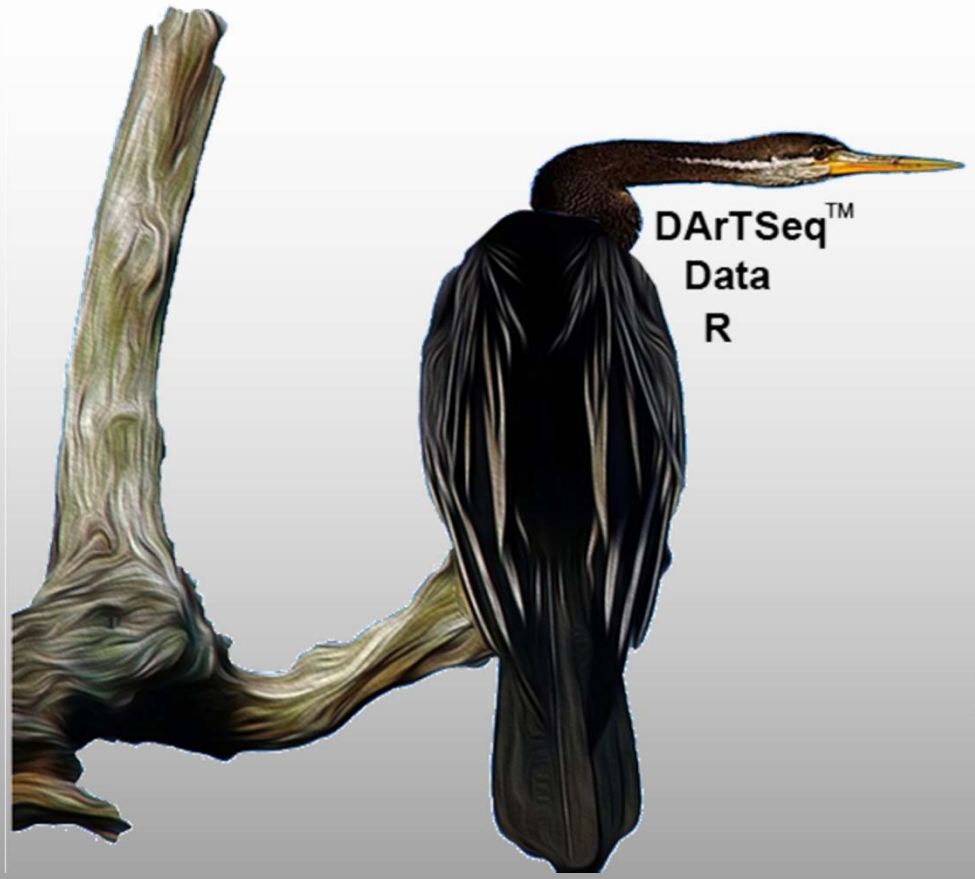


Analysing SNP data with dartR

Bernd Gruber and Arthur Georges
Institute for Applied Ecology
University of Canberra

Figure 1-1. R as it appears when it first starts. The R Console window and two other windows are visible. The Program Editor and R Graphics windows do not appear until required.

