UNIVERSITY OF CALIFORNIA, DAVIS
PUBLIC STRAWBERRY BREEDING PROGRAM

New genomic tools for studying cultivated and wild octoploid strawberry

JANUARY 12, 2019

Roadmap Forward…
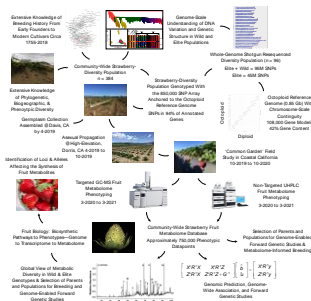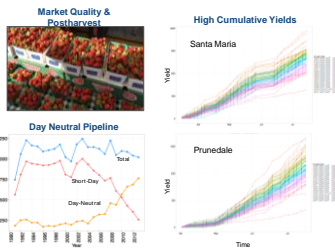
---

## UC Davis Strawberry Breeding Program

The last 30 years…

- Successful focus on yield and marketable quality for California field production
- Day neutral breeding has increased potential
- Flavor is valued, but not historically an trait for growers
- Virtually no introduction of new genetic program
- Limited use of genomic tools and breeding process

**Market Quality & Postharvest**

**High Cumulative Yields**

Santa Maria

Prunedale

**Day Neutral Pipeline**

Total
Short-Day
Day-Neutral



**Current Focus:**
- Germplasm knowledge
- Octoploid genomics tools
- Building foundation for genomic breeding

**Future Goals:**
- Increased fruit quality
- Genomic prediction
- Leveraging untapped diversity to improve flavor

---

## UC Davis Strawberry Breeding Program

The last 30 years…

- Successful focus on yield and marketable quality for California field production
- Day neutral breeding has increased potential
- Flavor is valued, but not historically an important trait for growers
- Virtually no introduction of new genetic program
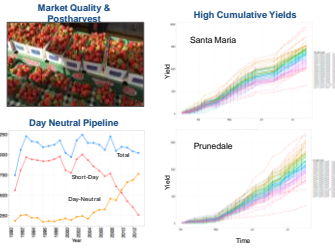- Limited use of genomic tools and breeding process

**Market Quality & Postharvest**

**High Cumulative Yields**

Santa Maria

Prunedale

**Day Neutral Pipeline**

Total
Short-Day
Day-Neutral

### A short history of strawberry genomic breeding resources

2011 – *Fragaria vesca* genome assembly (Shulaev et al)
- diploid progenitor genome, translational system for non-climacteric fruit genetic research

Wide variety of published marker genotyping strategies – SSRs, target capture, DArtSeq, ddRADseq

2014 – First pass at an octoploid assembly (Hirakawa et al)
- highly fragmented

2015 – Affymetrix iStraw90 Axiom array (Bassil et al)
- diploid-segregating markers
- iStraw35 spin-off, high quality subset

## A short history of strawberry genomic breeding resources

2011 – *Fragaria vesca* genome assembly (Shulaev et al)
- diploid progenitor genome, translational system for non-climacteric fruit genetic research

Wide variety of published marker genotyping strategies – SSRs, target capture, DArtSeq, ddRADseq

2014 – First pass at an octoploid assembly (Hirakawa et al)
- highly fragmented

2015 – Affymetrix iStraw90 Axiom array (Bassil et al)
- diploid-segregating markers
- iStraw35 spin-off, high quality subset

2016-present – push for a high-quality octoploid genome assembly
- collaboration at UC Davis, Michigan State
- multiple other groups

### Challenges for Octoploids

Total subgenomic ambiguity, fails to account for inter-species divergence and gene content across subgenomes

Low quality

<17K pass QC for 1500 UC Davis genotypes, subgenome bias, large physical gaps, ambiguous locations

---

## PacBio Assembly – Finding the Right Parameters

**Primary goal for assembly: Zero subgenome collapsing with haploid-level subgenomes**

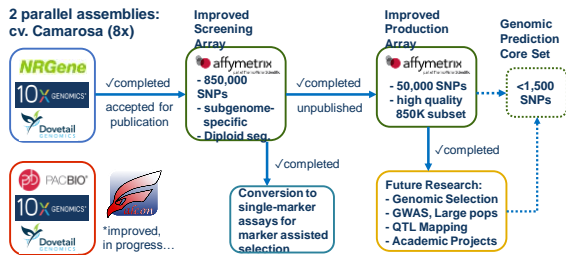**How can we estimate and avoid subgenome collapses during the assembly process?**

Approach 1. How many contigs stack up against the *F. vesca* genome?

