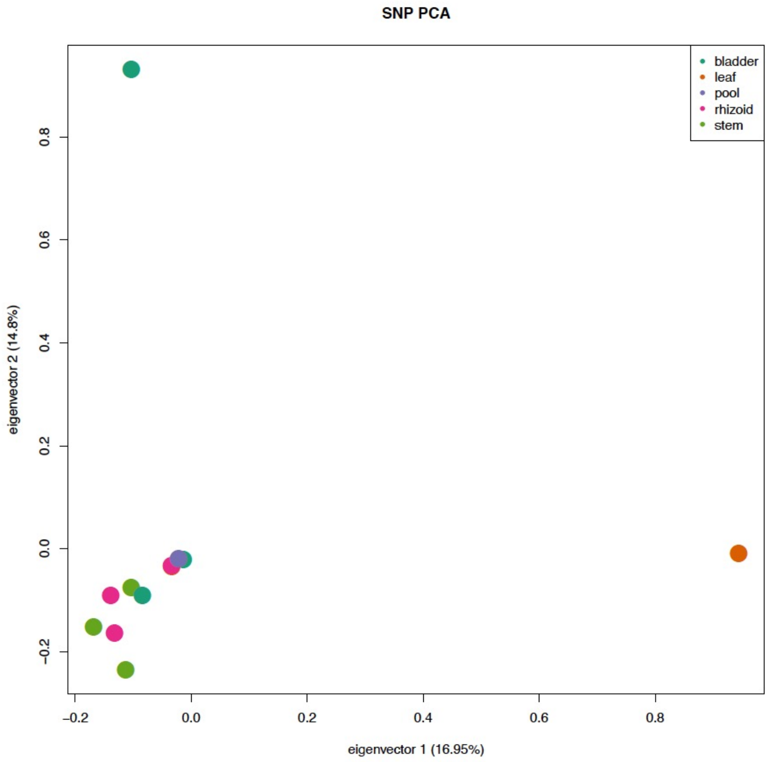
SNP Calling

Steps in module

- Map transcriptome reads to reference genome
  - BWA MEM
- Call SNPs with Stacks
  - Perl wrapper ref_map.pl
- Filter with VCFtools
  - Manual filtering to remove poor individuals and poor loci
- PCA using filtered SNP data
  - SNPRelate package in R



Step 1. Mapping reads to genome

```
cd /scratch/

mkdir SNP_calling_exercise

cd SNP_calling_exercise
```

```
cp /scratch/Botany2020NMGWorkshop/Genome_assembly/Completed_assemblies/Ugibba_pruned_assembly.fasta .
```

```
cp /scratch/Botany2020NMGWorkshop/raw_data/Ugibba/subset_fastq/*fastq.gz .

cp /scratch/Botany2021NMGWorkshop/6.DownstreamAnalyses/SNP_Calling_Scripts/BWA_Ugibba_afp.sh .
```

```
/opt/bwa-0.7.17/bwa index Ugibba_pruned_assembly.fasta
```

```
sudo chmod u+x /scratch/SNP_calling_exercise/BWA_Ugibba_afp.sh
```

```
/scratch/SNP_calling_exercise/BWA_Ugibba_afp.sh
```

Step 1. Mapping reads to genome

What is this script doing?
Let’s look inside the script:

```
less -S /scratch/SNP_calling_exercise/BWA_Ugibba_afp.sh
```

Step 2. Call SNPs with Stacks

```
sudo emacs /opt/stacks-2.55/scripts/ref_map.pl
```

Find the following line:

```
my $exe_path     = "_BINDIR_";
```

Change the line to:

```
my $exe_path     = “/opt/stacks-2.55";
```

```
cd /scratch
mkdir ref_SNPs
```

```
cp /scratch/Botany2021NMGWorkshop/6.DownstreamAnalyses/population_map_subset.txt ref_SNPs/
```

```
/opt/stacks-2.55/scripts/ref_map.pl --samples /scratch/SNP_calling_exercise/sorted_bam --popmap
ref_SNPs/population_map_subset.txt -o ref_SNPs/ -T 4
```

```
/opt/stacks-2.55/populations --batch-size 1 -P ref_SNPs/ -M ref_SNPs/population_map_subset.txt -t 4 --ordered-export --vcf
```

Step 3. Filter SNPs

For filtering, we’ll install VCFtools:

```
cd /scratch
mkdir Installed_programs
cd Installed_programs
git clone https://github.com/vcftools/vcftools.git

cd vcftools
./autogen.sh
./configure --prefix=/scratch/Installed_programs/vcftools
make
make install

/scratch/Installed_programs/vcftools/bin/vcftools --help
```

