1 March 2013

Description of SNP data and methods used for association mapping in *Medicago truncatula*

Medicago HapMap Project
University of Minnesota
www.medicagohapmap.org
John Stanton-Geddes, Roman Briskine, Kevin Silverstein, Jeremy Yoder, Tim Paape, Peter Tiffin, Nevin Young

SNP DATA

288 *Medicago truncatula* accessions were sequenced using Illumina and reads were aligned to the *M. truncatula* v3.5 reference genome (accession HM101 = A17, Young *et al* 2011). Twenty-six of the 288 accessions were sequenced to 15X averaged aligned depth. These 26 are HM001-HM016, HM019-HM021, HM023-HM028, and HM101. For these 26 accessions, a base was called to be different from the reference if it was covered by ≥ 2 uniquely aligned reads, ≥ 70% of the reads that aligned to that location called that base, and there were < 1,000 unique reads covering the site (Branca *et al* 2011). Sites covered by ≥ 2 uniquely aligned reads but with ≤ 30% of the reads calling a variant were assumed to be the same as the reference genome. The 70% requirement means that no heterozygous sites were identified, but this is expected to have only a minor effect on the data because of high selfing rates in natural populations and multiple generations of selfing prior to DNA extraction.

The remaining 262 accessions were sequenced to an average coverage of ~ 6X (Stanton-Geddes *et al* 2012). For these accessions, a base was considered to differ from the reference genome if it was covered by ≥ 1 unique read, ≥ 70% of the reads that aligned to that location called that base, and there were < 500 unique reads covering the site.

NOTE: the low stringency threshold of 1 unique read was used for completeness. Some of these may be false-positives. To reduce the number of false positives, we removed all positions where the SNP was present in only a single accession. Moreover, an implicit filter against most false-positives occurs when requiring a minimum allele frequency threshold in association studies. This is discussed further in the 'ASSOCIATION ANALYSES USING TASSEL' section below.

INPUT FILES

The SNP data are provided in a format compatible for GWAS [TASSEL polymorphism format] using TASSEL 3.0 (Bradbury *et al* 2007). We use this format because the files are large (up to 1.8GB) and this is the simplest format for storing these data. A separate file is provided for each of the 8 chromosomes, as well as for pseudomolecules "C" representing chloroplast (~ 125 kb), "U" representing unanchored BACs (~ 17.5 Mb) and "T" (~ 12.3 Mb) representing tentative consensus sequences (TCs) from the Dana Farber Cancer Institute Mt gene index v.10.0 (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago) that are absent from pseudomolecules 1 – 8 or "U". Pseudomolecules "U" and "T" were constructed by joining individual sequences (unanchored BACs and uncaptured TCs respectively) with insertion of 1 kb gaps. Map files that indicate position of each included sequence accompany the data for these pseudomolecules.

Characters A/C/G/T code SNPs, and missing data are coded as '?'.

The first row of each data file gives the number of accessions, the number of sites and the ploidy level. For example,

288 1508348:1

indicates there 288 accessions, 1,508,348 sites, and the data is haploid. The second row gives the position of each SNP. The following rows contain genotype information for each accession with one accession per row.

Since the TASSEL-compatible format mentioned above can be difficult to parse, data are also provided in an alternative format [SNP format] that offers accession data in columns rather than in rows. First line of the file lists the number of sites and the number of accessions. The second line contains column headers while all subsequent lines contain SNP data for each site. The first 9 columns list the following data: position, gene context (these values are the same as in gene context data files described below), reference allele, minor allele frequency (for multi-allelic sites, this is the frequency of the second most frequent allele), flag indicating whether the site is multi-allelic, and counts for each possible nucleotide call (A, C, T, and G). Missing data is denoted by "N".

GENE CONTEXT AND ANNOTATION

Gene context data is based on Mt3.5.1v4 annotation and included in a separate set of files, one for each pseudomolecule. The latest annotation files can be downloaded from JCVI website (http://jcvi.org/cgi-bin/medicago/download.cgi). However, there may be discrepancies between the newer versions of the annotation and the provided gene context files. Each gene context file contains the following columns.

