Readme file (CADL)

7 June 2016

Description of HM342 *Medicago sativa* Cultivated Alfalfa at the Diploid Level* (CADL) v0.95P genome data

Participants:

Joann Mudge, Nicholas P. Devitt , Diego A. Fajardo, Thiru Ramaraj, Andrew D. Farmer, Xinbin Dai, Zhaohong Zhuang, Peng Zhou, Joseph Guhlin, Christopher D. Town, Maria J. Monteros, Patrick X. Zhao, Jason R. Miller, Kevin A. T. Silverstein, Nevin D. Young

LIST OF FILES

medsa.CADL_HM342.v0.95P.fasta.gz

DISCLAIMER

This assembly is a pre-release version provided "as is" for the community with no claims on the quality or completeness of the sequence or gene coverage. Please be aware that:

1) SSPACE-Longread scaffolding is very aggressive and likely introduced some chimeras. No attempt has been made to identify and correct these chimeras.
2) *Medicago sativa* is a highly heterozygous organism with an expected haploid genome size of 800 Mb. The sequence similarity of the two haplotypes in the diploid CADL varies, often diverging enough from each other that they are assembled separately. This has resulted in an assembly size of ~1200 Mb rather than the expected 800 Mb, suggesting that at least half of the genome is represented by the assembly by two distinct haplotypes. In regions of the most

divergence, presence/absence differences in gene content can be seen between the haplotypes. This also implies that the gene content of the current assembly contains a significant proportion of allelic copies of genes, a supposition that is confirmed by both alignment to the related *Medicago truncatula* genome and by analysis of genes that are typically found in single copies in plant genomes.

RESTRICTIONS ON USE

The CADL assemblies available here, including the previous and current version, are made available to the research community by the *Medicago* HapMap consortium under the *Toronto Agreement* [http://www.nature.com/nature/journal/v461/n7261/full/461168a.html]. As producers of these data, we reserve the right to be the first to publish a genome-wide analysis of the data.

The pre-publication data released here is embargoed for publication except for analyses of single gene loci or small (< 10 kb) genome regions. Researchers are encouraged to contact us if there are queries about referencing or publishing analyses based on the pre-publication data obtained via this website. Researchers are also invited to consider collaborations with the *Medicago* Hapmap consortium for larger studies or if the limitations here restrict further work.

CADL SOURCE MATERIAL (Renamed HM342 as part of the Hapmap project)

A single plant was clonally propagated at the University of Minnesota. DNA was isolated by Amplicon Express in February, 2015.

CADL ASSEMBLY VERSION 0.95P

This assembly was generated with ~100X PacBio Reads, described in the following table.

HM Number	Name	Chemistry	Mean subread length	Subread N50	Subread total length	Number of subreads	Max length	Coverage
HM342	CADL	P6-C4	8,034	13,070	81,270,428,779	9,013,313	52,329	101.59

Falcon v. 0.2.2 was used for correction and assembly, utilizing Dazzler for the read alignments during the correction phase. Scaffolding was performed with SSPACE-Longread followed by Quiver polishing. The assembly statistics are in the table below.

	CADL Assembly
Assembler	Falcon-Quiver-SSPACE
Contigs	10,110
Max Contig	2,901,187
Mean Contig	122,611
Contig N50	349,080
Contig N90	70,351
Total Contig Length	1,239,601,791
Assembly GC	34.01
Scaffolds	6,545
Max Scaffold	4,047,589
Mean Scaffold	191,699
Scaffold N50	547,092
Scaffold N90	130,123
Total Scaffold Length	1,254,671,259
Captured Gaps	3,565
Max Gap	27,691
Mean Gap	4,227
Gap N50	6,479
Total Gap Length	15,069,468

* Bingham ET and McCoy TJ (1979) Cultivated Alfalfa at the Diploid Level: Origin, Reproductive Stability, and Yield of Seed and Forage. Crop Science 19: 97-100.