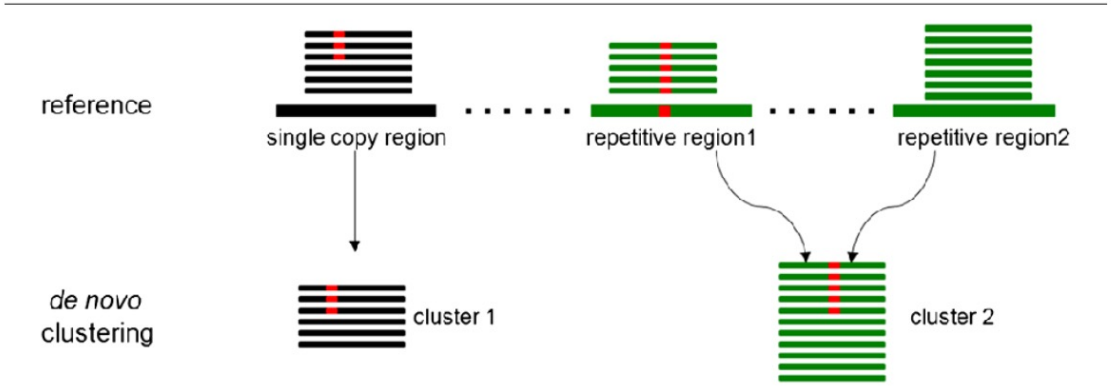
SNP Calling

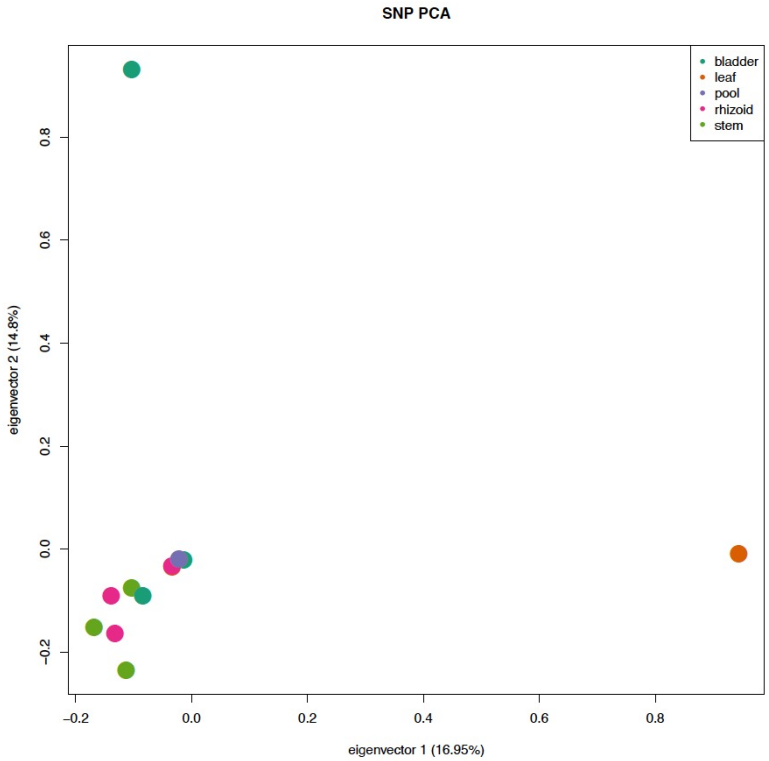
Population genetic studies

- Many options and different requirements for SNP calling
- RAD-Seq
  - Stacks, both reference guided or *de novo*
- Hyb-Seq, RNA-Seq or Whole Genome Sequencing
  - GATK, Freebayes, mpileup
  - Must have a reference genome or at the very least something to map reads to



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



Installing bwa (if necessary)

```
sudo apt update
sudo apt install bwa
```

Step 1. Mapping reads to genome

cd /scratch/Botany2020NMGWorkshop/Genome_assembly/SNP_calling

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

bwa index Ugibba_pruned_assembly.fasta

Step 1. Mapping reads to genome

/scratch/Botany2020NMGWorkshop/Genome_assembly/scripts/BWA_Uggiba.sh

What is this script doing?
An example of the 'bwa mem' command:

```
bwa mem -t 6 -R "@RG\tID:bladderR1\tSM:Rep1\tPL:HiSeq\tPU:HTNMKDSXX\tLB:RNA-Seq"
Ugibba_pruned_assembly.fasta
/scratch/Botany2020NMGWorkshop/raw_data/Ugibba/transcriptome/Ugibba_bladderR1.fastq.gz
> Ugibba_bladderR1.sam
```

Step 2. Call SNPs with Stacks

```
cd /scratch
mkdir ref_SNPs
```

```
/opt/stacks-2.55/scripts/ref_map.pl --samples /scratch/Botany2020NMGWorkshop/Genome_assembly/sorted_bam
 --popmap population_map.txt -o ref_SNPs/ -T 6
```

```
/opt/stacks-2.55/populations --batch_size 1 -P ref_SNPs/ -M population_map.txt -t 6 --ordered_export --vcf
```

Step 3. Filter SNPs

First, we'll install VCFtools:

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

./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/Botany2020NMGWorkshop/Genome_assembly/SNP_calling/ref_SNPs

/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

cp /scratch/Botany2020NMGWorkshop/Genome_assembly/scripts/SNPRelate.R .

Rscript SNPRelate.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