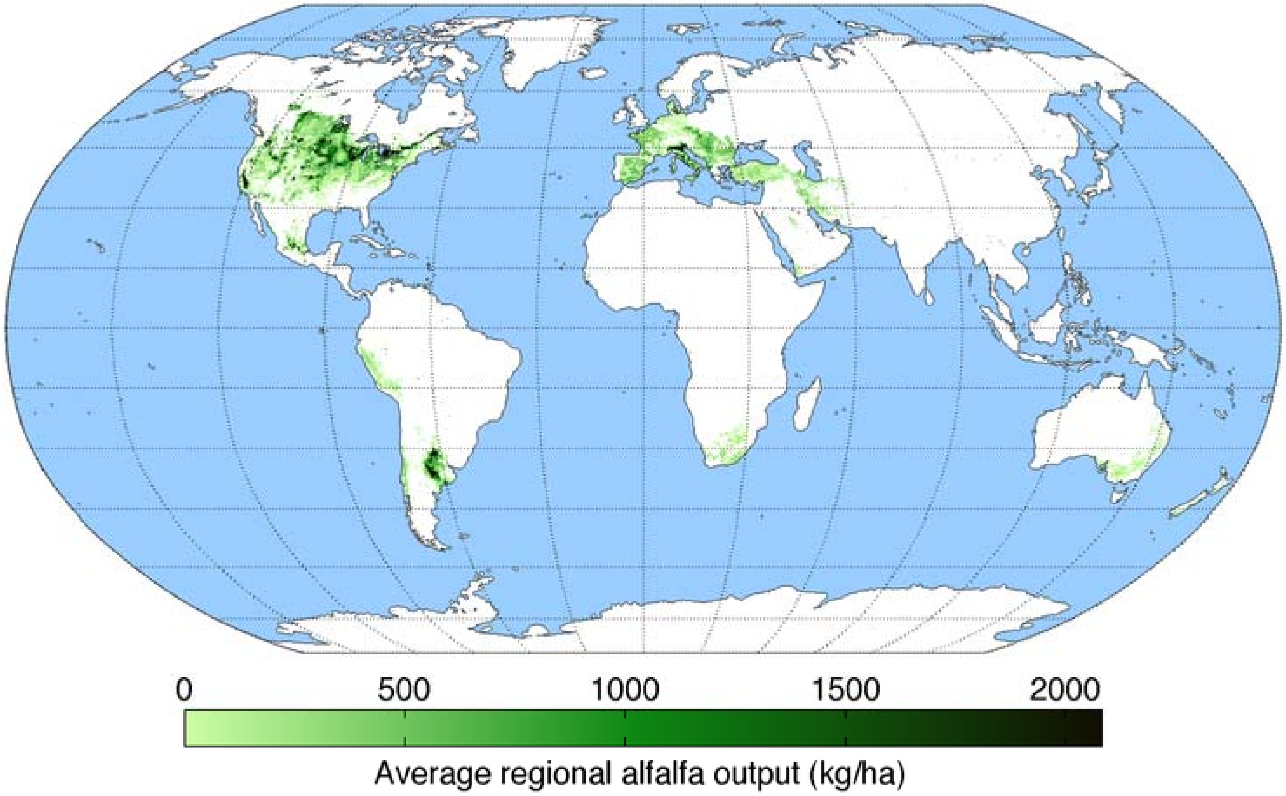
The Alfalfa Breeder's Toolbox Integrates Genomic, Genetic and Germplasm Resources for Alfalfa Improvement

Maria J. Monteros, Chunlin He, Jaeyoung Choi, Patrick Zhao,
Nadim Tayeh, Xinbin Dai, Andrew Farmer,
Joann Mudge, Haibao Tang, Junil Chang, Nick Krom, Justin Vaughn,
Perdeep Mehta, Christy Motes, Michael Trammell, Brian Motes,
Shawn Sullivan, Ivan Liachko, Charles Brummer, Nevin Young,
Chris Town, Michael Udvardi

NAAIC, Trifolium and Grass Breeders Conference
June 5, 2018
Logan, UT

NOBLE
RESEARCH
INSTITUTE

Alfalfa (*Medicago sativa* L.) is a Widely Cultivated Forage Legume Worldwide



Monfreda et al. 2008. G. Biogeo. Cycles.

- Autotetraploid
- $2n = 4x = 32$
- Genome size: ~840 Mbp
- Perennial
- Relative of *M. truncatula*
- Outcrossing
- Cultivars are multi-parent synthetics (heterozygous)

NOBLE RESEARCH INSTITUTE

Alfalfa - *Medicago sativa* Complex

Diploid
2X

Medicago sativa
subsp. *caerulea*

Medicago sativa
subsp. *falcata*

Medicago sativa
subsp. *hemicycla*

Tetraploid
4X

Medicago sativa
subsp. *sativa*

Warm weather

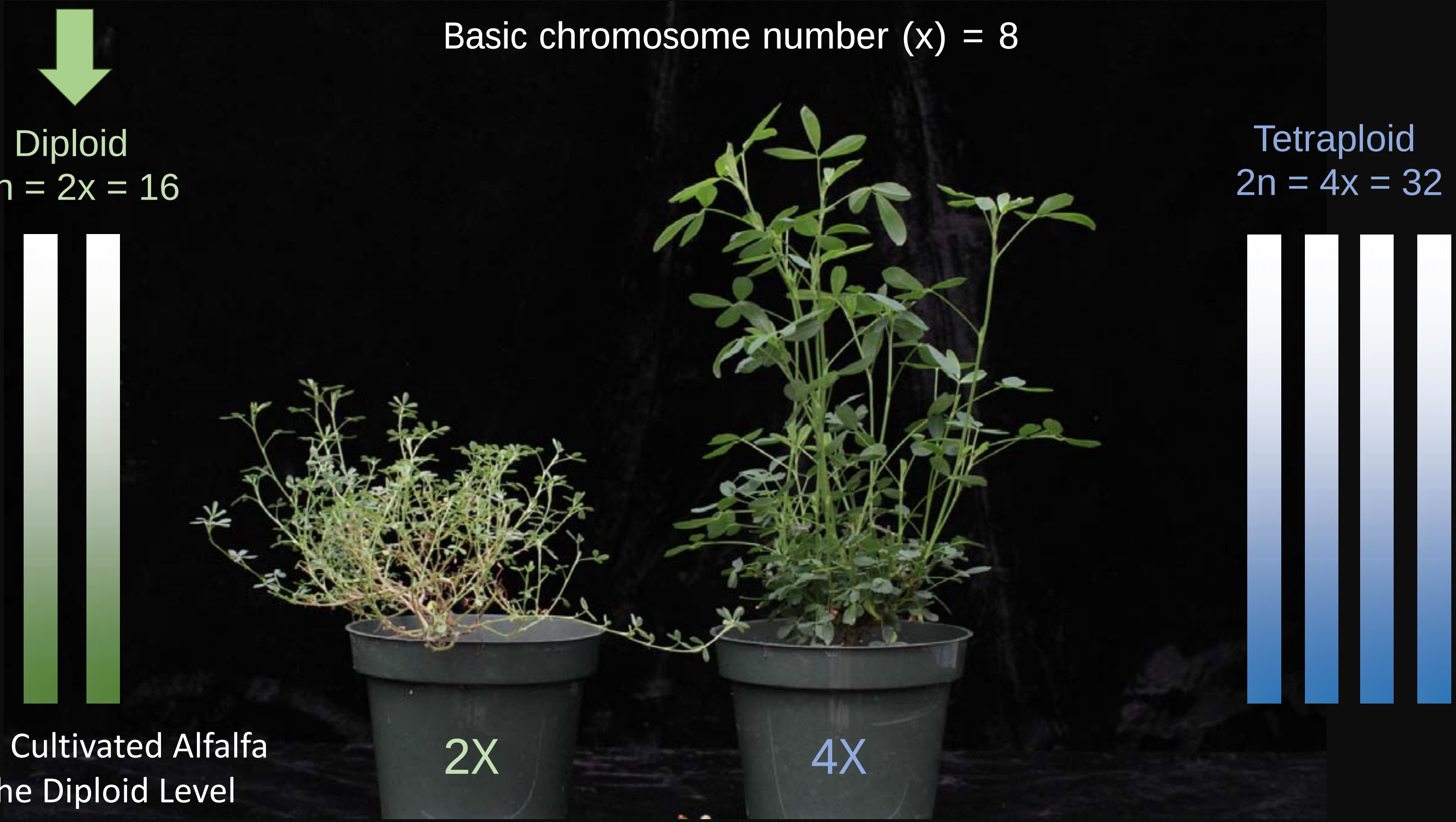
Medicago sativa
subsp. *falcata*

Cold tolerant

Medicago sativa
subsp. *varia*

NOBLE
RESEARCH
INSTITUTE

Alfalfa (*Medicago sativa*) Genetics



Basic chromosome number (x) = 8

Diploid
$2n = 2x = 16$

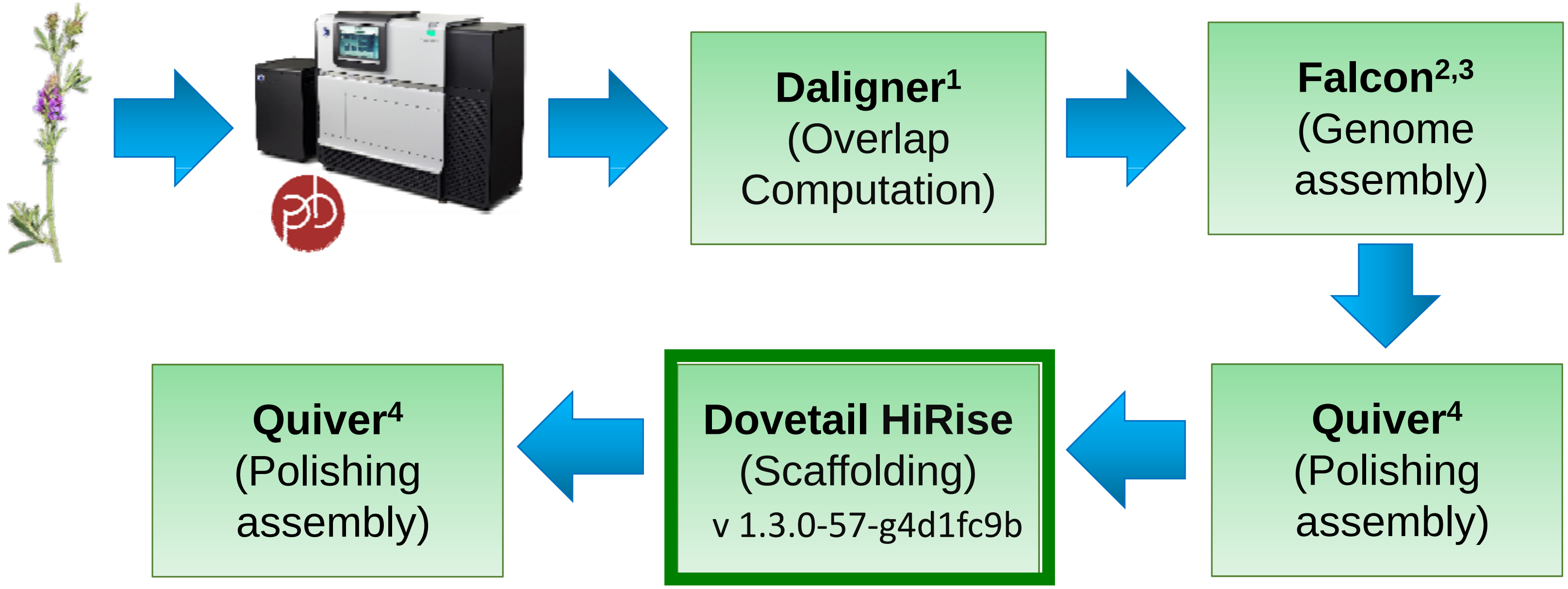
Tetraploid
$2n = 4x = 32$

CADL: Cultivated Alfalfa
at the Diploid Level

Developed by Ted Bingham from cultivated tetraploids

NOBLE RESEARCH INSTITUTE

CADL Genome Assembly Pipeline (V.1.0)



Daligner[1]
(Overlap Computation)

Falcon[2,3]
(Genome assembly)

Quiver[4]
(Polishing assembly)

Dovetail HiRise
(Scaffolding)
v 1.3.0-57-g4d1fc9b

Quiver[4]
(Polishing assembly)

1. Myers G. 2014. The Daligner Overlap Library. https://github.com/thegenemyers/DALIGNER.
2. Chin C, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 10:563–569.
3. Chin J. 2015. FALCON: experimental PacBio diploid assembler. https://github.com/PacificBiosciences/FALCON.
4. PacBio® variant consensus caller (Quiver algorithm). https://github.com/PacificBiosciences/GenomicConsensus.

Source: Joann Mudge, NCGR

NOBLE
RESEARCH
INSTITUTE

CADL Dovetail Assembly (V.1.0)

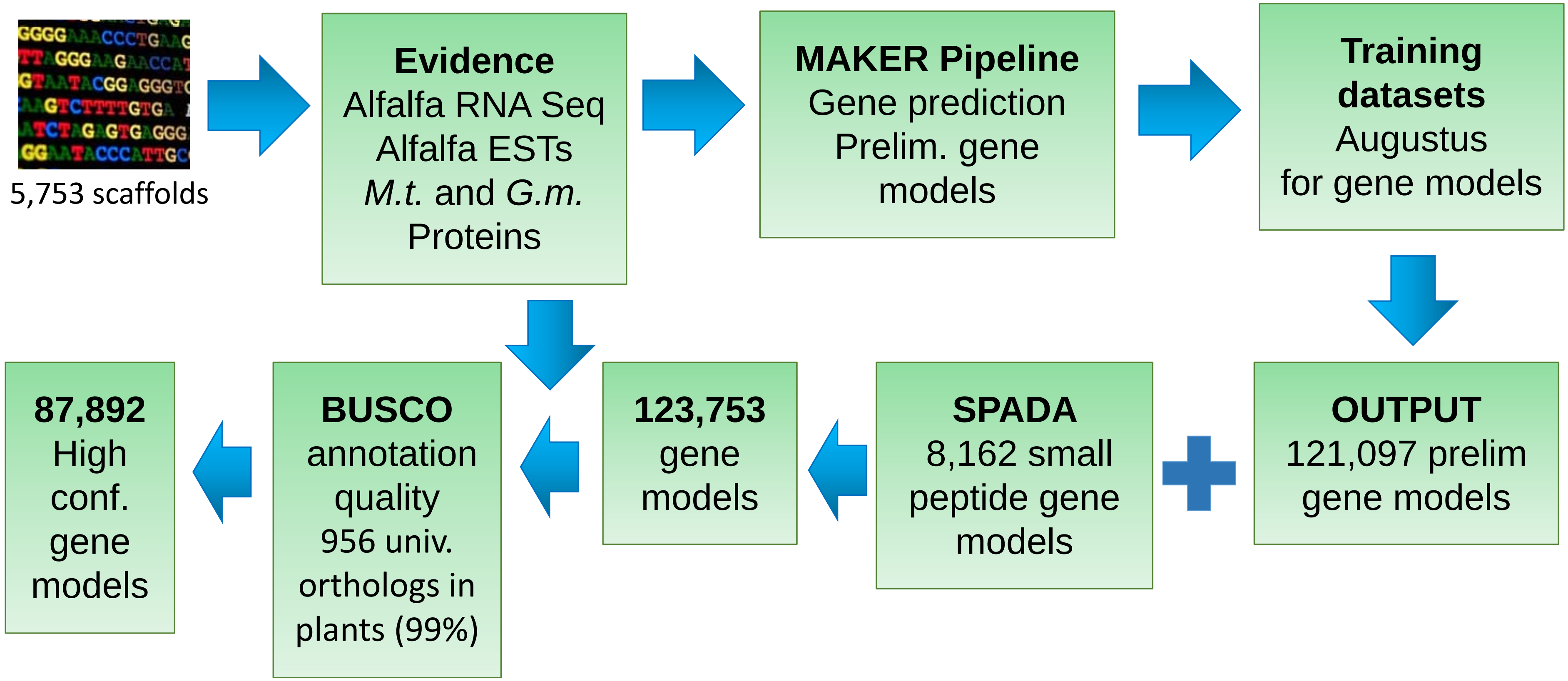
(Falcon 0.4 + Quiver + Dovetail + Quiver)

CADL 1.0 genome submitted to NCBI

	CADL Assembly
Contigs	6,921
Max Contig	2,901,187
Mean Contig	180,749
Contig N50	694,594 bp
Total Contig Length	1,250,961,487 bp
Scaffolds	**5,753**
Max Scaffold	6,073,685 bp
Mean Scaffold	217,463
Scaffold N50	1,271,357 bp
Total Scaffold Length	1,251,062,122

Source: Joann Mudge, NCGR

CADL Gene Annotation Pipeline (V.1.0)



5,753 scaffolds

Evidence
Alfalfa RNA Seq
Alfalfa ESTs
M.t. and *G.m.*
Proteins

MAKER Pipeline
Gene prediction
Prelim. gene
models

Training datasets
Augustus
for gene models

OUTPUT
121,097 prelim
gene models

\+

SPADA
8,162 small
peptide gene
models

123,753
gene
models

BUSCO
annotation
quality
956 univ.
orthologs in
plants (99%)

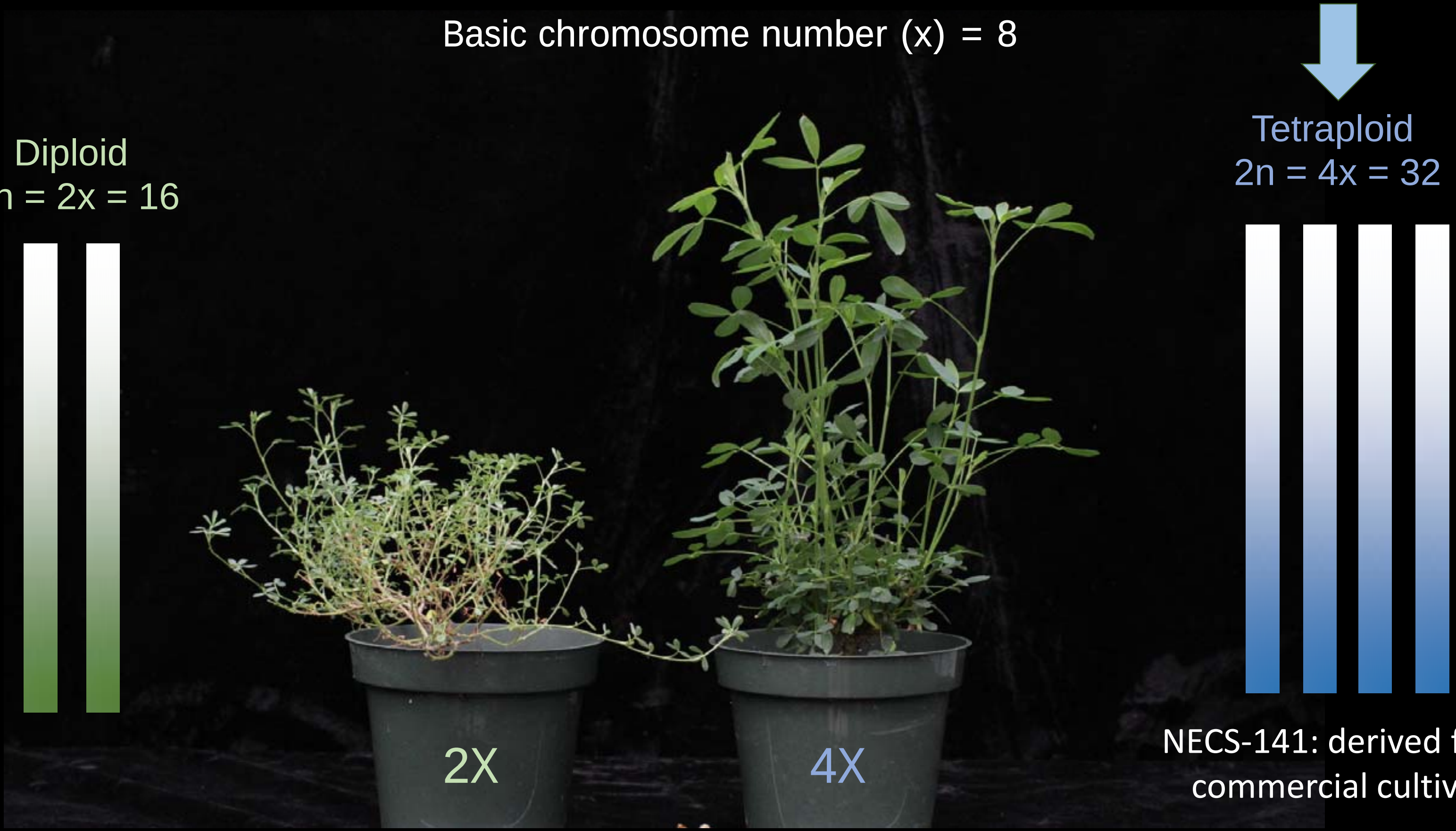
87,892
High
conf.
gene
models

Source: Xinbin Dai, Patrick Zhao

Campbell MS et al. 2014. Maker pipeline for genome annotation and curation. Curr. Prot. Bioinformatics. 48:1-39.

Alfalfa (*Medicago sativa*) Genetics

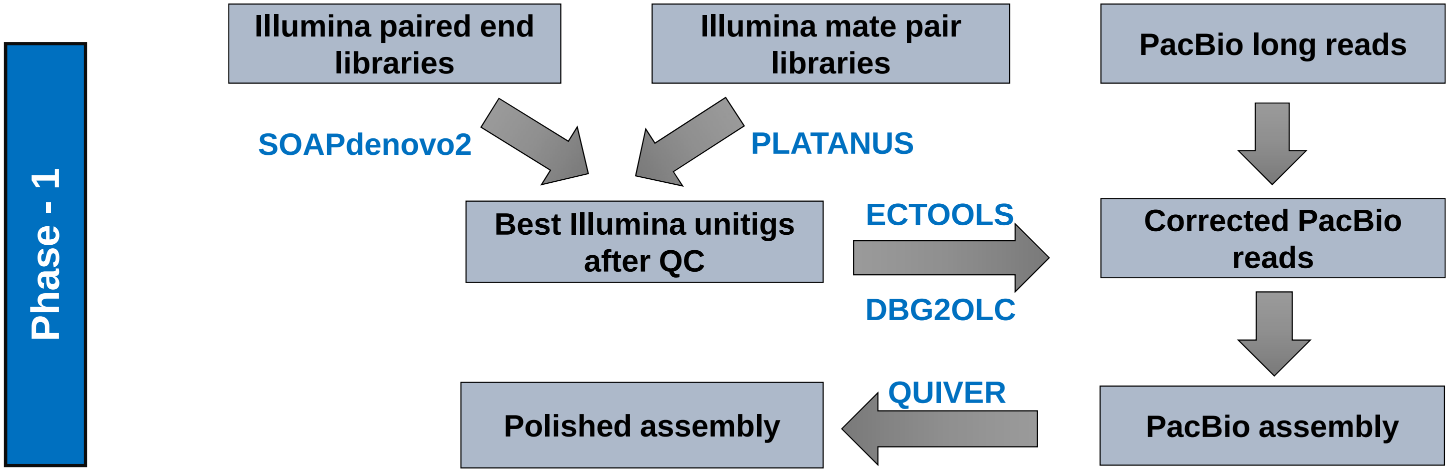
Basic chromosome number (x) = 8



Diploid
$2n = 2x = 16$

Tetraploid
$2n = 4x = 32$

NECS-141: derived from commercial cultivars



Phase - 1

Illumina paired end libraries

Illumina mate pair libraries

PacBio long reads

SOAPdenovo2

PLATANUS

Best Illumina unitigs after QC

ECTOOLS

DBG2OLC

Corrected PacBio reads

QUIVER

Polished assembly

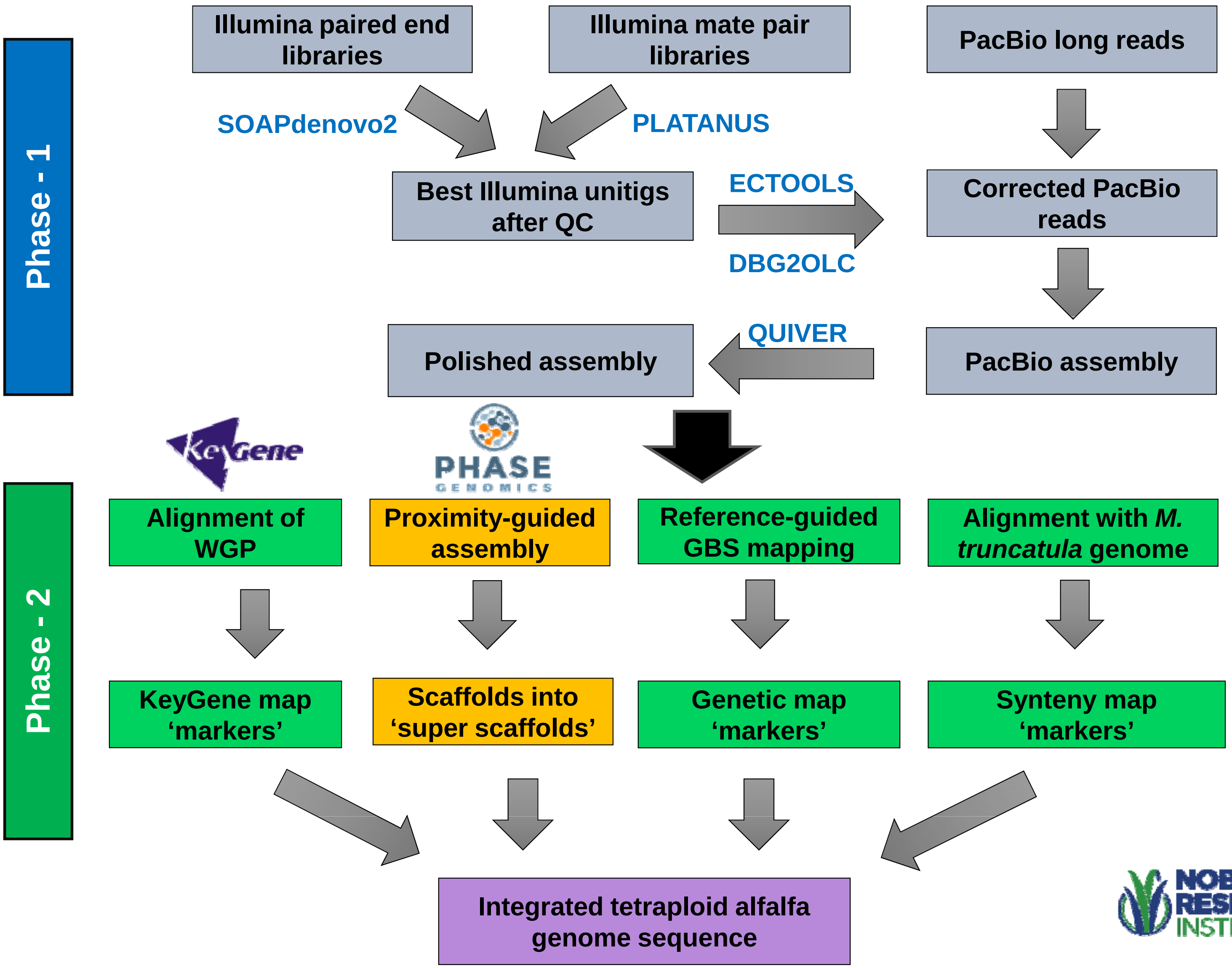
PacBio assembly

Assembly Strategies for Tetraploid Alfalfa Genome

Platform	Assembler	# of scaffolds	Total length	Maximum length	Contig N50	Scaffold N50	CEGMA complete genes	Transcript coverage
Illumina	ABYSS	263,746	897Mb	29,439,826	3,304	5,021	68.95%	86.6%
	SOAPdenovo2	359,556	1096Mb	79,123	4,662	6,303	81.05%	94.7%
	PLATANUS	780,257	863Mb	426,269	7,302	33,927	65.73%	91.8%
PacBio	**Falcon**	**50,072**	1607Mb	762,063	62,942	62,942	87.90%	93.6%
	Celera Assembler	**47,839**	2682Mb	3,289,947	81,910	81,910	**92.74%**	**95.4%**
	CANU	**12,893**	2349Mb	4,086,099	310,769	310,769	**92.34%**	**95.4%**

