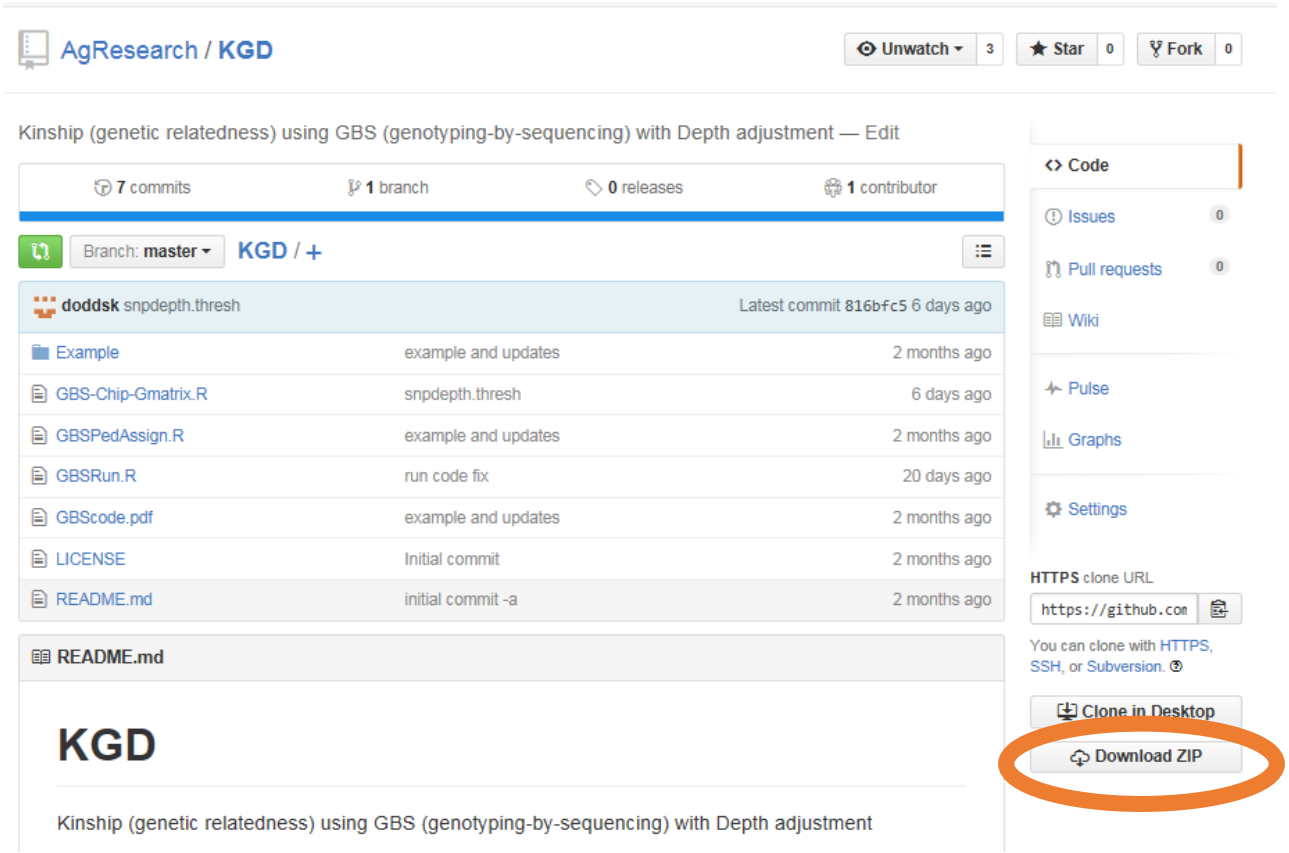
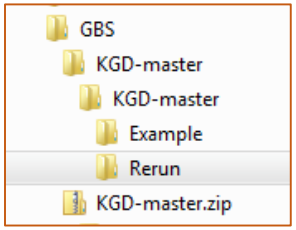
Estimation of relatedness from genotyping by sequencing data

Pre-workshop set-up:
Download zip from: https://github.com/AgResearch/KGD



Prepare directories



Unzip files

Create a new directory (e.g. Rerun) to re-run the example

KGD-master directory:

R code files:
GBS-Chip-Gmatrix.R
GBSPedAssign.R
GBSRun.R Copy to Rerun

Documentation: GBScode.pdf

Copy input files from Example to Rerun

HapMap.hmc.txt.gz

Ped-GBS.csv

Ped-Groups.csv

Prepare GBSRun.R

Open R (R Console used here)

Set working directory to Rerun

File > Change dir … or

setwd("*your-path*/Rerun")

Open GBSRun.R in editor

File > Open script … or …

Modify directory for sourcing code

..

```
genofile <- "HapMap.hmc.txt.gz"

source("../GBS-Chip-Gmatrix.R")
```

Run first 2 lines of code

sink("*R output file*") if you want (sink() to redirect to console)

Preliminary output

```
Read 102 items
Read 14709 records
Analysing 96 individuals and 14709 SNPs
Proportion of missing genotypes:  0.3640826
Mean sample depth: 9.281431
```

1st line: 102 items (# samples + 6)

14709 data rows (# SNPs)

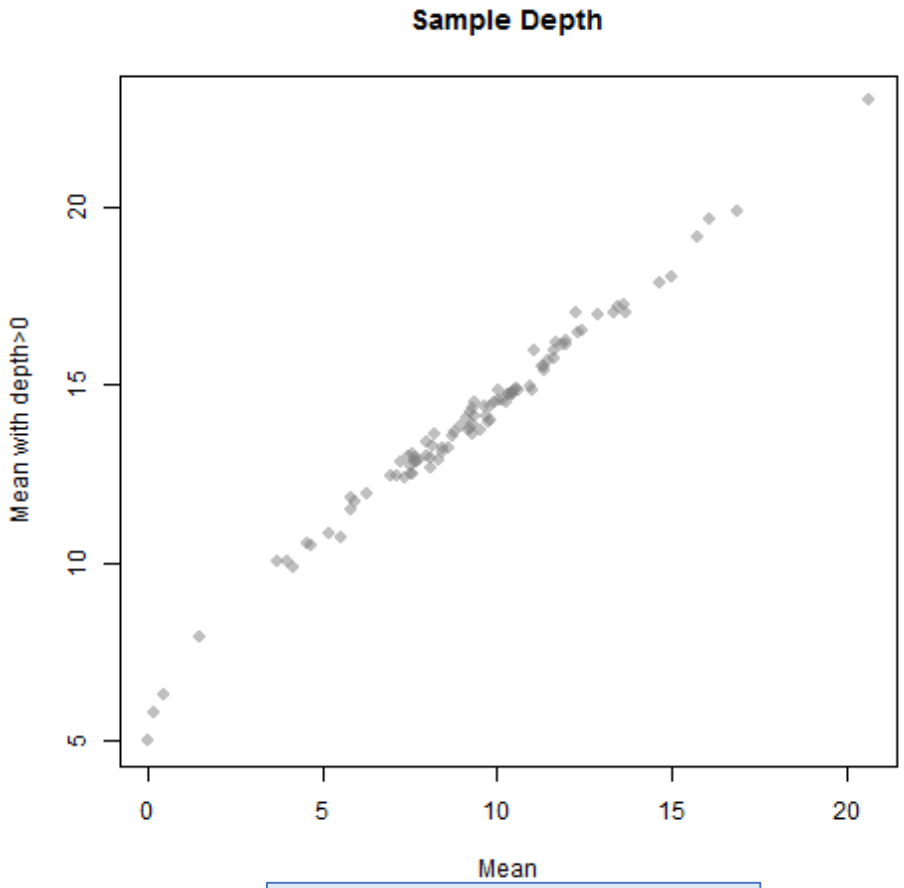
96 individuals (or genotypings)

…

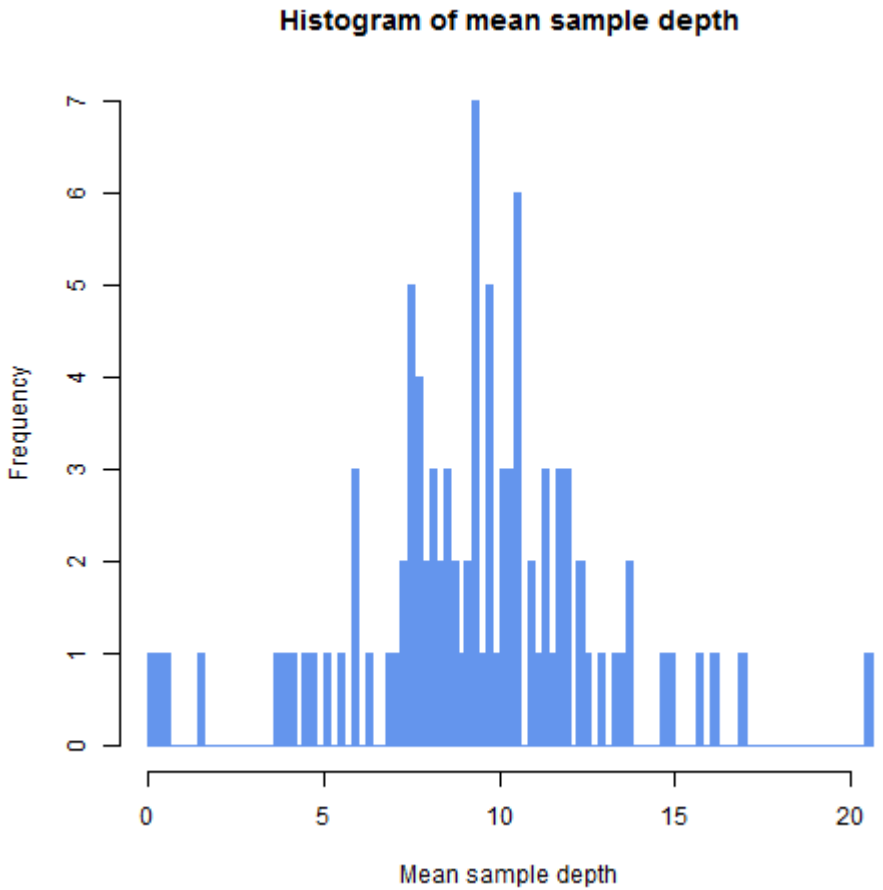
Can be other messages:

- Monomorphic SNPs dropped
- Individuals with very low call rate or depth dropped