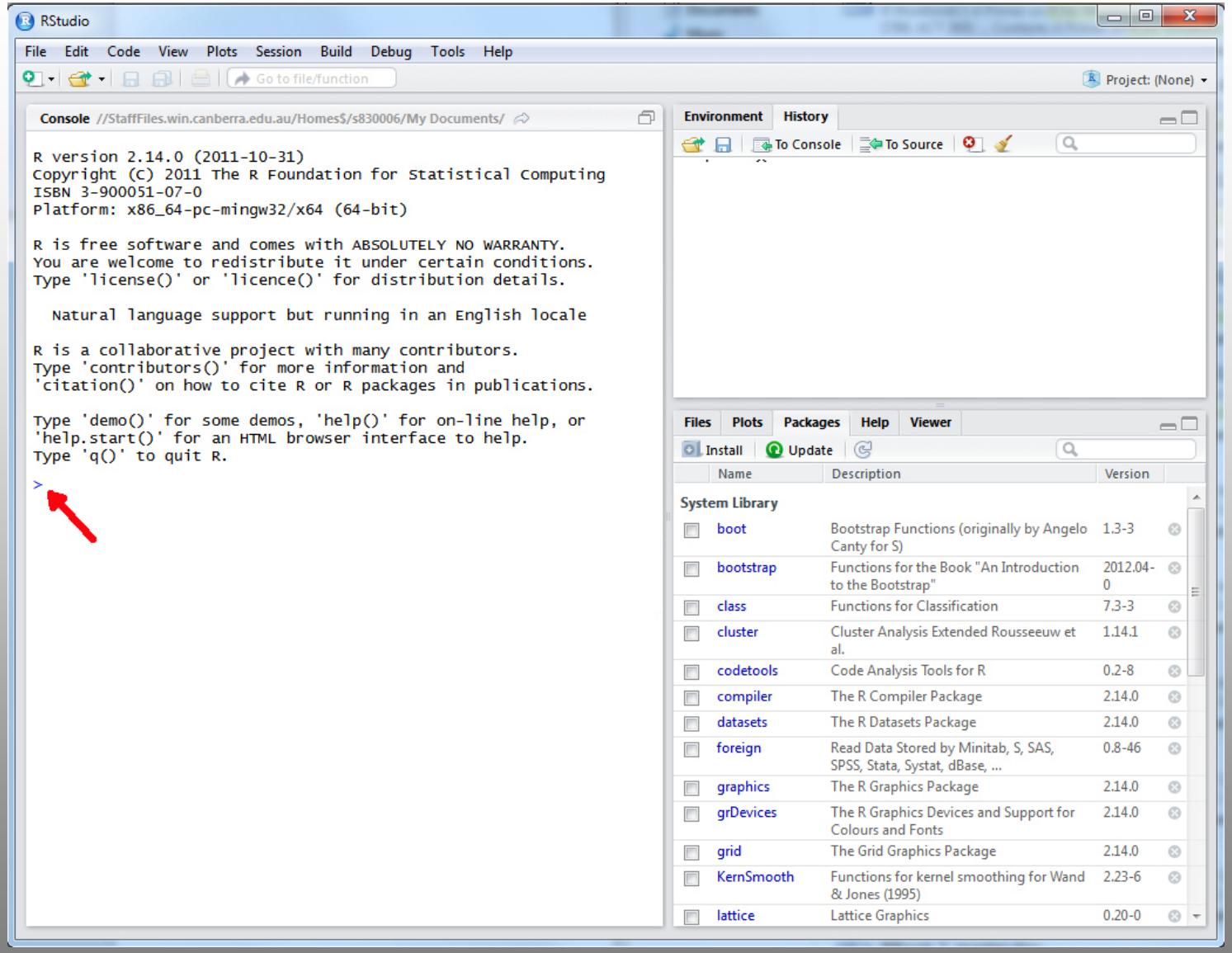


Figure 1-2. R as it appears after opening a new script window. The R Program Editor, Console window and two other windows are visible.

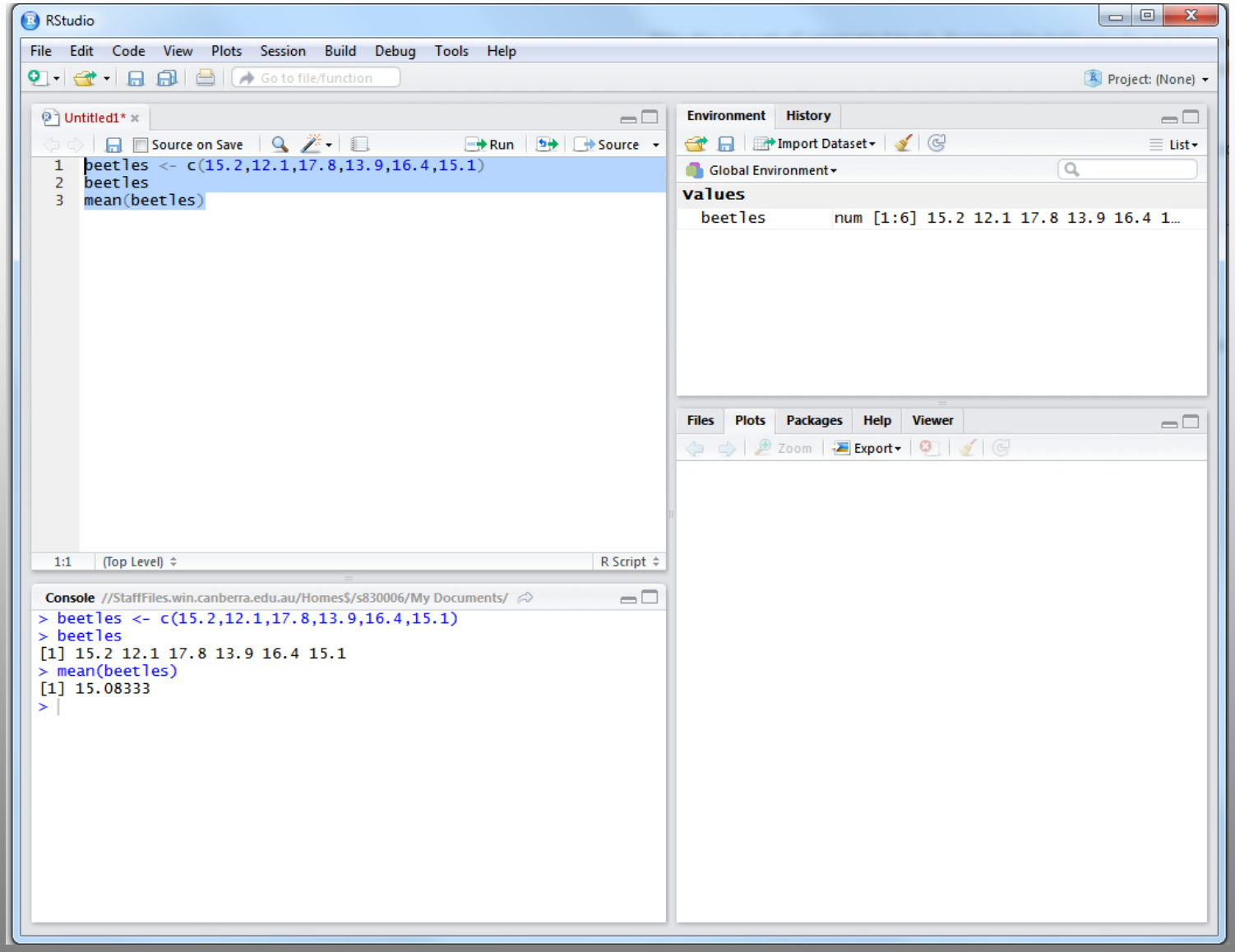


Figure 1-3. A diagram showing the steps in reduced genome representation and genotyping by sequencing.

