nrgene

IMPROVING GENETIC RESEARCH AND BREEDING THROUGH COMPARATIVE PANGENOME ANALYSIS

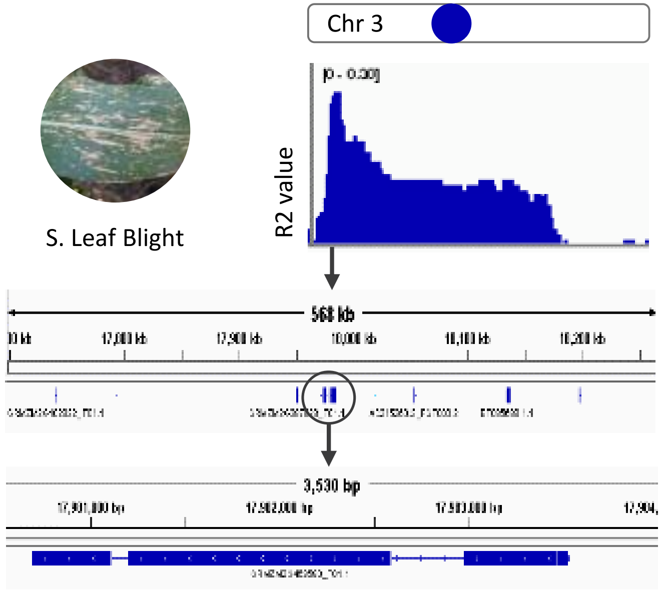
Paul Chomet, Ph.D., NRGene

Cotton Breeders Tour 2019

Genome Sequence: A Key for Crop Engineering & Improvement

Trait Discovery

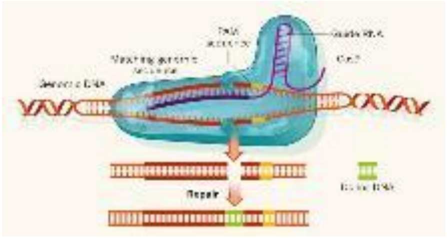
S. Leaf Blight



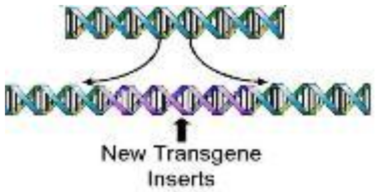
NBS-LRR Resistance Gene

Genome Modification

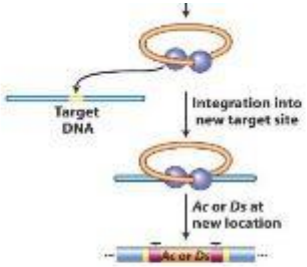
Editing



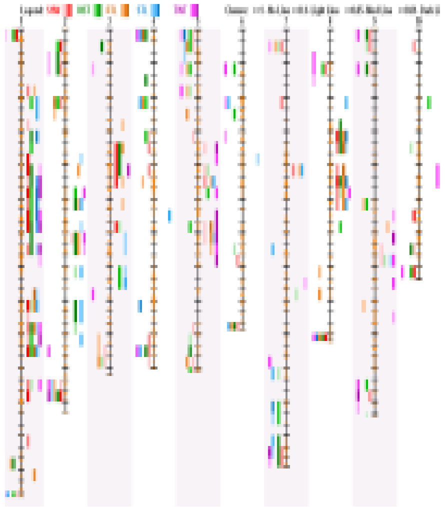
Transgenes



Mutagenesis



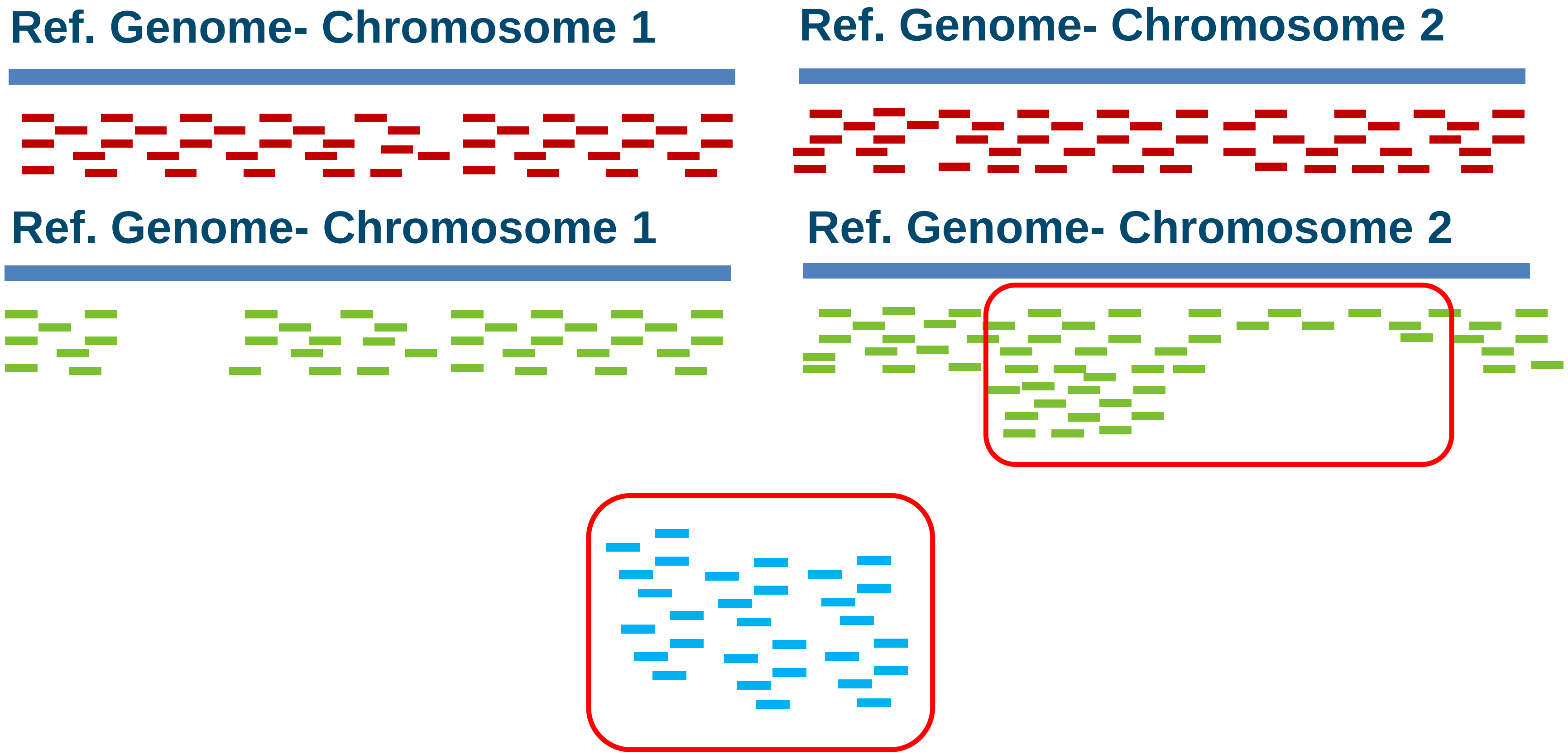
Marker Aided Breeding