Preliminary output – Sample Depth



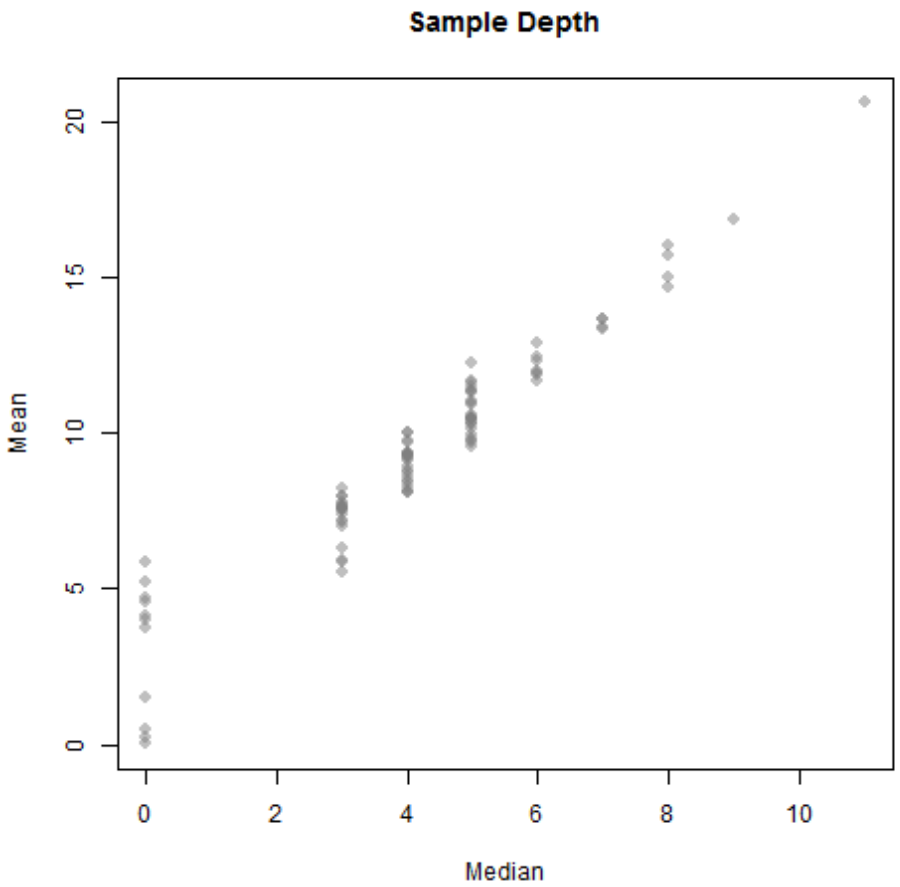
SampDepth-scored.png



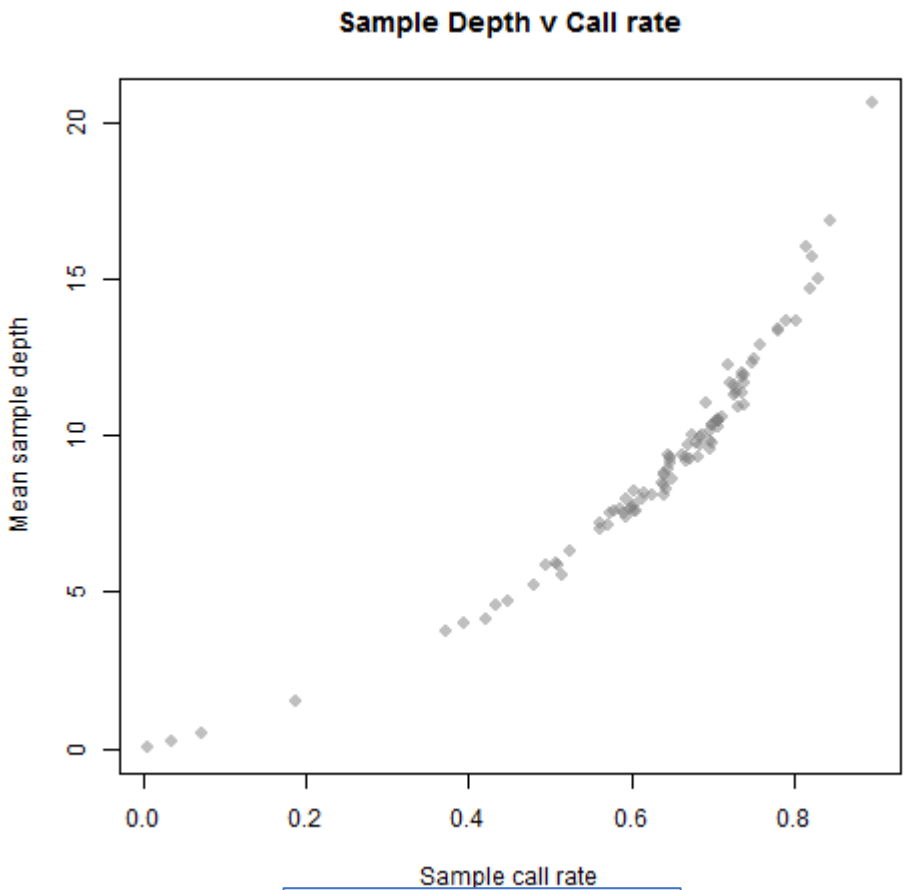
SampDepthHist.png

Some low call rate samples

Preliminary output – Sample Depth

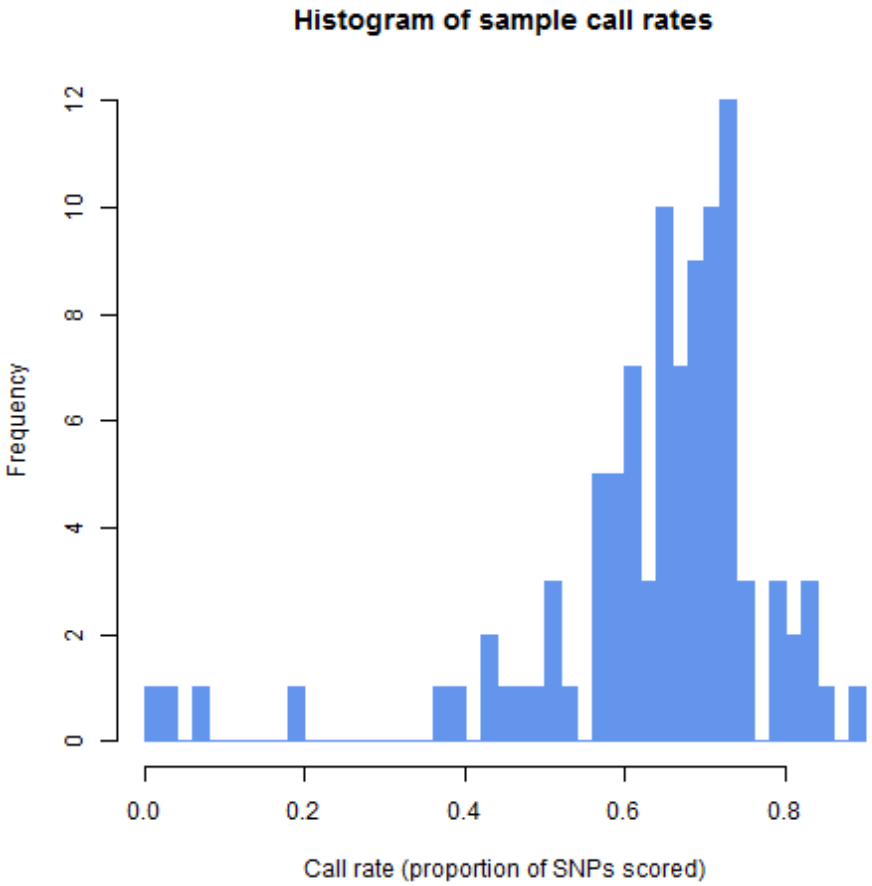


SampDepth.png



SampDepthCR.png

Preliminary output – Sample stats

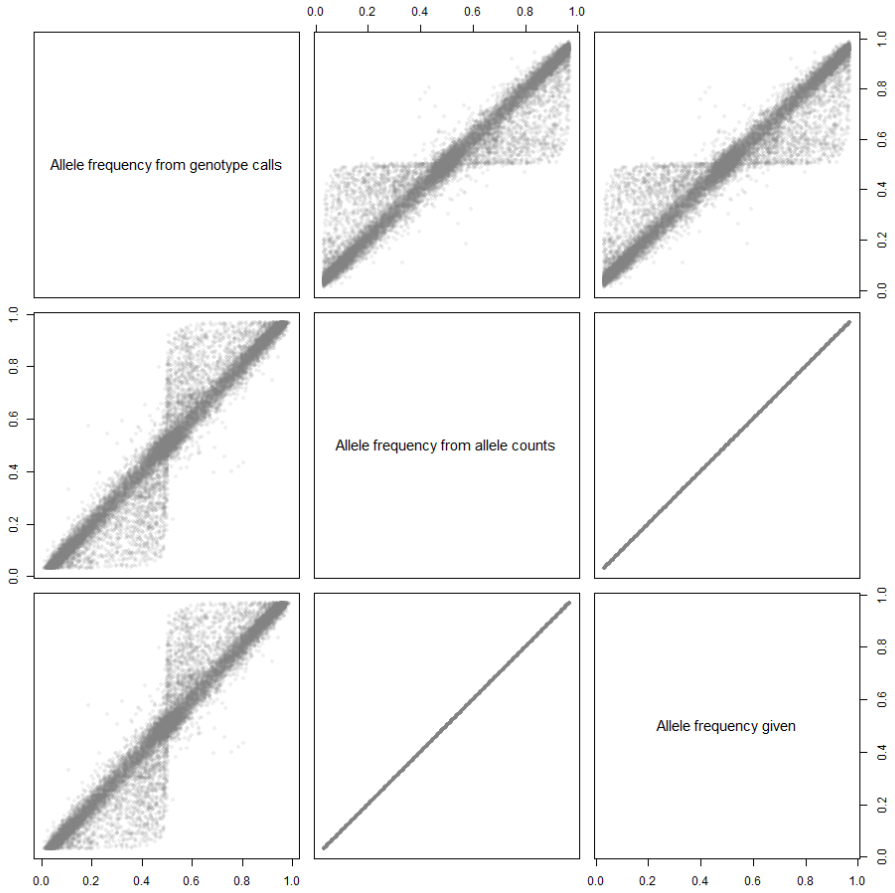


CallRate.png

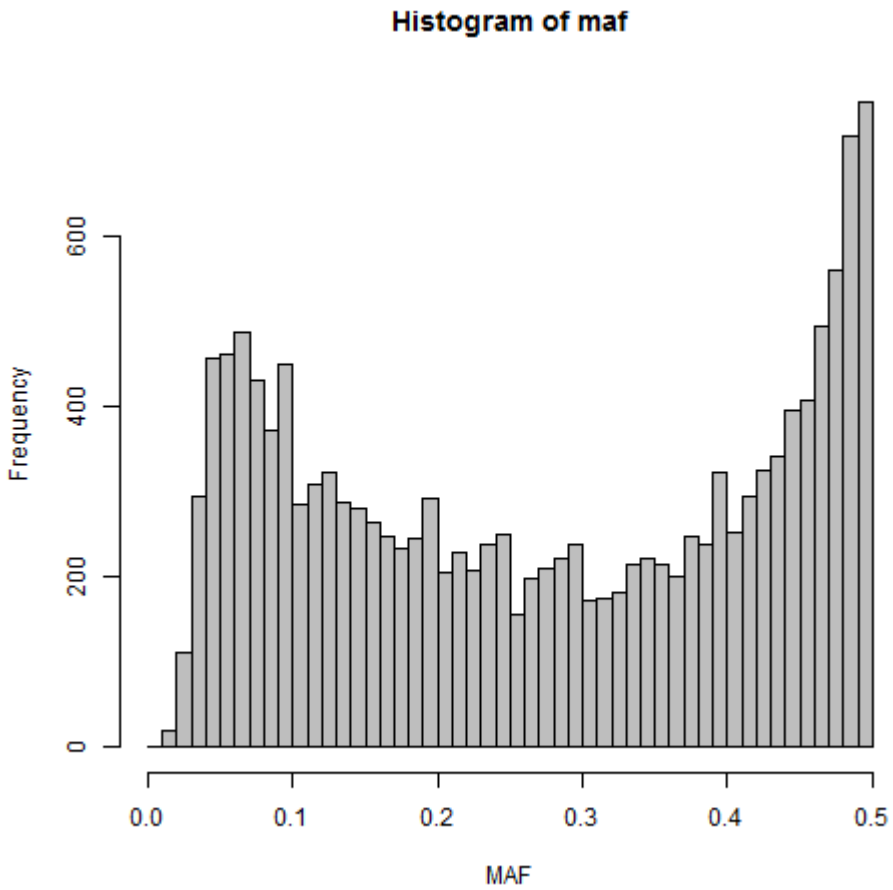
seqID	callrate	sampdepth
Ref	0.647631	9.208444
Seq68	0.494255	5.839486
Seq59	0.729621	11.44497
Seq55	0.479434	5.191311
⋮	⋮	⋮
Blank	0.005303	0.026582