Assemblies by Haibao Tang and Andrew Farmer

NOBLE RESEARCH INSTITUTE



Phase - 1

Illumina paired end libraries

Illumina mate pair libraries

PacBio long reads

SOAPdenovo2

PLATANUS

Best Illumina unitigs after QC

ECTOOLS

DBG2OLC

Corrected PacBio reads

PacBio assembly

QUIVER

Polished assembly

Phase - 2

Alignment of WGP

Proximity-guided assembly

Reference-guided GBS mapping

Alignment with *M. truncatula* genome

KeyGene map ‘markers’

Scaffolds into ‘super scaffolds’

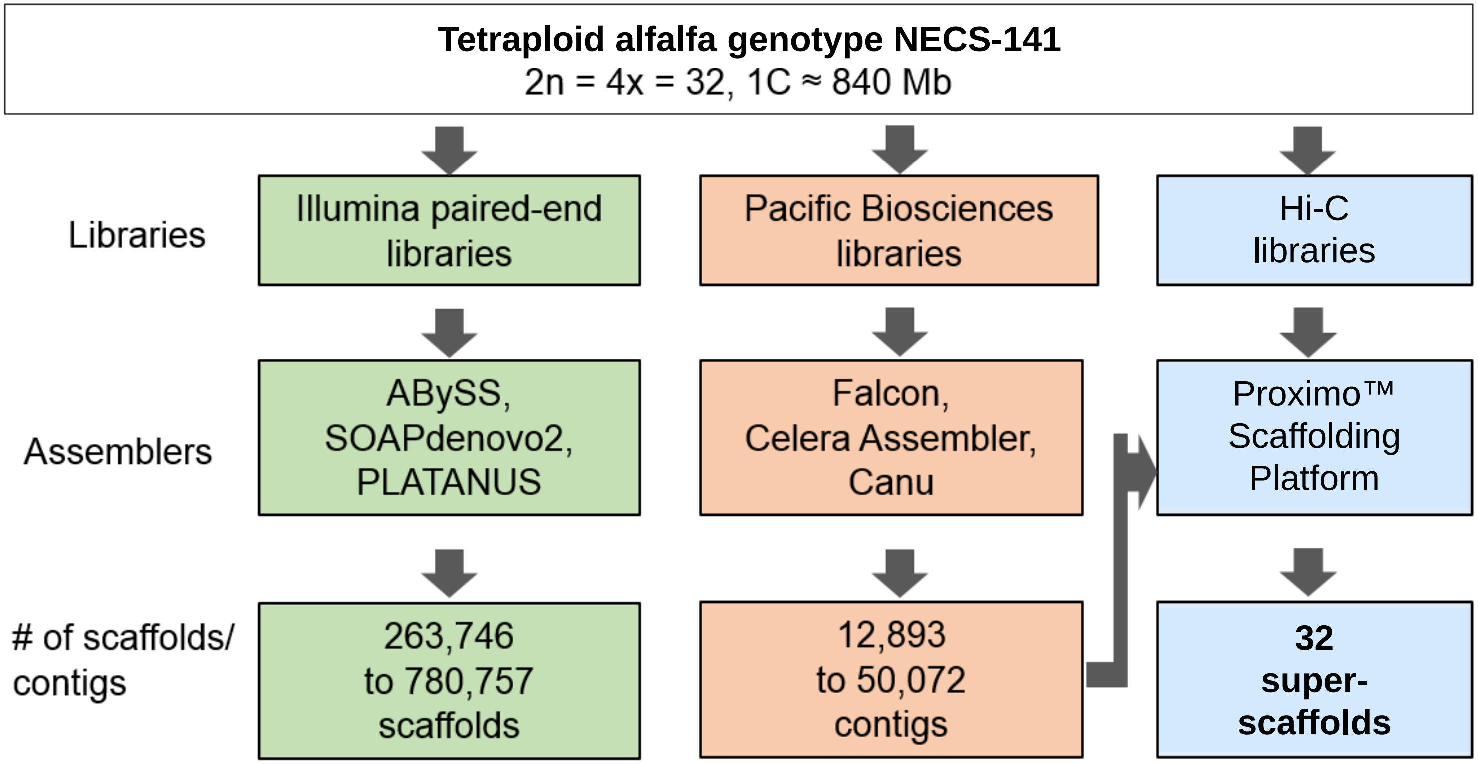
Genetic map ‘markers’

Synteny map ‘markers’

Integrated tetraploid alfalfa genome sequence

NOBLE RESEARCH INSTITUTE

Multiple Approaches for Sequencing the 4X Alfalfa Genome



Tetraploid alfalfa genotype NECS-141
2n = 4x = 32, 1C ≈ 840 Mb

Libraries

Illumina paired-end libraries

Pacific Biosciences libraries

Hi-C libraries

Assemblers

ABySS, SOAPdenovo2, PLATANUS

Falcon, Celera Assembler, Canu

Proximo™ Scaffolding Platform

# of scaffolds/ contigs

263,746 to 780,757 scaffolds

12,893 to 50,072 contigs

32 super-scaffolds

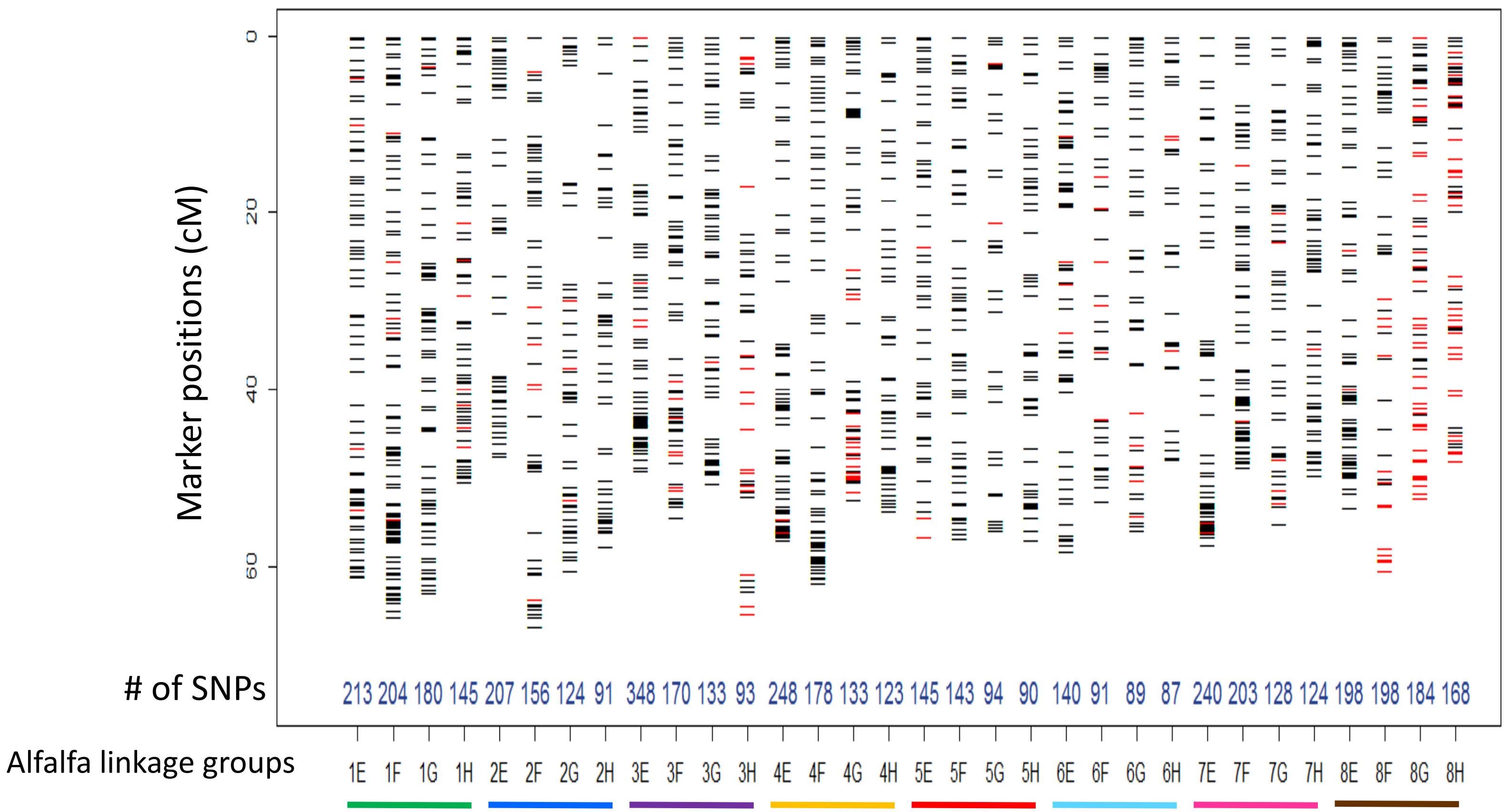
NCGR National Center for Genome Resources