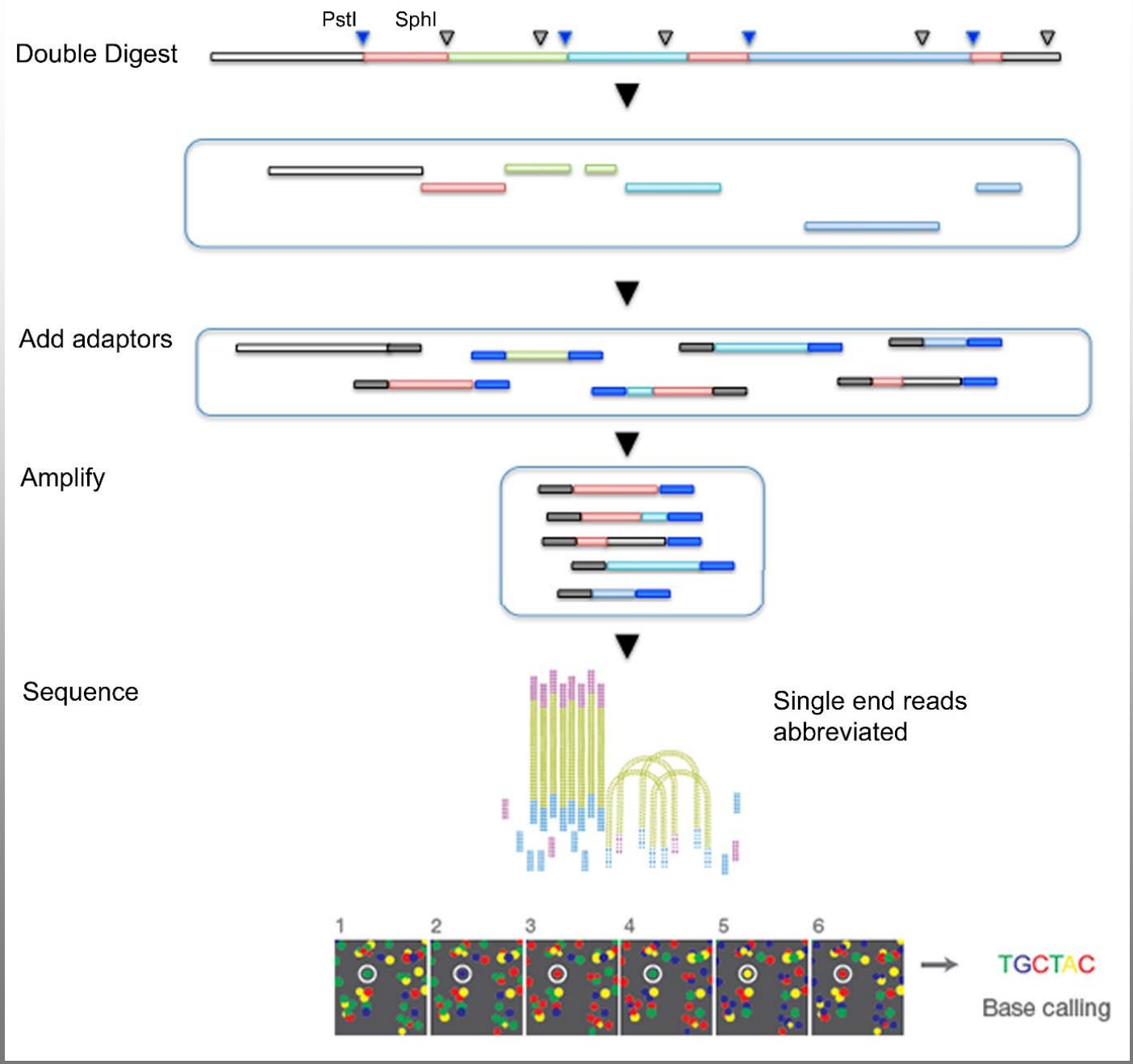


Figure 1-3. A diagram showing a single nucleotide polymorphism.

	Ind 01	Ind 02	Ind 03	Ind 04	Ind 05	Ind 06	Ind 07	Ind 08	Ind 09	Ind 10
Locus 1	A/A	A/A	A/A	A/A	A/G	A/A	A/A	A/A	A/A	-/-
Locus 2	C/C	C/C	C/C	C/C	C/C	C/C	C/T	C/C	C/C	C/C
Locus 3	C/G	G/G	G/G	G/G	G/G	C/C	C/C	C/C	C/C	C/C
Locus 4	A/A	A/T	A/A	A/T	T/T	A/A	A/A	A/A	A/A	A/A
Locus 5	A/A	A/A	A/A	A/A	-/-	A/G	A/A	A/A	A/A	A/A
Locus 6	C/C	C/C	C/C	C/C	C/C	C/C	C/T	C/C	C/C	C/C
Locus 7	C/G	G/G	G/G	G/G	G/G	C/C	C/C	C/C	C/C	C/C
Locus 8	A/A	A/T	A/A	A/T	T/T	A/A	A/A	A/A	A/A	A/A
Locus 9	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
Locus 10	C/C	C/C	C/C	C/C	C/C	C/C	C/T	C/C	C/C	C/C
Locus 11	C/G	G/G	G/G	G/G	G/G	C/C	C/C	C/C	C/C	C/C

Installing from CRAN

```
install.packages("dartR")
library(dartR)
```

Installing from GitHub (latest stable)

```
install.packages("devtools")
library(devtools)
install_github("green-striped-gecko/dartR")
library(dartR)
```

Preliminary pipeline

Typical Data Entry Pipeline

- Examine the data provided by DArT PL in Excel
- Read the data into dartR
  gl <- gl.read.dart(…..)
- Correct any errors
- Examine the final dataset
  nLoc, nPop, nLoc ……
- Filter (CallRate, RepAvg, MAF, monomorphs etc)
- Recalculate the locus metrics
  gl <- gl.recalc.metrics(gl)
- Save the genlight object for future use
  saveRDS(gl, file="mygl.Rdata")

Genlight data object

Dataframe
LOCUS METADATA
AlleleID, CloneID, AlleleSequence, SNP, SnpPosition, CallRate, OneRatioRef, OneRatioSnp, FreqHomRef, FreqHomSnp, FreqHets, PICRef, PICSnp, AvgPIC, AvgCountRef, AvgCountSnp, RepAvg

