Imputation and Meta-analysis

Sarah Medland – OHBM 14/06/2015

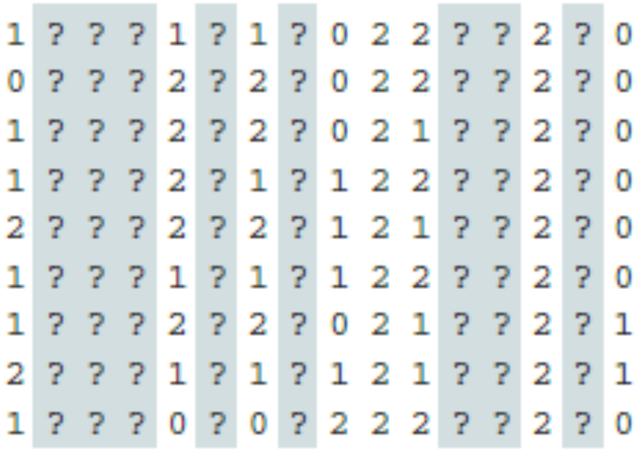
QIMR Berghofer
Medical Research Institute

Imputation

- Why do we impute
  - To allow *comparison* with other samples on other chips
  - To *fine map* – ie run association at variants we have not genotyped
  - To improve *call rate* – ie increase the number of variants available for poorly genotyped samples (not ideal)
  - To identify *genotyping errors*

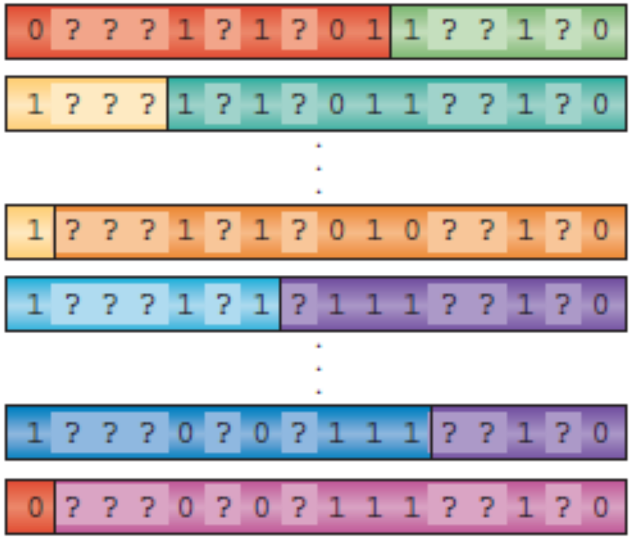
A quick conceptual theory of imputation

- Start with some genotype data



1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

- using LD the structure within your data phase your data to reconstruct the haplotypes



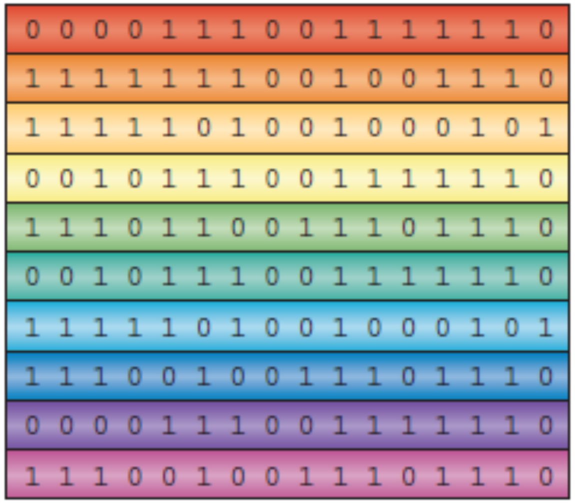
```
0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
⋮
1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0
⋮
1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
```

A quick conceptual theory of imputation

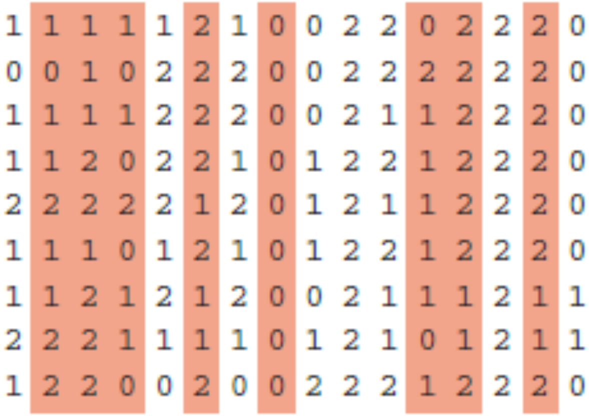
- Compare your phased data to the references

- Use the LD structure to Impute in the missing genotypes

(Marchini, J. and Howie, B. 2010. *Nat Rev Genet* 11 499-511.)



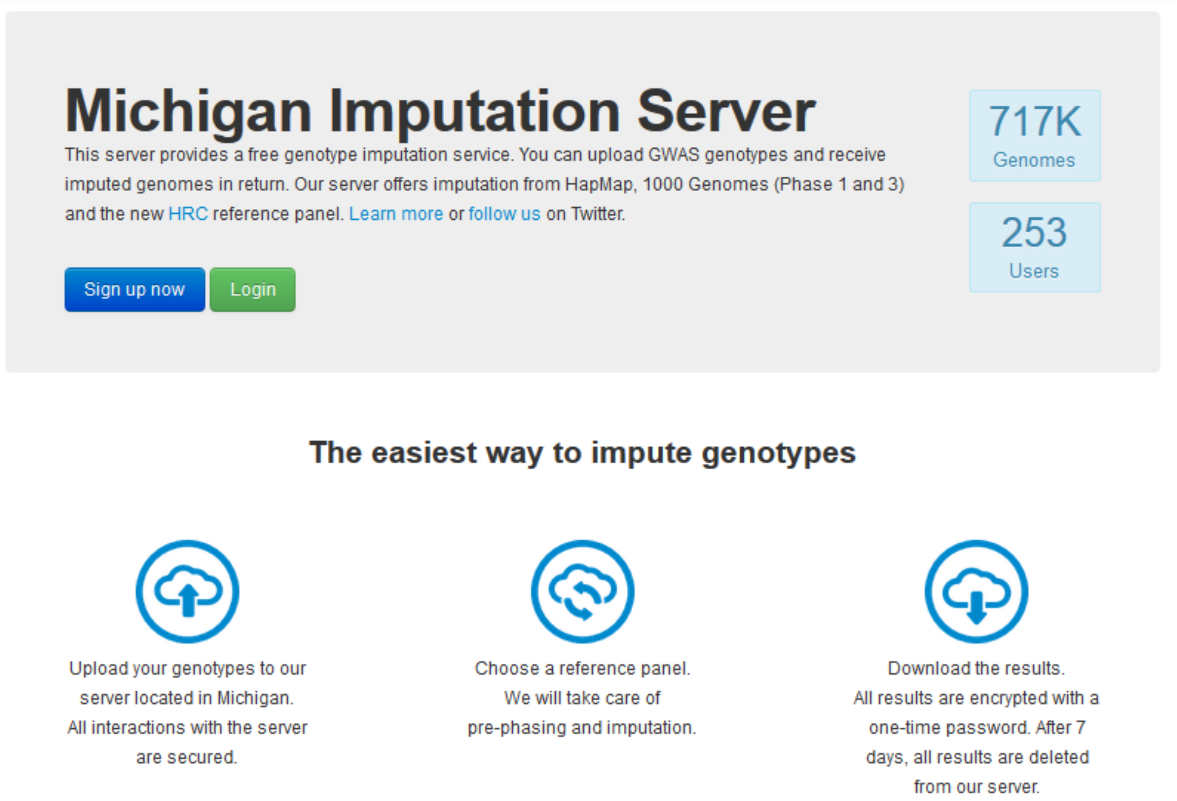
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0



1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

Easiest (and best) way of imputing

- Use the Imputation Servers
  - https://imputationserver.sph.umich.edu/
  - https://imputation.sanger.ac.uk/



Michigan Imputation Server

This server provides a free genotype imputation service. You can upload GWAS genotypes and receive imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and 3) and the new HRC reference panel. Learn more or follow us on Twitter.

717K Genomes

253 Users

The easiest way to impute genotypes

Upload your genotypes to our server located in Michigan. All interactions with the server are secured.

Choose a reference panel. We will take care of pre-phasing and imputation.

Download the results. All results are encrypted with a one-time password. After 7 days, all results are deleted from our server.

But I'm going to assume you have the time, computational capacity, storage space and desire to do this yourself…

Step 1 – Pick your references

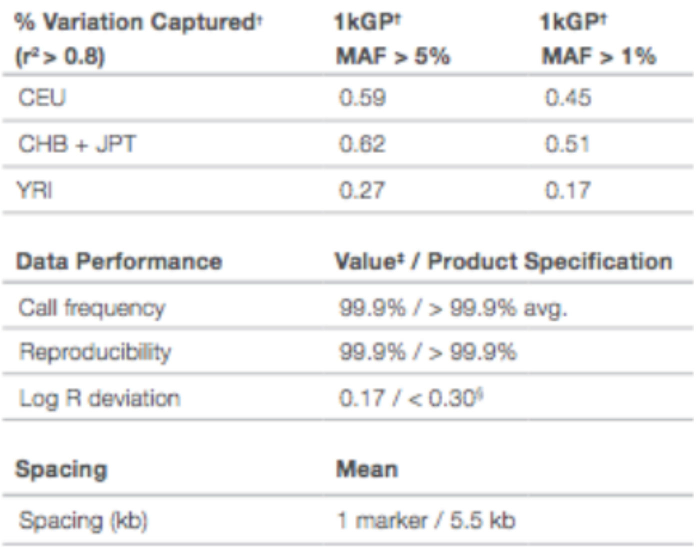
- HapMapII or HapMapIII
  - 2.4M and 1.3M variants respectively
  - Well imputed and well known set
  - Good for first imputation run – not commonly used anymore
- 1KGP aka 1000GP
  - Phase1v3 ~37M variants of these ~11M will be useable
    - 1,092 individuals
  - Phase3v5 ~82M variants of these ~12M will be useable
    - 2,504 individuals
- Haplotype reference consortium
  - Only from the Imputation servers
  - 39M variants 32,488 individuals of these ? useable…

Pick your references

- All Ethnicities vs Specific Ethnicity panels
  - Consider what the consortiums/collaborators you want to work with want to do
  - Case by case basis
  - All ethnicities panels are larger (and slower)
  - Can be more accurate – esp for a ‘cosmopolitan US’ sample
  - May not improve imputation for homogeneous populations or those with strong founder effects

Step 2- Genotype data

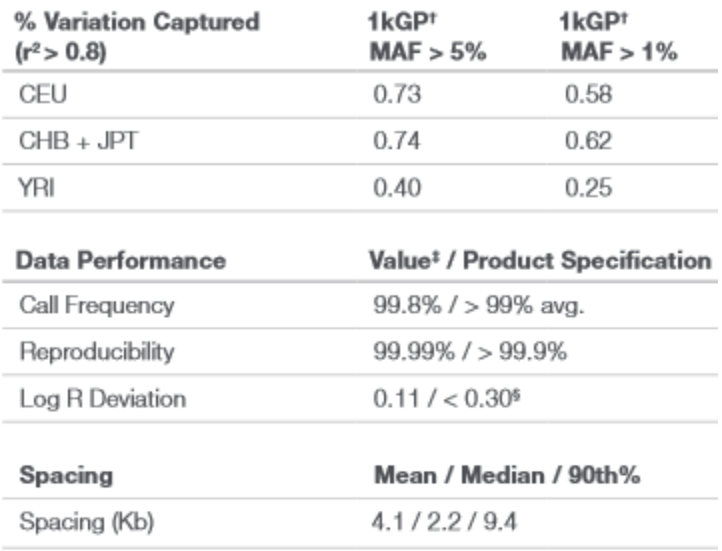
- Ideally use a chip designed for imputation
  - All chips have data sheets if you are obtaining genotyping make sure you check the sheet before choosing the chip!
  - Also look for papers on imputation using your preferred chip and ask authors who have published using that chip
  - Check the manifests and make sure your favourite genes are covered!



