“Genomic resource development and analysis in Apples & Pears”

Dr. Alan E. Yocca for the position of Genomics Data Scientist

Career Background

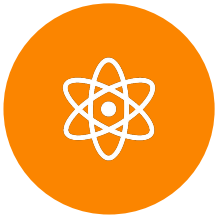
The Pennsylvania State University:
2014 – 2016

Undergraduate Research Assistant - dePamphilis lab

Michigan State University:
2017 – 2022

Graduate Assistant - Edger lab

Dissertation title: “Leveraging Angiosperm Pangenomics to Understand Genome Evolution”



HudsonAlpha Institute for Biotechnology: 2022 – present

Postdoctoral Research Associate - Harkess lab

USDA:
2023 – present

Postdoctoral Research Associate - Honaas lab

Evolution of Conserved No
thaliana
Origin and evolution
genome

Patrick P. Edger[1,2*], Thomas J. Poorten[3], R
Michael R. McKain[5], Ronald D. Smith[6],
Elizabeth I. Alger[1], Kevin A. Bird[1,2],
Avital Brodt[8], Kobi Baruch[8], Thomas
Jeffrey P. Mower[10], Kevin L. Childs[11,12]
and Steven J. Knapp[3*]

Uncovering a miltiradiene biosynthetic gen cluster in the Lamiaceae reveals a dynamic evolutionary trajectory

Received: 8 April 2022
Accepted: 4 January 2023

Abigail E. Bryson[1,7], Emily R. Lanier[1,7], Kin H. Lau[2,3], John P. Hamilton[2,4],
Brieanne Vaillancourt[2,4], Davis Mathieu[1], Alan E. Yocca[2,5], Garret P. Mille

Genome of the North American wild apple species *Malus angustifolia*

Ben N. Mansfeld, Shujun Ou, Erik Burchard, Alan Yocca, Alex Harkess, Ben Gutierrez, Steve van Nocker, Lisa Tang, Christopher Gottschalk

Machine learning appro
genes in pangenomes

Disease Resistance Genetics and Genomics in Octoploid Strawberry

Christopher R. Barbey,*[,†,1] Seonghee Lee,[‡] Sujeet Verma,[‡] Kevin A. Bird,[§,**] Alan E. Yocca,[††]
Whitaker,[‡] and Kevin M. Folta*[,†]

Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff

Robert VanBuren[1,2,9✉], Ching Man Wai[1,2,9], Xuewen Wang[3,9], Jeremy Pardo[1,2,4], Alan E. Yocca[1,4],
Hao Wang[3], Srinivasa R. Chaluvadi[3], Guomin Han[3], Douglas Bryant[5], Patrick P. Edger[1],
Joachim Messing[6], Mark E. Sorrells[7], Todd C. Mockler[5], Jeffrey L. Bennetzen[3] & Todd P. Michael[8✉]

n-genome dynamics of

Alan E. Yocca, Adrian Platts, Elizabeth Alger, Qin Qiao, Patrick P. Edger, Li Xue, La Qiong, Jie Lu[a], Yichen Zhang[a], Qiang Cao[a], Alan E. Yocca[b,e],
Michal Babinski[1], Maria Magallanes-Lundback[1], Philip, Adrian E. Platts[b], Steven J. Knapp[f], Marc Van Montagu[g,h,2], Yves Van de Peer[g,h,i,j,2], Jiajun Lei[c,2], and Ticao Zhang[k],
Nick W. Albert[11], Sara Montanari[12], Nicholi Vorsa[13],
Massimo Iorizzo[4,16] and Patrick P. Edger[1,3,17,*]

Postdoc Projects:

- Chromosome scale haplotype phased genome assemblies with PacBio HiFi
- *Malus* crop wild relatives assembly and annotation
- Detecting recombination events using *k*-mers
- When is the perfect time to pick an apple or pear?
- Current progress in graph-based pangenomes

Apples & Pears

- *Malus x domestica* (Apple) and *Pyrus communis* (European Pear)
- Pome fruits in Rosaceae
- Diverged ~8 MYA
- Share ancient WGD ~40 MYA
- Genome size ~600-750Mb
- 17 chromosomes
- >50% of genes found in duplicate from last WGD

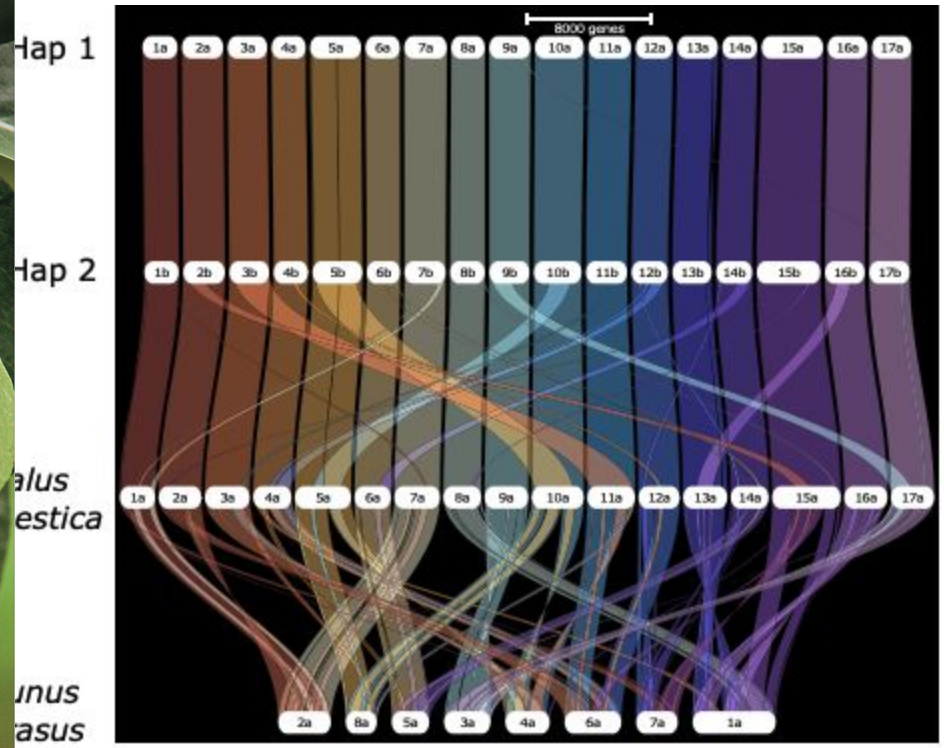
Postdoc Projects:

- **Chromosome scale haplotype phased genome assemblies with PacBio HiFi**
- *Malus* crop wild relatives assembly and annotation
- Detecting recombination events using *k*-mers
- When is the perfect time to pick an apple or pear?
- Current progress in graph-based pangenomes

American Campus Tree Genomes

- “American Campus Tree Genomes Project pushes equality in genomic science”
- Students sequence, assembly, annotate, and analyze a chromosome-scale genome
- Final project is a figure on a manuscript


G3
Genes | Genomes | Genetics
GSA
Genetics Society of America
MARCH 2024
VOLUME 14 • ISSUE 3
academic.oup.com/g3journal

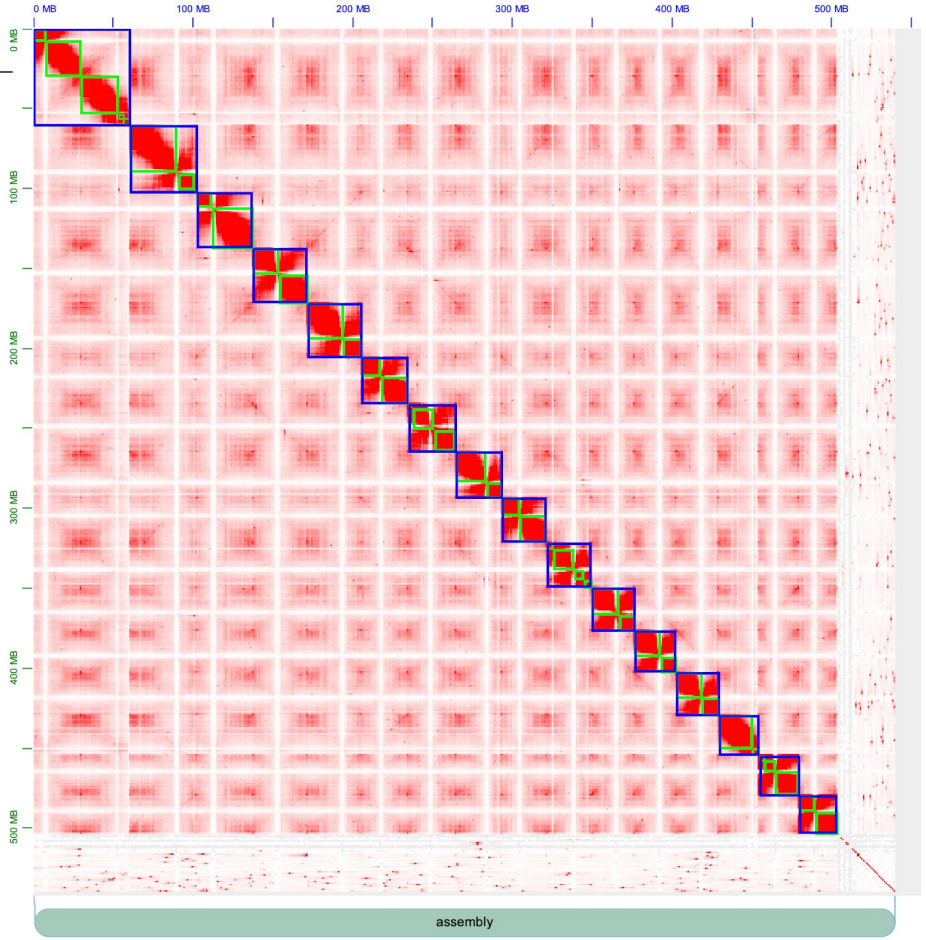

Hap 1
Hap 2
alus
estica
unus
asus
8000 genes
1a 2a 3a 4a 5a 6a 7a 8a 9a 10a 11a 12a 13a 14a 15a 16a 17a
1b 2b 3b 4b 5b 6b 7b 8b 9b 10b 11b 12b 13b 14b 15b 16b 17b
1a 2a 3a 4a 5a 6a 7a 8a 9a 10a 11a 12a 13a 14a 15a 16a 17a
2a 8a 5a 3a 4a 6a 7a 1a

