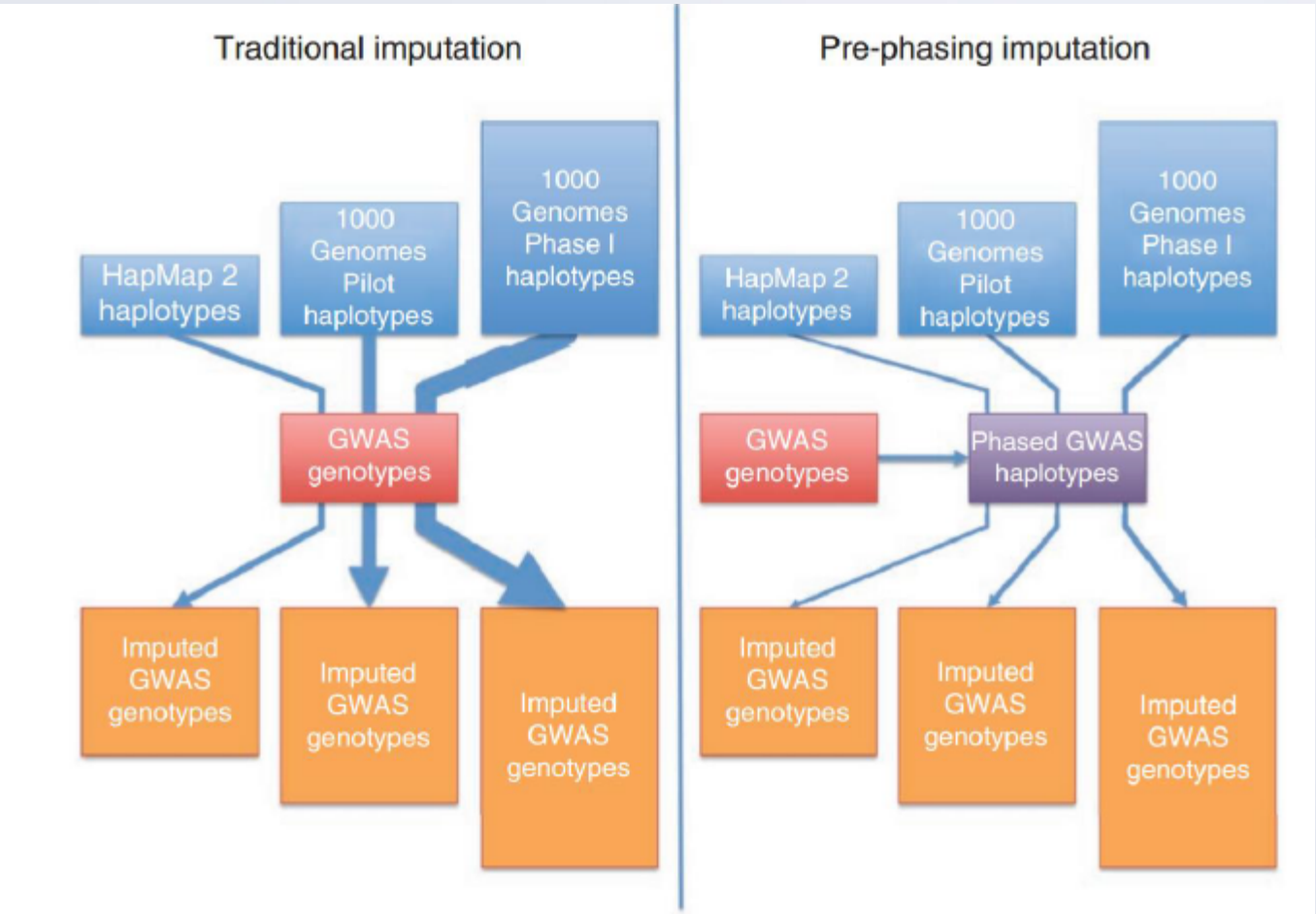
