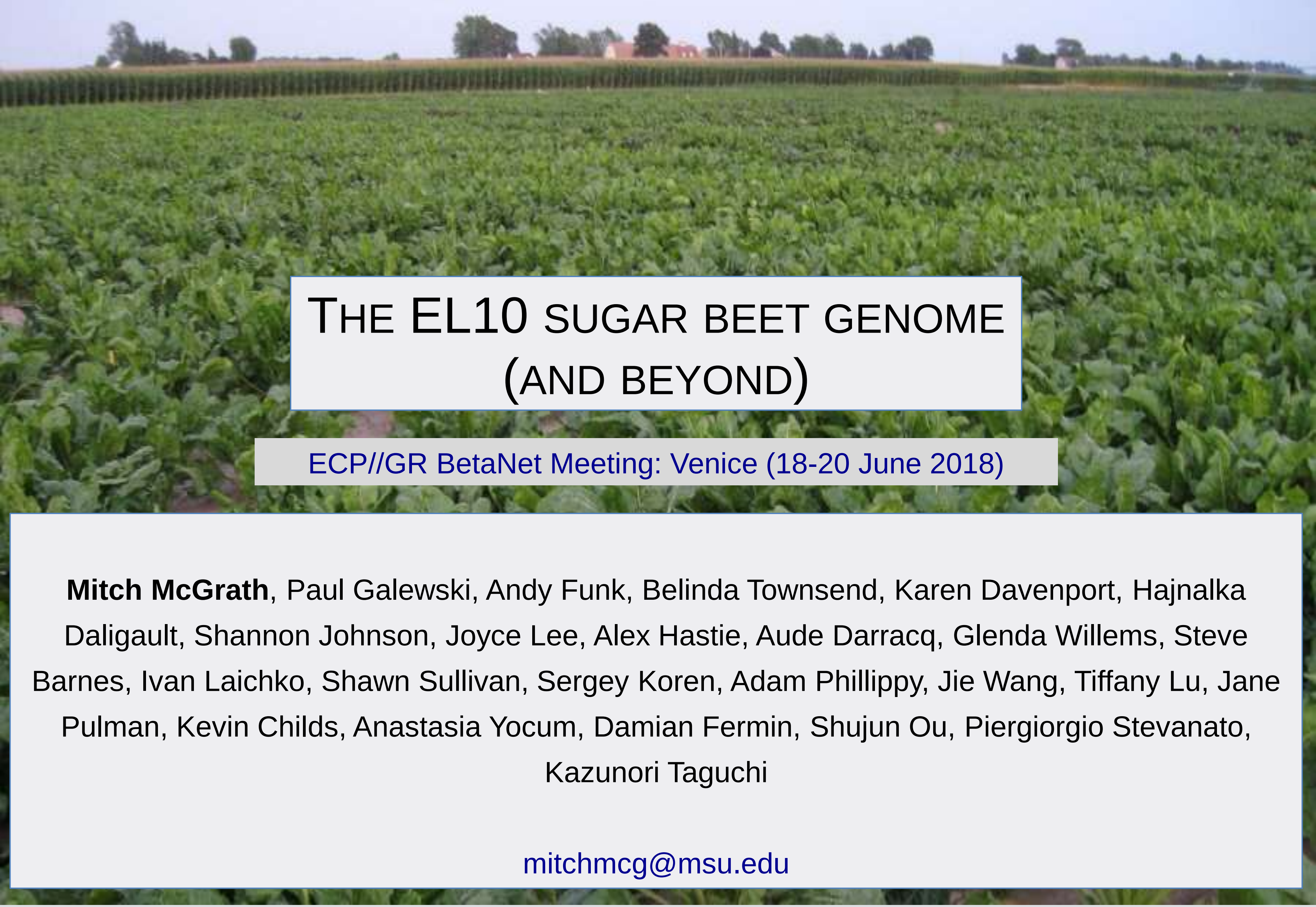


The EL10 sugar beet genome (and beyond)

ECP//GR BetaNet Meeting: Venice (18-20 June 2018)

Mitch McGrath, Paul Galewski, Andy Funk, Belinda Townsend, Karen Davenport, Hajnalka Daligault, Shannon Johnson, Joyce Lee, Alex Hastie, Aude Darracq, Glenda Willems, Steve Barnes, Ivan Laichko, Shawn Sullivan, Sergey Koren, Adam Phillippy, Jie Wang, Tiffany Lu, Jane Pulman, Kevin Childs, Anastasia Yocum, Damian Fermin, Shujun Ou, Piergiorgio Stevanato, Kazunori Taguchi

mitchmcg@msu.edu

Why assemble the EL10 sugar beet genome?

If you have one genome, you have one genome (*differences matter*).

Improved and 'affordable' technology for
(a more) *contiguous & accurate* genome assembly.

Chromosomes resolved to *single nucleotide* level (~ smallest unit of change).

An interesting technical challenge.

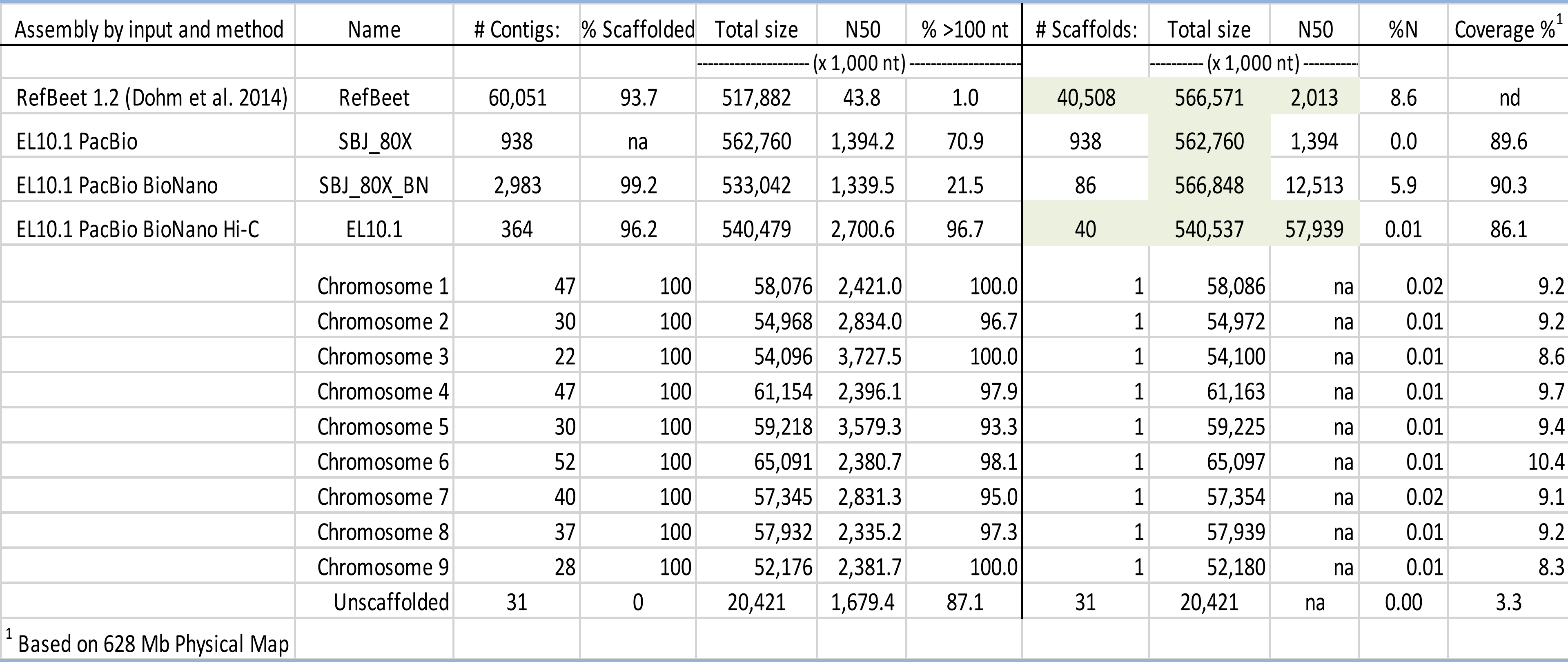
Progenitor of EL10 is parent to >23,000 hybrids and inbred derivatives,
>6,700 of these are inbred 6 to 9 generations (RILs).