Dataframe
INDIVIDUAL METADATA
Latitude Longitude
Maturity Sex

INDIVIDUALS / **LOCI**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AA010915	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0	0	1
UC_00126	2	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0
AA032760	0	0	-	0	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2
AA013214	0	2	0	0	0	2	2	0	0	0	1	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0
AA011723	0	2	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA012411	2	0	2	2	0	2	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0
AA019237	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0
AA019238	0	0	0	2	2	2	0	2	0	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2
AA019239	0	2	0	0	0	-	0	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0
AA019235	0	2	0	0	0	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0
AA019240	1	0	-	0	0	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0
AA019241	2	0	2	2	0	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0
AA019242	0	0	0	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2
AA019243	0	1	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0
AA019251	0	0	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2
AA019252	2	0	0	0	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA012405	2	-	0	1	0	0	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0
AA012406	0	0	0	2	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2
AA012409	0	0	2	0	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2
AA012499	0	2	2	2	2	0	0	0	2	2	0	0	0	1	-	-	2	0	0	0	2	2	0	0	0	1	-	2	-	2
AA012422	1	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0	0
AA012434	2	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA012469	0	0	0	2	2	2	0	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0
AA012500	2	0	1	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0	0	1
AA032799	2	0	0	2	2	2	1	2	0	0	2	0	0	0	2	0	2	2	2	1	1	0	0	0	2	2	0	0	0	1

0	Homozygous reference allele
1	Heterozygous
2	Homozygous alternate allele
-	Missing

Reading DArT PL data into a genlight object

Diversity Arrays Technology
IT'S A BULLSEYE WITH DArT

																		A5	B5	C5	D5	E5
																		9.01E+11	9.01E+11	9.01E+11	9.01E+11	9.01E+11
																		UC_1	UC_1	UC_1	UC_1	UC_1
AlleleID	CloneID	AlleleSequenc	SNP	SnpPosition	CallRate	OneRatioRef	OneRatioSnp	FreqHomRef	FreqHomSnp	FreqHets	PICRef	PICSnp	AvgPIC	AvgCountF	AvgCountS	RepAvg		AA010915	UC_00126	AA032760	AA013214	AA011723
100049687	20131003	TGCAGAAACA		12	0.98358	0.000982	0.999018	0.000982	0.999018	0	0.00196	0.00196	0.00196	22	11.32	1		0	0	-		0	0
100049687	20131003	TGCAGAAACA	12:A>G	12	0.98358	0.000982	0.999018	0.000982	0.999018	0	0.00196	0.00196	0.00196	22	11.32	1		1	1	-		1	1
100049698	20131004	TGCAGAAACC		16	0.44928	0.993548	0.030108	0.969892	0.006452	0.02366	0.01282	0.0584	0.03561	9.3304	7.8125	1		-	-	-	-	-	
100049698	20131004	TGCAGAAACC	16:C>T	16	0.44928	0.993548	0.030108	0.969892	0.006452	0.02366	0.01282	0.0584	0.03561	9.3304	7.8125	1		-	-	-	-	-	
100049728	20131006	TGCAGAAAGC		23	1	0.999034	0.016425	0.983575	0.000966	0.01546	0.00193	0.03231	0.01712	30.7213	17.6	1		1	1	1	1	1	1
100049728	20131006	TGCAGAAAGC	23:T>G	23	1	0.999034	0.016425	0.983575	0.000966	0.01546	0.00193	0.03231	0.01712	30.7213	17.6	1		0	0	0	0	0	0
100049805	20131010	TGCAGAAATT		56	0.99324	1	0.000973	0.999027	0	0.00097	0	0.00194	0.00097	7.0721	5	1		1	1	1	1	1	1
100049805	20131010	TGCAGAAATT	56:A>T	56	0.99324	1	0.000973	0.999027	0	0.00097	0	0.00194	0.00097	7.0721	5	1		0	0	0	0	0	0
100049816	11357138	TGCAGAAATT		51	0.98647	0.610186	0.400588	0.599412	0.389814	0.01077	0.47572	0.48023	0.47798	12.0158	9.17901	0.98995		1	1	1	1	1	1
100049816	11357138	TGCAGAAATT	51:C>T	51	0.98647	0.610186	0.400588	0.599412	0.389814	0.01077	0.47572	0.48023	0.47798	12.0158	9.17901	0.98995		0	0	0	0	0	0
100049839	20131014	TGCAGAACAA		39	0.90725	0.001065	0.998935	0.001065	0.998935	0	0.00213	0.00213	0.00213	7	4.32989	1		-		0	0	0	0
100049839	20131014	TGCAGAACAA	39:A>T	39	0.90725	0.001065	0.998935	0.001065	0.998935	0	0.00213	0.00213	0.00213	7	4.32989	1		-		1	1	1	1
100049926	20131016	TGCAGAACTG		33	0.89952	0.9087	0.105263	0.894737	0.0913	0.01396	0.16593	0.18837	0.17715	6.90047	4.47899	0.99327		1	1	1	1	1	1
100049926	20131016	TGCAGAACTG	33:C>T	33	0.89952	0.9087	0.105263	0.894737	0.0913	0.01396	0.16593	0.18837	0.17715	6.90047	4.47899	0.99327		0	0	0	0	0	0
100049990	20131018	TGCAGAAGAC		20	0.99614	0.974782	0.059166	0.940834	0.025218	0.03395	0.04917	0.11133	0.08025	7.90997	6	1		1	1	1	1	1	1
100049990	20131018	TGCAGAAGAC	20:G>T	20	0.99614	0.974782	0.059166	0.940834	0.025218	0.03395	0.04917	0.11133	0.08025	7.90997	6	1		0	0	0	0	0	0
100050079	11357397	TGCAGAAGGC		57	0.97778	1	0.000988	0.999012	0	0.00099	0	0.00197	0.00099	8.67041	8	1		1	1	-		1	1