NCGR National Center for Genome Resources

PHASE GENOMICS

NOBLE RESEARCH INSTITUTE

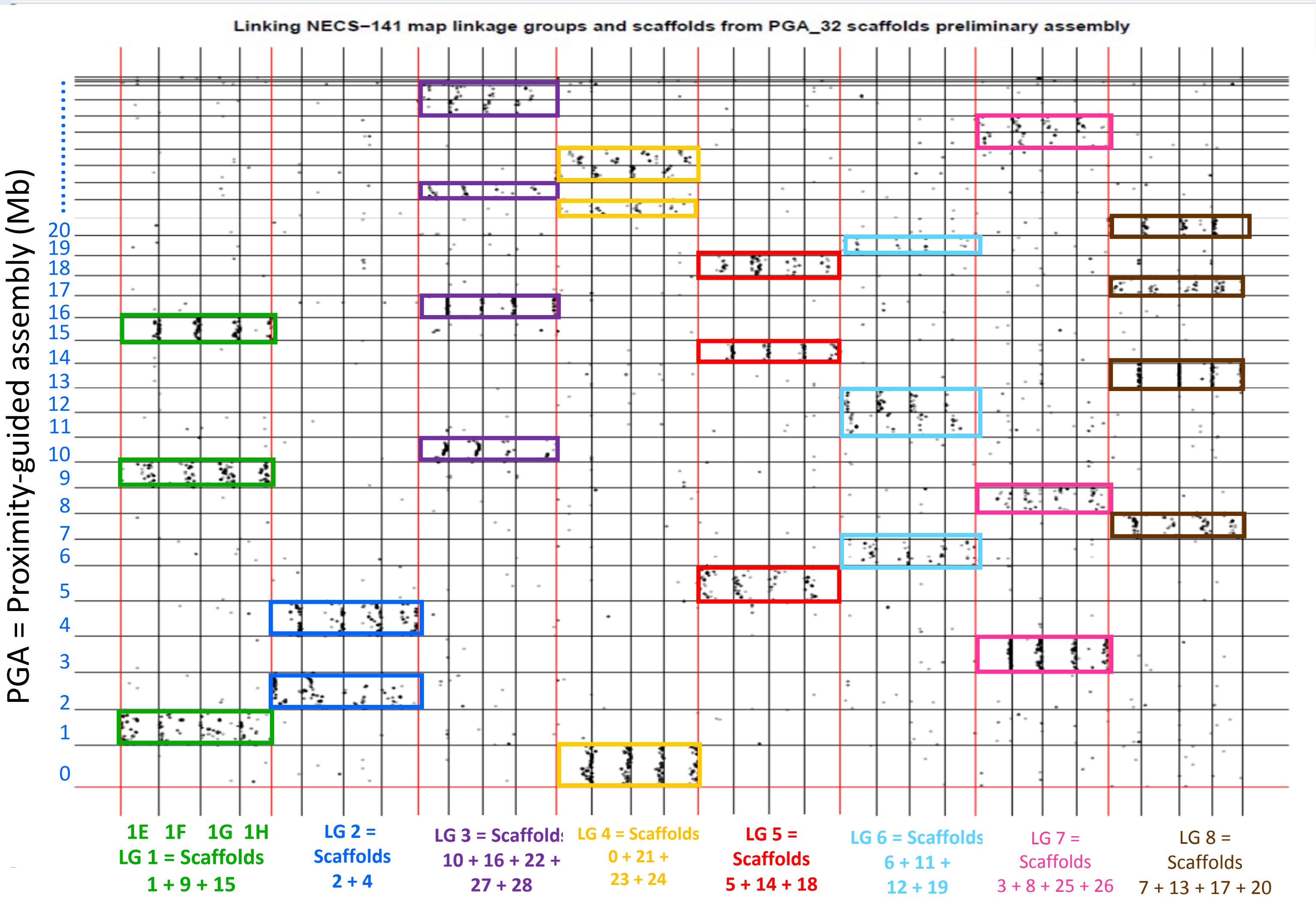
NECS-141 GBS-SNP Genetic Linkage Map



4 haplotypes per chromosome

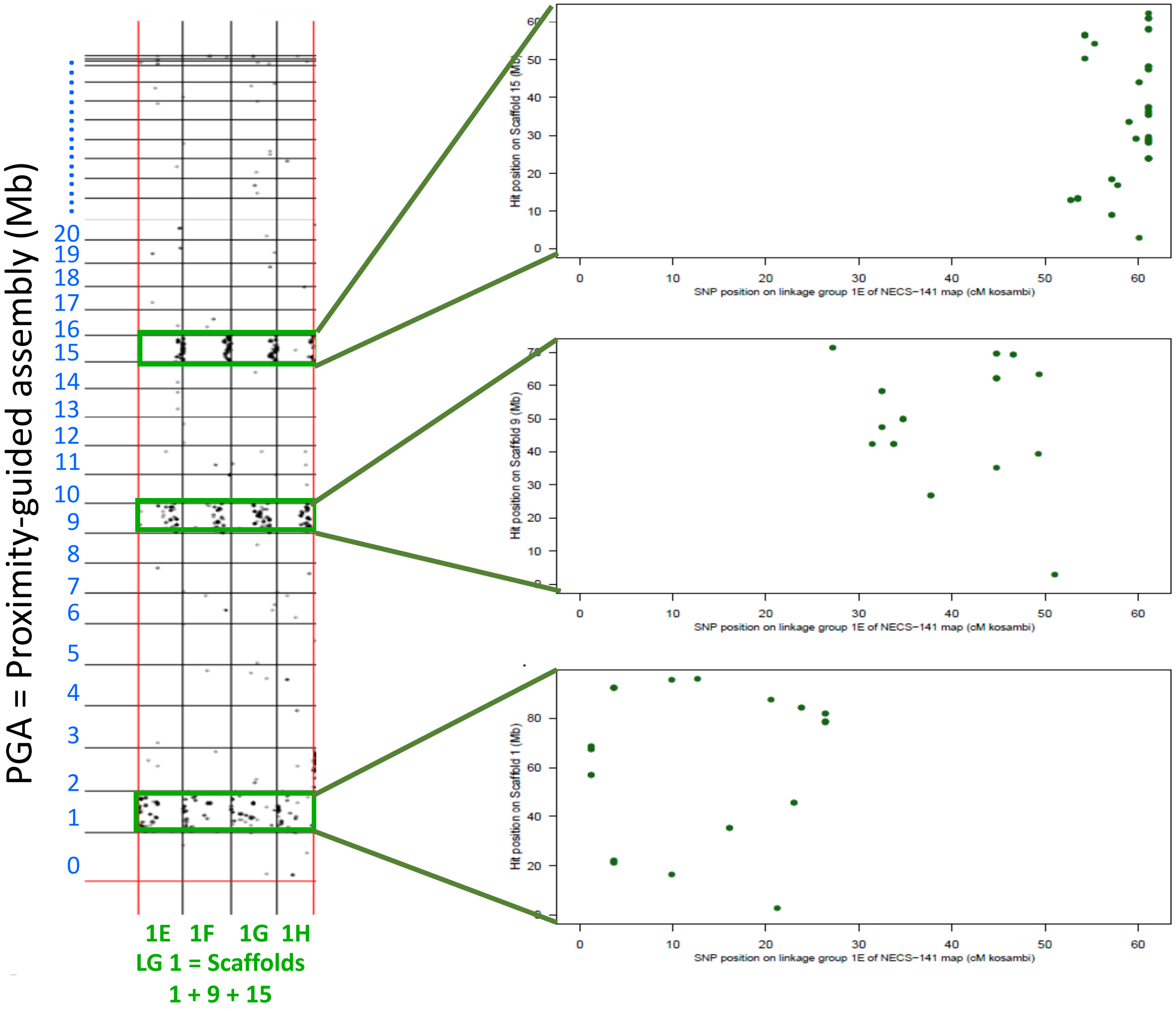
N. Tayeh, unpublished

Proximity-Guided Assembly + NECS-141 Map


Linking NECS−141 map linkage groups and scaffolds from PGA_32 scaffolds preliminary assembly

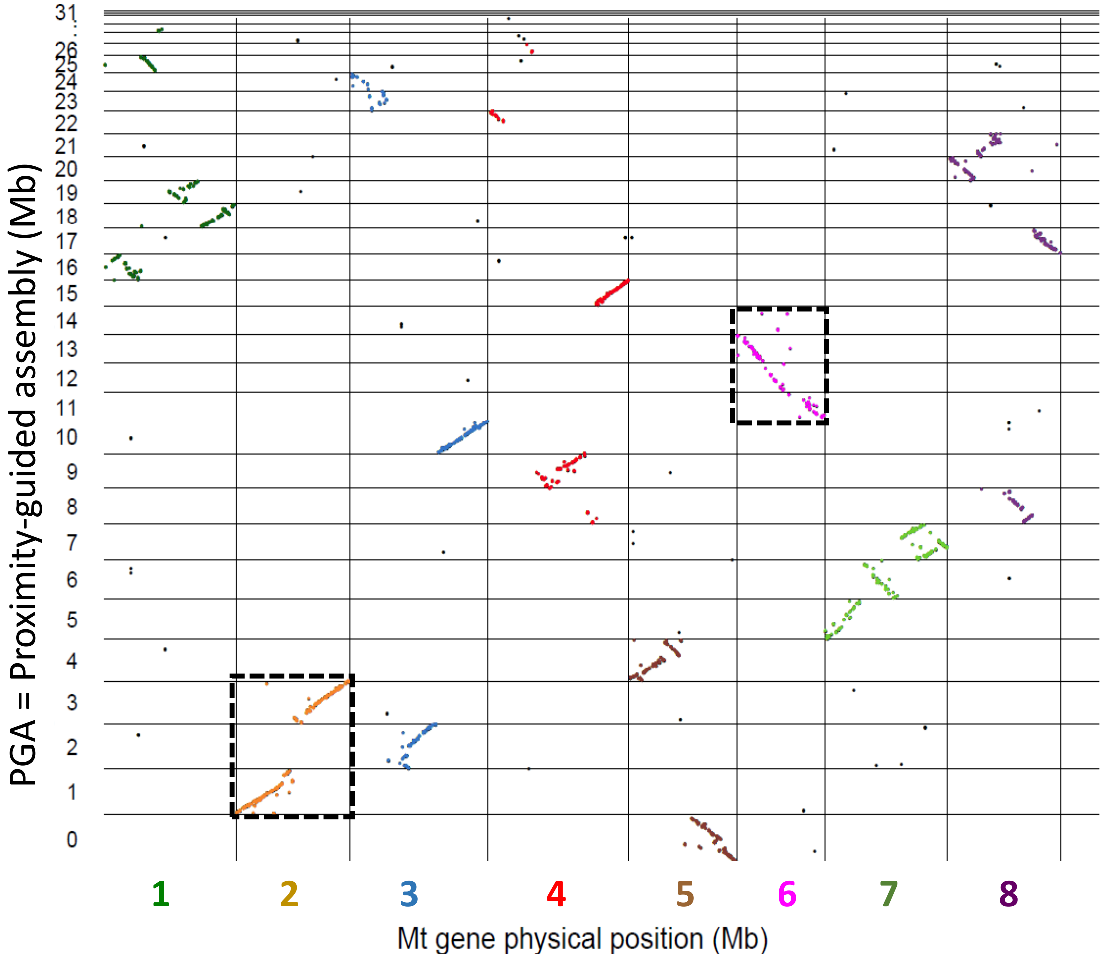
Each alfalfa chromosome is represented by only two to five 'super-scaffolds'

Proximity-Guided Assembly + NECS-141 Map



Alfalfa chromosome 1 is represented by three super-scaffolds: 1, 9 and 15.

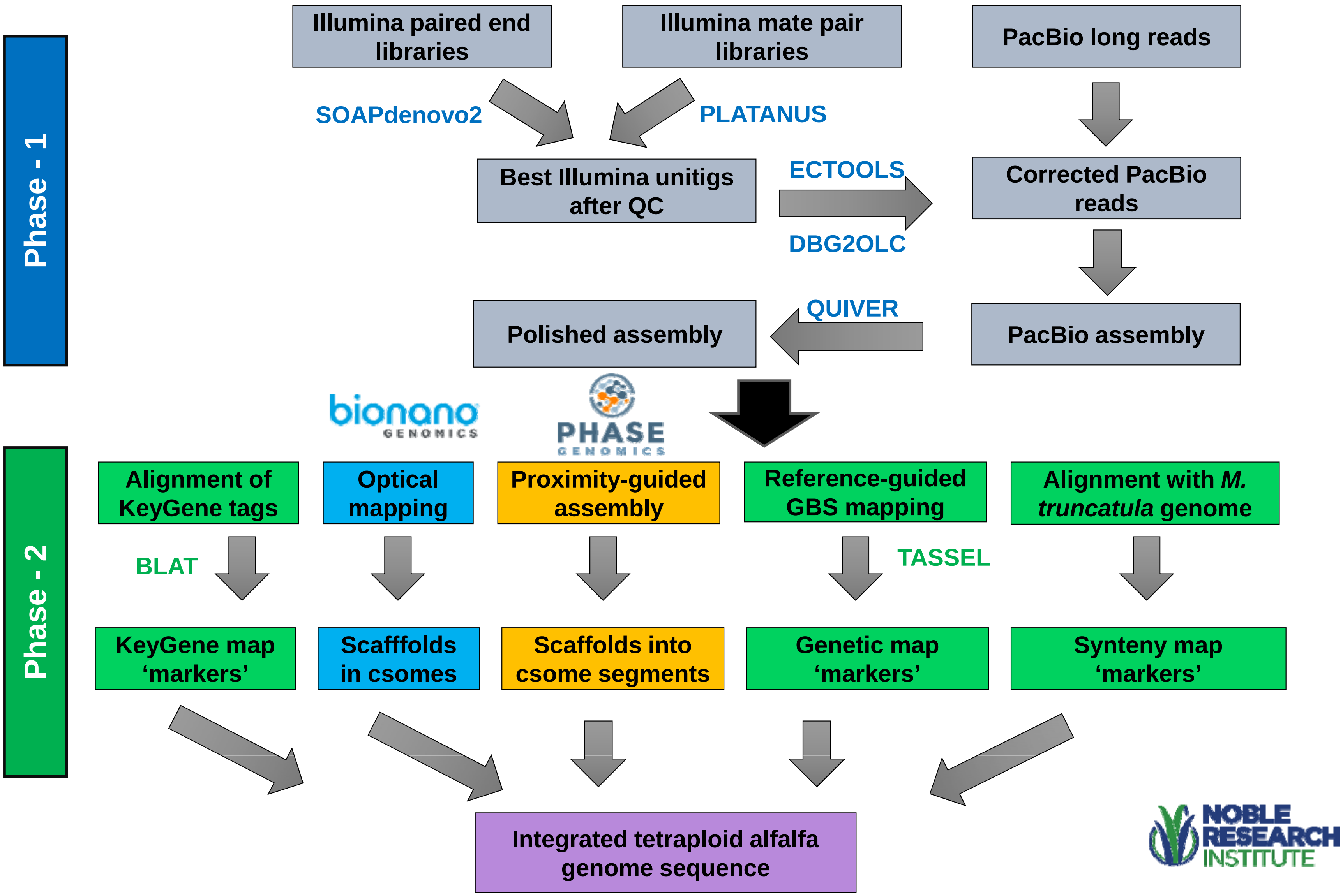
Alignment of the 4X Alfalfa Genome with *M. truncatula* (V4.0)



Chromo-some	Super-scaffolds
1	16, 18, 19, 25, 28
2	1, 3
3	2, 10, 23, 24
4	9, 15, 22, 26
5	0, 4
6	11, 12, 13, 14
7	5, 6, 7
8	8, 17, 20, 21

N. Tayeh, manuscript in preparation

NOBLE RESEARCH INSTITUTE



Phase - 1

Illumina paired end libraries

Illumina mate pair libraries

PacBio long reads

SOAPdenovo2

PLATANUS

Best Illumina unitigs after QC

ECTOOLS

DBG2OLC

Corrected PacBio reads

Polished assembly

QUIVER

PacBio assembly

Phase - 2

bionano GENOMICS

PHASE GENOMICS

Alignment of KeyGene tags

BLAT

KeyGene map 'markers'

Optical mapping

Scafffolds in csomes

Proximity-guided assembly

Scaffolds into csome segments

Reference-guided GBS mapping

TASSEL

Genetic map 'markers'

Alignment with *M. truncatula* genome

Synteny map 'markers'

