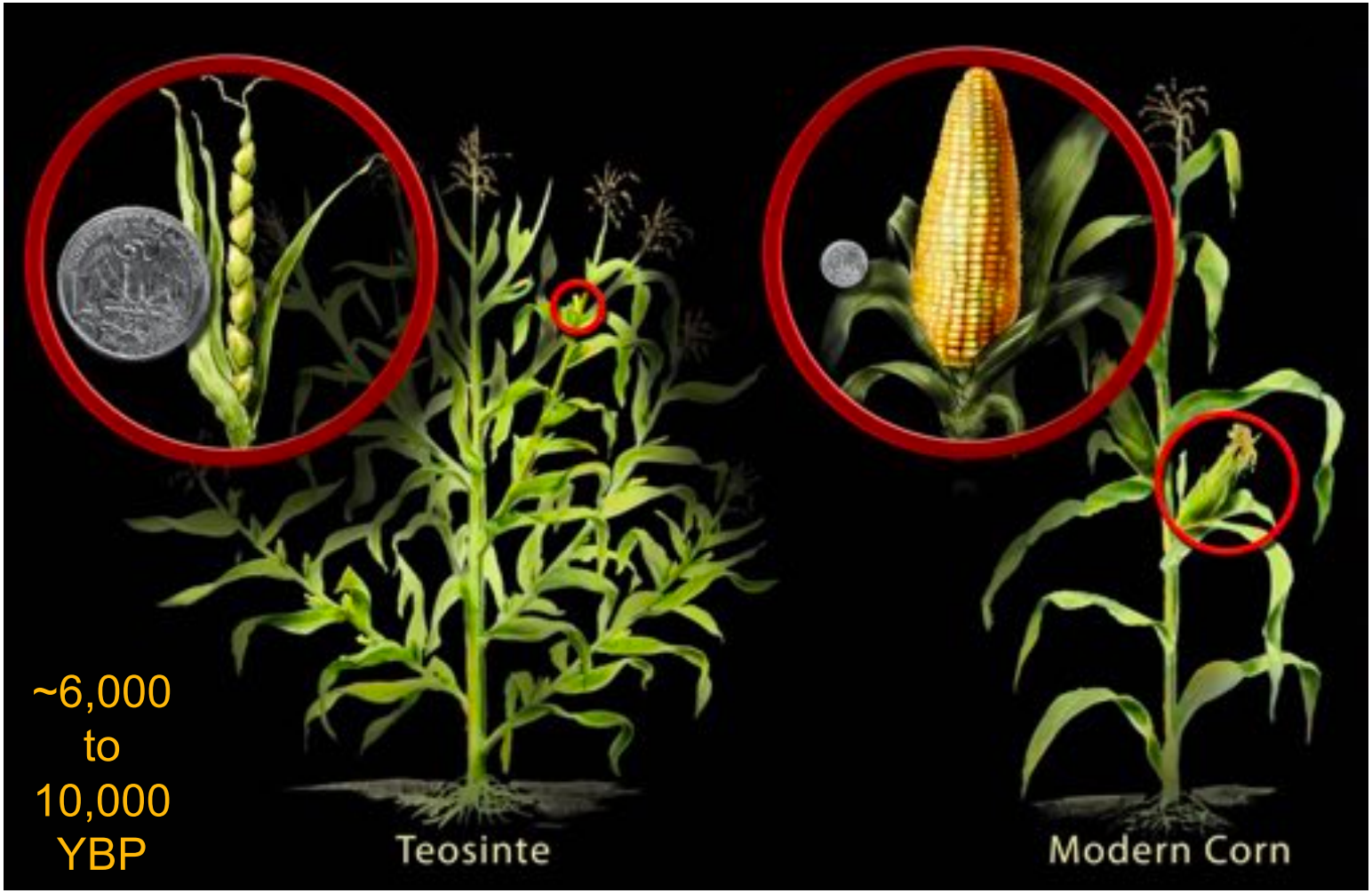
Plant Genomics

Michael Schatz

Nov 12, 2019

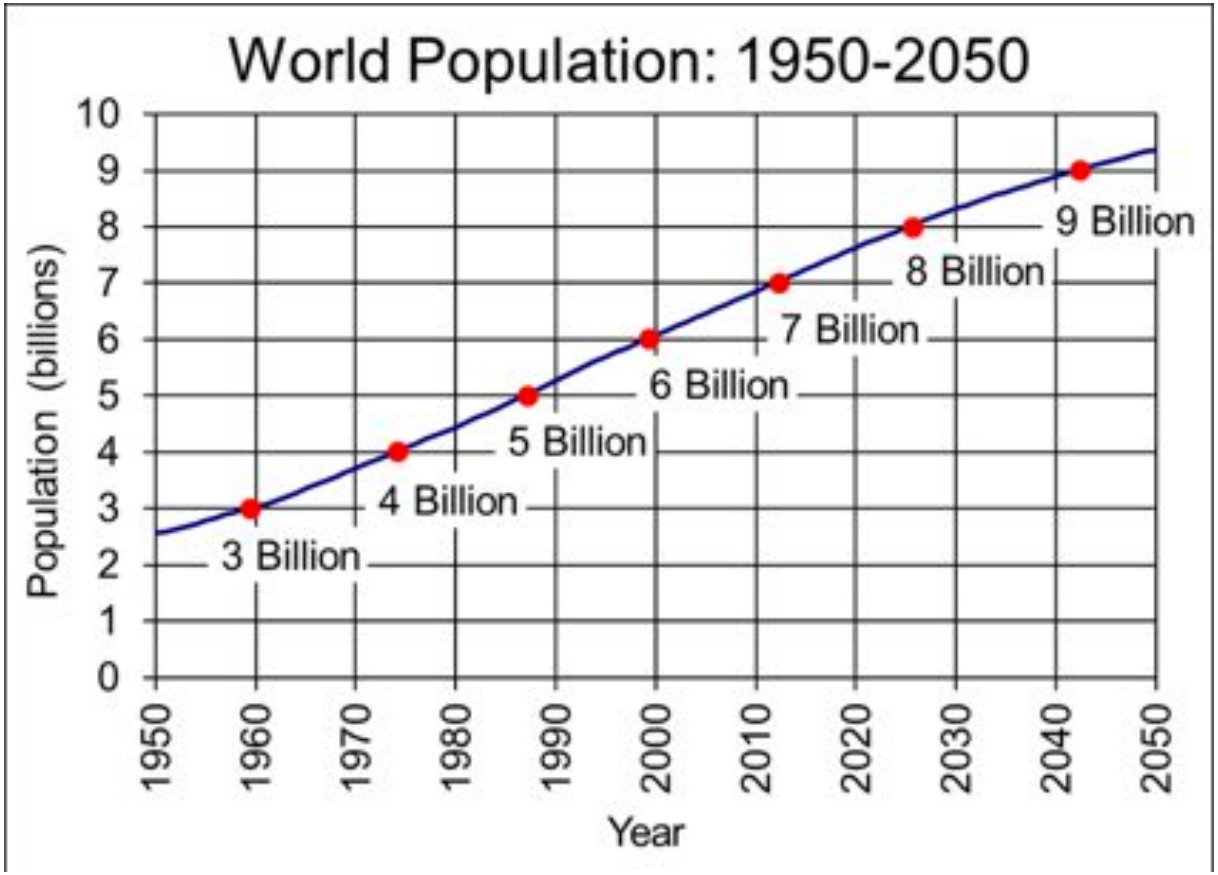
Lecture 22: Computational Biomedical Research

Earliest Genomics



Jumping Genes

The origin and behavior of mutable loci in maize.
McClintock, B. (1950) *PNAS*. 36(6):344–355.
Nobel Prize in Physiology or Medicine in 1983



Projected world population estimated by the United Nations

List of most valuable crops and livestock products

From Wikipedia, the free encyclopedia

The following list, derived from the statistics of the Food and Agriculture Organization (FAO) unless otherwise noted, lists the most important agricultural products produced by the countries of the world.[1][2]

The data in this article, unless otherwise noted, is for 2012. The value and production of individual crops varies substantially from year to year as prices fluctuate on the world and country markets and weather and other factors influence production.

Crop	Type	global value in billion US dollars (2012)	global production in metric tons (2011)	Top producing country and value billion US dollars (2012)
Rice, paddy	Cereals	$337	740,961,445	$117 (Mainland China) **2016 data**
Cattle, meat	Livestock	$336	65,973,820	$53 (United States) **2016 data**
Pig, meat	Livestock	$306	118,168,709	$167 (Mainland China) **2016 data**
Cow's milk, whole, fresh	Livestock Primary	$286	659,150,049	$35 (United States) **2016 data**
Chicken, meat	Livestock	$128	90,001,779	$24.4 (United States)
Wheat	Cereals	$84	701,395,334	$13.7 (Mainland China)
Soybeans		$65	262,037,569	$21.8 (United States)
Tomatoes	Vegetables	$58	159,347,031	$17.9 (Mainland China)
Sugarcane		$57	1,800,377,642	$23.9 (Brazil)

Rice in 2012 was the most valuable agricultural crop in the world. It was second to maize (corn) in the quantity produced of cereal products. This rice field is in Cambodia.

Tomato Domestication & Agriculture

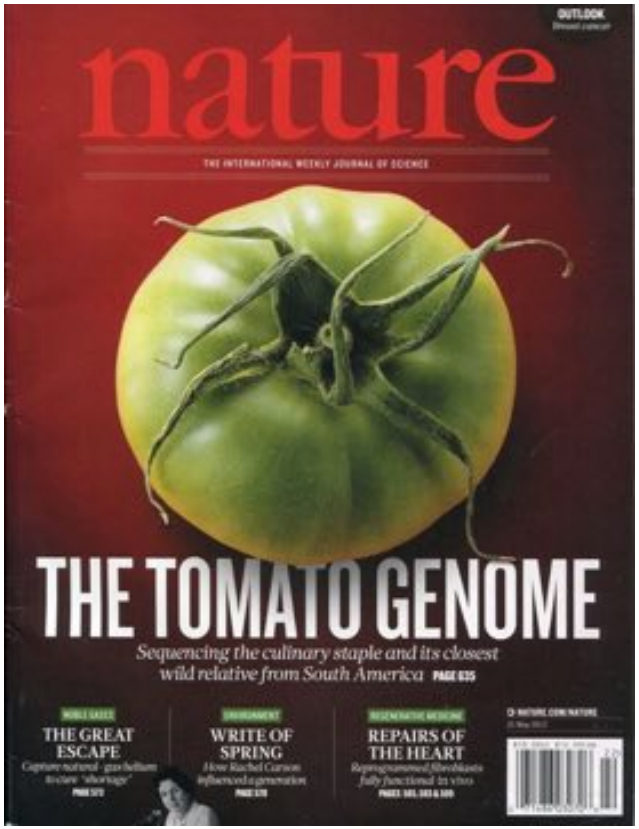
Tomatoes are one of the most valuable crops in the world

- Worldwide annual production >175 million tons & >$85B
- Major ingredient in many common foods:
  - Sauces, salsa, ketchup, soups, salads, etc

Tomatoes are an important plant model system

- Originally from South America, transported to Europe by early explorers in the 17th century, and then back to North America in the 18th century
- Extensive phenotypic variation: >15,000 named varieties
  - Model for studying fruiting and flowering
- Member of important Solanaceae family
  - Potato, pepper, eggplant, tobacco, petunia, etc

