Development of Resources for Alfalfa Improvement: An Assembly of the Genome Sequence and the Alfalfa Breeder's Toolbox

Maria J. Monteros[1], Chunlin He[1], Jaeyoung Choi [1], Patrick X. Zhao[1], Nadim Tayeh[1], Xinbin Dai[1], Andrew D. Farmer[2], Joann Mudge[2], Haibao Tang[3], Junil Chang[1], Nick Krom[1], Justin N. Vaughn[4], Perdeep Mehta[1], Christy M. Motes[1], Alyssa Nedley[1], Michael Trammell[1], Brian Motes[1], Shawn Sullivan[5], Ivan Liachko[5], E. Charles Brummer[6], Nevin D. Young[7], Christopher D. Town[8], Michael K. Udvardi[1]

[1]Noble Research Institute, Ardmore, OK; [2]NCGR, Santa Fe, NM; [3]University of Arizona, AZ, USA, [4]University of Georgia, Athens, GA; [5]Phase Genomics, Seattle, WA; [6]University of California, Davis, CA; [7]University of Minnesota, St Paul, MN, [8]J. Craig Venter Institute, Rockville, MD

Introduction

- Alfalfa (*Medicago sativa* L.) is a perennial forage legume with global agronomic importance and an estimated annual value of more than $8 billion in the USA alone (Bouton, 2007).

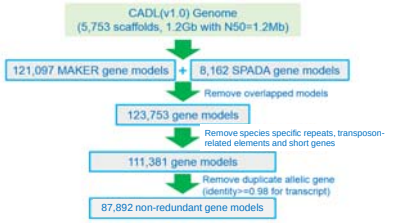
- The symbiotic nitrogen fixation capacity of alfalfa, high biomass yields and forage quality make alfalfa an excellent source of digestible energy and protein grown in pastures or harvested for hay and silage production.
- Genetic and genomic resources can be used to accelerate the genetic gains of plant breeding programs. Molecular markers available in alfalfa include those generated using high-resolution melting (HRM) and genotyping-by-sequencing (GBS) (Han et al. 2011; Khu et al. 2013; Li et al. 2012; Li et al. 2014). These resources can be used to implement marker-assisted selection (MAS) or genomics-assisted breeding (GAB) (Kole et al. 2015).
- The Alfalfa Breeder's Toolbox (ABT) was developed as a resource to integrate genomic, genetic and phenotypic data to advance scientific discoveries, integrate basic and applied knowledge and inform cultivar development activities.

Materials and Methods

Alfalfa genome sequences

- *Medicago truncatula* genome sequence (Mt4.0) (Young et al. 2011; Tang et al. 2014).
- Cultivated alfalfa at the diploid level (CADL) genome v1.0 (Fajardo et al. 2016). PacBio sequencing and Dovetail libraries were assembled into 5,753 scaffolds (N50=1.2Mb). Gene annotation was pursued using the MAKER pipeline and BUSCO (Campbell et al. 2014; Simao et al. 2015; Dai et al. 2017) as previously described (Dai et al. 2017) to generate high-confidence (HC) gene models (Fig. 1).

Fig. 1. Overview of the CADL (v1.0) genome sequence annotation pipeline.



- Sequencing of the tetraploid alfalfa genotype NECS-141 (2n=4x=32) was initially pursued using Illumina paired-end libraries (Bennett and Leitch, 2011) and PacBio sequencing (Eid et al. 2009) with an average read length of 10,200 bp (Monteros et al. 2015). Most recently, the Proximo™ Hi-C method (Phase Genomics) was used to generate a revised assembly of the PacBio Canu output using bwa aln (Li and Durbin, 2010). Overall, multiple approaches were used to assemble the sequences generated from different technologies (Simpson et al. 2009; Luo et al. 2012; Kajitani et al. 2014; Chin et al. 2016; Berlin et al. 2015; Koren et al. 2017) into 32 super-scaffolds (Table 1; Fig. 2; Fig. 3).

Alfalfa Breeder's Toolbox (ABT)

- Chado-Drupal-Tripal was used to develop the ABT (Fig. 4) and JBrowse was used to integrate and visualize genomic sequences, gene models, gene expression levels and allele frequencies (Fig. 5) (Cho et al. 2012, Ficklin et al. 2011, Jung et al. 2011, Sanderson et al. 2013, Buels et al., 2016).
- The features of the ABT allow access to information from multiple datasets to address practical questions to inform plant breeding decisions.

Genotypic data: Genotypic data from tetraploid alfalfa populations are aligned against the reference genome in JBrowse and can be visualized in the form of allele frequencies. Changes in allele frequencies due to selection from breeding can be visualized for different genotypes (Fig. 5) and tracked throughout the breeding process.