SampleStats.csv

Preliminary output – Allele frequency

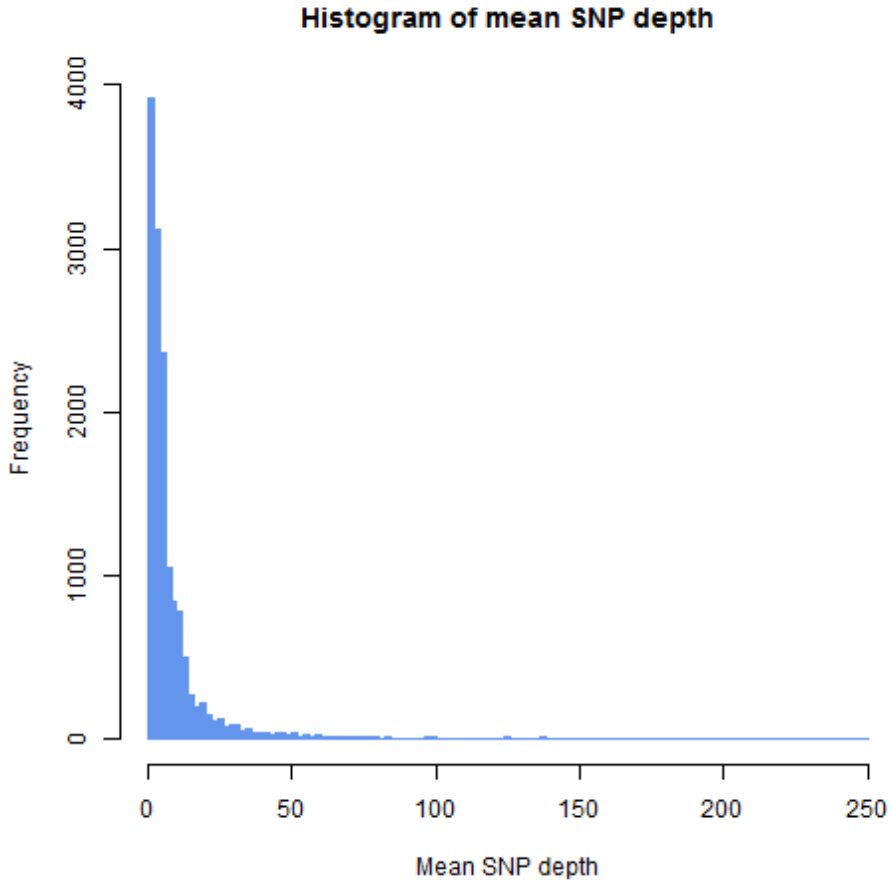


AlleleFreq.png

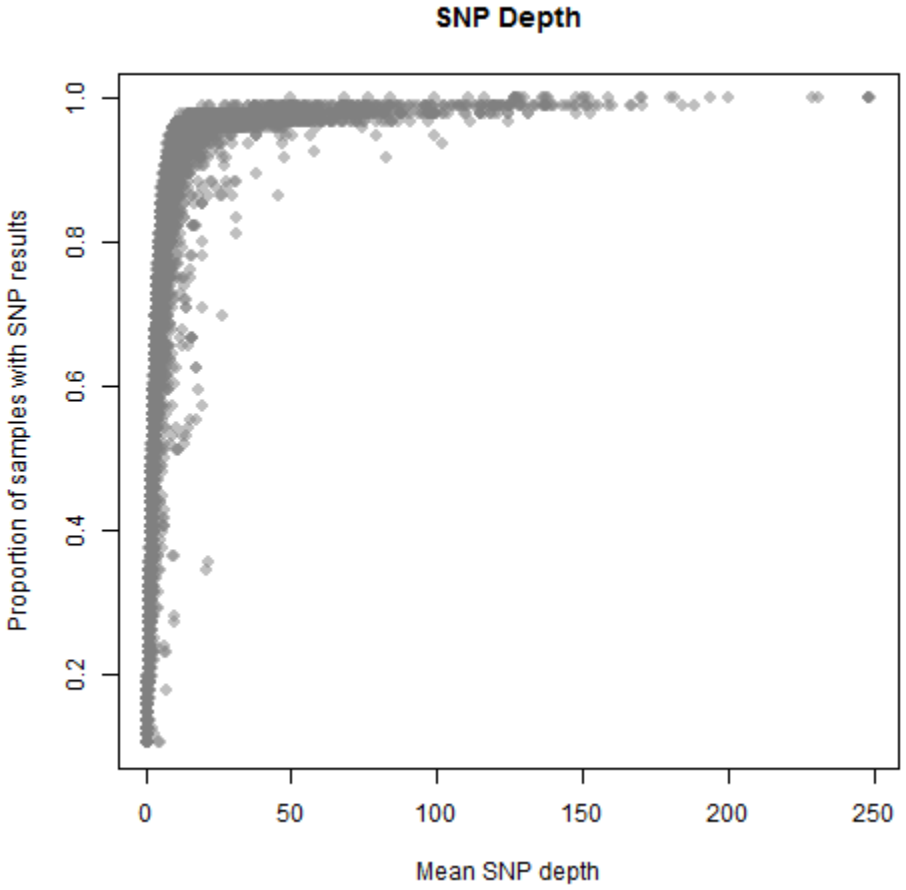


MAF.png

Preliminary output – SNP Depth

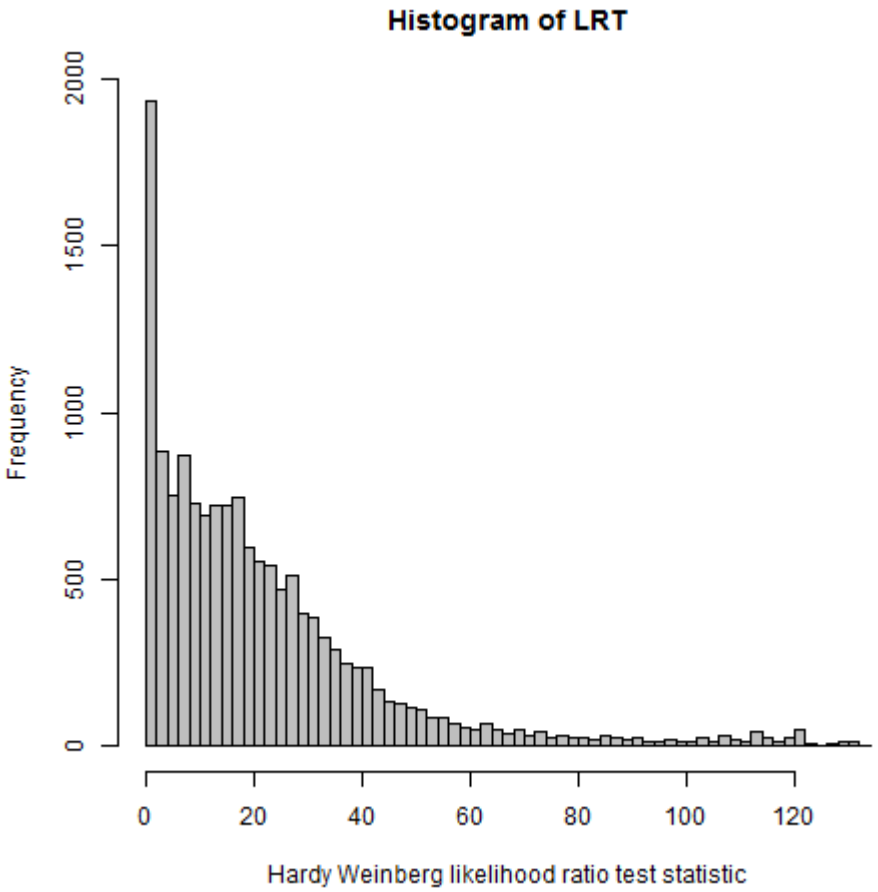


SNPDepthHist.png

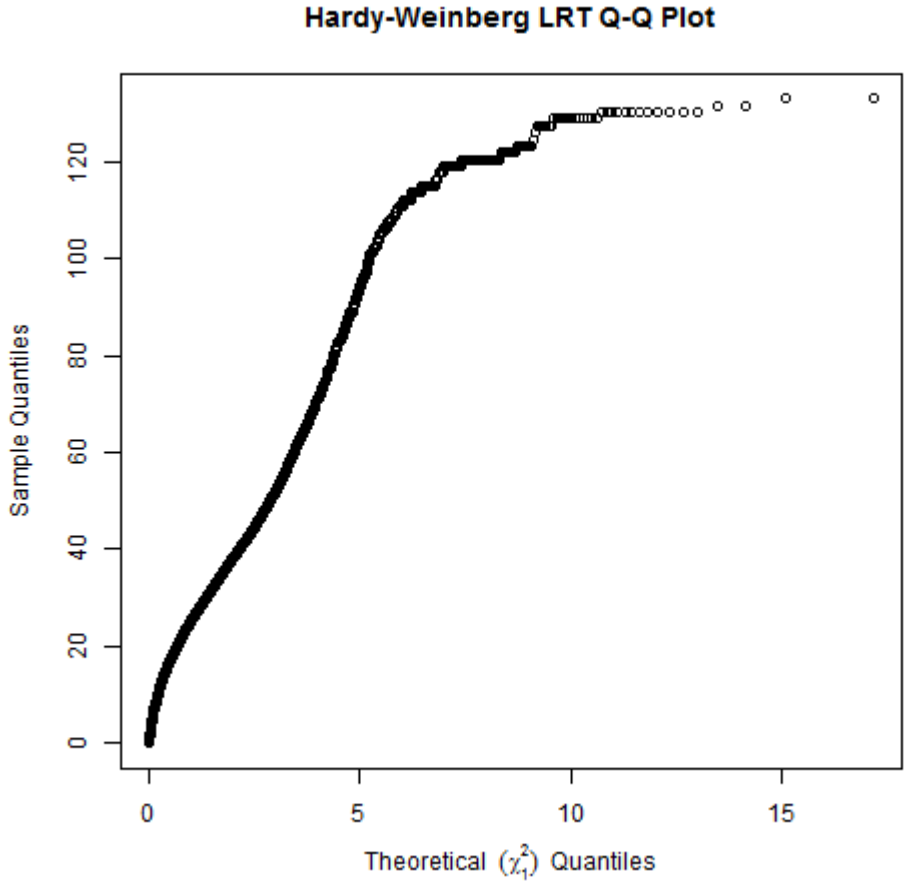


SNPDepth.png

Preliminary output – Hardy-Weinberg



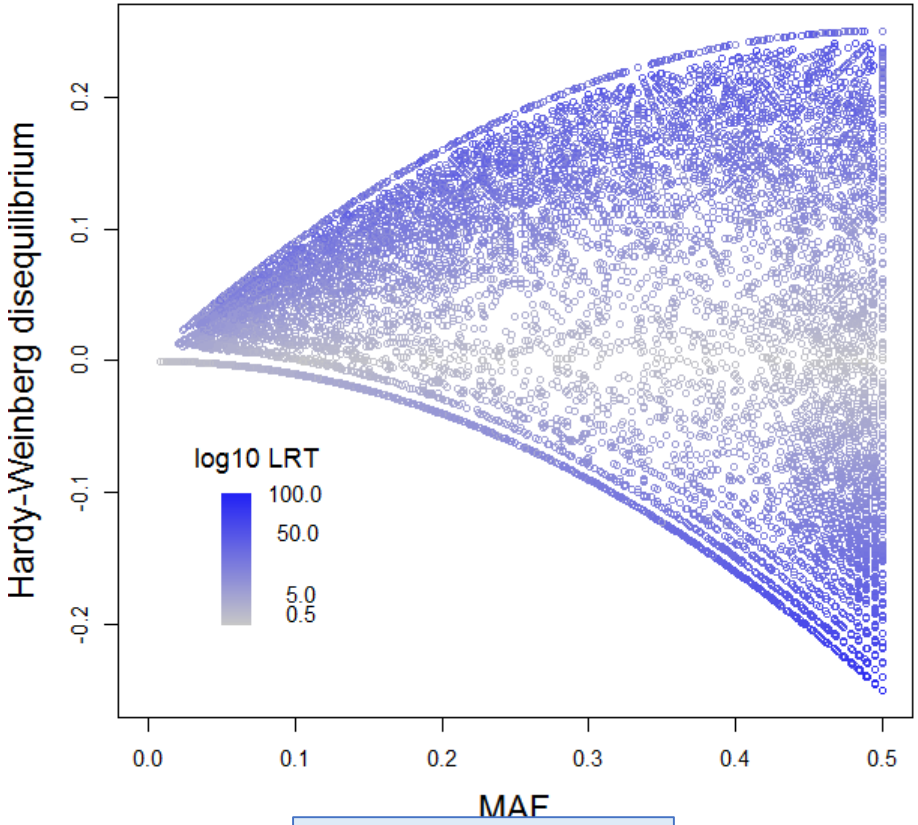
LRT-hist.png