HiFi genome assembly

- ~40x HiFi coverage per haplotype
- ~80x Omni-C coverage per haplotype
- Assembled with hifiasm
- Quality Control:
  - Omni-C contact map
  - Dot plots to existing assemblies and to itself
  - Remap HiFi data to the assembly

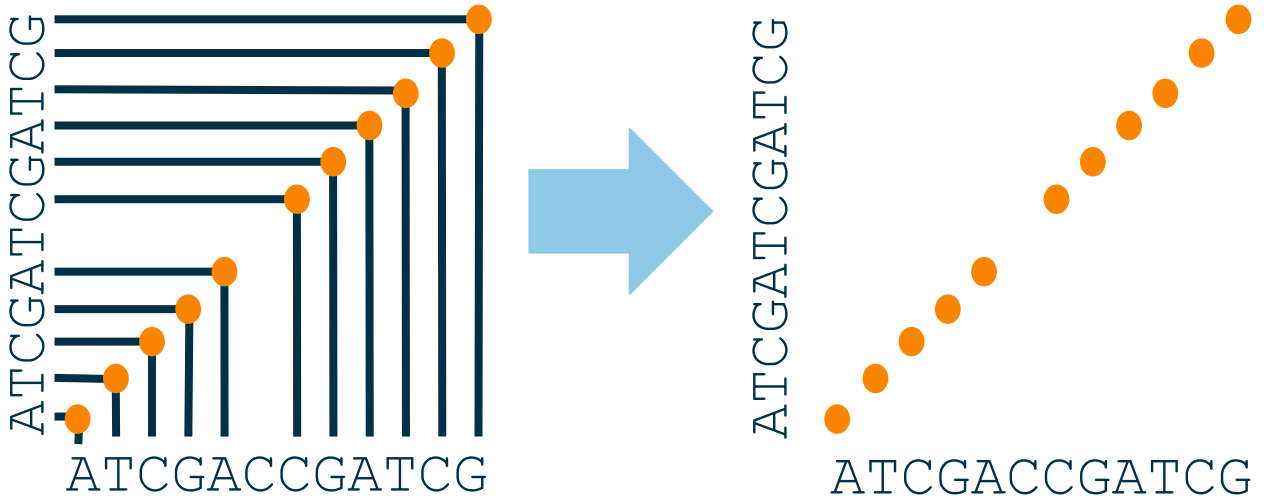
Omni-C contact map

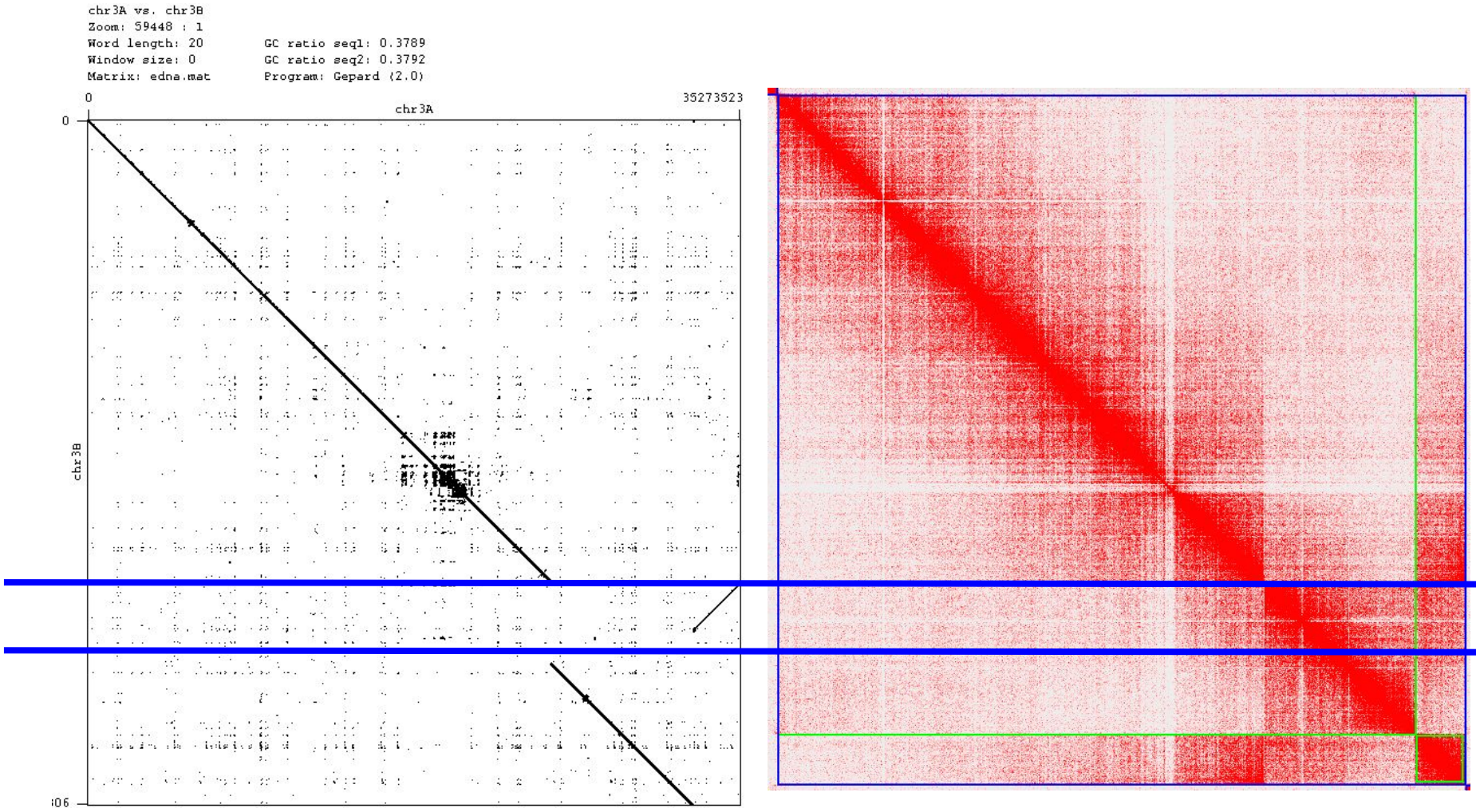
Chromosome fusion, not real!