Molecular markers: SSRs and SNPs distributed throughout the genome and those associated with QTLs for abiotic stress tolerance (Han et al. 2011; Khu et al. 2013; Li et al. 2014) were integrated into the ABT (Fig. 4; Fig. 5) and can be used to identify key genes and markers in a target genomic region to facilitate molecular breeding applications (Fig. 6).

Gene expression atlas (GEA): A web-based gene expression atlas that integrates RNA-sequence data using the DESeq2 and Salmon software (Love et al. 2014) from contrasting alfalfa genotypes (O'Rourke et al. 2015) and those grown under abiotic stress (pH, presence of aluminum and drought), enables the query and visualization of differentially and co-expressed target genes (Fig. 7; Fig. 8).

Phenotypic data: Diverse alfalfa germplasm including PI accessions from the alfalfa core collection (Basigalup et al. 1995) were evaluated in the field and used to collect biomass and other agronomic traits. Individuals can be sorted and ranked by multiple traits (Fig. 9) to identify the best performers for further characterization and population development.

Other functionalities: The ABT also allows users to search for specific markers, gene targets, genomic regions, perform BLAST searches using DNA sequences and/or candidate genes and perform *in silico* PCR.

Figure 2. Overview of sequencing platforms and assembly strategies pursued to generate the tetraploid alfalfa genome sequence.

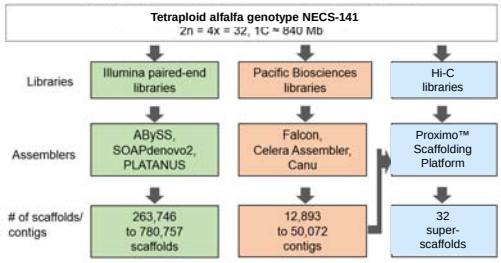
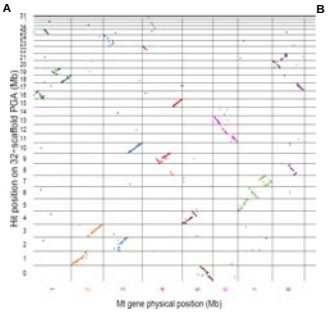


Table 1. Sequencing platforms and assembly strategies used to generate the tetraploid alfalfa genome sequence.

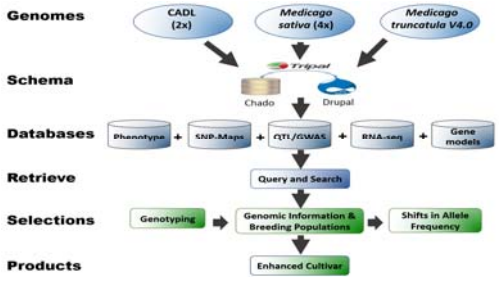
Platform	Assembler	# of scaffolds / contigs	Total length (Mb)	N50 (Kb)	Maximum length (Kb)	Complete core eukaryotic genes (%)	Transcript coverage (%)
Illumina	ABYSS	263,746	897	5	29	68.95	86.6
	SOAPdenovo2	359,556	1096	6	79	81.05	94.7
	PLATANUS	780,257	863	34	426	65.73	91.8
PacBio	Falcon	50,072	1607	63	762	87.9	93.6
	Celera Assembler	47,839	2682	82	3,289	92.74	95.4
	Canu	12,893	2349	311	4,086	92.34	95.4
Phase Genomics' Proximo™	Proximo™ Scaffolding Platform	32	2238	87,112	133,759	-	-

Figure 3. A. Dot plot comparing orthologous regions of Mt4.0 pseudomolecules and the 32 super-scaffolds from the proximity-guided assembly of the tetraploid alfalfa genome sequence. **B.** Chromosomal distribution of the 32 alfalfa super-scaffolds.



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

Figure 4. Overview of the ABT to integrate genomic, genetic and phenotypic data to advance practical plant breeding applications.



Summary

- We have generated assemblies of the diploid and tetraploid alfalfa genomes using multiple approaches.
- The alfalfa gene expression atlas (GEA) of the ABT enables users to query and assess gene expression of alfalfa genotypes tolerant vs. sensitive to abiotic stress factors to further understand stress tolerance mechanisms.
- Users can utilize the ABT to search and retrieve genomic and genetic information to implement molecular breeding strategies using SNPs targeting differentially expressed genes to increase the frequencies of favorable alleles associated with agriculturally-relevant traits.
- Access to phenotypic data from diverse alfalfa germplasm collected in the field facilitates ranking of genotypes based on multiple traits (selection index) and identification of potential parents to generate breeding populations.