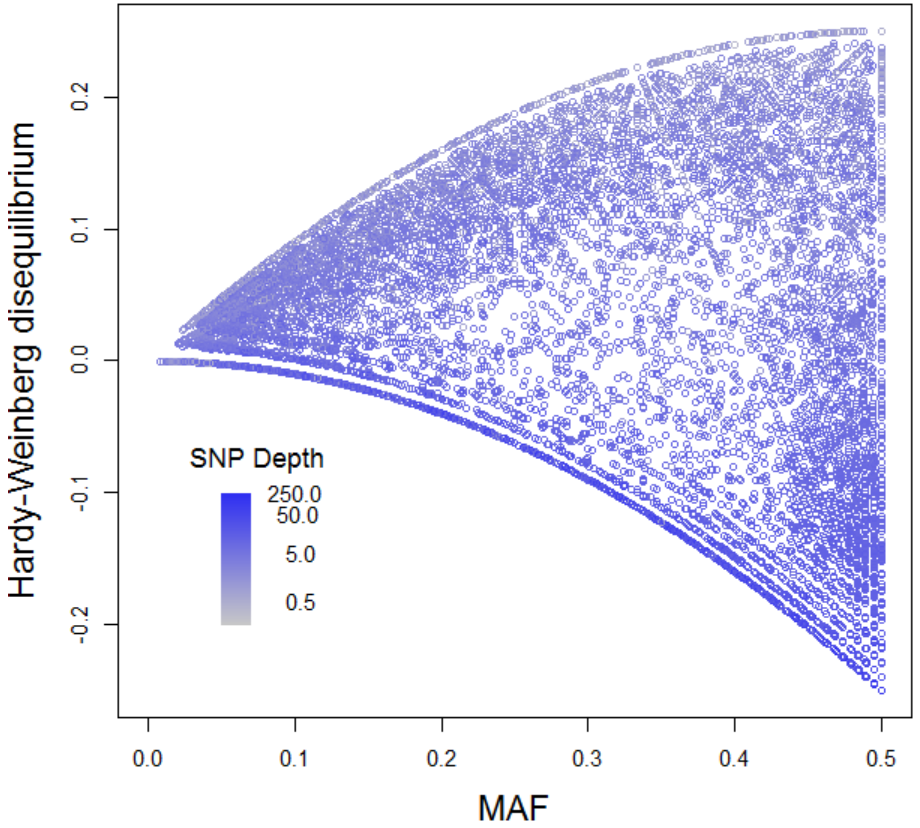
LRT-QQ.png

(Likelihood ratio tests) – not that useful ?

Preliminary output – Hardy-Weinberg



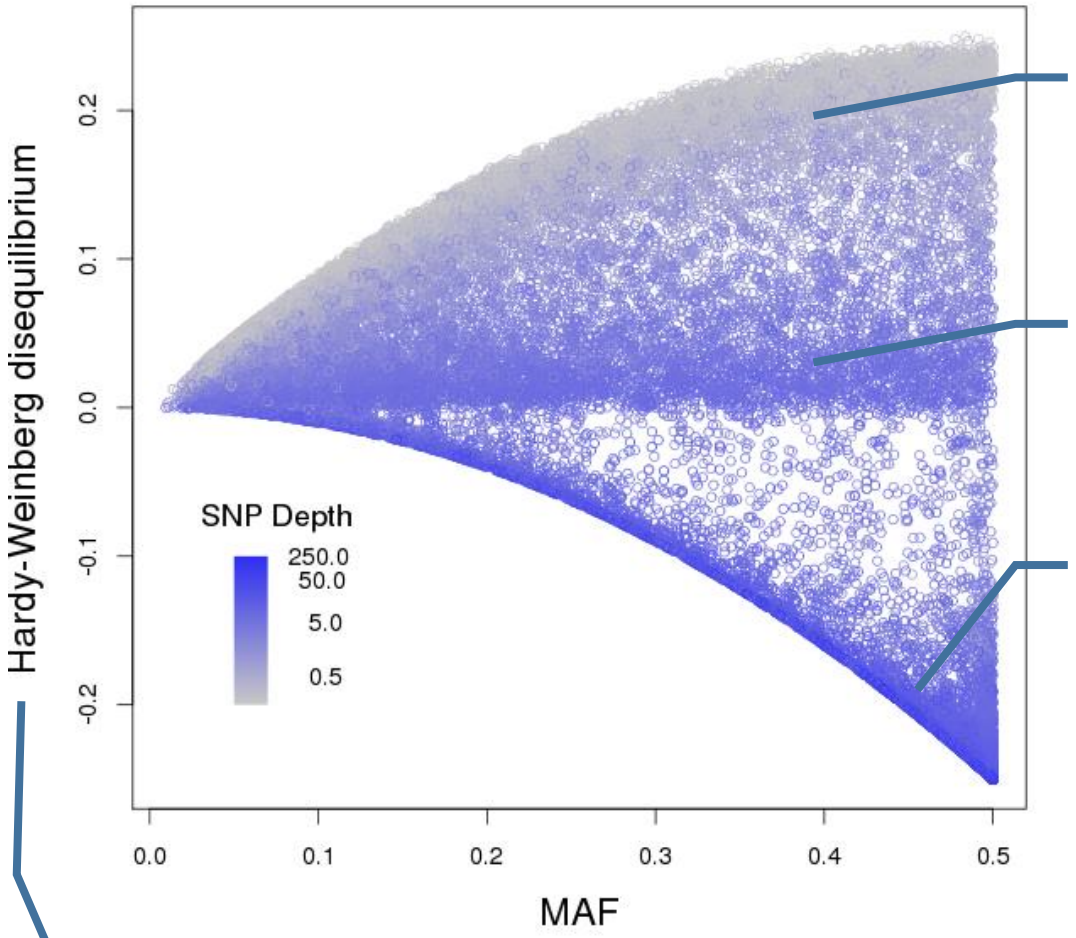
HWdisMAFsig.png



finplot.png

Fin plot

Fin plot – example from paper



Low depth SNPs,
appear homozygous

Medium depth SNPs,
appear in HWE

High depth SNPs,
appear in heterozygous
Possibly duplication regions
Many have high MAF