Integrated tetraploid alfalfa genome sequence

NOBLE RESEARCH INSTITUTE

Optical Mapping - Hybrid Scaffold Alignments

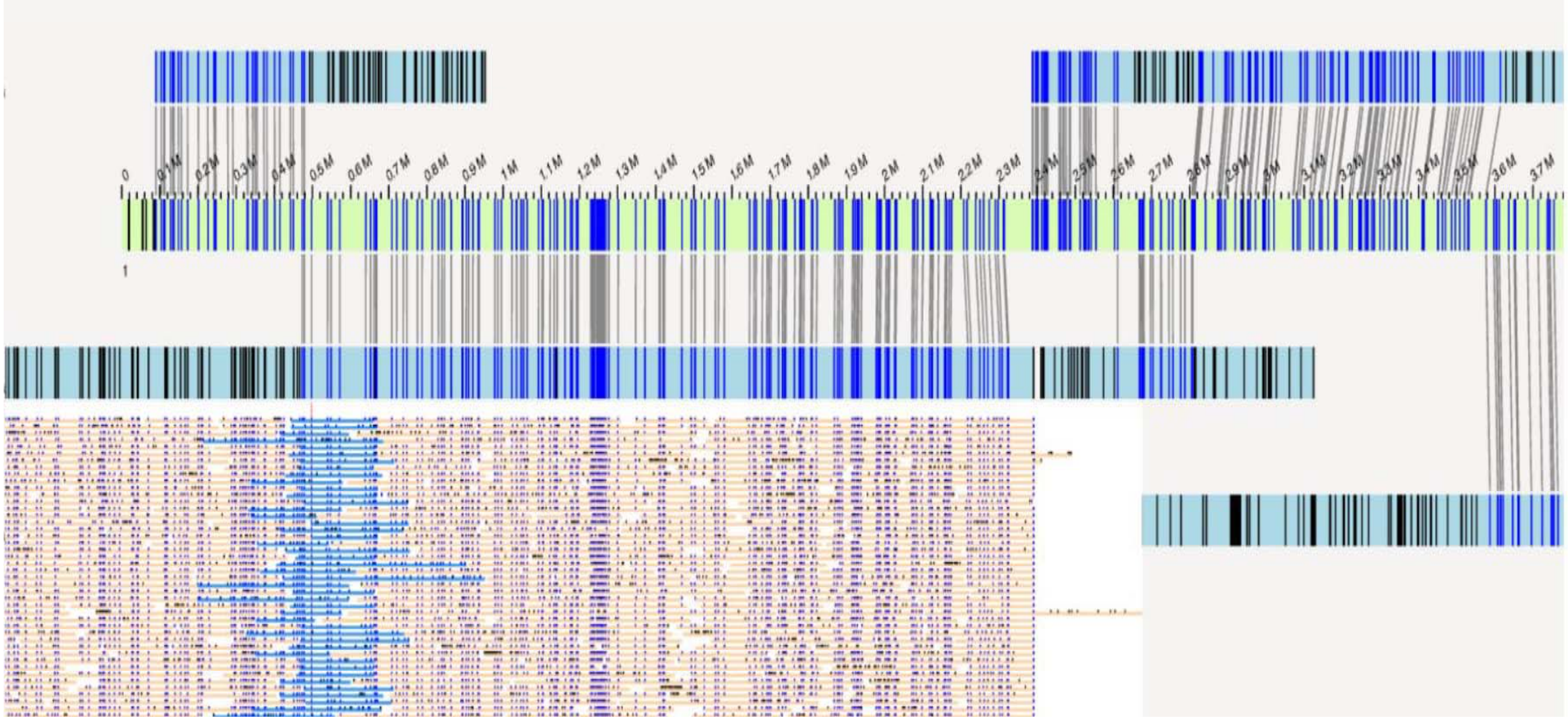
(PacBio-Canu + Optical Mapping)

Genome maps BspQI

Canu ctg1

Genome maps BspQI

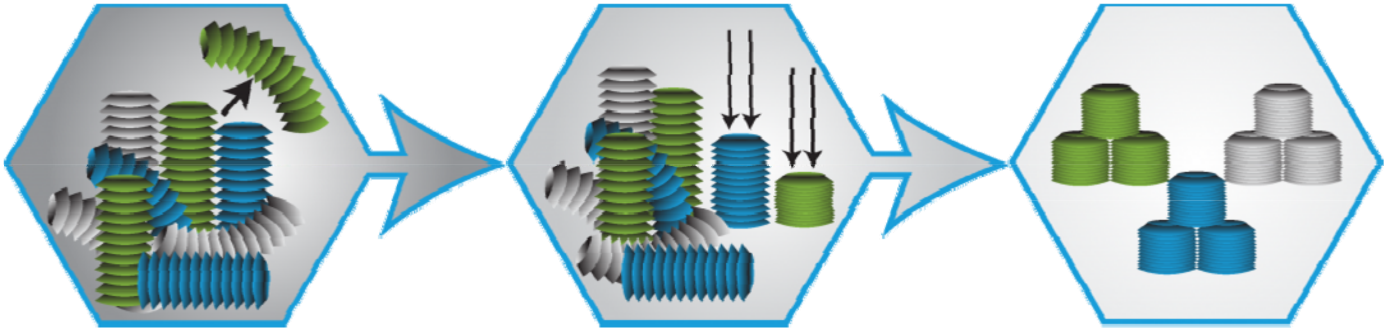
Optical mapping
Single molecules



2 haplotypes

3 haplotypes

Challenges with Big Datasets

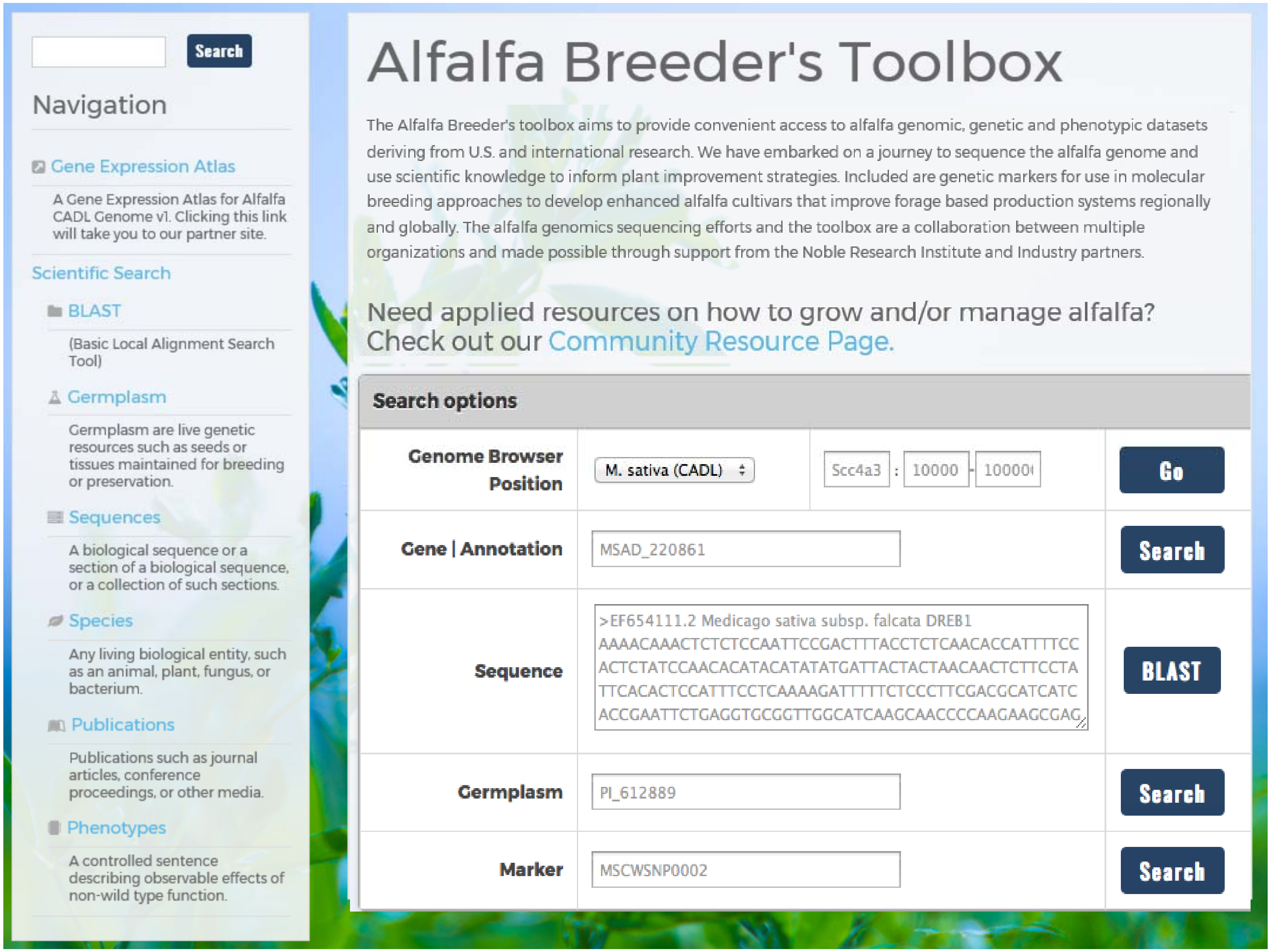


Store

Integrate

Organize

NOBLE RESEARCH INSTITUTE



Navigation

Gene Expression Atlas

A Gene Expression Atlas for Alfalfa CADL Genome v1. Clicking this link will take you to our partner site.

Scientific Search

BLAST

(Basic Local Alignment Search Tool)

Germplasm

Germplasm are live genetic resources such as seeds or tissues maintained for breeding or preservation.

Sequences

A biological sequence or a section of a biological sequence, or a collection of such sections.

Species

Any living biological entity, such as an animal, plant, fungus, or bacterium.

Publications

Publications such as journal articles, conference proceedings, or other media.

Phenotypes

A controlled sentence describing observable effects of non-wild type function.

Alfalfa Breeder's Toolbox

The Alfalfa Breeder's toolbox aims to provide convenient access to alfalfa genomic, genetic and phenotypic datasets deriving from U.S. and international research. We have embarked on a journey to sequence the alfalfa genome and use scientific knowledge to inform plant improvement strategies. Included are genetic markers for use in molecular breeding approaches to develop enhanced alfalfa cultivars that improve forage based production systems regionally and globally. The alfalfa genomics sequencing efforts and the toolbox are a collaboration between multiple organizations and made possible through support from the Noble Research Institute and Industry partners.

Need applied resources on how to grow and/or manage alfalfa? Check out our Community Resource Page.

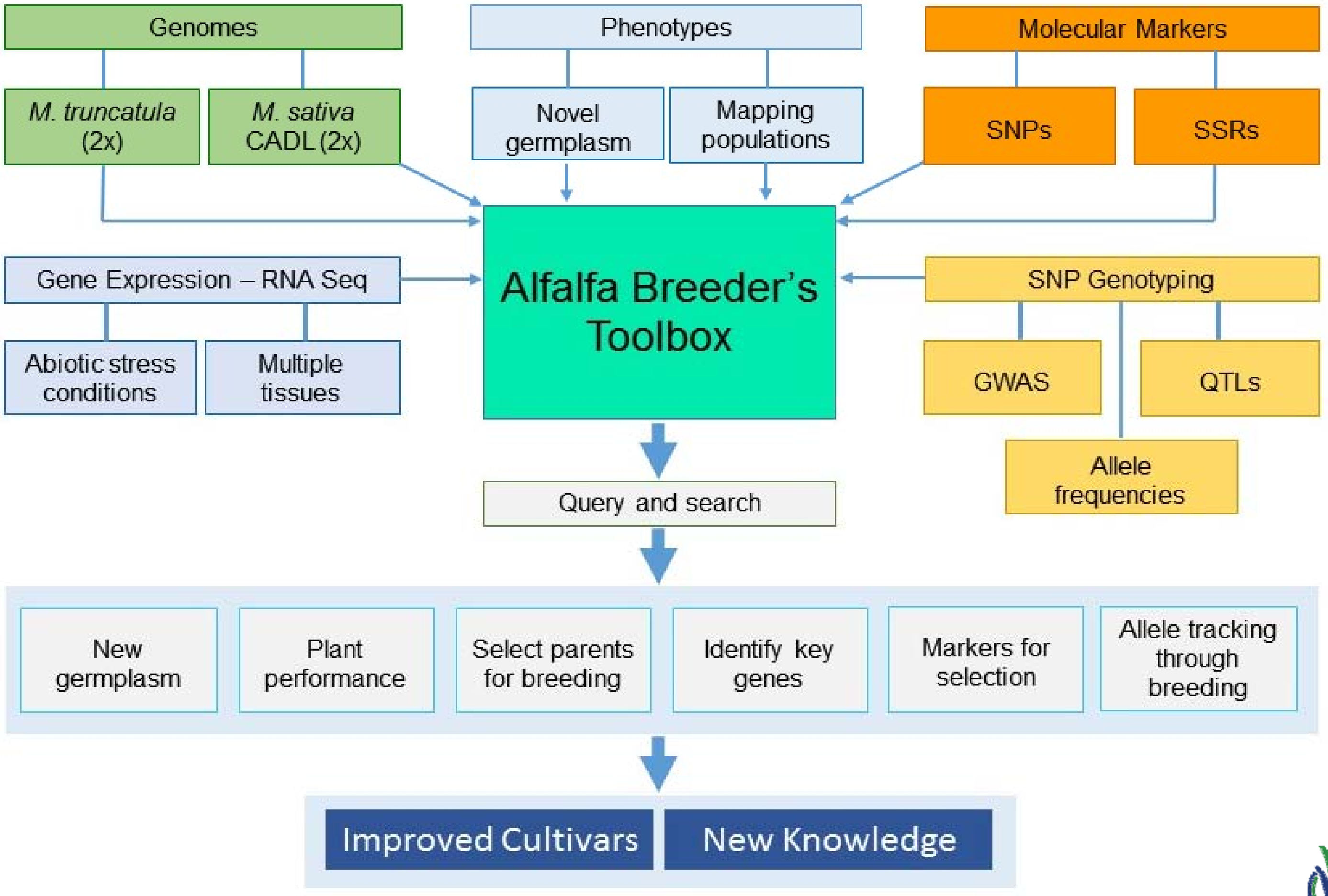
Search options

Genome Browser Position	M. sativa (CADL)	Scc4a3 : 10000 - 10000(	Go
Gene | Annotation	MSAD_220861		Search
Sequence	>EF654111.2 Medicago sativa subsp. falcata DREB1 AAAACAAACTCTCTCCAATTCCGACTTTACCTCTCAACACCATTTTCC ACTCTATCCAACACATACATATATGATTACTACTAACAACTCTTCCTA TTCACACTCCATTTCCTCAAAAGATTTTTCTCCCTTCGACGCATCATC ACCGAATTCTGAGGTGCGGTTGGCATCAAGCAACCCCAAGAAGCGAG		BLAST
Germplasm	PI_612889		Search
Marker	MSCWSNP0002		Search

http://alfalfatoolbox.org/

NOBLE RESEARCH INSTITUTE

Overview of the Alfalfa Breeder’s Toolbox (ABT)