% Variation Captured† ($r^2 > 0.8$)	1kGP† MAF > 5%	1kGP† MAF > 1%
CEU	0.59	0.45
CHB + JPT	0.62	0.51
YRI	0.27	0.17

Data Performance	Value‡ / Product Specification
Call frequency	99.9% / > 99.9% avg.
Reproducibility	99.9% / > 99.9%
Log R deviation	0.17 / < 0.30§

Spacing	Mean
Spacing (kb)	1 marker / 5.5 kb



% Variation Captured ($r^2 > 0.8$)	1kGP† MAF > 5%	1kGP† MAF > 1%
CEU	0.73	0.58
CHB + JPT	0.74	0.62
YRI	0.40	0.25

Data Performance	Value‡ / Product Specification
Call Frequency	99.8% / > 99% avg.
Reproducibility	99.99% / > 99.9%
Log R Deviation	0.11 / < 0.30§

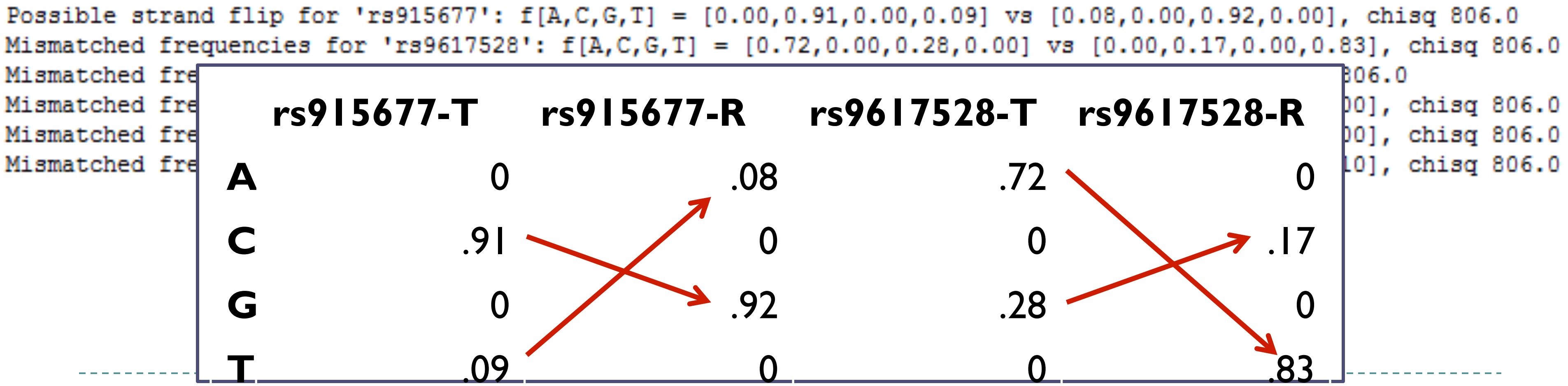
Spacing	Mean / Median / 90th%
Spacing (Kb)	4.1 / 2.2 / 9.4

Genotype Data

- Make sure your data are clean!
  - Convert to PLINK binary format
  - Exclude snps with:
    - excessive missingness (>5%)
    - low MAF (<1%)
    - HWE violations (~$P<10^{-4}$)
    - Mendelian errors
    - Exclude variants that are not in your reference panel (optional but recommended)

Genotype Data

- Make sure your data are clean!
  - Drop strand ambiguous snps (AT and CG snps)
    - □ Remember: DNA is composed of 2 antiparallel strands the complement of an A is a T and the complement of a C is G this makes it difficult to work out if the genotypes are strand aligned to the references. +ve and –ve strand is an arbitrary construct changes between builds and sources. Much better to drop these SNPs and reimpute them…
  - Align the strand of the non-ambiguous snps



```
Possible strand flip for 'rs915677': f[A,C,G,T] = [0.00,0.91,0.00,0.09] vs [0.08,0.00,0.92,0.00], chisq 806.0
Mismatched frequencies for 'rs9617528': f[A,C,G,T] = [0.72,0.00,0.28,0.00] vs [0.00,0.17,0.00,0.83], chisq 806.0
Mismatched fre                                                                                       306.0
Mismatched fre                                                                                      00], chisq 806.0
Mismatched fre                                                                                      00], chisq 806.0
Mismatched fre                                                                                      10], chisq 806.0
```

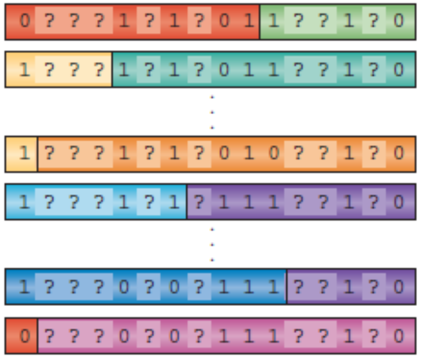
	rs915677-T	rs915677-R	rs9617528-T	rs9617528-R
A	0	.08	.72	0
C	.91	0	0	.17
G	0	.92	.28	0
T	.09	0	0	.83

Genotype Data

- Make sure your map (base pair positions) are on the correct build!
  - HapMap references were on hg18
  - 1KGP references are on hg19!
    - Distance and order of variants can change – absolutely critical that your data and the reference are on the same build!!!

Step 3 - Phase your data

- Phasing programs "use a hidden Markov model (HMM) to model the haplotypes underlying G as an imperfect mosaic of haplotypes in the set H. Compatible haplotypes are sampled for G using the forward-backward algorithm for HMMs"



0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0

1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0

⋮

1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0

1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0

⋮

1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0

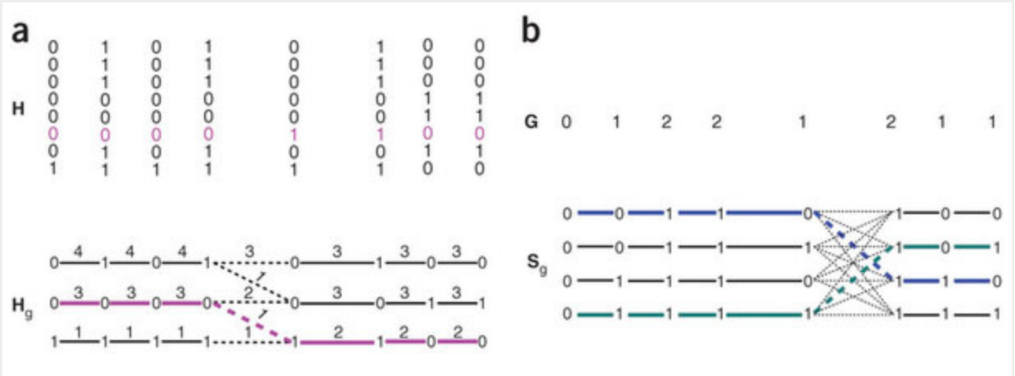
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0

- Problem: complexity is quadratic and scales with sample size and Nsnps $O(MK^2)$

Delaneau, O. et al. 2013. *Nat Meth* 10 5-6.

Phase your data

- Currently best program for phasing is SHAPEIT2
  - Delaneau, O., Zagury, J.-F. et al. 2013. *Nat Meth* 10 5-6.
- Avoids the quadratic bottle neck by:
  - "collapsing all *K* haplotypes in **H** into a graph structure, **H**$_g$, and then carrying out the HMM calculations on this graph."
  - Sampling pairs of haplotypes
- Transition accuracy is improved by drawing on surrogate family members

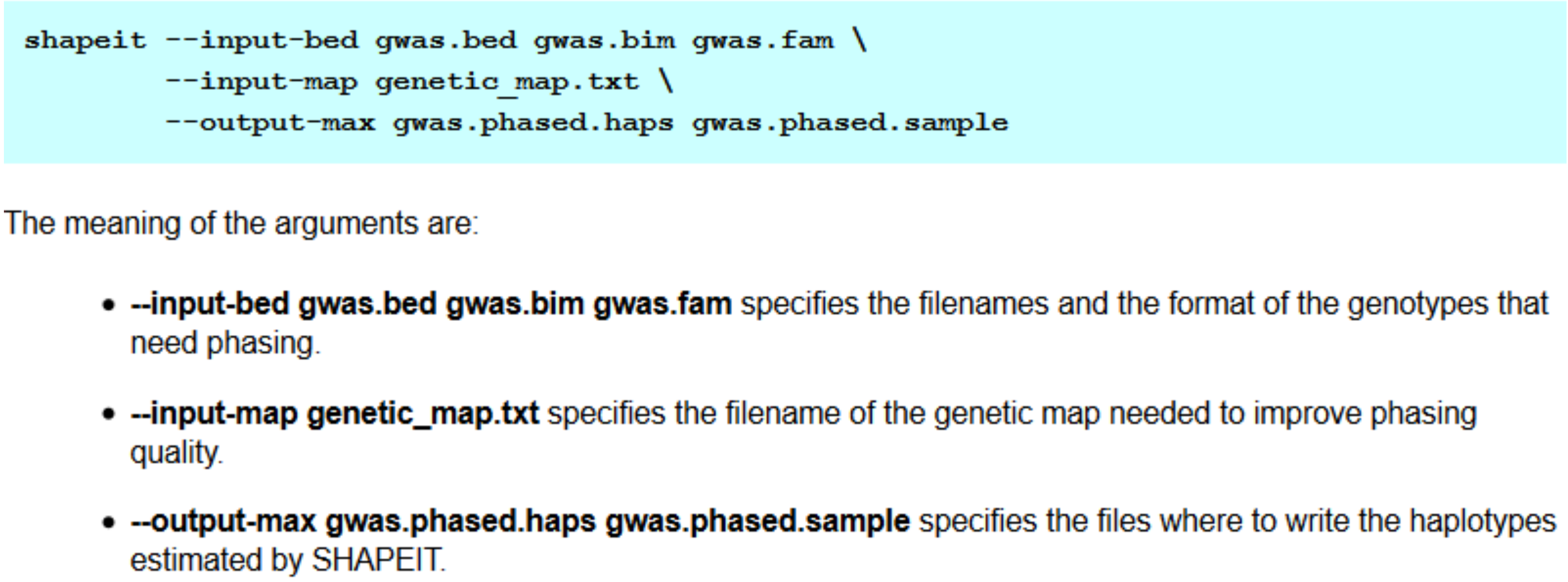


Phase your data

- SHAPEIT2
- Transition accuracy is improved by drawing on surrogate family members
  - restricts each phasing update to a set of k template haplotypes chosen separately for each individual at each iteration
  - The k templates are chosen by computing Hamming distances between an individual's current sampled haplotypes and each possible template haplotype.
  - the k templates with the smallest distances are refereed to as “surrogate family members”

SHAPEIT2

- https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
  - Can multi-thread



```
shapeit --input-bed gwas.bed gwas.bim gwas.fam \
        --input-map genetic_map.txt \
        --output-max gwas.phased.haps gwas.phased.sample
```

The meaning of the arguments are:

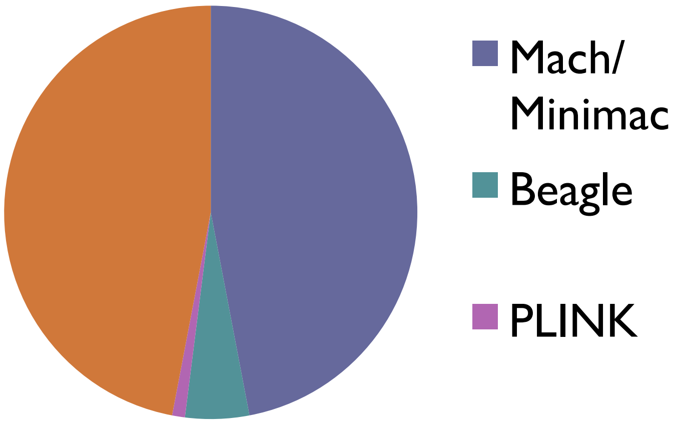
- **--input-bed gwas.bed gwas.bim gwas.fam** specifies the filenames and the format of the genotypes that need phasing.
- **--input-map genetic_map.txt** specifies the filename of the genetic map needed to improve phasing quality.
- **--output-max gwas.phased.haps gwas.phased.sample** specifies the files where to write the haplotypes estimated by SHAPEIT.

  - Note: this is a genetic map based on recombination (cM) not a physical map (BP)!

Step 4 – Impute your data

- Chose a program
  - Minimac3
  - IMPUTE2
  - Beagle
  - Never use PLINK

Imputation program popularity

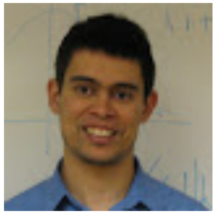


- Similar accuracy, features,

time frame

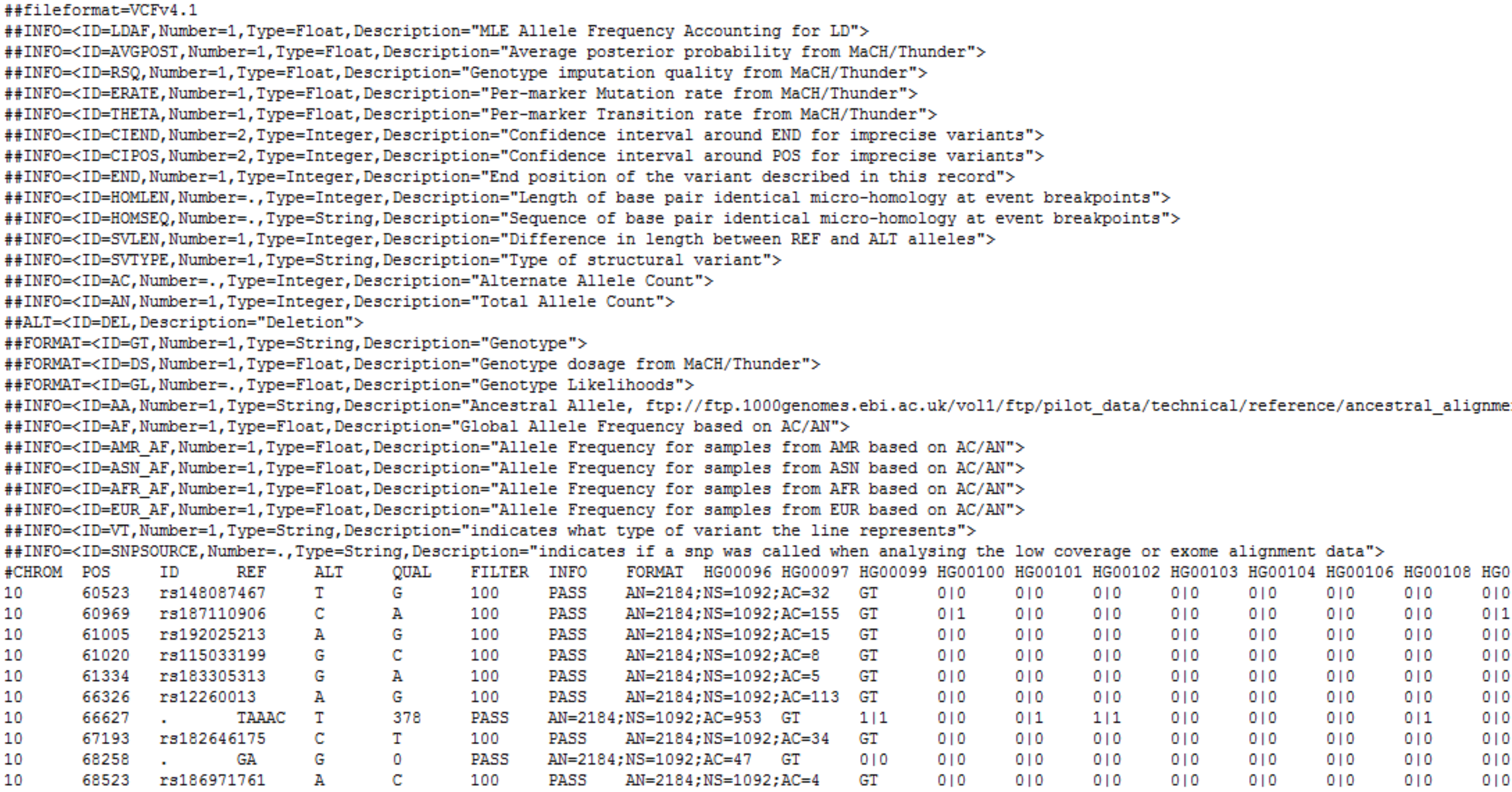
- Different output formats & downstream analysis options