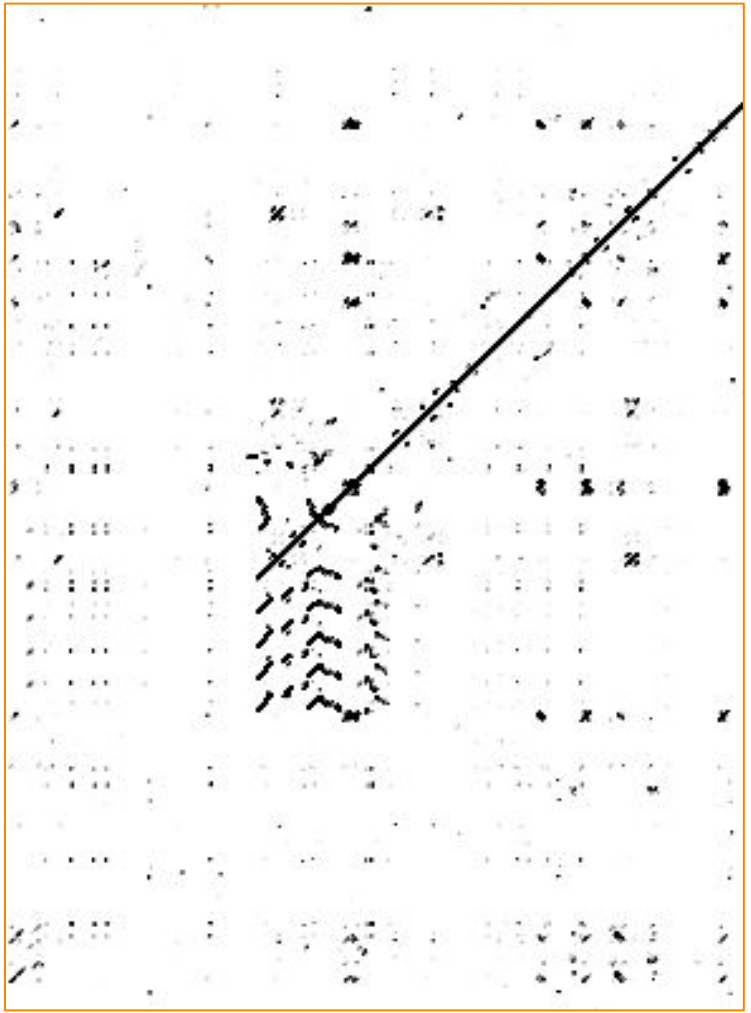
Dot plots

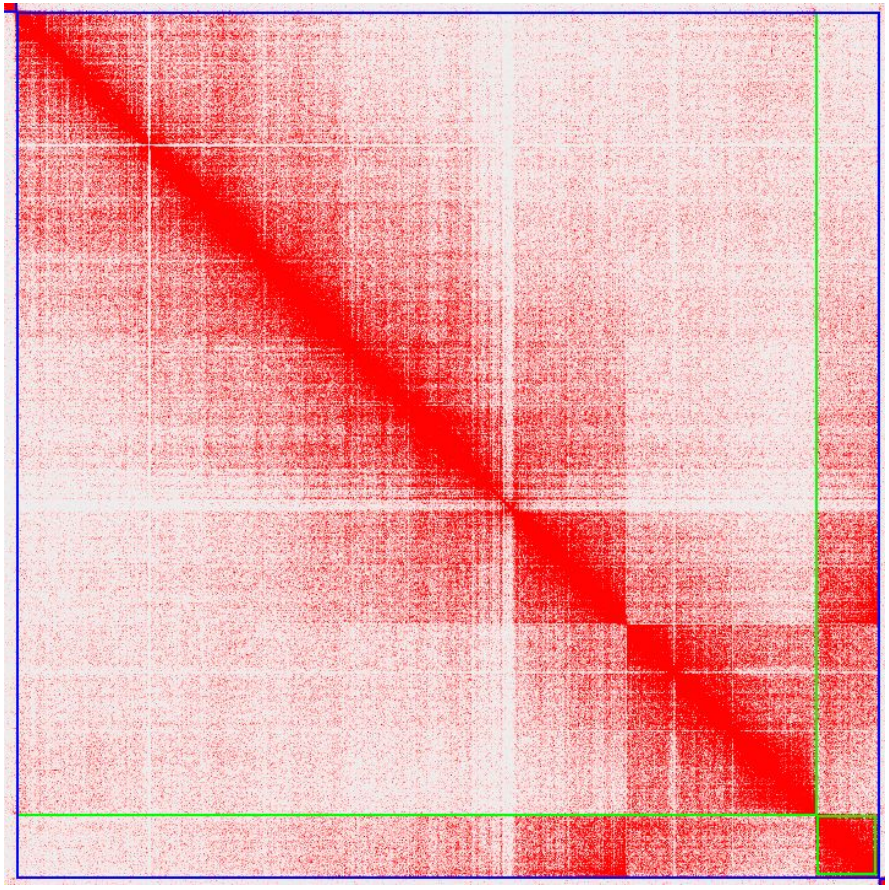
- Sequence alignment in 2-dimensions

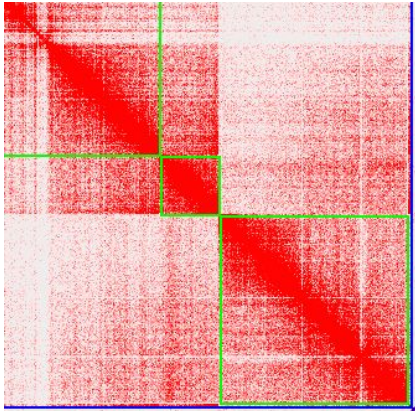


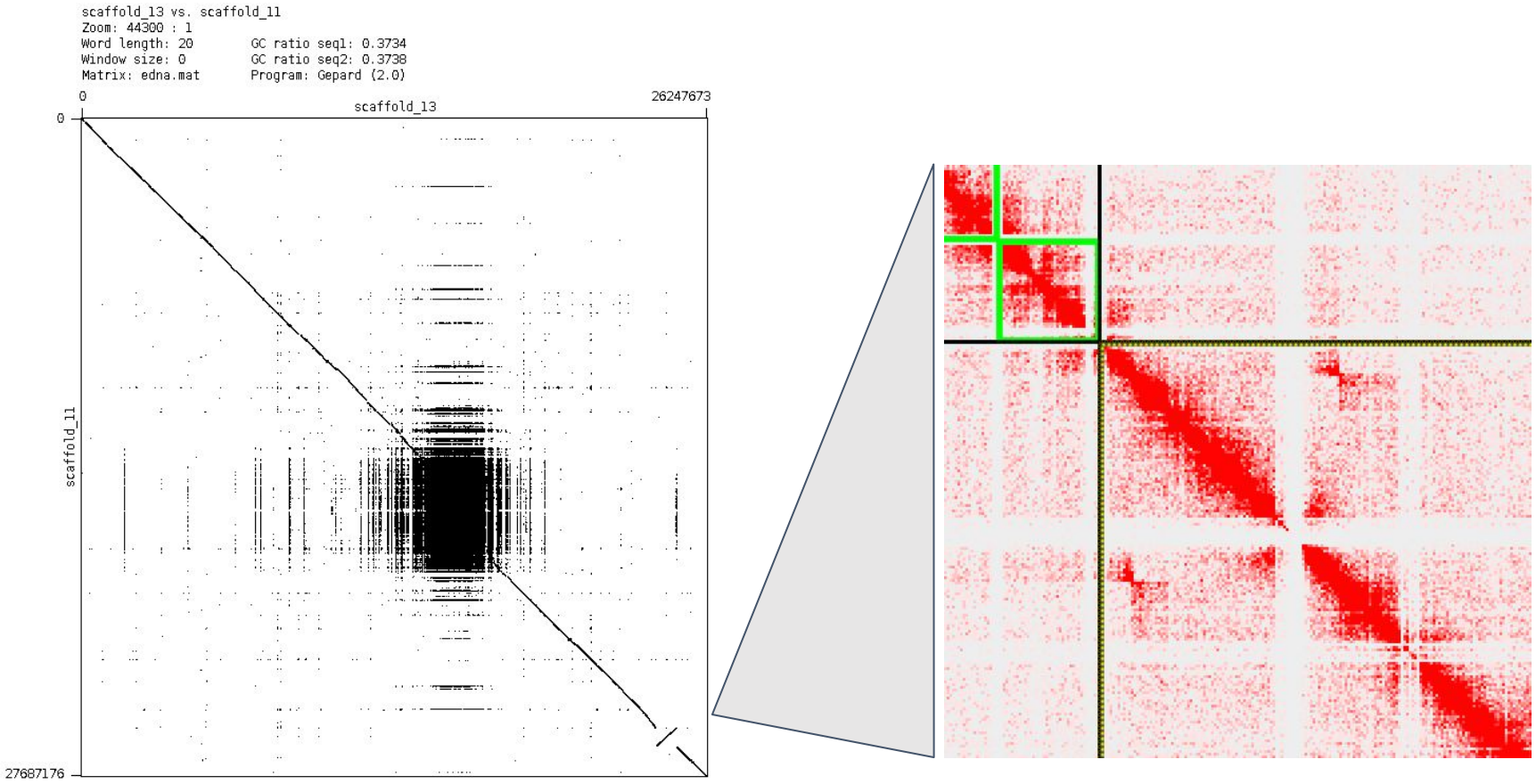

chr3A vs. chr3B
Zoom: 59448 : 1
Word length: 20
Window size: 0
Matrix: edna.mat
GC ratio seq1: 0.3789
GC ratio seq2: 0.3792
Program: Gepard (2.0)
0
35273523
chr3A
0
chr3B