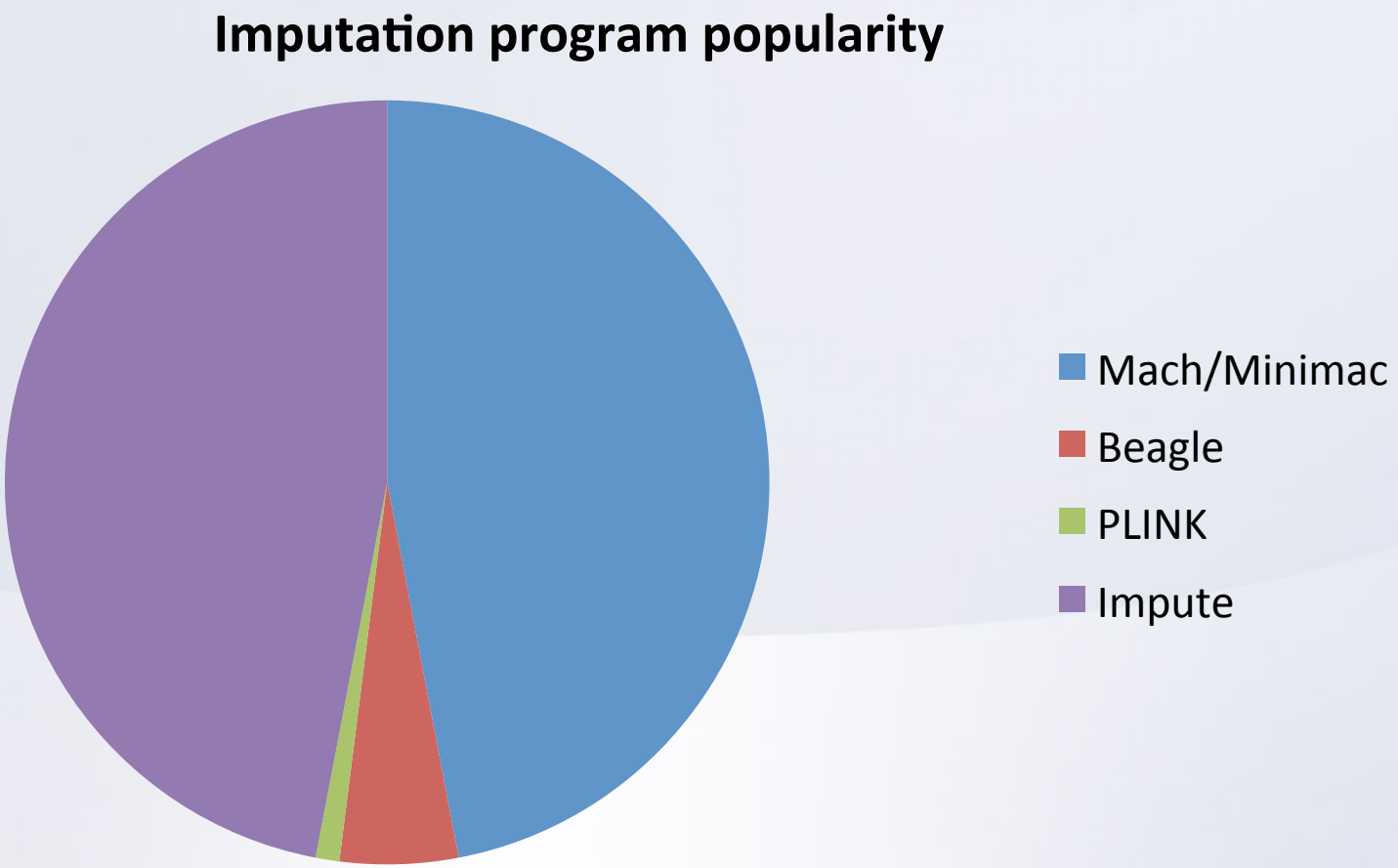
Imputation Practical