→

CROP
IMPROVEMENT

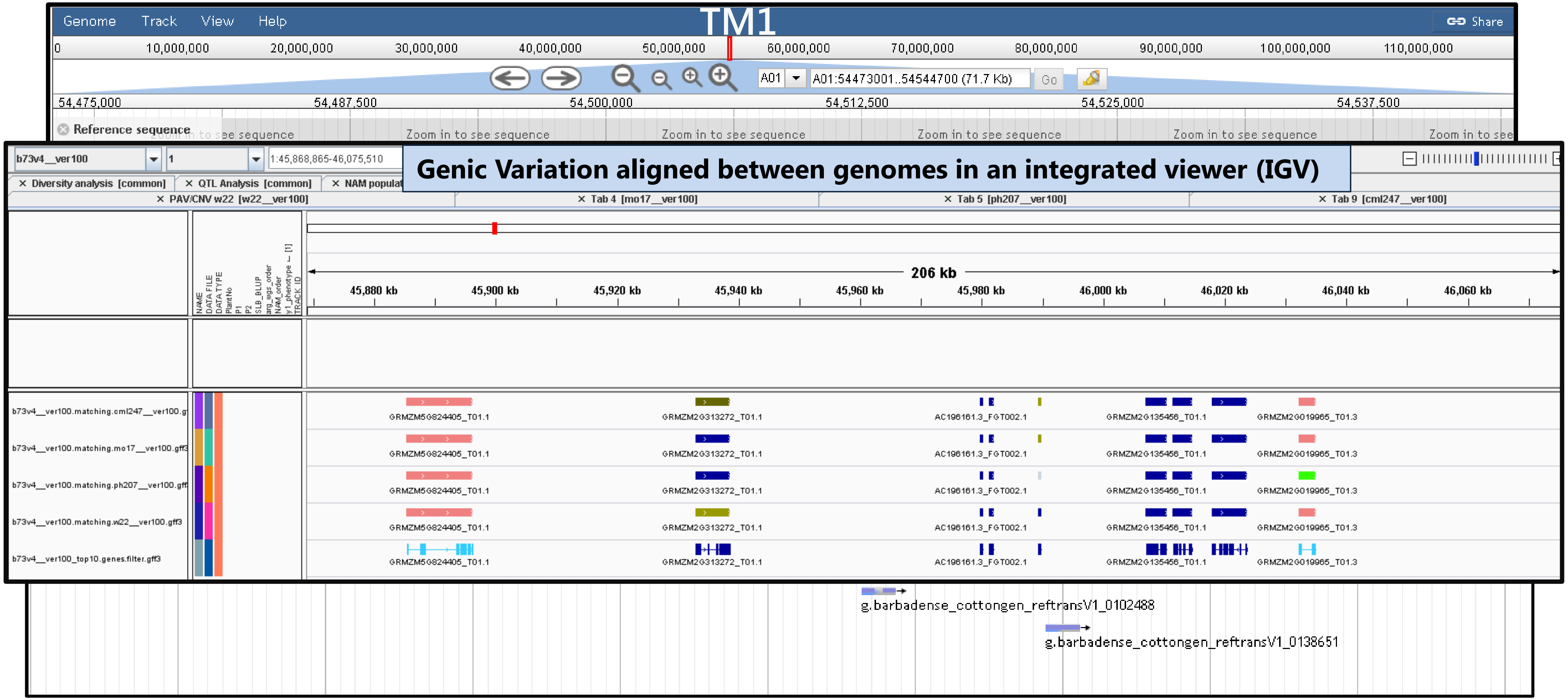
How Do You Analyze Across Genomes Data? Reference Genome Based Approach


Ref. Genome- Chromosome 1
Ref. Genome- Chromosome 2
Ref. Genome- Chromosome 1
Ref. Genome- Chromosome 2

- High rate of false discovery polymorphism due to misalignment
- High rate of undetected polymorphisms due to unmapped sequences
- Limited discovery of only part of the polymorphism: SNPs and small INDELs (no structural variation)

nrgene

More Genome Assemblies are Being Made Available How Can They Be Integrated for Analyses?