Zoomed in on breakpoints of the misplaced contig

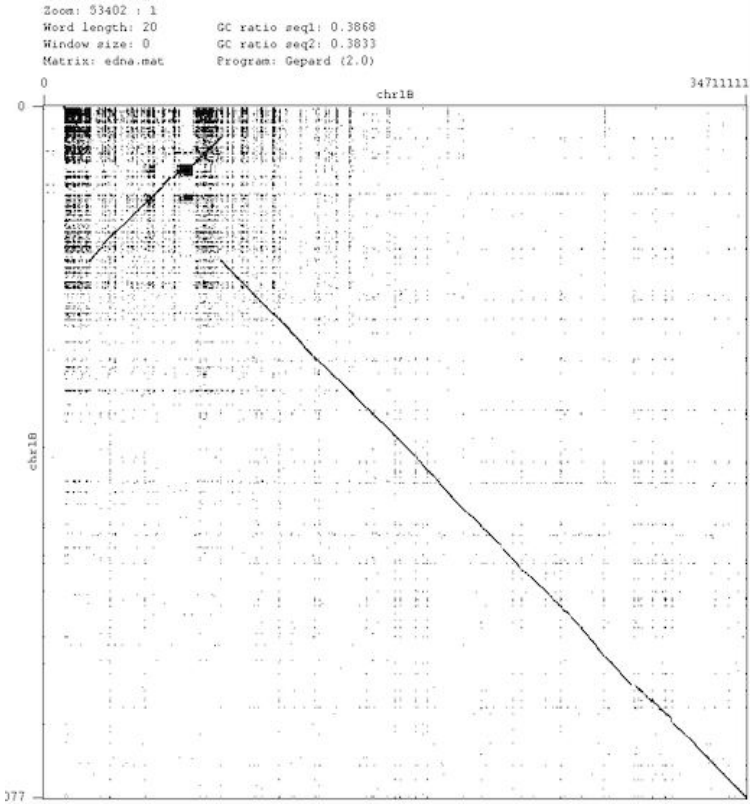


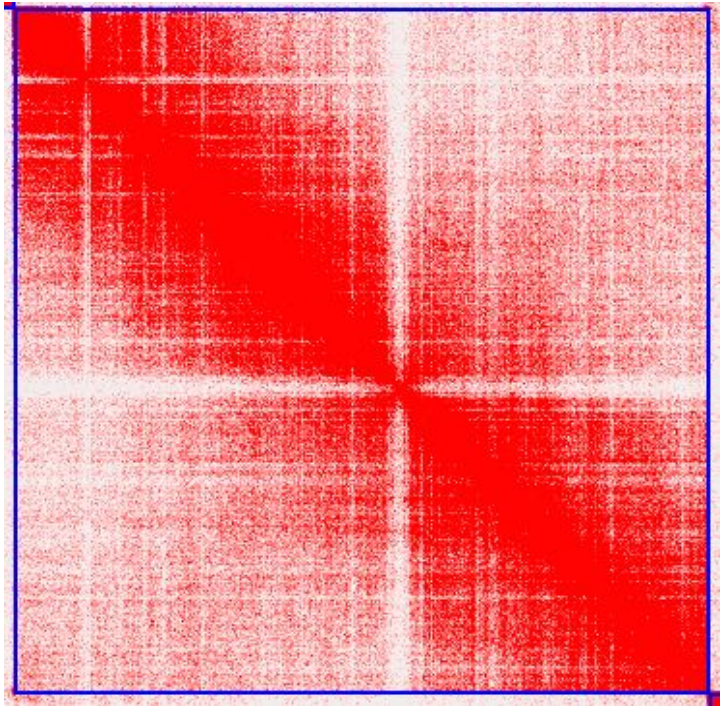




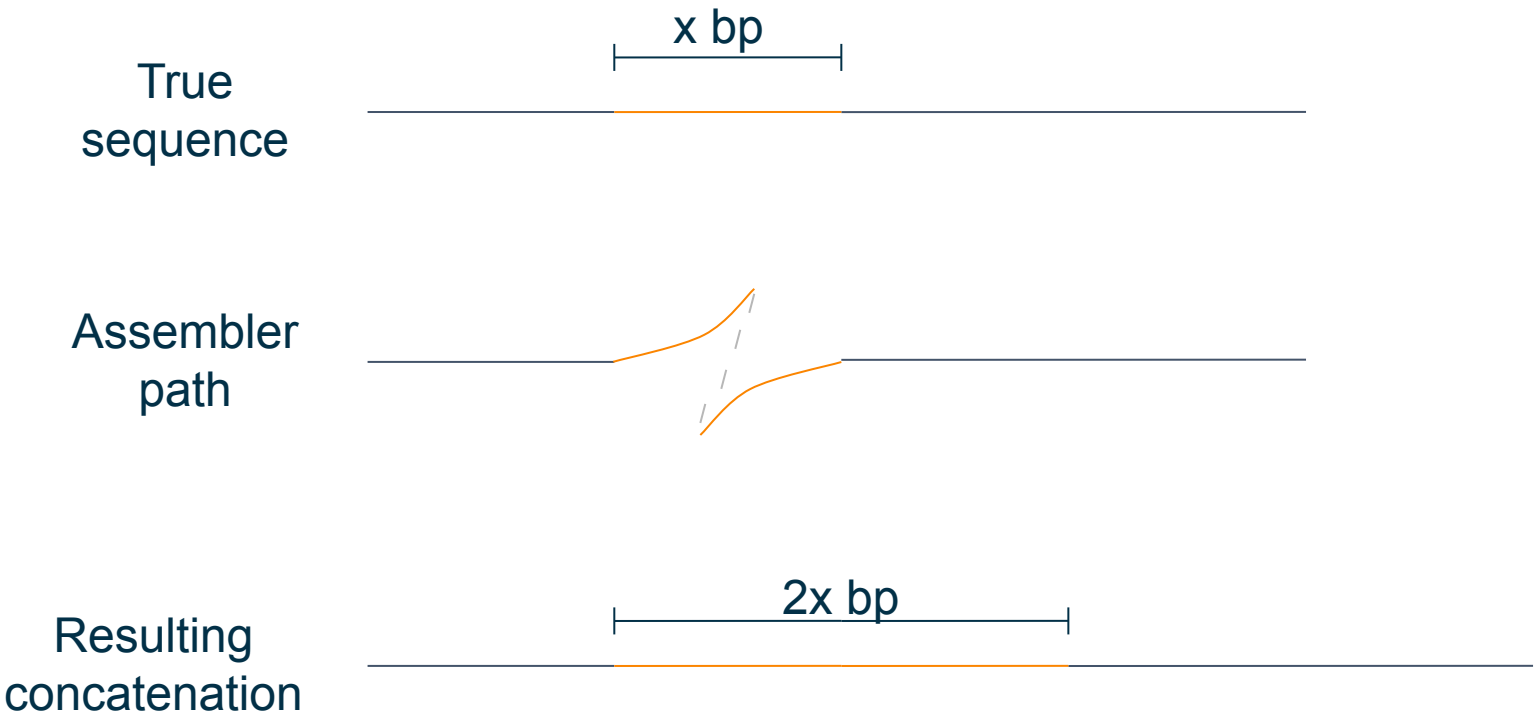

scaffold_13 vs. scaffold_11
Zoom: 44300 : 1
Word length: 20
Window size: 0
Matrix: edna.mat
GC ratio seq1: 0.3734
GC ratio seq2: 0.3738
Program: Gepard (2.0)
0
scaffold_13
26247673
0
scaffold_11
27687176

Some assembly issues are not as straightforward



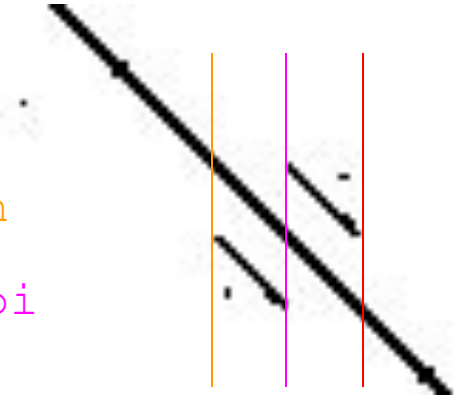


Alternate assembled haplotype

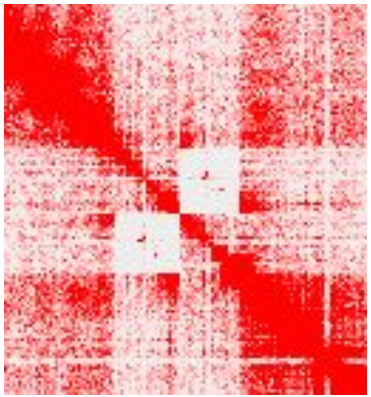


Alternate assembled haplotype

Dot plot “self vs. self”



Omni-C map



Postdoc Projects:

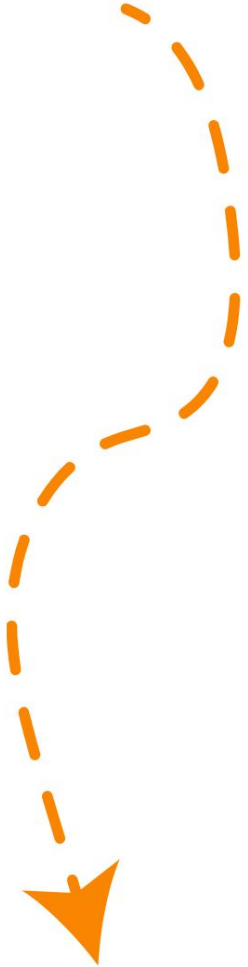
- Chromosome scale haplotype phased genome assemblies with PacBio HiFi
- ***Malus* crop wild relatives assembly and annotation**
- Detecting recombination events using *k*-mers
- When is the perfect time to pick an apple or pear?
- Current progress in graph-based pangenomes

Malus crop wild relatives

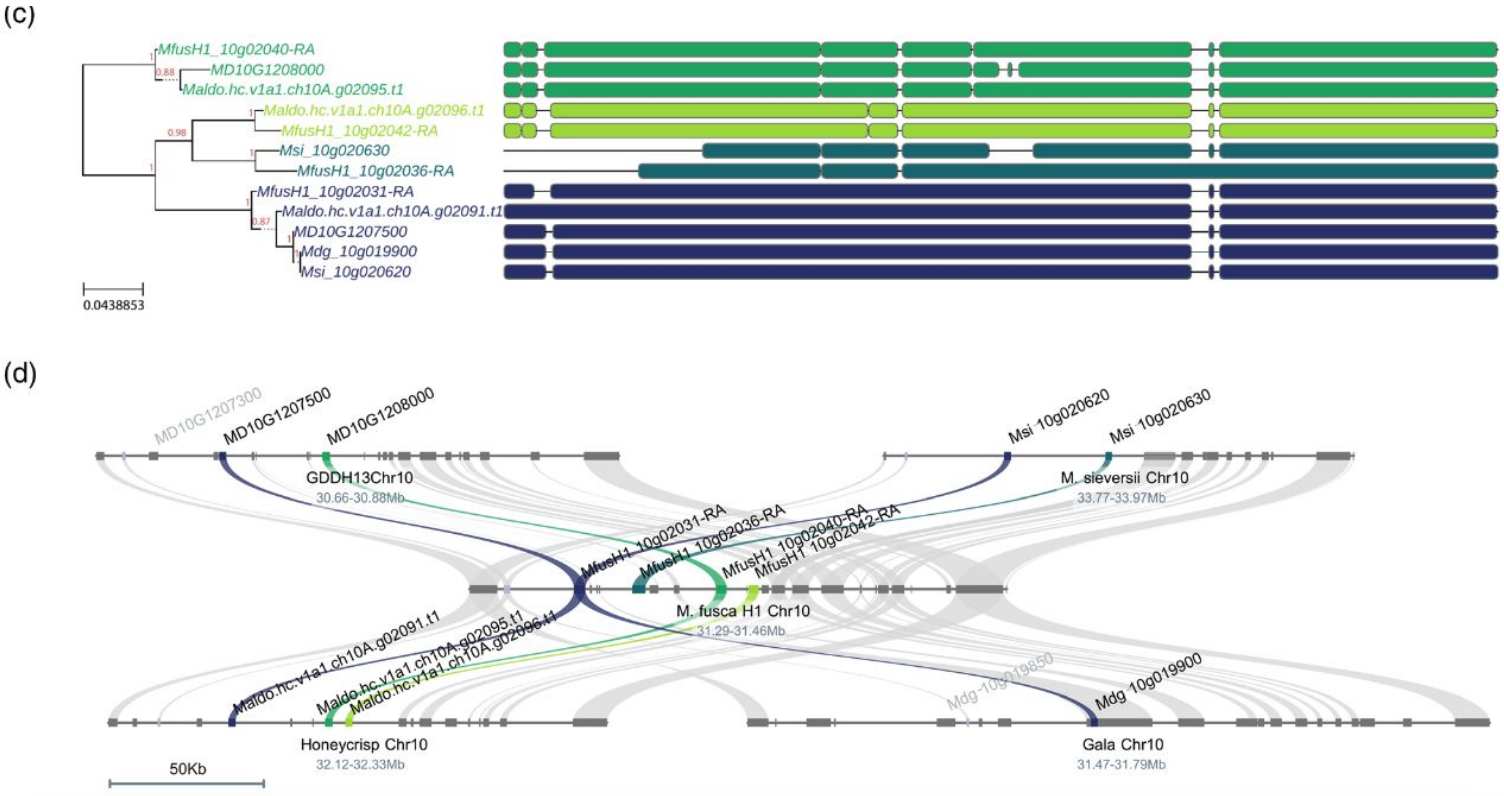
- 4 native North American *Malus* species
- Readily hybridize with domesticated Apple
- Native species exhibit abiotic and biotic resistance
  - Fire blight resistance in *M. fusca*