`gl.read.dart()`

Dataframe
LOCUS METADATA
AlleleID, CloneID, AlleleSequence, SNP, SnpPosition, CallRate, OneRatioRef, OneRatioSnp, FreqHomRef, FreqHomSnp, FreqHets, PICRef, PICSnp, AvgPIC, AvgCountRef, AvgCountSnp, RepAvg

Dataframe
INDIVIDUAL METADATA
Latitude Longitude
Maturity Sex

INDIVIDUALS	LOCI																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AA010915	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0	0	1
UC_00126	2	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0
AA032760	0	0	-	0	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	0	2	1	1	2	2	2	2	0	0	2
AA013214	0	2	0	0	0	2	2	0	0	0	1	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0
AA011723	0	2	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA012411	2	0	2	2	0	2	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0
AA019237	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	0	2	2	0
AA019238	0	0	0	2	2	2	0	2	0	0	2	0	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2
AA019239	0	2	0	0	0	-	0	-	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0
AA019235	0	2	0	0	0	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0
AA019240	1	0	-	0	0	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	0	2	1	1	2	2	2	2	0	0
AA019241	2	0	2	2	0	0	2	0	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0
AA019242	0	0	0	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2
AA019243	0	1	2	0	0	2	1	2	2	2	0	0	2	0	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA019251	0	0	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2
AA019252	2	0	0	0	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA012405	2	-	0	1	0	0	2	0	0	2	0	0	2	1	2	2	2	0	0	0	2	0	0	0	2	1	1	2	2	0
AA012406	0	0	0	2	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	0	2	1	1	2	2	2	2	0	0	2
AA012409	0	0	2	0	2	2	0	0	2	1	2	2	2	0	0	2	0	0	0	0	2	1	1	2	2	2	2	0	0	2
AA012499	0	2	2	2	2	0	0	0	2	2	0	0	0	1	-	-	2	0	0	0	2	2	0	0	0	1	-	2	-	2
AA012422	1	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0	0
AA012434	2	2	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0
AA012469	0	0	0	2	2	2	0	0	2	0	0	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0
AA012500	2	0	1	2	1	2	2	2	0	0	2	0	0	0	2	1	1	2	2	2	2	0	0	0	2	2	0	0	0	1
AA032799	2	0	0	2	2	2	1	2	0	0	2	0	0	0	2	0	2	2	2	1	1	0	0	0	2	2	0	0	0	1

0	Homozygous reference allele
1	Heterozygous
2	Homozygous alternate allele
-	Missing

Basic genlight commands

`nLoc(gl)`	number of loci
`locNames(gl)`	list of loci
`nInd(gl)`	number of individuals (specimens or samples)
`indNames(gl)`	list of individuals
`nPop(gl)`	number of populations
`popNames(gl)`	list of populations
`pop(x)`	list of population assignments for each individual
`as.matrix(gl)`	generate a matrix of the SNP scores, with 0 as homozygous reference, 2 as homozygous alternate, and 1 as heterozygous.
`glPlot(gl)`	a smear plot of individual against locus, useful for gross pattern identification and assessment of allelic dropout

gl[1:5,1:10] behaves like a data matrix

Basic genlight functions

`gl <- gl.drop.pop( )` remove listed populations from gl

`gl <- gl.keep.pop( )` keep only the listed populations

`gl <- gl.drop.ind( )` remove listed individuals

`gl <- gl.keep.ind( )` keep only the listed individuals

`gl <- gl.recalc.metrics( )` recalculate locus metrics,

`gl <- gl.filter.monomorphs( )` remove monomorphic loci, including all NAs

`gl <- gl.define.pop( )` create a new population for listed individuals

`gl <- gl.merge.pop( )` merge two populations under a new name, or if applied to one population, to rename it.

These scripts manage things in the background

Basic genlight functions

`gl <- gl.make.recode.pop( )`	make a recode table based on existing population labels. You will need to edit the second column of the recode table to specify the new labels to apply.
`gl <- gl.make.recode.ind( )`	make a recode table based on existing individual labels. You will need to edit the second column of the recode table to specify the new labels to apply. Individuals assigned the new label 'Delete' will be removed from the genlight object.
`gl <- gl.recode.pop( )`	apply the specified pop.recode table to the populations
`gl <- gl.recode.ind( )`	apply the specified ind.recode table to the individuals
`gl <- gl.edit.recode.pop( )`	edit population assignments, and apply the changes on closure
`gl <- gl.edit.recode.ind( )`	edit population assignments, and apply the changes on closure