*F. vesca* chromosome
PB contigs

**Approach 2. How many contigs align to conserved single-copy *Fragaria vesca* genes?**

*F. vesca* chromosome
*F. vesca* SCG
PB contigs

---

## Building Resources for Molecular Breeding

**2 parallel assemblies: cv. Camarosa (8x)**

NRGene
10x GENOMICS
Dovetail

√completed
accepted for publication

PacBio
10x GENOMICS
Dovetail

*improved, in progress…

**Improved Screening Array**
affymetrix
- 850,000 SNPs
- subgenome-specific
- Diploid seg.

√completed
unpublished

√completed

**Conversion to single-marker assays for marker assisted selection**

**Improved Production Array**
affymetrix
- 50,000 SNPs
- high quality 850K subset

√completed

**Future Research:**
- Genomic Selection
- GWAS, Large pops
- QTL Mapping
- Academic Projects

**Genomic Prediction Core Set**

<1,500 SNPs

---

## PacBio Assembly – Finding the Right Parameters
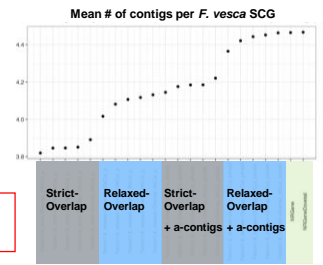
**Key Parameters:**

**read overlap** – aligner thresholds for error-correction, pre-assembly of long-reads
- **strict = 4800, 2400**
- **relaxed = 2500, 2000**

**length cutoff** – minimum length of input reads for pre-assembly, final assembly

**inclusion of alternate contigs**

**Strategy:**
Prioritize haplotype separation to minimize inter-subgenome collapses, address within-subgenome haplotype separation downstream (goal=haploid)
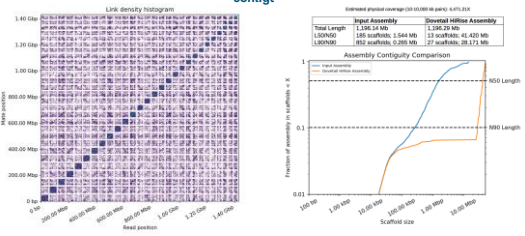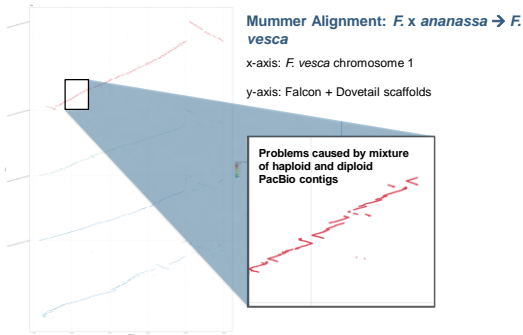
**Mean # of contigs per *F. vesca* SCG**

Strict-Overlap | Relaxed-Overlap | Strict-Overlap + a-contigs | Relaxed-Overlap + a-contigs

---

---

## PacBio Assembly – Dovetail HiC Scaffolding

**Dovetail scaffolding of Falcon assembly improved contiguity**

Link density histogram

Estimated physical coverage (10:10,000 kb pairs): 4,071.21X

| | Input Assembly | Dovetail HiRise Assembly |
|---|---|---|
| Total Length | 1,196.14 Mb | 1,196.29 Mb |
| L50N50 | 185 scaffolds, 1,544 Mb | 13 scaffolds, 41.420 Mb |
| L90N90 | 852 scaffolds, 0.285 Mb | 27 scaffolds, 28.171 Mb |