Annotation workflow

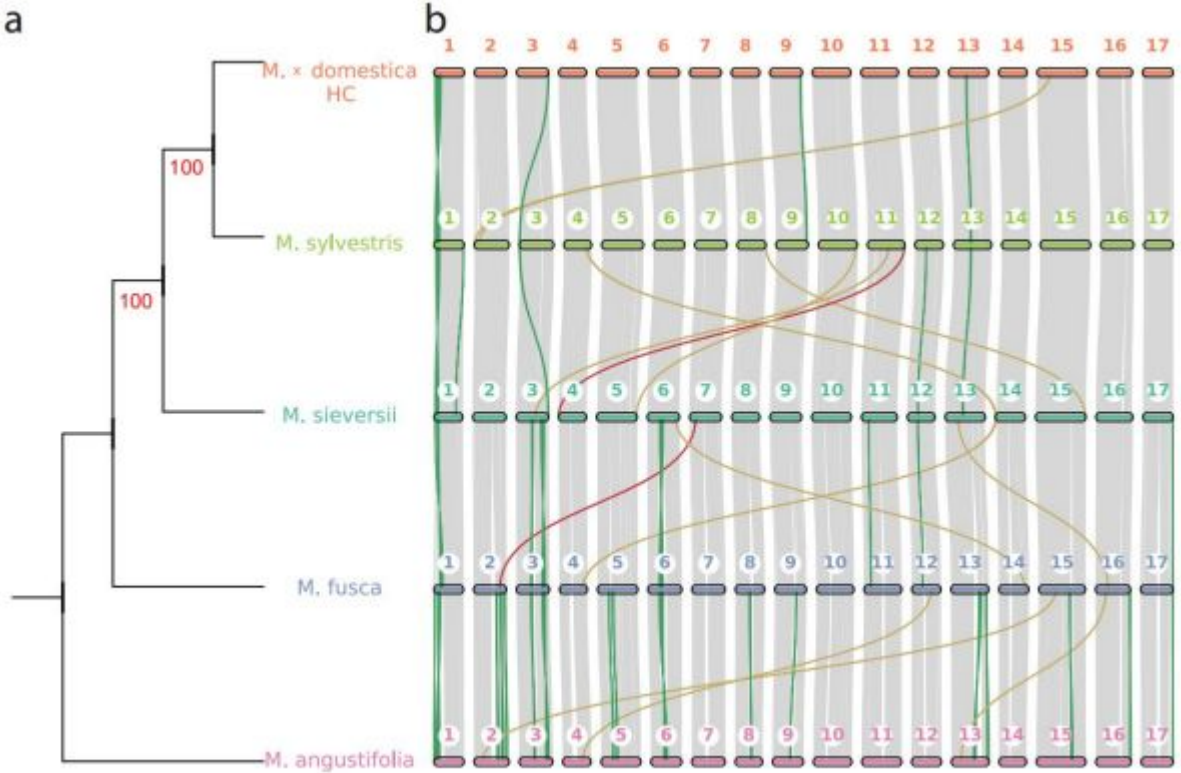
- MAKER pipeline
- Evidence-based annotation
  - Arabidopsis proteins
  - “Honeycrisp” apple proteins
  - Available genotype-specific RNA-seq of diverse tissue types
- Two-rounds of *ab*-initio gene prediction with AUGUSTUS and SNAP
- Quality control
  - Gene length distributions
  - BUSCO
  - Exon count distributions
  - Genome-browser



Malus fusca Fire Blight resistance locus



Malus angustifolia



Postdoc Projects:

- Chromosome scale haplotype phased genome assemblies with PacBio HiFi
- *Malus* crop wild relatives assembly and annotation
- **Detecting recombination events using *k*-mers**
- When is the perfect time to pick an apple or pear?
- Current progress in graph-based pangenomes

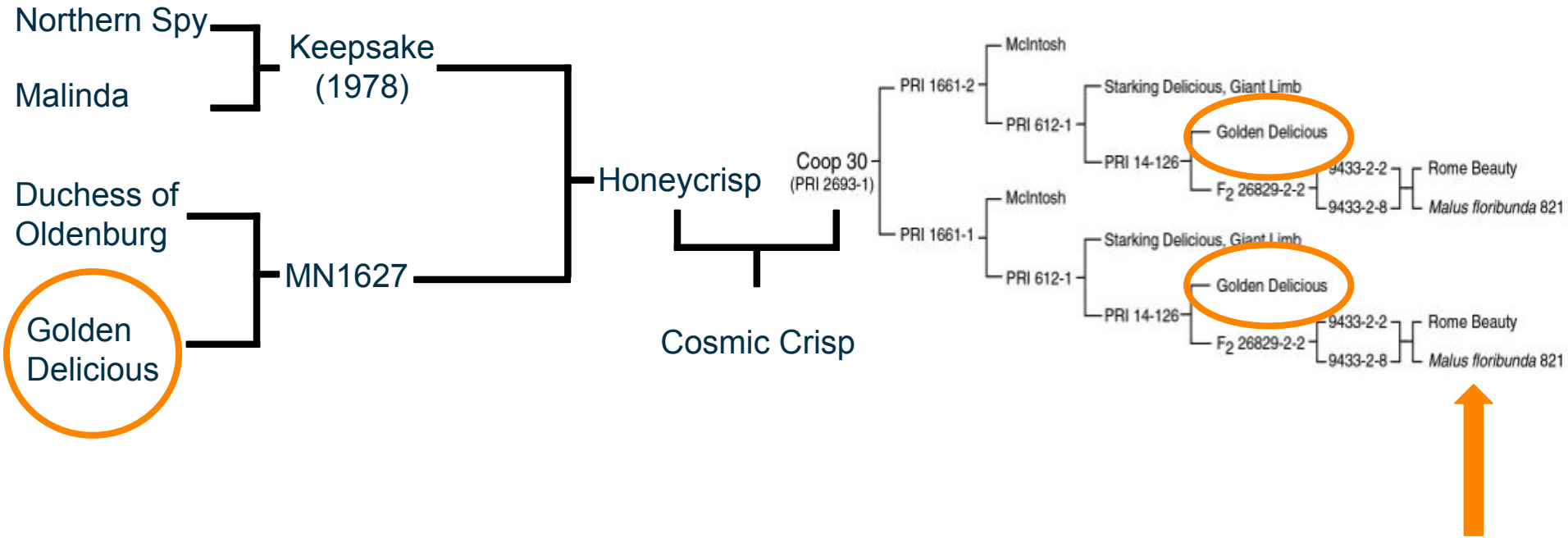
Experimental System

“Honeycrisp”

“Enterprise”

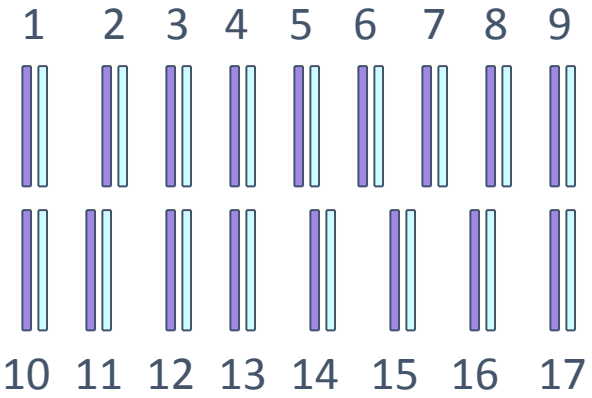
“Cosmic Crisp”

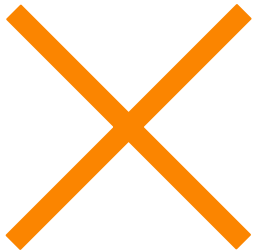
Experimental System



Experimental System

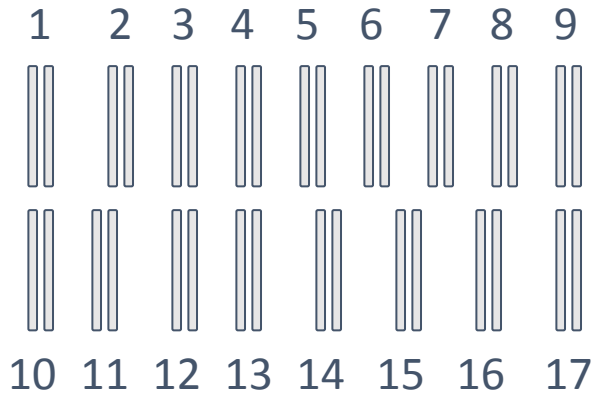
“Honeycrisp”





“Enterprise”

“Cosmic Crisp”



??

Which “Cosmic Crisp” chromosome was inherited from which parent?

- We have a phased “Cosmic crisp” and a phased “Honeycrisp” (HC from here on) which one parent of “Cosmic crisp” (CC from here on)
- Lets use kmers to find which copy in CC came from HC and which from the other parent “Enterprise”

Workflow

- Use meryl to find kmers specific to one CC haplotype not found in the other CC haplotype
- Intersect these with kmers found in both HC haplotypes
- The chromosome that came from HC should have a higher abundance of kmers shared than the other haplotype

Results based on *k*-mer abundance

Chromosome	HapA-specific shared *k*-mers	HapB-specific shared *k*-mers	Difference	Ratio HapA/HapB
1	2051610	5471251	-3419641	0.3749800548
2	3077972	9635778	-6557806	0.3194316017
3	7550411	3455526	4094885	2.185025087
4	6170561	1835473	4335088	3.361836976
5	2910988	6149547	-3238559	0.4733662496
6	6958204	4067368	2890836	1.710738738
7	1436727	8390383	-6953656	0.1712349722
8	7203946	3592405	3611541	2.005326794
9	1265578	3941492	-2675914	0.3210910995

Negative value reflect more CC hapB kmers matching HC than hapA kmers

Results based on kmer abundance (cont)

Chromosome	HapA-specific shared kmers	HapB-specific shared kmers	Difference	Ratio HapA/HapB
10	7976761	2412989	5563772	3.305759371
11	2105156	6412403	-4307247	0.3282944007
12	5573204	1887142	3686062	2.953251001
13	2326268	9489397	-7163129	0.2451439222
14	5953204	1372498	4580706	4.337495574
15	9694659	4465069	5229590	2.171222662
16	10090592	3721436	6369156	2.711478042
17	5902270	2390097	3512173	2.469468812