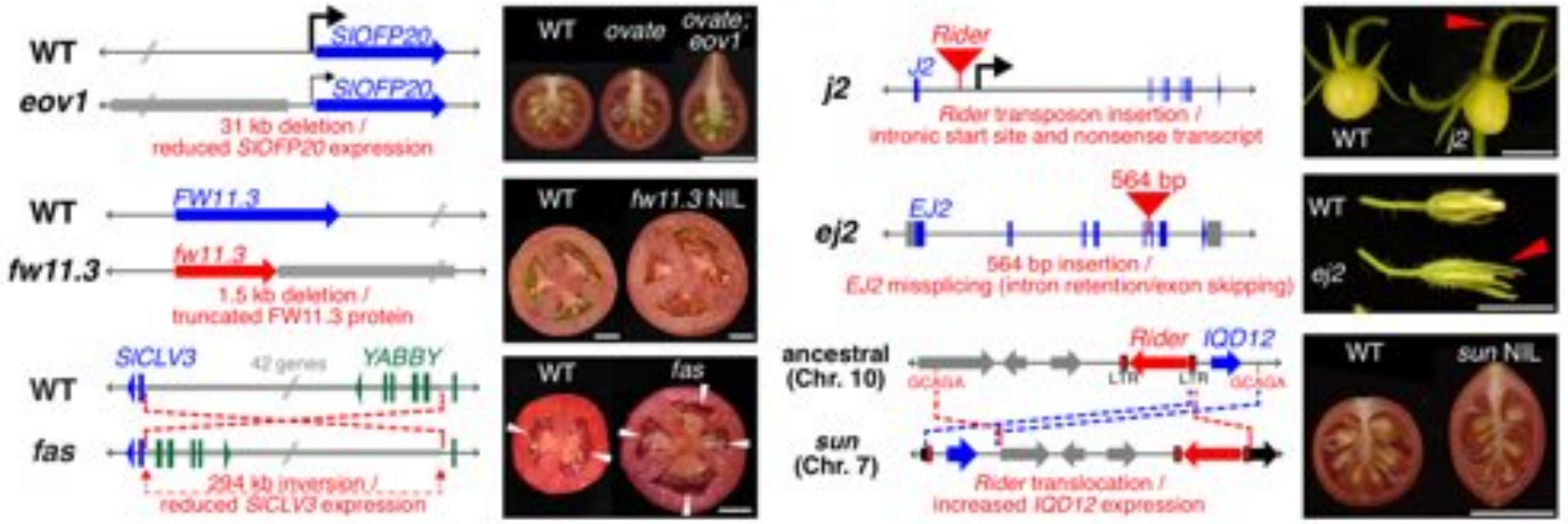
Tomato Genomics and Genetics



Tomato Reference Genome published in May 2012

- International consortium from 14 countries requiring years of effort and tens of millions of dollars
  - Sanger + 454 + fosmids + BAC-ends + genetic map + FISH
- 'Heinz 1706' cultivar (v3)
  - 12 chromosomes, 950 Mbp genome, diploid
  - 22,707 contigs, 133kbp contig N50, 80M 'Ns'
  - 20Mb on "chromosome 0"
- Resource for thousands of studies
  - Candidate SNPs for many traits identified through GWAS
  - Candidate genes and pathways through RNAseq
  - Extensive investment into agricultural traits:
    - ripening, flavor, fruit size, color, morphology

Structural Variations Are Drivers of Quantitative Variation



Recent results highlight structural variations to play a major role in phenotypic differences

- SV are any variants >50bp: insertions, deletions, inversions, duplications, translocations, etc
- Adds, removes, and moves exons, binding sites, and other regulatory sequences
- Notoriously difficult to identify using short reads: high false positive & false negative rate

Structural Variation Landscapes in Tomato Genomes and their role in Natural Variation, Domestication, and Crop Improvement

Zach Lippman
CSHL / HHMI

Joyce Van Eck
Boyce Thompson

Esther van der Knaap
Univ. of Georgia

Fritz Sedlazeck
Baylor

Sara Goodwin
CSHL

Project overview

1. Select diverse samples
2. Sequence their genomes
3. Find novel genetic variants
4. Identify and validate variants associated with agricultural and phenotypic traits

Step 1: Select tomatoes

Tomato origins

Wild tomatoes
Solanum pimpinellifolium
”Currant tomatoes”

Early domesticated
S. lycopersicum var cerasiforme
“Cherry tomatoes”

Modern domesticated
S. lycopersicum
“Processing tomatoes”

Azoychka Russian Heirloom

Makes for a great salsa because it's ripe and fleshy, lower juice than other red tomatoes

Brandywine

Heirloom cultivar with large pink beefsteak-shaped fruit, popularly considered among the best tasting available.

Floradade

Delicious, bright red variety that has a great ability to withstand heat and produce high yields

Kumato

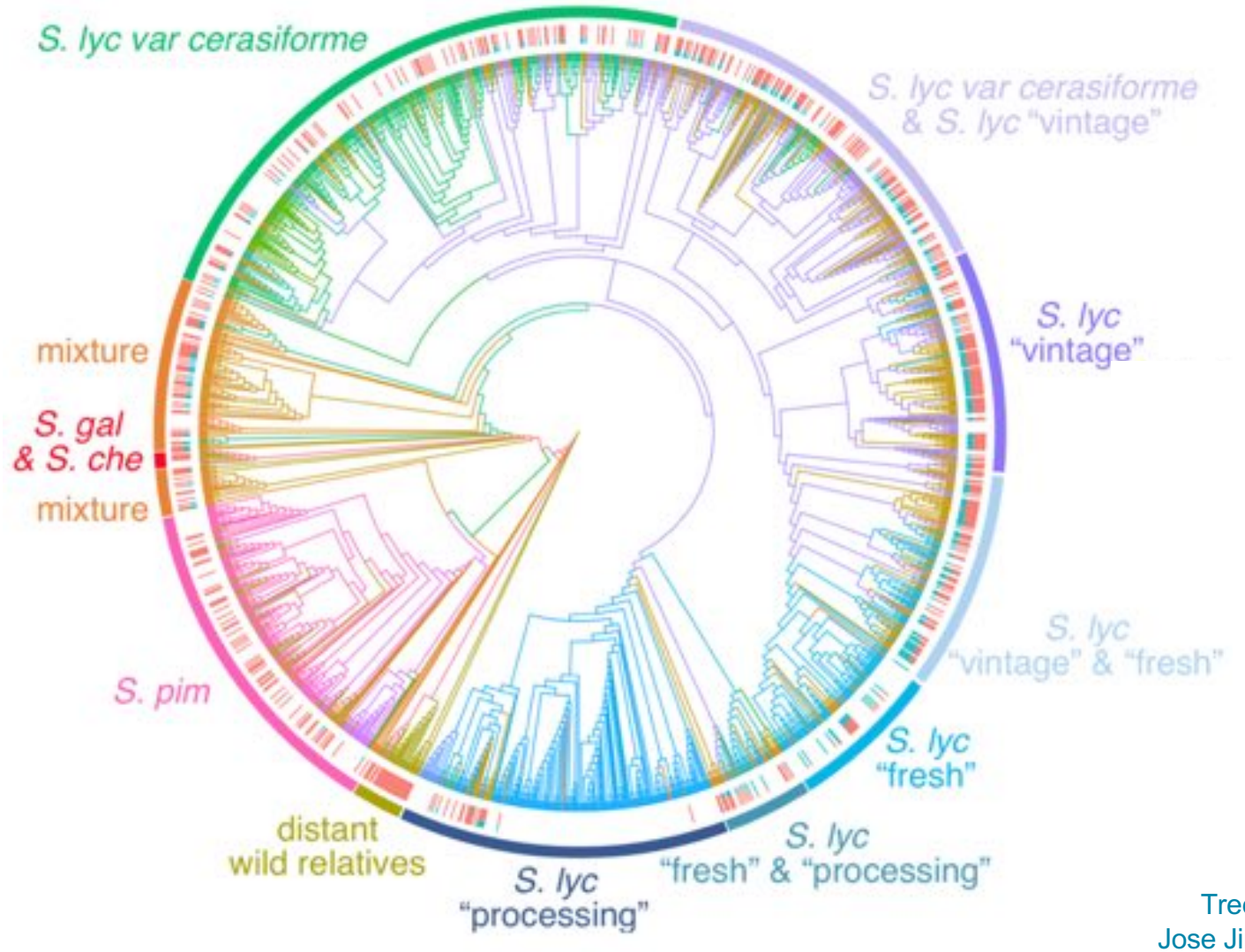
Trade name for a patented cultivar in Spain called 'Olmeca'; sweeter than most other varieties

Roma

Plum variety popularly used both for canning and pastes because of their slender and firm nature

Suncoast

A Florida crown jewel; carries the crimson gene, with an overall scarlet color inside and out.



Tree produced by:
Jose Jimenez-Gomez

Which samples to pick?

	Sample1	Sample2	Sample3	Sample4	Sample5
Var1	1	1	1	0	0
Var2	1	1	1	0	0
Var3	1	1	1	0	0
Var4	0	0	0	1	0
Var5	0	0	0	1	0
Var6	0	0	0	0	1

I can only afford to sequence 3 samples, but I want to maximize diversity

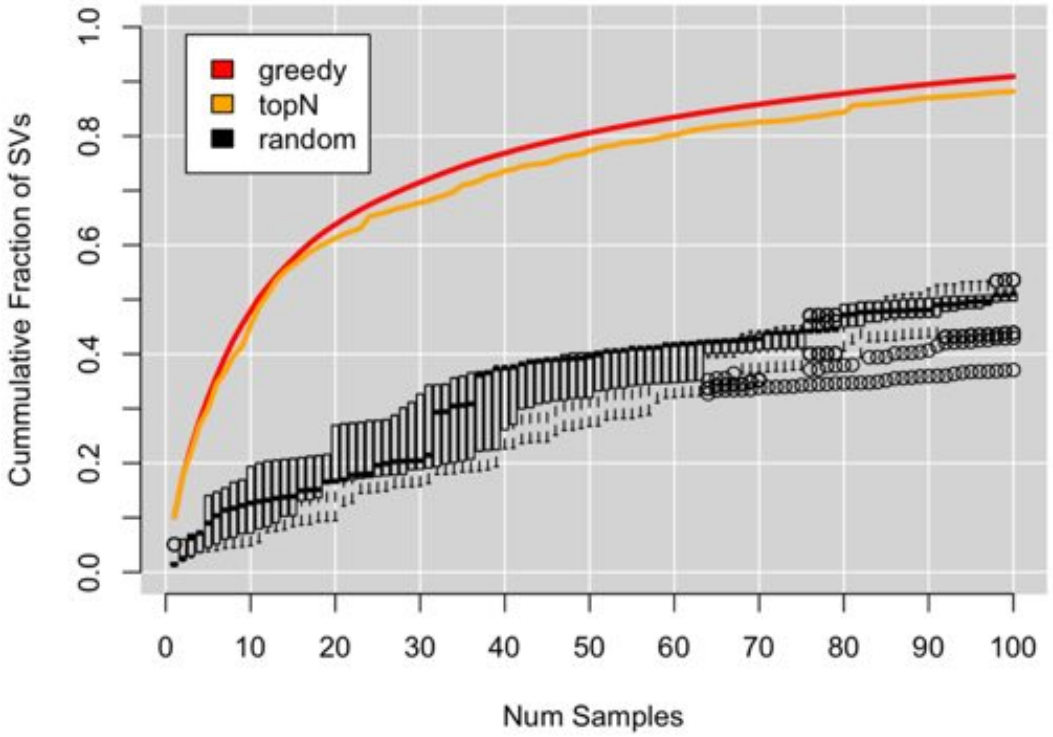
Random: Generally a good strategy, but high variance in diversity

Rank order: S1(3), S2(3), S3(3), S4(2), S5(1)
<- Starts good, but many overlapping variants

Diversity order: S1(3 new), S4(2 new), S5(1 new), S4(0 new), S5(0 new)
<- At every step lets bring in as much new diversity as possible

I can only afford to see 3 movies, and want to maximize diversity of actors/actresses

Optimized Sample Selection



Our goal is to select the 100 samples that *collectively* capture the most diversity

- Short-read based SVs will under-sample variants but still represents relative diversity
- Selecting 100 at ***random only recovers about 40%*** of the total known diversity
- ***Optimal strategy is NP-hard*** using a set-cover algorithm, we ***approximate using a greedy approach***
  - Ranking samples by number of variants picks diverse samples, although tends to pick siblings (nearly duplicate samples)