Ongoing Activities

- Pursue optical mapping of tetraploid alfalfa and integrate alfalfa datasets with other legume species leveraging existing databases and initiatives including the Legume Information System and Legume Federation.
- Generate genotypic data of additional alfalfa populations evaluated at multiple locations to integrate with sensor-based phenotypic data to facilitate selection of parental lines for population development.
- Expand content of the ABT by integrating curated data from collaborators and partners to advance opportunities for alfalfa improvement.

Figure 5. Integration of gene models, differential gene expression, SNPs and allele frequencies anchored to the reference sequence visualized with JBrowse.

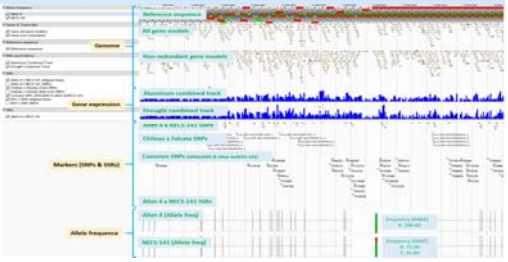


Figure 6. Identification of genes and molecular markers in a target genomic region.

Figure 7. Gene expression profiles of organic acid transporters in roots of Altet-4 (Al tolerant) and NECS-141 (Al sensitive) after 96 hrs of growth at pH 4 + Al.

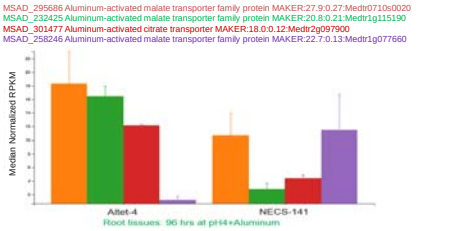


Figure 8. Differential gene expression of malate transporters in the Altet-4 and NECS-141 genotypes after 3 h and 96 hr at pH 7 and pH 4 + Al (pH 4).

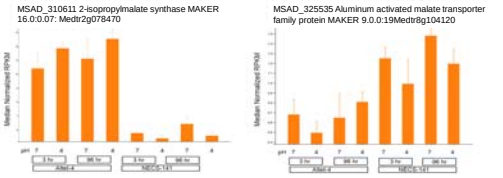


Figure 9. Sample output of phenotypic data (multiple traits in addition to biomass and crude protein) based on field evaluations of alfalfa germplasm.

References

- Basigalup et al. 1995. Crop Science 35: 1163-1168.
- Bennett and Leitch. 2011. Annals of Botany 107: 467-590.
- Berlin et al. 2015. Nature Biotechnology 33: 623-630.
- Bouton 2007. Euphytica 154: 263-270.
- Buels et al. 2016. Genome Biology 17: 66.
- Campbell et al. 2014. Plant Physiology 164: 513-524.
- Chin et al. 2016. Nature Methods 13: 1050-1054.
- Dai et al. 2017. Plant and Animal Genome XXV Conference.
- Eid et al. 2009. Science 323: 133-138.
- Fajardo et al. 2016. Plant and Animal Genome XXIV Conference.
- Han et al. 2011. BMC Genomics 12: 350.
- Jung et al. 2011. Database 2011: bar051.
- Kajitani et al. 2014. Genome Research 24: 1384-1395.
- Khu et al. 2013. Crop Science 53: 148-163.
- Kole et al. 2015. Frontiers in Plant Science 6:563.
- Koren et al. 2017. bioRxiv: 071282.
- Li and Durbin. 2010. Bioinformatics 26: 589-595.
- Li et al. 2012. BMC Genomics 13: 568.
- Li et al. 2014. G3: Genes| Genomes| Genetics 4: 1971-1979.
- Love et al. 2014. Genome Biology 15: 550.
- Luo et al. 2012. Gigascience 1: 18.
- Monteros et al. 2015. Plant and Animal Genome XXIII Conference.
- O'Rourke et al. 2015. BMC Benomics 16: 502.
- Sanderson et al. 2013. Bioinformatics: btv351.
- Simão et al. 2015. Bioinformatics: btv351.
- Simpson et al. 2009. Genome Research 19: 1117-1123.
- Tang et al. 2014. BMC Genomics 15: 312.
- Young et al. 2011. Nature 480: 520-524.

Acknowledgements

The tetraploid alfalfa genome and the ABT are funded by the Forage 365 initiative of The Noble Research Institute. The CADL genome sequencing was funded by an NSF grant to Nevin Young. We thank the CADL genome assembly team (Nicholas Devitt, Diego Fajardo, Thiru Ramaraj, Zhaohong Zhuang, and Peng Zhou) and Malay Saha, Yuhong Tang, Wenchao Zhang and Rokebul Anower for their contributions and suggestions to the project.