7 chromosome where hapB matches better than hapA: 1,2,5,7,9,11,13

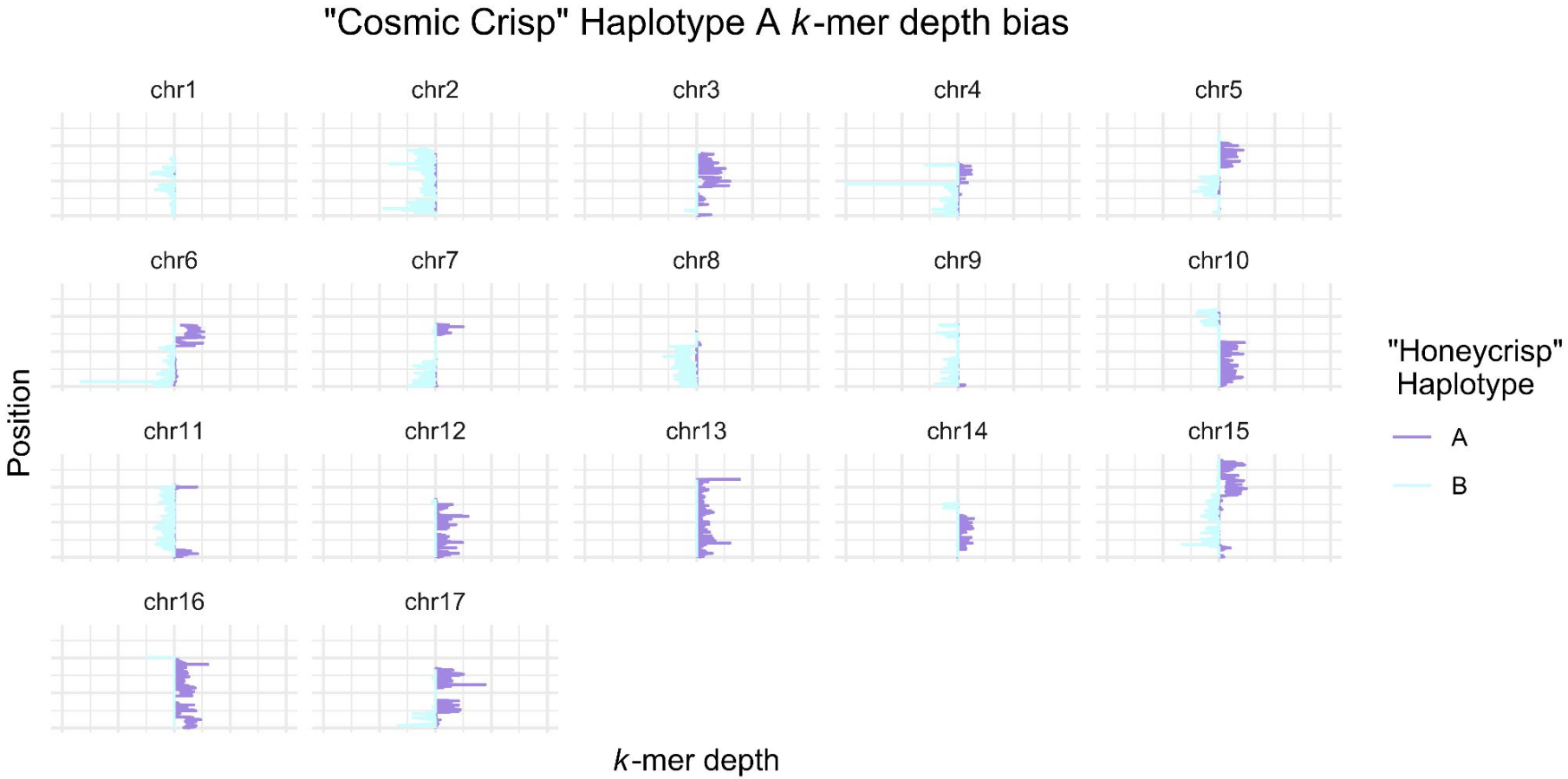
k-mers to detect recombination events

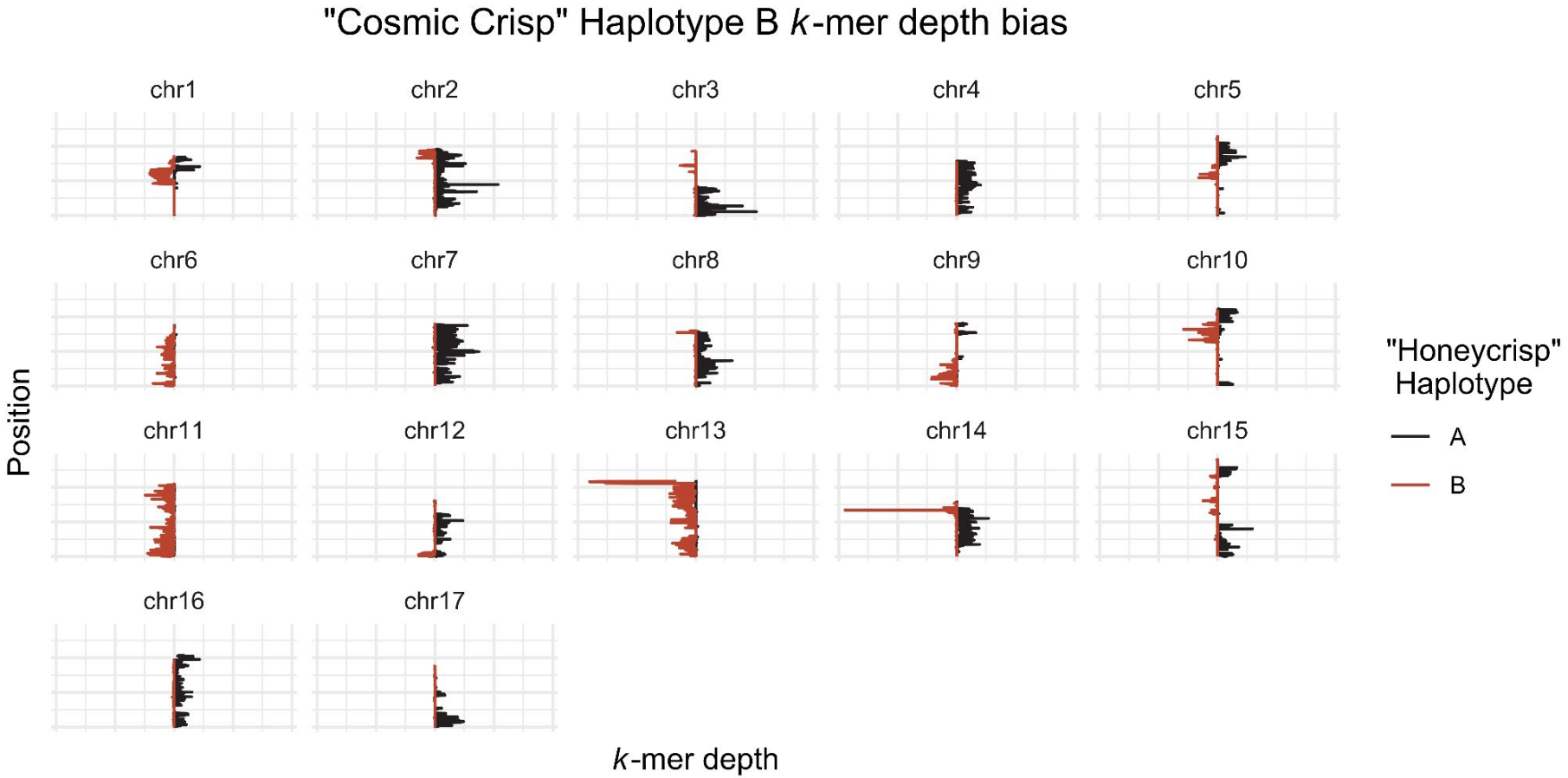
- Now we know which “Cosmic Crisp” haplotype came from which parent
- Can we detect recombination events between parental haplotypes?

Workflow

```
PROCESSING TREE #1 using 10 threads.
  opDifference
    opDifference
      opDifference
        /cluster/home/ayocca/01_Rosaceae_Evo/Malus/01_cosmic_phasing//03_meryl/Mdom_CC_hapA_chr8A.meryl/
        /cluster/home/ayocca/01_Rosaceae_Evo/Malus/01_cosmic_phasing//03_meryl/Mdom_CC_hapB_chr8B.meryl/
      opIntersect
        /cluster/home/ayocca/01_Rosaceae_Evo/Malus/01_cosmic_phasing//03_meryl/Mdom_HC_hapA_chr8A.meryl/
        /cluster/home/ayocca/01_Rosaceae_Evo/Malus/01_cosmic_phasing//03_meryl/Mdom_HC_hapB_chr8B.meryl/
    /cluster/home/ayocca/01_Rosaceae_Evo/Malus/01_cosmic_phasing//03_meryl/Mdom_EP_hapBoth_chr8Both.meryl/
    output to /cluster/home/ayocca/01_Rosaceae_Evo/Malus/01_cosmic_phasing//03_meryl/Mdom_CC_hapA_chr8A.meryl.diff_hapB.diff_EP/
```

- k-mers specific to one haplotype
- Remove any k-mers present in the other parent
- Remove any k-mers not present in the parent of interest
- Map k-mers back to the parent of interest for only exact matches


"Cosmic Crisp" Haplotype A k-mer depth bias
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
Position
k-mer depth
"Honeycrisp" Haplotype
A
B


"Cosmic Crisp" Haplotype B k-mer depth bias
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
Position
k-mer depth
"Honeycrisp" Haplotype
A
B

Motivating Questions:

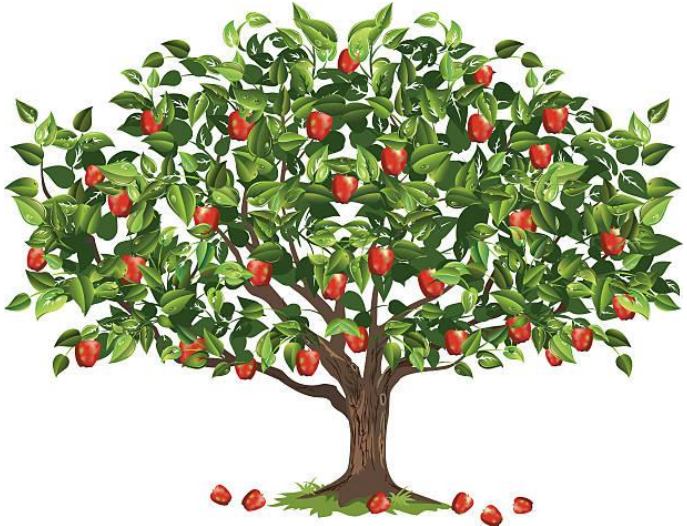
- How good is PacBio HiFi?
- Can we generate chromosome scale haplotype phased genome assemblies?
- What is the genetic diversity of crop wild relatives?
- Can we detect recombination events using *k*-mers?
- **When is the perfect time to pick an apple or pear?**
- Current progress in graph-based pangenomes

Apple Harvest

- Apples require cold storage after harvest to reach optimum maturity
- Response to cold storage and chilling hour requirements depend on maturity at harvest
- Current tests for maturity are unreliable especially across years and locations
- Is there a stage of physiological maturity measured by gene expression where post-harvest physiology and response to cold storage are predictable?