SVCollector: Optimized sample selection for validating and long-read resequencing of structural variants
Sedlazeck *et al* (2018) bioRxiv doi: https://doi.org/10.1101/342386

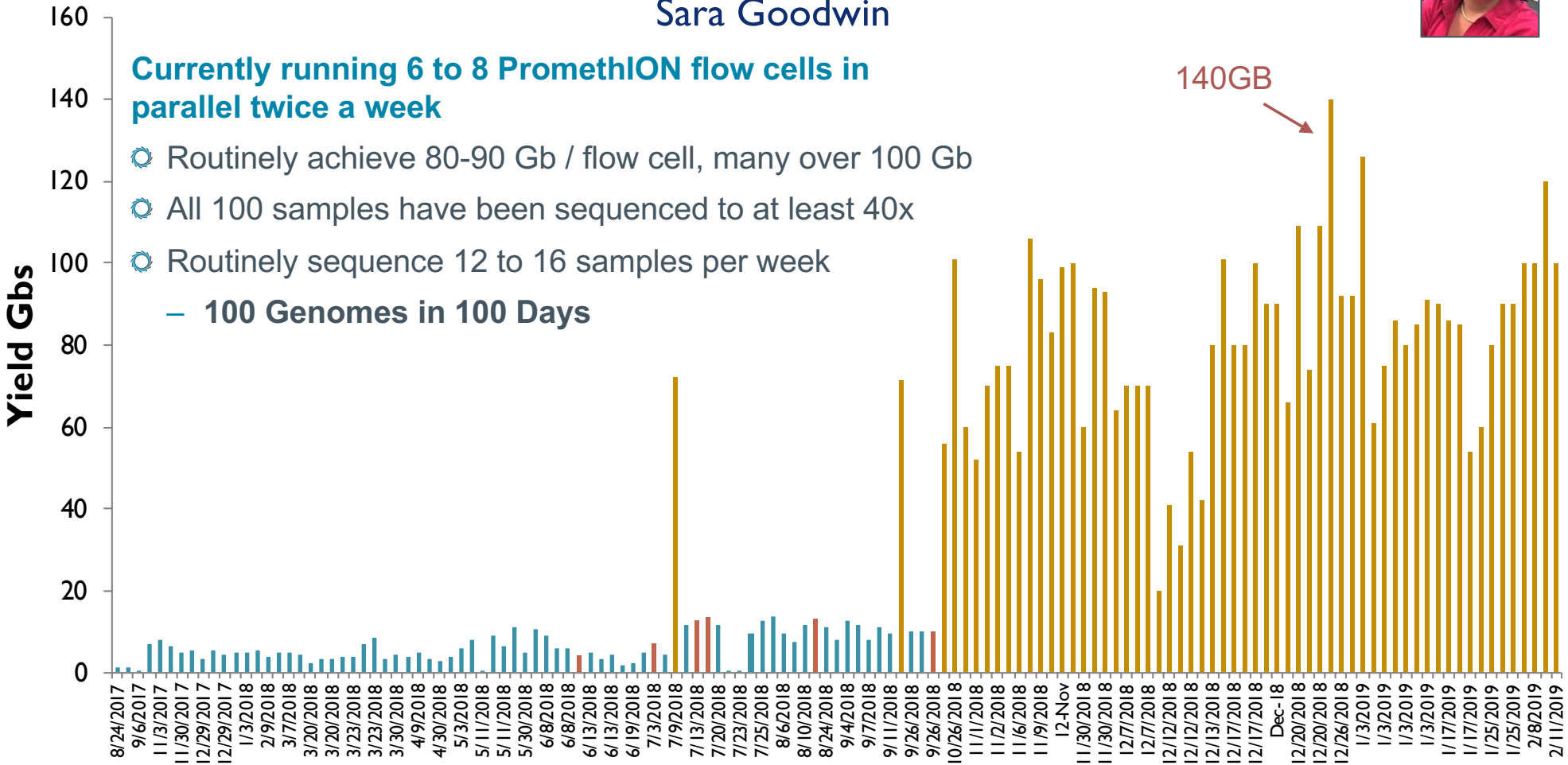
Step 2: Sequence genomes

Nanopore Performance at CSHL

Sara Goodwin

Currently running 6 to 8 PromethION flow cells in parallel twice a week

- Routinely achieve 80-90 Gb / flow cell, many over 100 Gb
- All 100 samples have been sequenced to at least 40x
- Routinely sequence 12 to 16 samples per week
  - **100 Genomes in 100 Days**



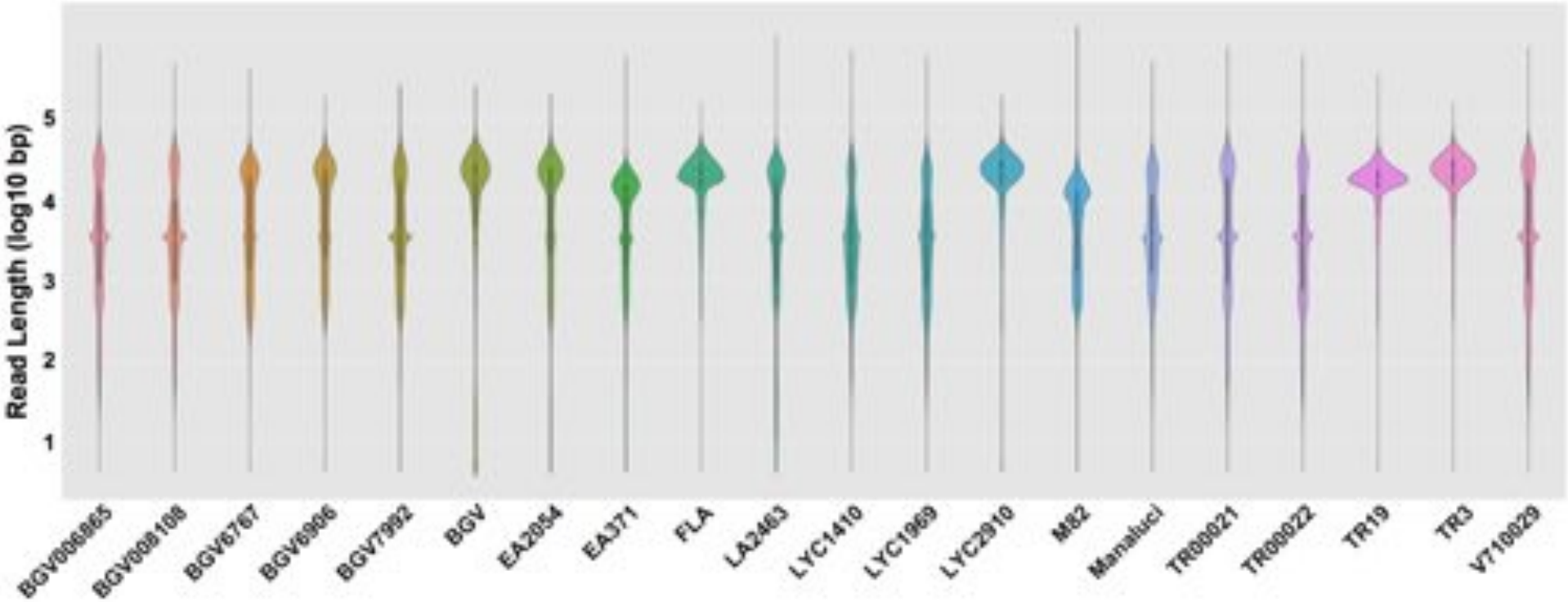
Nanopore Read Lengths

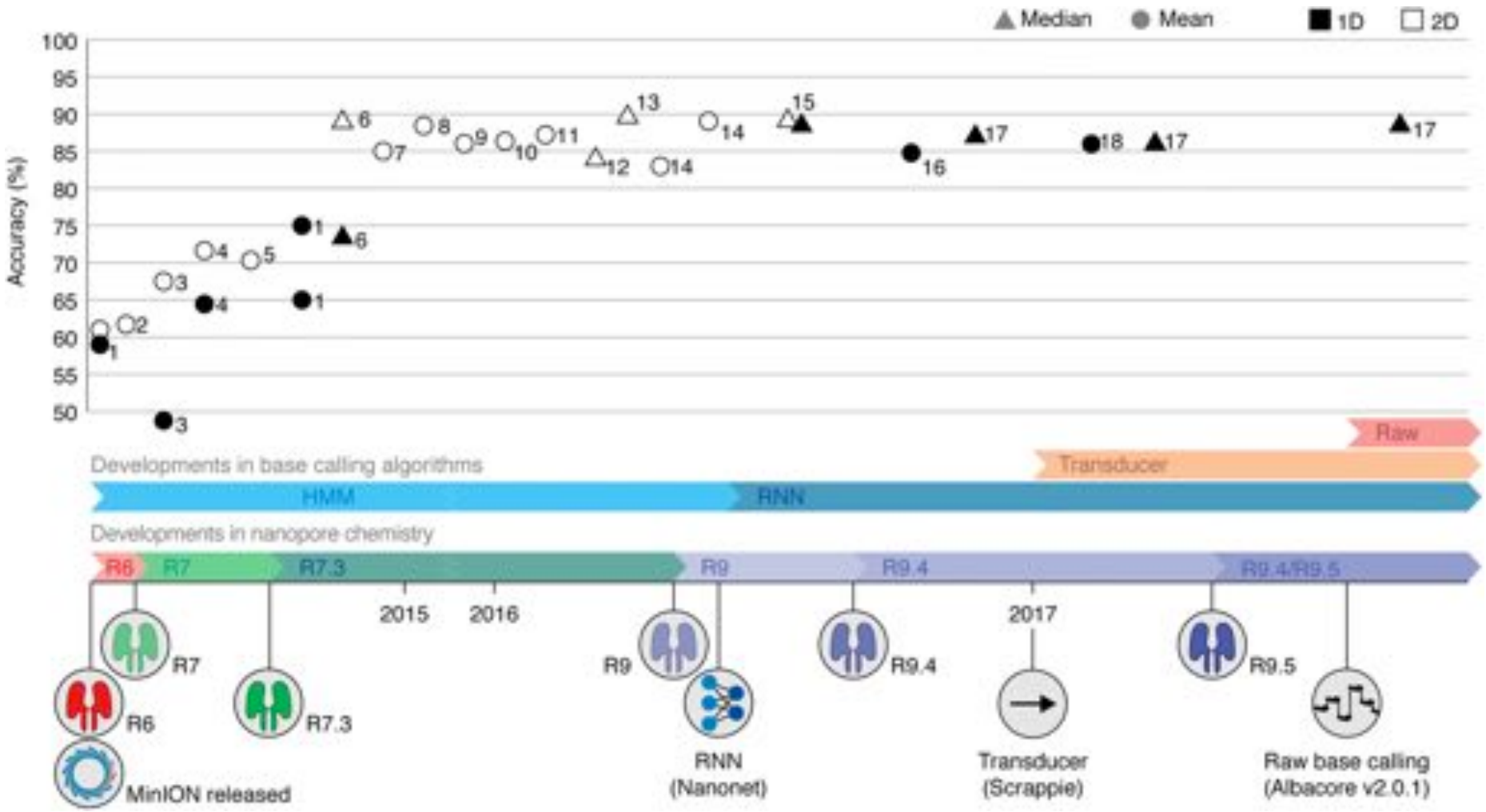
Optimized Sequencing strategy

- Fragmentation at 30kbp using the Megarupter
- 109 Ligation Sequencing Kit yields both long reads and high yield

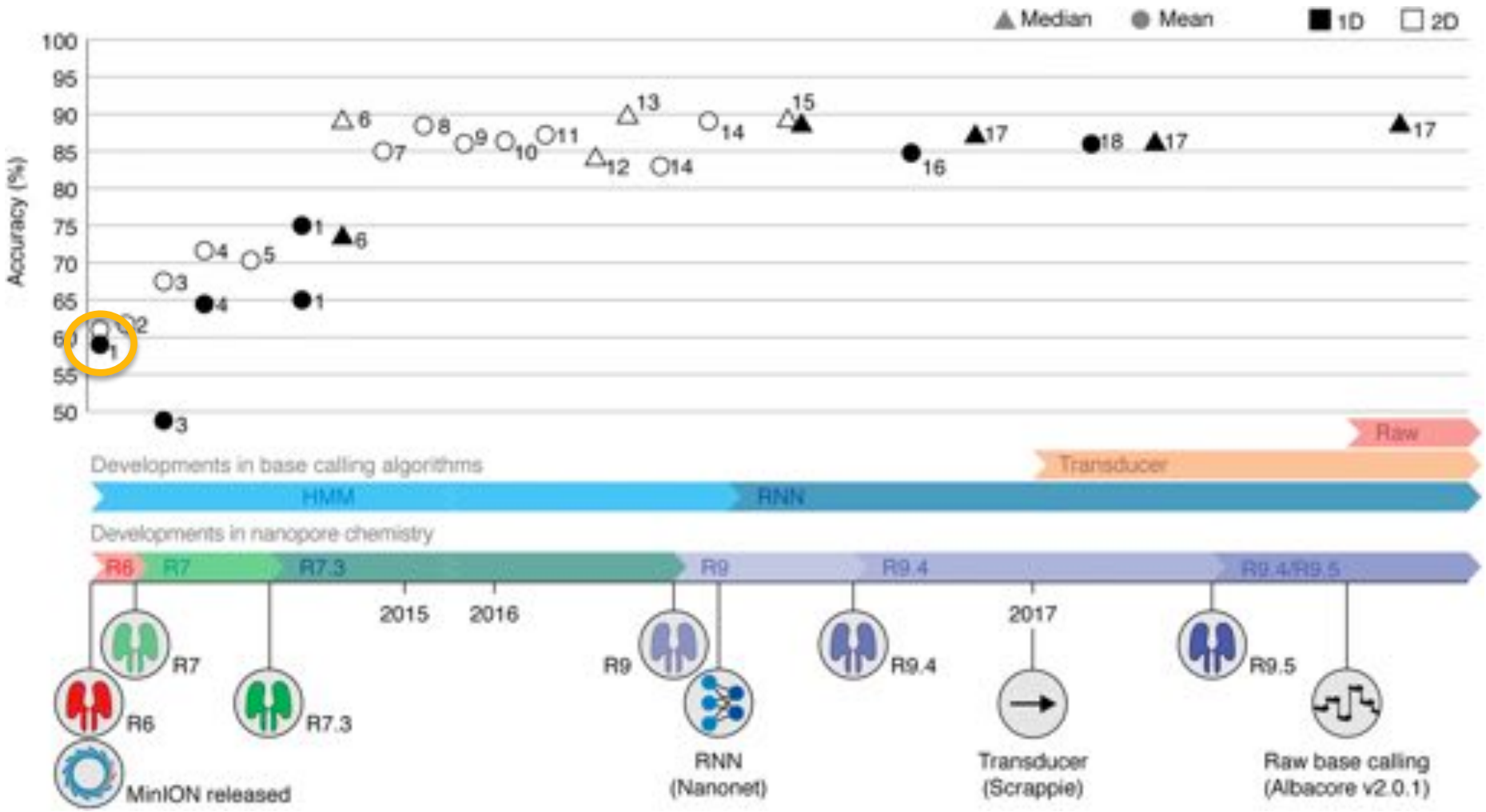
Very long reads with PromethION

- Read length N50: 25kbp – 30kbp
- Over 20x coverage of reads over 20kbp
- Even better results using the Circulomics SRX

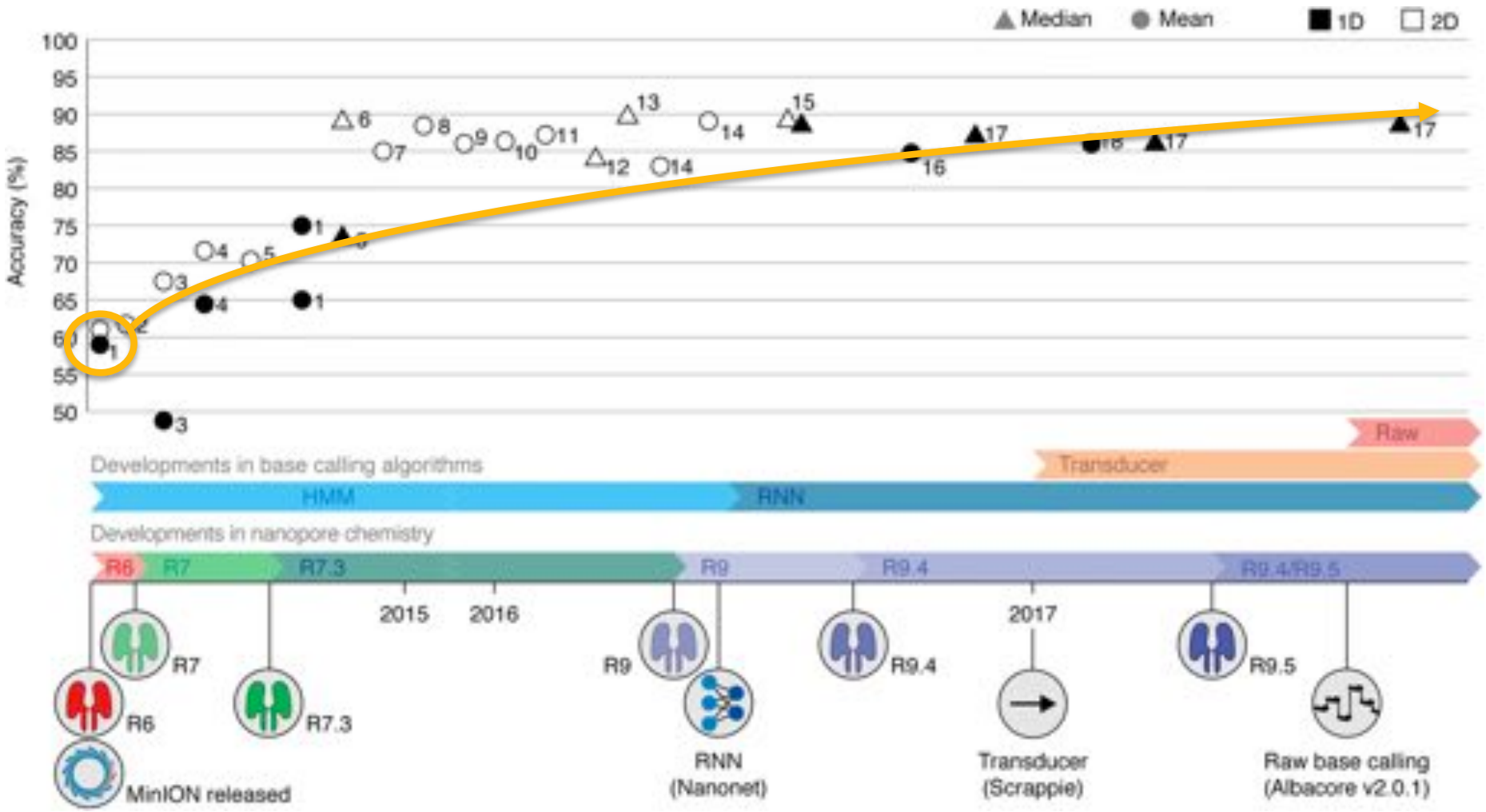




From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology.* https://doi.org/10.1186/s13059-018-1462-9



From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology*. https://doi.org/10.1186/s13059-018-1462-9



From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology*. https://doi.org/10.1186/s13059-018-1462-9

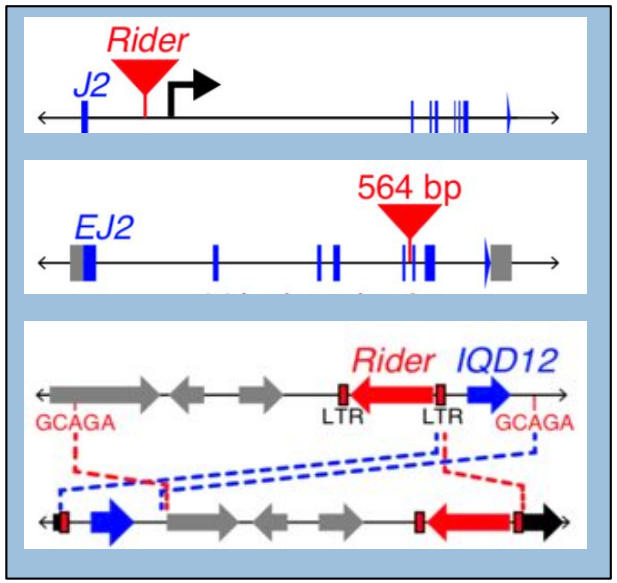
Data Management

High throughput of PromethION has introduced some new IT challenges

- Upgraded the fiber connection between the sequencing lab and the data center
- Substantial storage requirements
- Substantial load on filesystem to manage hundreds of millions of fast5 files

~ Part 3 ~

Structural Variation Identification

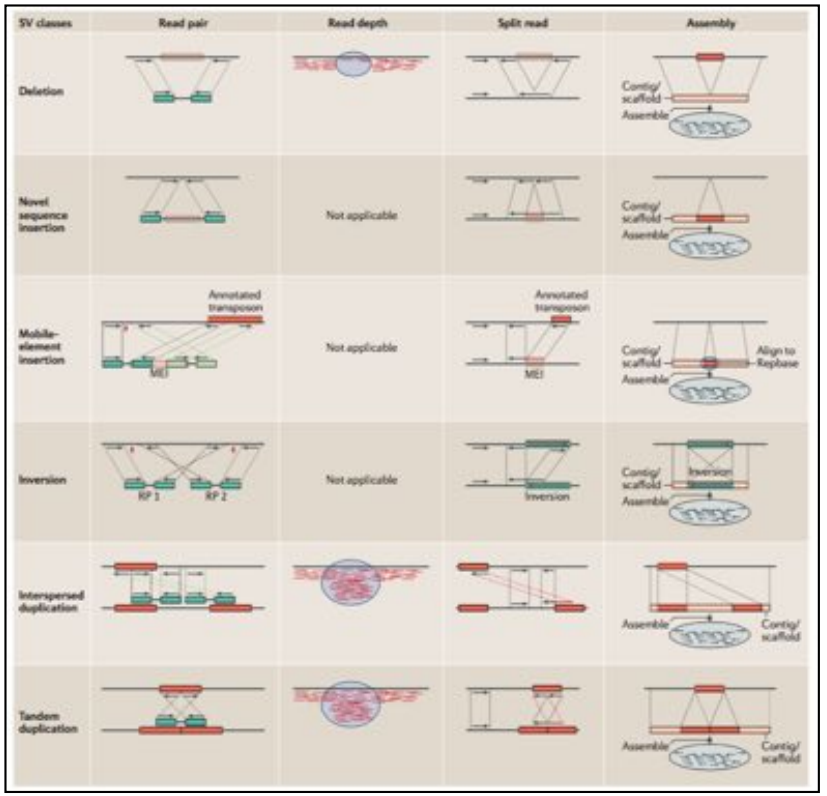


Structural Variation Identification

Two major strategies for detection

- ***Alignment-based detection***
  - Split-read alignment to detect the breakpoints of events
  - Fast, accurately identifies most variants, including heterozygous variants
  - Very long insertions may be incomplete

- ***Assembly-based detection***
  - De novo assembly followed by whole genome alignment
  - Can capture novel sequences and other complex variants
  - Slow, demanding analysis, limited by contig length, heterozygous variants challenging



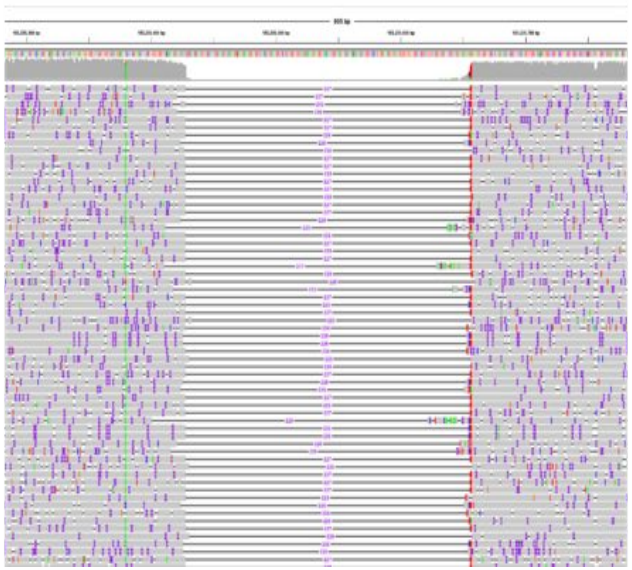
Genome structural variation discovery and genotyping

Alkan, C, Coe, BP, Eichler, EE (2011) *Nature Reviews Genetics.* May;12(5):363-76. doi: 10.1038/nrg2958.