Assembly Contiguity Comparison

## Mummer Alignment: *F.* x *ananassa* → *F. vesca*

x-axis: *F. vesca* chromosome 1

y-axis: Falcon + Dovetail scaffolds



**Problems caused by mixture of haploid and diploid PacBio contigs**

## NRGene and PacBio Genome – Assembly Metrics
(complete)    (incomplete)

Assembled Genome Lengths Close to Predicted Genome Size

| | NRGene + | PacBio + |
|---|---|---|
| Scaffolds | | |
| Total | 805,488,706 | 806,643,951 |
| # Seqs | 28 | 29 |
| Min | 20,034,018 | 3,437,851 |
| Max | 43,627,644 | 36,254,589 |
| n50 | 27,809,139 | 28,884,336 |
| n90 | 24,073,015 | 23,911,622 |
| Contigs | | |
| Total | 800,292,989 | 806,592,138 |
| # Seqs | 6,042 | 1,263 |
| Min | 76 | 1,082 |
| Max | 3,470,356 | 11,485,031 |
| n50 | 520,757 | 2,325,566 |
| n90 | 103,717 | 426,070 |
| # Gaps | 6,014 | 1,234 |

Chromosome-Scale Contiguity in Both Assemblies
28 Pseudo-Molecules

Gene Content = 42%

Significantly Longer Contigs & Fewer Gaps in the PACBIO Assembly

**New Challenge:**
**Reducing Assemblies to Haploid Level without Collapsing Subgenomes**



## NRGene and PacBio Genome – Genes and Repetitive Elements

**Repetitive DNA Content**



**BUSCO analysis – Complete predicted plant gene content**
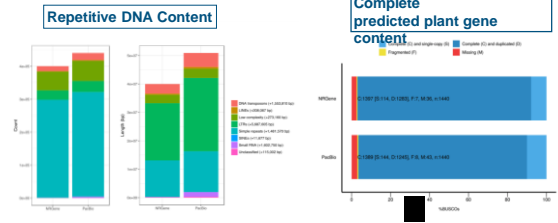


## Constructing a Haploid Assembly

1. Used HiC data to cluster chromosome-specific contig pools
2. Performed haplotype merging within 28 chr-specific pools

**After HaploMerging**
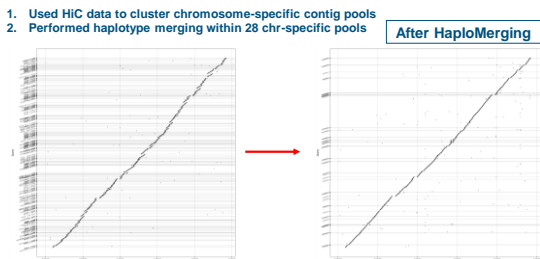


## California Beach Strawberry Genome

*Fragaria chiloensis* spp. *lucida* (Del Norte)

- Native California beach strawberry
- Characteristics
  - High salt tolerance
  - Small soft fruit with unique aroma
  - Pathogen resistance
- Genomic and breeding data collected
  - 70X PacBio SEQUEL data → **NEW assembly**
  - 160X PE 2x250 Illumina data
  - Mapping population (Camarosa x Del Norte)
    - 200 progeny
    - Illumina sequenced
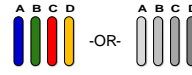    - Phenotypesd for multiple traits

Camarosa (mother)

Del Norte (father)
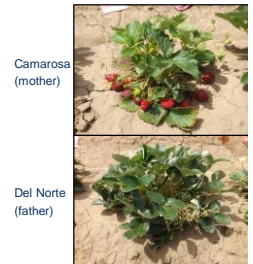
## WGS-based Octoploid Mapping