Sarah Medland

Ways to approach imputation



Programs used for imputation



- <u>NEVER</u> use PLINK for imputation!

How do they compare

- Similar accuracy
- Similar features
- Similar time frames
- Different data formats
  - Mach/minimac – individual=row snp=column
  - Impute – snp=row individual=column
  - Important for down stream analysis
- Different philosophies
  - Frequentist vs Bayesian

Mach/minimac

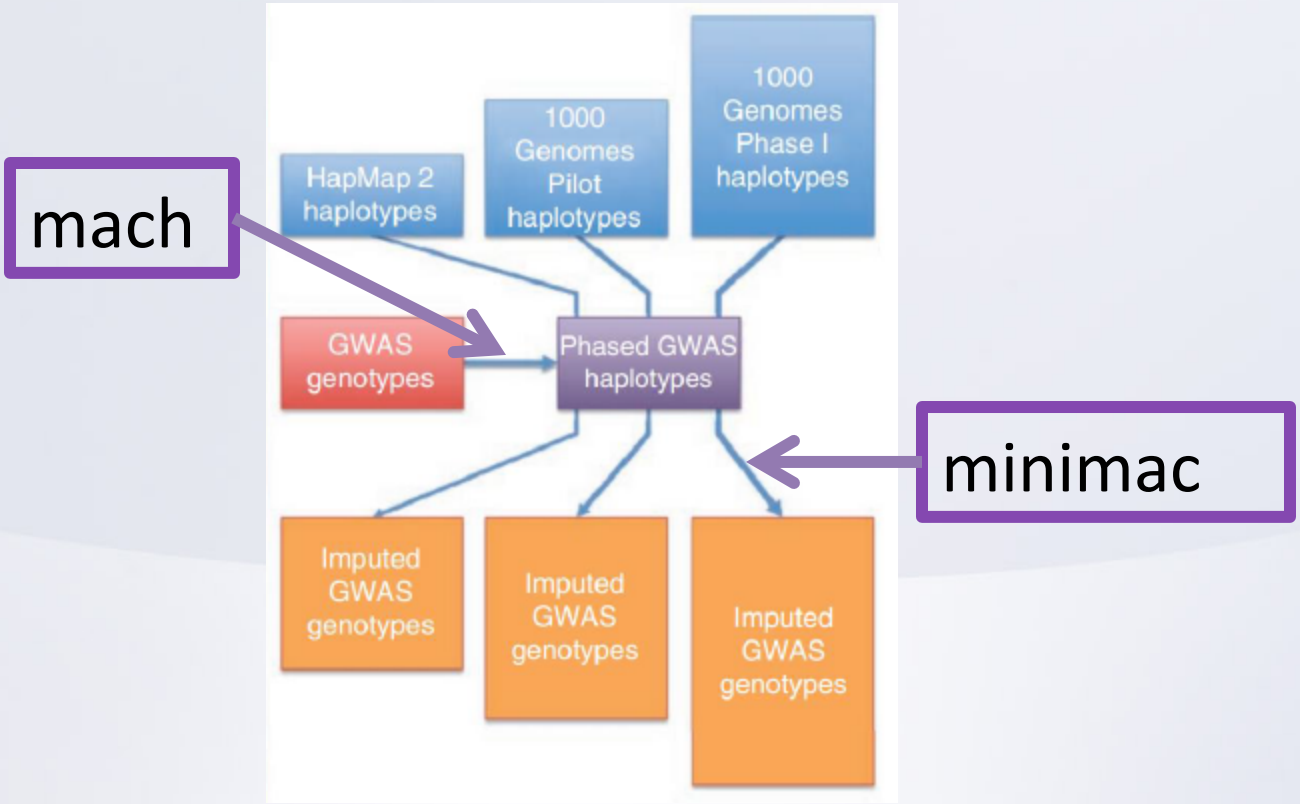


- http://genome.sph.umich.edu/wiki/Minimac
- http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook
- Built by Gonçalo Abecasis, Yun Li, Christian Fuchsberger and colleagues
- Downstream analysis options
  - Mach2qtl (continuous phenotypes)
  - Mach2dat (binary phenotypes)
  - Merlin-offline (family/twin based samples)

Impute2

- https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook
- Built by Jonathan Marchini, Bryan Howie and colleges
- Downstream analysis options
  - SNPtest
  - Quicktest

Today – mach/minimac



And yes, naming their software is not their strong point…

Related Pages

If you are looking to learn about small computers made by Apple, Inc., you have come to the wrong page. Try looking at http://www.apple.com/macmini/, instead.

If you are looking for a low calorie version of the Big Mac sandwich, you'll be sad to know the Mini Mac has been discontinued. However, you are not the only one who likes the idea of a Mini Mac and you'll probably find some company on the web [2].

Steps involved

1. Data Prep
   i. QC of own data
   ii. Selection of references and comparing own data against the reference
   iii. Updating build and alignment

Steps involved

1. Data Prep
2. Phasing in Mach
   i. Dividing the data sets into chunks
   ii. Reformatting data
   iii. Phasing each chunk

   (technically you can do this on a desktop in reality you will need access to a server)

Steps involved

1. Data Prep
2. Phasing in Mach
3. Imputing in minimac
   i. Setup the reference files
   ii. Impute

   (technically you can do this on a desktop in reality you will need access to a server)

Data Prep

1. QC of own data
   i. Convert to PLINK binary format
   ii. Exclude snps with excessive missingness (>5%), low MAF (<1%), HWE violations (~$P<10^{-4}$), Mendelian errors

(Derrek showed you how to do this)

Data Prep

2. Selection of references and comparing own data against the reference
   i. 1KGP phase1 v3 - http://www.sph.umich.edu/csg/abecasis/MaCH/download/
   - Note phase 2 references due out ~August
   - 1 vs all ethnicities

References are in 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_alignments/README">
##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 HG00109 HG00110 HG00111
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     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     0|0     0|0
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     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     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     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     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     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     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     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     0|0     0|0
```

(more about this format in Kwangsik's talk)

Data Prep

3. Updating build and alignment
   i. 1KGP phase1 v3 references are stored in build 37
   ii. Most GWAS data is stored in build 36
   iii. You must convert your data to build 37 prior to imputation
      http://genome.sph.umich.edu/wiki/LiftOver
      http://genome.ucsc.edu/cgi-bin/hgLiftOver
   iv. Use PLINK to update the positions of your snps and rename the snps into CHR:BP format
      ie rs148087467 becomes 10:60523

Data Prep

iv. Use PLINK to update the positions of your snps and rename the snps into CHR:BP format

v. Output your data in merlin format

```
plink --bfile lastQC --extract 1kgp.snps --update-map 1kgp.chr
--update-chr --flip flip.list --make-bed --out temp --noweb

plink --bfile temp --update-map 1kgp.bp  --geno 0.05  --mind 0.05
--make-bed --out lastQCb37 --noweb
```

```
for i in {1..22}
do
echo "plink --bfile lastQCb37 --chr "$i"  --recode --noweb --out
Mach/ready4mach."$i"" >> plink_writeout.sh
done
```

```
for i in {1..23}
do
echo "S dummy" > Mach/ready4mach."$i".dat
awk '{ print "M", $1 ":" $4}' Mach/ready4mach."$i".map >>
Mach/ready4mach."$i".dat
done
```

Merlin Format