Genomes

M. truncatula (2x)

M. sativa CADL (2x)

Phenotypes

Novel germplasm

Mapping populations

Molecular Markers

SNPs

SSRs

Gene Expression – RNA Seq

Abiotic stress conditions

Multiple tissues

Alfalfa Breeder’s Toolbox

SNP Genotyping

GWAS

QTLs

Allele frequencies

Query and search

New germplasm

Plant performance

Select parents for breeding

Identify key genes

Markers for selection

Allele tracking through breeding

Improved Cultivars

New Knowledge

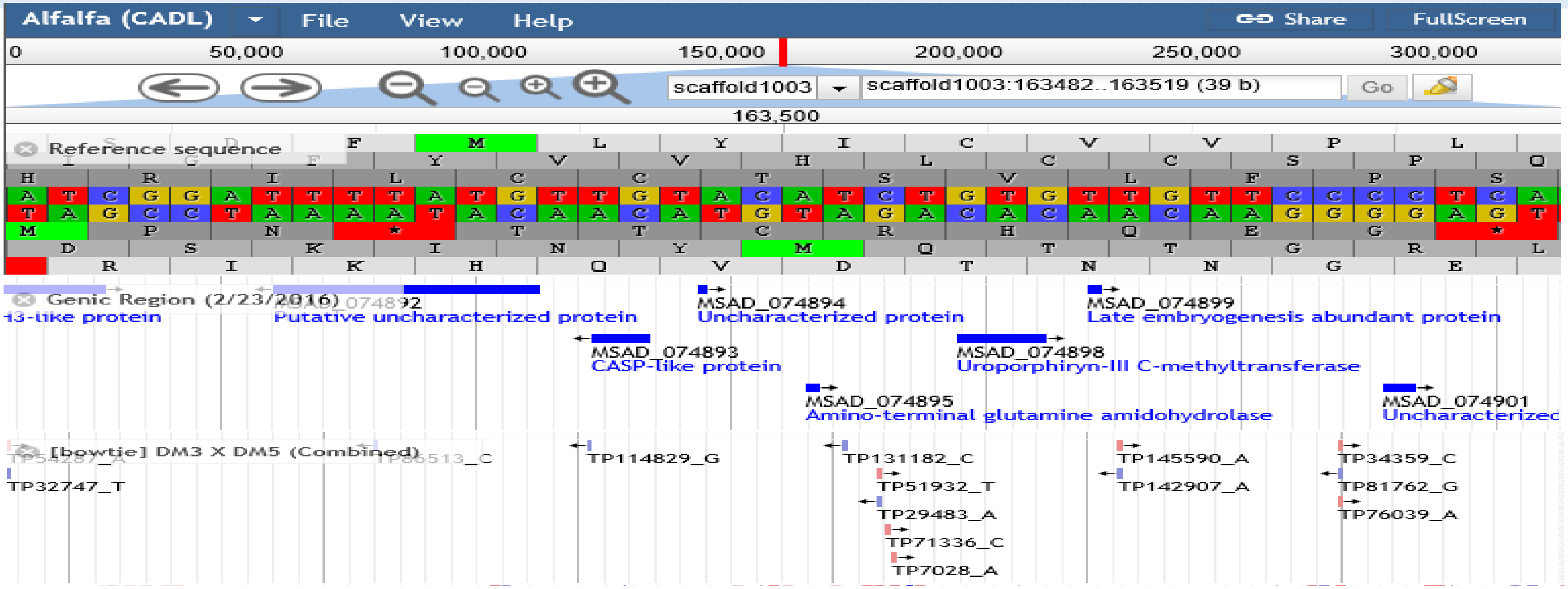
Tripal

Chado

Drupal

NOBLE RESEARCH INSTITUTE

Jbrowse: Integrated Alfalfa Datasets



Ref-seq

Gene models

SNPs

Jbrowse: Integrated Alfalfa Datasets

Ref-seq

Gene models

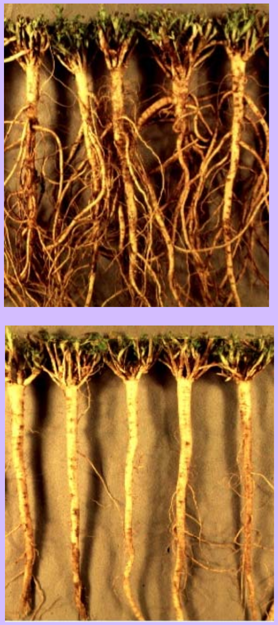
SNPs

Allele freq

A = Green C = Blue G = Yellow T = Red Gray = Monomorphic

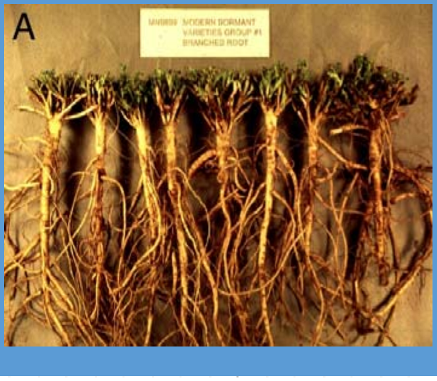
Genetic Differences Associated with Root Traits

Base Population



3 cycles of field-based selection

Selected for Branched Roots

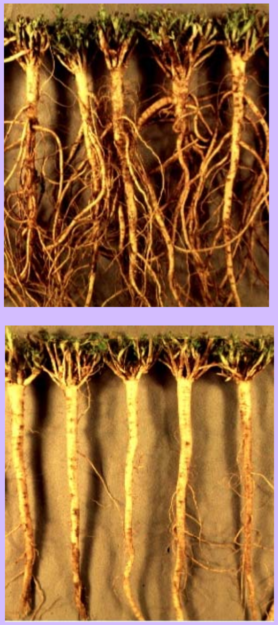


Selected for Tapped Roots

Images: Deborah Samac, USDA

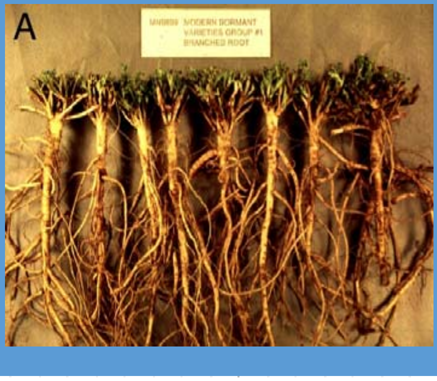
Genetic Differences Associated with Root Traits

Base Population



3 cycles of field-based selection

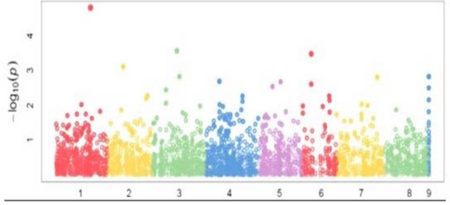
Selected for Branched Roots



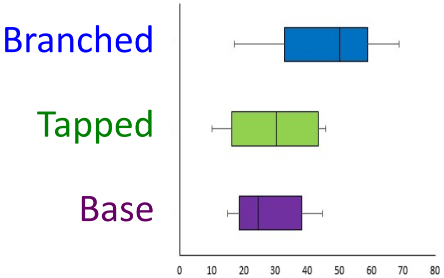
Selected for Tapped Roots

Images: Deborah Samac, USDA

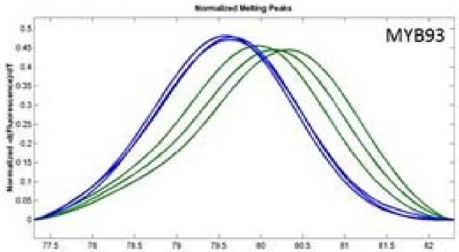
1. Genome-wide Association Study (GWAS)



2. Shifts in the populations



3. Genotyping assay for molecular breeding



Shifts in Allele Frequencies Resulting from Selection

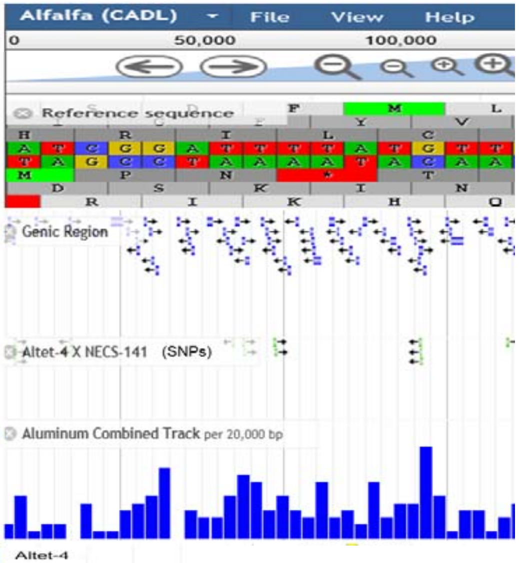
Ref-seq

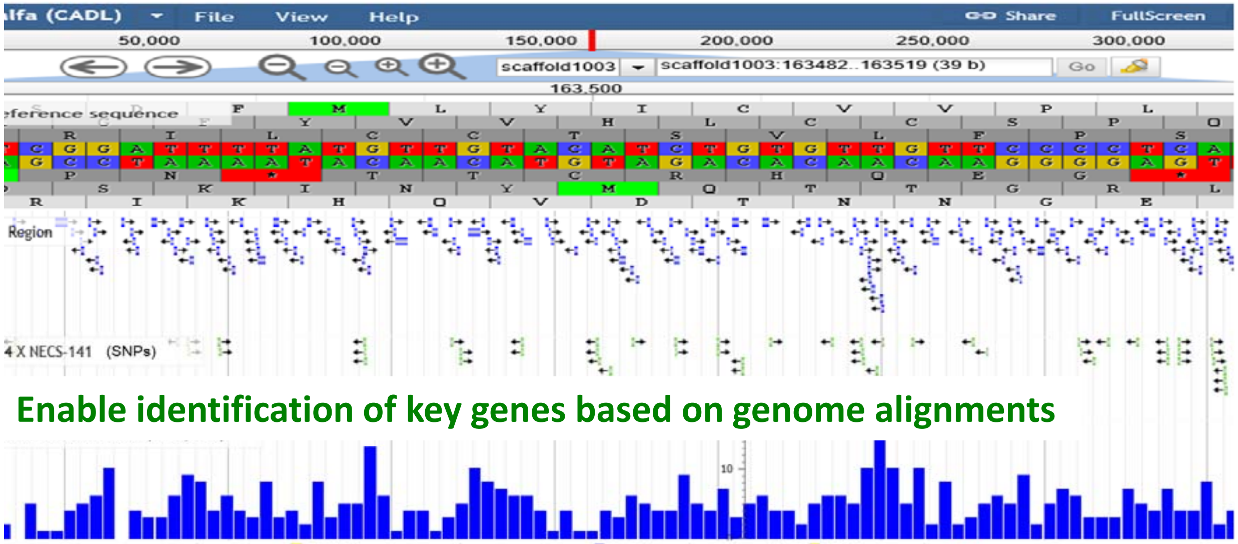
Gene model

SNPs

Gene expression

Allele freq





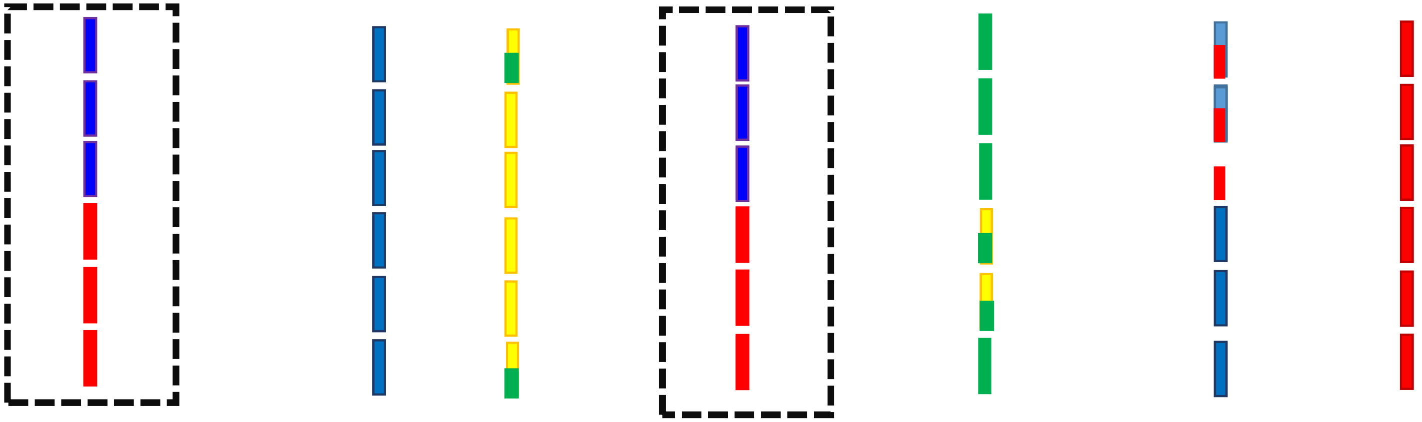
Enable identification of key genes based on genome alignments