F. chilensis mapping with Illumina data:
- 28 chromosome-scale linkage groups
- >2 million genome+map anchored markers
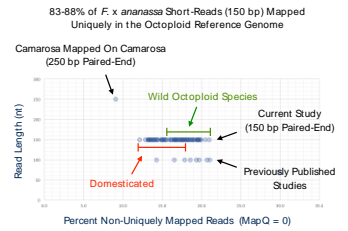- High-confidence chromosome scaffolding
- mapping multiple traits of interest
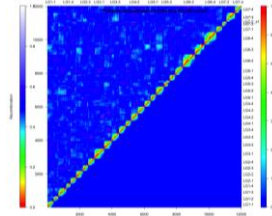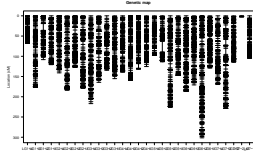
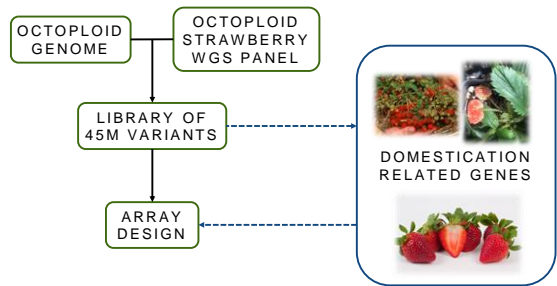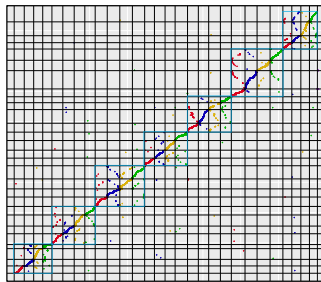

## Challenge for subgenome-specific alignment?



-OR-

**Sequence Panel**
- 145 octoploids at present
- Illumina sequence (PE 150nt)
- Used to identify genomic variants for SNP array
- ~~Mapped to NRGene~~
- **High rates of subgenome specific read mapping**



83-88% of *F. x ananassa* Short-Reads (150 bp) Mapped Uniquely in the Octoploid Reference Genome

Camarosa Mapped On Camarosa (250 bp Paired-End)

Wild Octoploid Species

Current Study (150 bp Paired-End)

Previously Published Studies

Domesticated

Percent Non-Uniquely Mapped Reads (MapQ = 0)

Read Length (nt)

---

*F. chiloensis* genetic map shows high concordance with *F.* x *ananassa* genome





OCTOPLOID GENOME

OCTOPLOID STRAWBERRY WGS PANEL

LIBRARY OF 45M VARIANTS

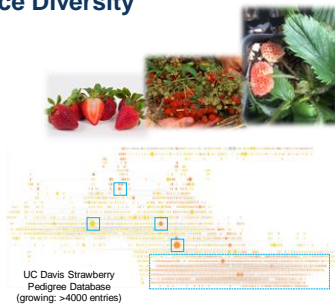ARRAY DESIGN

DOMESTICATION RELATED GENES
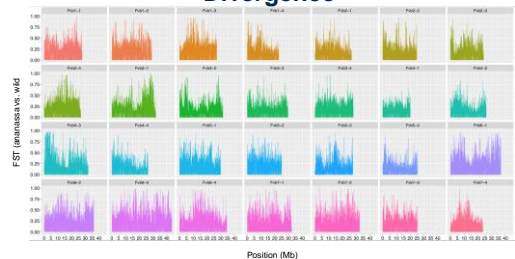
---

## Octoploid Sequence Diversity Panel
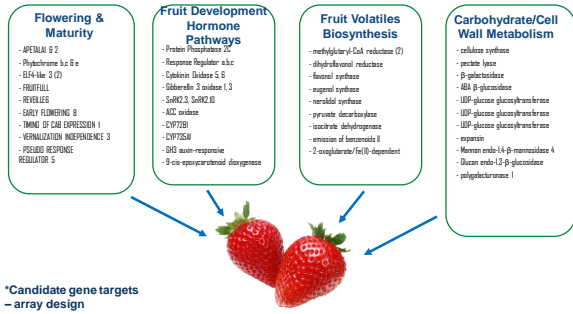
Assessing genetic diversity of the octoploid genome

**Selection Targets:**
1) Capture UC Davis germplasm diversity
2) Capture alleles from founders, high pedigree contribution heirlooms
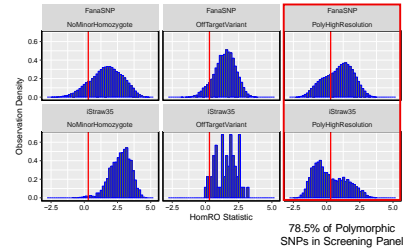3) Capture progenitor species diversity