```
S dummy
M 23:2700157
M 23:2732096
M 23:2743627
M 23:2772660
M 23:2789848
M 23:2813287
M 23:2822253
M 23:2825403
M 23:2847133
M 23:2849981
M 23:2928555
M 23:2944537
M 23:2951931
M 23:2996162
M 23:3002687
M 23:3012405
M 23:3025174
M 23:3028385
M 23:3030426
```

```
00359 0035902 0 0 2 -9 C C G G T T T C G G G G A C A A T T C C A G A G T T A A C T A A G
00359 8937101 0 0 2 -9 C C G G T T T C G G G G A C A A T T C C A G A G T T A A C T A A G
00955 0095501 0 0 1 -9 C C G G T T T C A G A G A C A A T T C C A G A G T T A A C T A A G
00955 0095502 0 0 2 -9 C C G G T T T C A G A G A A A A T T C C A G A G T T A A C T A A G
01149 0114901 0 0 2 -9 C C G G C T C C A A A G A C A A T T C C A G A G T T A A C T A A G
01149 8939202 0 0 2 -9 C C G G C C C C A A G G C C A A T T C C A A A A T T A A C C A A A
01160 0116001 0 0 2 -9 T C A G T T T C G G G G C C G A T T C C A G A G C T A A C T A A A
01160 0116002 0 0 2 -9 T C A G T T T C G G G G C C G A T T C C A G A G C T A A C T A A A
01168 0116801 0 0 2 -9 C C G G T T T T G G G G C C A A T T C C A A A A T T A A C C A A A
01168 0116802 0 0 2 -9 C C G G T T T C A G A G C C G A C T C C A G A G C T A A C T A A A
01376 0137601 0 0 1 -9 C C G G T T C C A A A A A C A A T T C C A G A G T T A A C T A A G
01376 0137602 0 0 1 -9 C C G G T T C C A A A A A C A A T T C C A G A G T T A A C T A A G
01376 6379550 0 0 2 -9 C C G G T T C C A A A G A C G A T T T C G G G G T T C A T T A A G
```

Phasing in Mach

1. Dividing the data sets into chunks
   i. http://genome.sph.umich.edu/wiki/ChunkChromosome
2. Phasing each chunk (bash)

```
for i in {1..23}
do
ChunkChromosome  -d ready4mach."$i".dat.gz -n 5000 -o 500
done
# loop over parts
for ((j=1; j<=40; j++))
do
for i in {1..22}
do
if test -f chunk"$j"-ready4mach."$i".dat.gz
then
echo  "mach1 -d chunk"$j"-ready4mach."$i".dat.gz -p
ready4mach."$i".ped.gz --prefix chunk"$j"-ready4mach."$i" --rounds 20
--states 200 --phase > chunk"$j"-ready4mach."$i".mach.log" >>
MaCH_phasing.sh
fi
```

./MaCH_phasing.sh

- Autochunk files

START	STOP	CORE_START	CORE_END	
2:18856	2:29793129	start	2:27405497	
2:23437969		2:52309494	2:27414134	2:50398712
2:48044772		2:79429714	2:50398902	2:77048211
2:75012327		2:121254949	2:77048896	2:118365400
2:115217554		2:154145266	2:118366904	2:151465458
2:147560832		2:184066858	2:151469069	2:181045431
2:178809106		2:218991005	2:181055597	2:216601700
2:213794754		2:243044147	2:216611057	stop

- Chunked dat files

```
M 13:51808360
M 13:51810716
M 13:51810953
M 13:51814527
M 13:51816665
M 13:51820344
M 13:51822189
M 13:51824328
S2 13:51828960
S2 13:51830473
S2 13:51831286
S2 13:51833944
S2 13:51839210
S2 13:51841291
S2 13:51842513
S2 13:51844429
S2 13:51845324
```

Phased data

```
00359->0035902 HAPLO1 TTATATAAGTCGTTGACGTGTCCAGCCCCAATGCCACCGAGTGTTACA
00359->0035902 HAPLO2 TTATGTAAGTTACTTGTACAGCCAGCCCCACCGCCACCGAGTGTTACG
00955->0095501 HAPLO1 TTATATAAGTCGTTGACGTGTTCAGCTCCACTGACGTAGAACACTACA
00955->0095501 HAPLO2 TGACATAAACCACTTGTGTATCCAGCCCCAATGCCACCGAGTGTTACG
01149->0114901 HAPLO1 TTATGTCAGTCGCCGGTGCATCTCAACCCACTGACGTAGAACACTATA
01149->0114901 HAPLO2 CGGCGCACACCACTTGCATATCCAGATCTCCTGATATAAGATATTACG
01160->0116001 HAPLO1 TTATGTCAGTCGCCGGTGCATCTCAACCCACTGACGTAGAACACTATA
01160->0116001 HAPLO2 CGGCGCACACCACTTGCATATCCAGCCCCAATGCCACCGAGTGTTACG
01168->0116801 HAPLO1 CGGCGCACACCACTTGCATATCTCAACCCACTGACGTAGAACACTATA
01168->0116801 HAPLO2 CGGCGCACACCACTTGCATATTCAGCTCCACTGACGTAGAACACTACA
01376->0137601 HAPLO1 CGGCGCACACCACTTGCATATCCAGCCCCAATGCCACCGAGTGTTACG
01376->0137601 HAPLO2 CGGCGCACACCACTTGCACAGCCAGACCCACTAATACAGAATACTACA
```

Imputing in minimac