Step 3. Filter SNPs

Then, we'll use VCFtools to filter our raw VCF from Stacks:

```
cd /scratch/SNP_calling_exercise/

cp /scratch/Botany2020NMGWorkshop/Genome_assembly/SNP_calling/populations.snps.vcf .

/scratch/Installed_programs/vcftools/bin/vcftools --vcf populations.snps.vcf --max-missing 0.6 --min-meanDP 3 --max-meanDP 100 --maf 0.05
  --mac 3 --recode --recode-INFO-all --out Ugibba_SNPs_filtered

#initial pass throwing out all SNPs that are missing 60%
/scratch/Installed_programs/vcftools/bin/vcftools --vcf populations.snps.vcf --max-missing 0.4 --min-alleles 2 --max-alleles 2 --recode
  --recode-INFO-all --out Ugibba_first_pass
```

Step 3. Filter SNPs

```
#gives missing proportion of loci for each individual
/scratch/Installed_programs/vcftools/bin/vcftools --vcf Ugibba_first_pass.recode.vcf --missing-indv
```

```
#average depth for each individual
/scratch/Installed_programs/vcftools/bin/vcftools --vcf Ugibba_first_pass.recode.vcf --depth
```

```
#observed and expected heterozygosity
/scratch/Installed_programs/vcftools/bin/vcftools --vcf Ugibba_first_pass.recode.vcf --het
```

Step 3. Filter SNPs

#Create a list of individuals with at least 50% missing data

```
awk '$5 > 0.50' out.imiss | cut -f1 > lowDP50.indv
```

```
/scratch/Installed_programs/vcftools/bin/vcftools --vcf Ugibba_first_pass.recode.vcf --max-missing
0.4 --remove lowDP50.indv --recode --recode-INFO-all --out Ugibba_filtered_SNPs
```

Step 4. PCA in R

```
# Make sure we are in SNP_calling_exercise:
cd /scratch/SNP_calling_exercise/
```

```
cp /scratch/Botany2021NMGWorkshop/6.DownstreamAnalyses/SNP_Calling_Scripts/SNPRelate_afp.R .
```

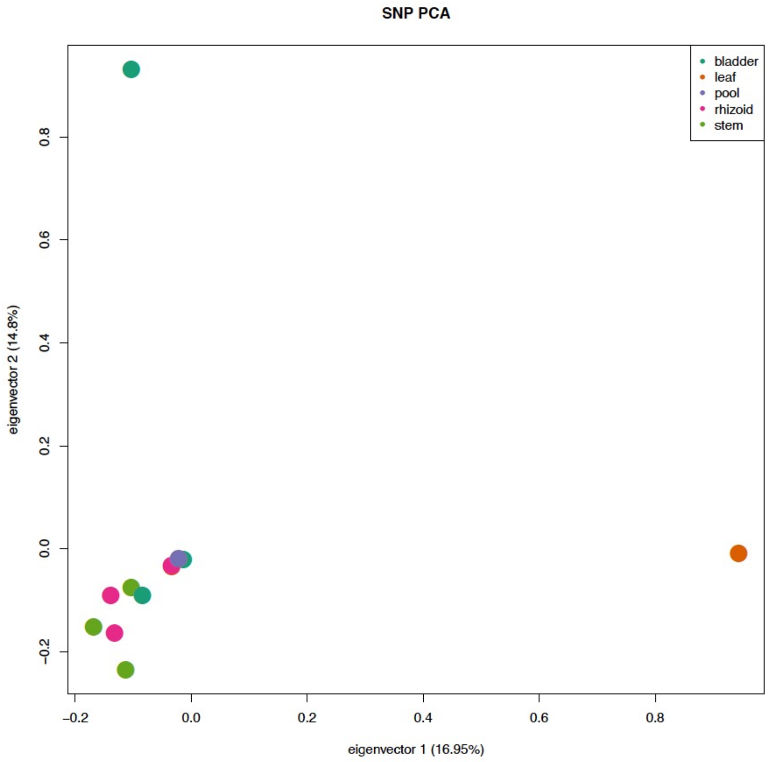
```
cp /scratch/Botany2020NMGWorkshop/Genome_assembly/SNP_calling/Ugibba_filtered_SNPs_renamed.recode.vcf .
```

Step 4. PCA in R

```
R
source("http://bioconductor.org/biocLite.R")
Would you like to use a personal library instead?  (y/n) y
Would you like to create a personal library
~/R/x86_64-pc-linux-gnu-library/3.4
to install packages into?  (y/n) y

biocLite("SNPRelate")
install.packages("RColorBrewer")

> quit()
Save workspace image? [y/n/c]: y

Rscript SNPRelate_afp.R
```

Steps in module

- Map transcriptome reads to reference genome
  - BWA MEM
- Call SNPs with Stacks
  - Perl wrapper ref_map.pl
- Filter with VCFtools
  - Manual filtering to remove poor individuals and poor loci
- PCA using filtered SNP data
  - SNPRelate package in R


SNP PCA
bladder
leaf
pool
rhizoid
stem
eigenvector 2 (14.8%)
eigenvector 1 (16.95%)