My recommendation



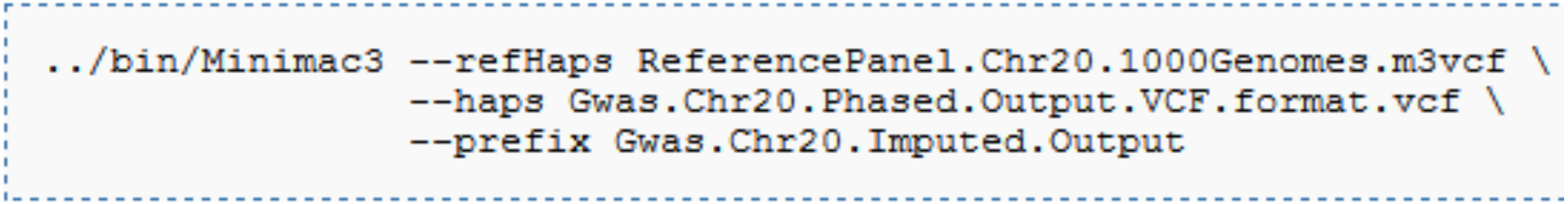
- MiniMac3
  - lower memory and more computationally efficient implementation
  - References are in a custom format (m3vcf) that can handle very large references with lower memory
  - Can read in the SHAPEIT2 references
  - Output is vcf format
  - Includes both SNP and individuals IDs – safest format to avoid errors
  - Downstream analysis with RAREMETALWORKER or other vcf input tools

vcf format



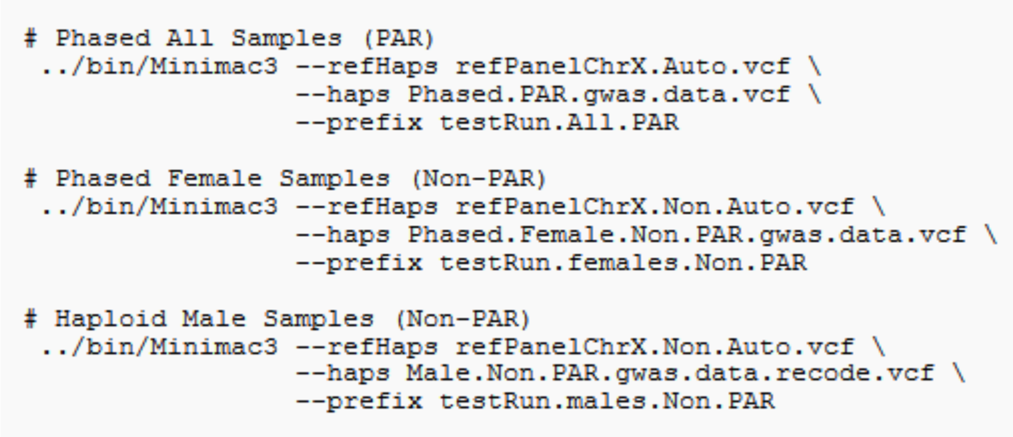
```
##fileformat=VCFv4.1
##INFO=<ID=LDAF,Number=1,Type=Float,Description="MLE Allele Frequency Accounting for LD">
##INFO=<ID=AVGPOST,Number=1,Type=Float,Description="Average posterior probability from MaCH/Thunder">
##INFO=<ID=RSQ,Number=1,Type=Float,Description="Genotype imputation quality from MaCH/Thunder">
##INFO=<ID=ERATE,Number=1,Type=Float,Description="Per-marker Mutation rate from MaCH/Thunder">
##INFO=<ID=THETA,Number=1,Type=Float,Description="Per-marker Transition rate from MaCH/Thunder">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate Allele Count">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total Allele Count">
##ALT=<ID=DEL,Description="Deletion">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage from MaCH/Thunder">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignme
##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele Frequency based on AC/AN">
##INFO=<ID=AMR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AMR based on AC/AN">
##INFO=<ID=ASN_AF,Number=1,Type=Float,Description="Allele Frequency for samples from ASN based on AC/AN">
##INFO=<ID=AFR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AFR based on AC/AN">
##INFO=<ID=EUR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from EUR based on AC/AN">
##INFO=<ID=VT,Number=1,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=SNPSOURCE,Number=.,Type=String,Description="indicates if a snp was called when analysing the low coverage or exome alignment data">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG00104 HG00106 HG00108 HG0
10      60523   rs148087467     T       G       100     PASS    AN=2184;NS=1092;AC=32   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      60969   rs187110906     C       A       100     PASS    AN=2184;NS=1092;AC=155  GT      0|1     0|0     0|0     0|0     0|0     0|0     0|0     0|1
10      61005   rs192025213     A       G       100     PASS    AN=2184;NS=1092;AC=15   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      61020   rs115033199     G       C       100     PASS    AN=2184;NS=1092;AC=8    GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      61334   rs183305313     G       A       100     PASS    AN=2184;NS=1092;AC=5    GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      66326   rs12260013      A       G       100     PASS    AN=2184;NS=1092;AC=113  GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      66627   .       TAAAC   T       378     PASS    AN=2184;NS=1092;AC=953  GT      1|1     0|0     0|1     1|1     0|0     0|0     0|0     0|1     0|0
10      67193   rs182646175     C       T       100     PASS    AN=2184;NS=1092;AC=34   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      68258   .       GA      G       0       PASS    AN=2184;NS=1092;AC=47   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      68523   rs186971761     A       C       100     PASS    AN=2184;NS=1092;AC=4    GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
```

Imputing in minimac3

- 

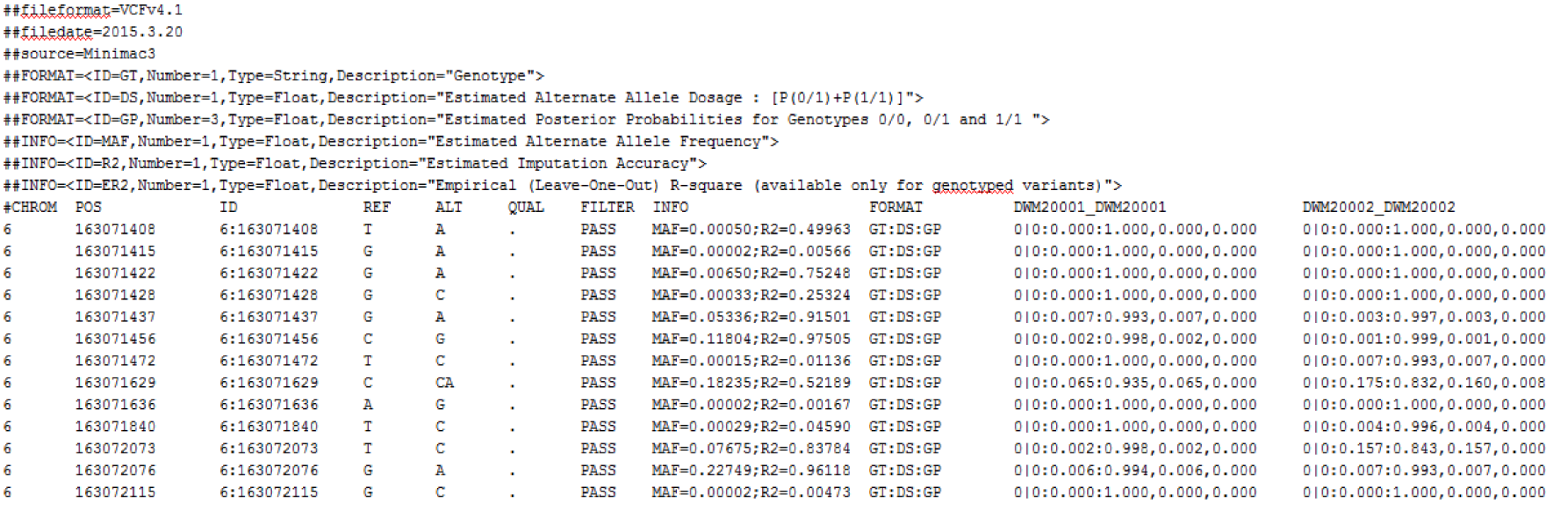
```
../bin/Minimac3 --refHaps ReferencePanel.Chr20.1000Genomes.m3vcf \
                --haps Gwas.Chr20.Phased.Output.VCF.format.vcf \
                --prefix Gwas.Chr20.Imputed.Output
```

- Can impute X
  - Impute Males & Females together for the pseudo Autosomal region (PAR)
  - Separately for the non-PAR



```
# Phased All Samples (PAR)
 ../bin/Minimac3 --refHaps refPanelChrX.Auto.vcf \
                 --haps Phased.PAR.gwas.data.vcf \
                 --prefix testRun.All.PAR

# Phased Female Samples (Non-PAR)
 ../bin/Minimac3 --refHaps refPanelChrX.Non.Auto.vcf \
                 --haps Phased.Female.Non.PAR.gwas.data.vcf \
                 --prefix testRun.females.Non.PAR

# Haploid Male Samples (Non-PAR)
 ../bin/Minimac3 --refHaps refPanelChrX.Non.Auto.vcf \
                 --haps Male.Non.PAR.gwas.data.recode.vcf \
                 --prefix testRun.males.Non.PAR
```

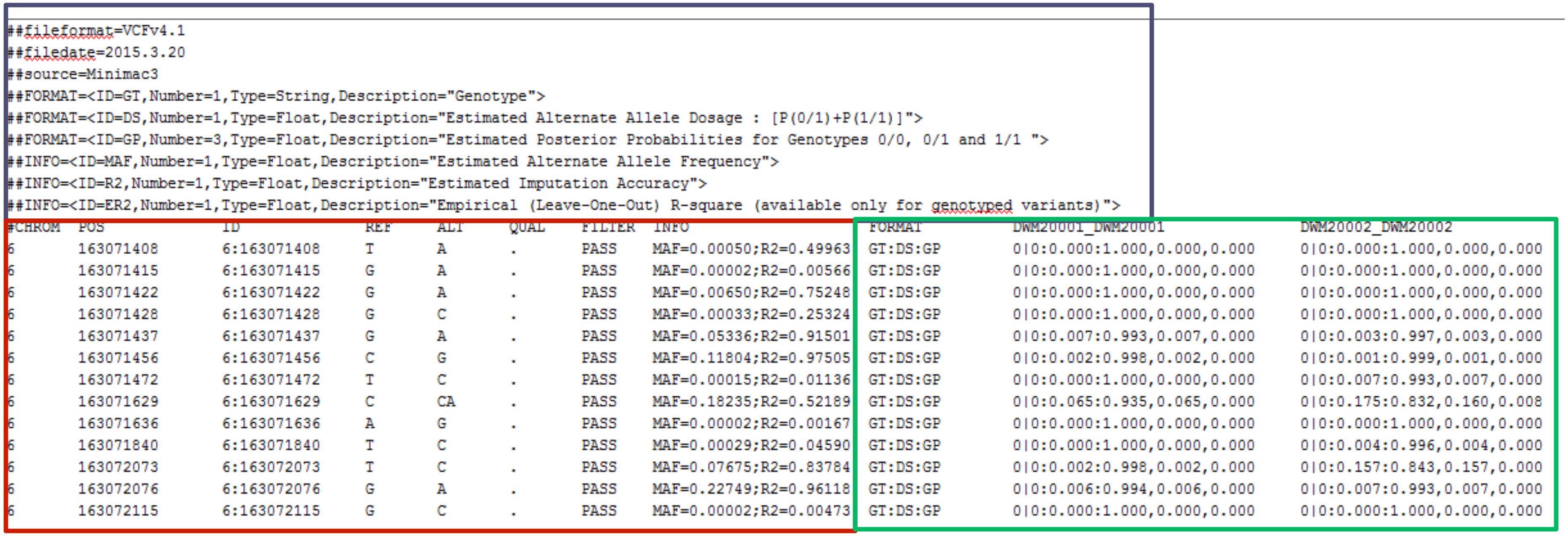
Output



```
##fileformat=VCFv4.1
##filedate=2015.3.20
##source=Minimac3
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1 ">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
#CHROM  POS        ID           REF  ALT  QUAL  FILTER  INFO                     FORMAT    DWM20001_DWM20001                DWM20002_DWM20002
6       163071408  6:163071408  T    A    .     PASS    MAF=0.00050;R2=0.49963   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.000:1.000,0.000,0.000
6       163071415  6:163071415  G    A    .     PASS    MAF=0.00002;R2=0.00566   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.000:1.000,0.000,0.000
6       163071422  6:163071422  G    A    .     PASS    MAF=0.00650;R2=0.75248   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.000:1.000,0.000,0.000
6       163071428  6:163071428  G    C    .     PASS    MAF=0.00033;R2=0.25324   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.000:1.000,0.000,0.000
6       163071437  6:163071437  G    A    .     PASS    MAF=0.05336;R2=0.91501   GT:DS:GP  0|0:0.007:0.993,0.007,0.000      0|0:0.003:0.997,0.003,0.000
6       163071456  6:163071456  C    G    .     PASS    MAF=0.11804;R2=0.97505   GT:DS:GP  0|0:0.002:0.998,0.002,0.000      0|0:0.001:0.999,0.001,0.000
6       163071472  6:163071472  T    C    .     PASS    MAF=0.00015;R2=0.01136   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.007:0.993,0.007,0.000
6       163071629  6:163071629  C    CA   .     PASS    MAF=0.18235;R2=0.52189   GT:DS:GP  0|0:0.065:0.935,0.065,0.000      0|0:0.175:0.832,0.160,0.008
6       163071636  6:163071636  A    G    .     PASS    MAF=0.00002;R2=0.00167   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.000:1.000,0.000,0.000
6       163071840  6:163071840  T    C    .     PASS    MAF=0.00029;R2=0.04590   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.004:0.996,0.004,0.000
6       163072073  6:163072073  T    C    .     PASS    MAF=0.07675;R2=0.83784   GT:DS:GP  0|0:0.002:0.998,0.002,0.000      0|0:0.157:0.843,0.157,0.000
6       163072076  6:163072076  G    A    .     PASS    MAF=0.22749;R2=0.96118   GT:DS:GP  0|0:0.006:0.994,0.006,0.000      0|0:0.007:0.993,0.007,0.000
6       163072115  6:163072115  G    C    .     PASS    MAF=0.00002;R2=0.00473   GT:DS:GP  0|0:0.000:1.000,0.000,0.000      0|0:0.000:1.000,0.000,0.000
```