Hwdis = $p_{AA} - p_A^2$

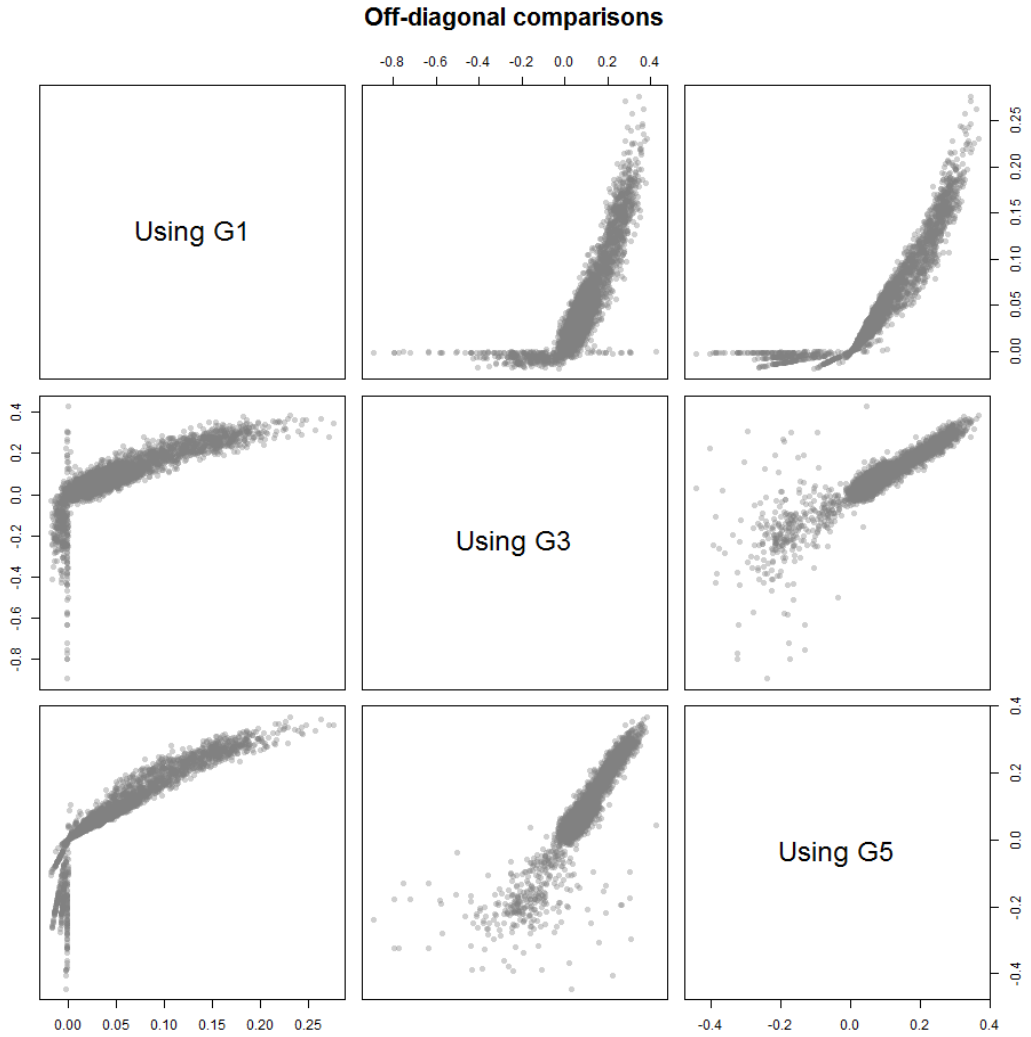
Calculate relatedness

Run next line of code

```
Gfull <- calcG()
```

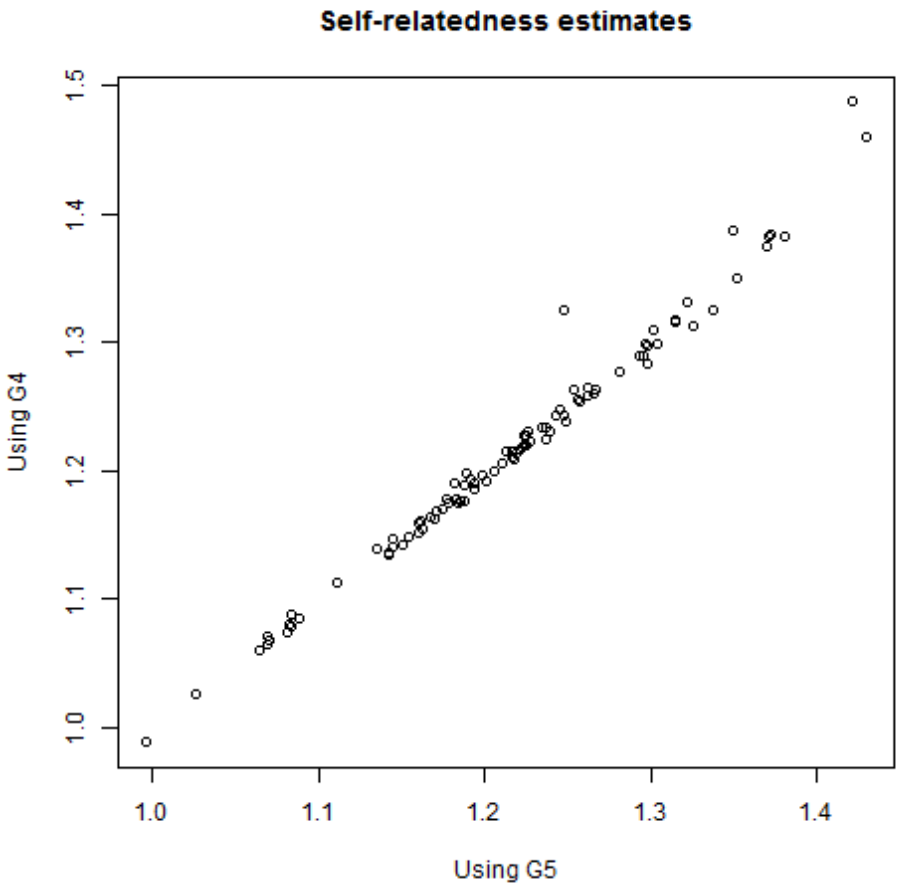
```
Calculating G matrix, analysis code:
# SNPs:  14709
# individuals:  96
Proportion of missing genotypes:  0.3640826
Mean sample depth: 9.281431
Mean self-relatedness (G5 diagonal): 1.215244
```

Comparison of methods (see paper)

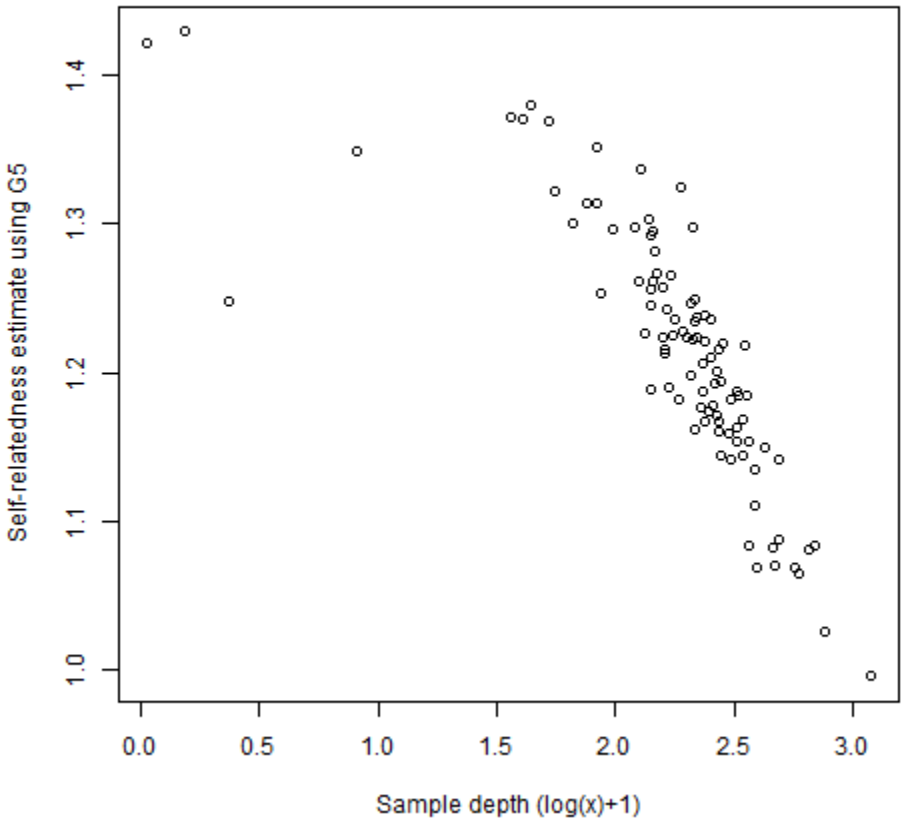


Gcompare.png

Comparison of diagonals (self-relatedness)

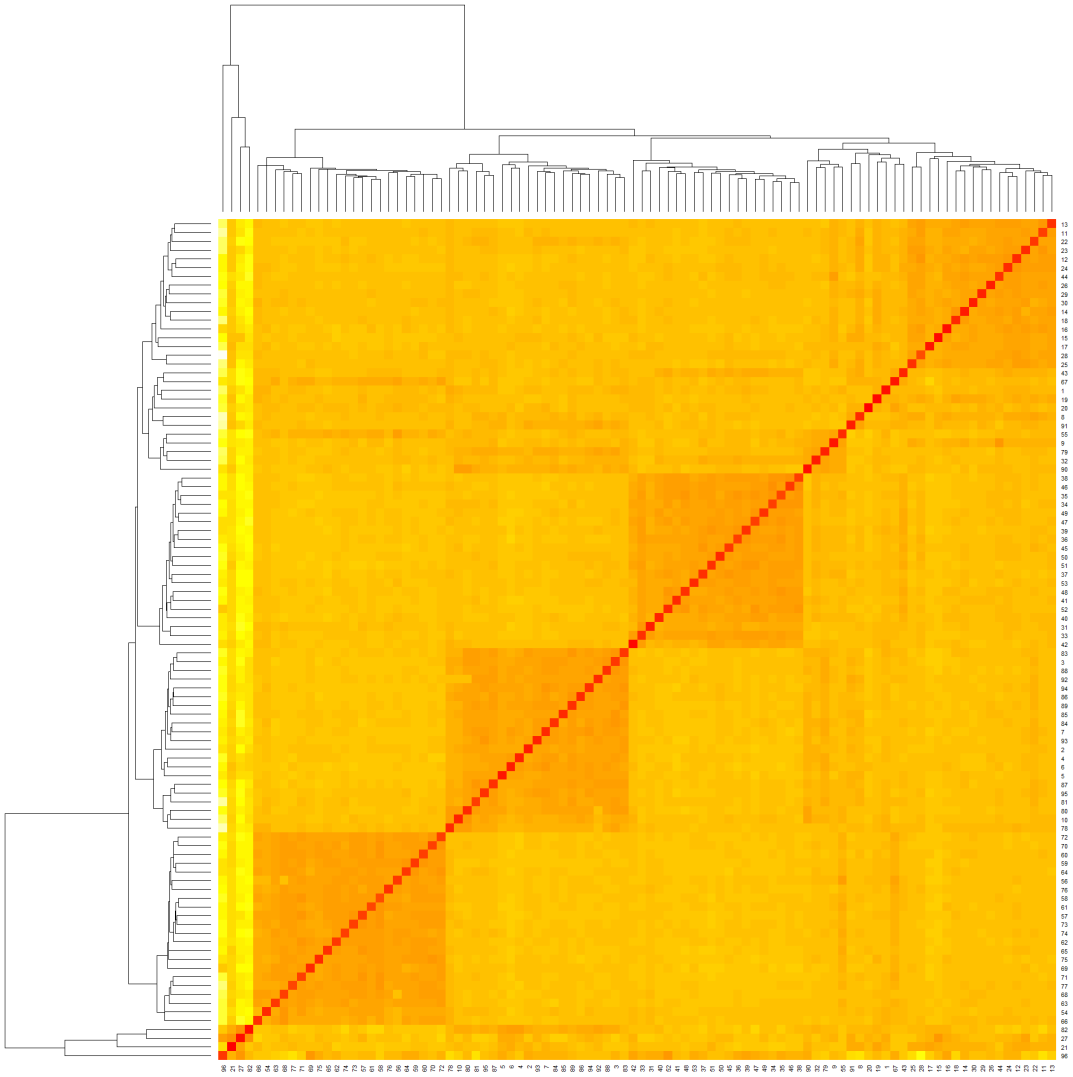


G-diag.png



Gdiagdepth.png

Heatmap



Heatmap-G5.png

calcG arguments

snpsubset	indices of SNPs to use
sfx	suffix for output file names
puse	allele frequencies to use in the calculations
indsubset	indices of individuals to include
depth.min	minimum depth for a genotype
depth.max	maximum depth for a genotype
npc	number of principal components to display

Calculate relatedness with SNP filter

Use Hardy-Weinberg disequilibrium cut-off of -0.05

```
GHWdgm.05 <- calcG(which(HWdis > -0.05),"HWdgm.05", npc=4)
```

```
Calculating G matrix, analysis code: HWdgm.05
# SNPs:  11500
# individuals:  96
Proportion of missing genotypes:  0.445558
Mean sample depth: 4.348901
Mean self-relatedness (G5 diagonal): 1.577778
minimum eigenvalue:  4.320261e-38
```

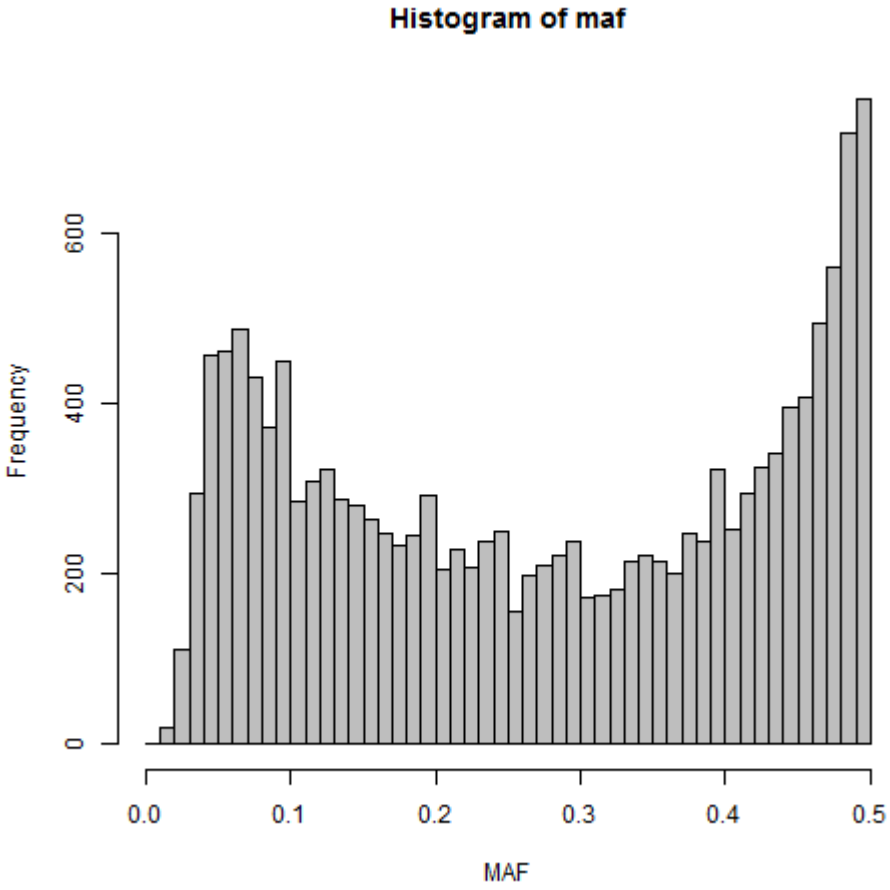
Previous
14709
96
0.364
9.28
1.22

Calculate relatedness with SNP filter

```
str(GHWdgm.05)
```

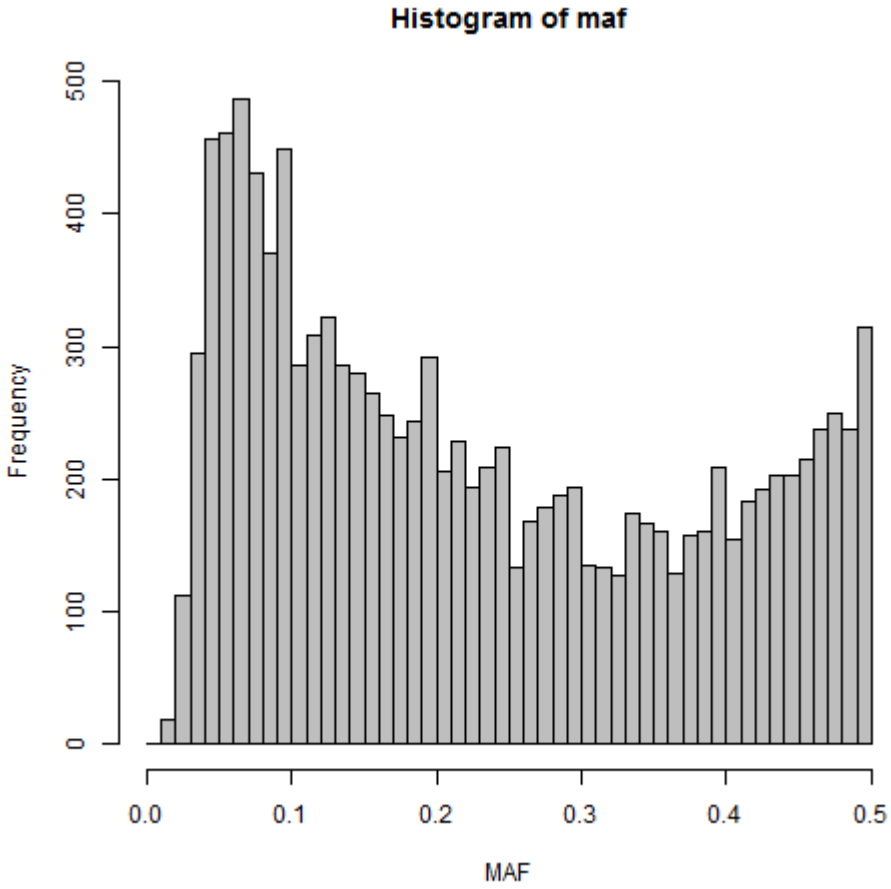
```
List of 4
 $ G1 : num [1:96, 1:96] 0.927 0.01208 0.01568 0.00432 -0.0069 ...
 $ G4d: num [1:96] 1.69 1.68 1.52 1.66 1.66 ...
 $ G5 : num [1:96, 1:96] 1.7133 0.0468 0.0366 0.0175 -0.0257 ...
 $ PC :List of 4
  ..$ d: num [1:96] 10.2 9.24 7.85 5.48 3.55 ...
  ..$ u: num [1:96, 1:4] -0.043 -0.11 -0.138 -0.195 -0.129 ...
  ..$ v: num [1:96, 1:96] -0.0503 -0.1164 -0.1449 -0.2032 -0.1359 ...
  ..$ x: num [1:96, 1:4] -0.438 -1.117 -1.409 -1.991 -1.318 ...
```

Comparison of MAFs



Unfiltered

MAF.png

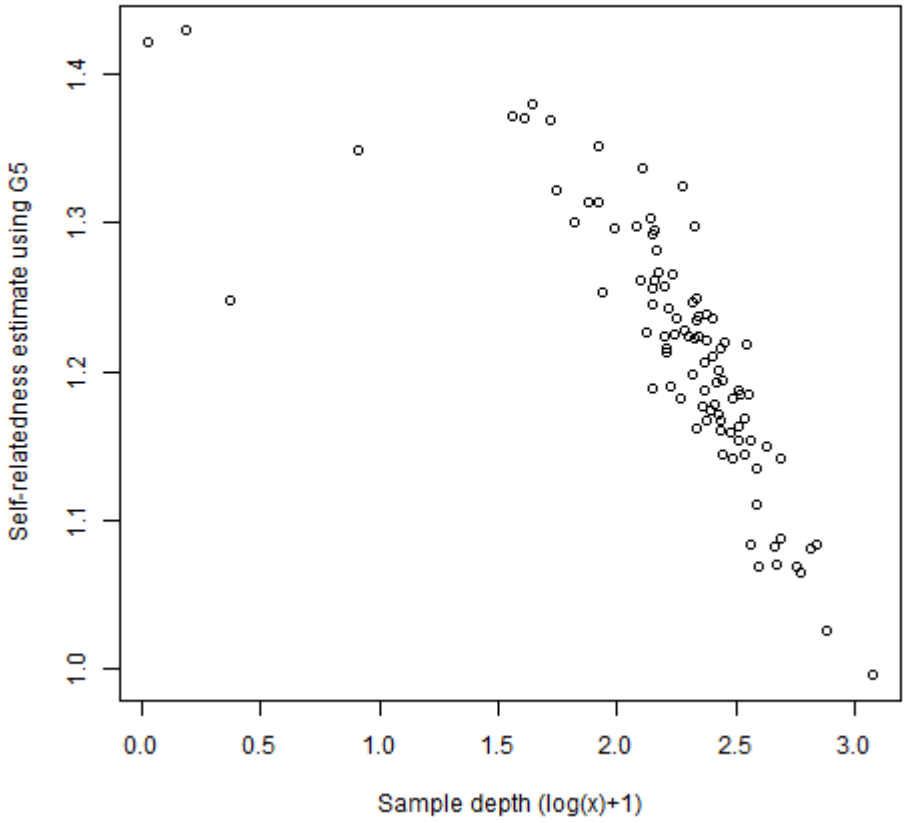


MAFHWdgm.05.png

Filtered

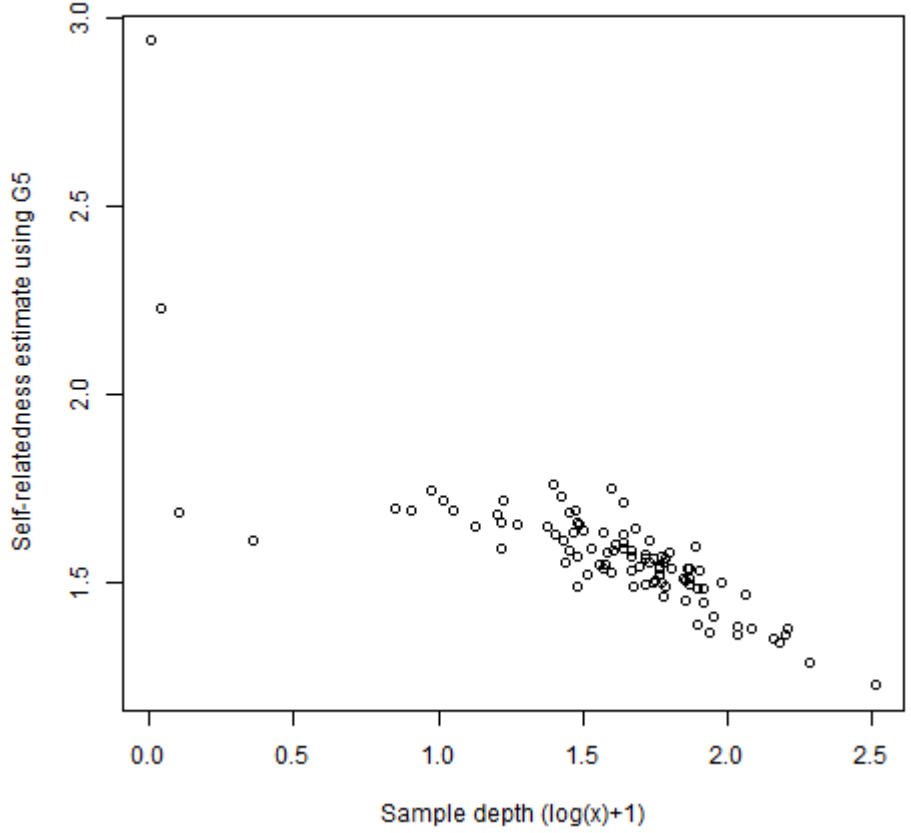
Comparison of diagonals (self-relatedness)

