1. Read GEMMAmanual.pdf
2. Format input files: GEMMA notes
   1. PLINK (binary ped)
      1. genotype
         1. 0 for missing genotypes
      2. phenotype
         1. -9 or NA for missing phenotypes
      3. 3 files <http://zzz.bwh.harvard.edu/plink/>
         1. \*.bed
            1. default SNP-major mode
            2. see PLINK instructions below for \*.ped to \*.bed
         2. \*.bim
            1. column 5, binary alleles: 1 = minor allele = effect allele
            2. column 6, binary alleles: 0 = major allele
         3. \*.fam
            1. GEMMA only needs column 1 (individual ID), column 6 (phenotype)

-n (num) with -n 1 means column 6 is pheno, -n 2 means 7… etc.

* + - * 1. if phenotype is disease status, set cases = 1, controls = 0.
  1. BIMBAM <http://stephenslab.uchicago.edu/software.html>
     1. mean genotype file
        1. mean genotype can be in gzip
        2. col 1 = SNP ID
        3. col 2, col 3 = allele types, minor first
        4. col 3 – x = posterior/ imputed mean genotypes of individuals numbered 0-2
        5. example, 2 SNPs and 3 individuals
           1. rs1, A, T, 0.02, 0.80, 1.50
           2. rs2, G, C, 0.98, 0.04, 1.00
        6. GEMMA ignores allele types in col 2, col 3
        7. minor allele is effect allele only if minor allele = 1, major allele = 0
        8. can generate BIMBAM mean genotype file from IMPUTE genotype file:
           1. <http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html>
           2. cat [impute filename] | awk -v s=[number of samples/individuals] ’{ printf $2 "," $4 "," $5; for(i=1; i<=s; i++) printf "," $(i\*3+3)\*2+$(i\*3+4); printf "\n" }’ > [bimbam filename]
           3. make sure ‘ symbols are correct
     2. phenotype file
        1. each line = phenotype value for each individual
        2. must keep same individual order as mean genotype file
        3. numeric ONLY
        4. NA for missing values
        5. multiple phenotypes as multiple columns. For association tests, select column with -n (num)
           1. -n 1 uses col 1
           2. -n 2 uses col 2, etc.
        6. if binary, controls = 0, cases = 1
     3. optional SNP-annotation file
        1. col 1 = SNP id
        2. col 2 = bp position
        3. col 3 = chromosome numbered
        4. must include all SNPs from mean genotype file but order is unimportant
     4. recommended if imputing: any real 0-2
  2. relatedness matrix
     1. can skip this if running association test with a linear model (see IV-A)
     2. can be in gzip
     3. can be estimated from genotypes using GEMMA (see example/demo.txt)
     4. can be original relatedness matrix
        1. n x n matrix
           1. -km 1 to specify format
           2. must use this if using BIMBAM
           3. row and column correspond to individuals in same order as \*.fam or mean genotype file. Each cell is relatedness value between pair of individuals.
        2. Or, id id value format
           1. -km 2 to specify format
           2. only works for PLINK binary ped
           3. first 2 columns = 2 individual id numbers, col 3 = relatedness value
           4. IDs need not be in \*.fam order
           5. missing relatedness values considered 0
     5. can be eigen values and eigen vectors of original relatedness matrix
        1. file 1: eigen values
           1. 1 column, n elements
        2. file 2: corresponding eigen vectors
           1. n x n matrix
  3. marginal z-scores file
     1. can be in gzip
     2. row 1 = header (SNP, N, Z, INC\_ALLELE, DEC\_ALLELE)
     3. column 1 = SNP id
     4. col 2 = total number of SNPs
     5. col 3 = marginal z-score
     6. col 4, 5 = SNP alleles
     7. SNP order does not matter
  4. SNP category file
     1. can be in gzip
     2. row 1 = header
     3. col 1 = chromosome number (optional)
     4. col 2 = bp position (optional)
     5. col 3 = SNP id
     6. col 4 = genetic distance on chr (optional)
     7. col 5 – x = non-overlapping categories (e.g. CODING, UTR, PROMOTER, INTRON, ELSE) with 0/1 in each (only 1 category per SNP)
     8. SNP order does not matter
  5. Covariates file format
     1. optional
  6. LD score file
     1. LD scores for all SNPs
     2. row 1 = header
     3. col 1 = chromosome number (optional)
     4. col 2 = SNP id
     5. col 3 = bp position (optional)
     6. col 4 = LD score of SNP