Basic genlight filters

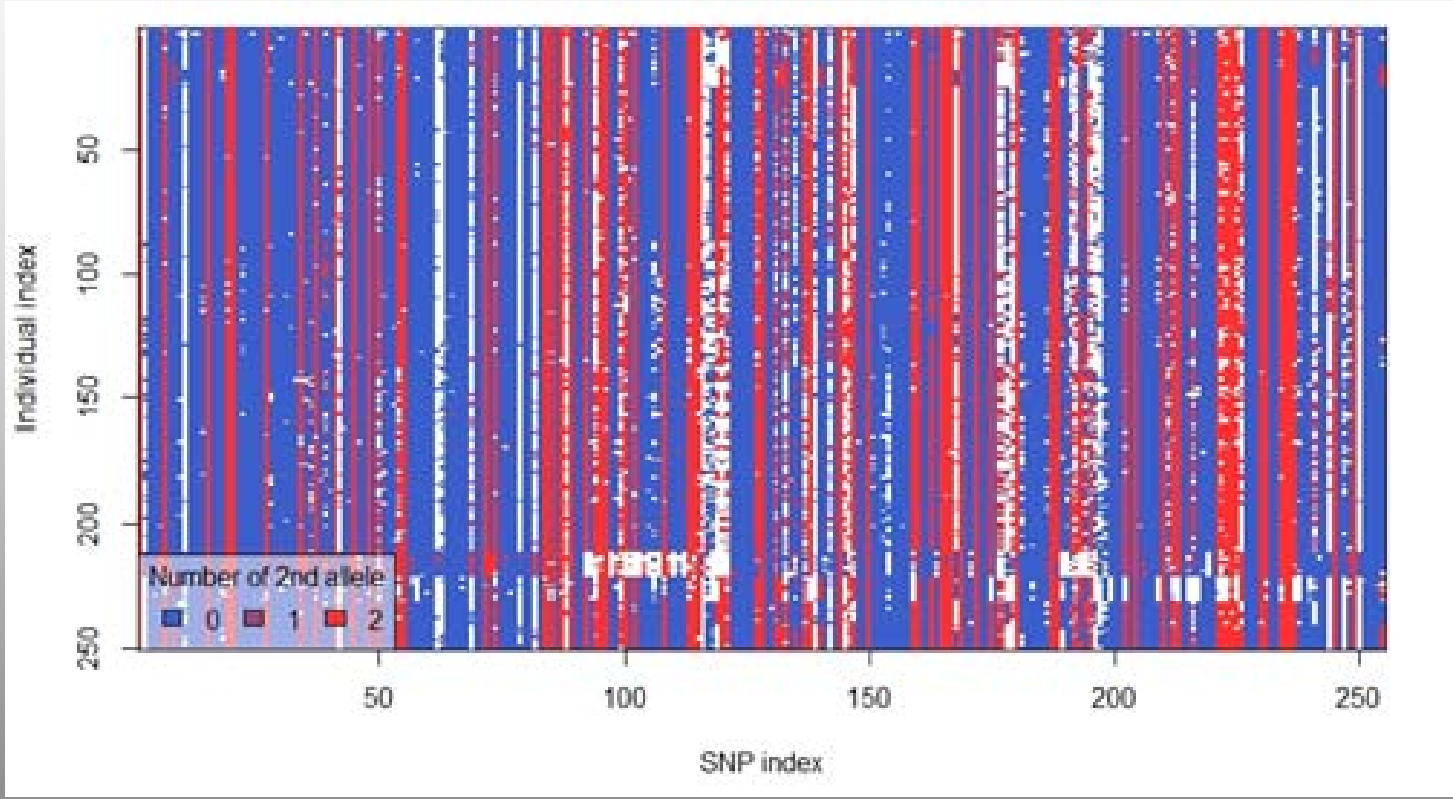
`gl <- gl.filter.repavg( )`	filter out loci for which the repeatability is less than a specified threshold, say threshold = 0.99
`gl <- gl.filter.callrate( )`	filter out loci for which the call rate (rate of non-missing values) is less than a specified threshold, say threshold = 0.95
`gl <- gl.filter.maf( )`	filter on minor allele frequency
`gl <-gl.filter.secondaries( )`	filter out SNPs that share a sequence tag, except one retained at random
`gl <- gl.filter.hamming( )`	filter out loci **that** differ from each other by less than a specified number of base pairs
`gl <- gl.filter.monomorphs( )`	filter out monomorphic loci and loci that are scored all NA

Note
These functions have companion gl.report.xxxx()
functions

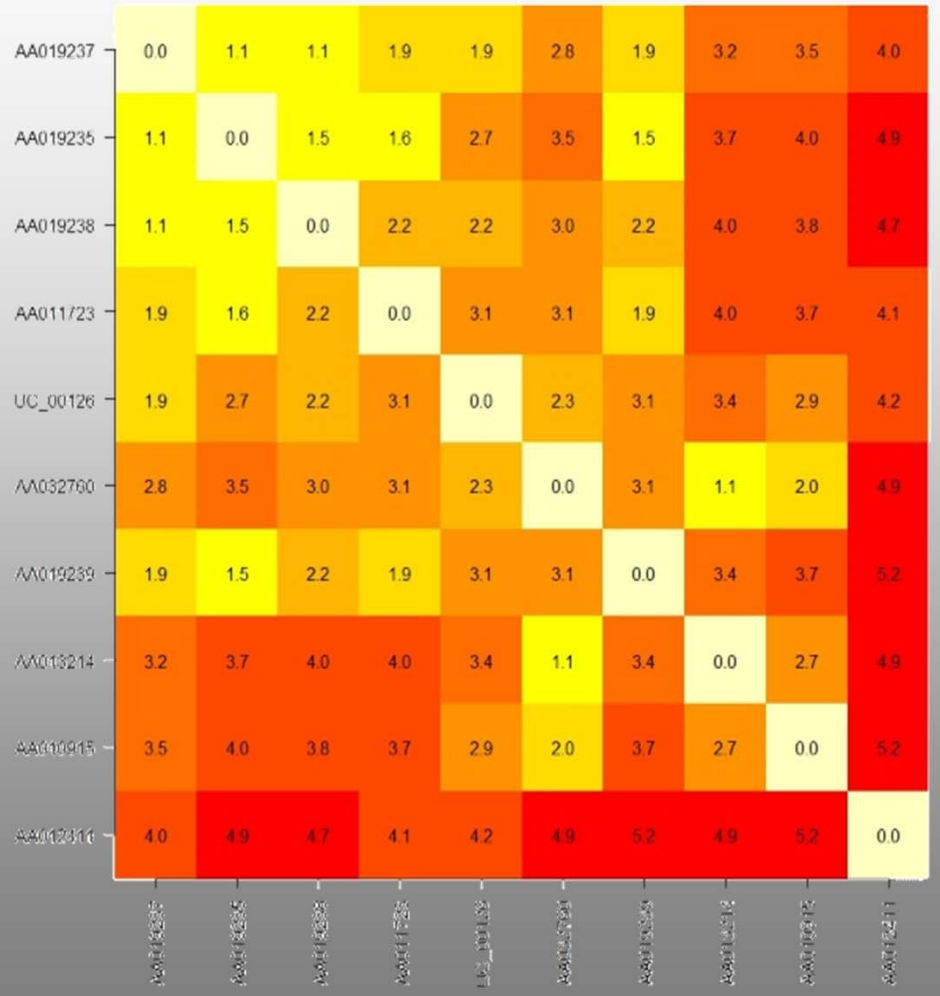
Visualisaation:
Smear Plot

```
glPlot(gl)
```