Genic Variation aligned between genomes in an integrated viewer (IGV)

nrgene

The PanMAGIC^{TM} Solution

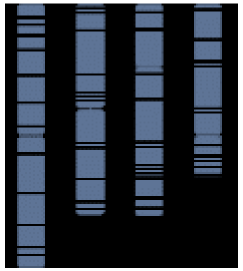
a method to capture genomic information and move across genomes

Select key lines

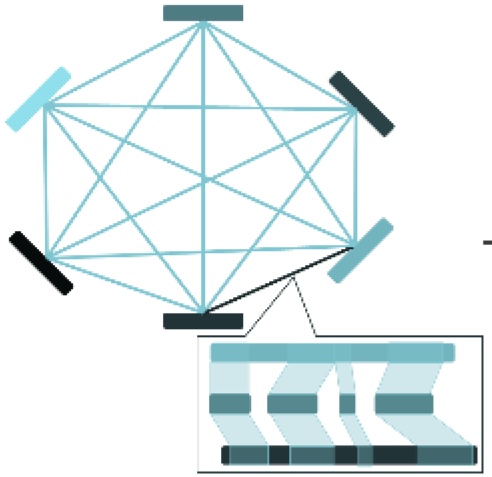
De-novo assembly of selected key lines

+

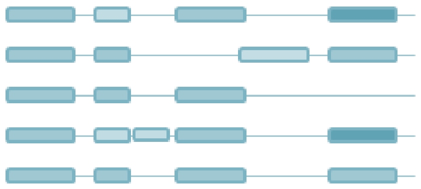
Supplied genetic or physical map



All to all genome mapping

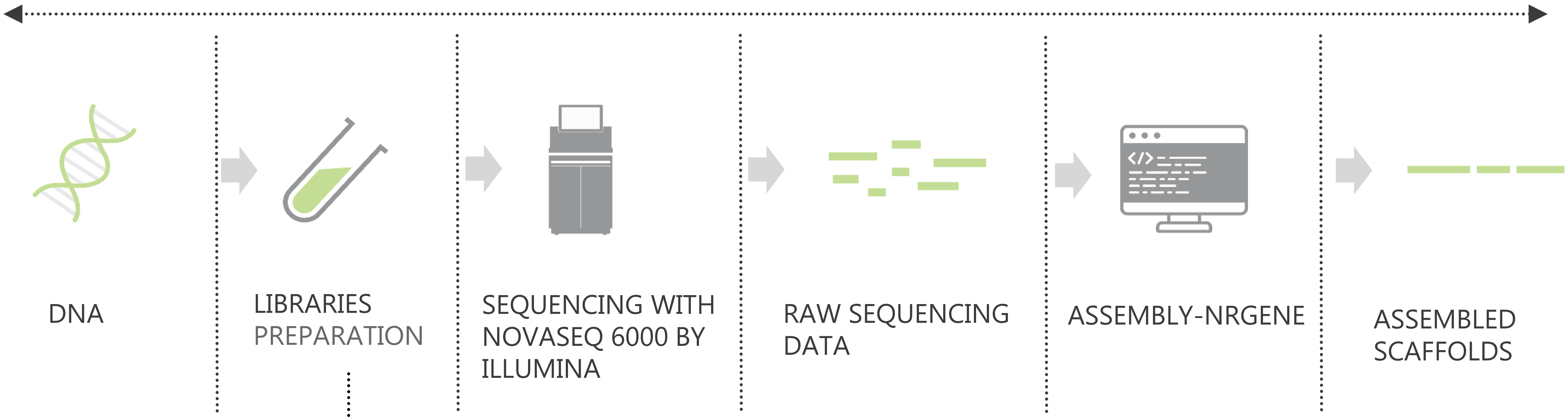


Transcript mapping PAV/ CNV and SV calling



nrgene

Comparative Genome Analyses Starts with High Quality Assemblies



DNA

LIBRARIES PREPARATION

SEQUENCING WITH NOVASEQ 6000 BY ILLUMINA

RAW SEQUENCING DATA

ASSEMBLY-NRGENE

ASSEMBLED SCAFFOLDS

Library type	Insert size	Reads length	Coverage
PCR-Free Shotgun Pair-end	470bp	250X2	40X
Nextera Mate-pair	2-4Kbp	150X2	30X
10X Chromium	50-100Kbp	150X2	25X
Total			**95X**

Library type	Insert size	Reads length	Coverage
PCR-Free Shotgun Pair-end	470bp	250X2	45X
PCR-Free Shotgun Pair-end	700bp	150X2	35X
Nextera Mate-pair	2-4Kbp	150X2	35X
Nextera Mate-pair	5-7Kbp	150X2	35X
Nextera Mate-pair	8-10Kbp	150X2	30X
10X Chrmoium	50-100Kbp	150X2	30X
Total			**210X**

NRGENE DeNovo3.0

NRGENE DeNovoMAX

nrgene

Improved Algorithms Have Allowed Lower Sequence Coverage While Maintaining High Quality Assemblies