UC Davis Strawberry Pedigree Database (growing: >4000 entries)

## Wild vs. Domesticated Population Divergence



FST (ananassa vs. wild)

Position (Mb)

4

## Flowering & Maturity
- APETALA1 G 2
- Phytochrome b,c & e
- ELF4-like 3 (2)
- FRUITFULL
- REVEILLE6
- EARLY FLOWERING 8
- TIMING OF CAB EXPRESSION 1
- VERNALIZATION INDEPENDENCE 3
- PSEUDO RESPONSE REGULATOR 5

## Fruit Development Hormone Pathways
- Protein Phosphatase 2C
- Response Regulator a,b,c
- Cytokinin Oxidase 5, 6
- Gibberellin 3 oxidase 1, 3
- SnRK2.3, SnRK2.10
- ACC oxidase
- CYP72B1
- CYP73SA1
- GH3 auxin-responsive
- 9-cis-epoxycarotenoid dioxygenase

## Fruit Volatiles Biosynthesis
- methylglutaryl-CoA reductase (2)
- dihydroflavonol reductase
- flavanol synthase
- eugenol synthase
- nerolidol synthase
- pyruvate decarboxylase
- isocitrate dehydrogenase
- emission of benzenoids II
- 2-oxoglutarate/Fe(II)-dependent

## Carbohydrate/Cell Wall Metabolism
- cellulase synthase
- pectate lyase
- β-galactosidase
- ABA β-glucosidase
- UDP-glucose glucosyltransferase
- UDP-glucose glucosyltransferase
- UDP-glucose glucosyltransferase
- expansin
- Mannan endo-1,4-β-mannosidase 4
- Glucan endo-1,3-β-glucosidase
- polygalacturonase 1

**Candidate gene targets – array design**

## Increased Likelihood of Subgenome Specific Probe Hybridization (HomRO >= 0.30)



78.5% of Polymorphic SNPs in Screening Panel

## Development of Marker Array Panel
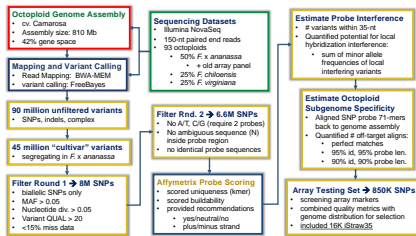


## Overview: Strawberry SNP Selection

✓ PHASE 1: **Identify and select candidate marker probes**
- based on variant calling in sequenced octoploid panel
- populated 850K screening array
- Selection priority
  - in/around genes (600K)
  - genome coverage (250K)
  - allele diversity in domesticated germplasm
  - low probe interference

✓ PHASE 2: Test probes on Affymetrix screening array
- 446,644 AxiomQC-passing polymorphic markers
  - 78.5% polymorphic hi-resolution (PHR)
  - 12,624 passing iStraw35 markers (>10K anchored)

✓ PHASE 3: Select markers to populate production array
- 49K probe capacity
- Numerous criteria: technical quality, diversity, coverage

Marker Analysis – Areas of Interest
- Genetic Diversity
  - Heterozygosity, allelic diversity across cultivated sub-populations
- Mapping quality
  - 2 mapping populations on screening array (n=48)
- Major QTL regions represented
- Candidate genes
  - Fruit texture and quality
  - Volatile biosynthesis
  - Hormone-mediated fruit development
  - Flowering and maturity
  - R genes
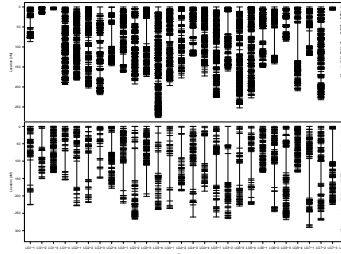- Physical genome coverage

