100 Genomes in 100 Days:
The Structural Variant Landscape of Tomato Genomes
Michael Schatz

January 15, 2019
PAG2019 Bioinformatics Workshop

@mike_schatz
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 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 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

Project overview
1. Select diverse samples
2. Long read sequencing
3. Per-sample SV identification
4. Pan-tomato SV landscape
5. Identify and validate SVs associated with agricultural and phenotypic traits
~ Part 1 ~
Sample Selection
Tomato Population Genetics

More than 900 varieties of tomato and related species have been sequenced with short reads
- Most are between 20x and 40x coverage
- SNP-based phylogenetic tree shows 10 major clades

Initial examination of SVs
- Identify variants using an aggregation of three individual methods to improve sensitivity
  - Lumpy, Manta & Delly
  - Retain calls supported by ≥ 2 callers
  - Reduces false-positives while not worsening false-negatives
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
Optimized Sample Selection

Our goal is to select the 100 samples that collectively capture the most diversity. Short-read based SVs will undersample variants but still represent relative diversity. Selecting 100 at random only recovers about 40% of the total known diversity. Ranking samples by number of variants picks diverse samples, although tends to pick siblings (nearly duplicate samples). Optimal strategy is NP-hard using a set-cover algorithm, we approximate using a greedy approach.

SVCollector: Optimized sample selection for validating and long-read resequencing of structural variants
~ Part 2 ~
Long Read Sequencing
Nanopore Performance at CSHL

Sara Goodwin

Sequencing strategy

- Initial proposal called for mix of short, long, and linked read sequencing
- MinION and GridION became feasible spring/summer 2018
- Encouraging PromethION yields on test runs mid-summer 2018 motivated switch in strategy

Yield Gbs

MinION + GridION

PromethION

74GB
Nanopore Performance at CSHL
Sara Goodwin

MinION + GridION  PromethION

74GB  130GB

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

- Routinely achieve 60-80 Gb / flow cell, many over 100 Gb
- First 62 samples have been sequenced to at least 40x
- Now at 12 to 16 samples per week
  - 100 samples 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**
- Mean read length: 10kbp – 20kbp
- Read length N50: 15kbp – 30kbp
- Over 20x coverage of reads over 20kbp
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
Alignment Based Analysis

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
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 faster options
- Miniasm (Li, Bioinformatics, 2016) runs in ~72 core hours (+1.5 days for consensus)
- Wtdbg2 (https://github.com/ruanjue/wtdbg2) runs in ~8 core hours (+1.5 days for consensus) although mixed results depending on sample
- Discussing cloud-enabled pipelines with DNAnexus

Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation
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

Fast and accurate reference-guided scaffolding of draft genomes
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

---

*Fast and accurate reference-guided scaffolding of draft genomes*
Assembly Based Analysis

RaGOO scaffolding yields essentially complete chromosomes
- Final polishing using Bowtie2 + Pilon
  - Substantially faster than Nanopolish, and modestly more accurate based on gene-analysis and alignment to reference
- Gene annotations using MAKER

Identify structural variants using Assemblytics
- Finds variants within and between alignments
- Especially important for large insertions of novel sequences
- Tens of thousands of SVs, widely distributed across chromosomes

Assemblytics: a web analytics tool for the detection of variants from an assembly
~ Part 4 ~
The Landscape of Structural Variation in Tomato Genomes
The landscape of structural variations

Landscape of the first 20 accessions
- Substantial variation between samples
  - 15 to 50 thousand structural variations each
  - Mostly insertions + deletions

Population Genetics
- Most variants specific to 1 sample
- Many variant shared by multiple samples, including some in all 20
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

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


Sebastian Soyk
Cold Spring Harbor Laboratory
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
Michael Kirsche
Sam Kovaka
T. Rhyker
Ranallo-Benavide
Rachel Sherman

*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
Thank you!

@mike_schatz
http://schatz-lab.org
From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
ONT data

PacBio data

Illumina data

Truncated reads:

Insertion detected by long reads

Missing pairs