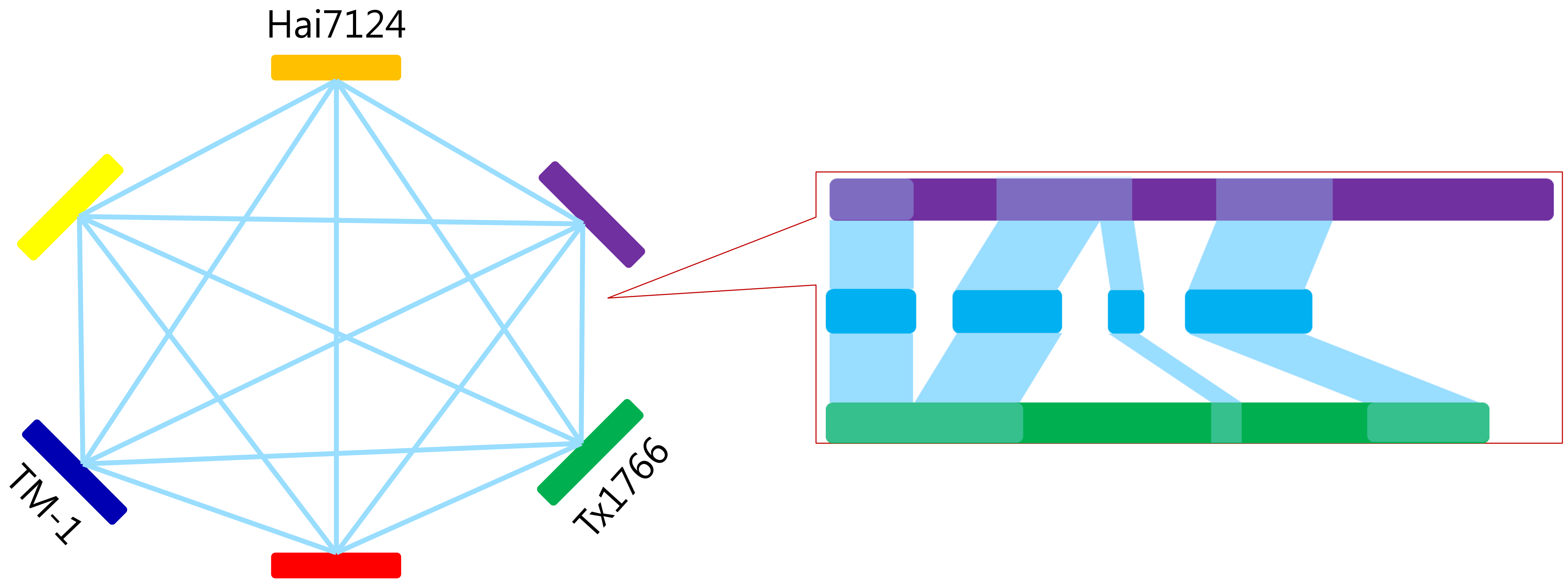
Crop	Pepper Diploid	Sunflower Diploid	Soy Diploid	Cotton Tetraploid	Sweet Corn Diploid	Bread Wheat Hexaploid
Total Assembly Size	3.26 Gbp	3.32 Gbp	1.01 Gbp	2.47 Gbp	2.36Gbp	14.58 Gbp
Scaffold N50 (# of scaffolds)	35.17 Mbp (25)	6.69 Mbp (149)	11.49 Mbp (29)	17.63 Mbp (43)	10.63 Mbp (66)	29.13 Mbp (129)
Scaffold N90 (# of scaffolds)	1.36 Mbp (211)	1.01 Mbp (571)	1.54 Mbp (114)	3.56 Mbp (154)	1.96 Mbp (258)	4.03 Mbp (619)
Unfilled Gaps (%N)	1.4%	0.8%	3.4%	1.4%	0.5%	0.9%
Complete BUSCO Genes	95.6%	91.0%	98.3%	96.1%	97.1%	98.3%
Short Reads Coverage	95X	95X	95X	95X	95X	95X

Available crops (Homozygote):

- Wheat
- Durum Wheat
- Barley
- Rye
- Oat
- Canola (Brassica Napus)
- Brassica oleracea
- Maize
- Soybean
- Common bean (phaseolus vulgaris)
- Chickpea
- Pepper
- Tomato
- Tobacco
- Melon
- Rice
- Sugar beet
- Cotton
- Sunflower
- Peanuts