## Genetic Diversity Across Programs
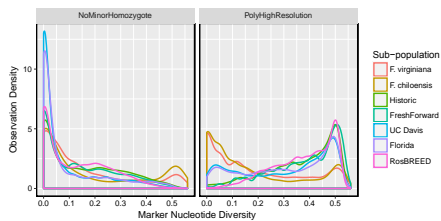


## Marker Genome Distribution
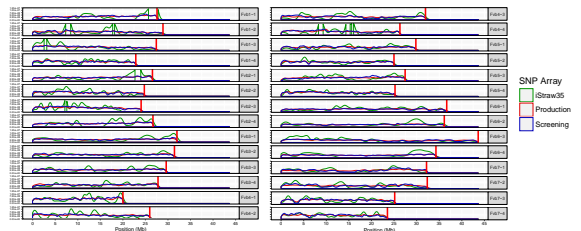
Early Array Applications

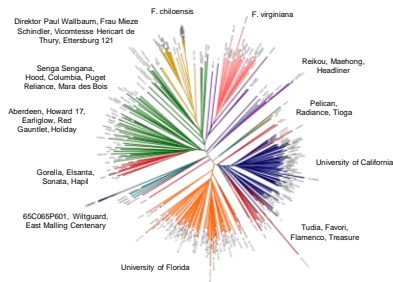## Genetic Mapping (49K array)



Camarosa (maternal) map
*F.* x *ananassa*
~11k QC-passing het x hom SNPs

~1.8k QC-passing het x het SNPs

Del Norte (paternal) map
*F. chiloensis* (wild)
~2.1k QC-passing hom x het SNPs
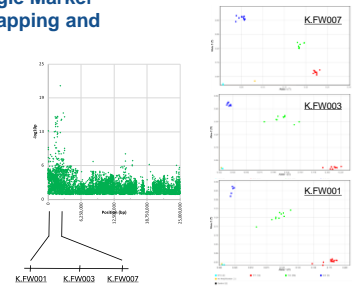
## Surveying Octoploid Germplasm Diversity



## Conversion to Single Marker Assays for Fine Mapping and Breeding
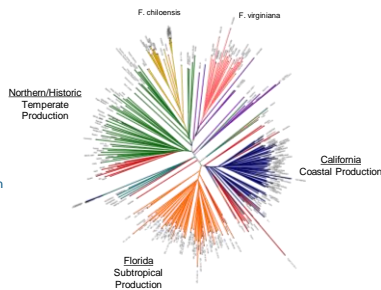
Fw1 Fusarium Resistance

- 850K array GWAS identified Fvb2-4 locus
- Subgenome-specific KASP assays flanking Fw1
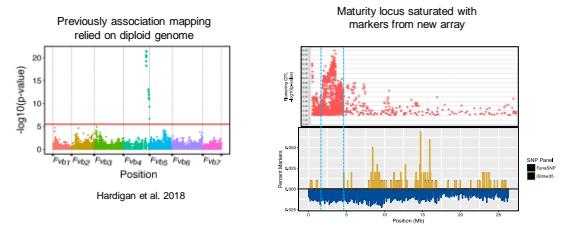
Nicolas Cobo



K.FW001  K.FW003  K.FW007

## Surveying Octoploid Germplasm Diversity
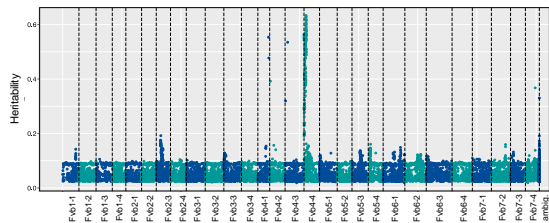
Future objectives: exploring loci that have been selected for performance in different US production environments



Northern/Historic Temperate Production

California Coastal Production

Florida Subtropical Production

## Strawberry Maturity Locus

Previously association mapping relied on diploid genome

Hardigan et al. 2018

Maturity locus saturated with markers from new array

## Strawberry Maturity Locus

## Strawberry Maturity Locus



Future work:
- Conversion of 850K panel markers to KASP assays
- Multiple segregating day neutral populations for fine mapping (2019)

STRAWBERRY BREEDING @ THE UNIVERSITY OF CALIFORNIA
# THANKS FOR LISTENING!





**Current Focus:**
- Germplasm knowledge
- Octoploid genomics tools
- Building foundation for genomic breeding

**Future Goals:**
- Increased fruit quality
- Genomic prediction
- Leveraging untapped diversity to improve flavor