Alignment Based Analysis

BWA-MEM

NGMLR

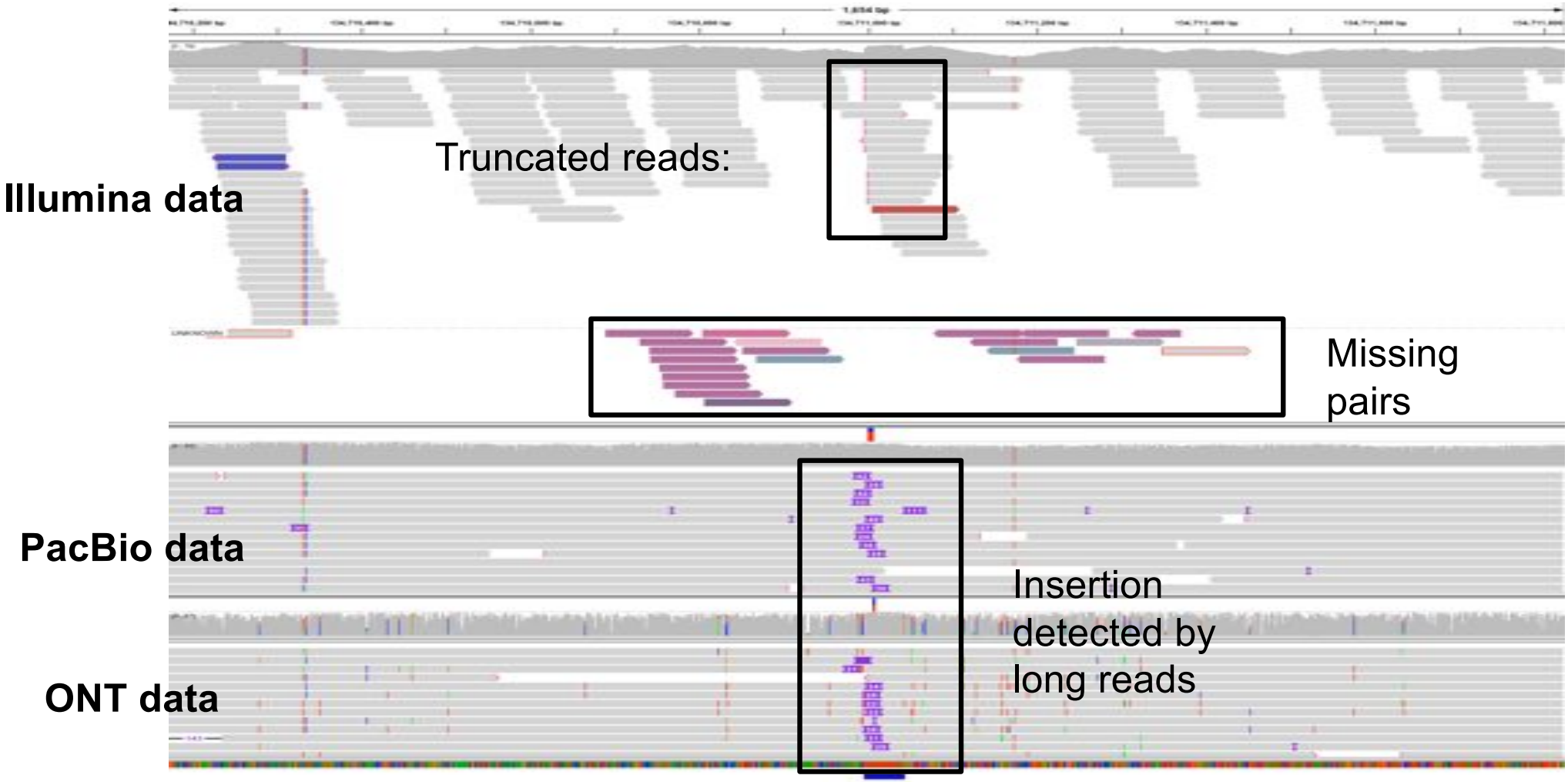


NGMLR: Dual mode scoring to accommodate indel errors plus SVs
CrossStitch: Local re-assembly across variants to improve breakpoints

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder, *et al* (2018) *Nature Methods*. doi:10.1038/s41592-018-0001-7

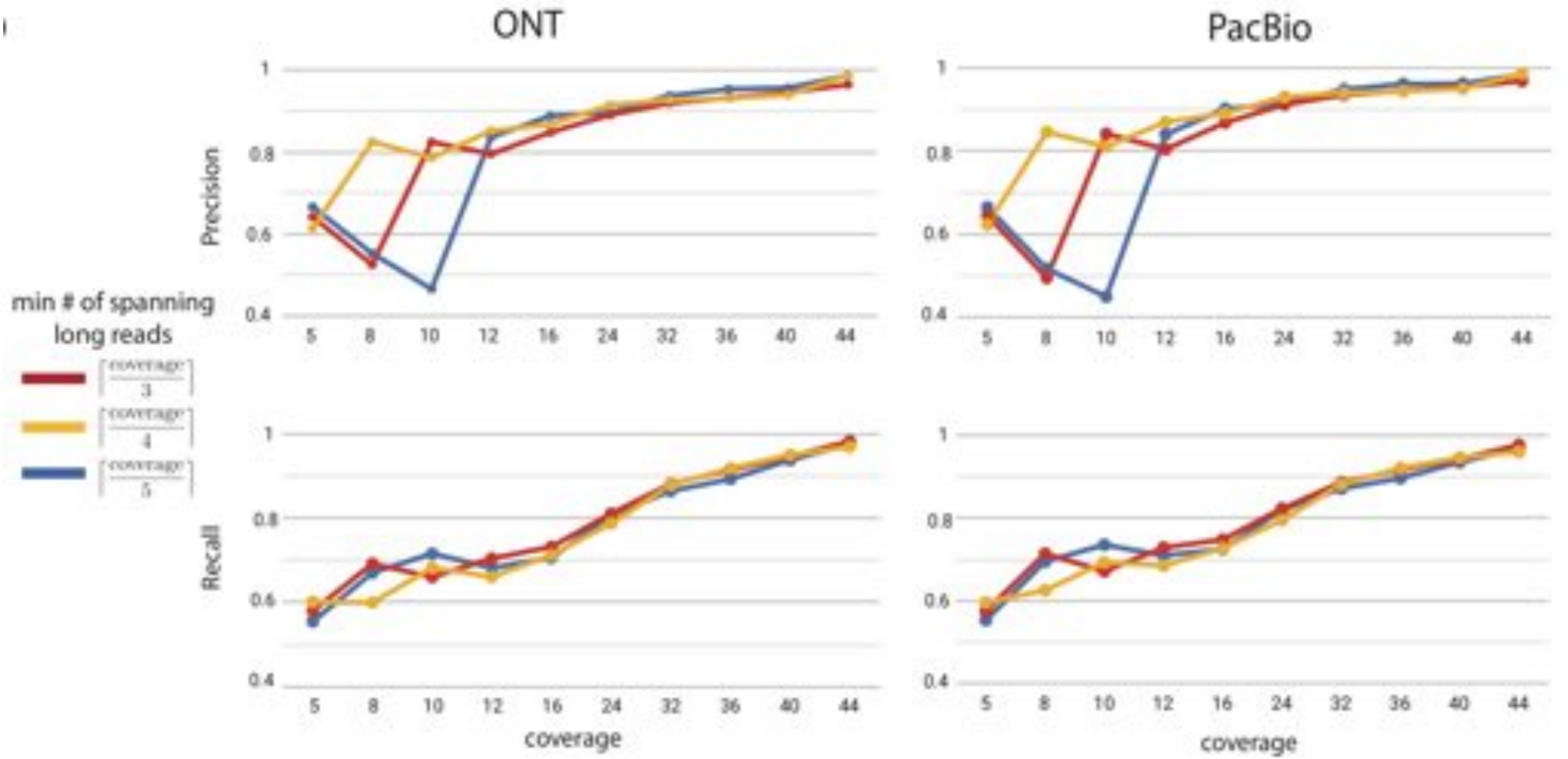
Illumina data

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

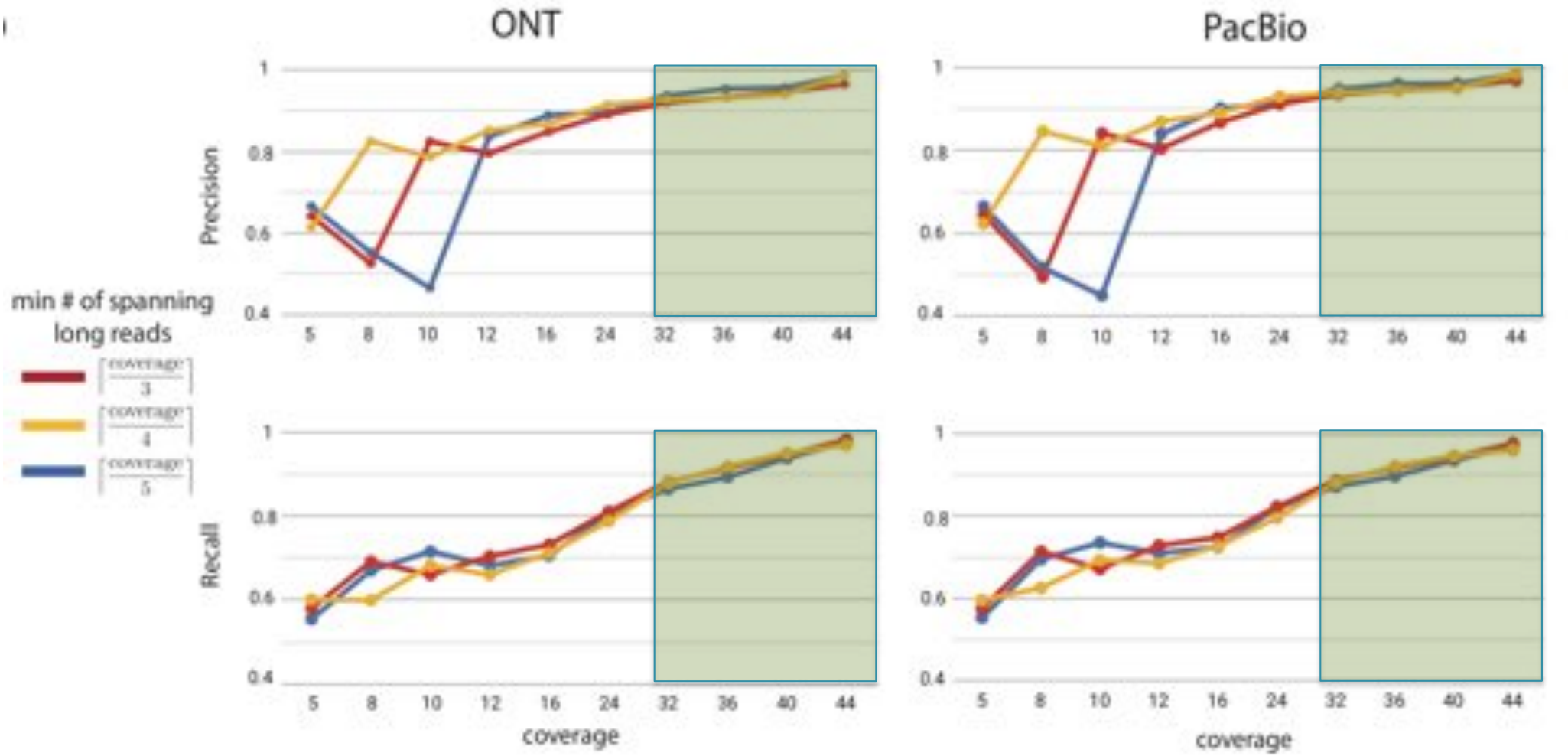


Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2017) *bioRxiv https://doi.org/10.1101/169557*

Coverage Requirements



Coverage Requirements

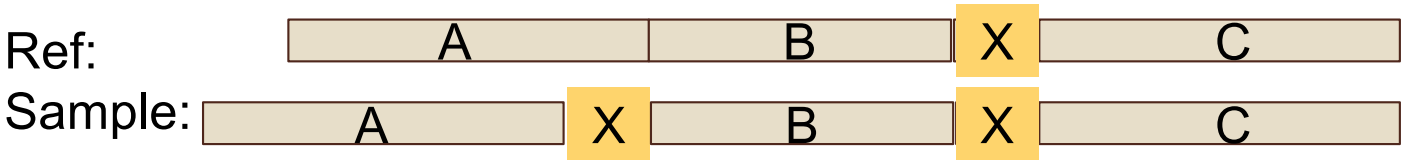


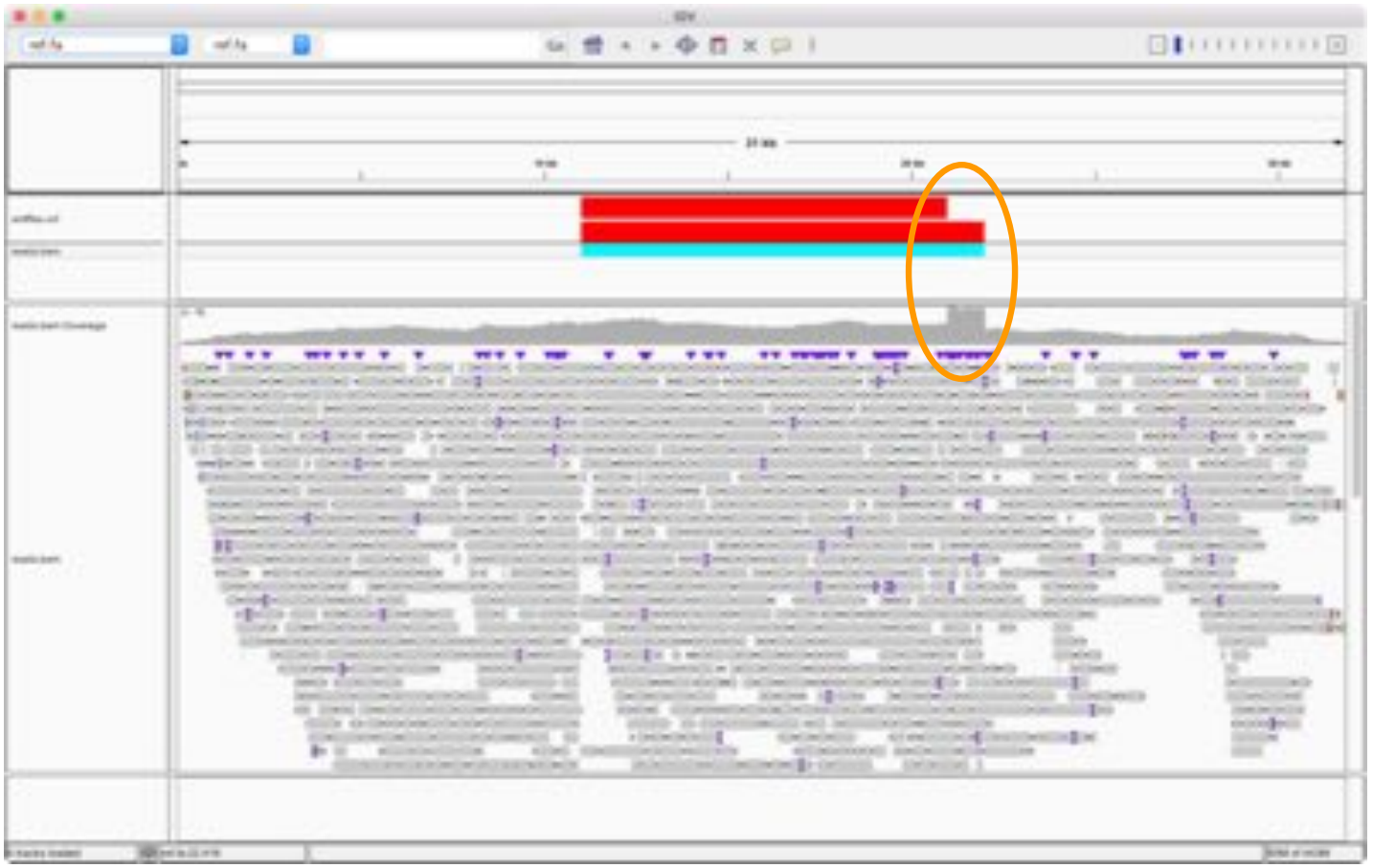
VCF Woes

TE Insertions are poorly represented in VCF

- In the syntax of VCF, TE insertions are represented as 2 separate breakpoints:
  - A “deletion” from A to X
  - A “duplication” of X jumping back to B

Be careful of how you interpret overlapping variants





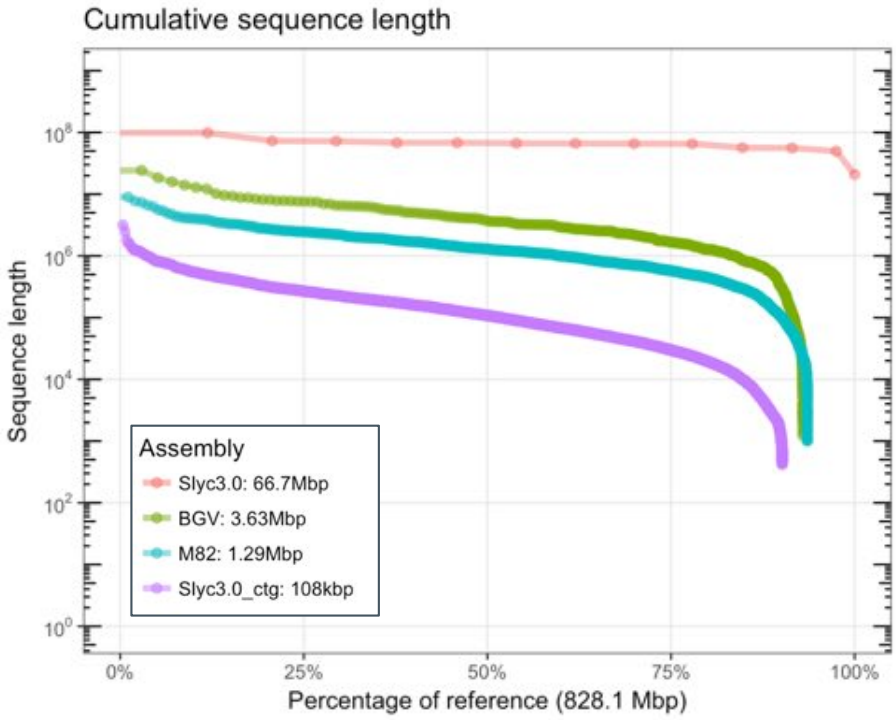
De novo Assembly

Gold level assemblies with Canu

- Well-established, integrated correction & assembly
- Contig N50 sizes >10-fold better than reference
- Main challenge is speed
  - ~2 weeks per assembly on ~320 cores

Exploring other options

- Wtdbg2 (https://github.com/ruanjue/wtdbg2) runs in ~8 core hours (+1.5 days for consensus) although mixed results depending on sample
- MaSuRCA (Zimin et al, Bioinformatics, 2013) hybrid assembler produces high quality consensus
- Discussing cloud-enabled pipelines with DNAnexus



Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation
Koren *et al* (2018) *Genome Research. doi: 10.1101/gr.215087.116*

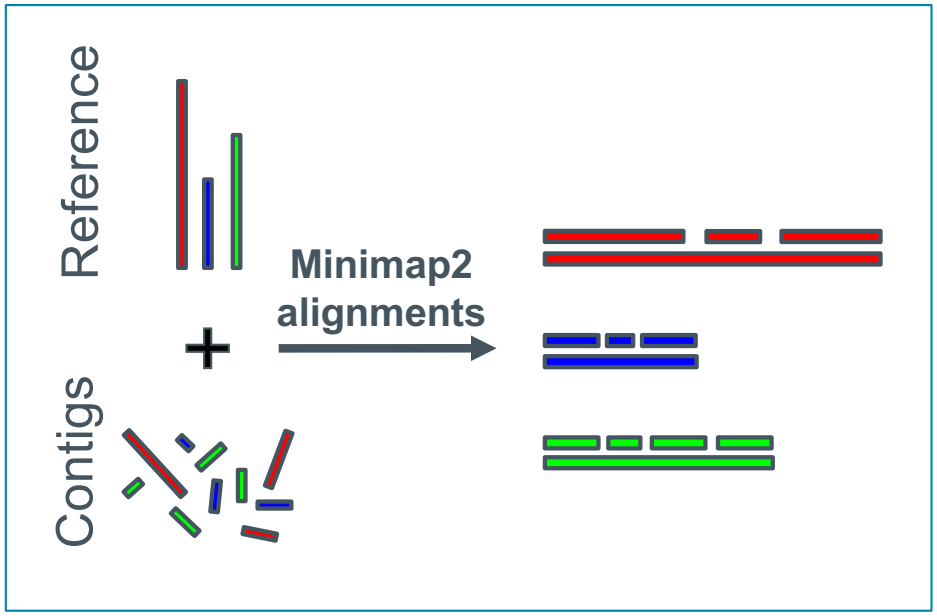
RaGOO: Fast and accurate reference-guided scaffolding

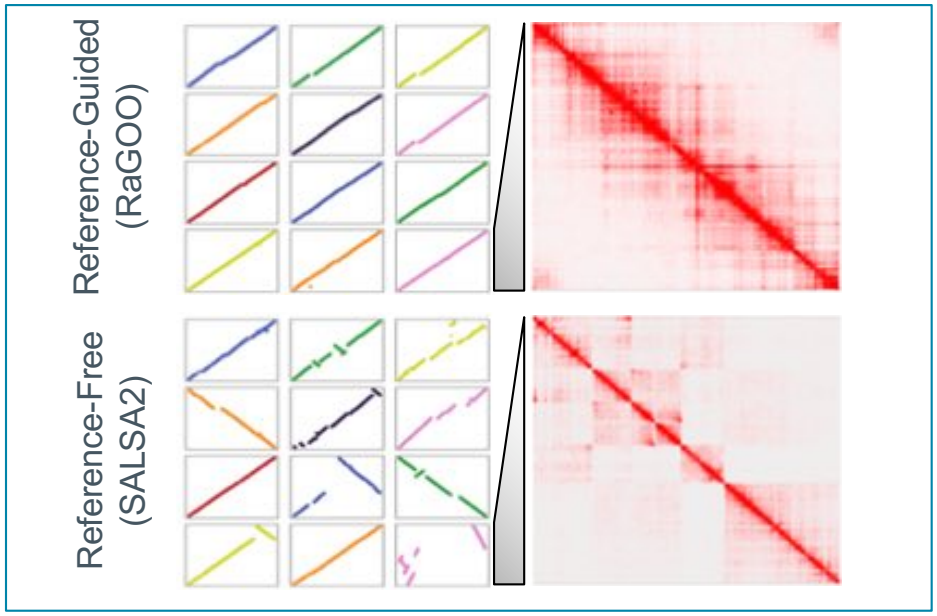
Reference guided scaffolding

- Use the reference genome as a “genetic map”
- Effective when sample is structurally similar to reference

Validation using Hi-C

- Reference-guided scaffolding leads to more complete and more accurate chromosomes





RaGOO: Fast and accurate reference-guided scaffolding of draft genomes
Alonge *et al* (2019) *Genome Biology*. *doi:10.1186/s13059-019-1829-6*

Sol Genomics Network

solgenomics.net/projects/tomato13/

Sol Genomics Network | Search | Maps | Genomes | Projects | Tools | About | Login

Chromosome-scale assembly and annotation of 13 diverse tomato accessions

Project overview

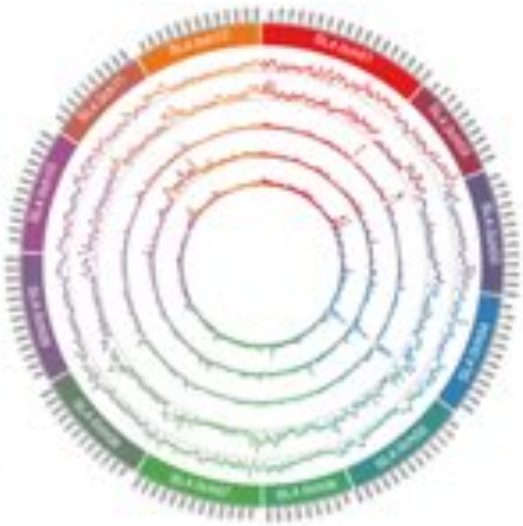
Michael Alonge, Srividya Ramakrishnan, Sebastian Soyk, Xingang Wang, Matthias Benoit, Zachary B. Lippman, Michael C. Schatz

The research groups of Michael C. Schatz and Zachary B. Lippman at Johns Hopkins University and Cold Spring Harbor Laboratory, respectively, have generated genome assemblies and associated gene annotations for 13 diverse tomato accessions. These assemblies and annotations, each with their own independent versioning, are beta pre-releases to the community.

This large dataset is being made available for research under the "Toronto statement", which outlines rules for pre-publication data sharing, under which we, the authors, reserve the right to publish the first analyses of the data, which includes descriptions of whole chromosome or genome-level analyses of genes, variants, gene families, repetitive elements, and comparisons with other organisms.

The following accessions have been assembled and annotated and are included in this release.

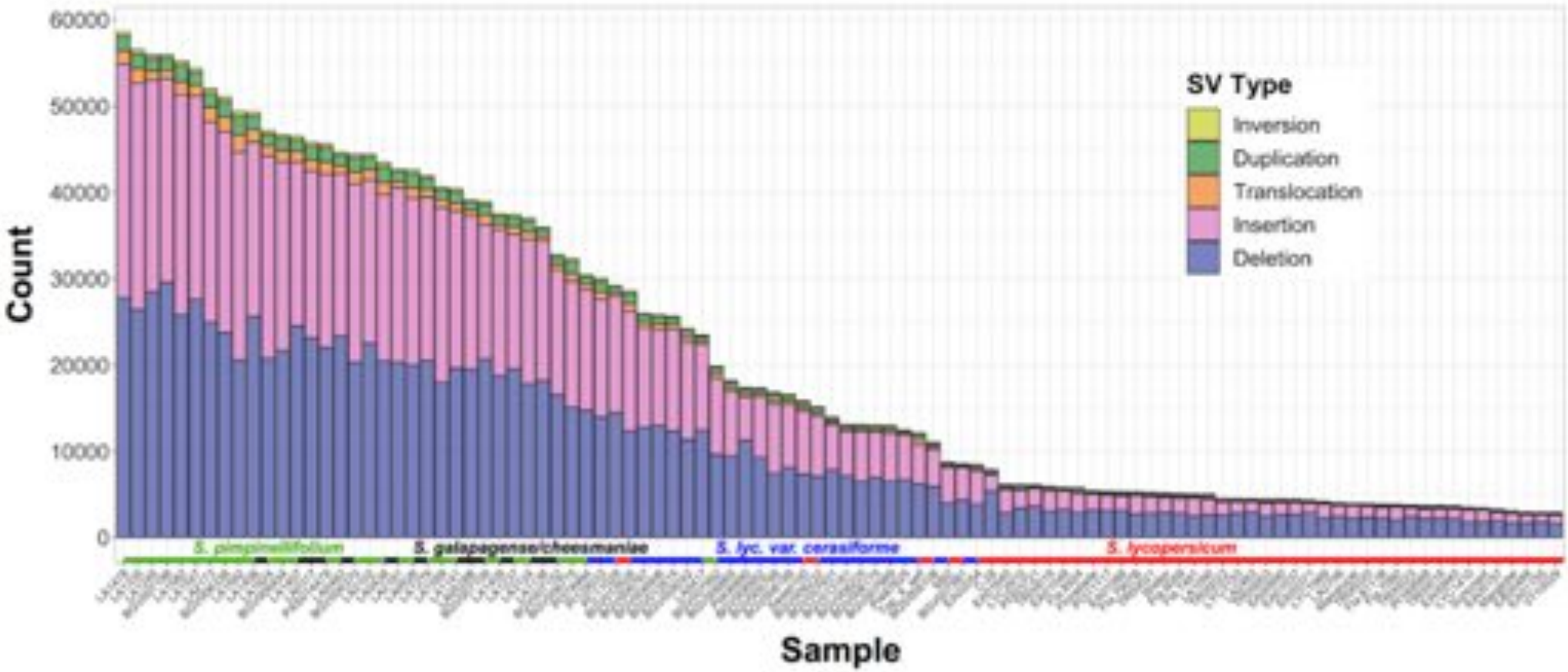
- Brandywine
- M82
- Floradade
- EA00371
- EA00990
- PI_S014479
- BGV006775
- BGV006865
- BGV007989
- BGV007931
- PI303721
- PI169588
- LYC1410



~ Part 4 ~

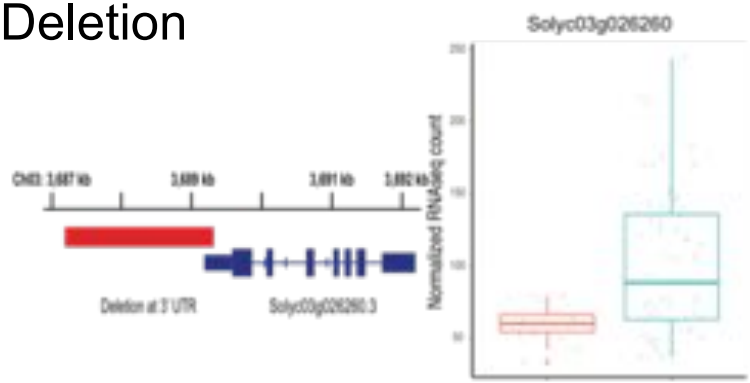
The Landscape of Structural Variation in Tomato Genomes

The landscape of structural variations



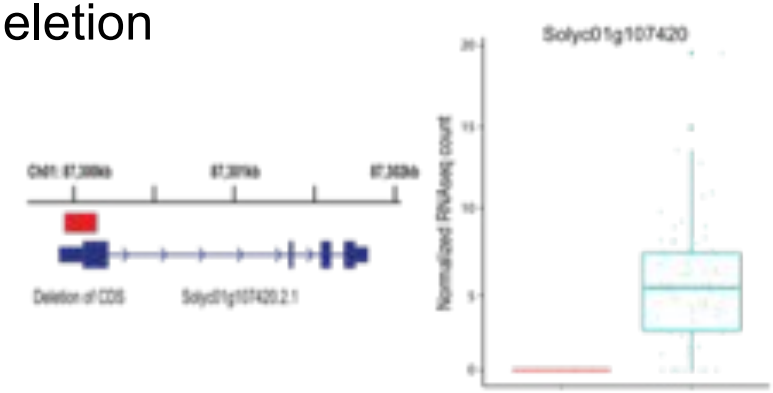
The Impact of SVs on Expression

3’ UTR Deletion



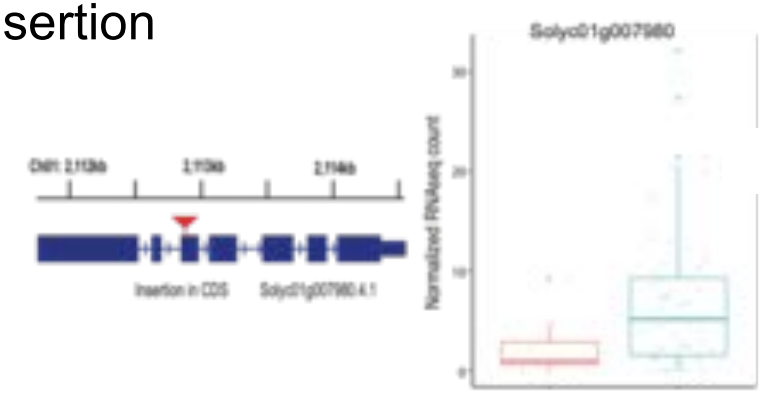
with SV(n=5) without SV(n=16)

CDS Deletion



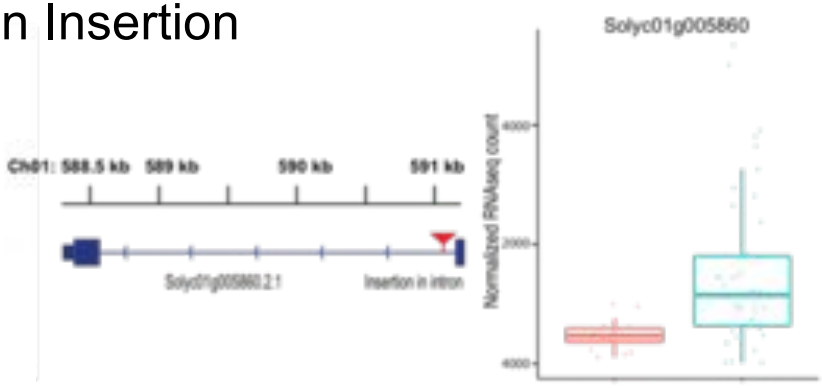
with SV(n=2) without SV(n=19)

CDS Insertion



with SV(n=2) without SV(n=19)

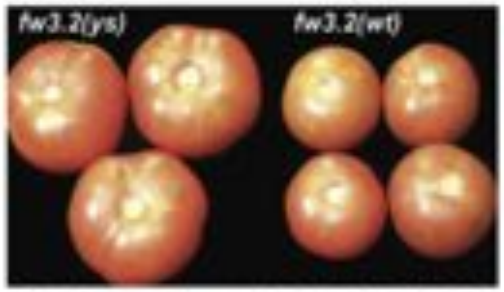
Intron Insertion



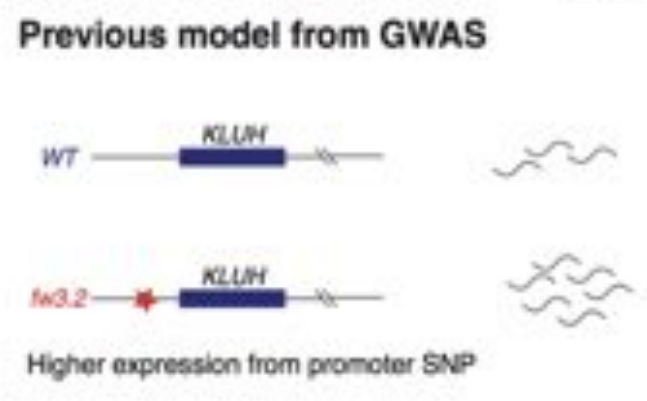
with SV(n=6) without SV(n=15)

Genetic Engineering

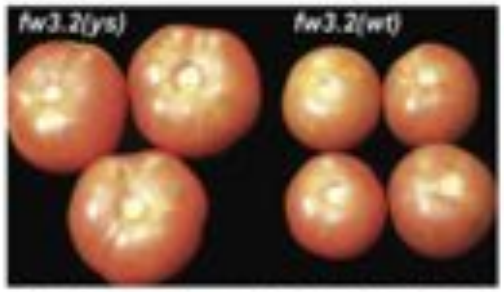
Uncovering mechanisms of increased fruit weight



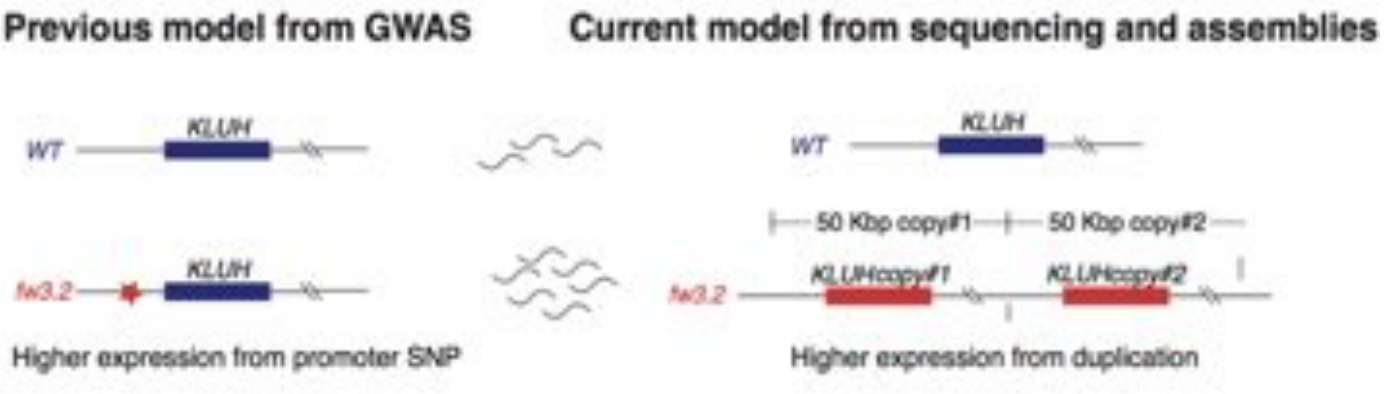
Chakrabarti, et al. 2013, PNAS



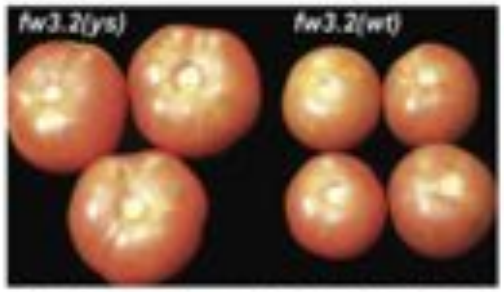
Uncovering mechanisms of increased fruit weight



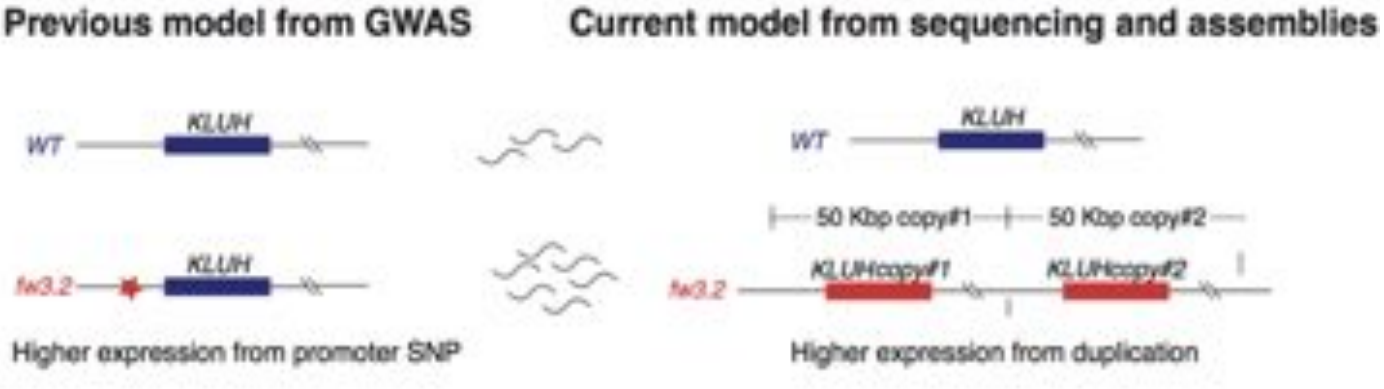
Chakrabarti, et al. 2013, PNAS

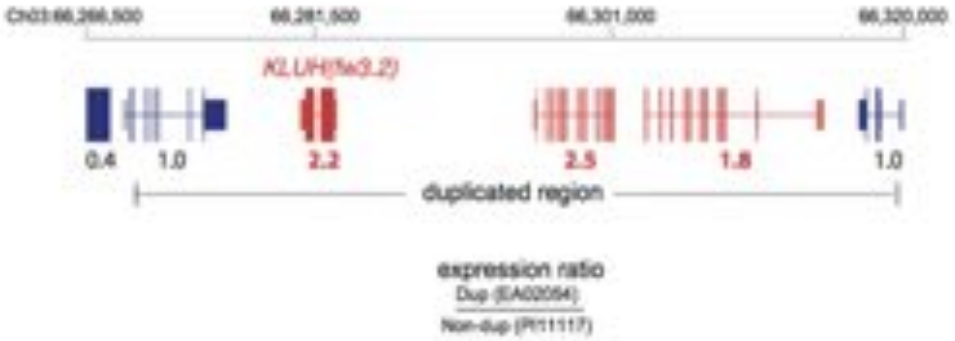


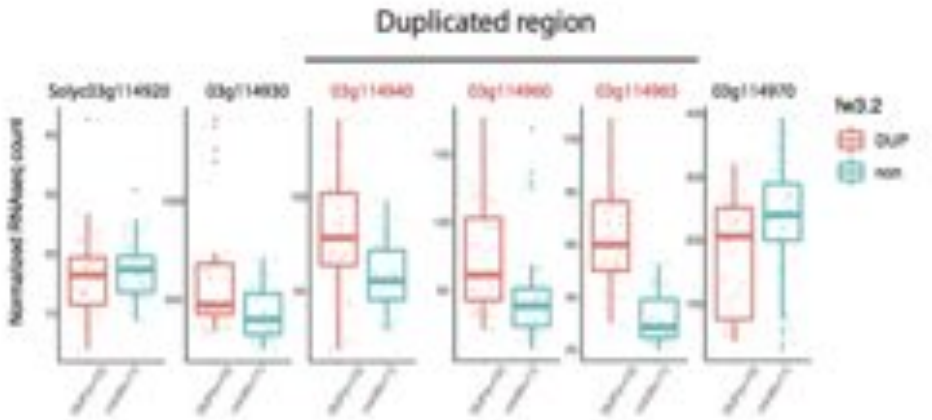
Uncovering mechanisms of increased fruit weight



Chakrabarti, et al. 2013, PNAS





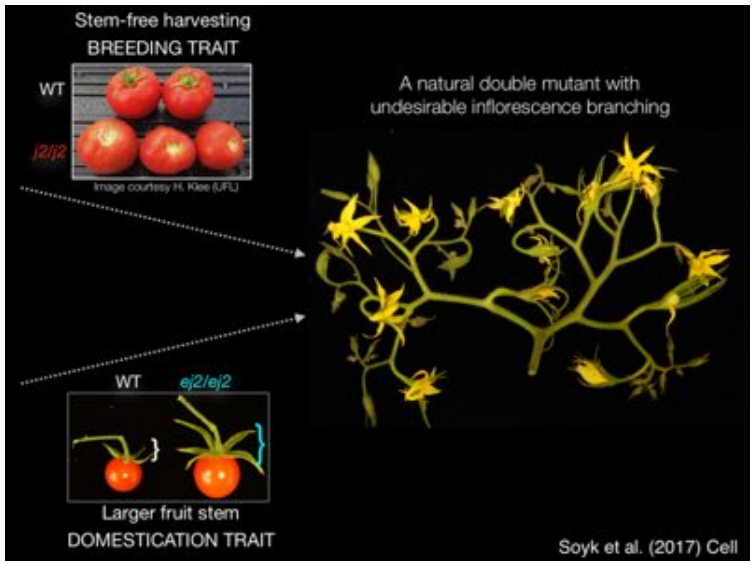


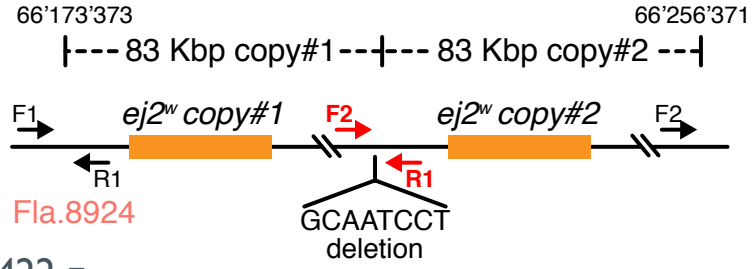
Normalized gene expression from 23 tomato accessions

Identification of the ej2 Tandem Duplication

Validation of our first SV association

- Crosses of tomato plants with a highly desirable breeding trait (j2: jointless2) and a desirable domestication trait (ej2: an enhancer for j2) are typically poorly producing plants – ***a negative epistatic interaction***
- However some breeding lines carry both alleles and yet have good yields through unknown means
  - One of our first samples was such a breeding line and revealed a 83kbp tandem duplication spanning ej2
  - Validated the duplication using Sanger, RNA-seq and quantitative genetics to conclude the duplication of the locus causes stabilization of branching and flower production
- **Now able to use CRISPR/cas9 to overcome the negative epistatic interaction to improve fruit yields**





Soyk *et al* (2019) *Nature Plants*. https://doi.org/10.1038/s41477-019-0422-z

What's next?

Rapid Domestication of Wild Species

http://technical.ly/baltimore/2016/11/01/johns-hopkins-genome-algorithm-wine/

Summary & Future Work

High throughput long read sequencing is unlocking the universe of structural variations

- Discovering tens of thousands of variants previously missed, as well as clarifying tens of thousands of false positives per sample
- Possible to rapidly characterize pan-genomes with >100 samples
- Throughput & accuracy rapidly improving, realtime direct alignment of nanopore signal data

Beyond mere structural variation identification

..... towards "Rules of Life" interaction maps and beyond

- Identify the specific pathways for many important traits
- Discovery and dissection of cis-regulatory epistasis
- Analysis of epigenetic modifications
- Engineering domestication traits in "wild" plants

Expect to see similar results in all other plant and animal species

Acknowledgements

Schatz Lab

Mike Alonge

Srividya Ramakrishnan

Sergey Aganezov

Charlotte Darby

Arun Das

Katie Jenike

Melanie Kirsche

Sam Kovaka

Bohan Ni

T. Rhyker Ranallo-Benavide

Rachel Sherman

Samantha Zarate

Your Name Here

Lippman Lab

Sebastian Soyk

Xingang Wang

Zachary Lemmon

Cold Spring Harbor Laboratory

Sara Goodwin

W. Richard McCombie

Baylor College of Medicine

Fritz Sedlazeck

Boyce Thompson

Joyce Van Eck

University of Georgia

Esther van der Knaap