Mt 4.0 HapMap Downloads

Last Updated: 26 June 2014
Created: 22 April 2014

Medicago HapMap Project (Version Mt4.0)
University of Minnesota and National Center for Genome Resources
http://www.medicagohapmap.org/
Joseph Guhlin, Peng Zhou, Andrew Farmer, Jeremy Yoder, Kevin Silverstein, John Stanton-Geddes, Roman Briskine, Peter Tiffin, Joann Mudge, Nevin Young

Due to the large size of the files, it is recommended that you use a command-line utility to download files, such as **wget**, which is available on Mac OS X, Windows, and Unix-like operating systems.

SNP Data

262 *Medicago truncatula* accessions were sequenced using Illumina. 55 accessions representing sister taxa and deeply derived lines were also sequenced. Reads were aligned to the *M. truncatula* v4.0 reference genome, representing the A17 genotype (HM101, Young et al, 2011). Twenty-six *M. truncatula* accessions (HM001-HM016, HM019-HM021, HM023-HM028, and HM101) were sequenced to 15X average aligned depth. The remaining accessions were sequenced to an average coverage depth of ~6X (Branca et al, 2011; Stanton-Geddes et al, 2013). The 262 *Medicago truncatula* accessions are the ones used for GWAS studies and are found in our “SNP Data”. The SNP calls for Sister Taxa lines are also available, see the section Sister Taxa below.

Alignment and SNP/indel calling was performed at NCGR using GSNAP and GATK. GSNAP (version 2013-03-31) was used to produce initial bam alignment files (alignment parameters:--max-mismatch 0.06 --terminal-threshold=1000 --npaths=1). The standard GATK best practices pipeline was used to process the aligned data (marking PCR/optical duplicates with Picard tools, realignment around indels, and two rounds of “iterative truth” base quality score recalibration). UnifiedGenotyper was used to produce a master VCF file on the processed bam files in multi-sample calling mode and assuming a diploid genome. Variant Quality Score Recalibration, the typical last step of the GATK best practices pipeline, was not applied, since a rich validation data set to use for training was not available.

The following variant calls were marked in the master VCF file and removed from the filtered data set: Non-SNPs (indels and multi-base substitutions), those having a combined sample read depth exceeding 5000, and those with multiple alternative alleles. Heterozygous calls were reset to homozygous reference (or homozygous variant) if the reported probability likelihood (PL) differed from its homozygous counterpart by fewer than 40 phred scale points, as reported by the Unified Genotyper. (This editing was performed with the assumption that, in a selfing species like Medicago, most of the heterozygous calls will be artifact due to low coverage depth, alignment error, or other non-biological causes.) Indeed, following expectation, the vast majority of genotypes that were altered in this manner had very low depth of coverage, and very little evidence supporting the heterozygous genotype call.

SNPs are provided in various formats: ***i)*** filtered set; ***ii)*** complete set -- includes variants (non-SNPs and low quality calls) that did not pass the filters; and ***iii)*** filtered SNPs on individual chromosomes. ChrU represents

all SNPs on unplaced scaffolds in the Mt 4.0 assembly. The complete set of files notes whether the variant call passed all additional filters or, if not, which filter(s) caused the SNP to fail.

The earlier *M. truncatula* HapMap v3.5 listed 288 *M. truncatula* accessions. Since that time, some accessions have been determined to not be true *M. truncatula* accessions and some are deeply derived and therefore not useful to include within GWAS studies. These can be retrieved via the **everything** file.

Table 1: ***File Descriptions for SNP Data***

chr*-filtered-set.bcf	Chromosome-specific SNP calls that have passed all filters.
filtered-set-2014Apr14.bcf	Complete set of SNPs that have passed all filters located on all chromosomes and scaffolds.
complete-set-2014Apr15.bcf	All variants detected by GATK, including those that have **not passed filters.**
everything-2014Apr18.bcf	Contains all lines, including sister taxa, lines not included in GWAS dataset, and results of filters. Variants that were called but did not pass filters are also included here.

Comparison to Mt 3.5 Release

A comparison of 3.5 and 4.0 SNP calls in a 40 kbp region on chromosome 5 is available for download. http://www.medicagohapmap.org/downloads/Mt40/Mt4.0_versus_Mt3.5_SNP_comparison.pdf

The previous release included 288 lines for GWAS studies. We are recommending 262 lines with the new release of the Mt 4.0 SNP calls. The table below provides the rationale for removal of these lines. Please note that all lines had SNPs called against them, and are available in the Sister Taxa files and in the everything file.

Table 2: ***Accessions removed from GWAS SNP set.*** *These accessions were previously in the Mt 3.5 SNP dataset but are now considered Sister Taxa or too deeply derived for use in GWAS analysis by subsequent phylogenetic analysis.*

Accession	**Reason**	**Reference**
HM017	Not M. truncatula	Yoder et al, 2013
HM018	Not M. truncatula	Yoder et al, 2013
HM022	Not M. truncatula	Yoder et al, 2013
HM029	Not M. truncatula	Yoder et al, 2013
HM030	Not M. truncatula	Yoder et al, 2013
HM216	Not M. truncatula	Stanton-Geddes et al, 2013
HM246	Not M. truncatula	Stanton-Geddes et al, 2013
HM247	Not M. truncatula	Stanton-Geddes et al, 2013

HM248	Not M. truncatula	Stanton-Geddes et al, 2013
HM249	Not M. truncatula	Stanton-Geddes et al, 2013
HM250	Not M. truncatula	Stanton-Geddes et al, 2013
HM251	Not M. truncatula	Stanton-Geddes et al, 2013
HM252	Not M. truncatula	Stanton-Geddes et al, 2013
HM254	Not M. truncatula	Stanton-Geddes et al, 2013
HM255	Not M. truncatula	Stanton-Geddes et al, 2013
HM257	Not M. truncatula	Stanton-Geddes et al, 2013
HM258	Not M. truncatula	Stanton-Geddes et al, 2013
HM261	Not M. truncatula	Stanton-Geddes et al, 2013
HM264	Not M. truncatula	Stanton-Geddes et al, 2013
HM273	Not M. truncatula	Stanton-Geddes et al, 2013
HM274	Not M. truncatula	Stanton-Geddes et al, 2013
HM275	Not M. truncatula	Stanton-Geddes et al, 2013
HM291	Not M. truncatula	Stanton-Geddes et al, 2013
HM292	Not M. truncatula	Stanton-Geddes et al, 2013
HM303	Not M. truncatula	Stanton-Geddes et al, 2013
HM317	Not M. truncatula	Stanton-Geddes et al, 2013

Sister Taxa

We excluded twenty-eight accessions from the association study SNP dataset because they represent other *Medicago* subspecies (Yoder et al., 2013). Five of those accessions (HM017, HM018, HM022, HM029, and HM030) were sequenced to 15X average aligned depth while the rest had the average depth of ~6X. These are designated as Sister Taxa and include those in the previous *Comparison to Mt 3.5 Release* table and also HM102 and HM318 - HM339. SNP calls for those accessions were generated using the same pipeline previously mentioned.

SNPs called for sister taxa and deeply derived lines of *M. truncatula* are provided in BCF format. However, due to their distant relatedness to *M. truncatula* ssp. *truncatula* accessions these data are not available in files set up for association analysis. The everything-2014Apr18.bcf file also includes all Sister Taxa lines, with the other lines used for association studies.

Table 3: ***File Descriptions for Sister Taxa Data***

chr*-filtered-set.bcf	Chromosome-specific SNP calls that have passed all filters.
sister-taxa-filtered-set-2014Apr14.bcf	Complete set of SNPs that have passed all filters located on all chromosomes and scaffolds.
sister-taxa-complete-set-2014Apr15.bcf	All variants detected by GATK, including those that have **not passed filters.**
everything-2014Apr18.bcf	Contains all lines, including sister taxa, lines not included in GWAS dataset, and results of filters. Variants that were called but did not pass filters are also included here. This file is in the download section “SNPS: Combined”.

File Formats

SNPs are provided as BCF files (Binary Variant Call Format) v2.1. A CSI file representing a fast-access index for the corresponding BCF file is provided with each BCF file. File format information is provided at the following link: https://github.com/samtools/hts-specs

BCF may be converted to VCF files using **bcftools, vcftools,** and other programs. BCF files are larger than previous formats hosted on the *Medicago* HapMap website, but provide more information. This format is becoming the standard file type for distributing variant data.