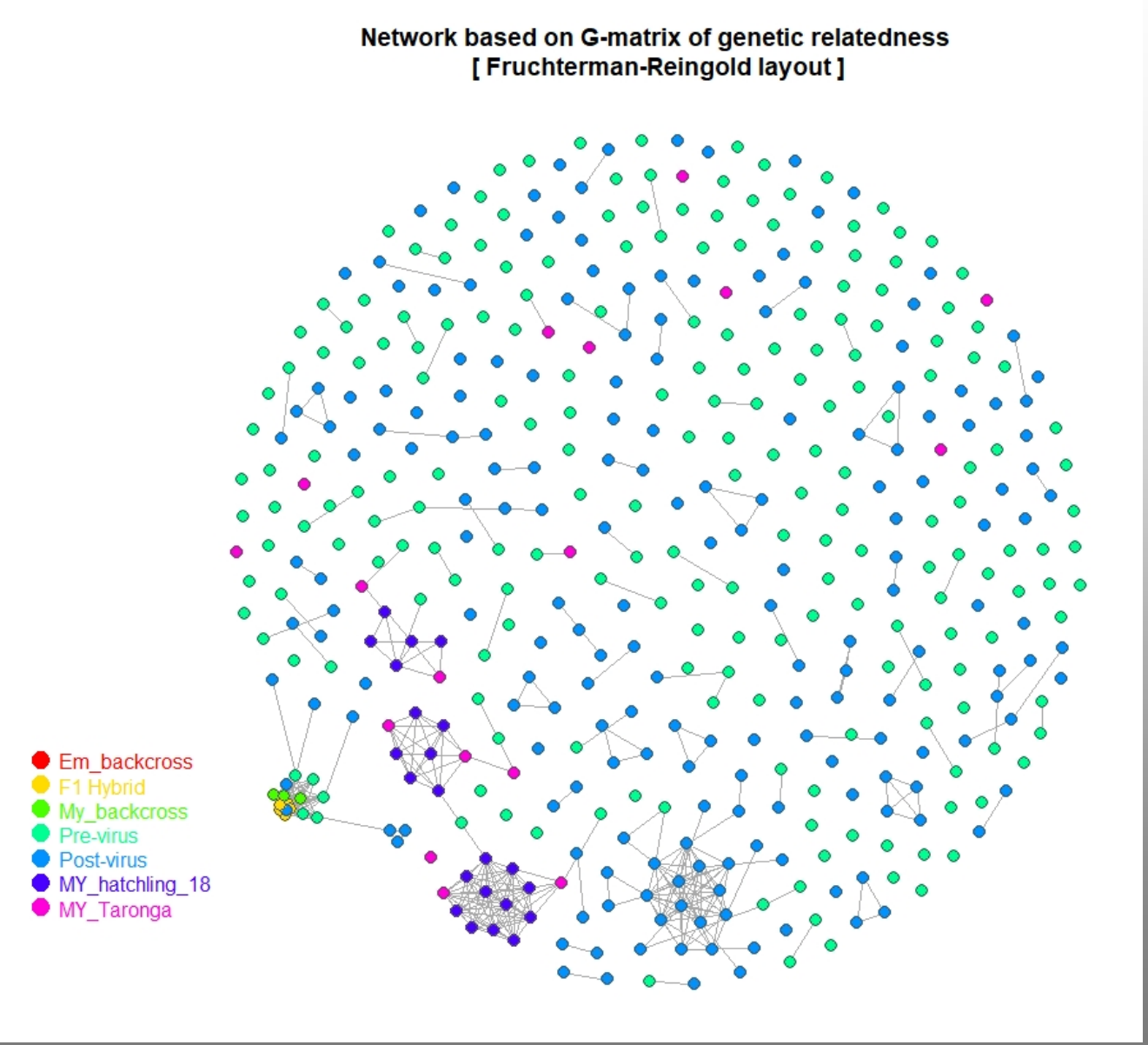
display a smear plot of individuals by loci