Unfiltered



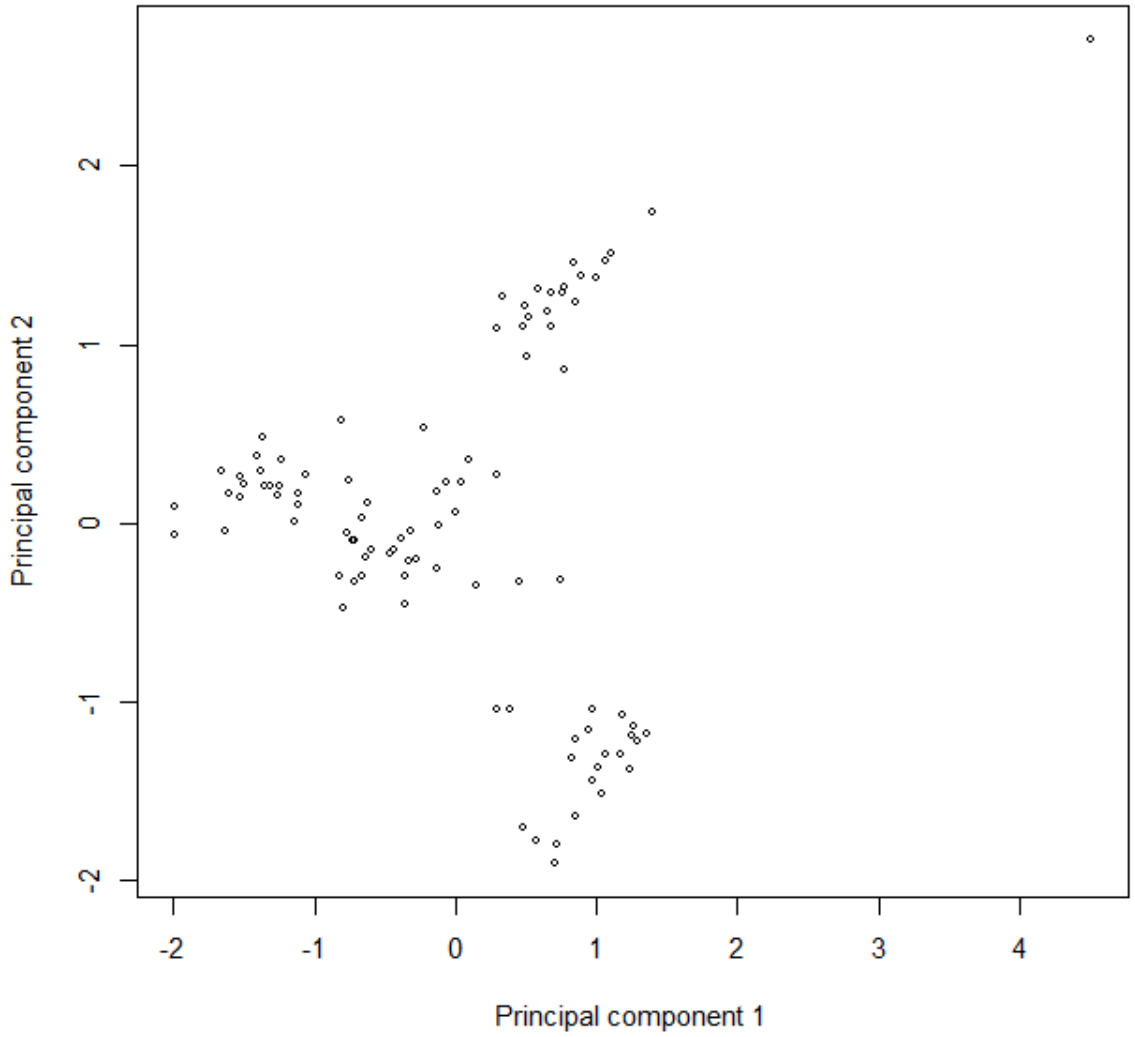
Gdiagdepth.png

Filtered



GHWdgm.05diagdepth.png

PCA



PC1v2G5HWdgm.05.png

Also see:
PCG5HWdgm.05.pdf

Check recorded pedigree

pedfile <- "Ped-GBS.csv"

Required		Ignored		Optional			
IndivID	seqID	Family	Relationship	FatherID	MotherID	Father Group	Mother Group
1	Seq1	1	Dam	5	3		
2	Seq2	1	Grand Dam				
⋮	⋮	⋮					
6	Seq6	1	Offspring	23	1		B
7	Seq7	1	Offspring	23	1		B
⋮	⋮	⋮					

Check recorded pedigree / Find parents

groupsfile <- "Ped-Groups.csv"

IndivID	ParGroup
1	B
24	B
48	B
72	B
2	B
⋮	⋮

Check recorded pedigree / Find parents

```
rel.thresh <- 0.2
GCheck <- "GHWdgm.05$G5"
source("../GBSPedAssign.R")
```

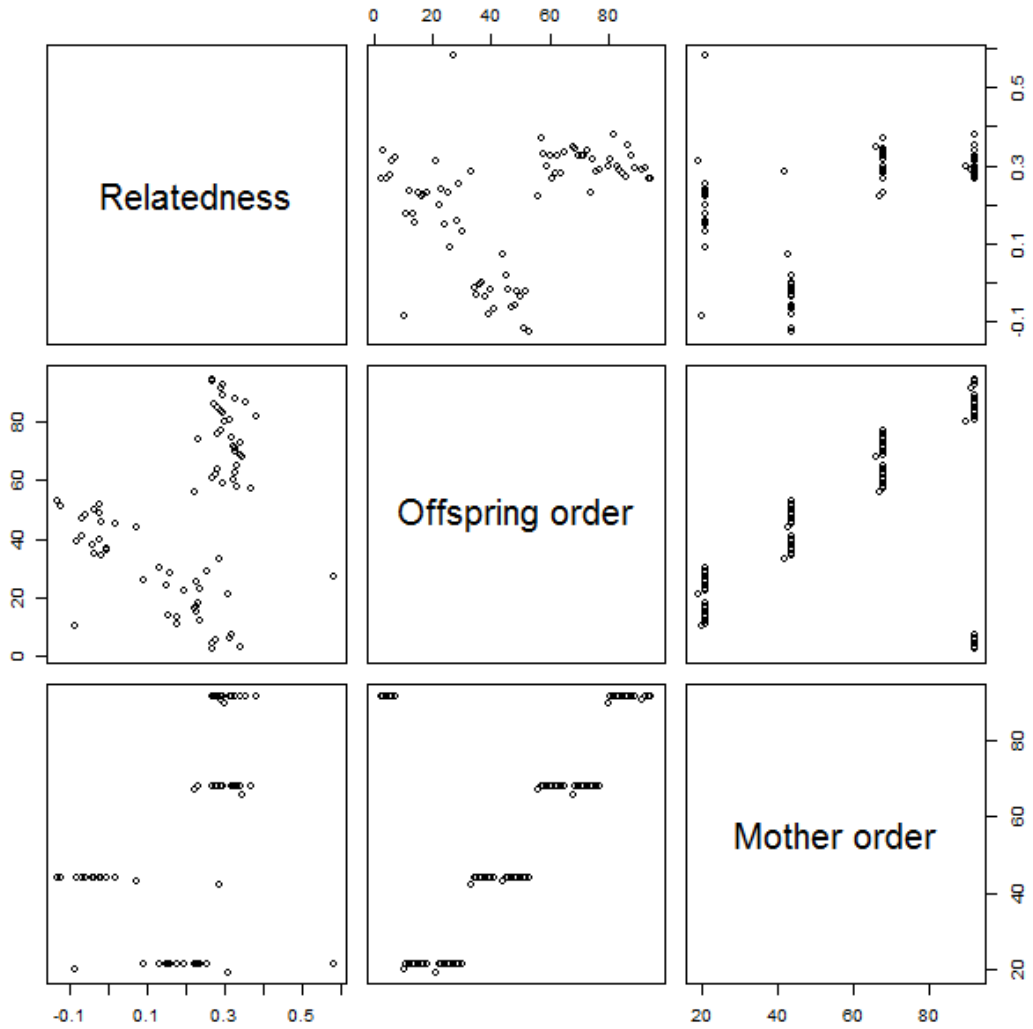
```
56 matches out of 78 Father comparisons: 71.8 %
Mean relatedness for Father matches 0.26
Mean relatedness for Father non-matches -0.0238
51 matches out of 78 Mother comparisons: 65.4 %
Mean relatedness for Mother matches 0.298
Mean relatedness for Mother non-matches 0.0181
Mean relatedness for full-sib families (as given)
  famfathers fammothers noffspring   meanrel
1         23         1        17 0.2286663
2         47        24        18 0.2653806
3         71        48        18 0.2921438
4         94        72        17 0.2499820
Mean relatedness within all full-sib families 0.2596066
Mean relatedness between individuals in full-sib families with different parents
-0.04724106
```

Check recorded pedigree

PedVerify.csv

IndivID	seqID	…	FatherRel	FatherMatch	MotherRel	MotherMatch
1	Seq1	…	0.01331	FALSE	0.071635	FALSE
2	Seq2		NA	NA	NA	NA
⋮	⋮					
6	Seq6		0.232647	TRUE	-0.01976	FALSE
7	Seq7		0.243087	TRUE	-0.06887	FALSE
⋮	⋮					

Check recorded pedigree



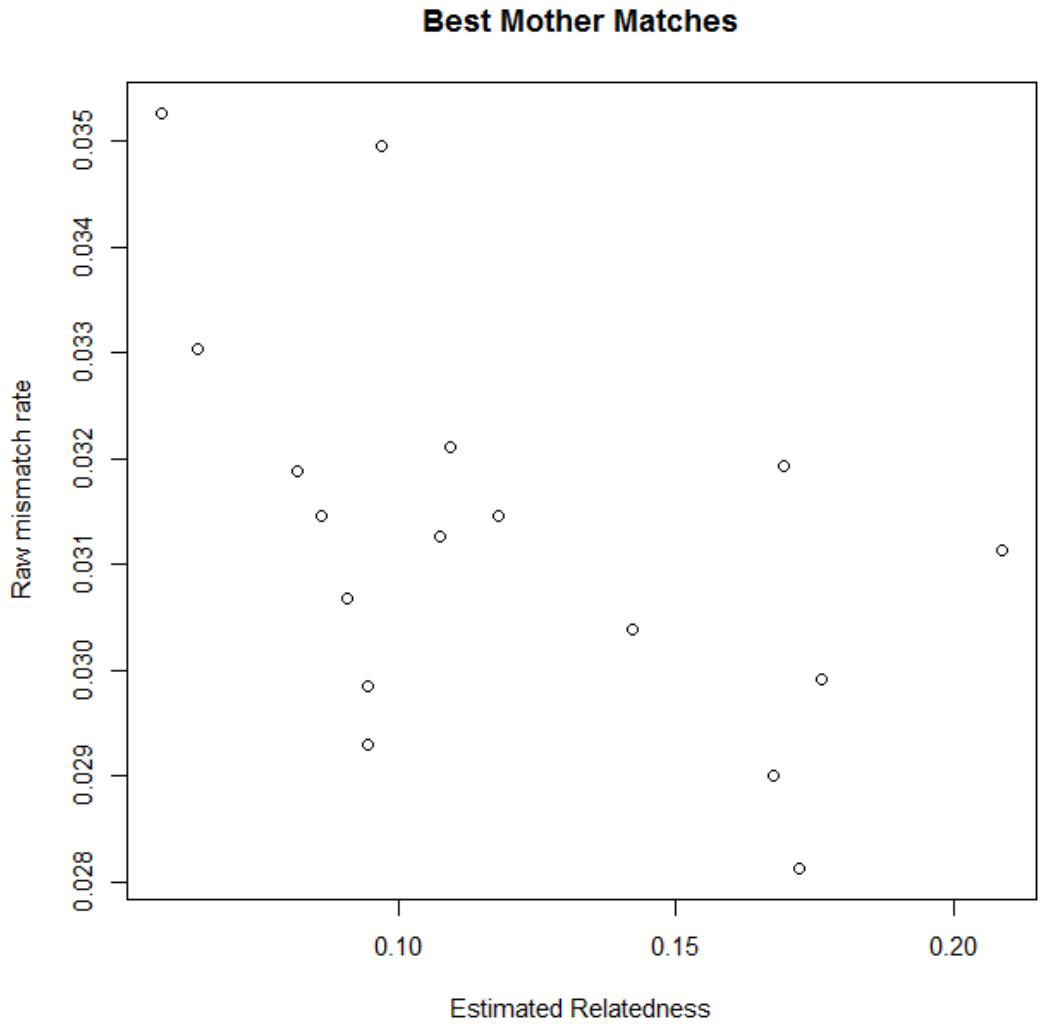
MotherVerify.png

Find parents

MotherMatches.csv

IndivID	seqID	BestMother Match	MotherMatch 2nd	Motherrel	Motherrel 2nd	mmrateMother
6	Seq6	2	3	0.107415	0.033553	0.031261
7	Seq7	2	3	0.057026	0.04448	0.035264
8	Seq8	2	3	0.167674	0.041685	0.029005
9	Seq9	3	2	0.081462	0.023037	0.031874
⋮	⋮					