Experimental Design

- 14 genotypes, 11 orchards, 3 years, and 6 time points relative to commercial harvest
- Fruit RNA-seq 3 bio replicates ~15 million read pairs
- 276 total libraries



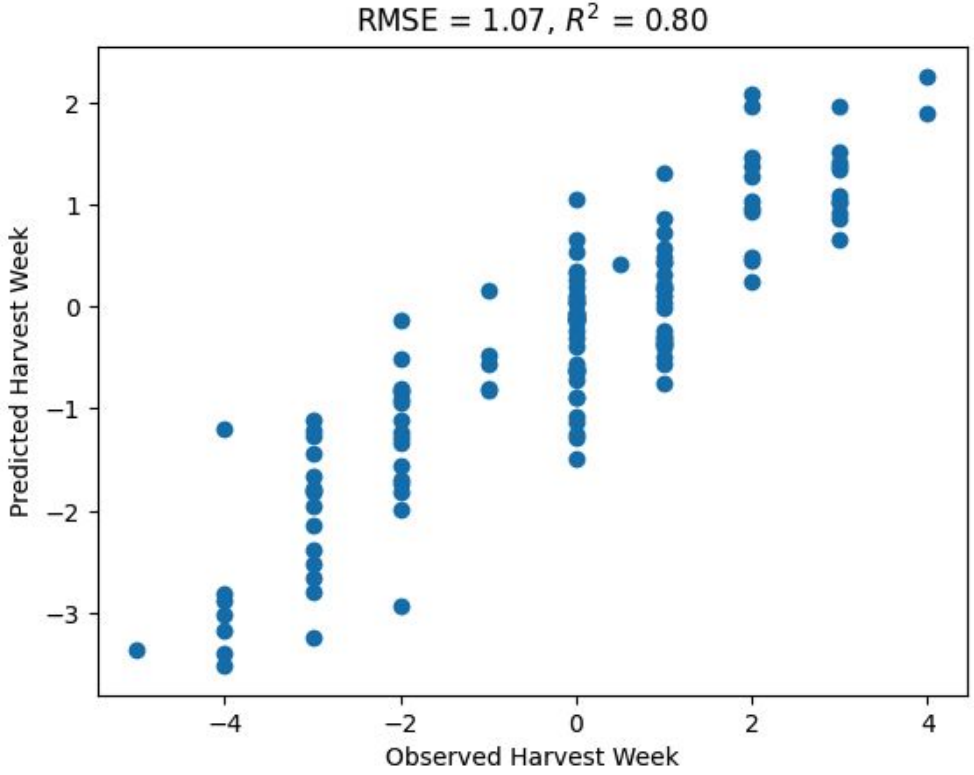
Workflow

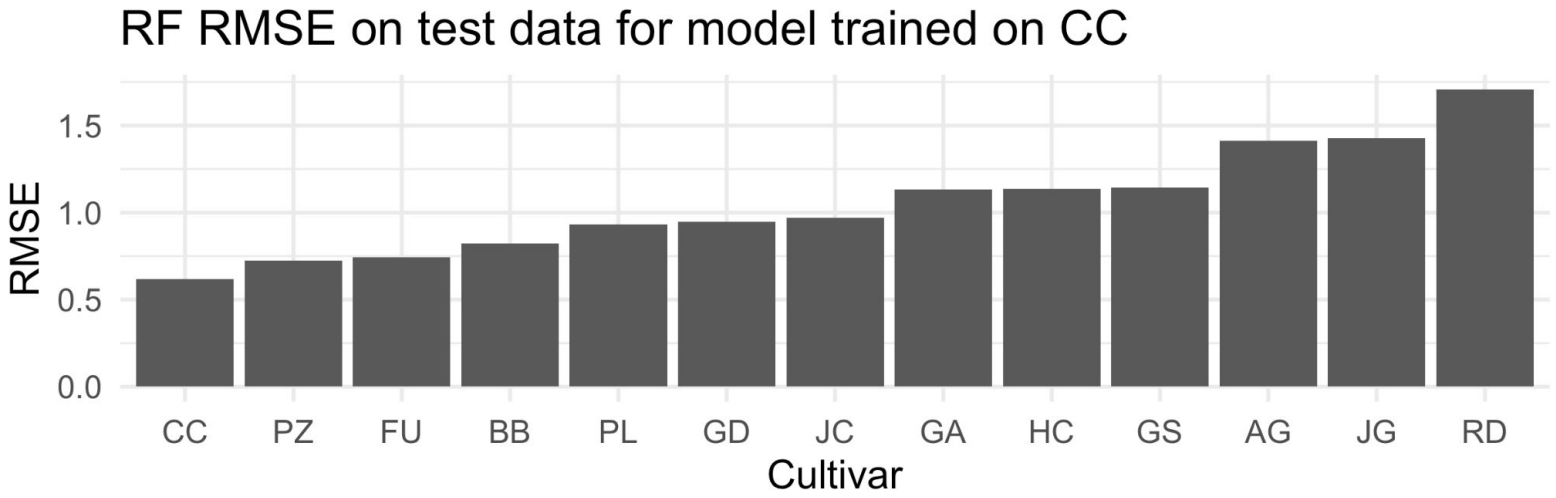
GEMmaker

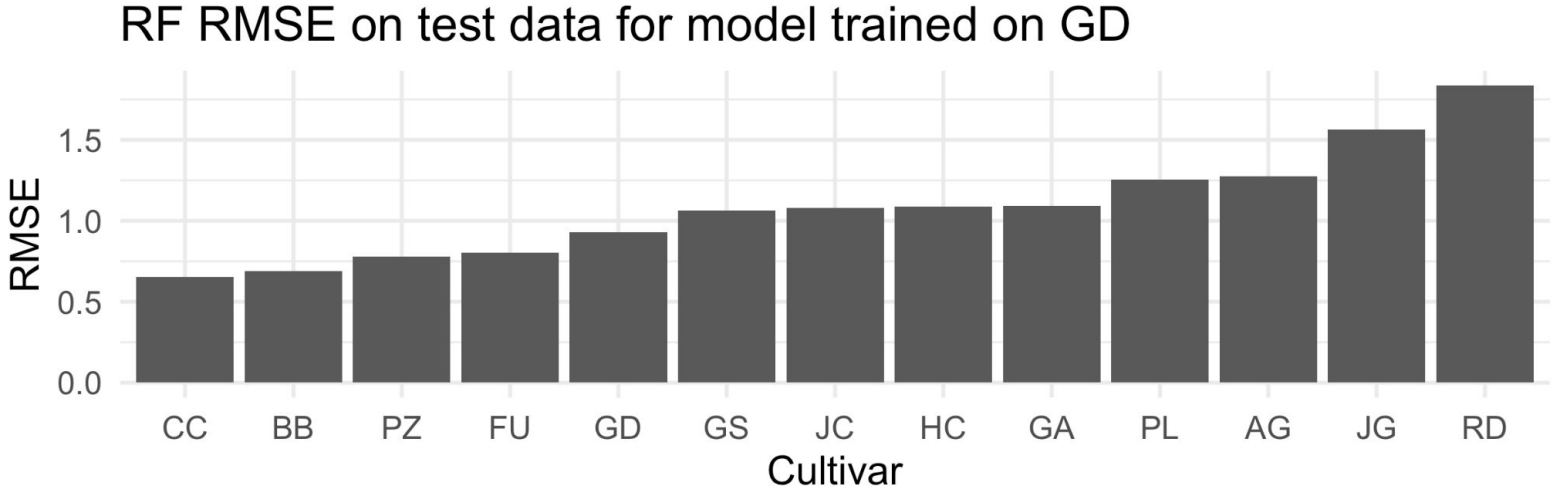
GEMMaker for RNA-seq QC and mapping with kallisto → Expression matrix TMM normalization → Additional QC on distribution of gene expression

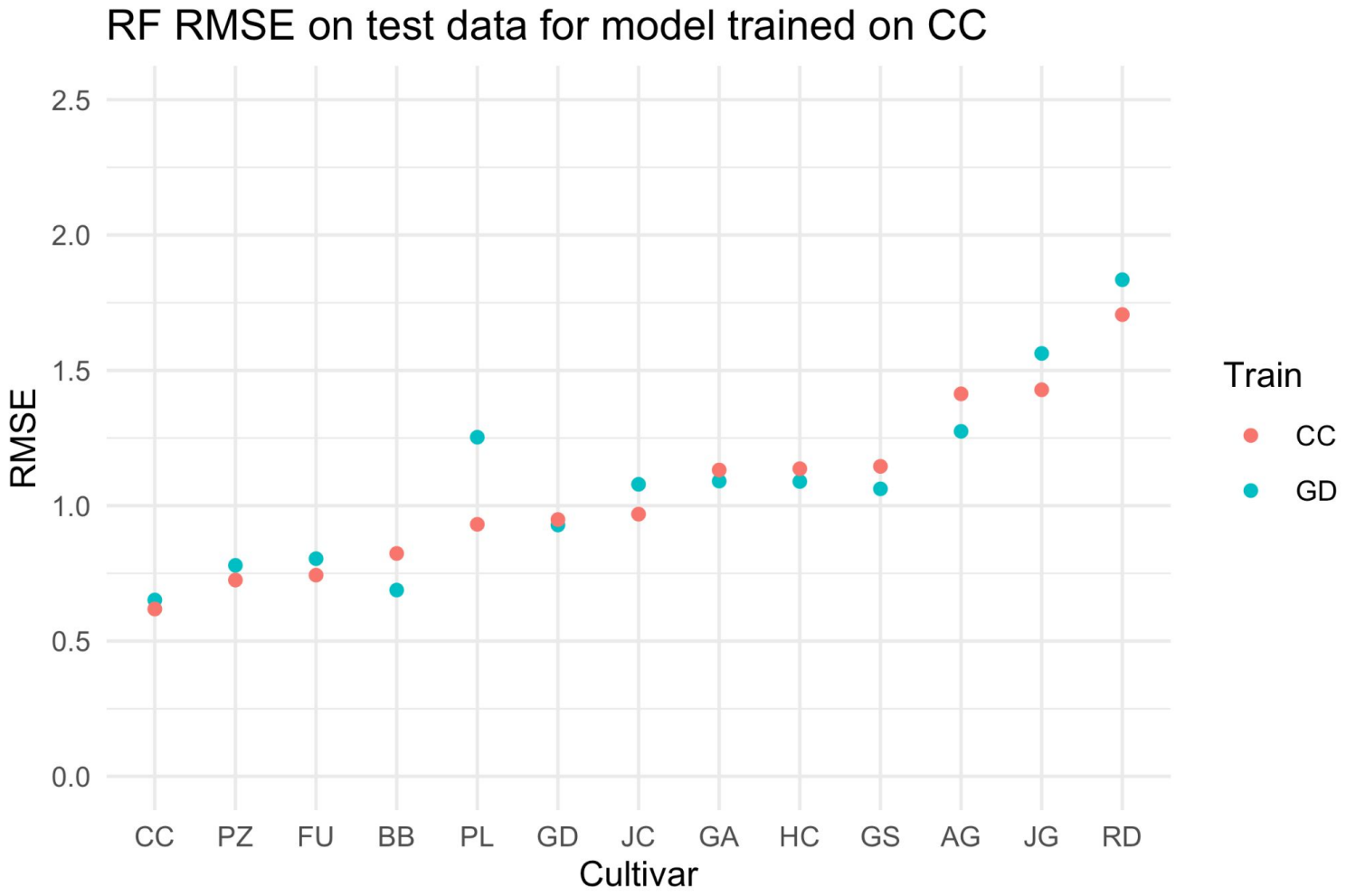
Split data into testing and training → Balance genotypes across test/train → Predict harvest week based on gene expression matrix