Distance metric -- Heatmap

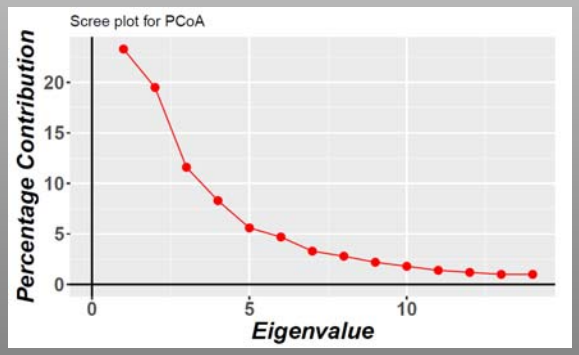


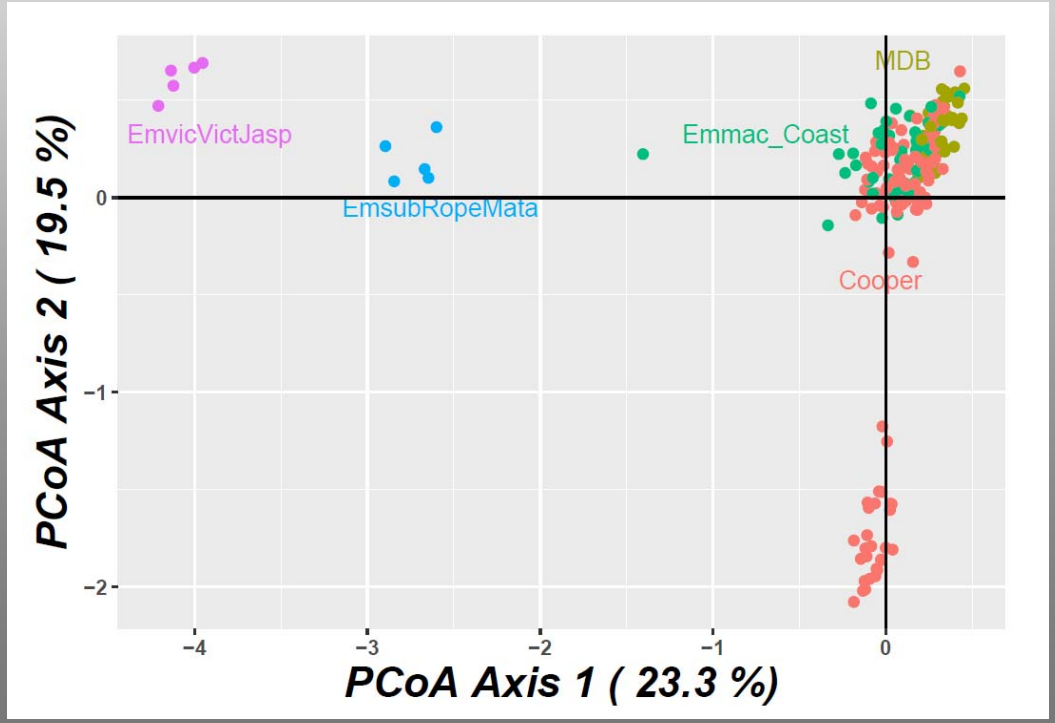
Distance metric -- Network



Principal Coordinates Analysis

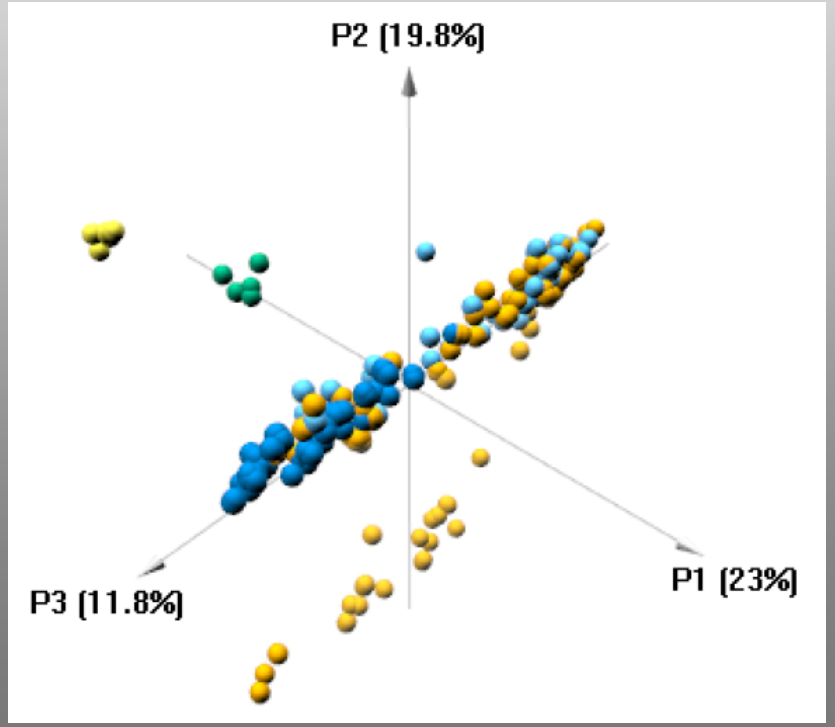
`pcoa <- gl.pcoa(gl)`	conduct the principal coordinates analysis
`gl.pcoa.scree(pcoa)`	plot eigenvalues for leading PCoA axes to enable assessment of the number of ordinated dimensions to examine
`gl.pcoa.plot(pcoa, gl)`	plot the individuals in the space defined by two specified PCoA axes
`gl.pcoa.plot.3d(pcoa,gl)`	plot the individuals in the space defined by three specified axes, and allow mouseover rotation





Principal Coordinates Analysis

`pcoa <- gl.pcoa(gl)`	conduct the principal coordinates analysis
`gl.pcoa.scree(pcoa)`	plot eigenvalues for leading PCoA axes to enable assessment of the number of ordinated dimensions to examine
`gl.pcoa.plot(pcoa, gl)`	plot the individuals in the space defined by two specified PCoA axes
`gl.pcoa.plot.3d(pcoa,gl)`	plot the individuals in the space defined by three specified axes, and allow mouseover rotation



Exercises

- Examine a Diversity Arrays Technology dataset
- Read the data into dartR

  gl <- gl.read.dart(…..)
- Interrogate the dataset

  nLoc, nPop, nLoc ……
- Filter the datset
- Save the genlight object for future use

  saveRDS(gl, file="mygl.Rdata")
- Visualization

Refer to the Workbook for details