Find parents



BestMotherMatches.png

Find parents

GroupsParentCounts.csv

IndivID	ParGroup	FatherFreq	MotherFreq
1	B	NA	NA
2	B	NA	13
3	B	NA	4
4	A	NA	NA
5	A	NA	NA
⋮	⋮		

Changing the default settings

gform	genotype input format: “uneak” (default), “Tassel” or “Chip”
sampdepth.thresh	Minimum mean sample depth (0.01)
snppdepth.thresh	Minimum mean SNP depth (0.01)
hirel.thresh	Threshold for reporting highly related pairs (0.9)

Exercise: Try a different threshold for sample depth

Customising

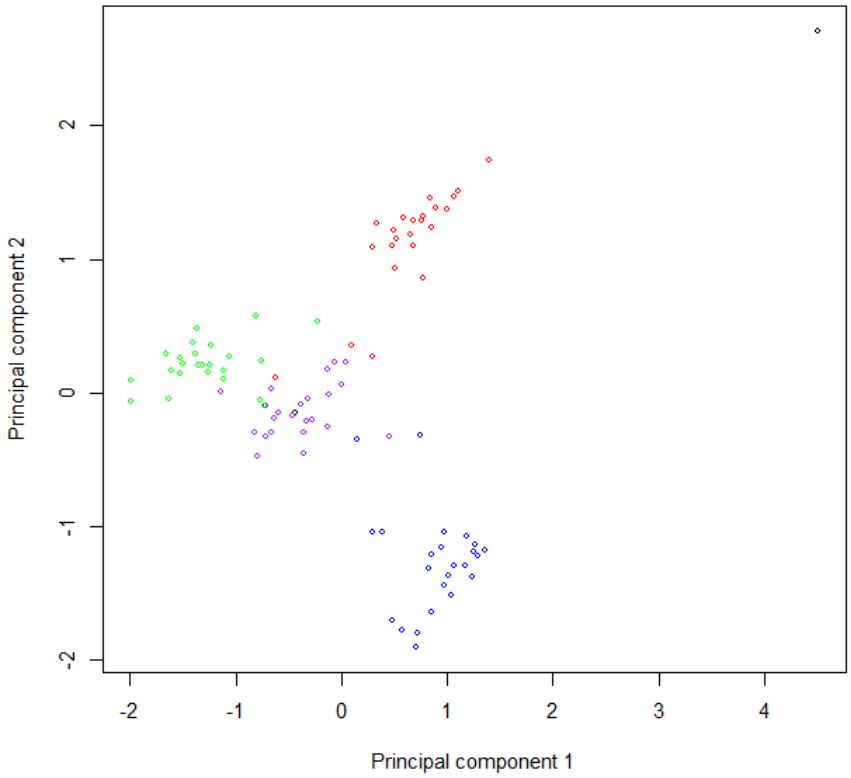
fcolo	“family colour” – colours to use for each individual in plots (black)
pedinfo	pedigree file (after using GBSPedAssign.R)

Exercises:

- Colour points in the PCA by family
- Redo the PC2 vs PC1 plot outside of calcG, with different symbols, sizes etc
- Plot the G matrix ordered by family (using image)

Customising
Colour points in the PCA by family

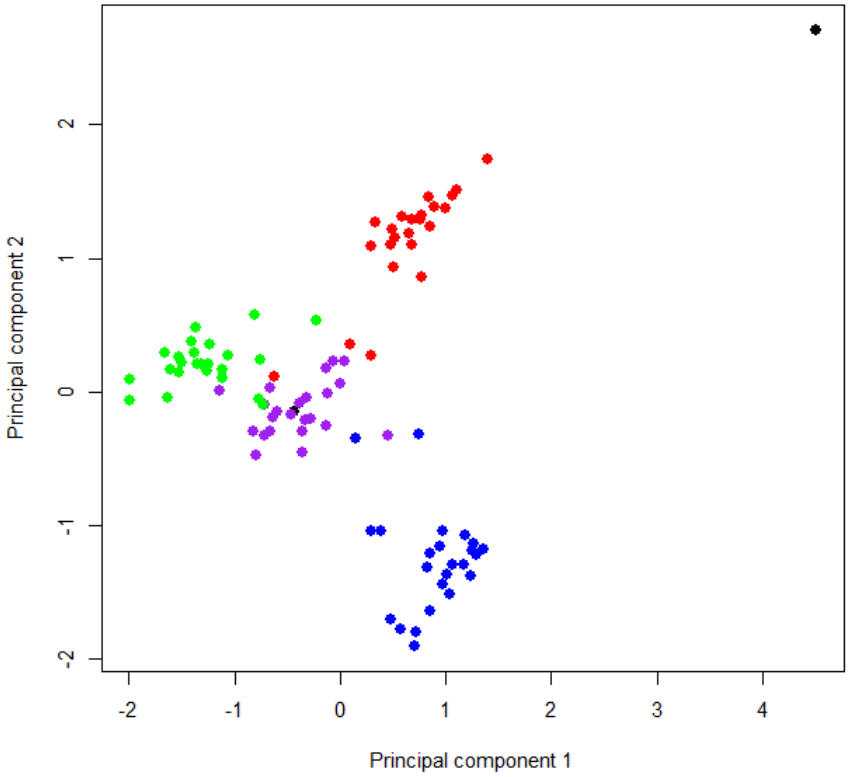
```
pedpos <- match(seqID,pedinfo$seqID)
fcolo <- c("red","blue","green","purple")[pedinfo$Family[pedpos]]
fcolo[is.na(fcolo)] <- "black"
GHWdgm.05 <- calcG(which(HWdis > -0.05),"HWdgm.05", npc=4)
```



PC1v2G5HWdgm.05.png

Customising
Redo the PC2 vs PC1 plot

```
png(paste0("PC1v2G5","a",".png"),width = 640, height = 640, pointsize=15)
   with(GHWdgm.05, plot(PC$x[,2] ~ PC$x[,1],cex=1.2,pch=16,col=fcolo,,
        xlab="Principal component 1",ylab="Principal component 2"))
   dev.off()
```

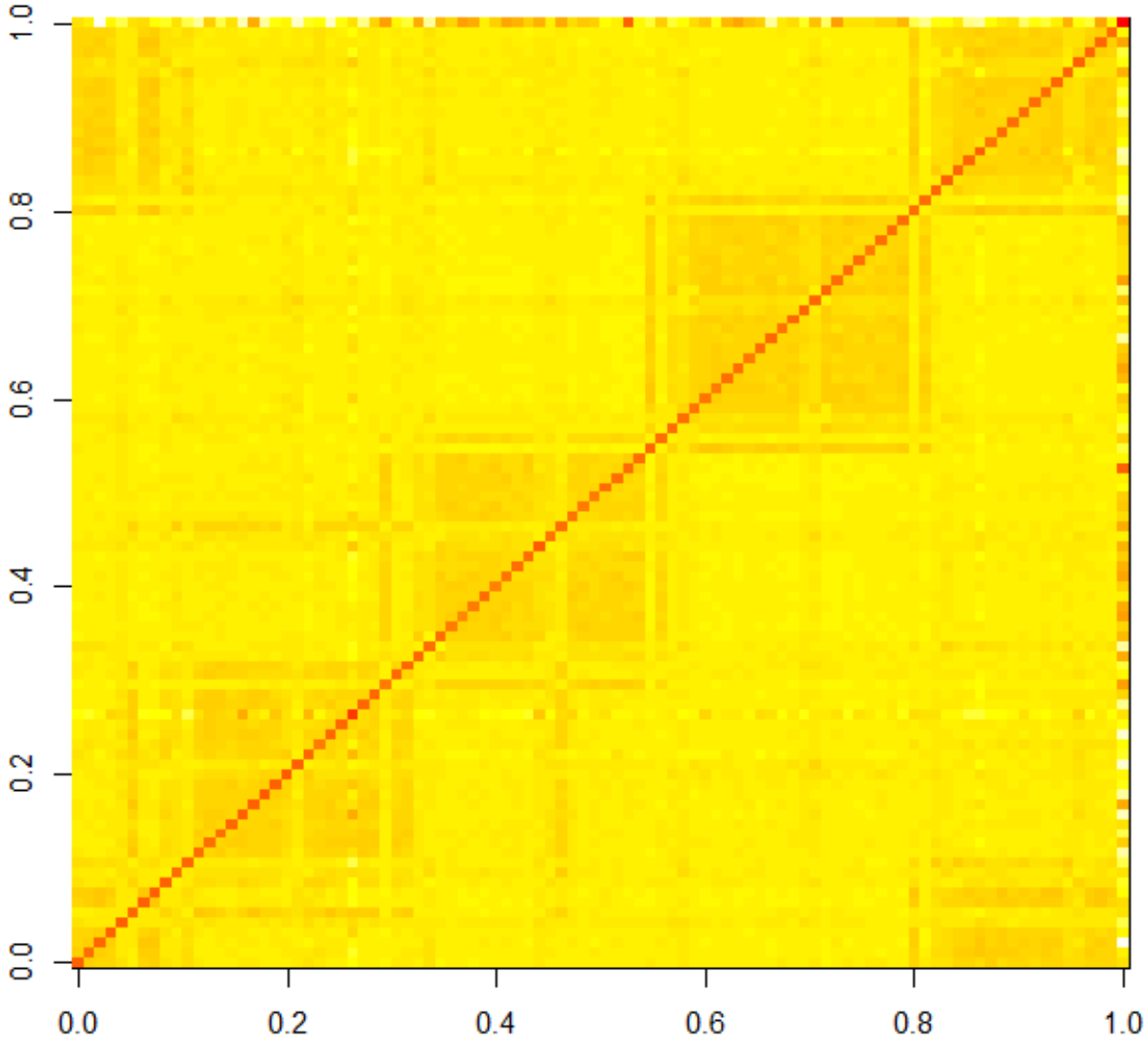


PC1v2G5a.png

Customising
Plot the G matrix ordered by family

```
relorder <- match(pedinfo$Relationship,c("Grand Dam","Grand
        Sire","Dam","Sire","Offspring","negative","reference"))
pedord <- order(pedinfo$Family,relorder)
png("GHWdgm.05.png",width = 640, height = 640, pointsize=15)
 with(GHWdgm.05, image(G5[pedord ,pedord ],col=rev(heat.colors(50))))
 dev.off()
```

Customising
Plot the G matrix ordered by family



GHWdgm.05.png

Resources

Dodds *et al.* (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *(BMC Genomics, submitted)*

- Simulations
- Worked example (Atlantic salmon)

Preprint

http://www.biorxiv.org/content/early/2015/08/24/025379

R code:

https://github.com/AgResearch/KGD

Acknowledgements

Genomics for Production &
Security in a Biological Economy

Ministry of Business, Innovation & Employment

Invermay GBS team:
Shannon Clarke, John McEwan,
Rudi Brauning, Rayna Anderson,
Tracey van Stijn, Rachael Ashby

agresearch

mapnet

Your species here