Sequencing the EL10 beet: inputs

Technology	Library		Coverage[1]
		PacBio passed reads	
PacBio long reads	RS II, P6-C4 chemistry (Los Alamos Nat'l Labs)	6,540,795	79.3
	Mean leangth = 9,096 nt (std.dev = 6,528)		
	> 40 kb initial mapping and pre-assembly	5,176	0.38
		BioNano passed labels	
Optical physical map	BioNano Genomics, *Bss* SI - *Bsp* PQ1 Hybrid Scaffold	121 Gb	161.3
	Bsp PQ1 (7.6 labels/100 kb)	40 Gb	
	Bss SI (10 labels/100 kb)	81 Gb	
		Illumina passed reads	
Paired-End short reads	HiSeq 2500, TruSeq Libraries, 125bp PE (MSU-RTSF)	447,211,041	149
Cross-linked *in vivo*	Phase Genomics Hi-C library, HiSeq 2500, TruSeq Libraries	355,892,798	118.6

[1] Using the published genome size of 758 Mb

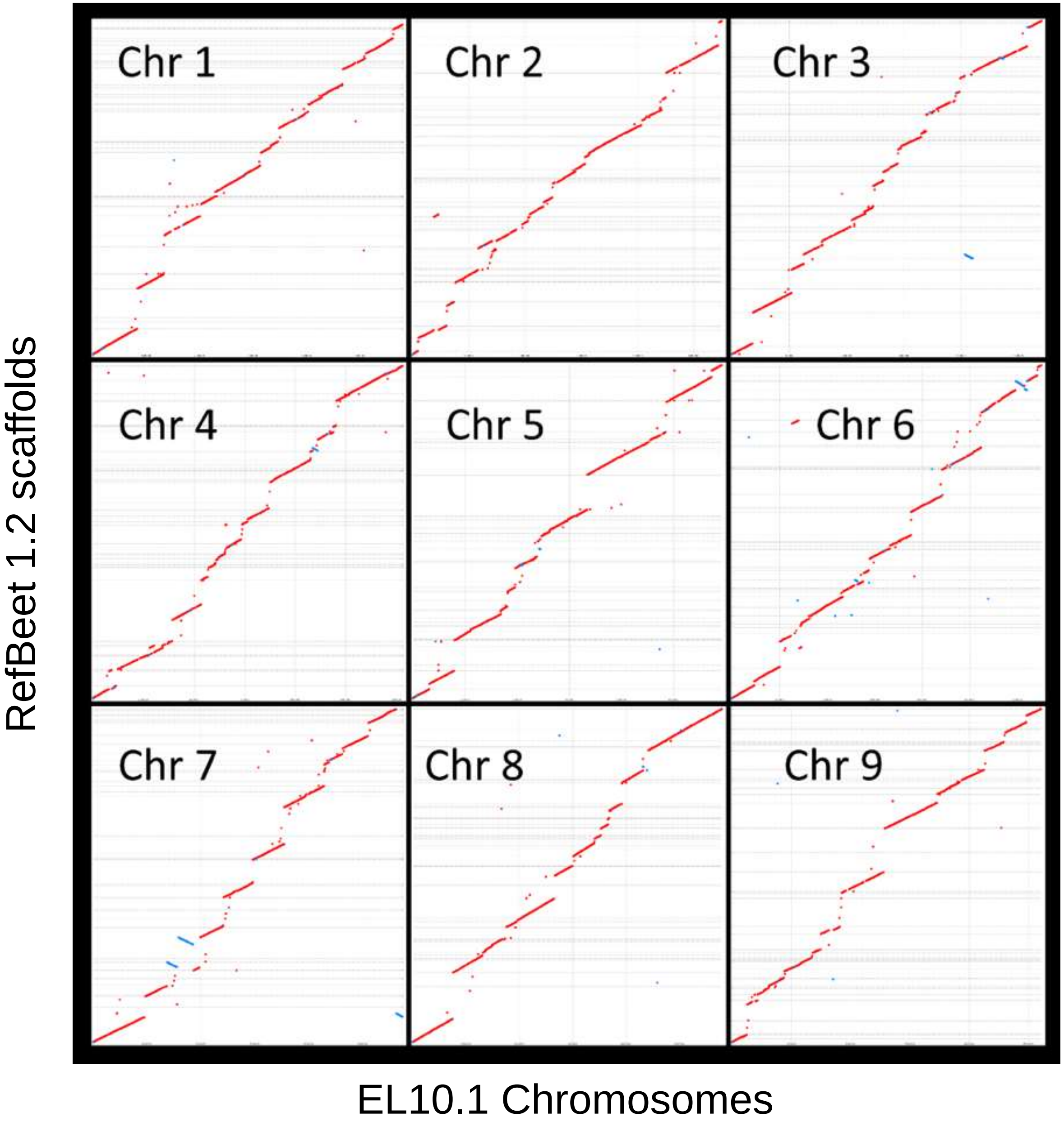
EL10 Assembly: outputs



Assembly by input and method	Name	# Contigs:	% Scaffolded	Total size	N50	% >100 nt	# Scaffolds:	Total size	N50	%N	Coverage %[1]
				-------------------- (x 1,000 nt) --------------------				---------- (x 1,000 nt) ----------			
RefBeet 1.2 (Dohm et al. 2014)	RefBeet	60,051	93.7	517,882	43.8	1.0	40,508	566,571	2,013	8.6	nd
EL10.1 PacBio	SBJ_80X	938	na	562,760	1,394.2	70.9	938	562,760	1,394	0.0	89.6
EL10.1 PacBio BioNano	SBJ_80X_BN	2,983	99.2	533,042	1,339.5	21.5	86	566,848	12,513	5.9	90.3
EL10.1 PacBio BioNano Hi-C	EL10.1	364	96.2	540,479	2,700.6	96.7	40	540,537	57,939	0.01	86.1
	Chromosome 1	47	100	58,076	2,421.0	100.0	1	58,086	na	0.02	9.2
	Chromosome 2	30	100	54,968	2,834.0	96.7	1	54,972	na	0.01	9.2
	Chromosome 3	22	100	54,096	3,727.5	100.0	1	54,100	na	0.01	8.6
	Chromosome 4	47	100	61,154	2,396.1	97.9	1	61,163	na	0.01	9.7
	Chromosome 5	30	100	59,218	3,579.3	93.3	1	59,225	na	0.01	9.4
	Chromosome 6	52	100	65,091	2,380.7	98.1	1	65,097	na	0.01	10.4
	Chromosome 7	40	100	57,345	2,831.3	95.0	1	57,354	na	0.02	9.1
	Chromosome 8	37	100	57,932	2,335.2	97.3	1	57,939	na	0.01	9.2
	Chromosome 9	28	100	52,176	2,381.7	100.0	1	52,180	na	0.01	8.3
	Unscaffolded	31	0	20,421	1,679.4	87.1	31	20,421	na	0.00	3.3