Branched

- 3233-6
- 3233-21
- 3233-25

Tapped

- 3234-3
- 3234-4
- 3234-27



A = green C = blue G = yellow T=red

Pipeline to Integrate Acid and Al Tolerance in Alfalfa

Develop populations

Trait integration

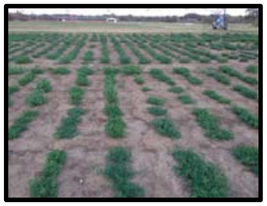
Marker-trait discovery

Breeding

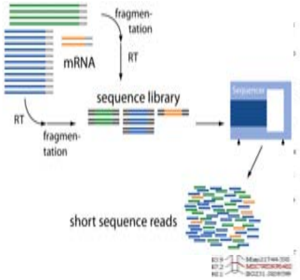
Select and intermate

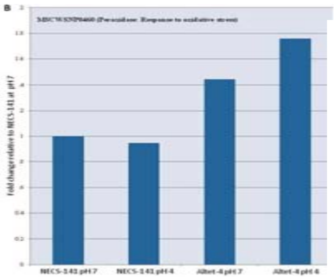
Germplasm:
Altet-4 (T1)
60T180-4 (T2)
NECS-141 (S)

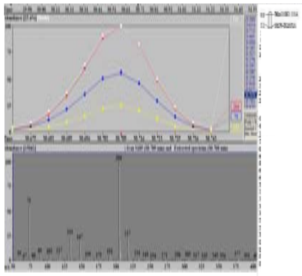
Populations
- Altet-4 (T1) × NECS-141 (S)
- NFAA (T2) × NECS-141(S)
- Altet-4 (T1) × NFAA (T2)

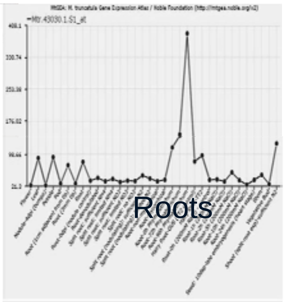


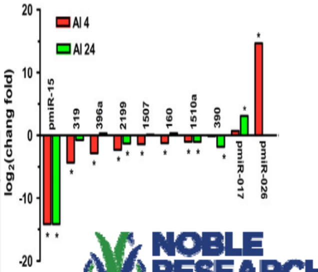
RNA Seq	**qRT-PCR**	**Metabolomics**	**Gene Atlas**	**miRNAs**
Gene discovery	Validation	Validation	AGED	Gene regulation











NOBLE RESEARCH INSTITUTE

Gene Models and Gene Expression Coverage

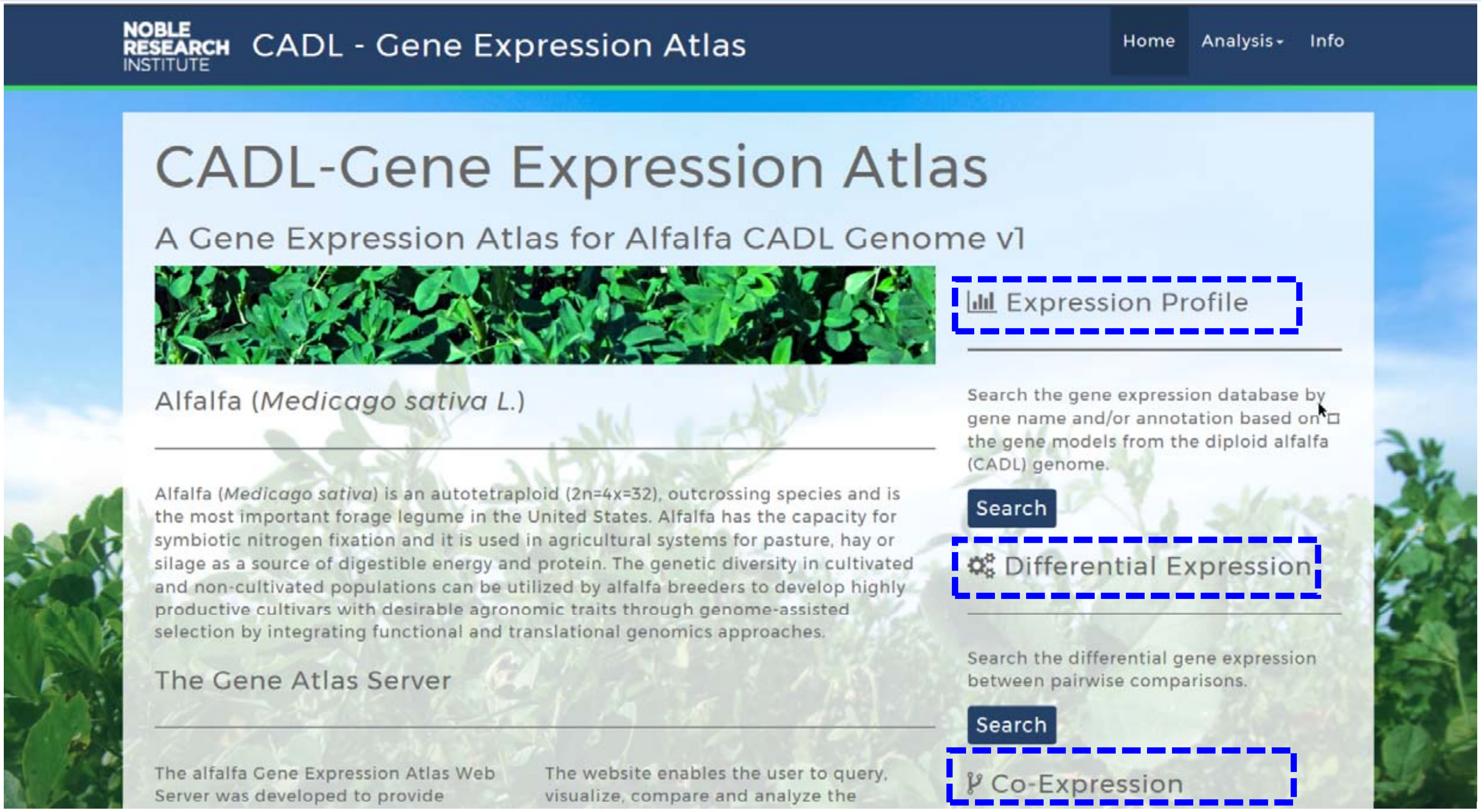
Al tolerant

Al sensitive

Al tolerant

Al sensitive

ABT: Alfalfa Gene Expression Atlas



CADL-Gene Expression Atlas

A Gene Expression Atlas for Alfalfa CADL Genome v1

Alfalfa (*Medicago sativa L.*)

Alfalfa (*Medicago sativa*) is an autotetraploid (2n=4x=32), outcrossing species and is the most important forage legume in the United States. Alfalfa has the capacity for symbiotic nitrogen fixation and it is used in agricultural systems for pasture, hay or silage as a source of digestible energy and protein. The genetic diversity in cultivated and non-cultivated populations can be utilized by alfalfa breeders to develop highly productive cultivars with desirable agronomic traits through genome-assisted selection by integrating functional and translational genomics approaches.

The Gene Atlas Server

The alfalfa Gene Expression Atlas Web Server was developed to provide

The website enables the user to query, visualize, compare and analyze the

Search the gene expression database by gene name and/or annotation based on the gene models from the diploid alfalfa (CADL) genome.

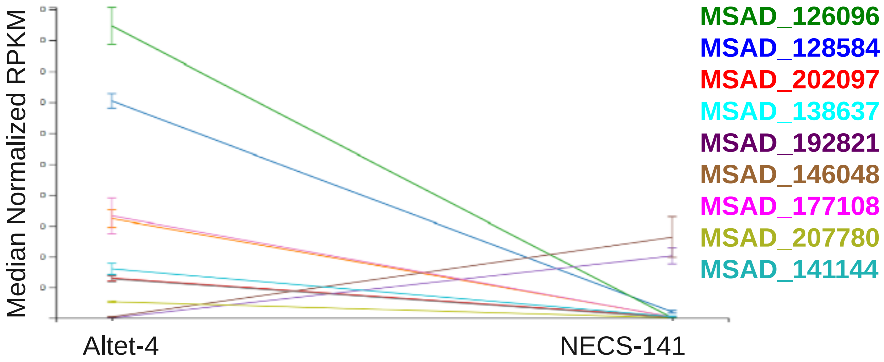
Search the differential gene expression between pairwise comparisons.

Tissues: *M. sativa* subsp. *sativa* and subsp. *falcata*: leaves, roots, flowers, N-fixing nodules, stem internodes

Abiotic stress: drought, low pH, Al toxicity

Alfalfa RNA-Seq Data in the Gene Expression Atlas (pH and Al)

1. Differential Expression (pH 4 + Al)

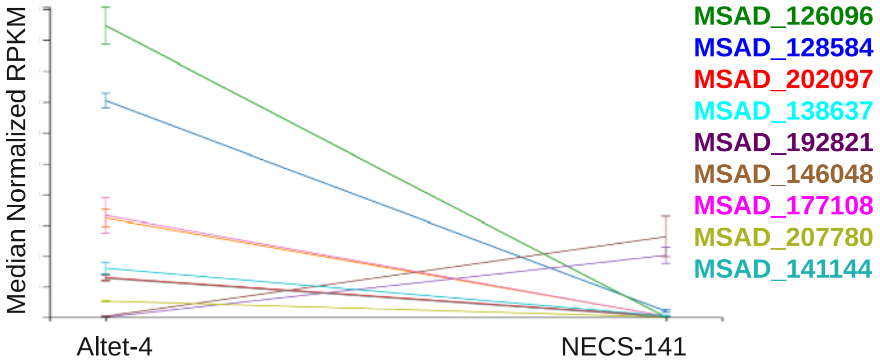


	Altet-4 (T)	NECS-141 (S)
MSAD_126098	141.0	4.0
MSAD_128584	64.7	0.7
MSAD_202097	189.3	0.2
MSAD_138637	26.1	0.7
MSAD_192821	0.1	40.6
MSAD_126887	0.7	52.8
MSAD_146048	66.4	0.7
MSAD_177108	25.5	0.2
MSAD_207780	10.6	0.2

***RPKM** = reads per kilobase per million

Alfalfa RNA Seq Data in the Gene Expression Atlas (pH and Al)

1. Differential Expression (pH 4 + Al)

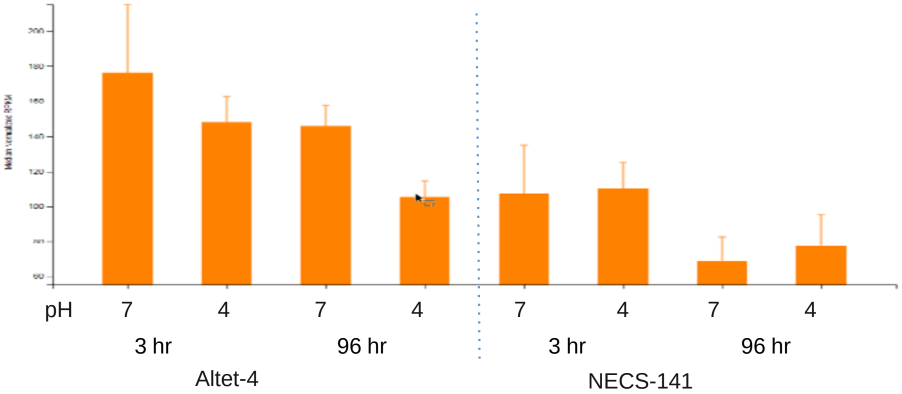


	Altet-4 (T)	NECS-141 (S)
MSAD_126098	141.0	4.0
MSAD_128584	64.7	0.7
MSAD_202097	189.3	0.2
MSAD_138637	26.1	0.7
MSAD_192821	0.1	40.6
MSAD_126887	0.7	52.8
MSAD_146048	66.4	0.7
MSAD_177108	25.5	0.2
MSAD_207780	10.6	0.2

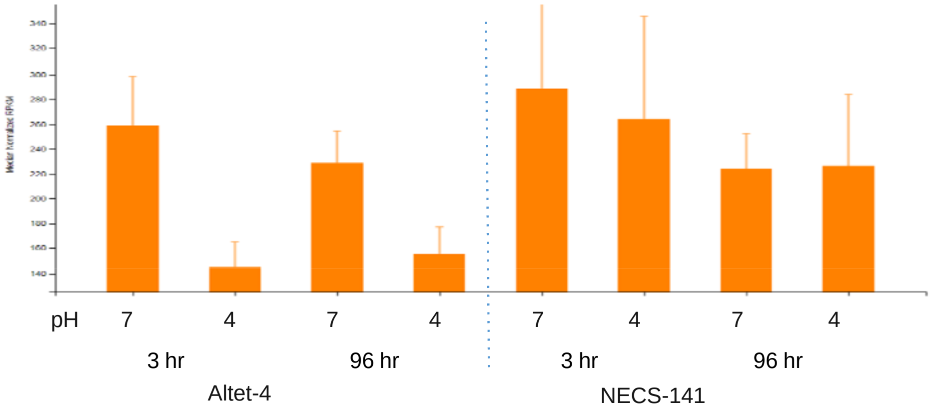
***RPKM** = reads per kilobase per million

2. Expression profile (Roots: keyword search)

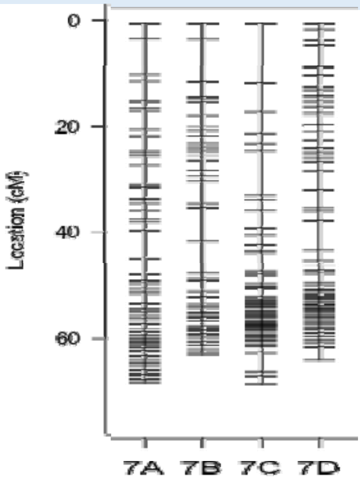
a. MSAD_182412: **Malate** dehydrogenase Medtr8g463760