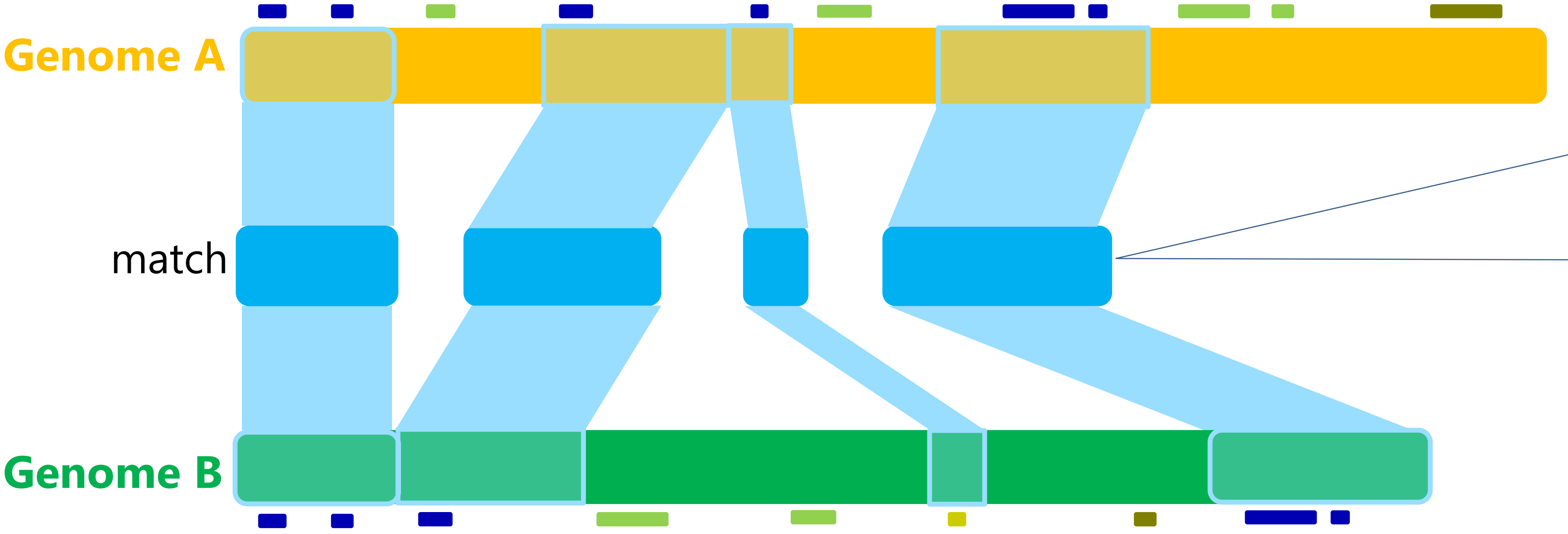
nrgene

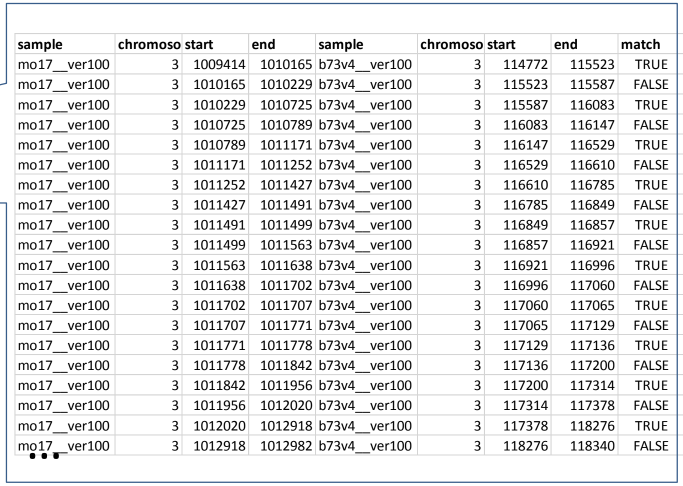
Pangenome Positioning is Enabled by All to All Mapping of Genome Coordinates



Input: a set of reference genomes
Output: all vs. all mappings depicting areas of homology and sequence polymorphism

Transcript Analysis and Structural Variants Calling





sample	chromoso	start	end	sample	chromoso	start	end	match
mo17__ver100	3	1009414	1010165	b73v4__ver100	3	114772	115523	TRUE
mo17__ver100	3	1010165	1010229	b73v4__ver100	3	115523	115587	FALSE
mo17__ver100	3	1010229	1010725	b73v4__ver100	3	115587	116083	TRUE
mo17__ver100	3	1010725	1010789	b73v4__ver100	3	116083	116147	FALSE
mo17__ver100	3	1010789	1011171	b73v4__ver100	3	116147	116529	TRUE
mo17__ver100	3	1011171	1011252	b73v4__ver100	3	116529	116610	FALSE
mo17__ver100	3	1011252	1011427	b73v4__ver100	3	116610	116785	TRUE
mo17__ver100	3	1011427	1011491	b73v4__ver100	3	116785	116849	FALSE
mo17__ver100	3	1011491	1011499	b73v4__ver100	3	116849	116857	TRUE
mo17__ver100	3	1011499	1011563	b73v4__ver100	3	116857	116921	FALSE
mo17__ver100	3	1011563	1011638	b73v4__ver100	3	116921	116996	TRUE
mo17__ver100	3	1011638	1011702	b73v4__ver100	3	116996	117060	FALSE
mo17__ver100	3	1011702	1011707	b73v4__ver100	3	117060	117065	TRUE
mo17__ver100	3	1011707	1011771	b73v4__ver100	3	117065	117129	FALSE
mo17__ver100	3	1011771	1011778	b73v4__ver100	3	117129	117136	TRUE
mo17__ver100	3	1011778	1011842	b73v4__ver100	3	117136	117200	FALSE
mo17__ver100	3	1011842	1011956	b73v4__ver100	3	117200	117314	TRUE
mo17__ver100	3	1011956	1012020	b73v4__ver100	3	117314	117378	FALSE
mo17__ver100	3	1012020	1012918	b73v4__ver100	3	117378	118276	TRUE
mo17__ver100	3	1012918	1012982	b73v4__ver100	3	118276	118340	FALSE

...

* illustration

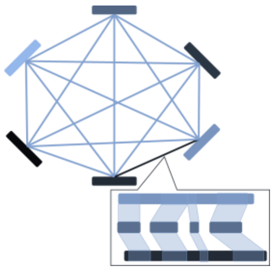
Locate transcript areas
Match annotation and indicate PAV/ CNV and translocations
Transcript analysis enables gene variation calling coupled with accurate mappings

MATCH	MAJOR TRANSLOCATION
PAV / CNV	MINOR TRANSLOCATION

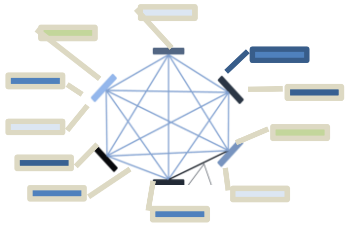
nrgene.

Immediate and Future Utility of Pangenome

Pangenome



Haplotype dB- GenoMAGIC system



Short term

- Cost effective Multi reference genome access- full sequence comparisons
- PAV/CNV Gene identification
- Allele sequence identification
- sequence positioning across genomes

Future capabilities

- Improved Genotyping Array development
- High Density seq based genotyping capability
- Marker imputation for genotyping cost savings
- Streamlined trait marker discovery/implementation
- Hap based mapping and functional allele discovery

Transcript Analysis and Structural Variants Calling

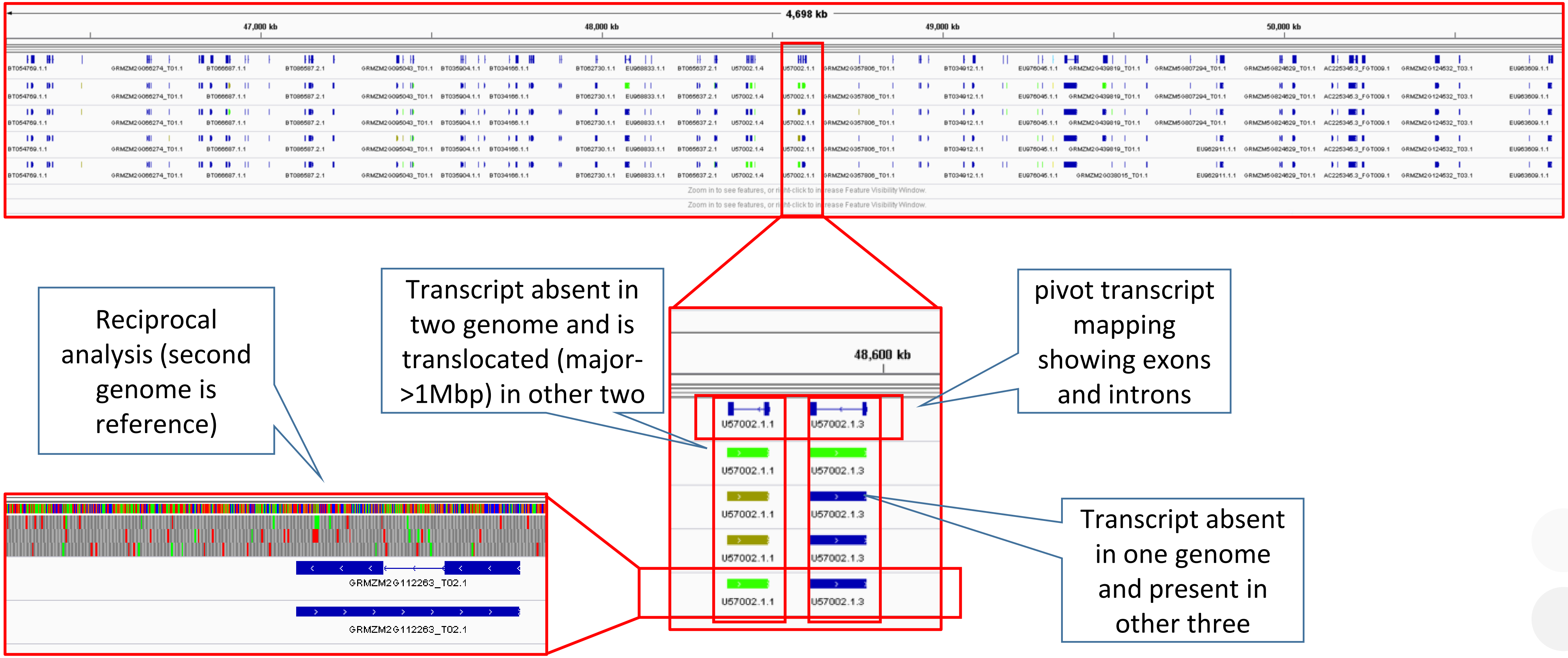
PanMAGIC™

MATCH

PAV / CNV

MAJOR TRANSLOCATION

MINOR TRANSLOCATION



Reciprocal analysis (second genome is reference)

Transcript absent in two genome and is translocated (major->1Mbp) in other two

pivot transcript mapping showing exons and introns

Transcript absent in one genome and present in other three

* Visualization using IGV browser (Broad institute)

nrgene

Polymorphisms Detected Across 3 Cotton Genomes

- Subsample of genome – chr 1

Chr1 comparison	*count of MNPs*	*count of SNPs*	*count of InDels*	*Polymorphism/kb*
TM1 vs Tx1766	4533 (2.0%)	174954 (78.6%)	42993 (19.3%)	1.88
TM1 vs Hai7124	7595 (1.3%)	495839 (83.6%)	89821 (15.2%)	5.13
Tx1766 vs Hai7124	5808 (0.9%)	540601 (85.8%)	83569 (13.3%)	5.33

- NRGene initial GenoMAGIC built with pangenome of TM1, Hai7124, Tx1766, + 16 lines to capture haplotype diversity
- Pangenome comparisons allows for identification of significant polymorphisms ~5.1 per 1000 bps
- ~15% of polymorphisms identified as Insertion/Deletions

Gene Editing Requires Discovery and Integrated Genomics Data

Functional Genomics/Gene target

- Pangenome can identify PAV/CNV for gene discovery
- Positions genes relative to QTL