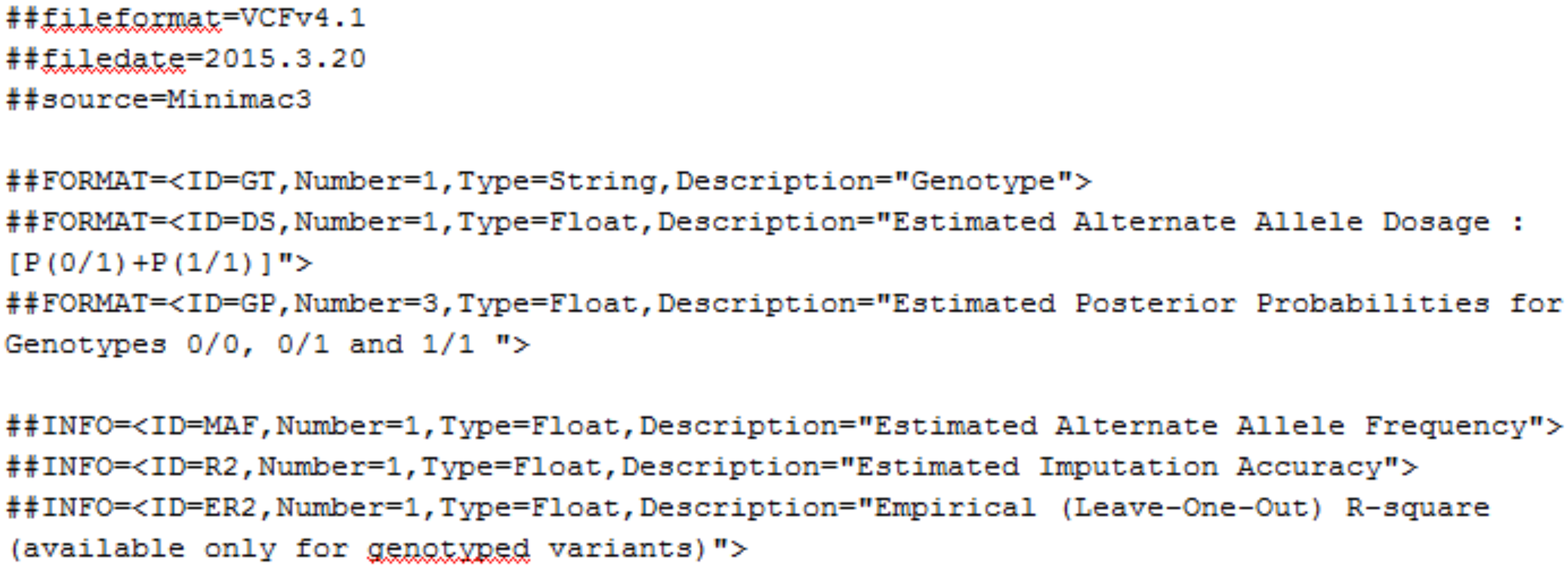
- Comments, info and genotypes in the 1 file
- 1 line per variant
- 1 column per person

Output



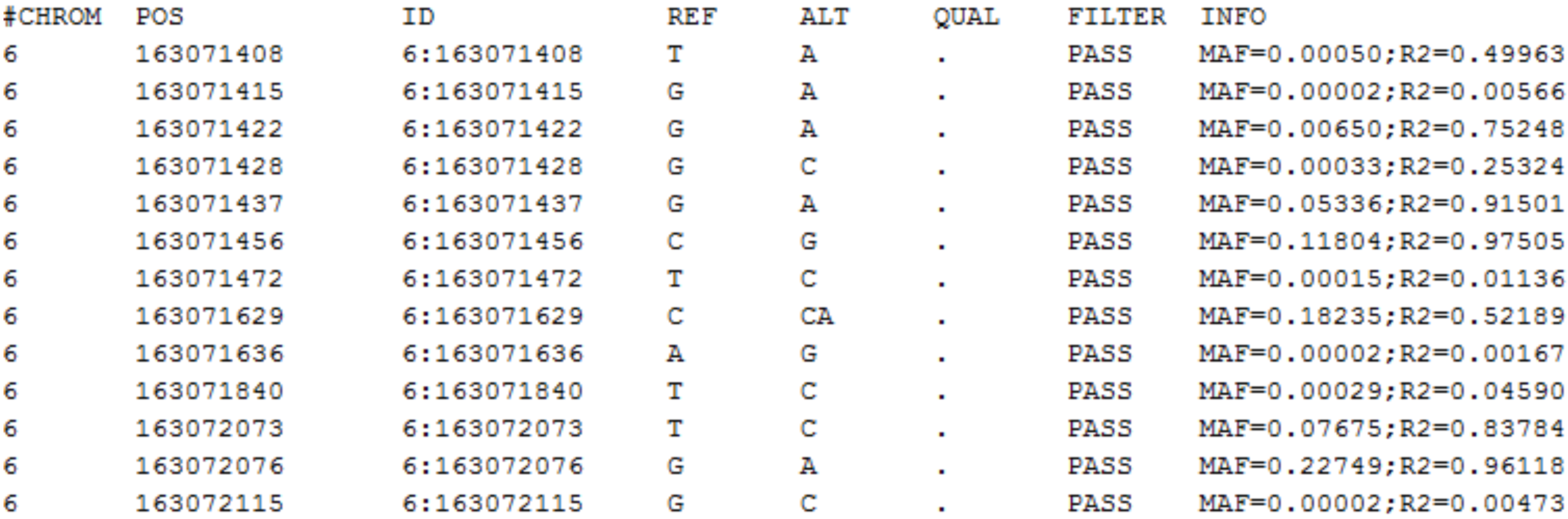
```
##fileformat=VCFv4.1
##filedate=2015.3.20
##source=Minimac3
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1 ">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
#CHROM  POS        ID           REF  ALT  QUAL  FILTER  INFO                       FORMAT    DWM20001_DWM20001                 DWM20002_DWM20002
6       163071408  6:163071408  T    A    .     PASS    MAF=0.00050;R2=0.49963     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.000:1.000,0.000,0.000
6       163071415  6:163071415  G    A    .     PASS    MAF=0.00002;R2=0.00566     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.000:1.000,0.000,0.000
6       163071422  6:163071422  G    A    .     PASS    MAF=0.00650;R2=0.75248     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.000:1.000,0.000,0.000
6       163071428  6:163071428  G    C    .     PASS    MAF=0.00033;R2=0.25324     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.000:1.000,0.000,0.000
6       163071437  6:163071437  G    A    .     PASS    MAF=0.05336;R2=0.91501     GT:DS:GP  0|0:0.007:0.993,0.007,0.000       0|0:0.003:0.997,0.003,0.000
6       163071456  6:163071456  C    G    .     PASS    MAF=0.11804;R2=0.97505     GT:DS:GP  0|0:0.002:0.998,0.002,0.000       0|0:0.001:0.999,0.001,0.000
6       163071472  6:163071472  T    C    .     PASS    MAF=0.00015;R2=0.01136     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.007:0.993,0.007,0.000
6       163071629  6:163071629  C    CA   .     PASS    MAF=0.18235;R2=0.52189     GT:DS:GP  0|0:0.065:0.935,0.065,0.000       0|0:0.175:0.832,0.160,0.008
6       163071636  6:163071636  A    G    .     PASS    MAF=0.00002;R2=0.00167     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.000:1.000,0.000,0.000
6       163071840  6:163071840  T    C    .     PASS    MAF=0.00029;R2=0.04590     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.004:0.996,0.004,0.000
6       163072073  6:163072073  T    C    .     PASS    MAF=0.07675;R2=0.83784     GT:DS:GP  0|0:0.002:0.998,0.002,0.000       0|0:0.157:0.843,0.157,0.000
6       163072076  6:163072076  G    A    .     PASS    MAF=0.22749;R2=0.96118     GT:DS:GP  0|0:0.006:0.994,0.006,0.000       0|0:0.007:0.993,0.007,0.000
6       163072115  6:163072115  G    C    .     PASS    MAF=0.00002;R2=0.00473     GT:DS:GP  0|0:0.000:1.000,0.000,0.000       0|0:0.000:1.000,0.000,0.000
```

The comments



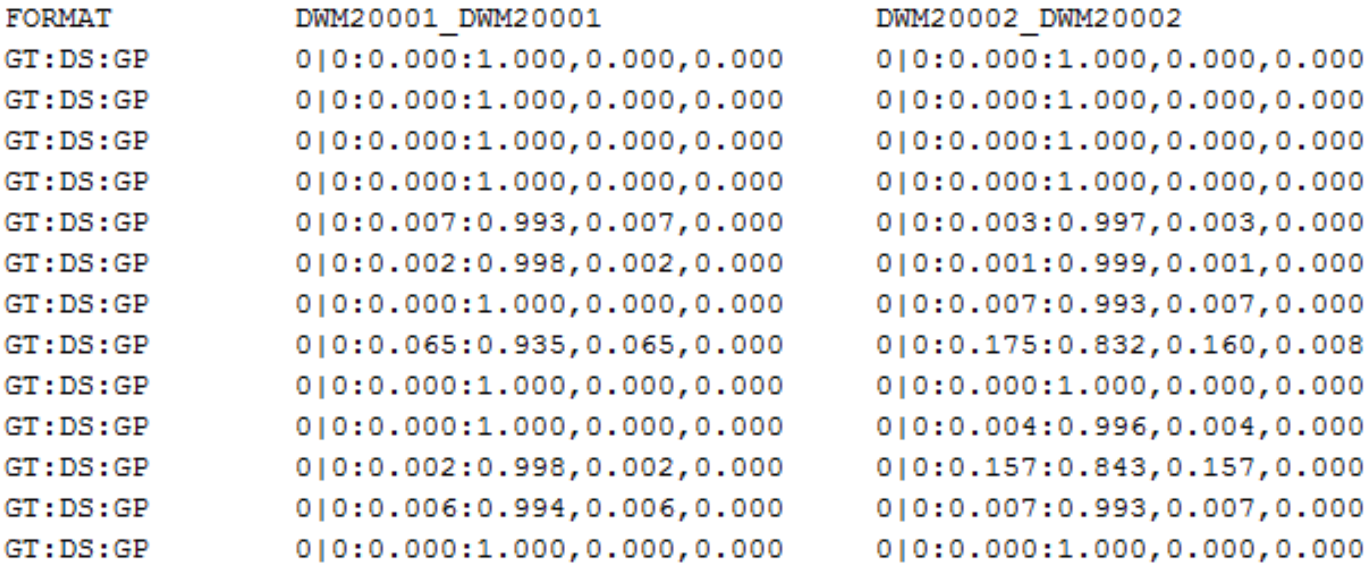
```
##fileformat=VCFv4.1
##filedate=2015.3.20
##source=Minimac3

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage :
[P(0/1)+P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for
Genotypes 0/0, 0/1 and 1/1 ">

##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square
(available only for genotyped variants)">
```

- The info



```
#CHROM  POS         ID           REF   ALT   QUAL   FILTER  INFO
6       163071408   6:163071408  T     A     .      PASS    MAF=0.00050;R2=0.49963
6       163071415   6:163071415  G     A     .      PASS    MAF=0.00002;R2=0.00566
6       163071422   6:163071422  G     A     .      PASS    MAF=0.00650;R2=0.75248
6       163071428   6:163071428  G     C     .      PASS    MAF=0.00033;R2=0.25324
6       163071437   6:163071437  G     A     .      PASS    MAF=0.05336;R2=0.91501
6       163071456   6:163071456  C     G     .      PASS    MAF=0.11804;R2=0.97505
6       163071472   6:163071472  T     C     .      PASS    MAF=0.00015;R2=0.01136
6       163071629   6:163071629  C     CA    .      PASS    MAF=0.18235;R2=0.52189
6       163071636   6:163071636  A     G     .      PASS    MAF=0.00002;R2=0.00167
6       163071840   6:163071840  T     C     .      PASS    MAF=0.00029;R2=0.04590
6       163072073   6:163072073  T     C     .      PASS    MAF=0.07675;R2=0.83784
6       163072076   6:163072076  G     A     .      PASS    MAF=0.22749;R2=0.96118
6       163072115   6:163072115  G     C     .      PASS    MAF=0.00002;R2=0.00473
```

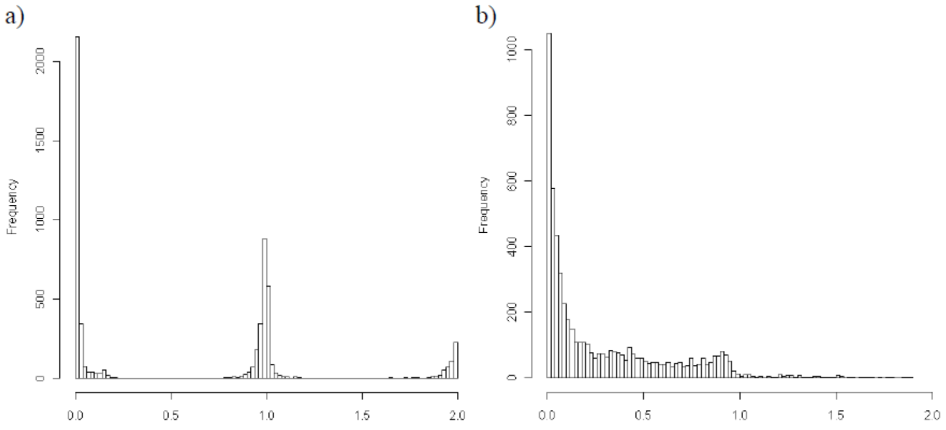
- The genotypes



```
FORMAT          DWM20001_DWM20001               DWM20002_DWM20002
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.000:1.000,0.000,0.000
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.000:1.000,0.000,0.000
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.000:1.000,0.000,0.000
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.000:1.000,0.000,0.000
GT:DS:GP        0|0:0.007:0.993,0.007,0.000     0|0:0.003:0.997,0.003,0.000
GT:DS:GP        0|0:0.002:0.998,0.002,0.000     0|0:0.001:0.999,0.001,0.000
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.007:0.993,0.007,0.000
GT:DS:GP        0|0:0.065:0.935,0.065,0.000     0|0:0.175:0.832,0.160,0.008
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.000:1.000,0.000,0.000
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.004:0.996,0.004,0.000
GT:DS:GP        0|0:0.002:0.998,0.002,0.000     0|0:0.157:0.843,0.157,0.000
GT:DS:GP        0|0:0.006:0.994,0.006,0.000     0|0:0.007:0.993,0.007,0.000
GT:DS:GP        0|0:0.000:1.000,0.000,0.000     0|0:0.000:1.000,0.000,0.000
```

Analyses…

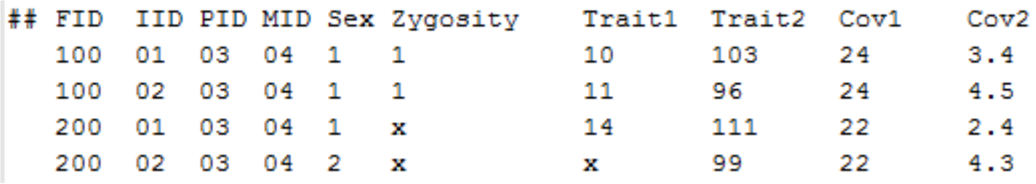
- DO NOT ANALYSE HARDCALL GENOTYPES!!!!!!
- Analyse the dosage or probabilities as this will account for the imputation uncertainty



Analyses in RAREMETALWORKER

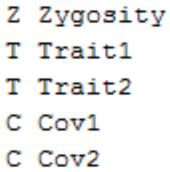
- Simple phenotype file formats
  - Can account for relatedness & twins
  - Can use GRM to account for relatedness (memory+++)
  - Ped file

(no header)



```
## FID   IID PID MID Sex Zygosity    Trait1   Trait2   Cov1     Cov2
   100   01  03  04  1   1           10       103      24       3.4
   100   02  03  04  1   1           11       96       24       4.5
   200   01  03  04  1   x           14       111      22       2.4
   200   02  03  04  2   x           x        99       22       4.3
```

  - Dat file



```
Z Zygosity
T Trait1
T Trait2
C Cov1
C Cov2
```

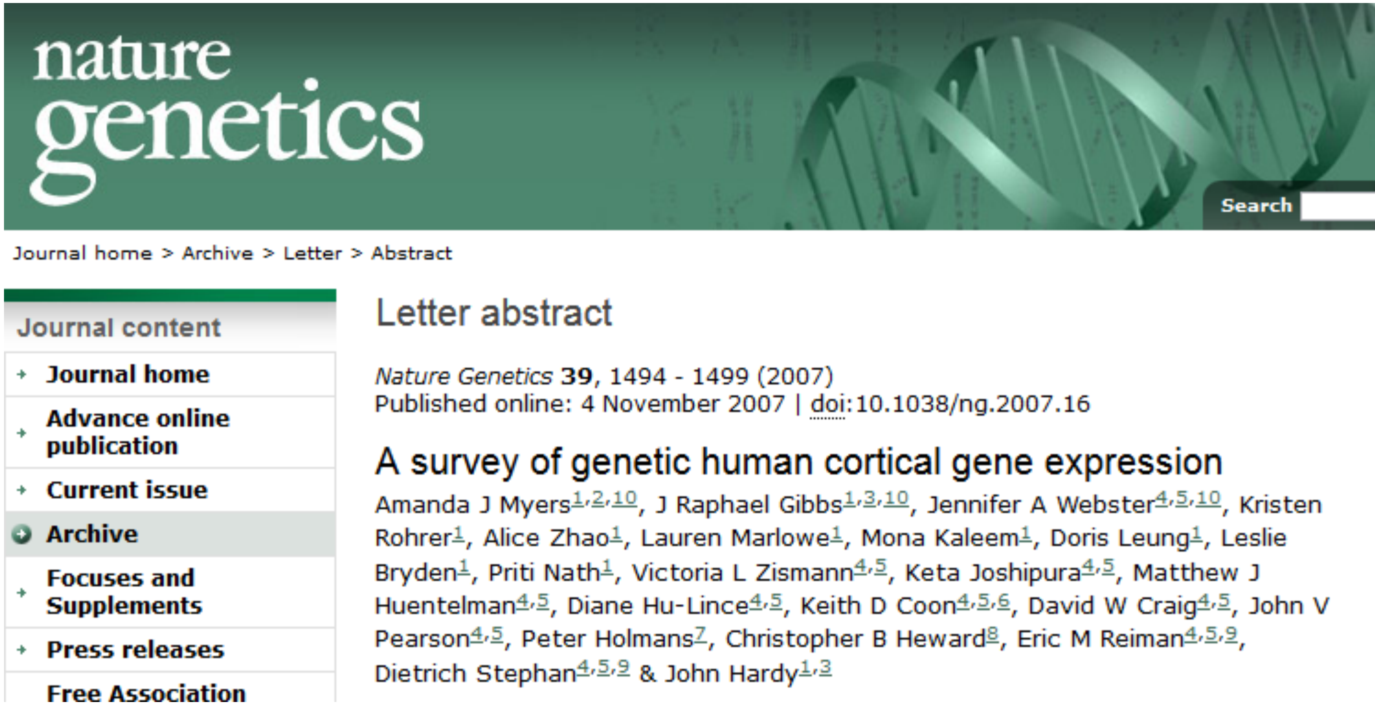
- raremetalworker --ped your.ped --dat your.dat --vcf your.vcf.gz --prefix example
- raremetalworker --ped your.ped --dat your.dat --vcf your.vcf.gz --kinPedigree --prefix example

Files to practice with

http://genome.sph.umich.edu/wiki/Minimac3_Imputation_Cookbook

- But really and truly consider using the Imputation Servers so that you can access the HRC references!
  - https://imputationserver.sph.umich.edu/

A practical example



Letter abstract

Nature Genetics **39**, 1494 - 1499 (2007)
Published online: 4 November 2007 | doi:10.1038/ng.2007.16

A survey of genetic human cortical gene expression

Amanda J Myers[1,2,10], J Raphael Gibbs[1,3,10], Jennifer A Webster[4,5,10], Kristen Rohrer[1], Alice Zhao[1], Lauren Marlowe[1], Mona Kaleem[1], Doris Leung[1], Leslie Bryden[1], Priti Nath[1], Victoria L Zismann[4,5], Keta Joshipura[4,5], Matthew J Huentelman[4,5], Diane Hu-Lince[4,5], Keith D Coon[4,5,6], David W Craig[4,5], John V Pearson[4,5], Peter Holmans[7], Christopher B Heward[8], Eric M Reiman[4,5,9], Dietrich Stephan[4,5,9] & John Hardy[1,3]

- http://labs.med.miami.edu/myers/LFuN/LFuN.html
- post-mortem gene expression in ‘brain’ tissue
- N=193

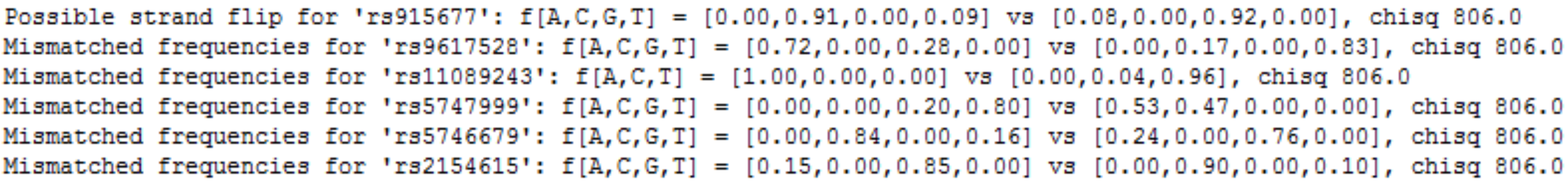
Imputation

- Chromosome 22 only – HapMapII- b36r22
- MaCH phasing
  - (In real life with a sample this size include the reference in the phasing)
- Minimac Imputation

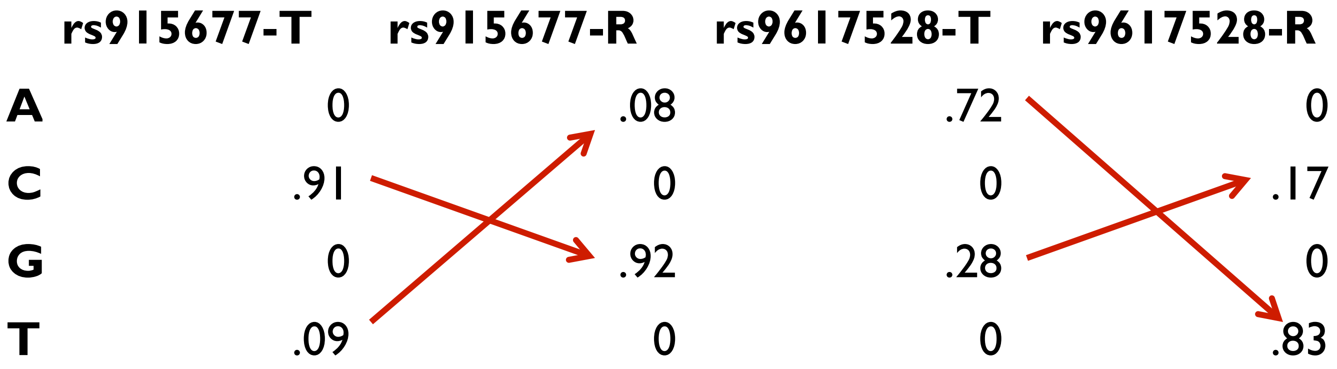
- Run twice
  - Once without stand alignment (badImp)
  - Once with strand alignment (goodImp)

How do we know there was no strand alignment from the output?

- No way of telling from the phasing log
  - B/c we didn’t include a reference
- Imputation log is FULL of errors



```
Possible strand flip for 'rs915677': f[A,C,G,T] = [0.00,0.91,0.00,0.09] vs [0.08,0.00,0.92,0.00], chisq 806.0
Mismatched frequencies for 'rs9617528': f[A,C,G,T] = [0.72,0.00,0.28,0.00] vs [0.00,0.17,0.00,0.83], chisq 806.0
Mismatched frequencies for 'rs11089243': f[A,C,T] = [1.00,0.00,0.00] vs [0.00,0.04,0.96], chisq 806.0
Mismatched frequencies for 'rs5747999': f[A,C,G,T] = [0.00,0.00,0.20,0.80] vs [0.53,0.47,0.00,0.00], chisq 806.0
Mismatched frequencies for 'rs5746679': f[A,C,G,T] = [0.00,0.84,0.00,0.16] vs [0.24,0.00,0.76,0.00], chisq 806.0
Mismatched frequencies for 'rs2154615': f[A,C,G,T] = [0.15,0.00,0.85,0.00] vs [0.00,0.90,0.00,0.10], chisq 806.0
```



	rs915677-T	rs915677-R	rs9617528-T	rs9617528-R
A	0	.08	.72	0
C	.91	0	0	.17
G	0	.92	.28	0
T	.09	0	0	.83

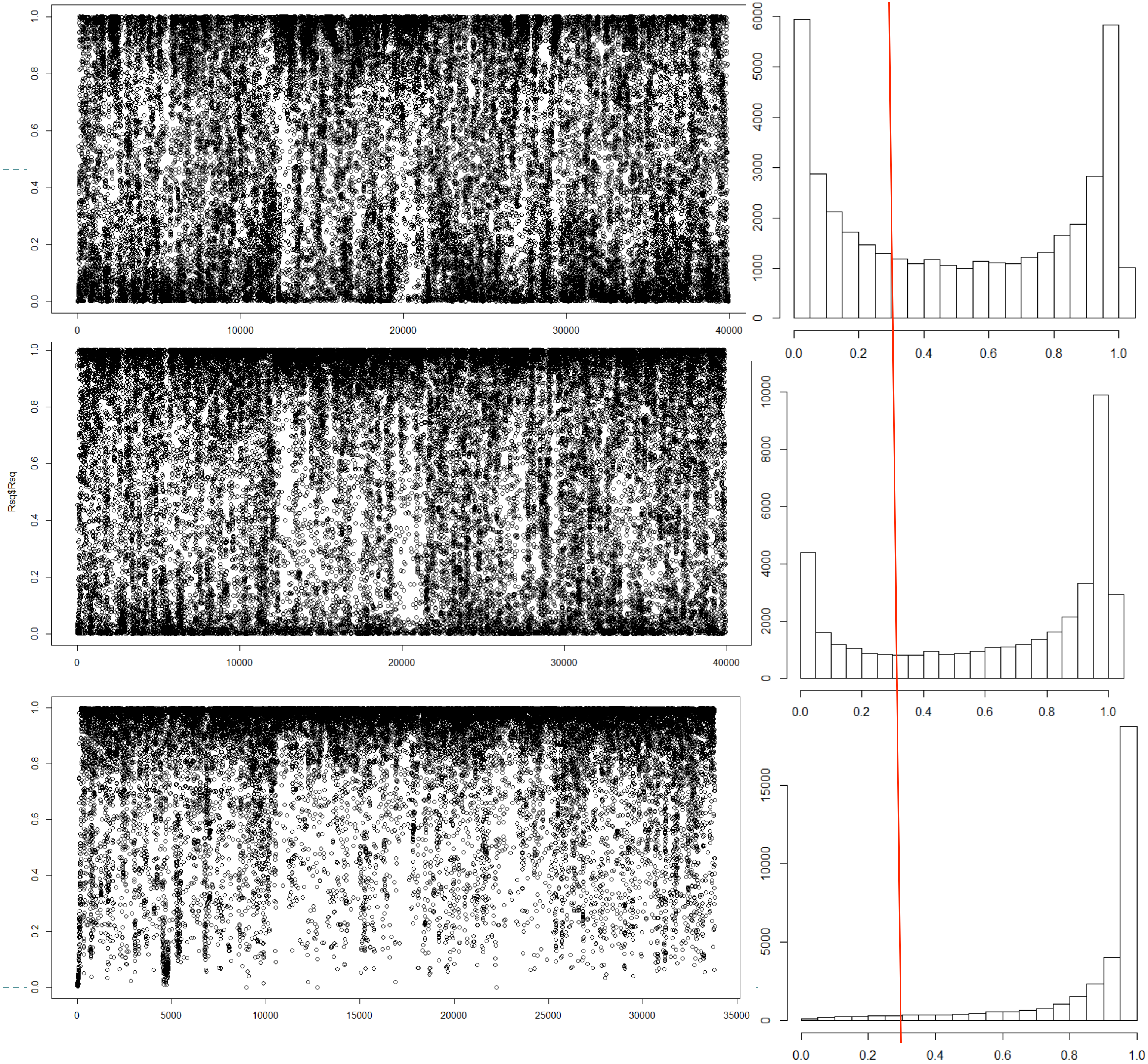
Plot the r2 for the 2 imputation runs

- How do they compare?
- badImp 17,908/39905 with r2 >=.6
- goodImp 24,685/39905 with r2 >=.6
  - still quite bad b/c of small N
  - Should have compensated by including ref data in the phasing step
- In a QIMR dataset N=19k 32296/33815