[1] Based on 628 Mb Physical Map

Contiguity

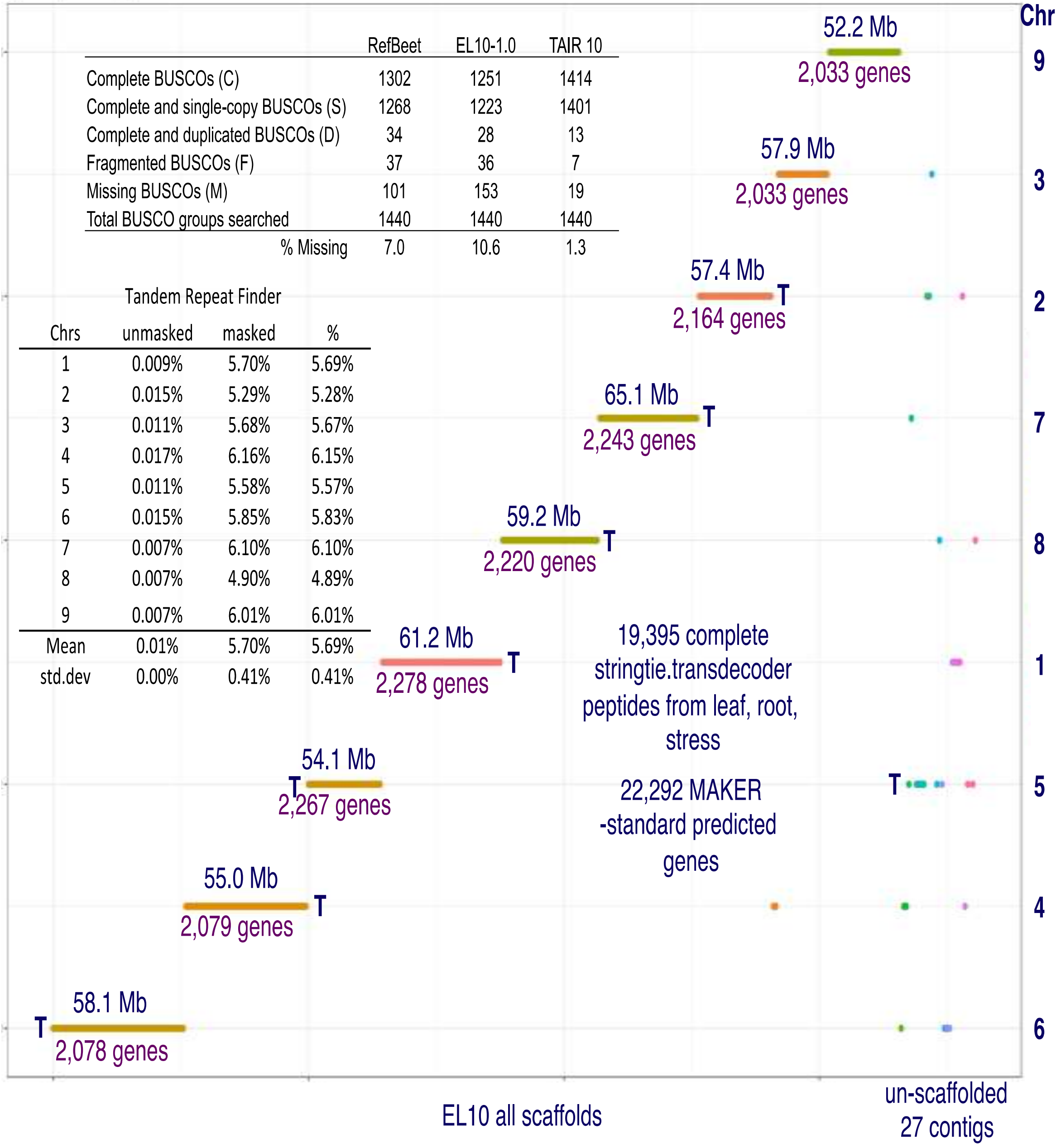

Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
RefBeet 1.2 scaffolds
EL10.1 Chromosomes

Completeness

	RefBeet	EL10-1.0	TAIR 10
Complete BUSCOs (C)	1302	1251	1414
Complete and single-copy BUSCOs (S)	1268	1223	1401
Complete and duplicated BUSCOs (D)	34	28	13
Fragmented BUSCOs (F)	37	36	7
Missing BUSCOs (M)	101	153	19
Total BUSCO groups searched	1440	1440	1440
% Missing	7.0	10.6	1.3

Tandem Repeat Finder

Chrs	unmasked	masked	%
1	0.009%	5.70%	5.69%
2	0.015%	5.29%	5.28%
3	0.011%	5.68%	5.67%
4	0.017%	6.16%	6.15%
5	0.011%	5.58%	5.57%
6	0.015%	5.85%	5.83%
7	0.007%	6.10%	6.10%
8	0.007%	4.90%	4.89%
9	0.007%	6.01%	6.01%
Mean	0.01%	5.70%	5.69%
std.dev	0.00%	0.41%	0.41%



19,395 complete stringtie.transdecoder peptides from leaf, root, stress

22,292 MAKER -standard predicted genes

EL10 is among the best plant assemblies created …
what can be done with it?

Develop new genetic markers for breeding.

Correlate markers with traits.

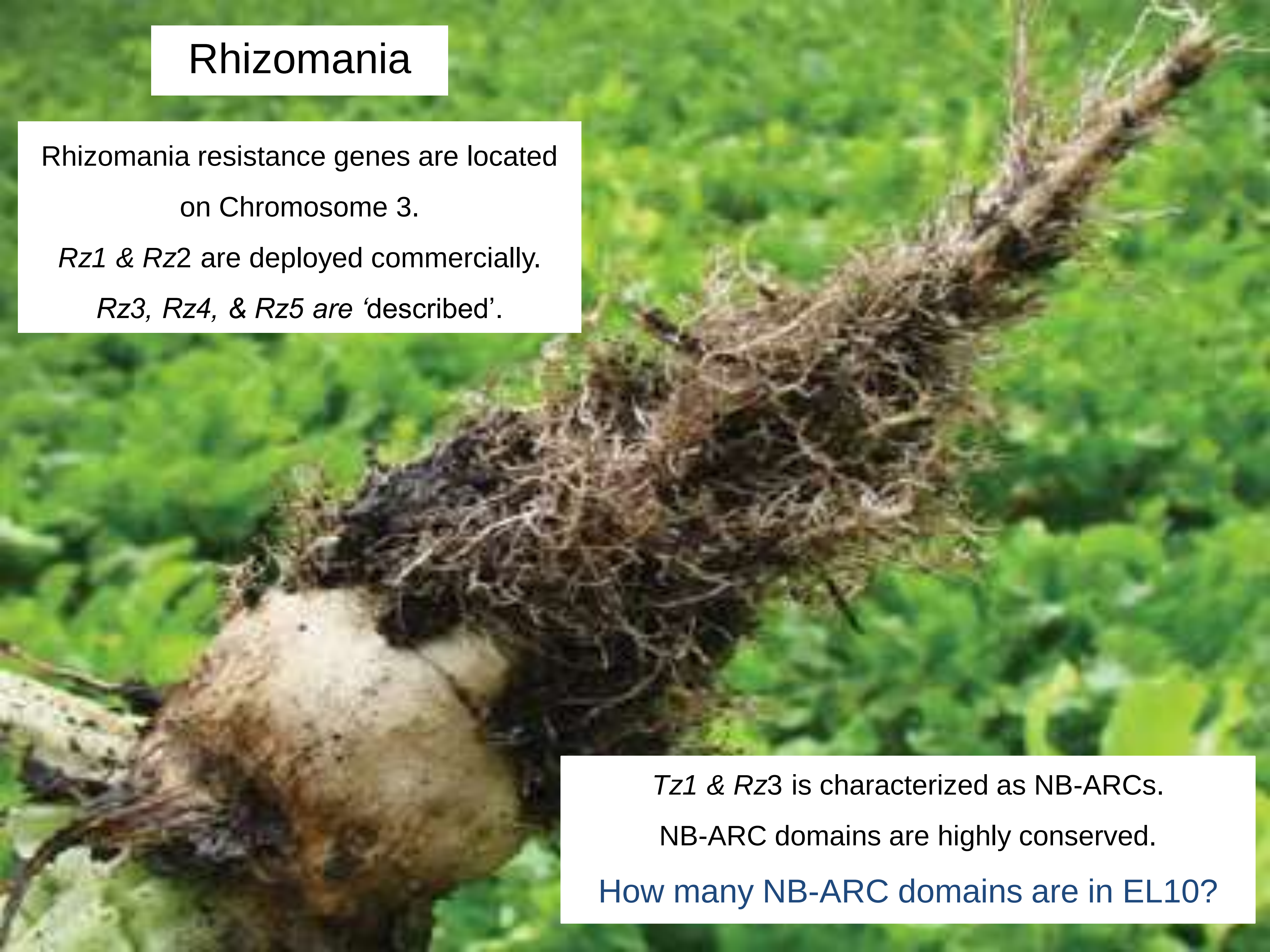
Examine occurrence and distribution of gene families.

Genome-wide evaluation of cultivars, breeding lines, and germplasm.

Evaluate distribution of genetic diversity.

Predict genes involved in agronomic traits.

AT SINGLE NUCLEOTIDE RESOLUTION



Rhizomania

Rhizomania resistance genes are located on Chromosome 3.

*Rz1 & Rz*2 are deployed commercially.

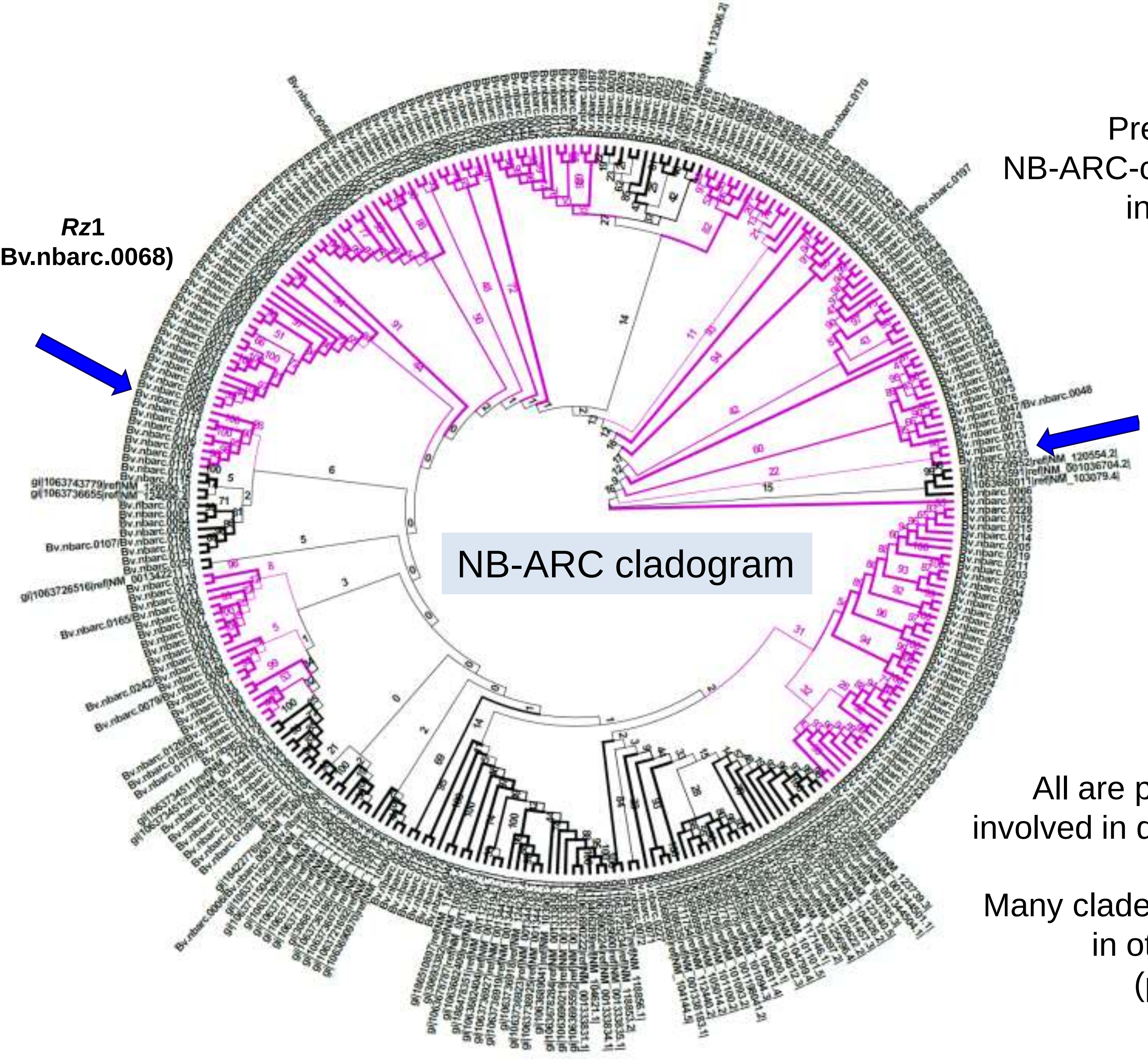
Rz3, Rz4, & Rz5 are ‘described’.

*Tz1 & Rz*3 is characterized as NB-ARCs.

NB-ARC domains are highly conserved.

How many NB-ARC domains are in EL10?

Predict 231
NB-ARC-containing genes
in EL10.1



***Rz*1**
(Bv.nbarc.0068)

***Rz*2 (Bv.nbarc.0121)**

NB-ARC cladogram

All are potential genes
involved in disease resistance.

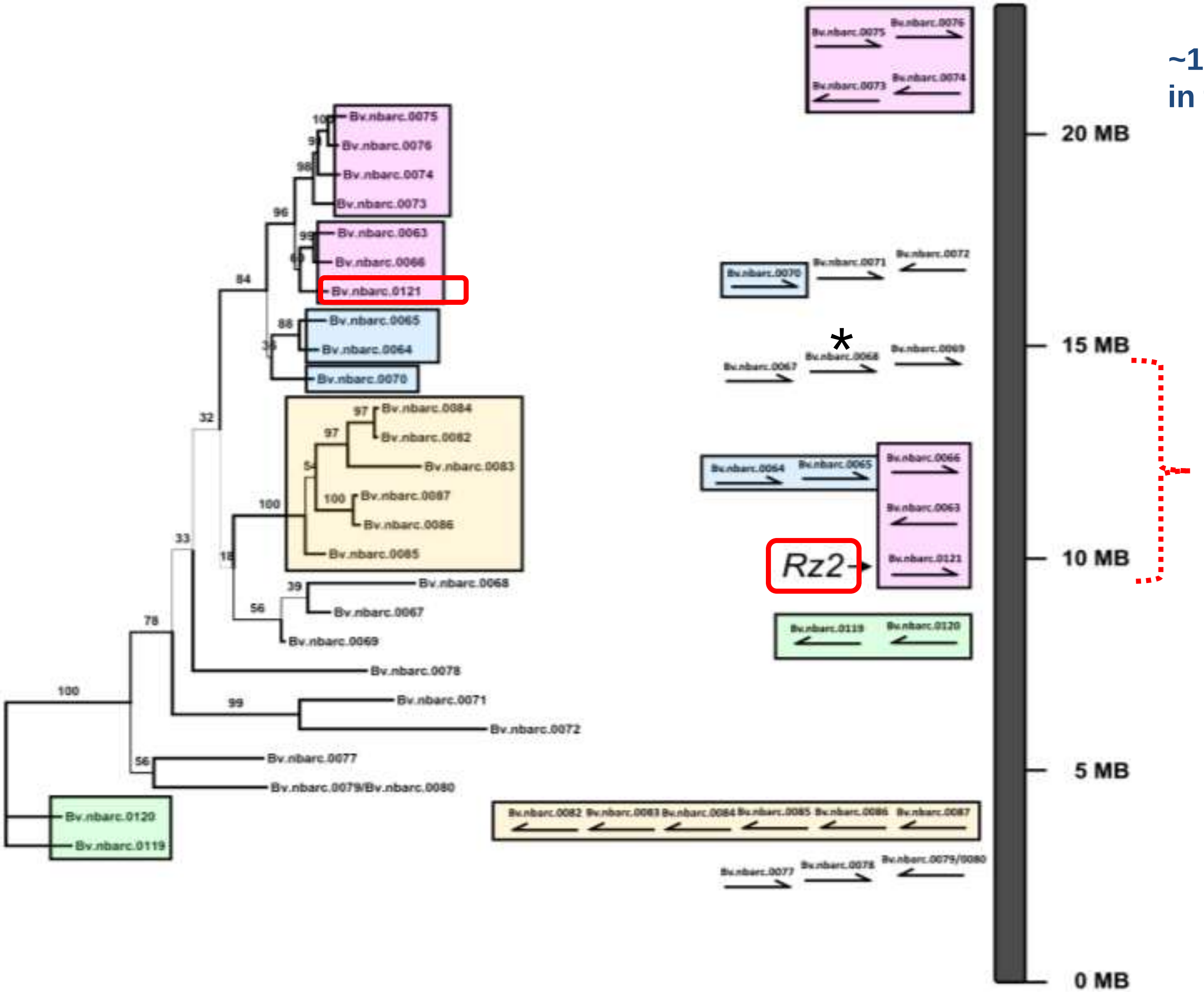
Many clades with no cognate
in other plants
(purple).

Clades of NB-ARCs on Chromosome 3

Position of NB-ARCs on Chromosome 3

26 NB-ARCs

~100 genes predicted in this ~ 20 Mb region



***Rz*-region**

rhizomania resistance markers are located within the Rz-region

Gene discovery

Breeding done with populations (most populations are variable).

Gene frequency estimates allow detection of past selection.

Approach :

Illumina sequence 25 individuals to 80X total depth of genome coverage.

Map reads to genome(s).

Determine polymorphic sites (~ 14 million SNPs).

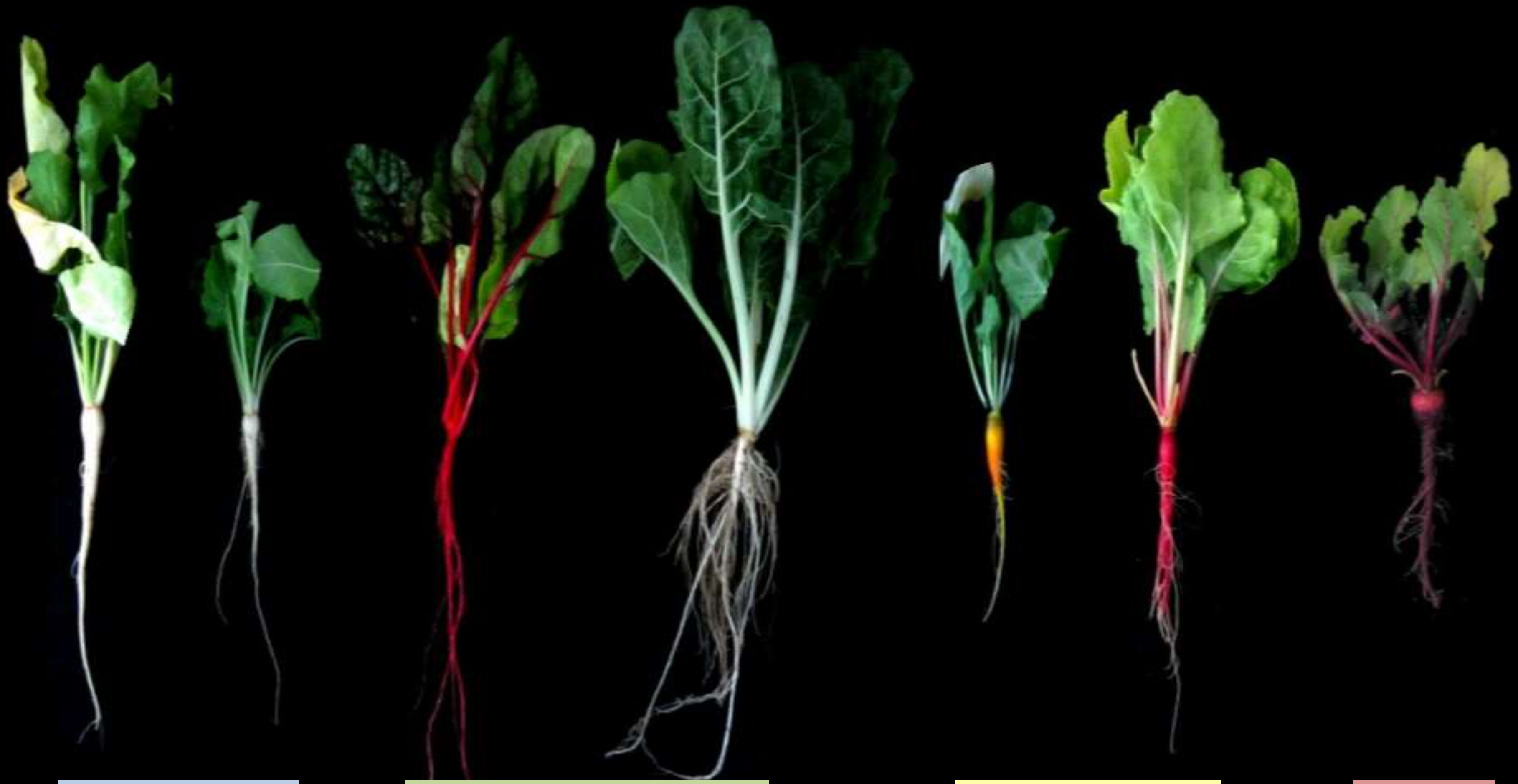
Filter to bi-allelic SNPs.

Calculate F_{ST} and 2pq.

Plot values across the genome assembly.

Interpret.

Crop type genes:
22 populations examined.
9 EL sugar, 7 table, 4 chard, 2 fodder.

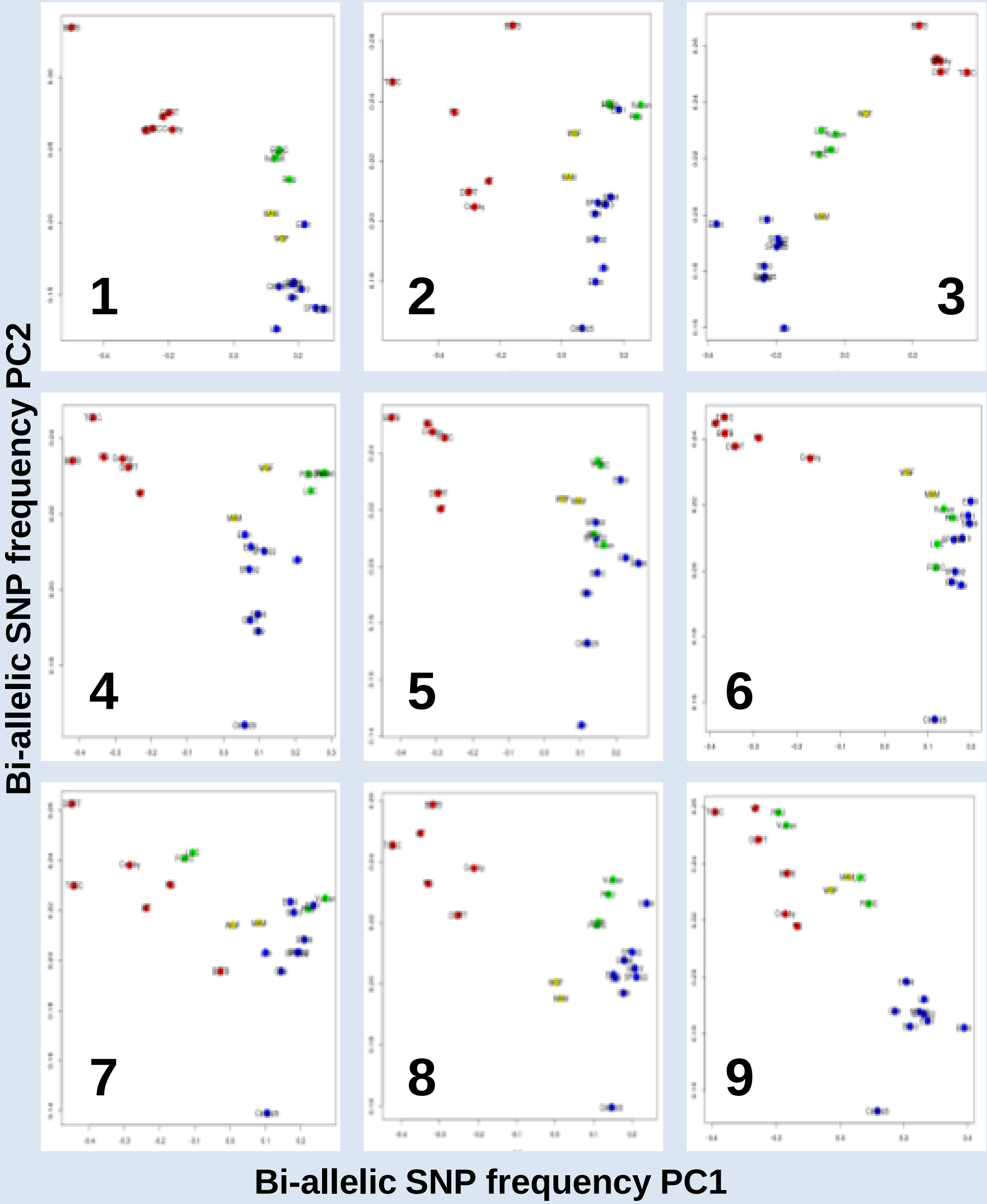


Sugar Chard Fodder Table

9 week old plants

Bi-allelic SNP frequency principal components

by Chromosome



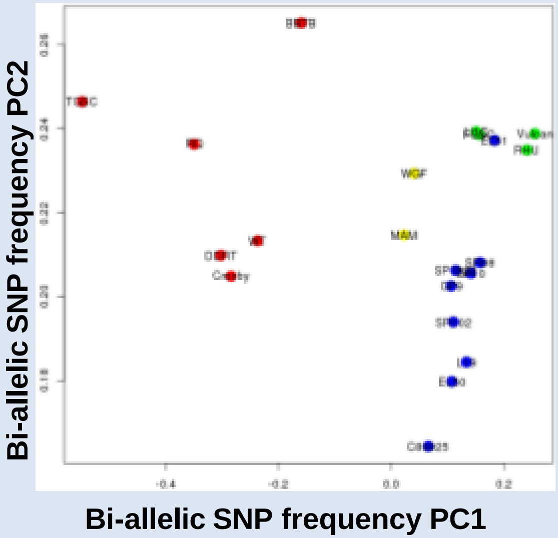
Sugar

Fodder

Chard

Table

Genome-wide



Population genomics of Chromosome 9

4 crop types, 2-9 accessions, 25 plants per accession, 80X PE Illumina sequences

FST

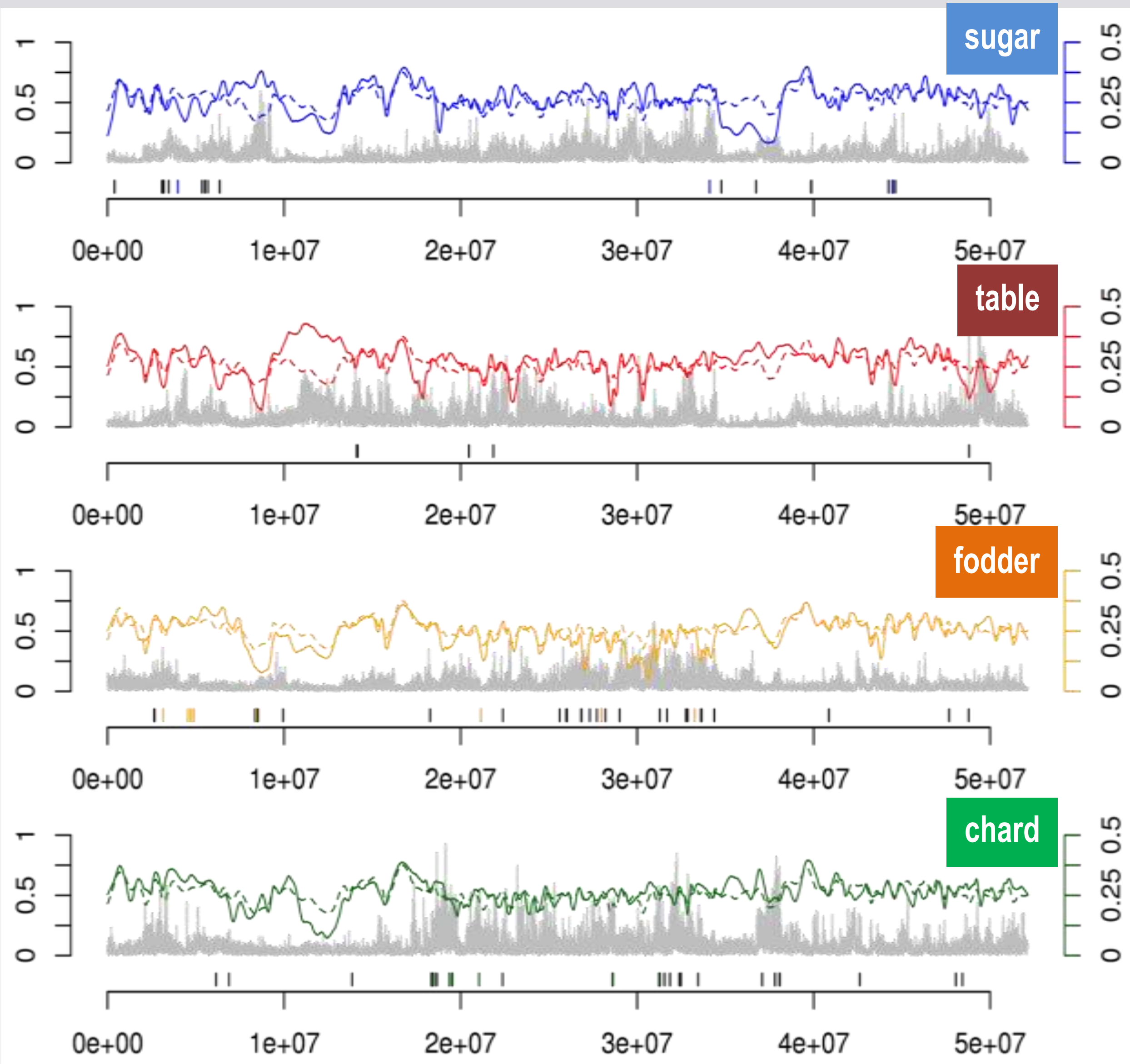
(histogram)

FST fixation index, ~ probability genes are identical by descent

uses average nucleotide differences

$$\frac{\text{(between pops – within pops)}}{\text{between pops}}$$

ranges between 0 and 1



nucleotide position

2pq

(lines)

(dashed line is all populations averaged)

ranges between 0 and 0.5

2pq = heterozygosity

$$p^2 + 2pq + q^2 = 1$$

Population genomics of Chromosome 9

4 crop types, 2-9 accessions, 25 plants per accession, 80X PE Illumina sequences

FST

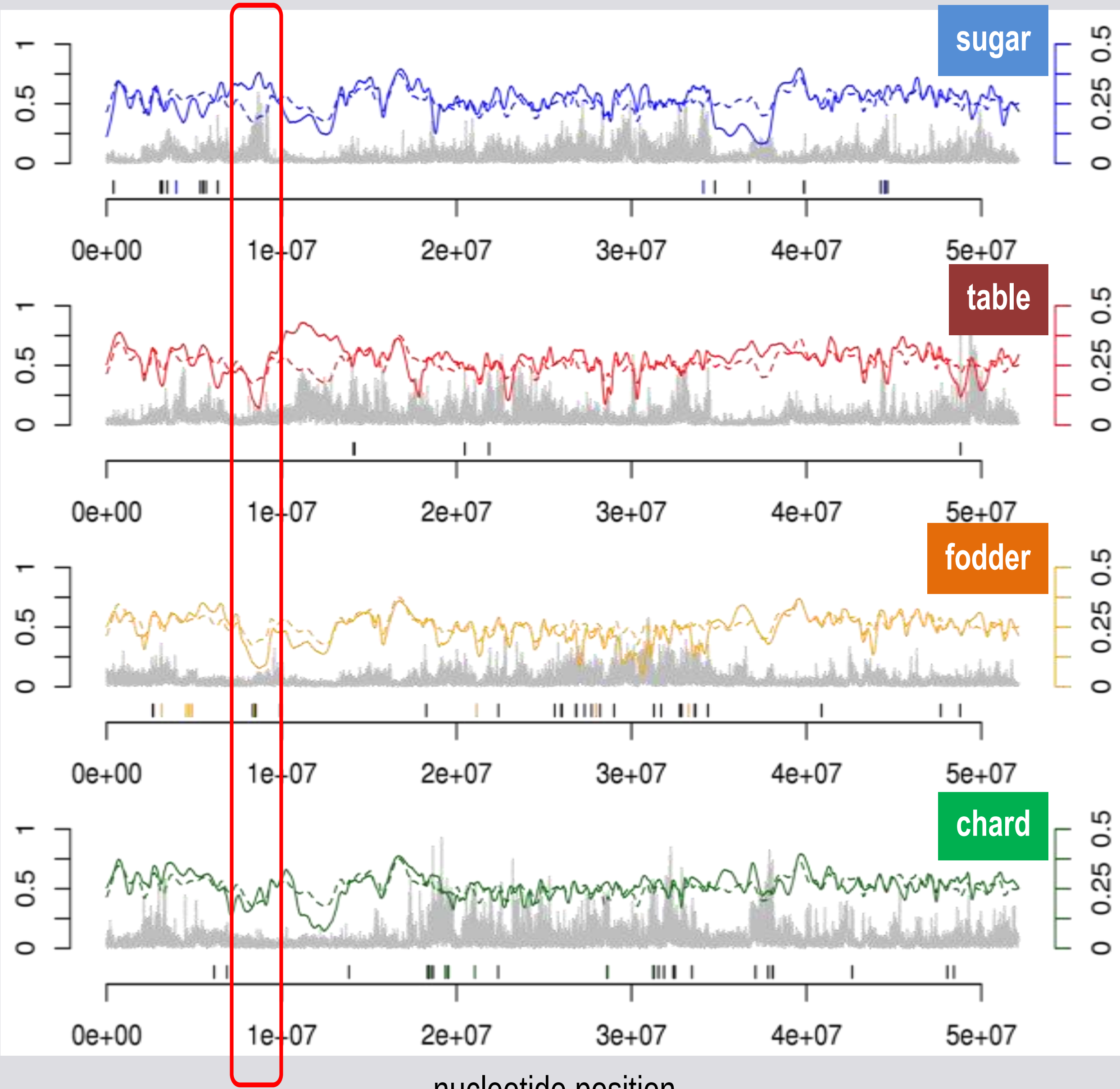
(histogram)

FST fixation index, ~ probability genes are identical by descent

uses average nucleotide differences

$$\frac{\text{(between pops – within pops)}}{\text{between pops}}$$

ranges between 0 and 1



nucleotide position

2pq

(lines)

(dashed line is all populations averaged)

ranges between 0 and 0.5

2pq = heterozygosity

$p^2 + 2pq + q^2 = 1$

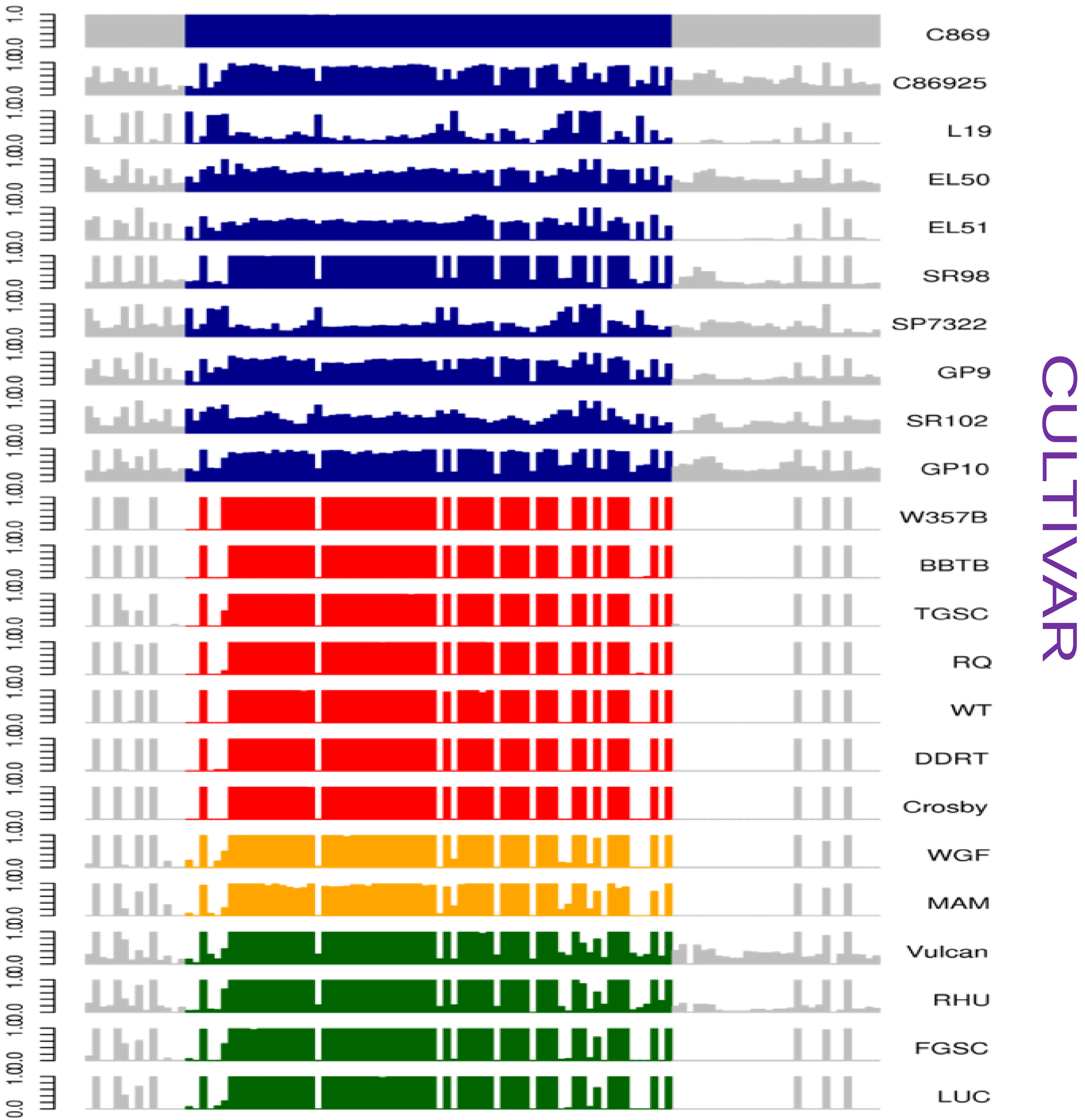
Zoom in on one gene in one FST peak of Chromosome 9

colored = coding
grey = non-coding

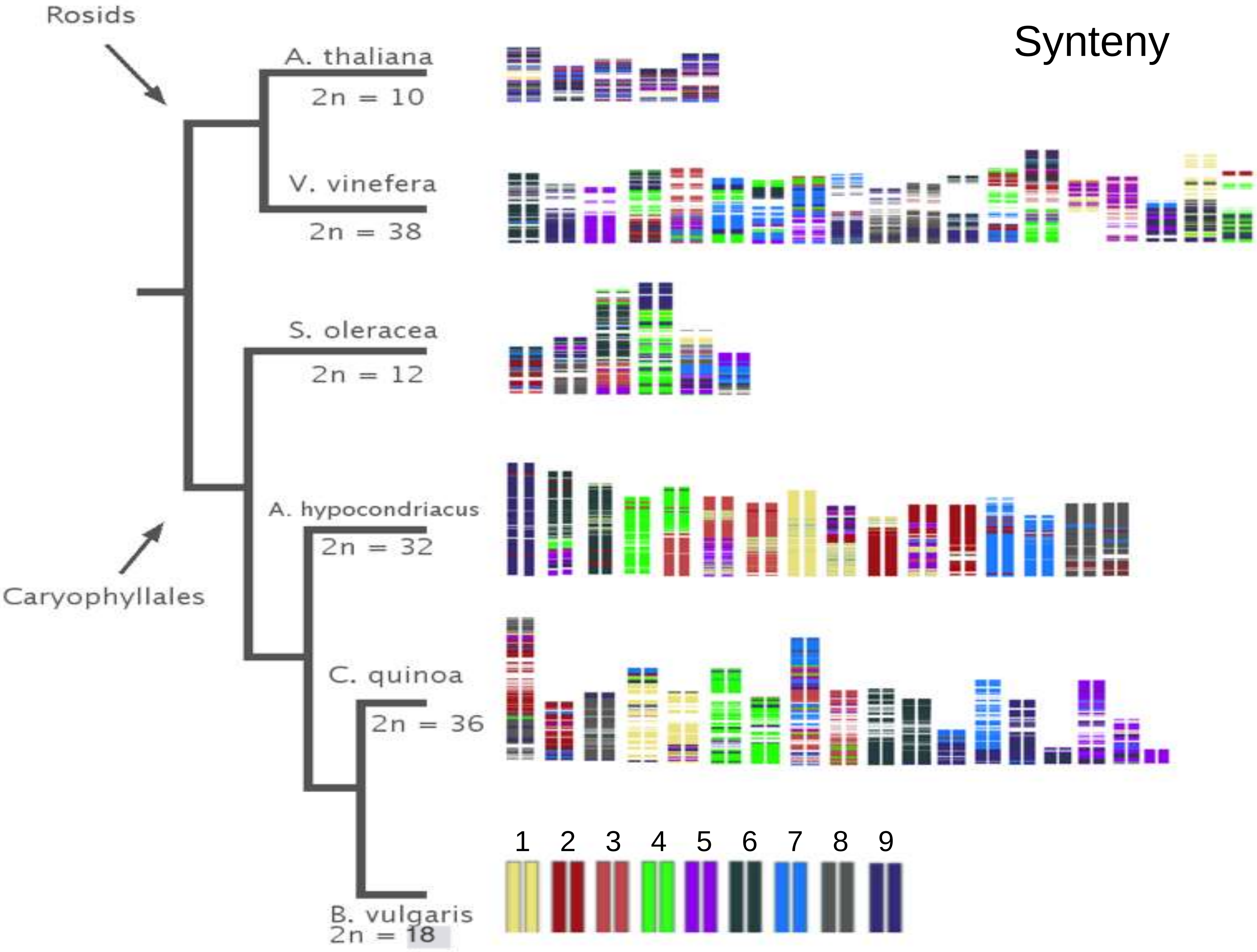
FST

Use to develop testable hypotheses,

~ 5,000 crop-wide FST peaks



Synteny


Rosids
A. thaliana
2n = 10
V. vinefera
2n = 38
S. oleracea
2n = 12
A. hypocondriacus
2n = 32
Caryophyllales
C. quinoa
2n = 36
B. vulgaris
2n = 18
1
2
3
4
5
6
7
8
9

Conclusion

High-quality genome assembly.

Comprehensive targeting of specific genes and gene families of interest.

Genome-wide assessment of differentiation possible.

Single nucleotide resolution allows specific genetic hypotheses
to be developed and tested.

Thank you for your comments, questions, and your attention!