```
#loop over parts
for ((j=1; j<=40; j++))
do
# Impute into phased haplotypes
for i in {1..22}
do
if test -f chunk"$j"-ready4mach."$i".dat.gz
then
echo "minimac --vcfReference  --rounds 5 --states 200 --refHaps
../chr"$i".phase1_release_v3.20101123.snps_indels_svs.genotypes.refpa
nel.EUR.nosingles.vcf.gz  --haps chunk"$j"-ready4mach."$i".gz --snps
chunk"$j"-ready4mach."$i".dat.gz.snps --autoClip
autoChunk-ready4mach."$i".dat.gz --gzip --prefix
chunk"$j"-ready4mach."$i".imputed >
chunk"$j"-ready4mach."$i"-minimac.log" >> MiniMac-impute.sh
fi
done
```

Output

- Dosage data

00359->0035902	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
00955->0095501	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
01149->0114901	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
01160->0116001	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
01168->0116801	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
01376->0137601	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02035->0203501	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02038->0203801	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02045->0204501	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02047->8942701	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02052->8710701	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02054->0205402	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02064->0206401	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02144->0214401	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02233->0223302	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
02917->0291701	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082
03066->0306602	DOSE	1.805	1.984	1.876	1.884	1.971	1.852	1.082

Output

- Info files

SNP	Al1	Al2	Freq1	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose1	Dose2
1:10583	G	A	0.79288	0.20712	0.79288	-0.00000	-	-	-	-	-	-
1:10611	C	G	0.97889	0.02111	0.97889	0.00000	-	-	-	-	-	
1:13302	C	T	0.86280	0.13720	0.86280	-0.00000	-	-	-	-	-	-
1:13327	G	C	0.96042	0.03958	0.96042	-0.00000	-	-	-	-	-	-
1:95207182	T	C	0.99547	0.00453	0.99547	0.10108	-	-	-	-	-	-
1:95207382	T	T	1.00000	0.00000	1.00000	0.00000	-	-	-	-	-	-
1:95207442	C	T	0.62754	0.37246	0.99999	1.00507	Genotyped	0.98810	0.99822	0.99645	0.99484	0.00421
1:95207524	G	A	0.78061	0.21939	1.00000	1.00511	Genotyped	1.00059	1.00000	1.00000	0.99924	0.00083
1:95207532:TG_T	R	D	0.78620	0.21380	0.99441	0.97729	-	-	-	-	-	-
1:95207558	C	T	0.99399	0.00601	0.99399	0.05165	-	-	-	-	-	-
1:95207633	A	C	0.93366	0.06634	0.99998	1.00482	Genotyped	0.94847	0.99901	0.99802	0.99621	0.00372
1:95207846	G	T	0.98937	0.01063	0.98942	0.31316	-	-	-	-	-	-

Imputation quality evaluation

Minimac hides each of the genotyped SNPs in turn and then calculates 3 statistics:

- looRSQ - this is the estimated rsq for that SNP (as if SNP weren't typed).
- empR - this is the empirical correlation between true and imputed genotypes for the SNP. If this is negative, the SNP alleles are probably flipped.
- empRSQ - this is the actual R2 value, comparing imputed and true genotypes.

These statistics can be found in the *.info file

Be aware that, unfortunately, imputation quality statistics are not directly comparable between different imputation programs (MaCH/minimac vs. Impute vs. Beagle etc.).

```
Mach2Qtl V1.1.0 (2011-05-23) -- QTL Association Mapping with Imputed Allele Counts
(c) 2007 Goncalo Abecasis, Yun Li


The following parameters are in effect:

Available Options
         Phenotypic Data : --datfile [], --pedfile []
   Imputed Allele Counts : --infofile [], --dosefile [], --probfile []
       Analysis Options : --useCovariates [ON], --quantileNormalization,
                          --dominant, --recessive, --additive [ON]
                 Output : --samplesize
```

```
mach2dat 1.0.21 -- Disease-snp Association Tests with Imputed Dosages
(c) 2008 Yun Li, Wei Chen, Goncalo Abecasis


The following parameters are in effect:

Available Options
         Phenotypic Data : --datfile [pheno.dat], --pedfile [pheno.ped]
   Imputed Genotype Data : --infofile [sample.mlinfo],
                           --dosefile [sample.mldose]
       Analysis Options : --useCovariates [ON], --likelihoodratio [ON],
                           --samplesize [ON], --verboseSampleSize,
                           --nrrounds [20], --rsqcutoff [1.0e-04],
                           --method [newton]
                 Output : --frequency
```

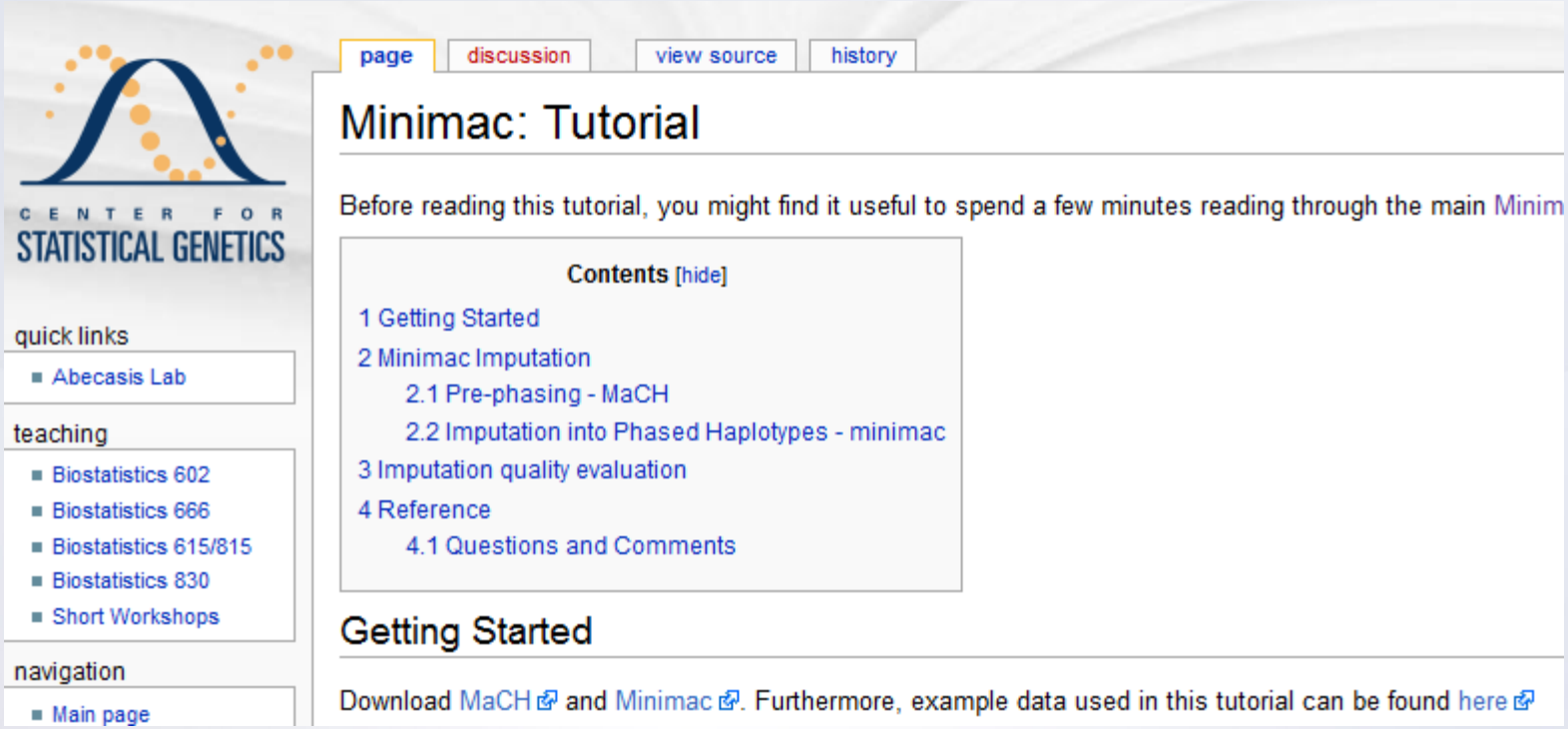
```
MERLIN -- Offline Association Analysis
          (c) 2006-2007 Goncalo Abecasis


The following parameters are in effect:
                    Data File :      merlin.dat (-dname)
                 Pedigree File :      merlin.ped (-pname)
                      Map File :      merlin.map (-mname)
                Frequency File :     merlin.freq (-fname)