Bad Imputation

Better Imputation

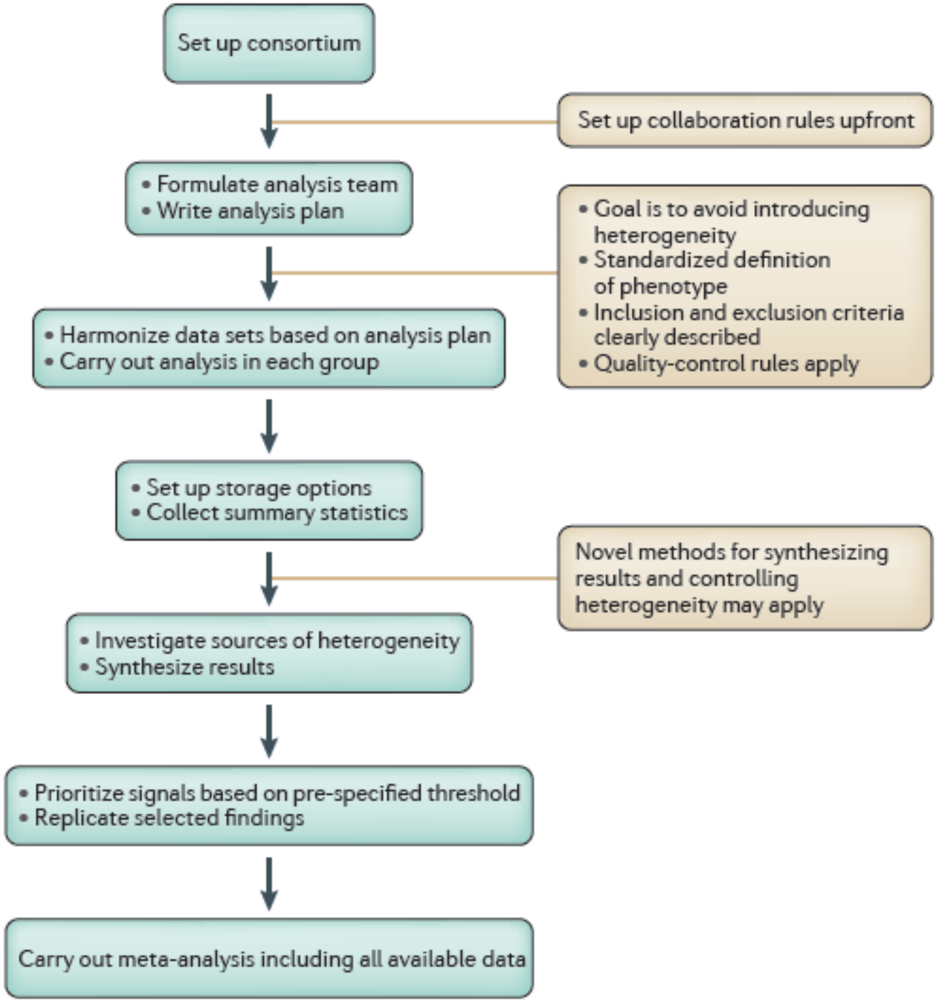
Good Imputation



Meta-analysis

Setting up a Meta-analysis

- Managing the personal and social connections is extremely important
- meta-analyses are usually unfunded
  - Time line is too short and budget is too small for a grant
- Meta-analyses do not work top down – to be successful they MUST be led by analysts who know what they are doing


Set up consortium
Set up collaboration rules upfront
• Formulate analysis team
• Write analysis plan
• Goal is to avoid introducing heterogeneity
• Standardized definition of phenotype
• Inclusion and exclusion criteria clearly described
• Quality-control rules apply
• Harmonize data sets based on analysis plan
• Carry out analysis in each group
• Set up storage options
• Collect summary statistics
Novel methods for synthesizing results and controlling heterogeneity may apply
• Investigate sources of heterogeneity
• Synthesize results
• Prioritize signals based on pre-specified threshold
• Replicate selected findings
Carry out meta-analysis including all available data

Evangelou, E. 2013. *Nat Rev Genet* 14 379-389.

Approaches to GWAS meta-analysis

- Fixed effects
  - Most common - most powerful approach for discovery under the model that the true effect of each risk allele is the same in each data set
    - Inverse variance weighted most common
    - N weighted also common
- Random effects
  - Uncommon - more appropriate when the aim is to consider the generalizability of the observed association and estimate the average effect size of the associated variant and its uncertainty across different populations
- Bayesian
  - Very uncommon – mainly MAs from the Welcome Trust

Quality control of data going into MA is critical!

- Exclude rare variants
  - Typically 1% or .5% MAF with large samples (5000+) can consider going lower
- Exclude poorly imputed variants
  - Imputation accuracy metric depends on the software used
    - Mach/minimac r^2
    - IMPUTE properinfo/info
    - BEAGLE ovarimp
  - Typically calculated as observed variance/expected – can empirically go over 1 usually capped at 1
  - Threshold .6
- Plot
  - QQ, Manhattan, SE vs N, SE vs MAF, SE vs Rsq, P vs Z

GWAS-MA

- Most commonly used software for common variant analysis – METAL
  - Automatic strand flipping of non-ambiguous SNPs
  - Calculation of max min mean allele frequency
  - Inverse variance & N weightings
  - Automatic genomic control correction
  - Herogeniety tests

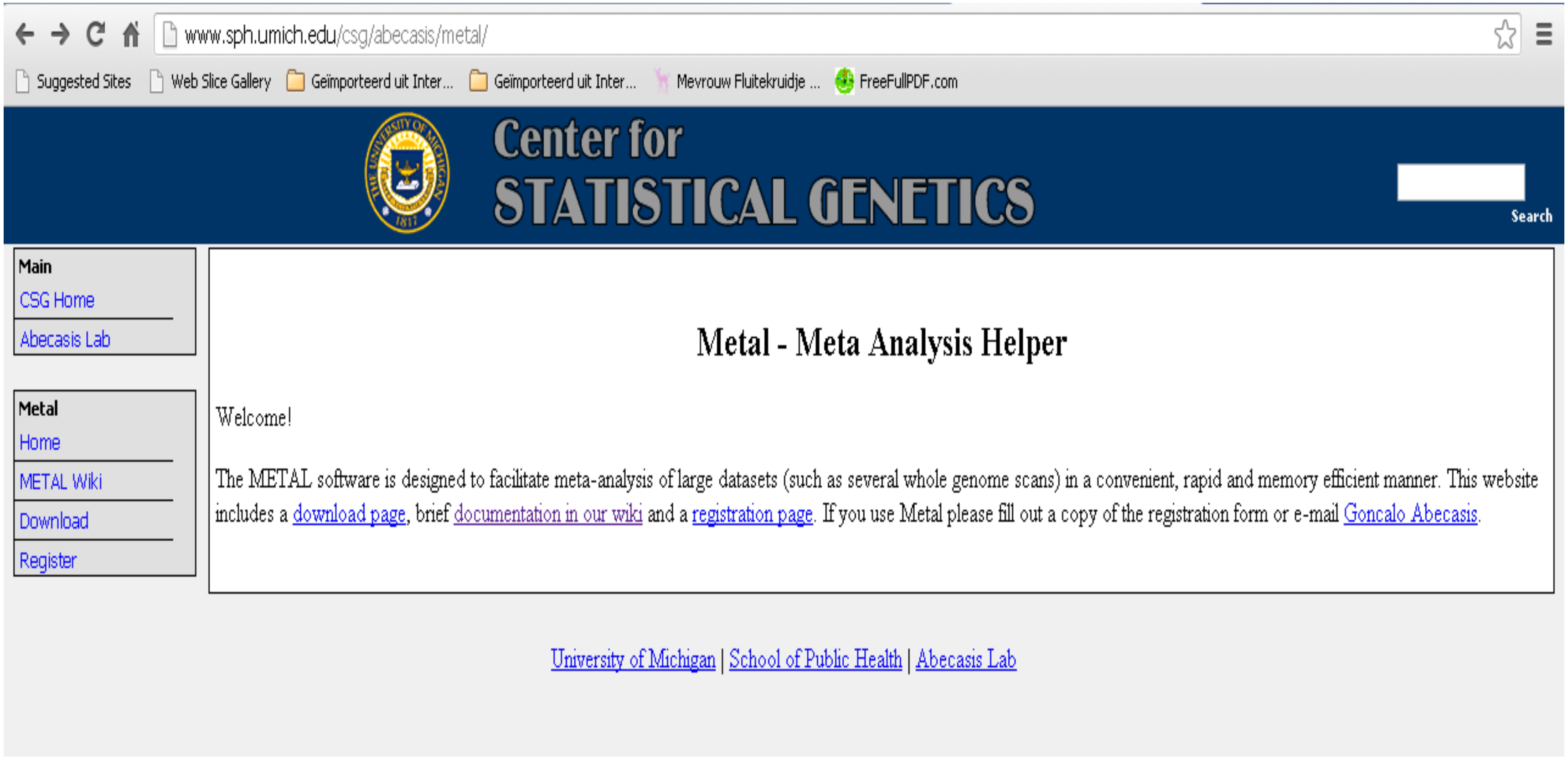
- Most commonly used software for rare variant analysis - RAREMETAL

METAL

http://www.sph.umich.edu/csg/abecasis/metal/

Documentation can be found at the metal wiki:

http://genome.sph.umich.edu/wiki/Metal_Documentation



Metal - Meta Analysis Helper

Welcome!

The METAL software is designed to facilitate meta-analysis of large datasets (such as several whole genome scans) in a convenient, rapid and memory efficient manner. This website includes a download page, brief documentation in our wiki and a registration page. If you use Metal please fill out a copy of the registration form or e-mail Goncalo Abecasis.

University of Michigan | School of Public Health | Abecasis Lab

METAL

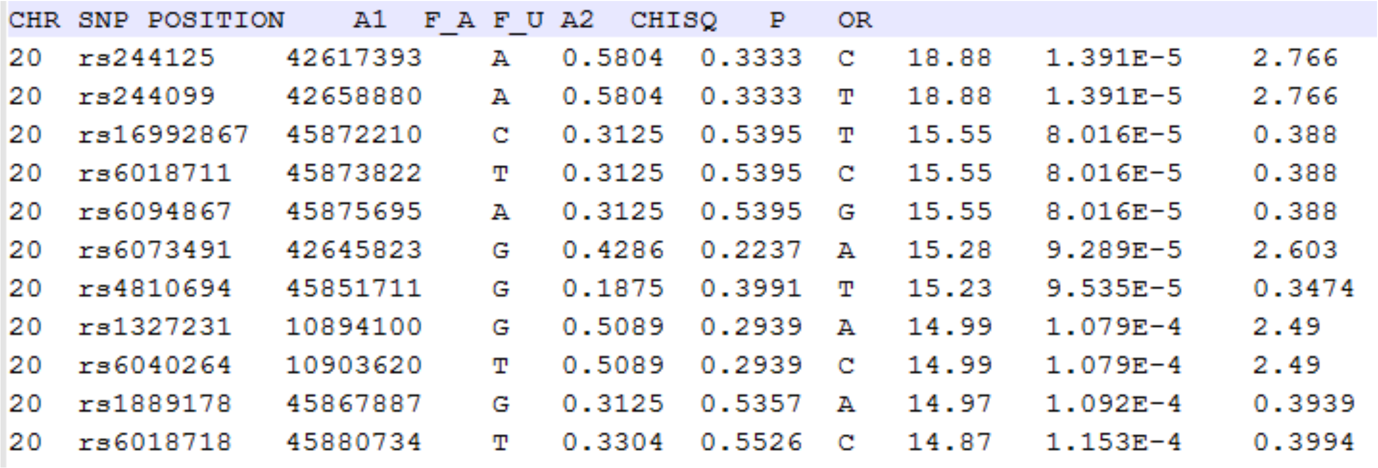
- Requires results files
- ‘Script’ file
  - Describes the input files
  - Defines meta-analysis strategy
  - Names output file

Steps

1. Check format of results files
   1. Ensure all necessary columns are available
   2. Modify files to include all information
2. Prepare script file
   1. Ensure headers match description
   2. Crosscheck each results file matches Process name
3. Run metal

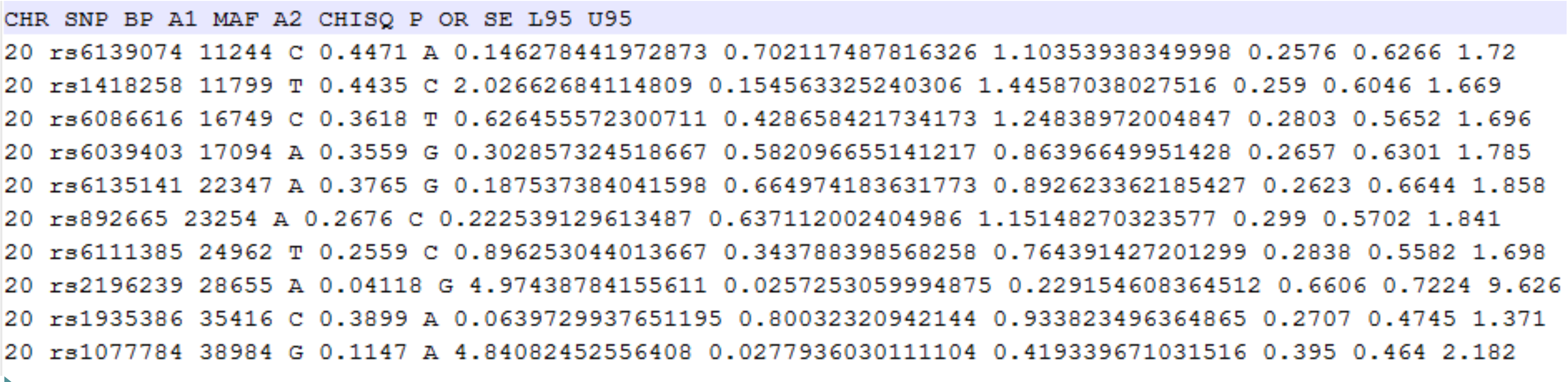
INPUT FILES

- Results1.txt



```
CHR SNP POSITION     A1  F_A F_U A2  CHISQ   P    OR
20  rs244125     42617393     A    0.5804  0.3333  C    18.88    1.391E-5     2.766
20  rs244099     42658880     A    0.5804  0.3333  T    18.88    1.391E-5     2.766
20  rs16992867   45872210     C    0.3125  0.5395  T    15.55    8.016E-5     0.388
20  rs6018711    45873822     T    0.3125  0.5395  C    15.55    8.016E-5     0.388
20  rs6094867    45875695     A    0.3125  0.5395  G    15.55    8.016E-5     0.388
20  rs6073491    42645823     G    0.4286  0.2237  A    15.28    9.289E-5     2.603
20  rs4810694    45851711     G    0.1875  0.3991  T    15.23    9.535E-5     0.3474
20  rs1327231    10894100     G    0.5089  0.2939  A    14.99    1.079E-4     2.49
20  rs6040264    10903620     T    0.5089  0.2939  C    14.99    1.079E-4     2.49
20  rs1889178    45867887     G    0.3125  0.5357  A    14.97    1.092E-4     0.3939
20  rs6018718    45880734     T    0.3304  0.5526  C    14.87    1.153E-4     0.3994
```

- Results2.txt



```
CHR SNP BP A1 MAF A2 CHISQ P OR SE L95 U95
20 rs6139074 11244 C 0.4471 A 0.146278441972873 0.702117487816326 1.10353938349998 0.2576 0.6266 1.72
20 rs1418258 11799 T 0.4435 C 2.02662684114809 0.154563325240306 1.44587038027516 0.259 0.6046 1.669
20 rs6086616 16749 C 0.3618 T 0.626455572300711 0.428658421734173 1.24838972004847 0.2803 0.5652 1.696
20 rs6039403 17094 A 0.3559 G 0.302857324518667 0.582096655141217 0.86396649951428 0.2657 0.6301 1.785
20 rs6135141 22347 A 0.3765 G 0.187537384041598 0.664974183631773 0.892623362185427 0.2623 0.6644 1.858
20 rs892665 23254 A 0.2676 C 0.222539129613487 0.637112002404986 1.15148270323577 0.299 0.5702 1.841
20 rs6111385 24962 T 0.2559 C 0.896253044013667 0.343788398568258 0.764391427201299 0.2838 0.5582 1.698
20 rs2196239 28655 A 0.04118 G 4.97438784155611 0.0257253059994875 0.229154608364512 0.6606 0.7224 9.626
20 rs1935386 35416 C 0.3899 A 0.0639729937651195 0.80032320942144 0.933823496364865 0.2707 0.4745 1.371
20 rs1077784 38984 G 0.1147 A 4.84082452556408 0.0277936030111104 0.419339671031516 0.395 0.464 2.182
```

Columns METAL uses

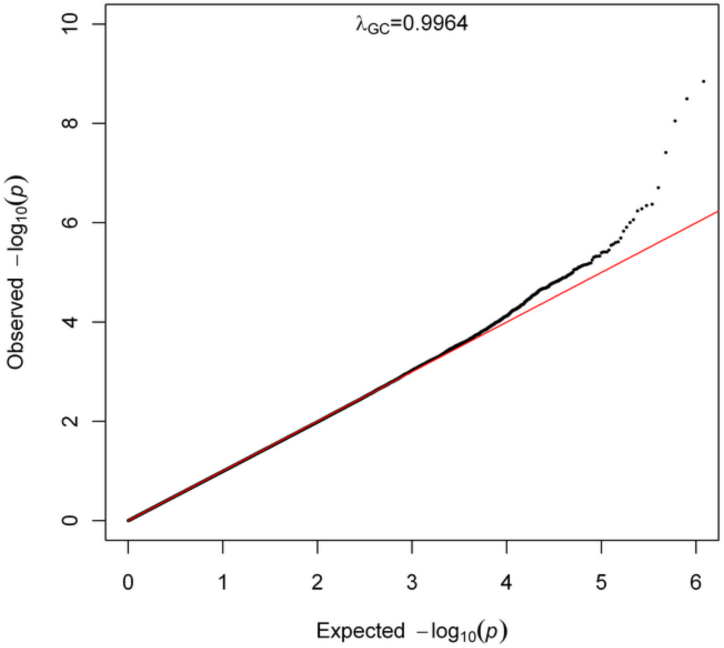
- SNP
- Effect allele & non-effect allele
- Frequency of effect allele
- OR/Beta
- SE [for standard error meta-analysis]
- P-value [for Z-score meta-analysis]
  - IMPORTANT – you can not use FDR controlled or adaptively permuted P values!
- N/weight column [for Z-score meta-analysis]

Effect allele

- Differs for different programs and analysis options
  - Minor/major allele
  - Alphabetical
  - 1st listed
- DO NOT ASSUME YOU KNOW ALWAYS DOUBLE CHECK!

Genomic control

- λ (lambda)
- Median test statistic/ expected median test stat
- Should be one



Strand Ambiguous SNPs

- When you get data from different studies is not always aligned the same way
- Remember A<>T & C<>G
- If a SNP is A/C or then the reverse strand is T/G
  - No ambiguity, regardless of strand we know which allele is which
  - A/G, T/C & T/G also non ambiguous
  - METAL can align you non ambiguous SNPs

Strand Ambiguous SNPs

- Remember A<>T & C<>G
- If a SNP is A/T then the reverse strand is T/A
  - AMBIGUOUS!!! Need to check allele freq to make sure samples are aligned
  - C/G SNPs are also ambiguous!
  - METAL can not align ambiguous SNPs

Meta-analysis running

- We will run meta-analysis based on effect size and on test statistic
- For the weights of test statistic, I've assumed that the sample sizes are the same
  - METAL defaults to weight of 1 when no weight column is supplied

Step 2: script file: meta_run_file

```
# PERFORM META-ANALYSIS based on effect size and on test statistic
# Loading in the input files with results from the  participating samples
# Note: Order of samples is …[sample size, alphabetic order,..]
# Phenotype is ..
# MB March 2013

MARKER  SNP
 ALLELE  A1 A2
 PVALUE  P
 EFFECT  log(OR)
 STDERR  SE                                  specifies column names

PROCESS results1.txt
PROCESS results2.txt                         processes two results files

OUTFILE meta_res_Z .txt                      Output file naming

ANALYZE                                      Conducts Z-based meta-analysis from test statistic
CLEAR                                        Clears workspace
SCHEME STDERR                                Changes meta-analysis scheme to beta + SE

PROCESS results1.txt
PROCESS results2.txt                         processes two results files

OUTFILE meta_res_SE .txt                     Output file naming
ANALYZE                                      Conducts effect size meta-analysis
```

Larger Consortia

```
# PERFORM META-ANALYSIS on P-values

module load metal

metal << EOT

# Loading in the inputfiles with results from the  participating samples
# Note: Order of samples is alpahabetic
# Phenotype is WB

# 1.AGES_HAP
MARKER SNPID
ALLELE coded_all noncoded_all
EFFECT Beta
PVALUE Pval
WEIGHT n_total
GENOMICCONTROL ON
COLUMNCOUNTING LENIENT
PROCESS AGES_HAP.txt

# 2.ALSPAC_HAP
MARKER SNPID
ALLELE coded_all noncoded_all
EFFECT Beta
PVALUE Pval
WEIGHT n_total
GENOMICCONTROL ON
COLUMNCOUNTING LENIENT
PROCESS ALSPAC_HAP.txt
```

AND SO ON (in this case 40 files)

Running metal

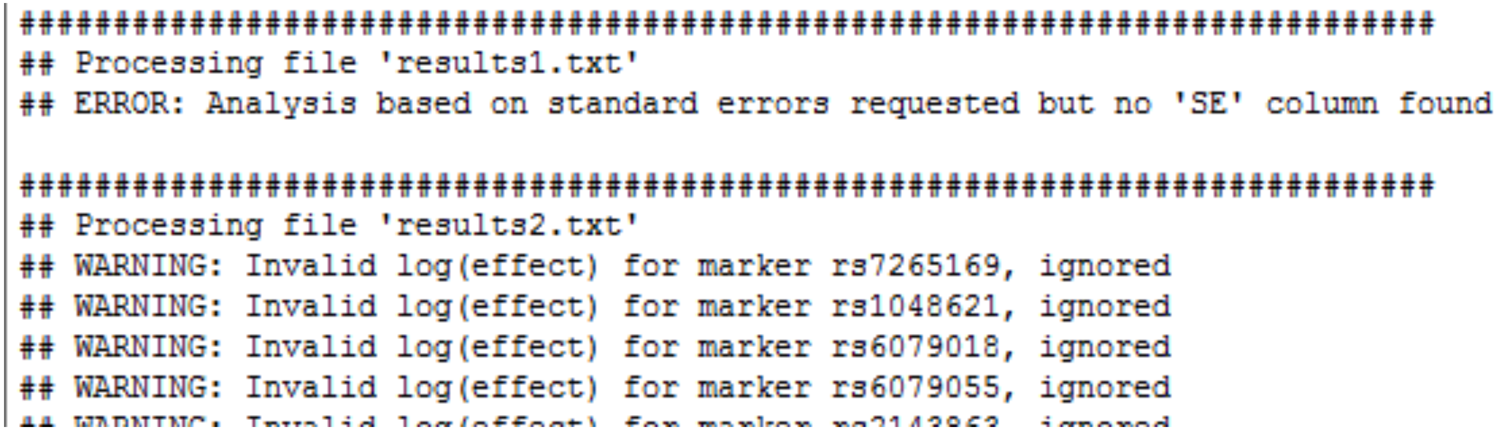
- metal < metal_run_file > metal_run.log

- metal is the command
- metal_run_file is the script file
- This will output information on the running of METAL things to standard out [the terminal]
- It will spawn 4 files:
  - 2 results files: meta_res_Z1.txt + meta_res_SE1.txt
  - 2 info files: meta_res_Z1.txt.info + meta_res_SE1.txt.info