We can predict relative harvest week!




RF RMSE on test data for model trained on CC
RMSE
1.5
1.0
0.5
0.0
CC
PZ
FU
BB
PL
GD
JC
GA
HC
GS
AG
JG
RD
Cultivar


RF RMSE on test data for model trained on GD
RMSE
1.5
1.0
0.5
0.0
CC
BB
PZ
FU
GD
GS
JC
HC
GA
PL
AG
JG
RD
Cultivar

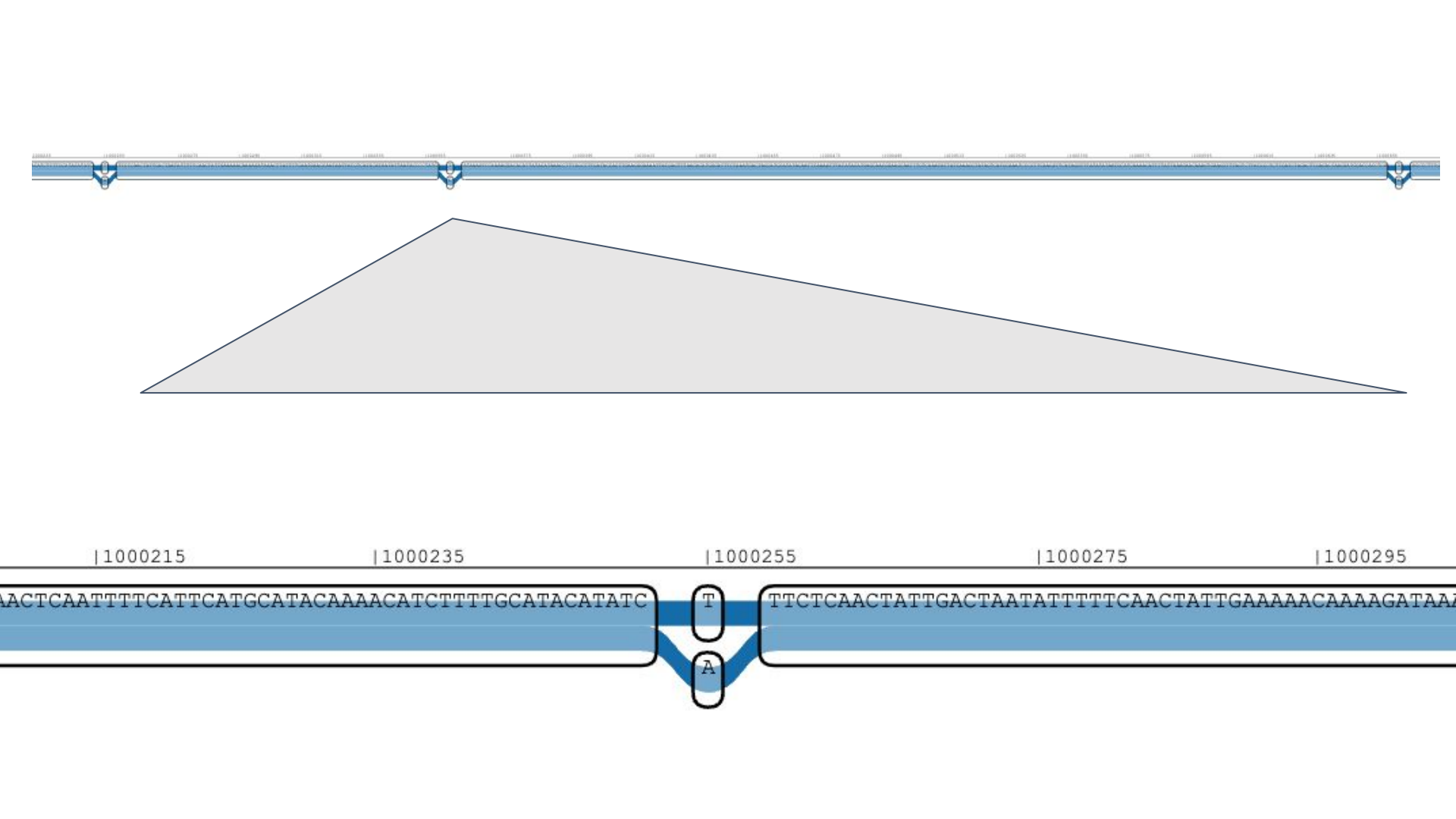

RF RMSE on test data for model trained on CC
RMSE
2.5
2.0
1.5
1.0
0.5
0.0
CC
PZ
FU
BB
PL
GD
JC
GA
HC
GS
AG
JG
RD
Cultivar
Train
CC
GD

Motivating Questions:

- How good is PacBio HiFi?
- Can we generate chromosome scale haplotype phased genome assemblies?
- What is the genetic diversity of crop wild relatives?
- Can we detect recombination events using *k*-mers?
- When is the perfect time to pick an apple or pear?
- **Current progress in graph-based pangenomes**

Graph-based pangenome

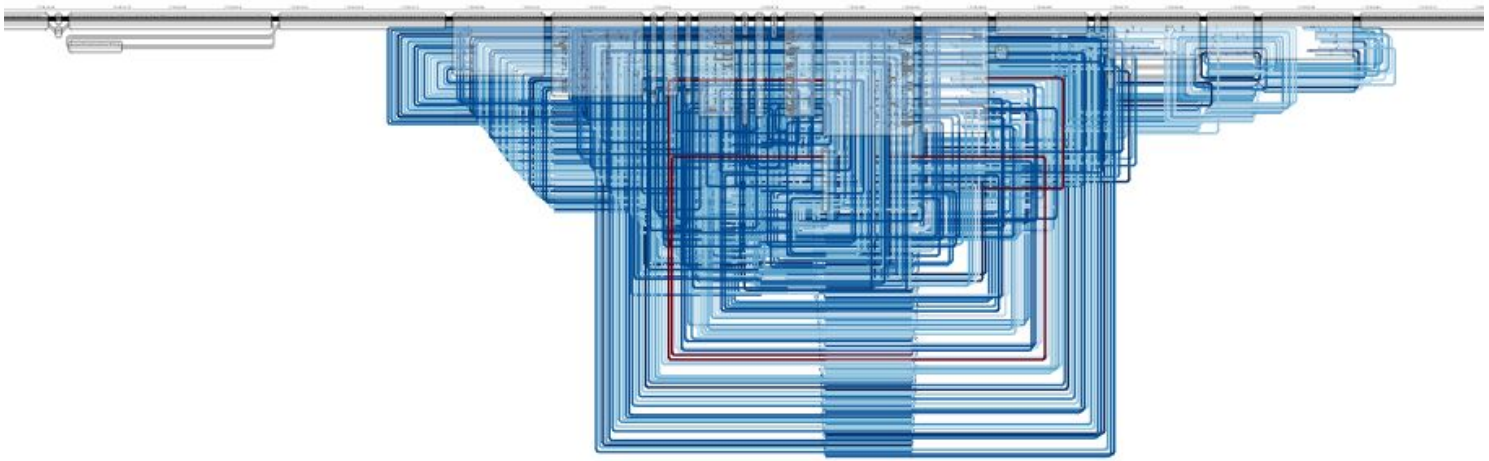
- Build haplotype graph between “Cosmic Crisp” haplotype A and haplotype B
- Visualize in sequenceTubeMap
- Identify alleles between haplotypes
- Map RNA-seq to the graph


|1000215
|1000235
|1000255
|1000275
|1000295
AACTCAATTTTCATTCATGCATACAAAACATCTTTTGCATACATATC
T
A
TTCTCAACTATTGACTAATATTTTTCAACTATTGAAAAACAAAAGATAAA

Identifying alleles between haplotypes

- MCScanX
  - Blast, identify syntenic blocks
  - 56% of genes between hapA/B “orthologs”
- Cactus-pangenome
  - Align haplotypes with minimap2
  - Intersect genome annotations on the pangenome graph
  - 73.3% of genes between hapA/B “orthologs”

Mapping RNA-seq… a work in progress



PROPOSED DURATION: 2 Years

Project Title: Risk Assessment for Loss of Firmness During Storage for Gala

PI:	Dr. Alan Yocca	**Co-PI:**	Dr. S Ticklin
Organization:	USDA-ARS Tree Fruit Research Lab	**Organization:**	te University
Telephone:	5096642280	**Teleph**	
Email:	alan.yocca@usda.gov; ayocca@hudsonalpha.org		edu
Co-PI:	Dr. Alex Harkess		as
Organization:	HudsonAlpha Institute of		ree Fruit Research Lab
Telephone:	2563270475		642280
Email:	aharkess@hudso		loren.honaas@usda.gov

Cooperators:

Year 1:	**Year 2:**	**Year 3:**	**Total Budget:**
$ 115,000.00	$ 90,000.00		$ 205,000.00

FUNDED