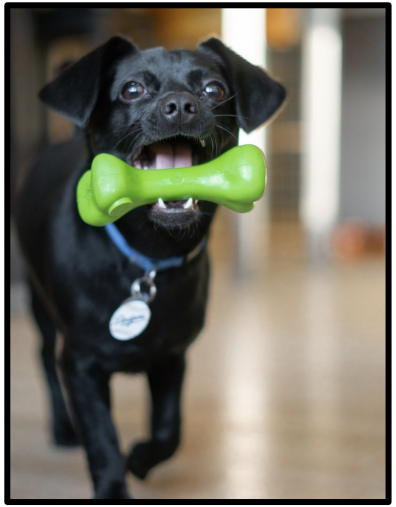
Variation of Genome Structure

Michael Alonge

Applied Comparative Genomics: EN.601.749

2-24-20


THIS IS BIRDLAND



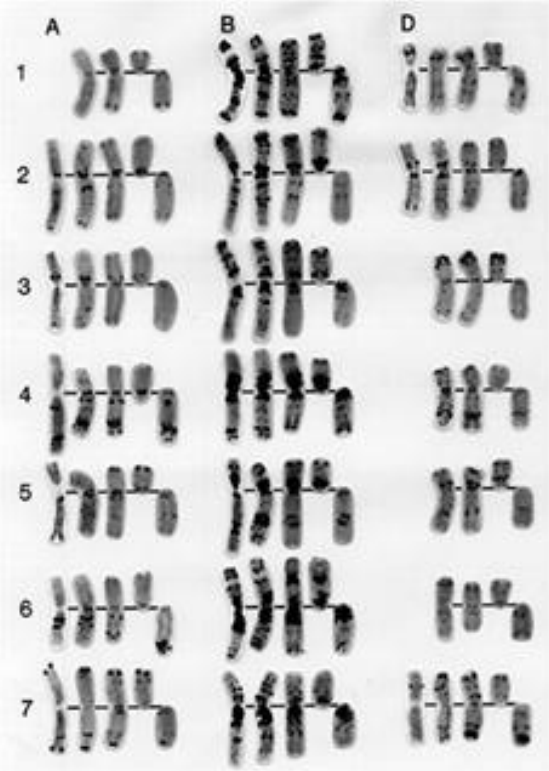
Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

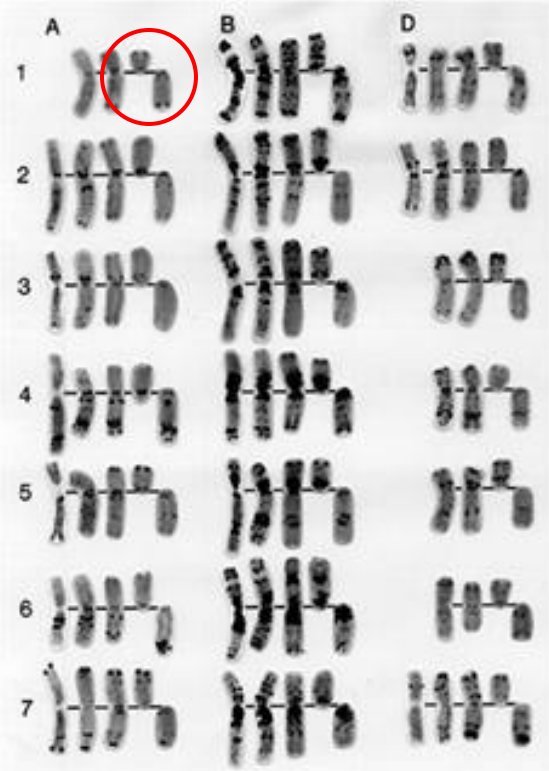
What is Genome Structure?



Bread wheat (*Triticum aestivum*)

https://ebrary.net/26390/environment/laying_foundation_cytogenetic_genome_mapping_wheat_aneuploid_stocks

What is Genome Structure?



Bread wheat (*Triticum aestivum*)

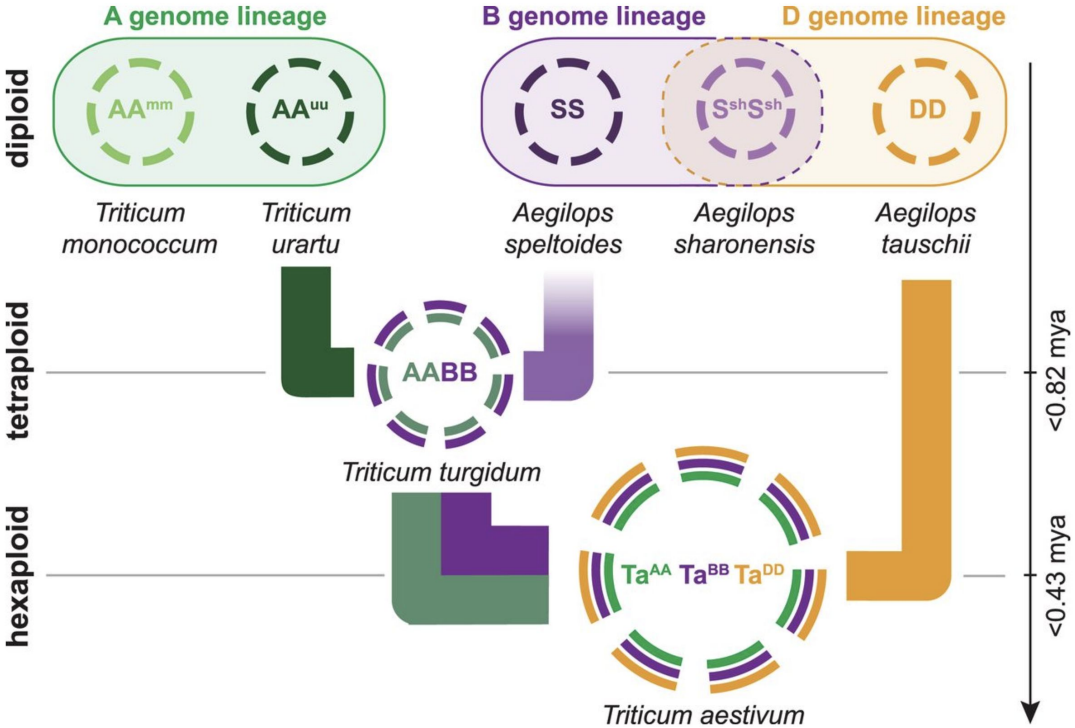
https://ebrary.net/26390/environment/laying_foundation_cytogenetic_genome_mapping_wheat_aneuploid_stocks

What is Genome Structure?

- **My definition**: Genome structure refers to large-scale genomic sequence composition.

- **Example**:
  - A genome assembly contig that accurately represents a whole chromosome would be considered “structurally accurate”.
  - If that contig was missing a large chunk of the chromosome, it would be considered “structurally inaccurate”.

* Sometimes, genome “structure” refers to 3D organization/characteristics of the genome. That is **not** what we will be discussing today.



Plant genomes have dynamic structure:

- Polyploidization
- TE activity
- Gene loss/duplication

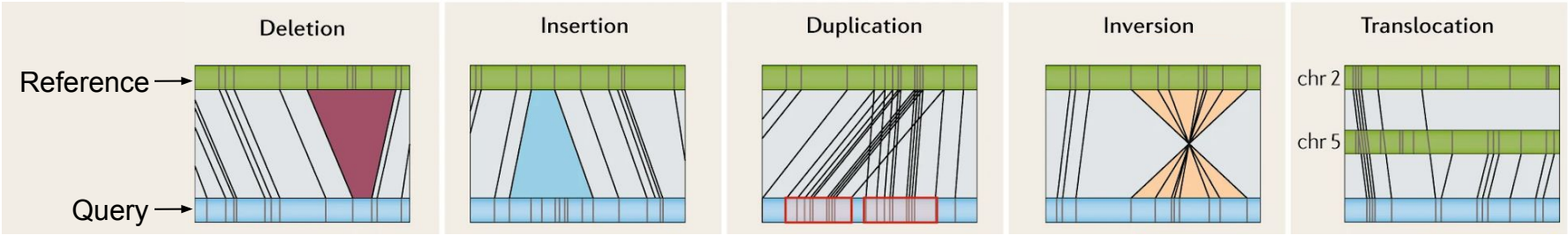
International Wheat Genome Sequencing Consortium. "A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome." *Science* 345.6194 (2014): 1251788.

What is Structural Variation?

- **My definition**: Variation of genome structure in a population.
- A **Structural Variant** (SV) is a particular variable locus.

- **Examples**:
  - Mike has two copies of the *DODGERSFAN* gene, while Bob only has 1.
  - “Brandywine” (an heirloom tomato variety) has a rare transposable element insertion in the *FLAVOR* gene.

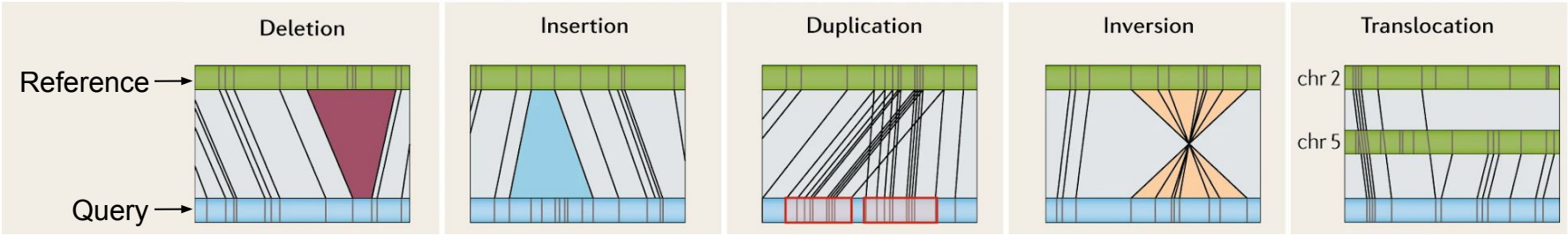
How is Structural Variation Classified?



- Just as with small variant calling, we typically classify structural variation by comparing two individuals to each other.
  - There are many ways to do this “comparison” which will be covered later.
- One individual is designated as the “reference” and the other the “query”.
- We then define query structural variants “with respect to” the reference
  - Mike has an insertion with respect to the reference
  - Bob has a deletion with respect to the reference
- SVs are usually defined as being longer than 50 bp.

Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." *Nature Reviews Genetics* (2019): 1-19.

How is Structural Variation Classified?



Deletion	Insertion	Duplication	Inversion	Translocation
A reference locus is not found in the query	The query locus is not found in the reference	The reference has 1 copy of a locus and the query has >1 copy of that locus.	Loci in the reference and query are reverse complements of one another.	A locus in the query moved to a different part of the genome

Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." *Nature Reviews Genetics* (2019): 1-19.

How is Structural Variation Classified?

Copy Number Variants (CNVs)

- A distinct but related SV classification
- Refers to variation in the copy number of a locus
- Example: The *CUTE* gene is copy number variable in dogs
  - The reference genome has 1 copy of the *CUTE* gene.
  - Rover has 0 copies of the *CUTE* gene (A.K.A a “deletion”).
  - Baxter has 2 copies of the *CUTE* gene (A.K.A a “duplication”).
  - Tupper has 15 copies of the *CUTE* gene.

How Is Structural Variation Created?

- **Faulty repair of DNA damage**
- Transposable element activity
- Non-disjunction
- DNA replication errors
- Unequal crossing-over

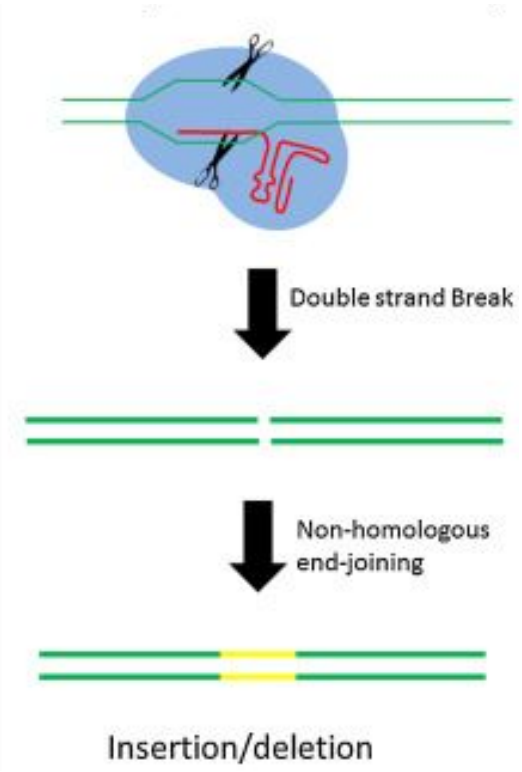
* SVs are usually created or mediated by repeats

How Is Structural Variation Created?

- **Faulty repair of DNA damage**
- Transposable element activity
- Non-disjunction
- DNA replication errors
- Unequal crossing-over

* SVs are usually created or mediated by repeats

CRISPR/Cas9 Gene Editing



Schaeffer, Scott M., and Paul A. Nakata. "CRISPR/Cas9-mediated genome editing and gene replacement in plants: transitioning from lab to field." *Plant Science* 240 (2015): 130-142.

