Last update: 7/18/04

MMAP Import and Export File Options

MMAP has commands to import data from Plink, Minimac, IMPUTE2 directly into a binary genotype file and commands to export to Mach and Beagle format. Some of these options are being beta tested or under development.

Plink

MMAP imports Plink binary format files into an SxM or MxS genotype binary file, depending on the Plink format, which is automatically detected.

--plink_bfile2mmap –plink_bfile <prefix> -- binary_output_prefix <mmap prefix>
Converts files <prefix>.bim, <prefix>.bed, <prefix>.fam into binary genotype file <mmap prefix>.bin and MMAP pedigree <mmap prefix>.ped.csv extracted from the <prefix>.fam.

Mach/MiniMac

MMAP imports Mach info and dosage files into an SxM binary genotype file. Since the map information is not contained in the info file, the Mach map file is required. Dosage files can be compressed or uncompressed. Options for reading in the probability file to create dominant and recessive dosages are under development.

--mach_dose2mmap –mach_info_filename <info file> --mach_dose_filename <dose file> imputation_map_filename <map file> --binary_output_filename <SxM binary gentype file> --genotype_dosage_short
<info file> and <dose file> are a list of Mach output files to be used. The chromosomes must be in the same order in both files. The option --genotype_dosage_short stores the dosage as 2 bytes with precision 4-5 decimal places. The --genotype_dosage_char option will store the dosage as 1 byte with 2-3 decimal place precision, reducing file size by half. It is recommended to create a single binary file containing all the chromosomes for flexibility of analysis even though the file will be large.

IMPUTE2

The IMPUTE2 import assumes that the probability *.imputed* and the information *.imputed_info* files are available. Since subject id information is not contained in the output files, this information must be included in the command line. This option is now set up to combine files per chromosome to manage large files. Thus, the chromosome is required input. The default coding is to use 1 byte. Since the probabilities are 3 decimal places, the 2 byte option is recommended. The following conventions are used to handle the different variant types found in the current imputation panels.
Conventions:

1. **SNPNAME** is coded using the **RSNUM** value. If the marker was typed (2 in the output column), the SNPNAME is the same as **RSNUM**, otherwise an “i” is appended. So rs123 becomes irs123
2. **RSNUM** is coded using the rs_id value in the .imputed or .imputed_info file, expect if it is missing (dot in the output), then it is coded as <chr>:<position>. Use RSNUM when reporting results.

3. **STRAND** is set to + by default as no strand information is available.
4. **ALLELE** as a coded as single characters using the nucleotides is both alleles are single characters. Otherwise, R is used for the non-coded allele and I or D as the effect allele depending on if the non-coded allele is a substring of the effect allele (I), or a superstring (D). Marker with <DEL> are coded as R/R. MMAP outputs a file that contains the original alleles and the codding. MMAP also supports multiallelic options where the alleles are as the original but truncated beyond a maximum length.

The following are **required** options
--impute2_prob2mmap
--impute2_prob_filename <file1> <file2> … <fileN>
--**impute2_info_filename** <file1> <file2> …<fileN>
--**chromosome** <chr>
--**subject_id_filename** <file>
--binary_output_filename <file>
--**csv_output_filename** <file>

The prob and info files should be in the same chunk order. The prob files can be gzip'd. The subject id file does NOT have a header. The imputation quality score *info* is embedded in the binary genotype file and can be outputted when running the single variant analysis. The default is to create an additive dosage.

--genotype_dosage_short add option to increase accuracy of stored dosage. Doubles file size.

Additional options to create alternative dosage files
--dominant_dosage creates dosage= 0*prob(AA) + 1*prob(AB) + 1*prob(BB)
--recessive_dosage creates dosage= 1*prob(AA) + 1*prob(AB) + 0*prob(BB)
--**het_dosage** creates dosage= 0*prob(AA) + 1*prob(AB) + 0*prob(BB)

NOTE: If you run –dom with the binary_genotype_file created using –dominant_dosage you will get a message that the option is not supported. For imputed data there is actually no model since all genotypes have a value. The –dom for observed data tells MMAP how to combine genotypes into dominant dosages and to fill in missing values. These do not apply for in this case. Thus, no model statement is needed.

Once the chromosome-specific files are created, they can then be combined into a single MMAP binary using
--combine_binary_genotype_files <file1> <file2> … <fileN>
--binary_output_filename <file>
The input files and output file are MMAP marker-by-subject binary genotype files.

Recommendation: To reduce the size of the combined binary genotype file, once the chromosome specific files are created, run the allele frequency option. The output will contain the minor allele frequency and imputation quality score, which can be used to extract a marker set based on minor allele frequency and/or imputation quality threshold. This marker set can

then be used to create reduced binary genotype files before combining to the full file. See MMAP.genotype.pdf for details on these options.

VCF

Under development

Exporting Data

Plink

--subject_by_marker_mmap2plink --binary_input_filename <SxM binary gentype file> --plink_output_prefix <prefix>
Creates <prefix>.map and <prefix>.ped which can then be converted into Plink binary format. Currently no support of export directly into binary.

--marker_by_subject_mmap2tped --binary_input_filename <SxM binary genotype file> --plink_output_prefix <prefix>
Creates <prefix>.fam, <prefix>.bim and <prefix>.tped

Mach/Merlin

To be documented

Beagle

To be documented

MSMS

To be documented

ForSim

To be documented

Idcoeffs

To be documented