1. format input files: PLINK notes <http://zzz.bwh.harvard.edu/plink/>
   1. can also use for quantitative trait association
   2. need PED and MAP files as input: \*.ped \*.map
      1. visual of file formats: <http://www.shapeit.fr/pages/m02_formats/pedmap.html>
         1. also here <http://www.gwaspi.org/?page_id=145>
   3. To convert myPlinkTextData.ped and myPlinkTextData.map in Plink binary format, use Plink as follows:
      1. plink --file myPlinkTextData --make-bed --out myPlinkBinaryData
   4. my code
      1. downloaded plink-1.07-x86\_64/ from bwh site to GitRepos/
      2. cd to Documents/GitRepos/BcAt\_RNAGWAS/data/B05\_GEMMA
      3. ./plink –noweb --file 01\_PLINK/dpcharMAF20NA10 --maf 0.2  –[missing-genotype](https://www.cog-genomics.org/plink/1.9/input#missing_genotype) 0 -make-bed –out 01\_PLINK/binMAF20NA10
         1. CODE 1 on myGEMMA\_todo.txt
         2. getting a really stupid error where PLINK thinks genotype=0 is an allele so all sites are more than biallelic
            1. ERROR: Locus "SNP39118" has >2 alleles:

individual "FAM1" "1.01.02" has genotype [ "A" "A" ]

but we've already seen [ "0" ] and [ "C" ]

* + - * 1. instructions: <http://zzz.bwh.harvard.edu/plink/dataman.shtml>

plink --file mydataname --recode12 --out mydataname\_allele12

./plink --noweb –file 01\_PLINK/dpcharMAF20NA10 –recode12 –out 01\_PLINK/dpcharMAF20NA10\_allele12

CODE 2 on myGEMMA\_todo.txt

gives same dumb error

plink --file mydataname --allele1234 --make-bed

./plink --noweb --file 01\_PLINK/dpcharMAF20NA10 --maf 0.2 --allele1234 –make-bed --out 01\_PLINK/binMAF20NA10\_allele1234

CODE 3 on myGEMMA\_todo.txt

* + - * 1. also trying: removing isolate 1.01.06.1 which is a duplicate. IDK if this will help

rerunning CODE 2…

rerunning CODE 3...

1. run GEMMA
   1. If running association test with a linear model, does not require a relatedness matrix
   2. missingness: default omit SNPs <5% missingness
      1. -miss (num)
      2. -miss 0.1 = 10% threshold
   3. MAF: default omit SNPs with MAF < 1%
      1. -maf (num)
      2. -maf 0.05 = 5% threshold
   4. basic usage with PLINK or BIMBAM
      1. ./gemma -bfile [prefix] -lm [num] -o [prefix]
         1. “-bfile [prefix]” specifies PLINK binary ped file prefix
      2. ./gemma -g [filename] -p [filename] -a [filename] -lm [num] -o [prefix]
         1. “-g [filename]” specifies BIMBAM mean genotype file name
         2. “-p [filename]” specifies BIMBAM phenotype file name
         3. “-a [filename]” (optional) specifies BIMBAM SNP annotation file name
      3. “-lm [num]” option specifies which frequentist test to use
         1. “-lm 1” performs Wald test, “-lm 2” performs likelihood ratio test, “-lm 3” performs score test, and “-lm 4” performs all the three tests
      4. “-o [prefix]” specifies output file prefix
   5. my code here:
      1. ./gemma -bfile 01\_PLINK/binMAF20NA10 -miss 0.1 -maf 0.2 -lm 4 -o 02\_GEMMA/binMAF20NA10\_PLINK
2. GEMMA for population structure
   1. ./gemma -bfile [prefix] -gk [num] -o [prefix]
      1. where the \-gk [num]" option speci\_es which relatedness matrix to estimate,
         1. i.e. \-gk 1" calculates the centered relatedness matrix while \-gk 2" calculates the standardized relatedness matrix;
      2. \-b [prefix] specifies PLINK binary ped file prefix
      3. if SNPs with lower minor allele frequency tend to have larger effects (which is inversely proportional to its genotype variance), then the standardized genotype matrix is preferred. If the SNP effect size does not depend on its minor allele frequency, then the centered genotype matrix is preferred. In our previous experience based on a limited examples, we typically find the centered genotype matrix provides better control for population structure in lower organisms