Outline

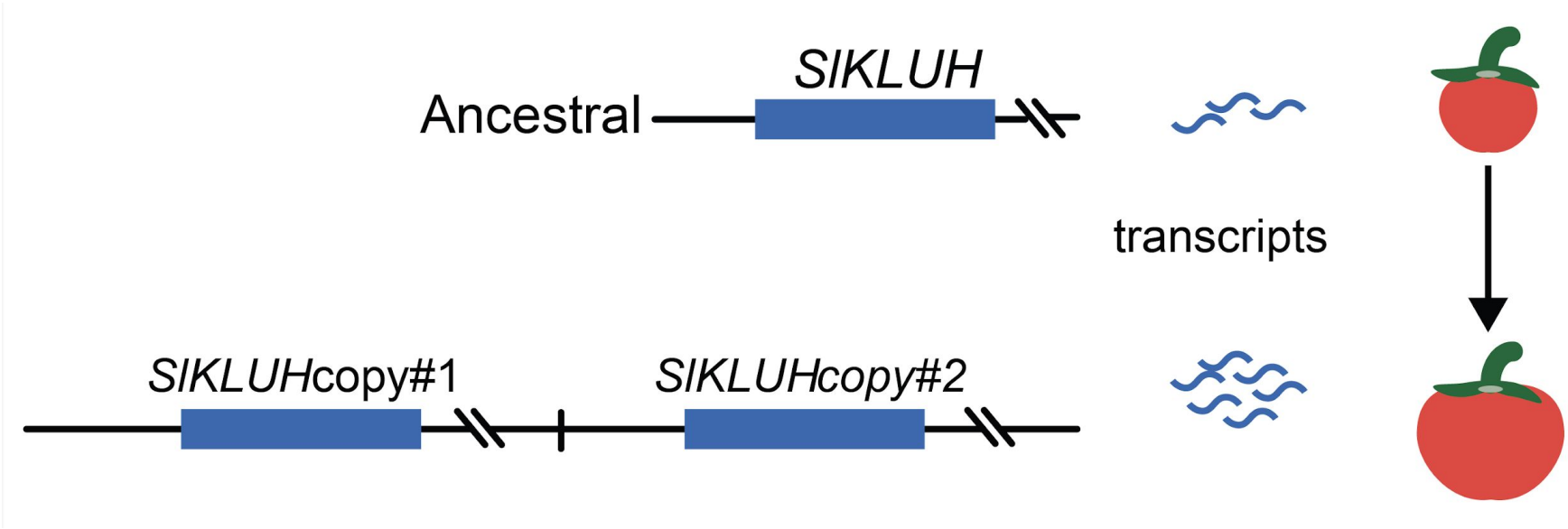
- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Outline

- Introduction to genome “structure”
- **Functional importance of genome structure**
- The Bioinformatics of SV calling
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- Applications in Tomato

SVs Impact Function

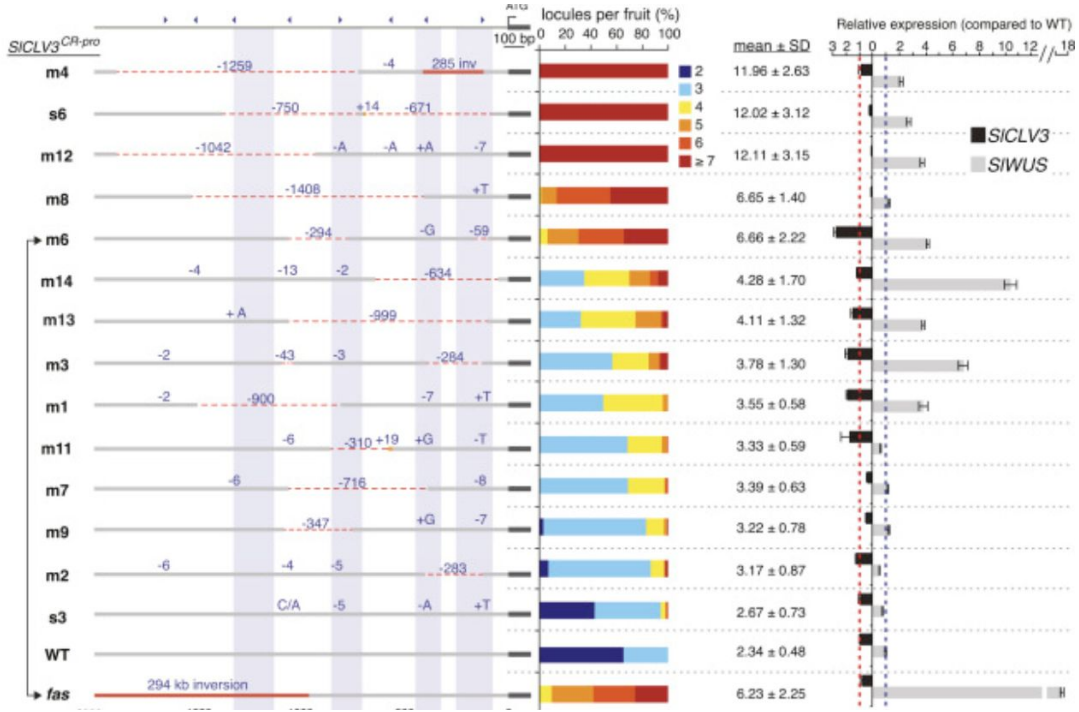
1. Protein coding genes



Alonge and Wang *et al*., unpublished

SVs Impact Function

2. Gene regulatory elements

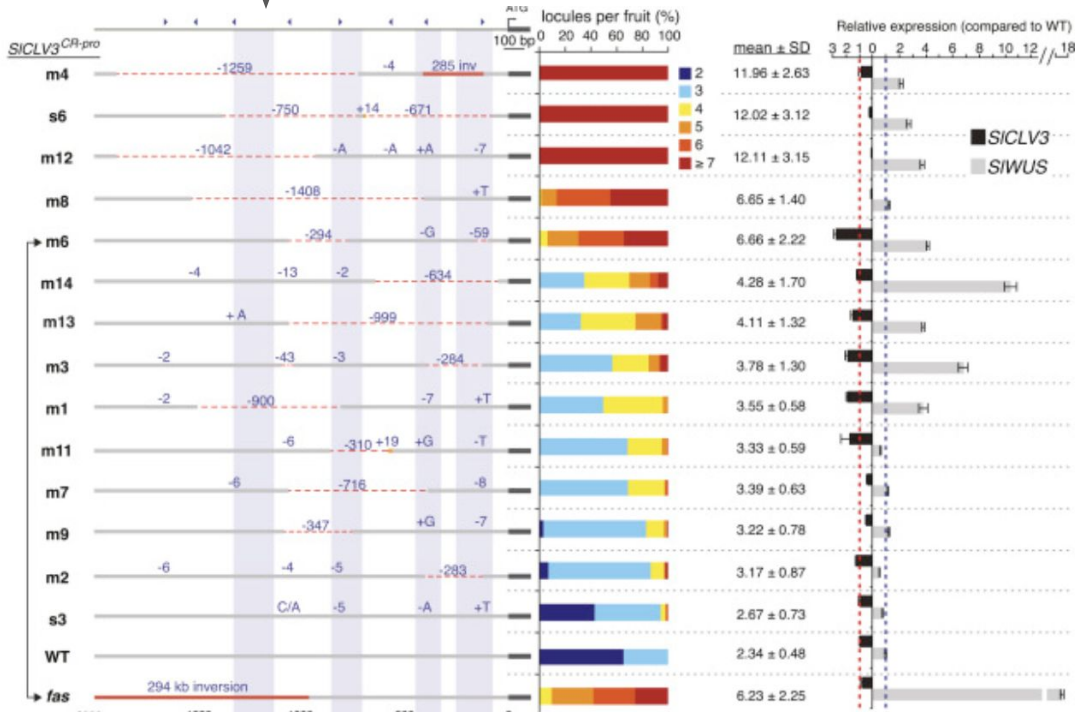


Rodríguez-Leal, Daniel, et al. "Engineering quantitative trait variation for crop improvement by genome editing." *Cell* 171.2 (2017): 470-480.

SVs Impact Function

2. Gene regulatory elements

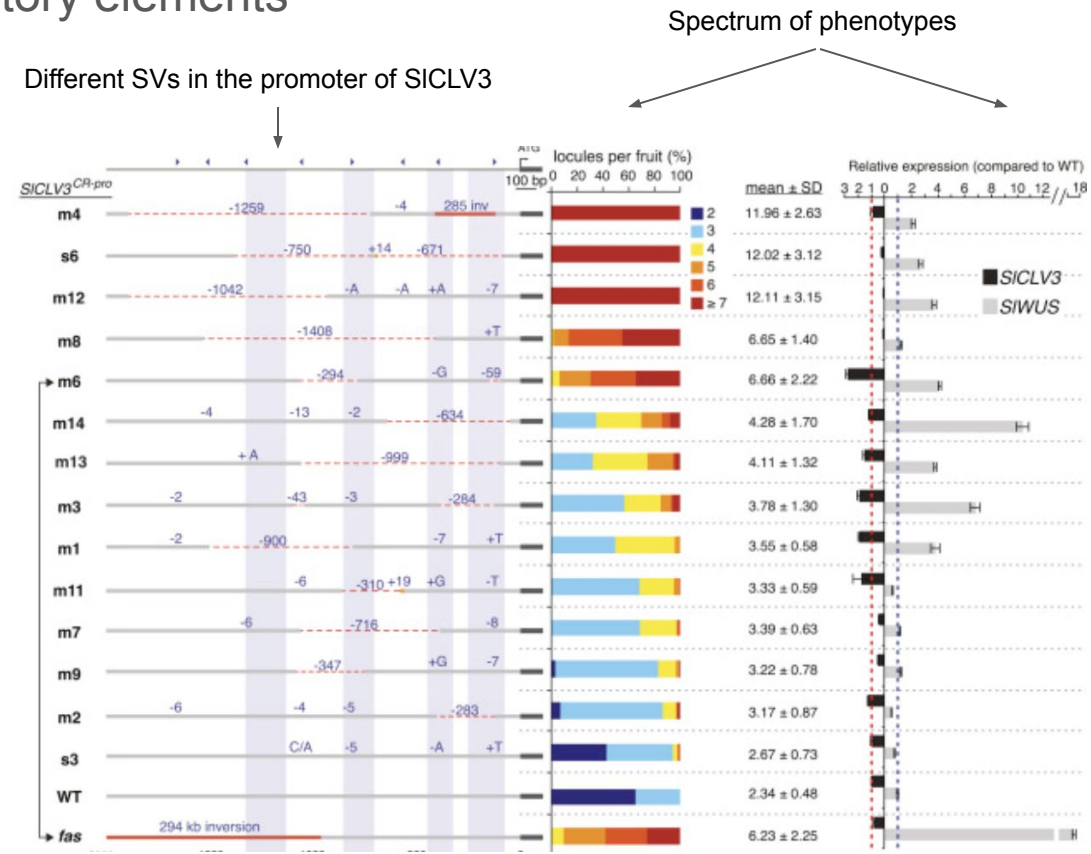
Different SVs in the promoter of SlCLV3



Rodríguez-Leal, Daniel, et al. "Engineering quantitative trait variation for crop improvement by genome editing." *Cell* 171.2 (2017): 470-480.

SVs Impact Function

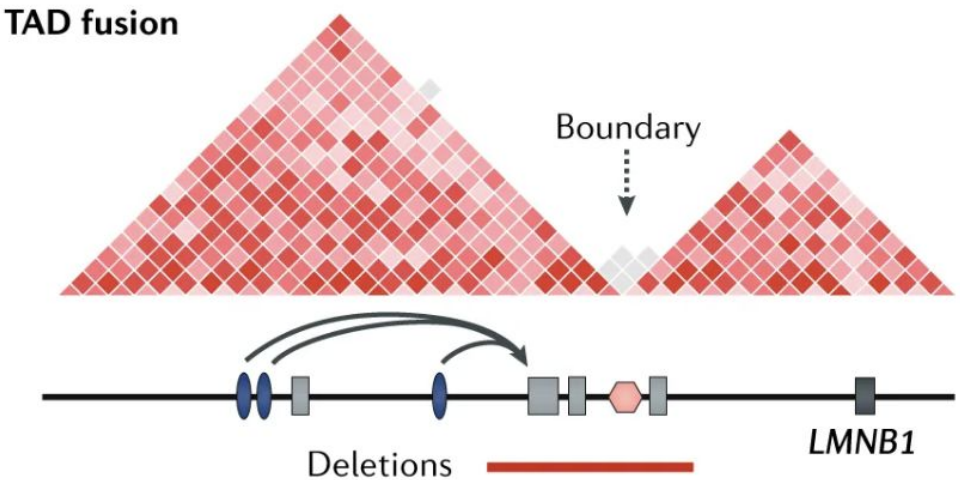
2. Gene regulatory elements

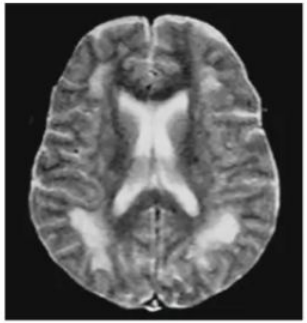


Rodríguez-Leal, Daniel, et al. "Engineering quantitative trait variation for crop improvement by genome editing." *Cell* 171.2 (2017): 470-480.

SVs Impact Function

3. 3D structure





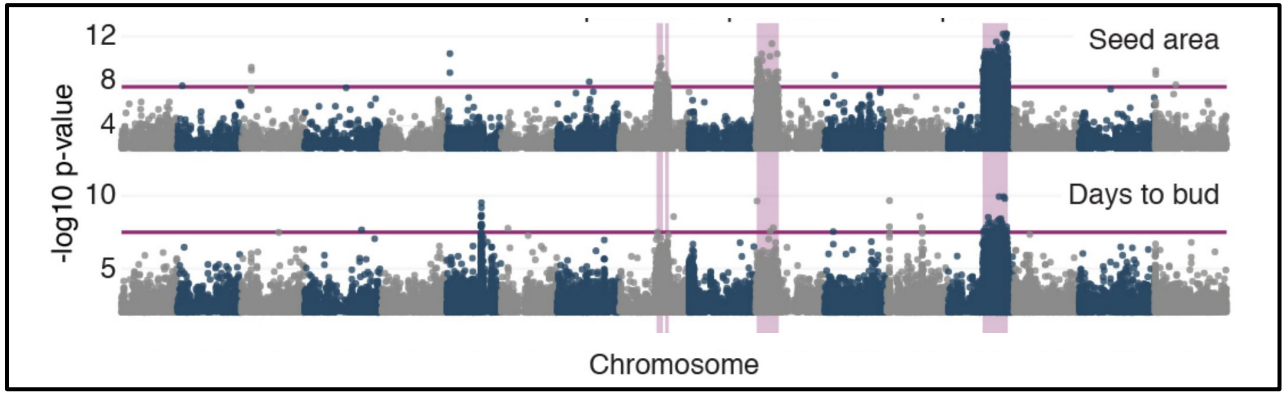
Adult-onset demyelinating leukodystrophy

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. "Structural variation in the 3D genome." *Nature Reviews Genetics* 19.7 (2018): 453-467.