Allele identification

- allelic variants across the pangenome are identified

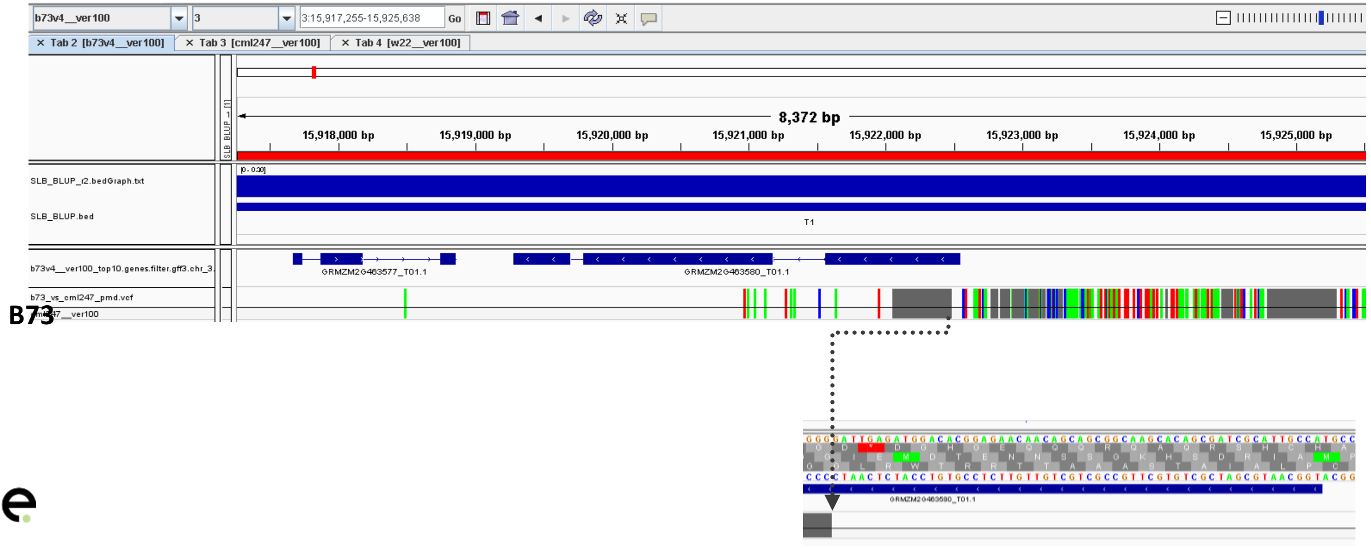
Editing and QC

- Pangenome improves precision of targets for edits
- Multi-reference improves off target detection

Testing and deployment

- Pangenome improves quality of markers for breeding

Polymorphism
CML247 genome vs. B73



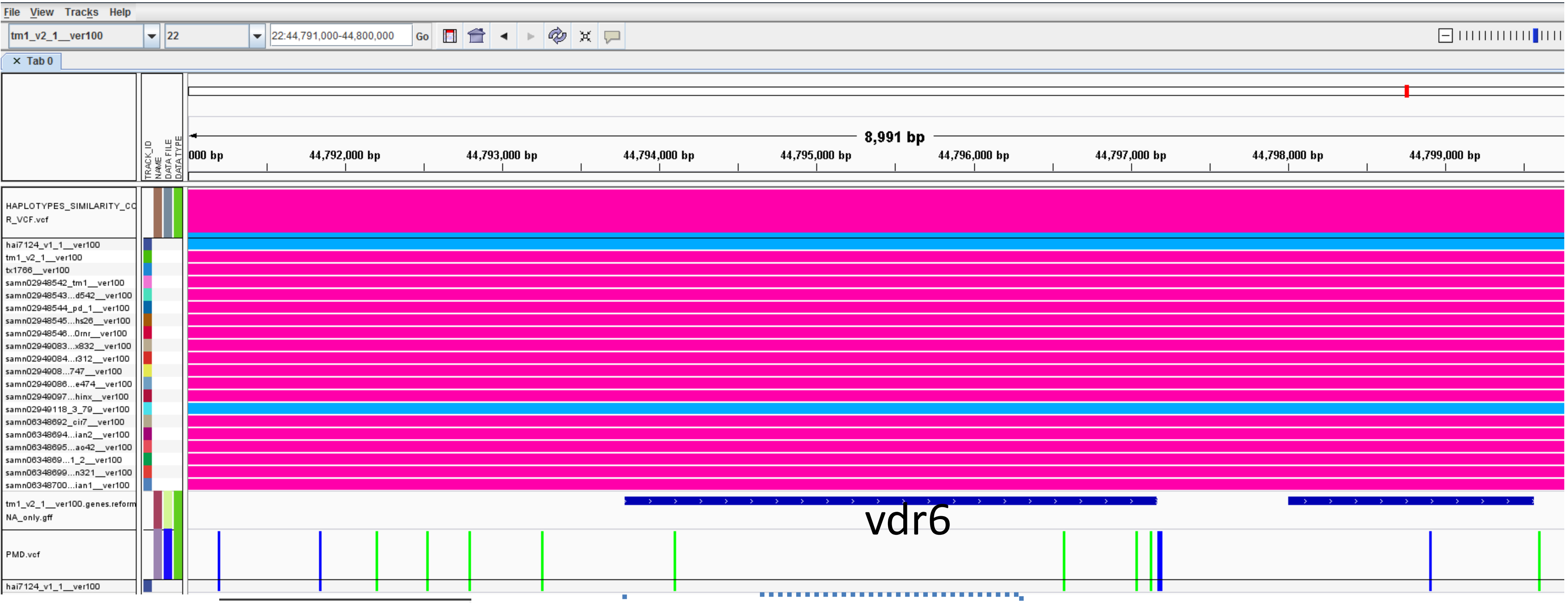
color	name
green	SNP
red	small MultiNP
blue	small indel
grey	large indel

Gbvdr6, a gene encoding a receptor-like protein of cotton (G. barbadense, confers resistance to verticillium wilt in Arabidopsis and upland cotton, Yang et al, Front. Plant Sci 2017.