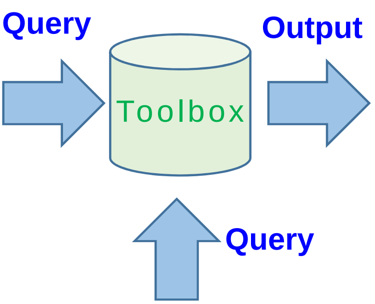


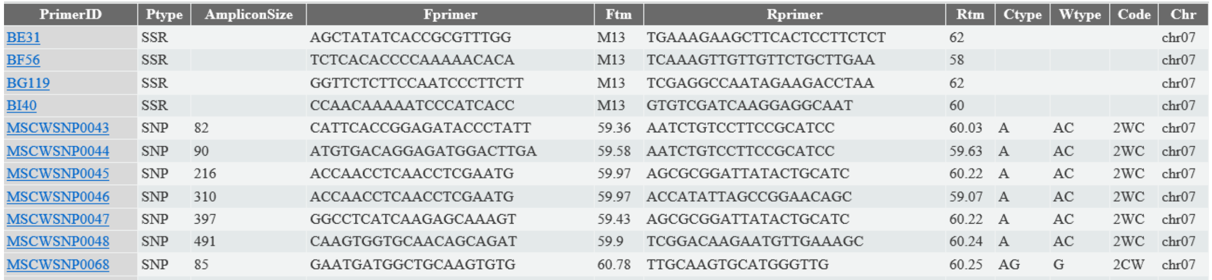
b. MSAD_136562: **Malate** dehydrogenase Medtr1g043040



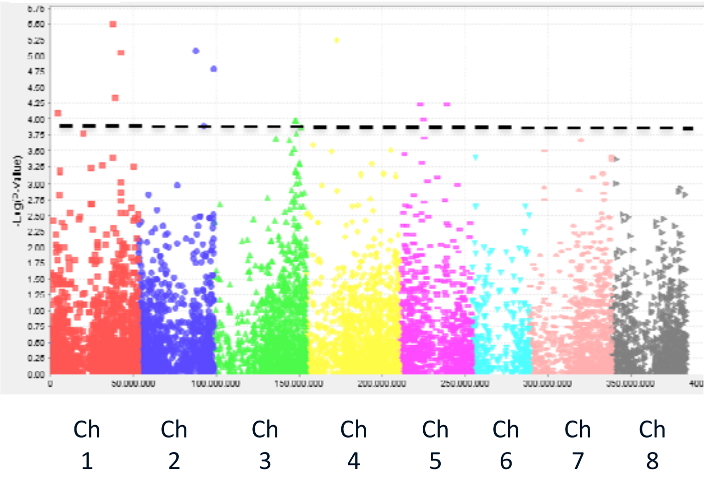
Integrate QTL, GWAS, SNP, Genome and Expression

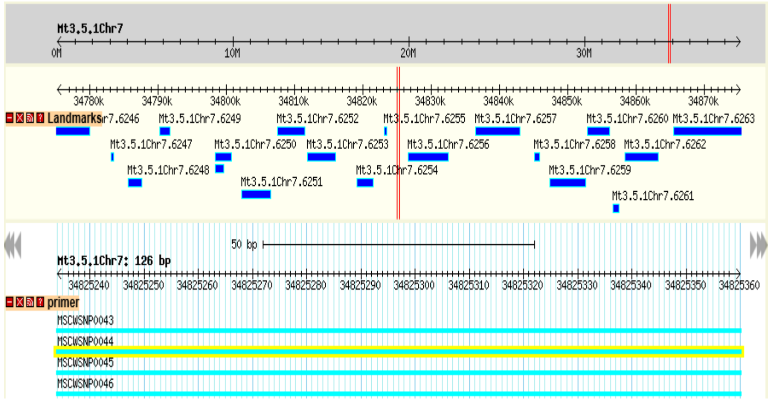


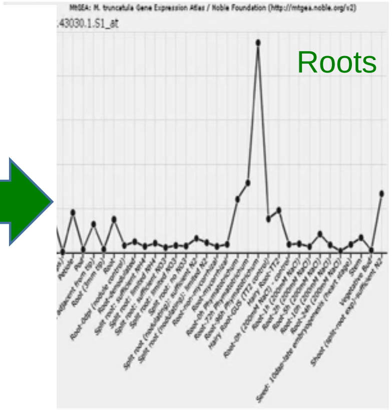




PrimerID	Ptype	AmpliconSize	Fprimer	Ftm	Rprimer	Rtm	Ctype	Wtype	Code	Chr
BE31	SSR		AGCTATATCACCGCGTTTGG	M13	TGAAAGAAGCTTCACTCCTTCTCT	62				chr07
BF56	SSR		TCTCACACCCCAAAAACACA	M13	TCAAAGTTGTTGTTCTGCTTGAA	58				chr07
BG119	SSR		GGTTCTCTTCCAATCCCTTCTT	M13	TCGAGGCCAATAGAAGACCTAA	62				chr07
BI40	SSR		CCAACAAAAATCCCATCACC	M13	GTGTCGATCAAGGAGGCAAT	60				chr07
MSCWSNP0043	SNP	82	CATTCACCGGAGATACCCTATT	59.36	AATCTGTCCTTCCGCATCC	60.03	A	AC	2WC	chr07
MSCWSNP0044	SNP	90	ATGTGACAGGAGATGGACTTGA	59.58	AATCTGTCCTTCCGCATCC	59.63	A	AC	2WC	chr07
MSCWSNP0045	SNP	216	ACCAACCTCAACCTCGAATG	59.97	AGCGCGGATTATACTGCATC	60.22	A	AC	2WC	chr07
MSCWSNP0046	SNP	310	ACCAACCTCAACCTCGAATG	59.97	ACCATATTAGCCGGAACAGC	59.07	A	AC	2WC	chr07
MSCWSNP0047	SNP	397	GGCCTCATCAAGAGCAAAGT	59.43	AGCGCGGATTATACTGCATC	60.22	A	AC	2WC	chr07
MSCWSNP0048	SNP	491	CAAGTGGTGCAACAGCAGAT	59.9	TCGGACAAGAATGTTGAAAGC	60.24	A	AC	2WC	chr07
MSCWSNP0068	SNP	85	GAATGATGGCTGCAAGTGTG	60.78	TTGCAAGTGCATGGGTTG	60.25	AG	G	2CW	chr07







http://plantgrn.noble.org/AGED

NOBLE RESEARCH INSTITUTE

Evaluation of Alfalfa Germplasm – Sensors & UAVs

Red River Farm (Burneyville, OK)

Dupy Farm (Gene Autry, OK)

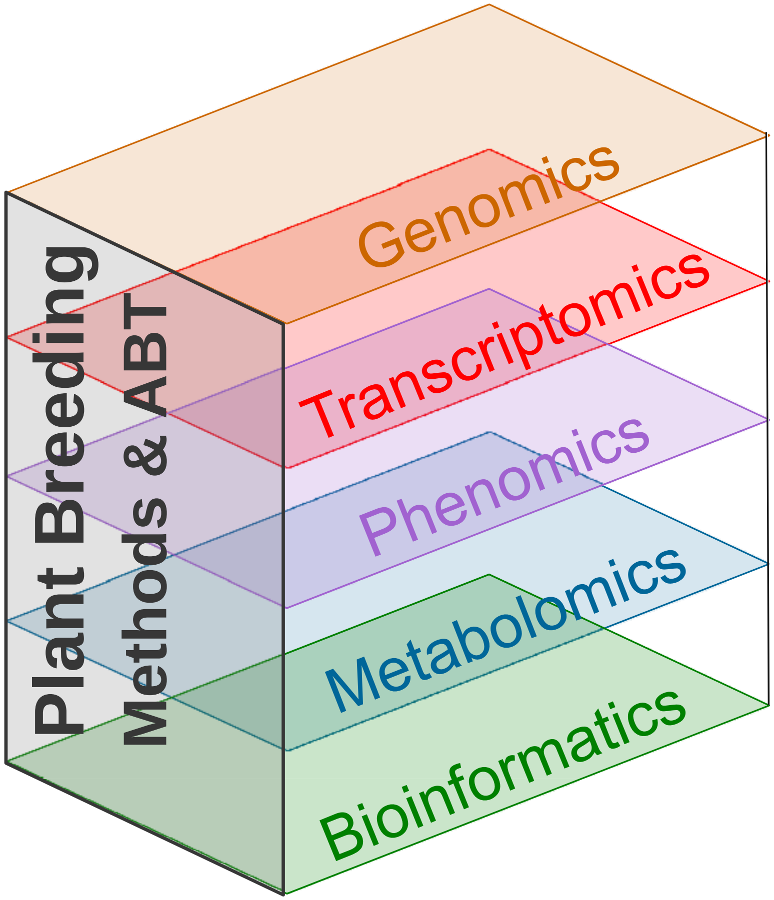
Manual

UAVs

Phenomobile: Spider

NOBLE
RESEARCH
INSTITUTE

Summary: Genome and Alfalfa Breeder’s Toolbox



Genome sequence
Gene annotation

Molecular markers
RNA-Sequencing
Gene expression

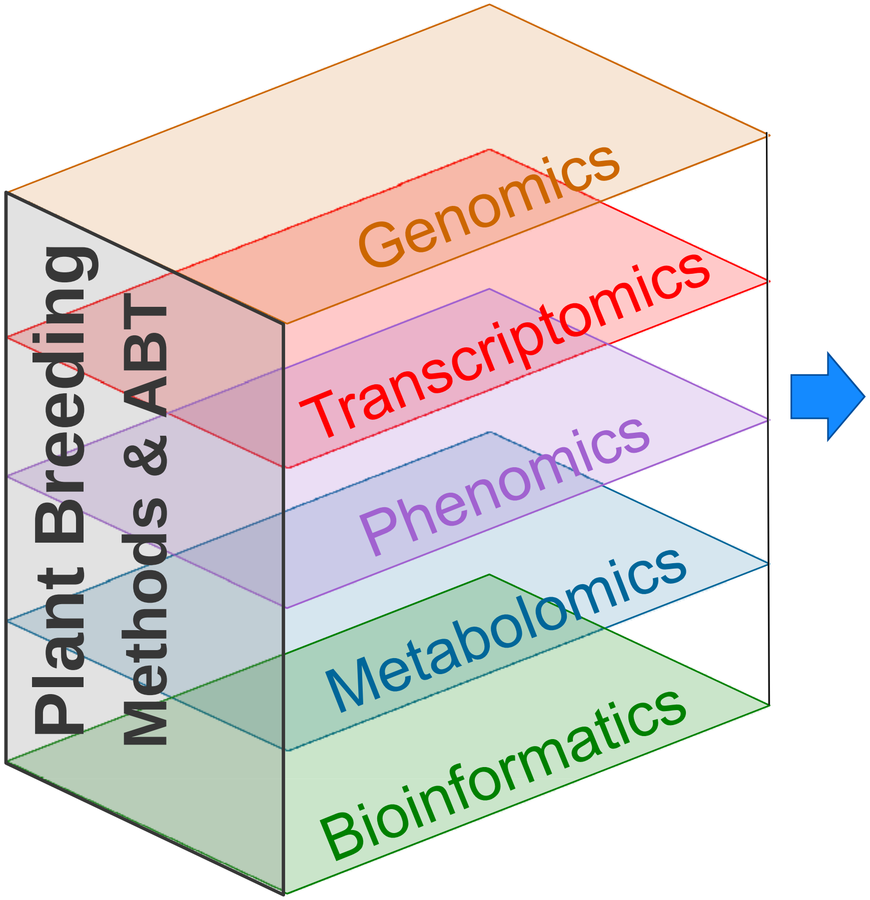
Identify genetic variation for key traits

Metabolic profiling

Integrated databases

NOBLE RESEARCH INSTITUTE

Summary: Genome and Alfalfa Breeder’s Toolbox



Plant Breeding Methods & ABT

Genomics

Transcriptomics

Phenomics

Metabolomics

Bioinformatics

NOBLE RESEARCH INSTITUTE

Acknowledgements

NCGR
Joann Mudge
Andrew Farmer
Diego Fajardo
Nicolas Devitt
Thiru Ramaraj
Univ. of Arizona
Haibao Tang
JCVI
Chris Town
UC Davis
Charlie Brummer
Univ. of Minnesota
Nevin Young
USDA-ARS
Deborah Samac
Joann Lamb
Justin Vaughn
Ethy Cannon
Steve Cannon
Phase Genomics
Shawn Sullivan
Ivan Liachko
Univ. of Wisconsin
Ted Bingham

Noble Research Institute
Xinbin Dai
Jaeyoung Choi
Patrick Zhao
Nick Krom
Junil Chang
Perdeep Mehta
Yuhong Tang
Mike Trammell
Suresh Bhamidimarri
Brian Motes
Michael Udvardi
Chunlin He
Nadim Tayeh
Rokebul Anower
Silvas Prince
Christy Motes
Tim Hernandez
Brice Cazenave
Denis Jaquez
Yanina Alarcon
Laura Putman
Na Li

NPGS (GRIN)


Genome sequencing


Alfalfa Breeder's Toolbox


Monteros Lab

J. Craig Venter® INSTITUTE
KeyGene
NCGR National Center for Genome Resources
NSF
USDA
University of Minnesota Driven to Discover℠
bionano GENOMICS
PHASE GENOMICS
NOBLE RESEARCH INSTITUTE

Thank you!

https://www.alfalfatoolbox.org/