SVs Impact Function

4. Recombination and cellular processes

SVs suppress recombination

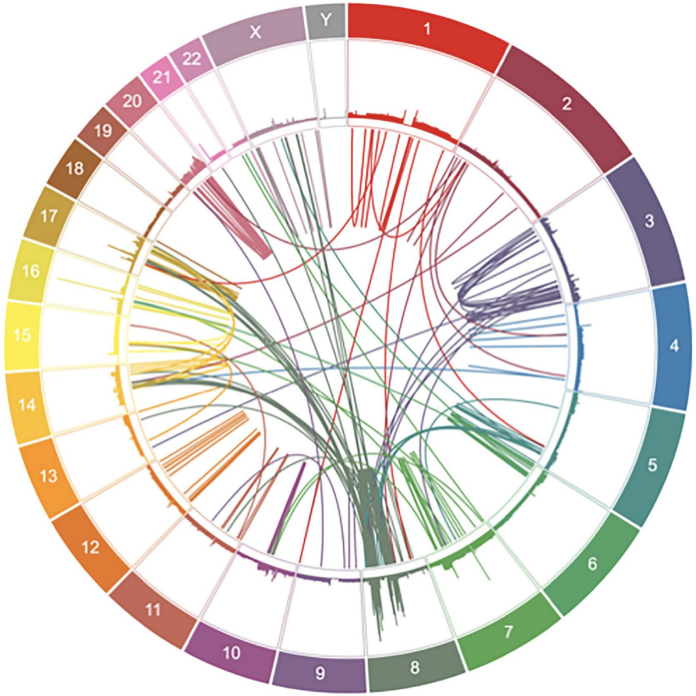


Todesco, Marco, et al. "Massive haplotypes underlie ecotypic differentiation in sunflowers." *bioRxiv* (2019): 790279.

SVs Impact Function

4. Recombination and cellular processes

Translocations in and SK-BR-3 breast cancer cell line



*SVs are prevalent in cancer cells

Nattestad, Maria, et al. "Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line." *Genome research* 28.8 (2018): 1126-1135.

Outline

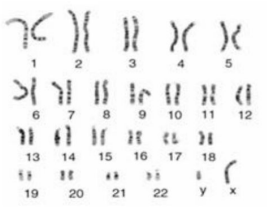
- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Outline

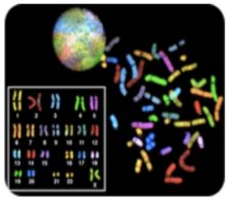
- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

How to Find SVs

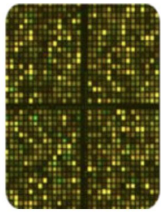
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

Today
High throughput
DNA sequencing

Long Read
DNA sequencing

Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Whole genome alignment

1. Assemble a “query genome”

Nattestad, Maria, and Michael C. Schatz. "Assemblytics: a web analytics tool for the detection of variants from an assembly." *Bioinformatics* 32.19 (2016): 3021-3023.

Whole genome alignment

1. Assemble a “query genome”

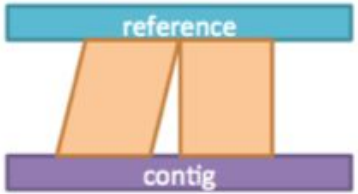
2. Align the query to a reference genome with Nucmer or Minimap2

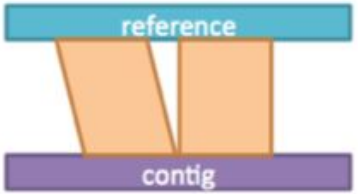
Nattestad, Maria, and Michael C. Schatz. "Assemblytics: a web analytics tool for the detection of variants from an assembly." *Bioinformatics* 32.19 (2016): 3021-3023.

Whole genome alignment

1. Assemble a “query genome”

2. Align the query to a reference genome with Nucmer or Minimap2





Nattestad, Maria, and Michael C. Schatz. "Assemblytics: a web analytics tool for the detection of variants from an assembly." *Bioinformatics* 32.19 (2016): 3021-3023.

Whole genome alignment

1. Assemble a “query genome”

2. Align the query to a reference genome with Nucmer or Minimap2

3. Infer SVs directly from the alignments

- **Tools**
  - **Assemblytics**
  - **Paftools.js**
  - **SyRI**

* Orange parallelograms represent alignments

Nattestad, Maria, and Michael C. Schatz. "Assemblytics: a web analytics tool for the detection of variants from an assembly." *Bioinformatics* 32.19 (2016): 3021-3023.

Whole genome alignment

Downsides

Nattestad, Maria, and Michael C. Schatz. "Assemblytics: a web analytics tool for the detection of variants from an assembly." *Bioinformatics* 32.19 (2016): 3021-3023.

Whole genome alignment

Downsides

- Assembly-to-assembly alignment is fallible
  - Sensitivity vs. specificity is hard to get right
  - Confounded by repeats
  - Alignment heuristics don’t always produce the best results (especially for plant genomes)
- Blind to some heterozygous SVs in unphased assemblies
- Genome assembly is hard!
  - Imperfections in assemblies lead to imperfections in the SV calls

Nattestad, Maria, and Michael C. Schatz. "Assemblytics: a web analytics tool for the detection of variants from an assembly." *Bioinformatics* 32.19 (2016): 3021-3023.

Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

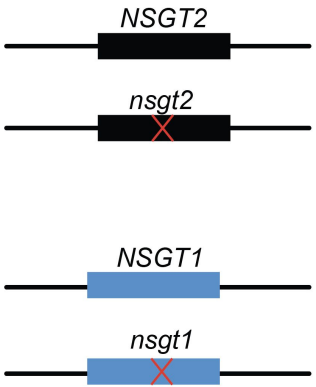
Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Alignment Free SV Calling

1. Instead of aligning whole genomes, just align smaller genomic elements, like genes. (e.g. with BLAST)

Various NSGT alleles

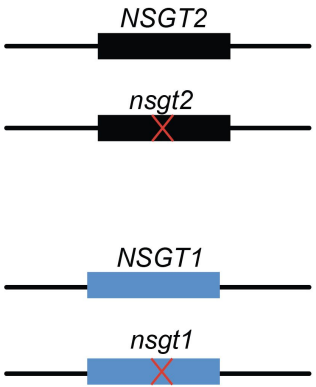


Alonge and Wang *et al*., unpublished

Alignment Free SV Calling

1. Instead of aligning whole genomes, just align smaller genomic elements, like genes. (e.g. with BLAST)

Various NSGT alleles



BLAST

5 different tomato assemblies

PAS014479

M82

BGV006775

Fla.8924

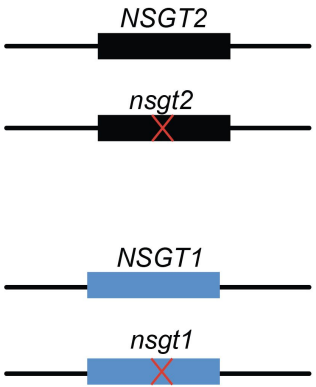
PI129033

Alonge and Wang *et al*., unpublished

Alignment Free SV Calling

1. Instead of aligning whole genomes, just align smaller genomic elements, like genes. (e.g. with BLAST)

Various NSGT alleles



BLAST

5 different tomato assemblies

PAS014479

M82

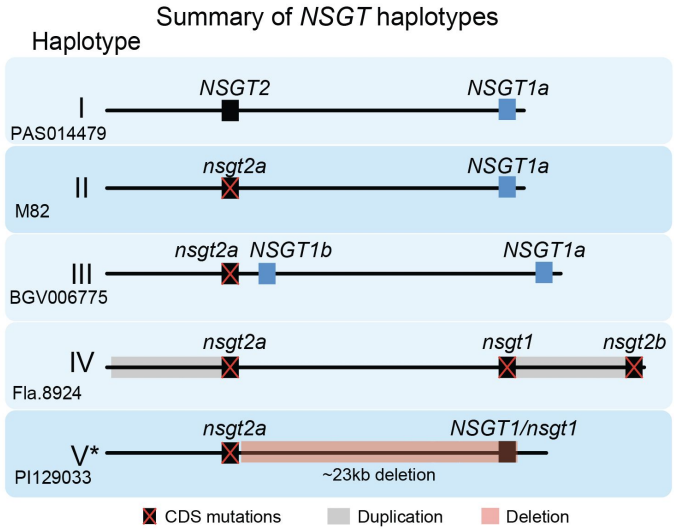
BGV006775

Fla.8924

PI129033

Manually Infer Haplotypes

NSGT CNVs



Alonge and Wang *et al*., unpublished

Alignment Free SV Calling

1. Instead of aligning w … (e.g. with BLAST)

Various NSGT alleles

NVs

SGT haplotypes

*** Aligning individual elements is time-consuming.**

This is usually done when you have a reason to suspect that a particular element may be structurally variable.

Alonge and Wang *et al.*, unpublished

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC

AAC

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC
AAC
 ACC

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC
AAC GAT ACA TCG ATT CAT
ACC ATT CAA CGG TTA ATG
CCG TTA AAT GGA TAC TGT
CGA TAC ATC GAT ACA GTC

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC

AAC GAT ACA TCG ATT CAT

ACC ATT CAA CGG TTA ATG

CCG TTA AAT GGA TAC TGT

CGA TAC ATC GAT ACA GTC

2. Store k-mer counts

AAC, 1	AAT, 1
ACC, 1	ATC, 1
CCG, 1	TCG, 1
CGA, 1	CGG, 1
GAT, 2	GGA, 1
ATT, 2	CAT, 1
TTA, 2	ATG, 1
TAC, 2	TGT, 1
ACA, 2	GTC, 1
CAA, 1	

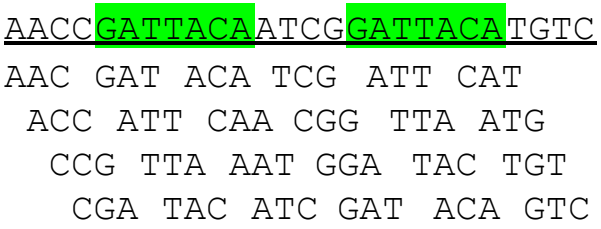
Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

2. Store k-mer counts

3. Assign a count to the k-mer starting at every offset



AAC, 1	AAT, 1
ACC, 1	ATC, 1
CCG, 1	TCG, 1
CGA, 1	CGG, 1
GAT, 2	GGA, 1
ATT, 2	CAT, 1
TTA, 2	ATG, 1
TAC, 2	TGT, 1
ACA, 2	GTC, 1
CAA, 1	

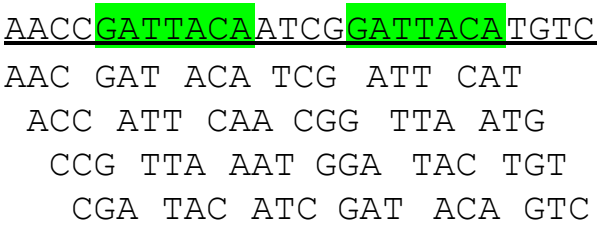
Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

2. Store k-mer counts

3. Assign a count to the k-mer starting at every offset

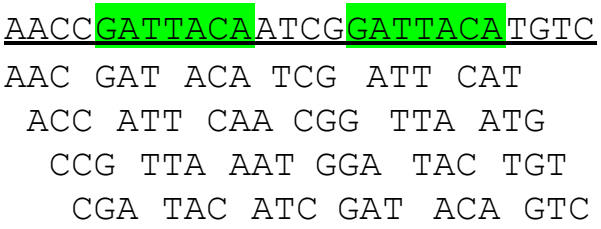


AAC, 1	AAT, 1
ACC, 1	ATC, 1
CCG, 1	TCG, 1
CGA, 1	CGG, 1
GAT, 2	GGA, 1
ATT, 2	CAT, 1
TTA, 2	ATG, 1
TAC, 2	TGT, 1
ACA, 2	GTC, 1
CAA, 1	

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer



2. Store k-mer counts

AAC, 1	AAT, 1
ACC, 1	ATC, 1
CCG, 1	TCG, 1
CGA, 1	CGG, 1
GAT, 2	GGA, 1
ATT, 2	CAT, 1
TTA, 2	ATG, 1
TAC, 2	TGT, 1
ACA, 2	GTC, 1
CAA, 1	

3. Assign a count to the k-mer starting at every offset

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC

AAC GAT ACA TCG ATT CAT

ACC ATT CAA CGG TTA ATG

CCG TTA AAT GGA TAC TGT

CGA TAC ATC GAT ACA GTC

2. Store k-mer counts

AAC, 1	AAT, 1
ACC, 1	ATC, 1
CCG, 1	TCG, 1
CGA, 1	CGG, 1
GAT, 2	GGA, 1
ATT, 2	CAT, 1
TTA, 2	ATG, 1
TAC, 2	TGT, 1
ACA, 2	GTC, 1
CAA, 1	

3. Assign a count to the k-mer starting at every offset

111122222111111222221111

AACCGATTACAATCGGATTACATGTC

Alignment Free SV Calling

2. Count k-mers to find duplications

1. Go through each k-mer

AACCGATTACAATCGGATTACATGTC

AAC GAT ACA TCG ATT CAT

ACC ATT CAA CGG TTA ATG

CCG TTA AAT GGA TAC TGT

CGA TAC ATC GAT ACA GTC

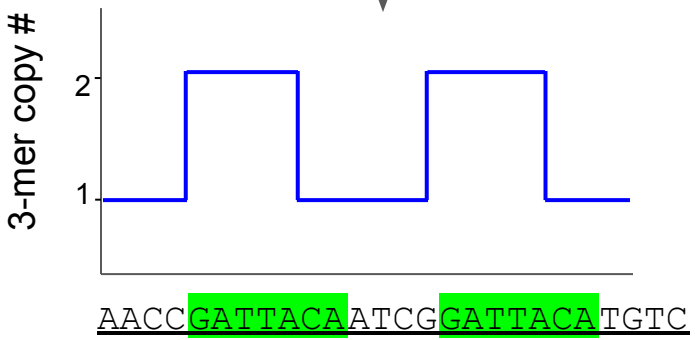
2. Store k-mer counts

AAC, 1	AAT, 1
ACC, 1	ATC, 1
CCG, 1	TCG, 1
CGA, 1	CGG, 1
GAT, 2	GGA, 1
ATT, 2	CAT, 1
TTA, 2	ATG, 1
TAC, 2	TGT, 1
ACA, 2	GTC, 1
CAA, 1	