- Verticilium wilt resistance has a unique haplotype derived from G. barbadense, all other upland lines have similar haplotype.

haplotypes

Polymorphism Hai7124 vs TM1



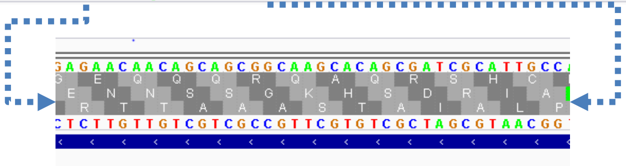
tm1_v2_1__ver100 | 22 | 22:44,791,000-44,800,000

HAPLOTYPES_SIMILARITY_CC R_VCF.vcf
hai7124_v1_1__ver100
tm1_v2_1__ver100
tx1766__ver100
samn02948542_tm1__ver100
samn02948543...d542__ver100
samn02948544_pd_1__ver100
samn02948545...hs26__ver100
samn02948546...0rnr__ver100
samn02949083...x832__ver100
samn02949084...r312__ver100
samn0294908...747__ver100
samn02949086...e474__ver100
samn02949097...hinx__ver100
samn02949118_3_79__ver100
samn06348692_cir7__ver100
samn06348694...ian2__ver100
samn06348695...ao42__ver100
samn0634869...1_2__ver100
samn06348699...n321__ver100
samn06348700...ian1__ver100
tm1_v2_1__ver100.genes.reform NA_only.gff
PMD.vcf
hai7124_v1_1__ver100

G. barb

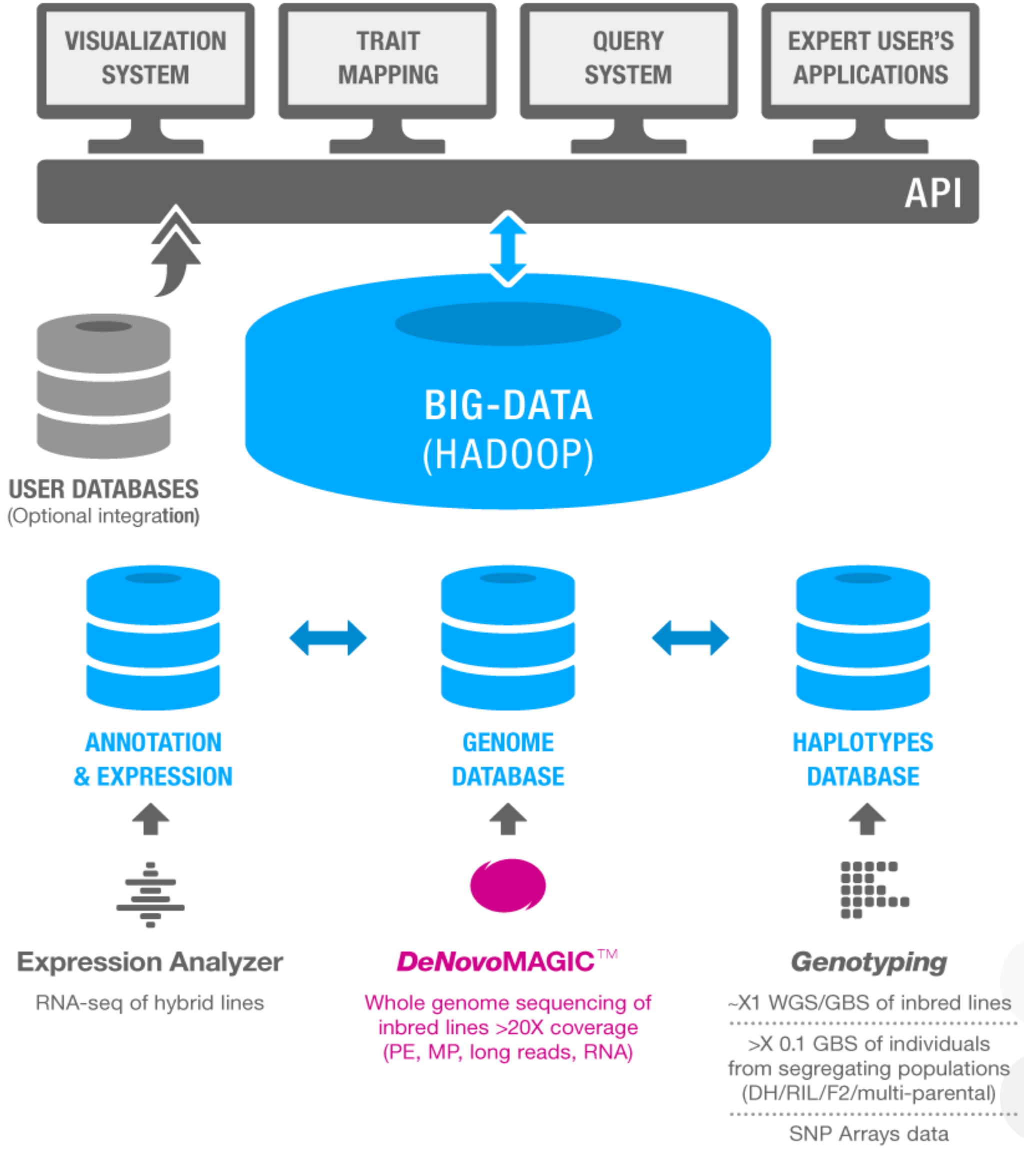
G. hirst

color	name
green	SNP
blue	small indel


GAGAACAACAGCAGCGGCAAGCACAGCGATCGCATTGCC
G E Q Q Q R Q A Q R S H C
E N N S S G K H S D R I A
R T T A A A S T A I A L P
CTCTTGTTGTCGTCGCCGTTCGTGTCGCTAGCGTAACGG

Summary:

- New method of capturing sequence based diversity in cotton using pan-genome positioning.
- Pangenome utilized multiple reference level assembled genomes
- See bulletin, flyer and www.nrgene.com for additional info and contacts
- Pangenome integrates into GenoMAGIC to offer additional capabilities
    - trait mapping, genotyping, marker development, diversity analyses



nrgene

THANK YOU

nrgene

info@nrgene.com | www.nrgene.com | @NRGene