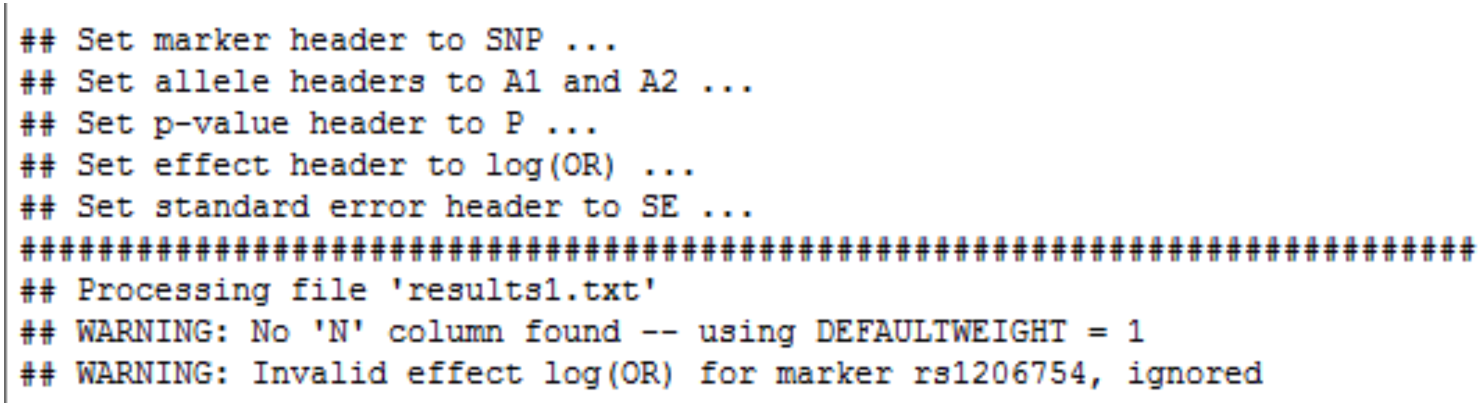
Output you'll see

- Overview of METAL commands
- Any errors
- And your best hit from meta-analysis

Common Errors

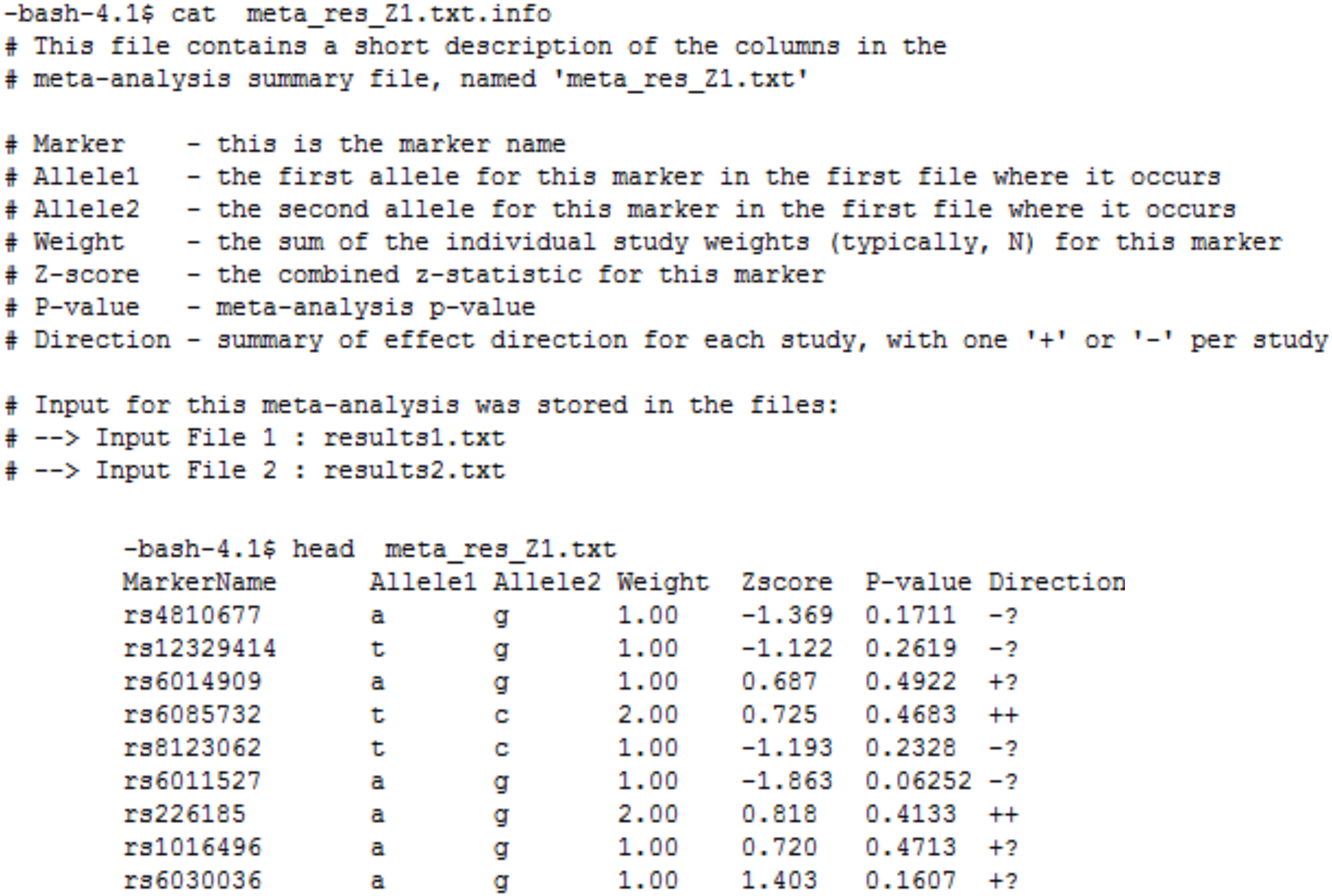


```
###########################################################################
## Processing file 'results1.txt'
## ERROR: Analysis based on standard errors requested but no 'SE' column found

###########################################################################
## Processing file 'results2.txt'
## WARNING: Invalid log(effect) for marker rs7265169, ignored
## WARNING: Invalid log(effect) for marker rs1048621, ignored
## WARNING: Invalid log(effect) for marker rs6079018, ignored
## WARNING: Invalid log(effect) for marker rs6079055, ignored
## WARNING: Invalid log(effect) for marker rs2143863, ignored
```



```
## Set marker header to SNP ...
## Set allele headers to A1 and A2 ...
## Set p-value header to P ...
## Set effect header to log(OR) ...
## Set standard error header to SE ...
###########################################################################
## Processing file 'results1.txt'
## WARNING: No 'N' column found -- using DEFAULTWEIGHT = 1
## WARNING: Invalid effect log(OR) for marker rs1206754, ignored
```

Output



```
-bash-4.1$ cat  meta_res_Z1.txt.info
# This file contains a short description of the columns in the
# meta-analysis summary file, named 'meta_res_Z1.txt'

# Marker    - this is the marker name
# Allele1   - the first allele for this marker in the first file where it occurs
# Allele2   - the second allele for this marker in the first file where it occurs
# Weight    - the sum of the individual study weights (typically, N) for this marker
# Z-score   - the combined z-statistic for this marker
# P-value   - meta-analysis p-value
# Direction - summary of effect direction for each study, with one '+' or '-' per study

# Input for this meta-analysis was stored in the files:
# --> Input File 1 : results1.txt
# --> Input File 2 : results2.txt
```

```
        -bash-4.1$ head  meta_res_Z1.txt
        MarkerName      Allele1 Allele2 Weight  Zscore  P-value Direction
        rs4810677       a       g       1.00    -1.369  0.1711  -?
        rs12329414      t       g       1.00    -1.122  0.2619  -?
        rs6014909       a       g       1.00    0.687   0.4922  +?
        rs6085732       t       c       2.00    0.725   0.4683  ++
        rs8123062       t       c       1.00    -1.193  0.2328  -?
        rs6011527       a       g       1.00    -1.863  0.06252 -?
        rs226185        a       g       2.00    0.818   0.4133  ++
        rs1016496       a       g       1.00    0.720   0.4713  +?
        rs6030036       a       g       1.00    1.403   0.1607  +?
```

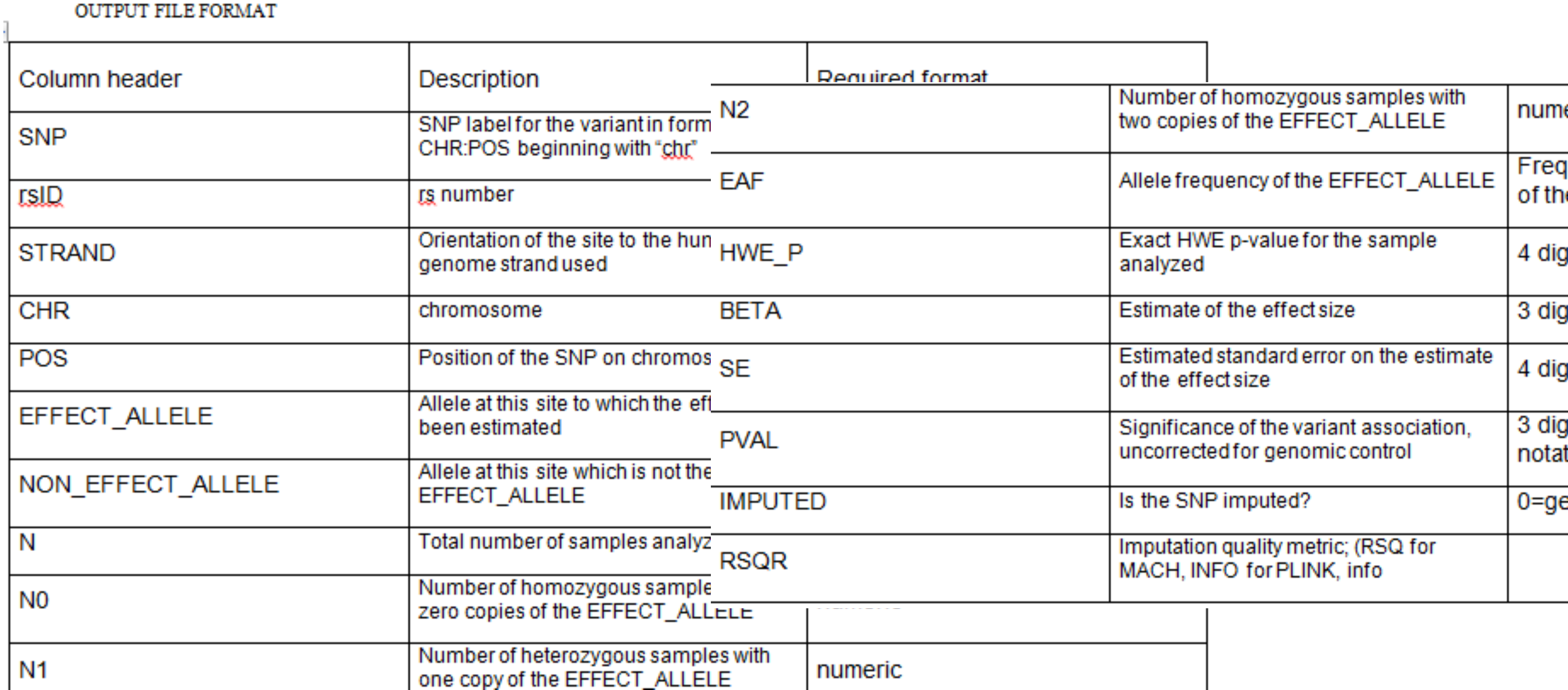
Important considerations for MA

- Duplicate QC sites
- Always check the input data
- Make sure you double check results
  - QQ plots
  - Manhattan plots
  - Allele frequencies etc
- Consider allowing cohorts to ignore variants with MAF <. 5% and low r2 – it will save you a lot of time and save a lot of storage space!

Don’t ask for stuff you don’t need

(Its annoying & adding extra columns*30M lines is a waste of space…)

- You need:
  - SNP, CHR:BP, EffectAllele, NonEffectAllele, EA_Freq, Ntotal, Beta, SE, P, Rsq
- Not

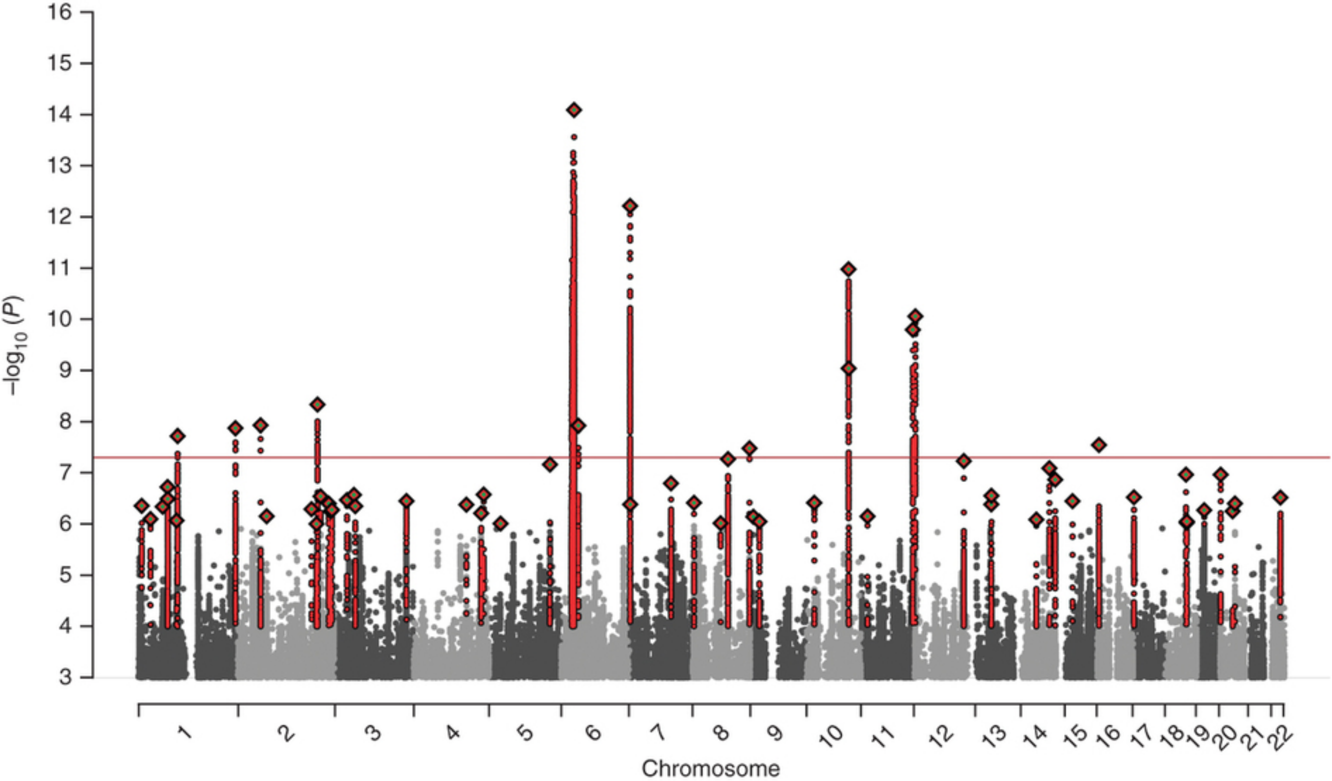


OUTPUT FILE FORMAT

Column header	Description	Required format
SNP	SNP label for the variant in form CHR:POS beginning with “chr”	
rsID	rs number	
STRAND	Orientation of the site to the hun genome strand used	
CHR	chromosome	
POS	Position of the SNP on chromos	
EFFECT_ALLELE	Allele at this site to which the eff been estimated	
NON_EFFECT_ALLELE	Allele at this site which is not the EFFECT_ALLELE	
N	Total number of samples analyz	
N0	Number of homozygous sample zero copies of the EFFECT_ALLELE	
N1	Number of heterozygous samples with one copy of the EFFECT_ALLELE	numeric

N2	Number of homozygous samples with two copies of the EFFECT_ALLELE	nume
EAF	Allele frequency of the EFFECT_ALLELE	Freq of th
HWE_P	Exact HWE p-value for the sample analyzed	4 dig
BETA	Estimate of the effect size	3 dig
SE	Estimated standard error on the estimate of the effect size	4 dig
PVAL	Significance of the variant association, uncorrected for genomic control	3 dig notat
IMPUTED	Is the SNP imputed?	0=ge
RSQR	Imputation quality metric; (RSQ for MACH, INFO for PLINK, info	

Questions?


16
15
14
13
12
11
10
9
8
7
6
5
4
3
$-\log_{10}(P)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome