20 May 2015

Description of Medicago R108 (HM340) v0.95 genome data

Medicago HapMap project
University of Minnesota, NCGR, and JCVI
www.medicagohapmap.org
Joann Mudge, Jason Miller, Thiru Ramaraj, Diego Fajardo, Peng Zhou, Joseph Guhlin, Kevin Silverstein and Nevin Young

LIST OF FILES
Medicago_R108_HM340_v0.95_assembly.fasta.gz
Medicago_R108_HM340_v0.95_annotation.gff3.gz

RESTRICTIONS ON USE

The R108 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.

R108 SOURCE MATERIAL (Renamed HM340 as part of the Hapmap project)
Seeds were obtained from Pascal Ratet derived from *M. truncatula* R108-1 C3 in July 2012. This is the stock propagated in Gif from the original regenerable line and used by the community for transformation (i.e., to produce also the Tnt1 line originally). The same seeds were distributed to all the labs who wanted to have R108 plants or produce transgenics.

R108 ASSEMBLY VERSION 0.95
An initial draft assembly was created using ALLPATHS (allpathslg-49962; Gnerre et al., 2011) in December 2014 using default parameters. The following are the assembly statistics obtained:
Assembly Statistics:

Contigs	22,475
Max Contig	297,747
Mean Contig	15,909
Contig N50	41,079
Contig N90	7,311
Total Contig Length	357,564,362
Assembly GC	32.98

Scaffolds	3,890
Max Scaffold	7,121,670
Mean Scaffold	99,791
Scaffold N50	1,120,750
Scaffold N90	200,551
Total Scaffold Length	388,188,365
Captured Gaps	18,585
Max Gap	10,508
Mean Gap	1,648
Gap N50	3,800
Total Gap Length	30,624,003

INPUT DATA

- 33.2Gb (66 X) of short insert paired end data (TruSeq), 2 x 100 bp from the HiSeq 2000, insert size 146 bp
- 28.8 Gb (58 X) of long insert paired end data (Nextera) (i.e. mate pair), 2 x 100 bp from the HiSeq 2000 with insert size 9000 bp

ANNOTATION

AUGUSTUS was used to make ab initio gene predictions for each genome assembly with both RNA-Seq expression evidence and HM101 (A17; Mt4.0 reference) homology evidence. RNA-Seq reads from HM340 were mapped to the *de novo* assembly using Tophat to generate intron hints for AUGUSTUS. We also transferred HM101 annotation to HM340 using synteny block mapping information and generated exon hints for AUGUSTUS. Predicted protein sequences were scanned for PFAM domains (Pfam-A.hmm) using HMMER and processed using custom scripts.

REFERENCES

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011 Jan 25;108(4):1513-8. doi: 10.1073/pnas.1017351108. Epub 2010 Dec 27. http://www.ncbi.nlm.nih.gov/pubmed/21187386

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*, 62. doi:10.1186/1471-2105-7-62 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1409804/