1. SNP position
2. Nearest gene call
3. Gene context (Table 1)
4. Distance to the nearest gene (Fig 1)
5. Reference allele
6. Reference amino acid (for SNPs in CDS, according to Mt3.5.1v3 annotation)
7. Average quality
8. Maximum quality

Please note that nearest gene call, gene context, and gene distance fields are meaningless for pseudomolecule "T". Hence, the first field in "T" is always empty, the second is set to 0, while the third shows the distance from the beginning of the corresponding TC. Reference amino acid data are also not available for pseudomolecule "T". Some unanchored BACs lack gene calls. In those cases, pseudomolecule "U" file have empty values for the nearest gene call and gene context fields while nearest gene distance is set to 0.

Since the gene context data only provides the ID of the nearest gene call for each SNP, a separate set of files are needed for gene annotation and structure. There is one file for pseudomolecules chr1-8, and one each for pseudomolecules "T" and "U". These files are in standard gff format (details available at http://www.sequenceontology.org/gff3.shtml). Briefly, each file contains 9 columns: seqid, source, type, start, end, score, strand, phase and attributes. The "attributes" column is of primary interest as it provides the gene ID and annotation for each mRNA. Note that because of alternative splicing, there can be multiple mRNA records for any given SNP.

Transposable elements are provided in separate files using the same format.

For pseudomolecules “T” and “U”, files that map tentative consensus sequences and unanchored BACs to their respective positions are also provided.

ASSOCIATION ANALYSES USING TASSEL

For the GWAS reported in Stanton-Geddes *et al* we included only SNPs that had been genotyped in > 100 accessions and had a minor allele frequency (MAF) > 2% (a base had to be called in ≥ 3 accessions to meet this criterion). The criterion of 100 accessions was made for reasons of statistical power. The 2% MAF requirement was made for two reasons: i) because we had used a lenient criteria for calling SNPs, it is likely that SNPs called in only a single accession would have high error rates, and ii) with < 280 accessions, there is little statistical power to robustly associate very rare SNPs with phenotypic variation.

We also found that results of GWAS are highly dependent upon the distribution of phenotypic values. Therefore, we strongly recommend transforming phenotypic data so that the distribution approximates a normal distribution.

As part of our GWAS analysis, we used a kinship (K) that accounts for relatedness among individual accessions. We derived this K matrix (available as “Mt3.5.1_var288_K-matrix_20120606.txt.gz”) using TASSEL with 5,000 randomly sampled SNPs from each chromosome (results rescaled with a minimum of 0 and maximum of 2) to remove potentially confounding demographic effects from biasing results. Our analyses indicated that a K matrix sufficiently accounted for demographic effects. The inclusion of other covariates (e.g. Q) was not necessary.

These are large data files and require substantial computation resources (~64GB RAM) to perform MLM analyses in TASSEL. To expedite analysis, GWAS can be performed independently for each chromosome/pseudomolecule, with chromosome 5 divided into two parts (5.1 and 5.2) at approximately the centromere. The results from all chromosomes/pseudomolecules can then combined and analyzed in R.

In the GWAS reported in Stanton-Geddes *et al*, 21 accessions – HM216, HM246-HM252, HM254-HM255, HM257-HM258, HM261, HM264, HM273-HM275, HM291-HM292, HM303, HM317 – were not included because they are members of a distinct lineage (see neighbor-joining tree in supplement of Stanton-Geddes *et al*) and 6 accessions (HM280, HM282-286) were also excluded due to high similarity.

OUTGROUP ACCESSIONS

We excluded twenty-eight accessions from the association study because they represented other subspecies of *Medicago* (Yoder et al., 2013). Five of those accessions (HM017, HM018, HM022, HM029, and HM030) were sequenced to 15X averaged aligned depth while the rest (HM102, HM318 – HM339) had the average depth of ~6X. Henceforth, we will refer to them as “outgroup accessions.” We generated SNP calls for those accessions using the same criteria as before (see SNP Data section.) However, we did not compare the results to the data set we used in the association study and we did not remove any SNPs that were present in a single outgroup accession. Thus, the false positive rate for the outgroup data set may be much higher than for the association study data set. We provide the

outgroup data in the alternative SNP format (see Input Data section for more information.) Only chromosomes 1-8 and the chloroplast data are available.

REFERENCES

Bradbury PJ, Zhang Z, Kroon DE*, et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635.

Branca A, Paape TD, Zhou P*, et al.* (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences of the United States of America* **108:** E864-870.

Stanton-Geddes J, Paape T, Epstein E, *et al.* (submitted). Sequence-based association genetics of the *Medicago truncatula – Sinorhizobium* symbiosis.

Yoder JB, Briskine, R, Mudge J, et al. (2013) Phylogenetic signal variation in the genomes of the genus *Medicago* (Fabaceae). Systematic Biology: doi:10.1093/sysbio/syt009.

Young ND, Debellé F, Oldroyd GED*, et al.* (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-524.

LIST OF CHANGES

21 September 2012

- Sites with degenerate bases in the reference genome were removed from the SNP files. (They remain in TASSEL files.) The change affected chromosomes 2, 4, 5, 5.2, 7, 8, and T (see Table 2.)
- Replaced 'N' with '?' in the first column (position 34) of the chromosome U TASSEL file.
- Replaced all lower case nucleotides with upper case nucleotides in the context files and SNP files. The change affected chromosomes 5, 5.1, 5.2, and 7.

1 March 2013

- Added the outgroup data set with the following accessions: HM017, HM018, HM022, HM029, HM030, HM102, and HM318 – HM339.

LIST OF FILES

- Annotation
  - Mt3.5v4_chr1-8_genes.gff.gz (6Mb) - gene annotation chr 1-8
  - Mt3.5v4_chr1-8_TEs.gff.gz (1Mb) - transposable elements chr 1-8
  - Mt3.5v4_chrT_genes.gff.gz (328kb) - gene annotation pseudomolecule T
  - Mt3.5v4_chrT_posmap.txt.gz (75kb) - maps tentative consensus sequences to positions on pseudomolecule T
  - Mt3.5v4_chrU_genes.gff.gz (377kb)- gene annotation pseudomolecule U
  - Mt3.5v4_chrU_TEs.gff.gz (102kb) - transposable elements pseudomolecule U
  - Mt3.5v4_chrU_posmap.txt.gz (2kb) - maps unanchored BACs to positions on pseudomolecule U
  - checksums_md5.txt
- GeneContext
  - Mt3.5_var288_chr1_context_20120606.txt.gz (14Mb)
  - Mt3.5_var288_chr2_context_20120606.txt.gz (18Mb)
  - Mt3.5_var288_chr3_context_20120606.txt.gz (23Mb)
  - Mt3.5_var288_chr4_context_20120606.txt.gz (21Mb)
  - Mt3.5_var288_chr5.1_context_20120921.txt.gz (15Mb)
  - Mt3.5_var288_chr5.2_context_20120921.txt.gz (15Mb)
  - Mt3.5_var288_chr5_context_20120921.txt.gz (30Mb)
  - Mt3.5_var288_chr6_context_20120606.txt.gz (11Mb)
  - Mt3.5_var288_chr7_context_20120921.txt.gz (21Mb)
  - Mt3.5_var288_chr8_context_20120606.txt.gz (15Mb)
  - Mt3.5_var288_chrC_context_20120606.txt.gz (10kb)
  - Mt3.5_var288_chrT_context_20120606.txt.gz (799kb)
  - Mt3.5_var288_chrU_context_20120606.txt.gz (8Mb)
  - checksums_md5.txt
- Outgroup
  - Mt3.5_outgroup_chr1_snp_20130227.txt.gz (29Mb)
  - Mt3.5_outgroup_chr2_snp_20130227.txt.gz (37Mb)
  - Mt3.5_outgroup_chr3_snp_20130227.txt.gz (44Mb)
  - Mt3.5_outgroup_chr4_snp_20130227.txt.gz (41Mb)
  - Mt3.5_outgroup_chr5_snp_20130227.txt.gz (57Mb)
  - Mt3.5_outgroup_chr6_snp_20130227.txt.gz (16Mb)
  - Mt3.5_outgroup_chr7_snp_20130227.txt.gz (41Mb)
  - Mt3.5_outgroup_chr8_snp_20130227.txt.gz (30Mb)
  - Mt3.5_outgroup_chrC_snp_20130227.txt.gz (62Kb)
  - checksums_md5.txt
- SNP
  - Mt3.5_var288_chr1_snp_20120606.txt.gz (108Mb)
  - Mt3.5_var288_chr2_snp_20120921.txt.gz (142Mb)
  - Mt3.5_var288_chr3_snp_20120606.txt.gz (182Mb)
  - Mt3.5_var288_chr4_snp_20120921.txt.gz (161Mb)
  - Mt3.5_var288_chr5.1_snp_20120921.txt.gz (118Mb)
  - Mt3.5_var288_chr5.2_snp_20120921.txt.gz (121Mb)
  - Mt3.5_var288_chr5_snp_20120921.txt.gz (238Mb)
  - Mt3.5_var288_chr6_snp_20120606.txt.gz (88Mb)
  - Mt3.5_var288_chr7_snp_20120921.txt.gz (165Mb)
  - Mt3.5_var288_chr8_snp_20120921.txt.gz (121Mb)
  - Mt3.5_var288_chrC_snp_20120606.txt.gz (70kb)
  - Mt3.5_var288_chrT_snp_20120921.txt.gz (7Mb)
  - Mt3.5_var288_chrU_snp_20120606.txt.gz (61Mb)
  - checksums_md5.txt
- TASSEL

    - Mt3.5_var288_chr1_tassel_20120606.txt.gz (136Mb)
    - Mt3.5_var288_chr2_tassel_20120606.txt.gz (177Mb)
    - Mt3.5_var288_chr3_tassel_20120606.txt.gz (220Mb)
    - Mt3.5_var288_chr4_tassel_20120606.txt.gz (199Mb)
    - Mt3.5_var288_chr5_tassel_20120606.txt.gz (289Mb)
    - Mt3.5_var288_chr6_tassel_20120606.txt.gz (100Mb)
    - Mt3.5_var288_chr7_tassel_20120606.txt.gz (203Mb)
    - Mt3.5_var288_chr8_tassel_20120606.txt.gz (147Mb)
    - Mt3.5_var288_chrC_tassel_20120606.txt.gz (72kb)
    - Mt3.5_var288_chrT_tassel_20120606.txt.gz (8Mb)
    - Mt3.5_var288_chrU_tassel_20120921.txt.gz (72Mb)
    - checksums_md5.txt
- Mt3.5_var288_K-matrix_20120606.txt.gz (575kb)

NOTE – chr5 also provided in two parts : chr5.1 and chr5.2

Figure 1. Distance to the nearest gene. The distance is positive if the gene is downstream and negative if the gene is upstream from the SNP.

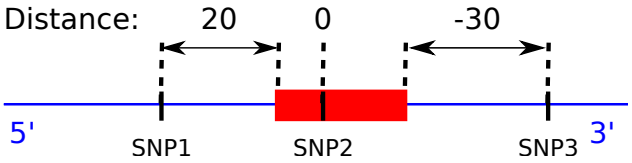

Table 1. Gene context abbreviations.

C	Coding sequence
I	Intron
3	3’ UTR
5	5’ UTR
0	Intergenic

Table 2. Sites with degenerate bases in the reference genome. These sites were removed from SNP and context files as of September 21, 2012, but they remain in TASSEL files.

Chr	Sites
Mt1	0
Mt2	55
Mt3	0
Mt4	2
Mt5	4
Mt5.1	0
Mt5.2	4
Mt6	0
Mt7	31
Mt8	2
MtC	0
MtT	3
MtU	0