Additional Options
   Inferred Genotypes : --datinfer [merlin-infer.dat],
                        --pedinfer [merlin-infer.ped]
     Analysis Options : --inverseNormal, --useCovariates, --filter,
                        --custom [covars.tbl]
        Output Files : --prefix [merlin], --pdf, --tabulate
```

Files to practice with

http://genome.sph.umich.edu/wiki/Minimac:_Tutorial



CENTER FOR STATISTICAL GENETICS

quick links
- Abecasis Lab

teaching
- Biostatistics 602
- Biostatistics 666
- Biostatistics 615/815
- Biostatistics 830
- Short Workshops

navigation
- Main page

page | discussion | view source | history

Minimac: Tutorial

Before reading this tutorial, you might find it useful to spend a few minutes reading through the main Minim

Contents [hide]

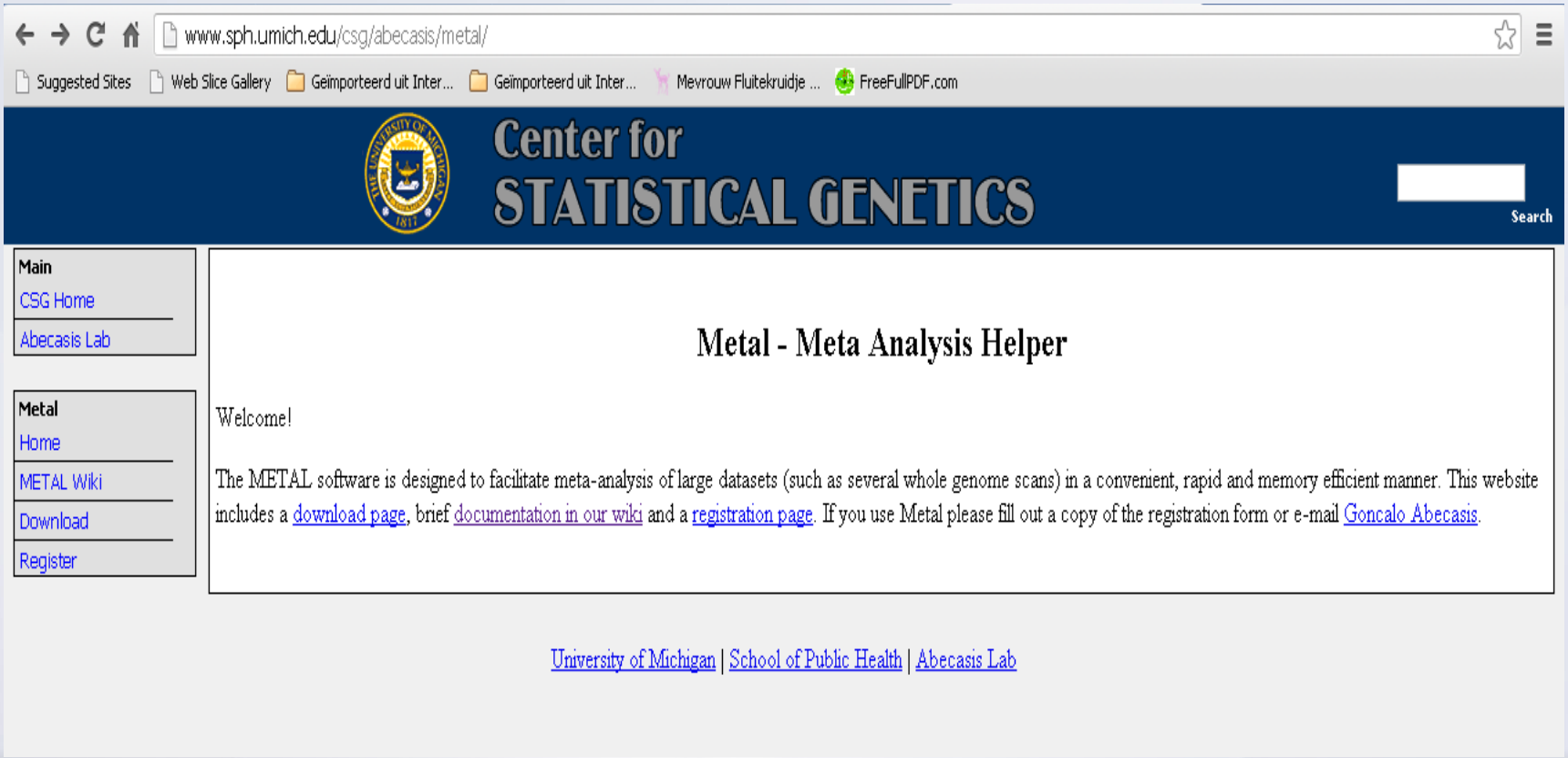
Getting Started

Download MaCH and Minimac. Furthermore, example data used in this tutorial can be found here

Meta-analysis Practical

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

- Metal is flexible
  - It can run fixed effects meta-analysis
  - Heterogeneity tests
  - Effect size, Sample Size, or Weighted meta-analysis

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

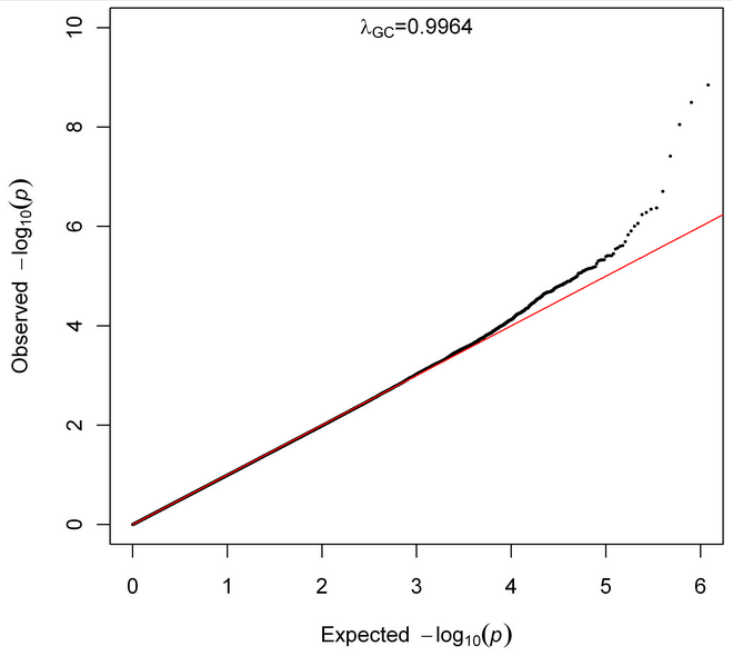
- SNP
- Effect allele
- OR/Beta
- SE [for standard error meta-analysis]
- P-value [for Z-score meta-analysis]