Files for Association Studies (when available) are in HapMap format, and are confirmed to work with GAPIT (**VERSION)** and TASSEL (**VERSION)**. Missing genotypes are recorded as **NN** (note that GAPIT requires SNP be called for all lines and will automatically infer the missing state using a fast but not necessarily accurate approach). Commands and script used to generate the HapMap format from BCF is included at the bottom of this document.

Association Analysis using TASSEL

SNPs for use in association analysis with TASSEL are provided in the hapmap.tgz file. This includes all 8 chromosomes. SNPs must be genotyped in > 100 accessions and have a minor allele frequency (MAF) > 2%. This is consistent with the methodology used in association studies for Mt 3.5.

We recommend TASSEL 5.0. Each chromosome should be run separately and can be run in parallel, up to the memory limit of the machine. The memory settings below have worked for us previously, this command can be run in a bash shell before executing TASSEL.

```
export _JAVA_OPTIONS="-Xms5632m -Xmx5632m”
```

Gene Context

SNPs in the GWAS panel have been analyzed by the program SnpEff with regards to the genic and transposable elements annotations provided by JCVI (available on their website) and the output file is available as a compressed VCF file. Additionally, we have converted the output from SnpEff to a tab-delimited file, and also provided files separated by chromosome. These files are available individually and gzip compressed, or together as gzip compressed tar file (.tgz).

Because SNPs may affect multiple genes, a single SNP may have multiple lines reporting possible effects. SNPs within 1000bp upstream or downstream from a gene are reported as potential modifiers. Additionally, SnpEff assumes that all genes and TEs are protein coding.

The columns in this file are as follows:

1. Chromosome Name
2. Position
3. SNP Quality (assigned by GATK)

4. Reference Allele
5. Alt. Allele
6. Effect - Sequence Ontology term that is further explained by SnpEff's documentation.
7. Effect Impact - See the link in #6
8. Functional Class - None, Silent, Missense, Nonsense
9. Codon Change / Distance - This is the codon change sequence (reference / alternative) OR the distance to the indicated gene
10. Amino Acid Change
11. Gene Name
12. Transcript Name
13. Exon/Intron Rank
14. Warnings - Any warnings generated by SnpEff

The full manual for SnpEff is available here. A custom built database was used with information from JCVI, including protein sequences.

References

Branca A, Paape T, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C, Denny R, Sadowsky, MJ, Ronfort J, Bataillon T, Young ND, Tiffin P (2011) Whole-genome nucleotide diversity, recombination, and linkage-disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. USA* 108: E864-870. doi:10.1073/pnas.1104032108.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118*. (2012). Fly (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J., ... & Tiffin P (2013) Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. *PLoS One*, *8*(5), e65688.

Yoder, J. B., Briskine, R., Mudge, J., Farmer, A., Paape, T., Steele, K., ... & Tiffin, P. (2013). Phylogenetic signal variation in the genomes of Medicago (Fabaceae). *Systematic biology*, *62*(3), 424-438.

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, *480*(7378), 520-524.

Commands and Scripts

Please change the file names OUTPUT_FILE_NAME and INPUT_FILE_NAME appropriately

Convert BCF to VCF

```
bcftools view -f PASS -O v -o OUTPUT_FILE_NAME.vcf INPUT_FILE_NAME.bcf
```

Convert BCF to tab-delimited file

This will output the SNPs and the called genotypes

```
bcftools query -H -f \
```

```
"%CHROM\t%POS[\t%TGT]\n" \
INPUT_FILE_NAME.bcf > OUTPUT_FILE_NAME.txt
```

Export a region from BCF to tab-delimited file

You must have also downloaded or regenerated the .csi file to perform these types of operations
Replace REGION below as appropriate, some examples are included below:
chr5
chr5:104932-107932

```
bcftools query -H -r REGION -f \
"%CHROM\t%POS[\t%TGT]\n" \
INPUT_FILE_NAME.bcf > OUTPUT_FILE_NAME.txt
```

BCF to HapMap Format

Copy and paste the commands below.
BCFtools can export VCF files to a nearly compliant HapMap format. The perl script finishes the process.
BCFtools is available here: https://github.com/samtools/bcftools
The chloroplast and scaffolds are removed below.

```
bcftools query -H -f \
"%CHROM:%POS\t.\t%CHROM\t%POS\t.\t.\t.\t.\t.\t.\t.[\t%TGT]\n" \
filtered-set-2014Apr14.bcf > snps.hapmap.unprocessed

perl hapmap.pl | grep -v chl | grep -v scaffold > snps.hapmap
```

hapmap.pl

```
open(my $fh, "<snps.hapmap.unprocessed");

my $first = 1;

while (<$fh>) {
        $output = $_;
        if ($first) {
                $output =~ s/^#?\s+//g; # Remove first # and spaces after
                $output =~ s/\[\d+\]//g;
                $output =~ s/\-I//g;
                $output =~ s/:GT//g;
                print $output;
                $first = 0;
        } else {
                $unknowns =()= $output =~ /\.\/\./g;
                $output =~ s/\///g;
                $output =~ s/\t\.\./\tNN/g;
                $output =~ s/\t\./\tNA/g;
                $output =~ s/chr//g;
                if ($unknowns <= 162) { # 262 lines, so 262 - 100 = 162
                        # only if at least 100 accessions make a
                        # genotype call
```

```
                        print $output;
                }
        }

}
```

Lines in Genome-wide Association Study SNP Dataset

Table 4: ***Lines in Mt 4.0 SNP GWAS dataset***

HM001	HM059	HM121	HM184	HM239
HM002	HM060	HM122	HM185	HM240
HM003	HM061	HM124	HM186	HM241
HM004	HM062	HM125	HM187	HM242
HM005	HM063	HM126	HM188	HM243
HM006	HM064	HM127	HM189	HM244
HM007	HM065	HM128	HM190	HM245
HM008	HM066	HM129	HM191	HM253
HM009	HM067	HM130	HM192	HM256
HM010	HM068	HM131	HM193	HM259
HM011	HM069	HM133	HM194	HM260
HM012	HM070	HM134	HM195	HM262
HM013	HM071	HM135	HM196	HM266
HM014	HM072	HM138	HM197	HM267
HM015	HM073	HM139	HM198	HM268
HM016	HM074	HM141	HM199	HM269
HM019	HM075	HM143	HM200	HM270
HM020-I	HM076	HM145	HM201	HM271
HM021	HM077	HM146	HM202	HM276
HM023	HM078	HM147	HM203	HM277
HM024	HM079	HM148	HM205	HM278
HM025	HM080	HM149	HM206	HM279
HM026	HM081	HM150	HM207	HM280
HM027	HM082	HM151	HM208	HM287
HM028	HM083	HM152	HM209	HM288
HM031	HM084	HM153	HM210	HM289
HM032	HM085	HM154	HM211	HM290
HM033	HM086	HM155	HM212	HM293
HM034	HM087	HM156	HM213	HM294
HM035	HM088	HM157	HM214	HM295
HM036	HM089	HM159	HM215	HM296
HM037	HM091	HM160	HM217	HM297
HM038	HM092	HM161	HM218	HM298
HM039	HM093	HM162	HM219	HM299
HM040	HM095	HM163	HM220	HM300
HM041	HM096	HM164	HM221	HM301

HM042	HM097	HM165	HM222	HM302
HM043	HM098	HM166	HM223	HM304
HM044	HM099	HM167	HM224	HM305
HM045	HM101	HM168	HM225	HM306
HM046	HM105	HM169	HM226	HM307
HM047	HM106	HM170	HM227	HM308
HM048	HM107	HM172	HM228	HM309
HM049	HM108	HM173	HM229	HM310
HM050	HM109	HM175	HM230	HM311
HM051	HM111	HM176	HM231	HM312
HM052	HM112	HM177	HM232	HM313
HM053	HM114	HM178	HM233	HM314
HM054	HM115	HM179	HM234	HM315
HM055	HM117	HM180	HM235	HM316
HM056	HM118	HM181	HM236	
HM057	HM119	HM182	HM237	
HM058	HM120	HM183	HM238	