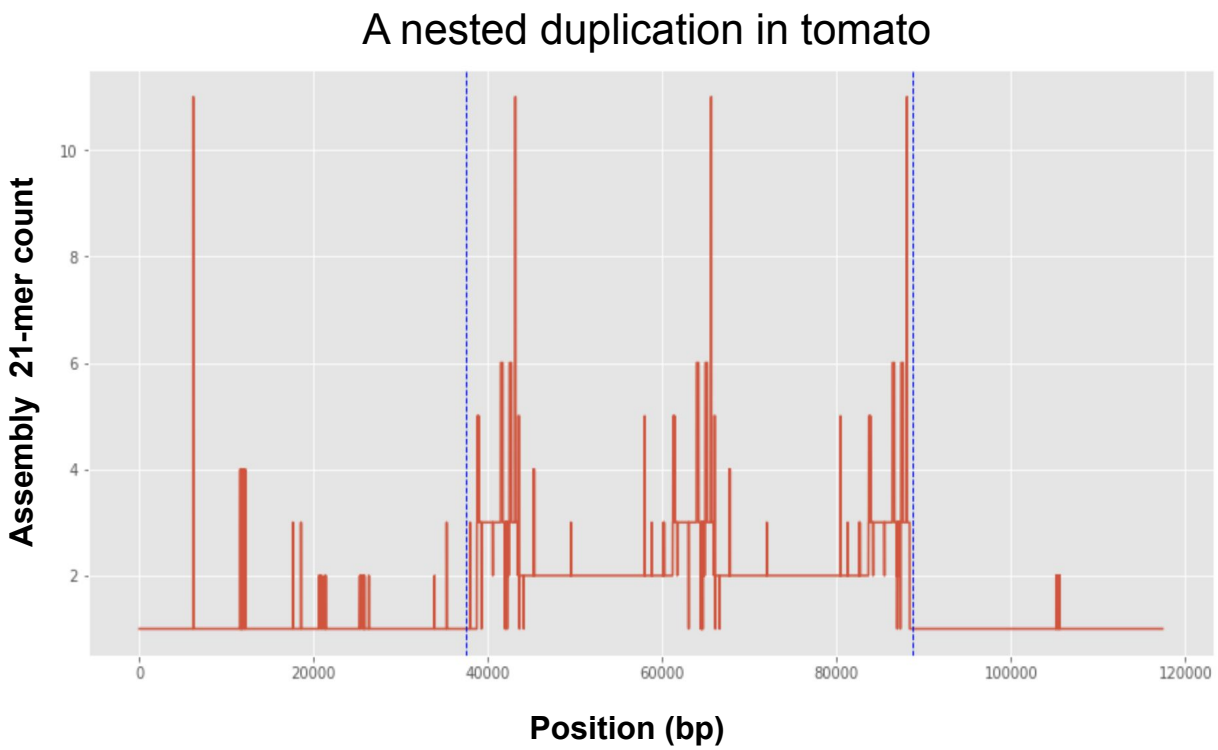
3. Assign a count to the k-mer starting at every offset

111122222111111222221111

AACCGATTACAATCGGATTACATGTC



Alignment Free SV Calling



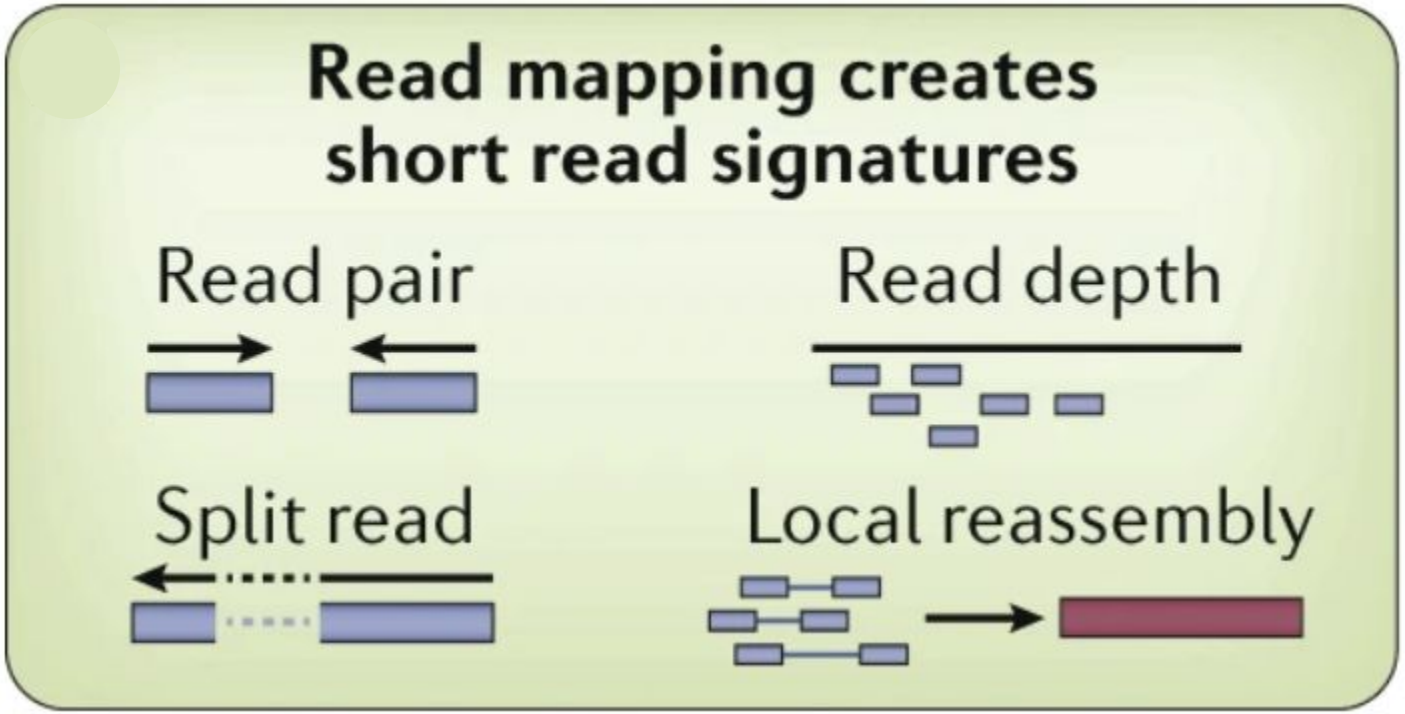
Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a ref reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Outline

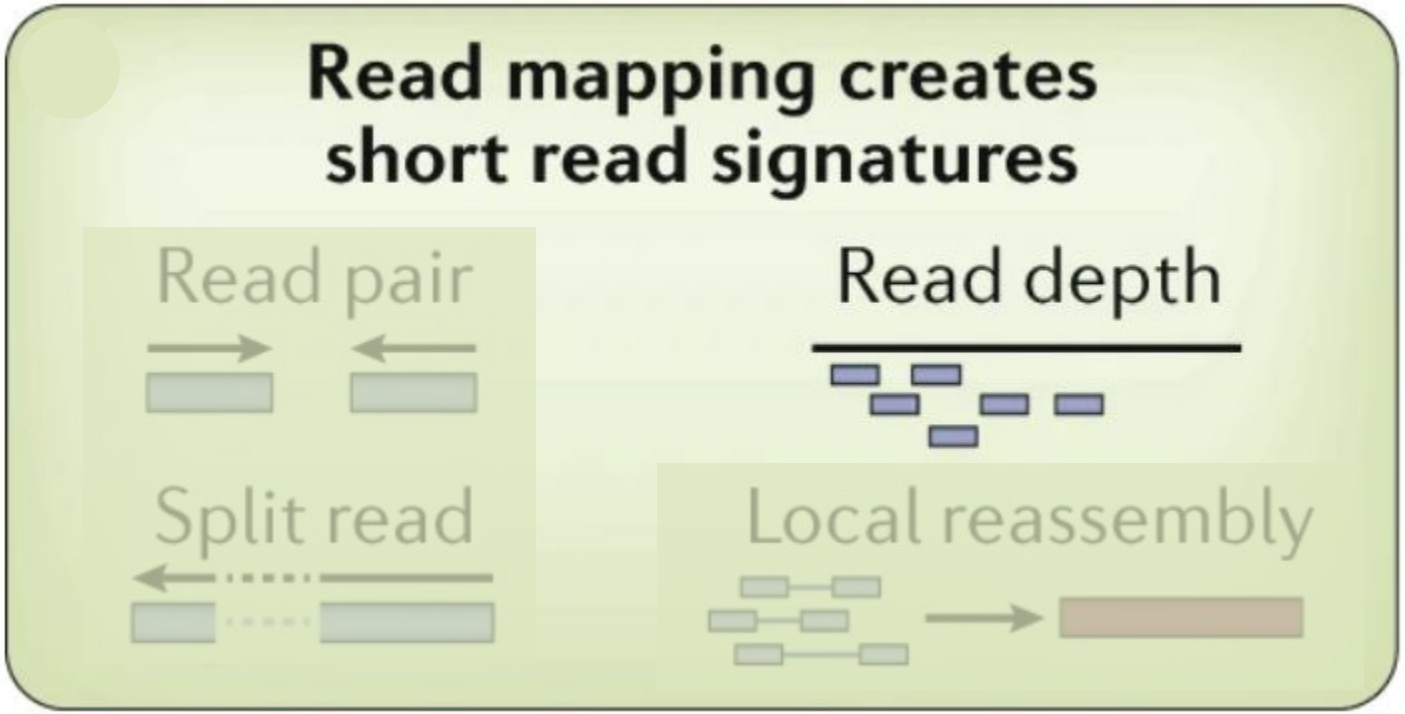
- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a ref reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Short-read Mapping



Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." *Nature Reviews Genetics* (2019): 1-19.

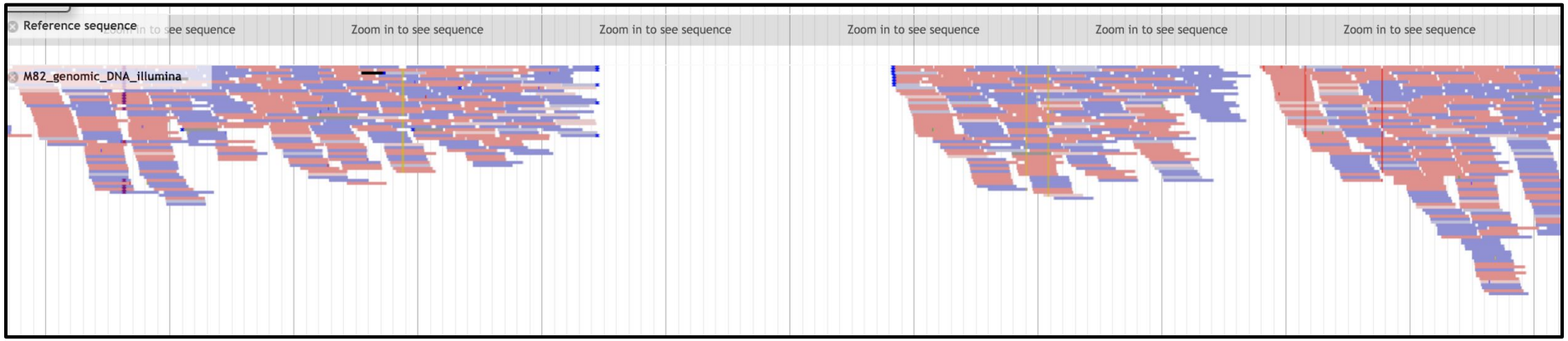
Short-read Mapping: coverage



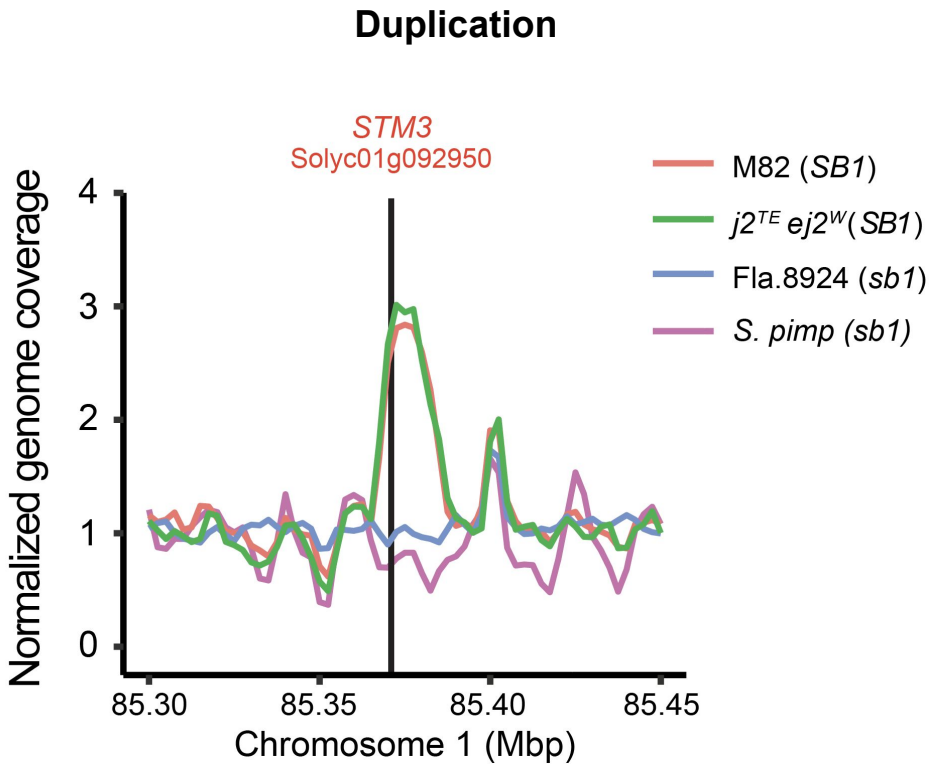
Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." *Nature Reviews Genetics* (2019): 1-19.

Short-read Mapping: Coverage

Deletion

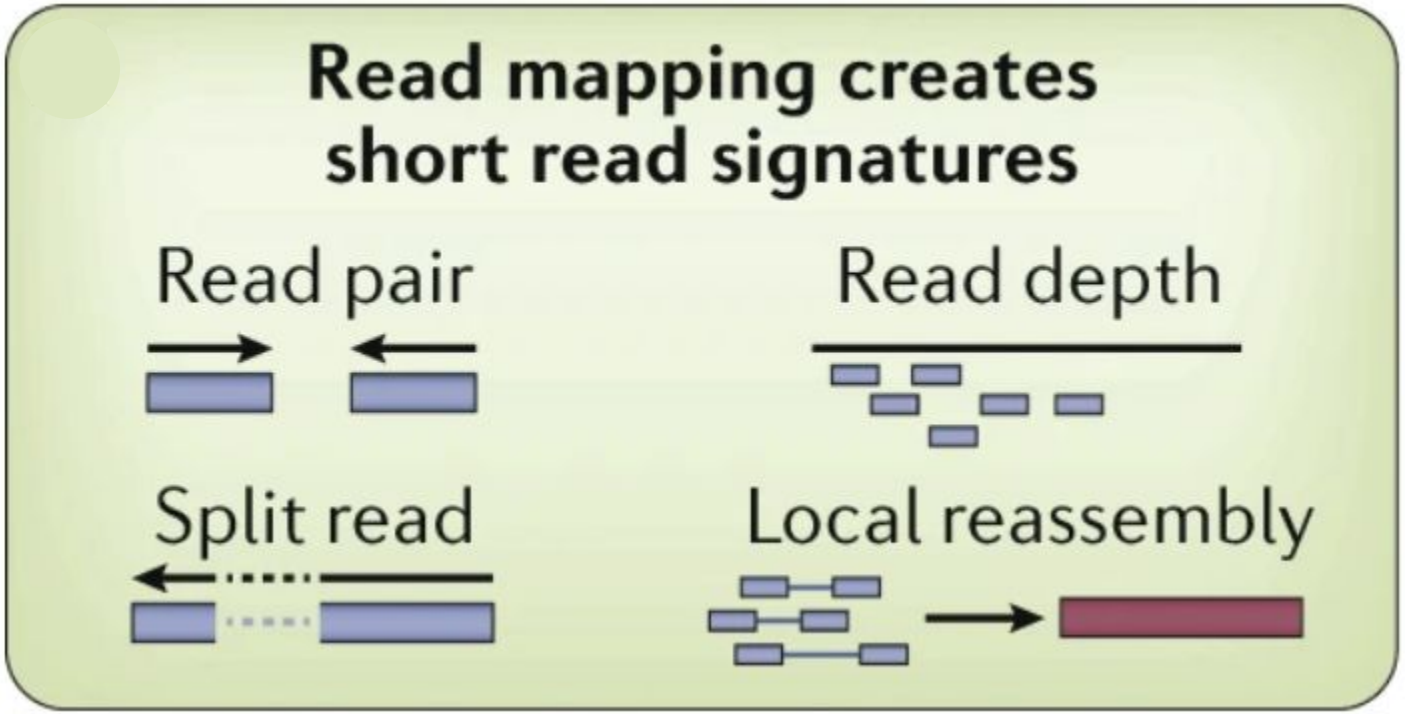


Short-read Mapping: Coverage



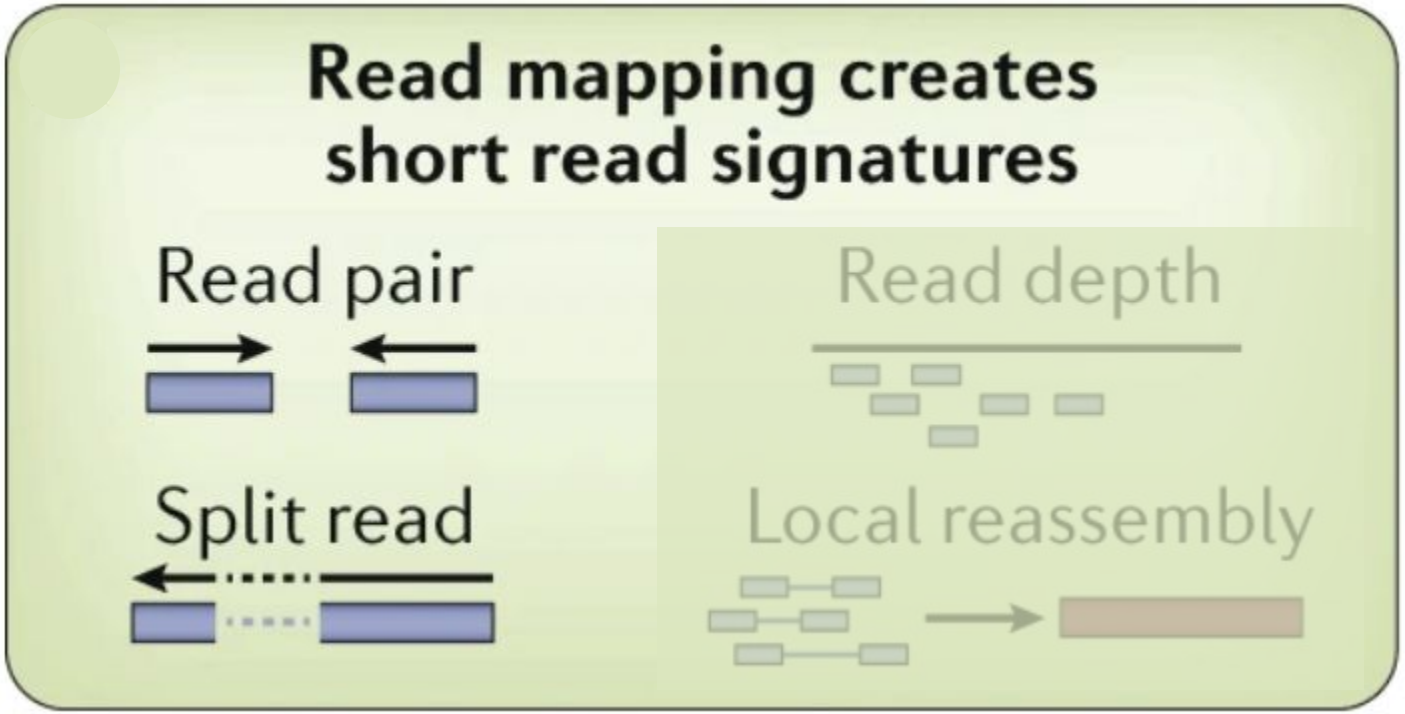
Alonge and Wang *et al*., unpublished

Short-read Mapping



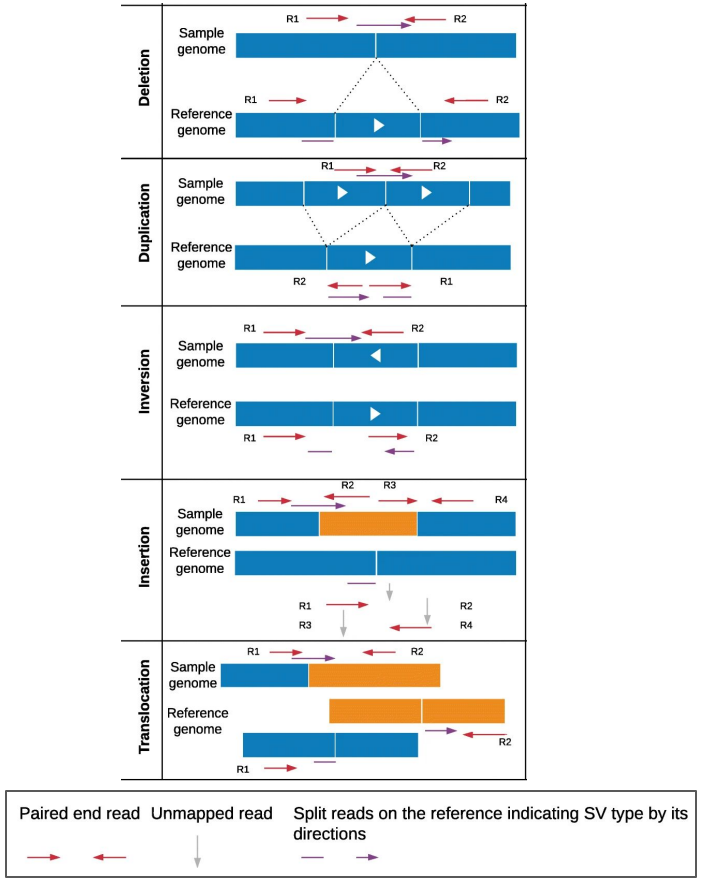
Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." *Nature Reviews Genetics* (2019): 1-19.

Short-read Mapping: Mates and Split-reads



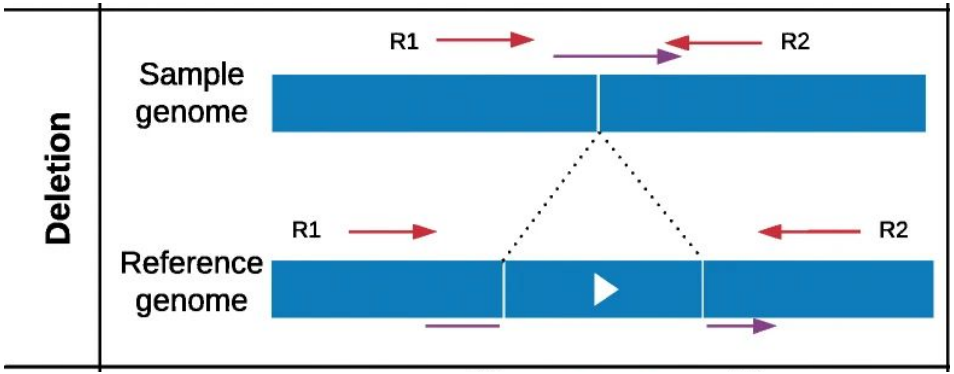
Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." *Nature Reviews Genetics* (2019): 1-19.

Short-read Mapping: Mates and Split-reads



Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." *Genome biology* 20.1 (2019): 246.

Short-read Mapping: Mates and Split-reads



Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." *Genome biology* 20.1 (2019): 246.

Short-read Mapping: Downfalls

Coverage

Short-read Mapping: Downfalls

Coverage

- Coverage can be affected by other factors aside from SVs
  - Reference bias
  - Repeats

Short-read Mapping: Downfalls

Coverage

- Coverage can be affected by other factors aside from SVs
  - Reference bias
  - Repeats

Split-read/paired-end

- Reads are already short, so they must be split into very short fragments to produce split-read alignments.
  - These short fragments can produce unreliable alignments
- Discordant mate-pair alignments are often misleading

Short-read Mapping: Downfalls

Insertions

- Reads are usually too short to contain an insertion and anchor it to flanking sequence
- Supporting insertion reads (if you can find them) are hard to assemble into a proper insertion sequence
- Many short-read SV callers don’t even bother trying to call insertions

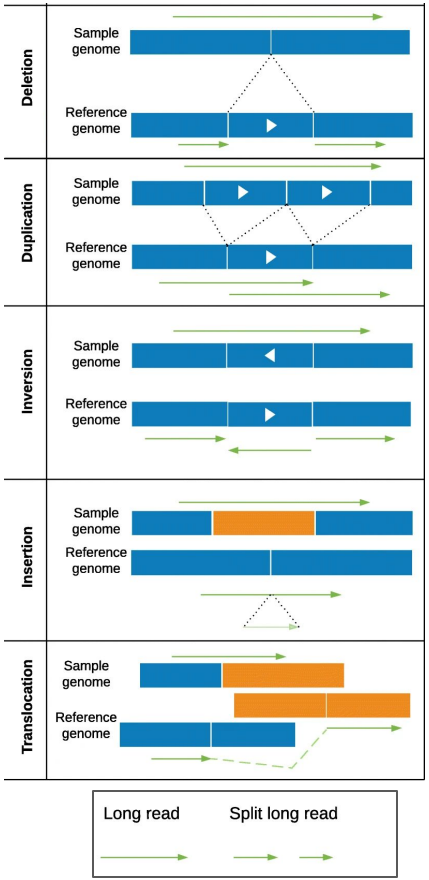
Outline

- Introduction to genome “structure”
- Functional importance of genome structure
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a ref reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- Applications in Tomato

Outline

- Introduction to genome “structure”
- Functional importance of genome structure
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a ref reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- Applications in Tomato

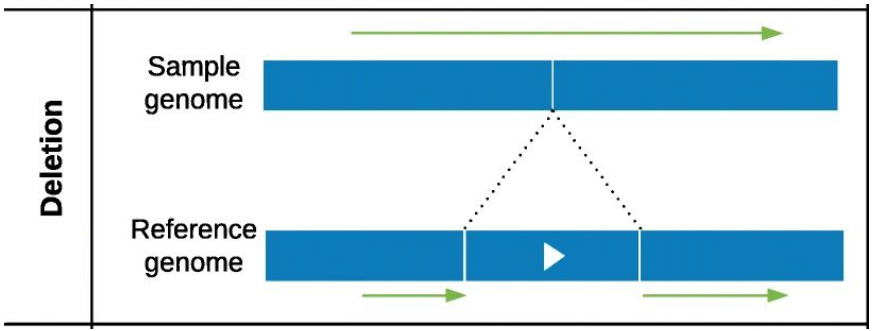
Long-read Mapping



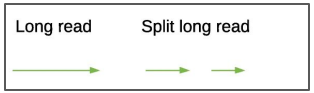
* Each green arrow is the same long read (or a portion of that long read)

Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." *Genome biology* 20.1 (2019): 246.

Long-read Mapping: Deletions

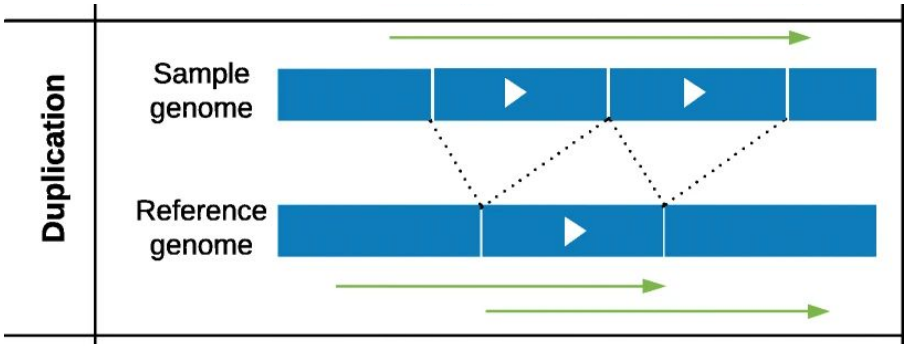


* Each green arrow is the same long read (or a portion of that long read)

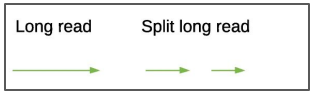


Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." *Genome biology* 20.1 (2019): 246.

Long-read Mapping: Duplications

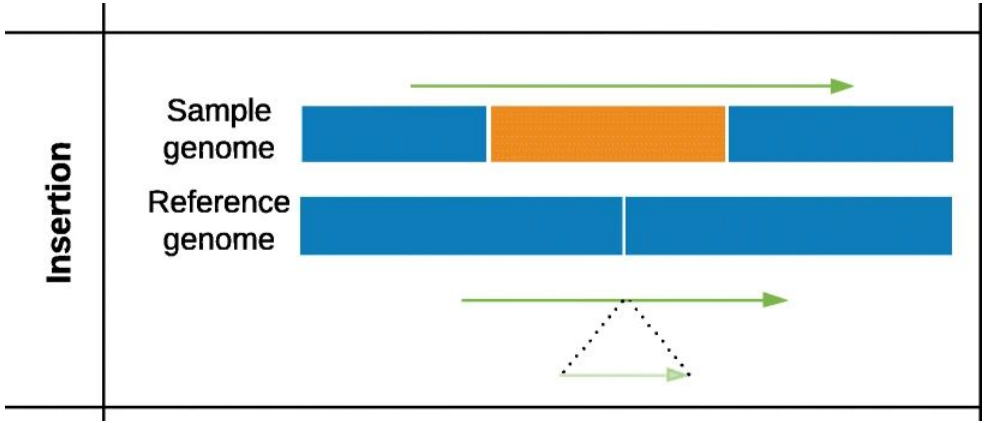


* Each green arrow is the same long read (or a portion of that long read)

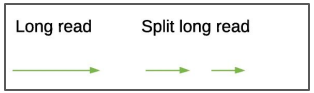


Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." *Genome biology* 20.1 (2019): 246.

Long-read Mapping: Insertions



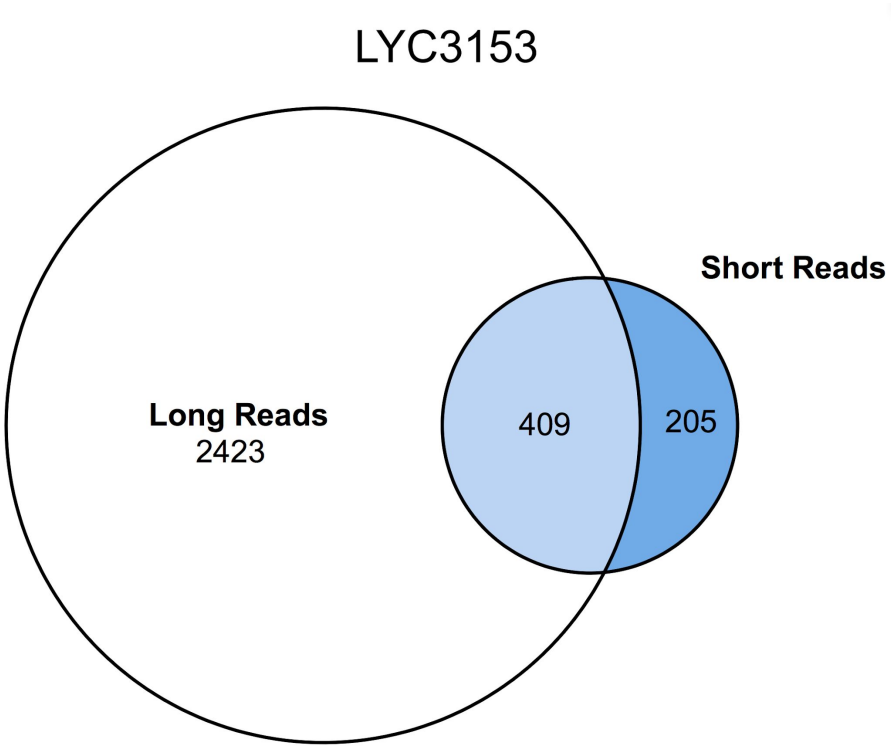
* Each green arrow is the same long read (or a portion of that long read)



Mahmoud, Medhat, et al. "Structural variant calling: the long and the short of it." *Genome biology* 20.1 (2019): 246.

Long-Read Mapping > Short-Read Mapping

- Long-reads provide more sensitive SV calls

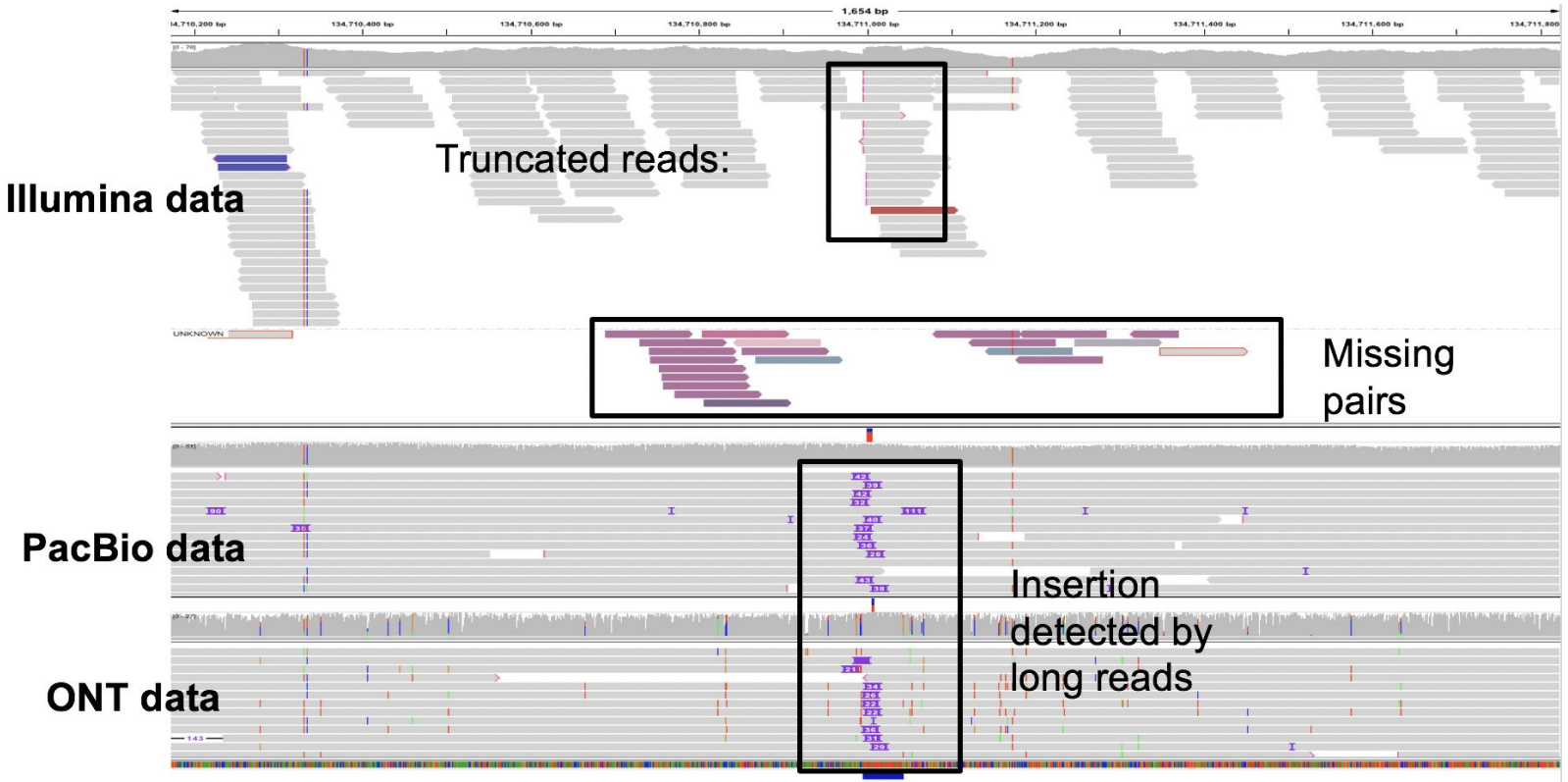


Alonge and Wang *et al*., unpublished

Long-Read Mapping > Short-Read Mapping

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

Long-Read Mapping > Short-Read Mapping



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

Outline

- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a ref reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Outline

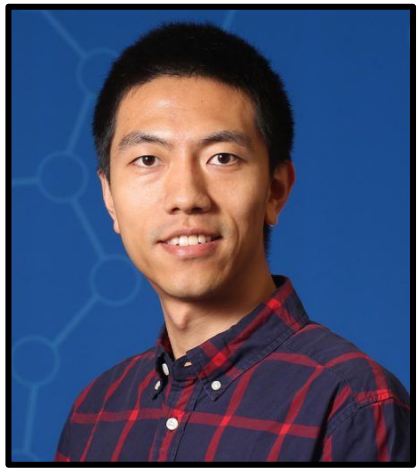
- **Introduction to genome “structure”**
- **Functional importance of genome structure**
- **The Bioinformatics of SV calling**
  - Assembly
    - Whole genome alignment to a reference
    - (Whole genome) Alignment free analysis
  - Read Mapping
    - Short-read mapping
    - Long-read mapping
- **Applications in Tomato**

Applications: Tomato

Zach Lippman
CSHL/HHMI

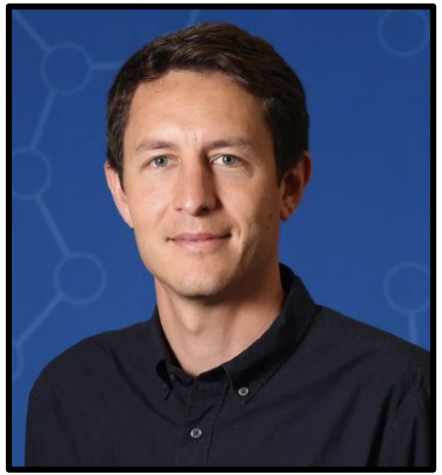
Matthias Benoit
Postdoc, CSHL

CSH Cold Spring Harbor Laboratory



Xingang Wang
Postdoc, CSHL

Sebastian Soyk
Asst. Professor, UNIL
(formerly postdoc, CSHL)



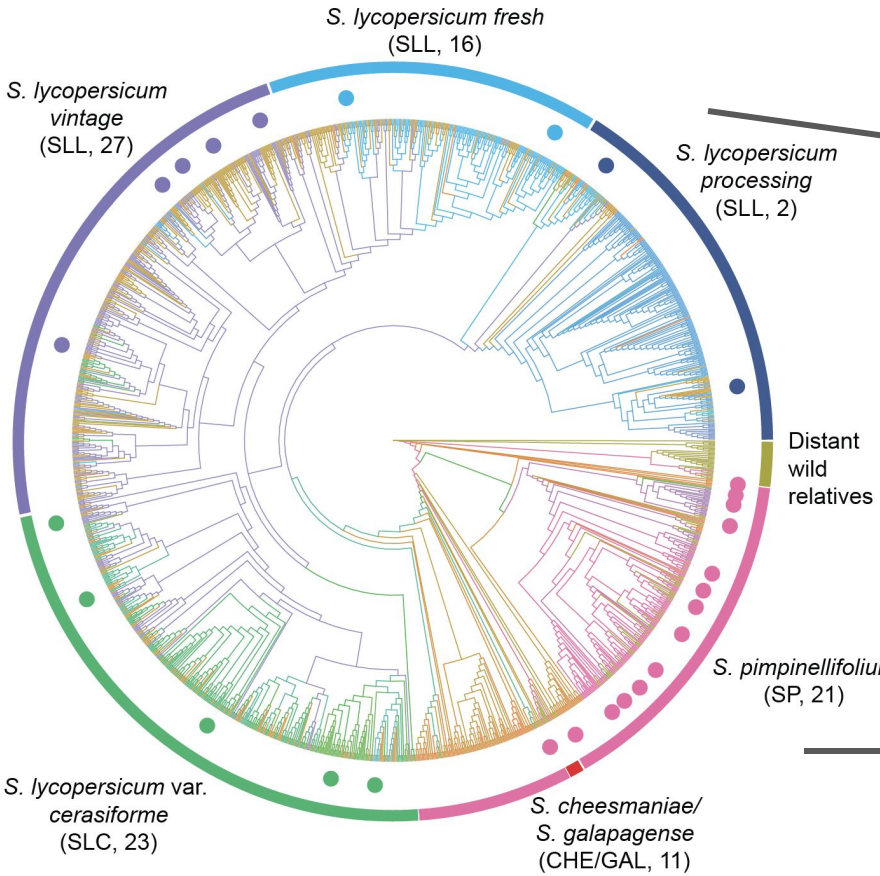
Tomato is an Important Model and Application

- **Naturally self-fertilizing**
- **Diploid**
- **Amenable to transformation**
  - **Gene editing with Cas9 well demonstrated.**
- Medium genome size (1 Gbp)
- Short life cycle (90 - 100 days)
- Amenable to cross-hybridization
  - Introgression Line (ILs) Populations
- Robust genetic/genomic resources
  - High-quality reference genome
  - Population-scale DNA and RNA seq databases.
  - Extensive mutant germplasm

- $50 billion industry
- Major source of nutrients

Achieving sustainable cultivation of tomatoes, Mattoo & Handa (2017)

Tomato Domestication

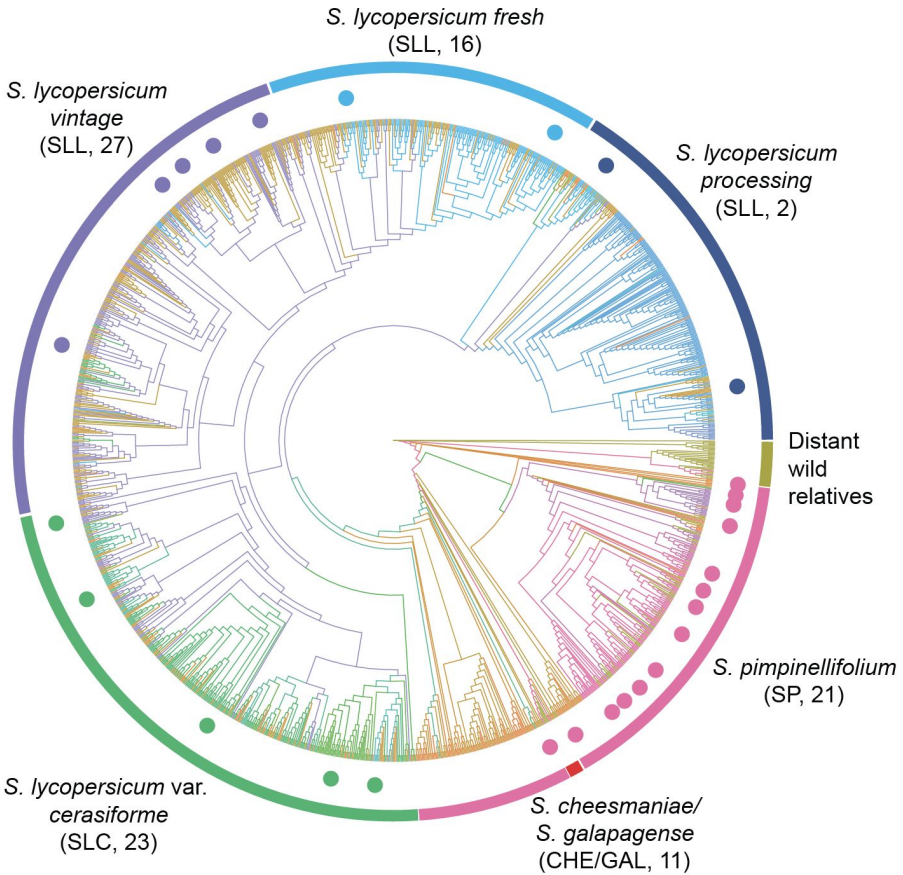


Alonge and Wang *et al*., unpublished

Modern

Wild Progenitor

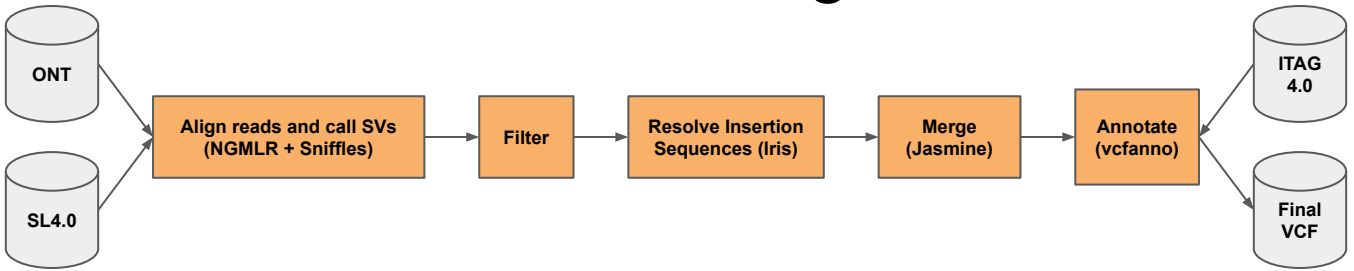
Sample Selection and Sequencing

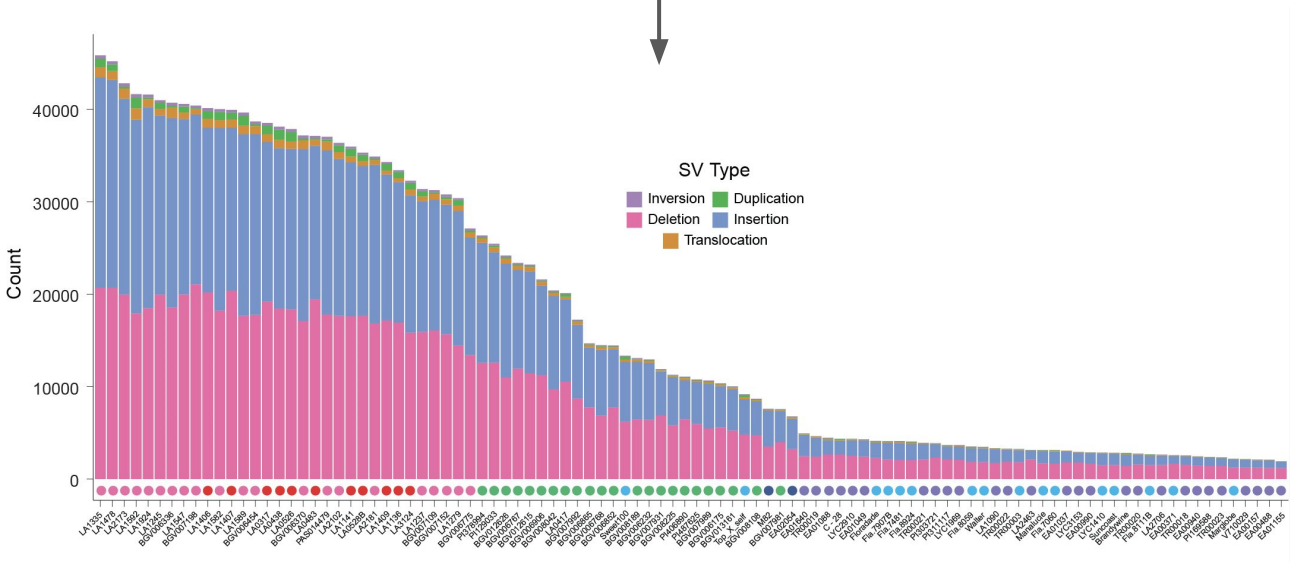


Alonge and Wang *et al*., unpublished

x 100

SV Calling

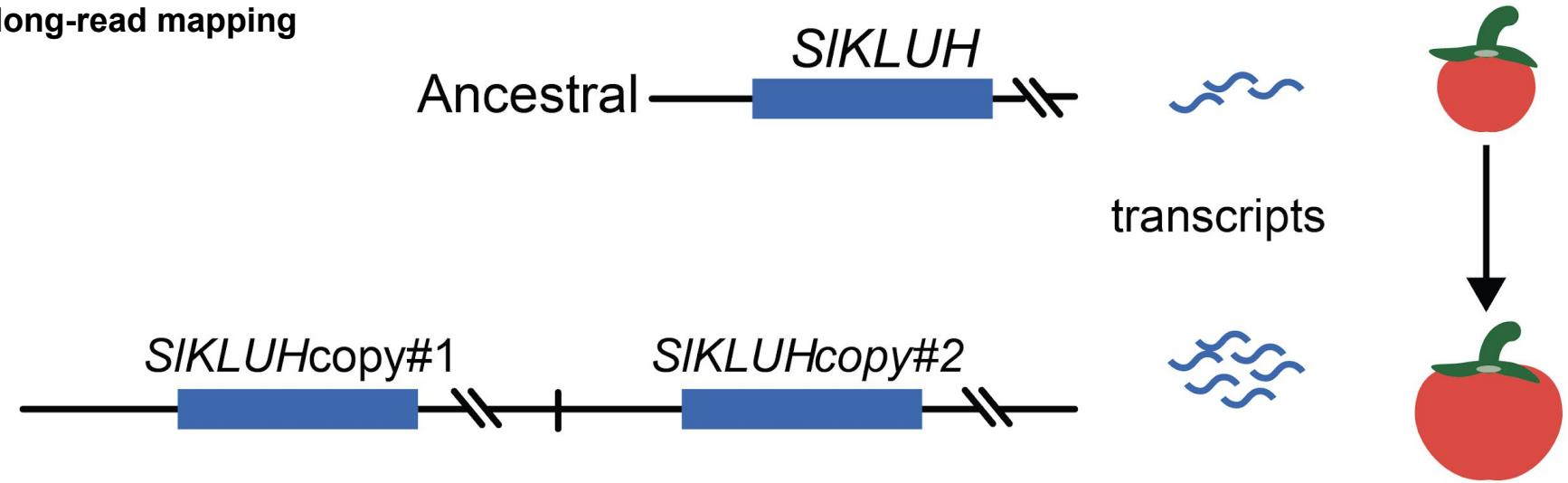




Alonge and Wang *et al*., unpublished

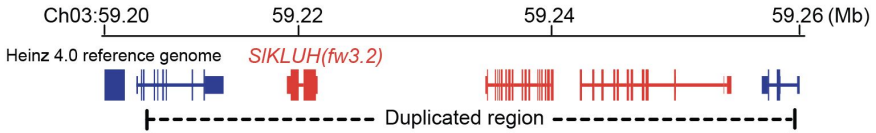
A Duplication Underlies a Fruit Weight QTL

*** Found with long-read mapping**



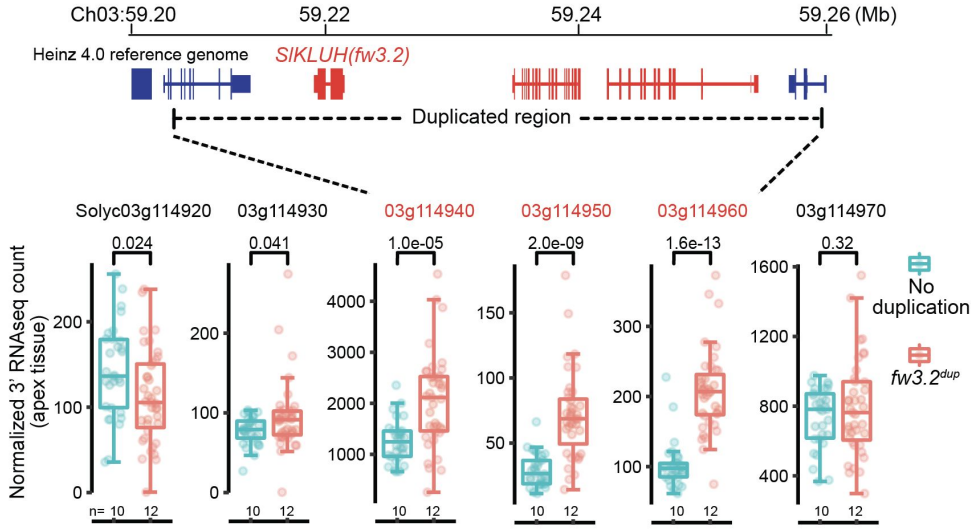
Alonge and Wang *et al*., unpublished

SVs Impact Gene Expression



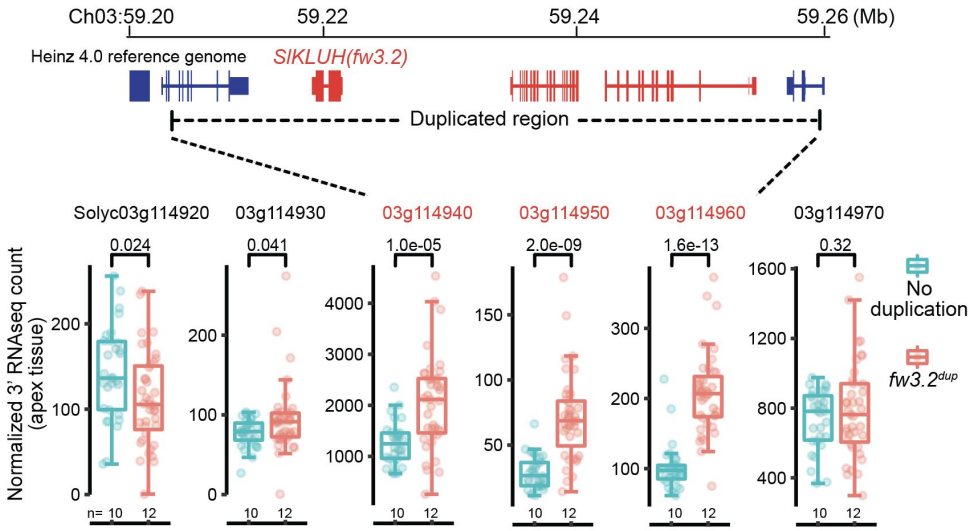
Alonge and Wang *et al*., unpublished

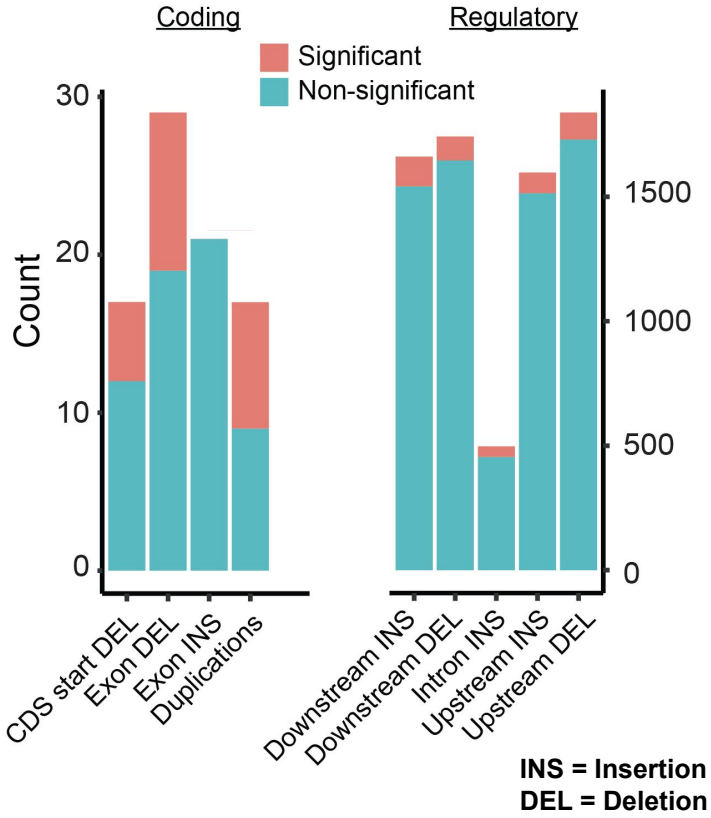
SVs Impact Gene Expression



Alonge and Wang *et al.*, unpublished

SVs Impact Gene Expression

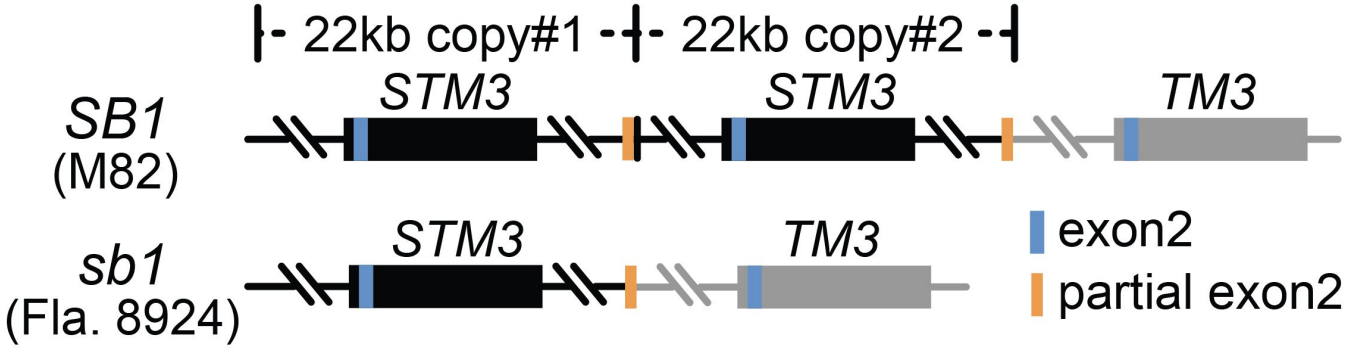




Alonge and Wang *et al.*, unpublished

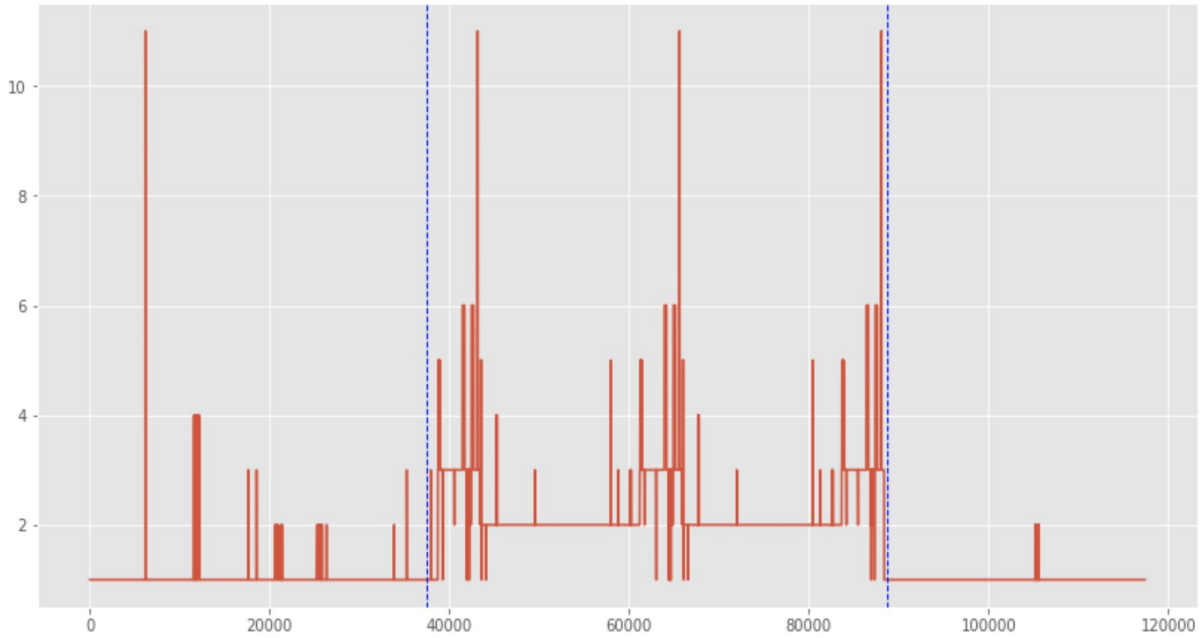
A Nested Duplication Influences Flowering

* Detected with
“alignment free”
assembly methods



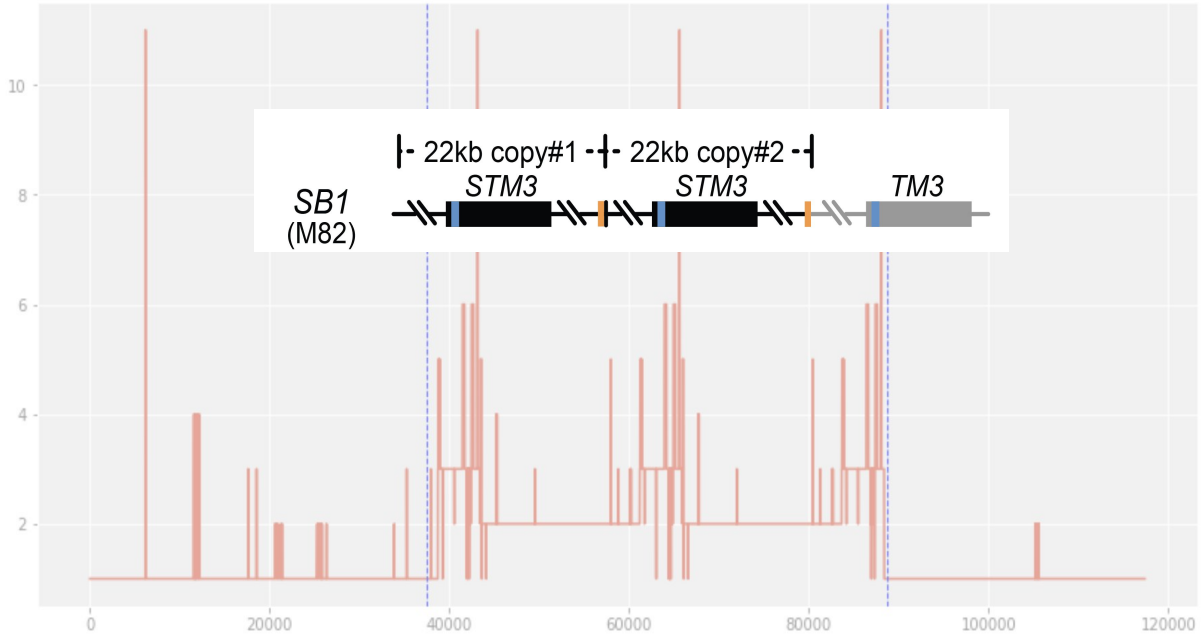
A Nested Duplication Influences Flowering

* Detected with “alignment free” assembly methods



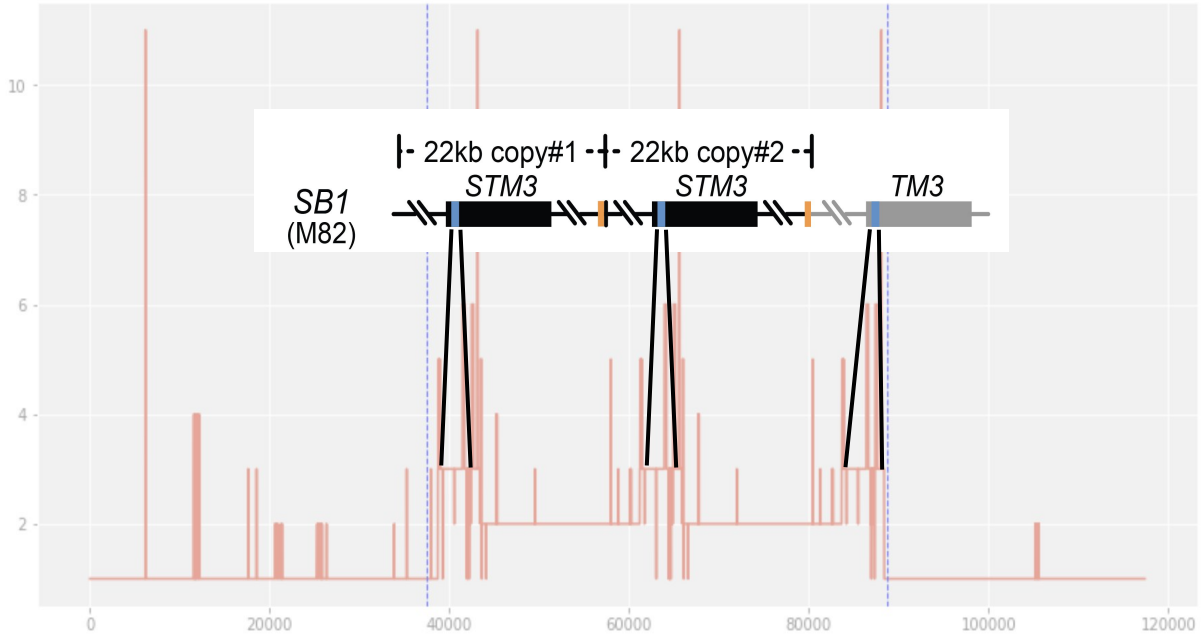
A Nested Duplication Influences Flowering

* Detected with “alignment free” assembly methods



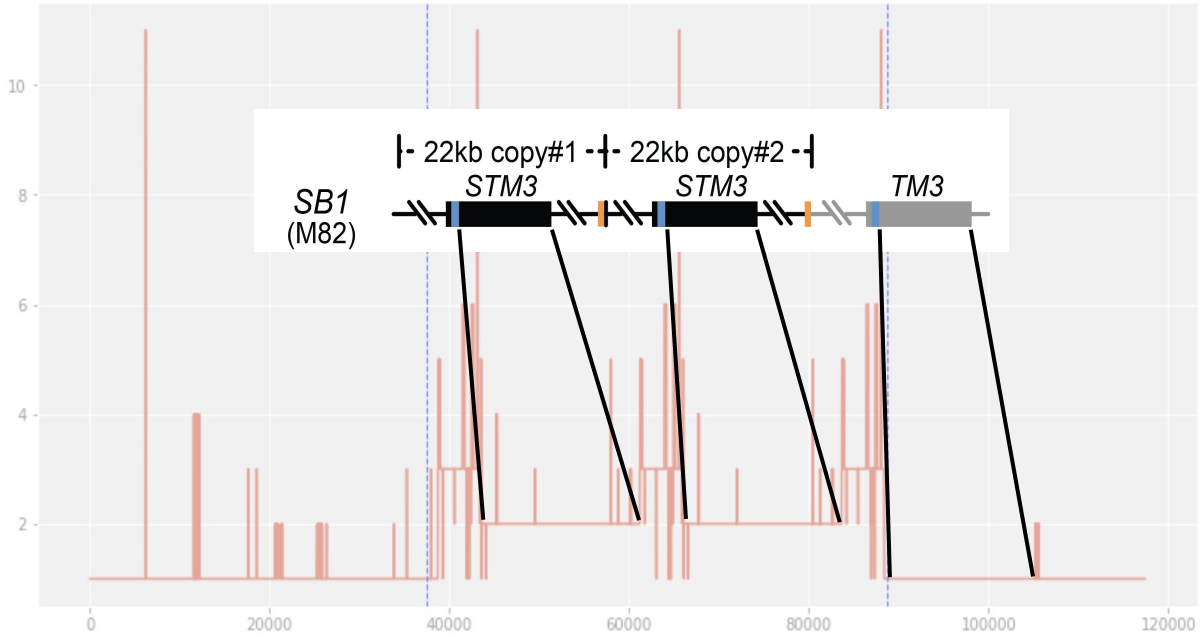
A Nested Duplication Influences Flowering

* Detected with “alignment free” assembly methods



A Nested Duplication Influences Flowering

* Detected with “alignment free” assembly methods



Conclusions: Why does this matter?

- SVs comprise a substantial portion of the natural genetic variation that we see in eukaryotes, both in terms of count and total bp.
- SVs underlie many of the traits that we care about:
  - Plant domestication and breeding QTL
  - Human diseases
- SVs impact cellular function and can shape evolution

Conclusions: Pan-genomics

- There is a lot of genomic “structure” in a population that is not captured by a single reference genome or a few reference genomes.
- Uncovering this structure has broad research impact. E.g.:
  - Helping the utility of resequencing experiments with a pan-genome graph-like datastructure.
    - Reduce reference bias
  - Discover more natural alleles!!!!
    - More assemblies reveal more potentially functional alleles
    - Assembly will probably replace WGS resequencing experiments for variant calling.