  - If we had two samples of different sizes & we wanted to do a p-value MA we would have to add an N/weight column

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
```

```
## 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

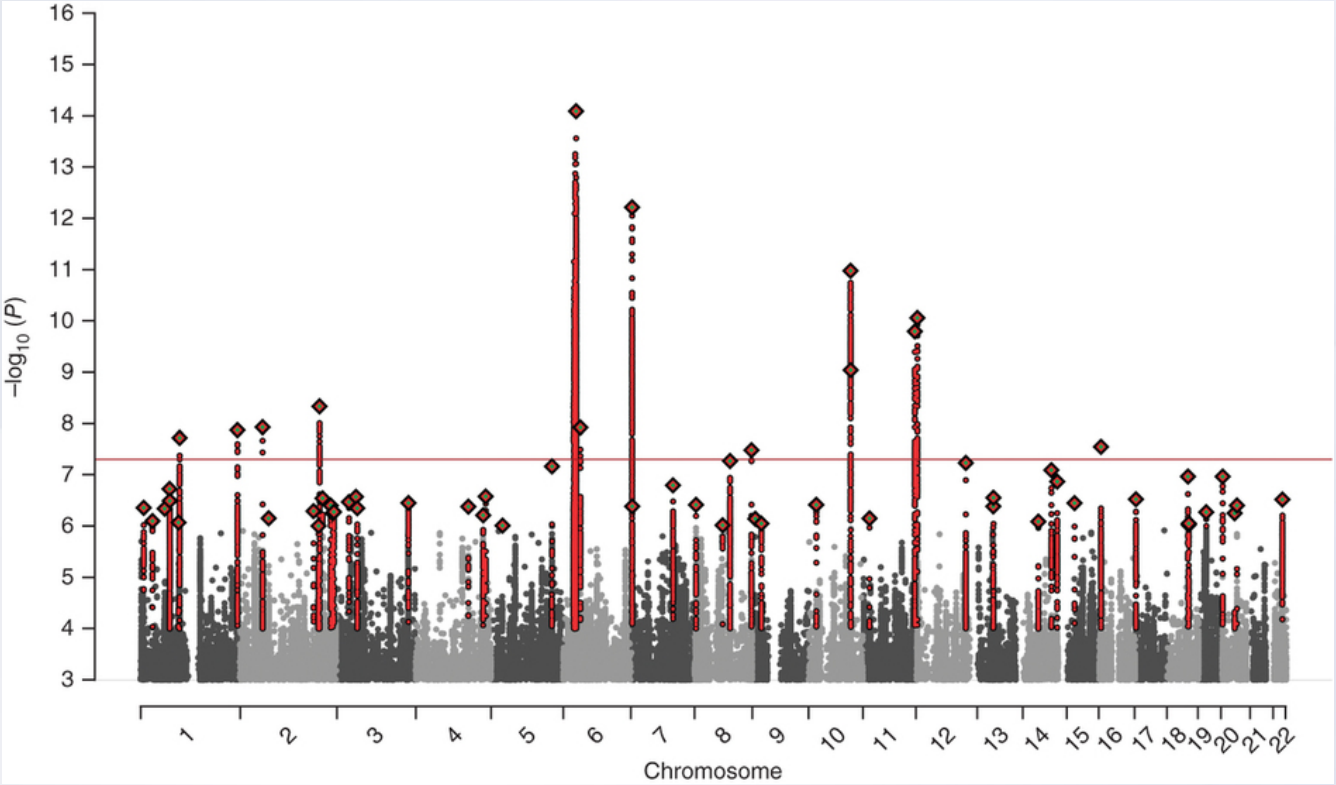
Don’t ask for stuff you don’t need

(It makes you look stupid & its annoying)

OUTPUT FILE FORMAT

Column header	Description	Required format
SNP	SNP label for the variant in format CHR:POS beginning with “chr”	CHR:POS
rsID	rs number	rs number if available
STRAND	Orientation of the site to the human genome strand used	+ or -
CHR	chromosome	
POS	Position of the SNP on chromosome	
EFFECT_ALLELE	Allele at this site to which the effect has been estimated	
NON_EFFECT_ALLELE	Allele at this site which is not the EFFECT_ALLELE	
N	Total number of samples analyzed	
N0	Number of homozygous samples with zero copies of the EFFECT_ALLELE	
N1	Number of heterozygous samples with one copy of the EFFECT_ALLELE	
N2	Number of homozygous samples with two copies of the EFFECT_ALLELE	
EAF	Allele frequency of the EFFECT_ALLELE	
HWE_P	Exact HWE p-value for the sample analyzed	
BETA	Estimate of the effect size	
SE	Estimated standard error on the estimate of the effect size	
PVAL	Significance of the variant association, uncorrected for genomic control	
IMPUTED	Is the SNP imputed?	
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
−log10 (P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome