Wildlife Genomics@UC

UNIVERSITY OF CANBERRA

SNP Analysis using dartR

How dartR stores your data and data input

Version 4

Copies of the latest version of this tutorial are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: **arthur.georges@biomatix.com.au**

Copyright @ 2025 Arthur Georges

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, including electronic, mechanical, photographic, or magnetic, without the prior written permission.

Such permission would normally be granted for educational purposes, to be used with or without modification, provided that due acknowledgement is given.

Citation: Georges, A., Gruber, B. and Mijangos, J.L(2025). How dartR stores your data and data input. Version 4. Institute for Applied Ecology, University of Canberra, Canberra, ACT, 2617.

dartR is a collaboration between the University of Canberra, CSIRO and Diversity Arrays Technology, and is supported with funding from the ACT Priority Investment Program, CSIRO and the University of Canberra.

Contents

Session 1: Introduction to DArTSeq 4

Sequencing 4

The SNP dataset 5

SilicoDArT 7

Where have we come? 7

Further reading 8

Session 2: Getting data into dartR 9

A sensible workflow 9

How dartR stores SNP data 9

Locus metadata 10

Individual metadata 12

Flags 13

History 13

How dartR stores SilicoDArT data 13

Reading DArT files into a genlight object 14

SNP genotypes 14

SilicoDArT genotypes 15

Reading non-DArT files into a dartR genlight object 16

Saving a genlight object 16

Tidy up the workspace 17

Where have we come? 17

Exercises 17

Exercise 1: 2-Row Format 17

Exercise 2: 1-Row Format 18

Exercise 3: SilicoDArT 18

References 19

Session 1: Introduction to DArTSeq

Sequencing

Diversity Arrays Technology Pty Ltd (DArT) is a private company that specializes in genotyping by sequencing. Their approach is one of genome complexity reduction. But what does this mean?

Basically, DArTSeq is a method that extracts reproducible genomic variation across the genomes of many individuals at an affordable cost. The technique digests genomic DNA using pairs of restriction enzymes (cutters) (Figure 1). When the DNA is cut at two locations within a reasonable distance of each other, the fragment is available for sequencing using the Illumina short-read platforms. Hence, the data are representational in the sense that they are generated for a random but reproducible selection of small fragments of sequence only, fragments that exhibit variation at the level of single base pairs (SNPs).

The first step in the process involves the selection of restriction enzymes that provide the best balance between getting adequate fraction of the genome represented, an adequate read depth for each fragment, and adequate levels of polymorphism. This is species specific and so requires some initial optimization.

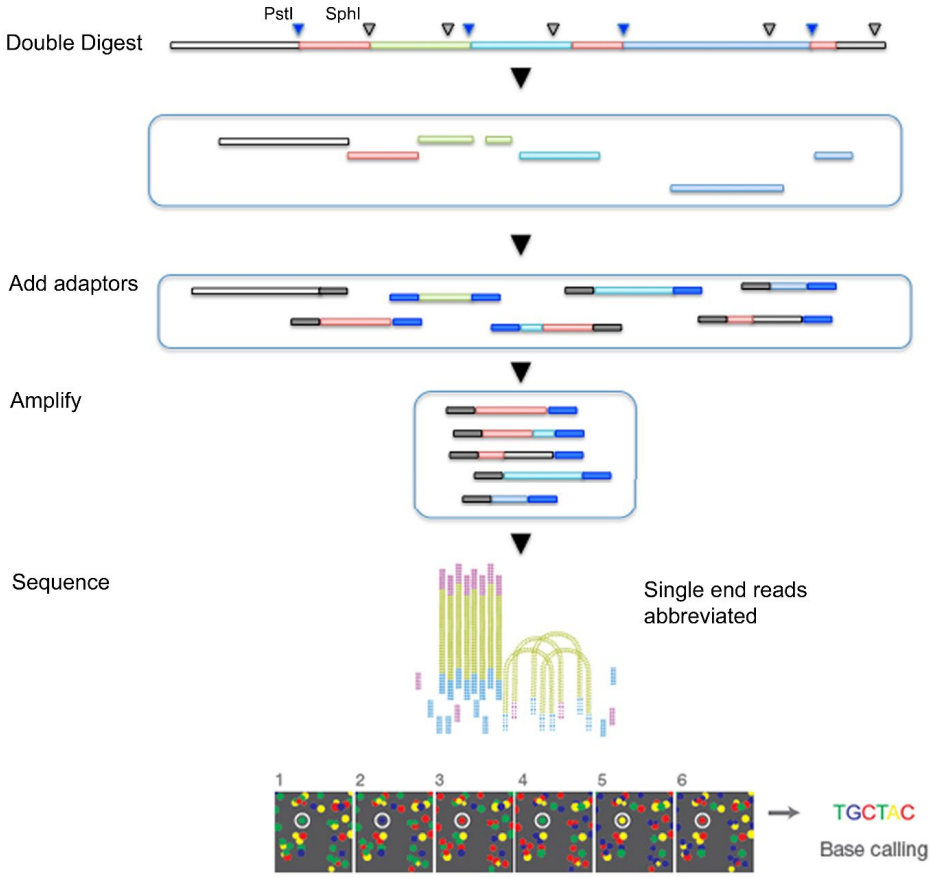


Figure 1. A diagram showing the workflow for representational sequencing using the services of Diversity Arrays Technology.

Once the best restriction enzymes are selected, say PstI (recognition sequence 5'-CTGCA|G-3') and SphI (5'-GCATG|C-3'), then the DNA is digested, and various adaptors added to the sequence fragments to allow Illumina short-read sequencing to proceed. These additional terminal sequences include a barcode to allow disaggregation of the sequences for each sample during later analysis.

The fragments of DNA selected by this process are sequenced in an abbreviated process to yield a set of raw "sequence tags" each of around 75 bp. They are filtered on sequence quality, particularly in the barcode region, truncated to 69 bp and stacked based on sequence similarity. A series of proprietary filters are then applied to select those sequence tags that include a reliable SNP marker.

In particular, one third of samples are processed twice as technical replicates, from DNA and using independent adaptors, through to allelic calls. Scoring consistency (repeatability) is used as the main selection criterion for high quality/low error rate markers.

These DArT analysis pipelines have been tested against hundreds of controlled crosses to verify mendelian behaviour of the resultant SNPs as part of their commercial operations.

When you come to publish, you may receive requests to be more elaborative than you are able to, because of the proprietary nature of the pipelines. Diversity Arrays Technology Pty Ltd is a private company and needs to hold some of its proprietary analyses inhouse. Note that other companies with whom you interact, including Illumina, do the same. The work is reproducible in that using the same service/equipment on the same samples will yield the same result. Most journals accept this.

The SNP dataset

SNPs, or single nucleotide polymorphisms, are single base pair mutations at a nuclear locus (Figure 2). That nuclear locus is represented in the dataset by two sequence tags which, at a heterozygous locus, take on two allelic states, one referred to as the reference state, the other as the alternate or SNP state.

Figure 2. A diagram illustrating what is meant by a SNP (single point polymorphism)

Because it is extremely rare for a mutation to occur twice at the same site in the genome (perhaps with the exception of Eucalypts), the SNP data are considered to be effectively biallelic. Sites with more than two states that occur rarely are typically eliminated in the quality control steps as they are bundled with

multiallelic sites arising from multiple copy sequences (e.g. as would arise from gene duplications) removed during preliminary filtering.

The data can be represented in a table of SNP bases (A, T, C or G), with two states for each individual at each locus in a diploid organisms.

Table 1. A table of genotypes for 10 individuals scored at 11 loci. The data show the base pairs on each haplotype of the sequence tag associated with the locus. The haplotypes are not phased with the reference and alternate alleles chosen arbitrarily.

	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

Alternatively, because the data are biallelic, it is computationally convenient to code the data as 0 for homozyogotes for one allele, 1 for heterozygotes, and 2 for homozygotes of the other allele.

The reference allele is arbitrarily taken to be the most common allele, so 0 is the score for homozygous reference, and 2 is the score for homozygous alternate or SNP state. NA indicates that the SNP could not be scored.

Table 2. A table of genotypes for 10 individuals scored at 11 loci. The data show scores of 0, 1 or 2 representing homozygous reference allele, heterozygous and homozygous alternate allele respectively. NA is used to represent where the locus could not be scored for a particular individual.

	Ind01	Ind02	Ind03	Ind04	Ind05	Ind06	Ind07	Ind08	Ind09	Ind10
Locus 1	0	0	0	0	1	0	0	0	0	NA
Locus 2	0	0	0	0	0	0	1	0	0	0
Locus 3	1	2	2	2	2	0	0	0	0	0
Locus 4	0	1	0	1	2	0	0	0	0	0
Locus 5	0	0	0	0	NA	1	0	0	0	0
Locus 6	0	0	0	0	0	0	1	0	0	0
Locus 7	1	2	2	2	2	0	0	0	0	0
Locus 8	0	1	0	1	2	0	0	0	0	0
Locus 9	0	0	0	0	0	0	0	0	0	0
Locus 10	0	0	0	0	0	0	1	0	0	0
Locus 11	1	2	2	2	2	0	0	0	0	0

This is the form the data are stored in in dartR, <u>though note that it departs from the coding arrangement used by Diversity Arrays Technology</u>.

Some sequence tags might contain more than one SNP, in which case they are likely to be closely linked when passed from parent to offspring. These may need consideration when preparing your data for analysis. Note that multiple SNPs occurring in the one sequence tag are each represented as a separate data record in the dataset.

The SNP data are provided in two forms by DArT, which will be described later.

SilicoDArT

As well as individuals varying in allelic composition at SNP sites, they can vary at the restriction sites used to pull the representation from the genome. A mutation at one or both of the restriction sites will result in allelic drop-out or null alleles. The presence or absence of particular sequence tags across individuals provides a source of information additional to the SNP data.

Broadly, SilicoDArT markers can be considered analogous to AFLPs (Amplified Fragment Length Polymorphisms).

Diversity Arrays Technology provide this second dataset, the presence or absence of scored sequence tags across individuals in what it calls the SilicoDArT dataset. The filtering pipeline applied to generate these data has been highly optimized for reliability, so do not be tempted to use the null alleles (missing data) present in the SNP dataset.

Table 3. A table of genotypes for 10 individuals scored at 11 loci. The data show scores of 0 or 1 representing the absence or presence of a sequenced tag at the locus for a each individual. NA is used to represent where the presence or absence could not be determined at a locus for a particular individual.

	Ind01	Ind02	Ind03	Ind04	Ind05	Ind06	Ind07	Ind08	Ind09	Ind10
Locus 01	0	1	0	0	0	0	1	1	0	1
Locus 02	1	0	1	1	1	1	1	0	0	1
Locus 03	1	1	0	1	0	0	0	0	1	0
Locus 04	1	1	0	1	1	0	1	0	0	0
Locus 05	0	0	0	0	1	NA	1	0	1	0
Locus 06	1	1	1	0	1	1	1	0	1	1
Locus 07	1	1	NA	0	0	1	1	0	1	0
Locus 08	1	0	1	0	0	1	1	0	1	1
Locus 09	1	1	0	0	0	1	1	1	0	1
Locus 10	0	0	1	1	0	1	1	1	NA	1
Locus 11	0	1	1	0	1	1	0	0	1	1

Note that unlike the SNP data, NA represents a truly missing value, in that the state, presence or absence of the sequence tag could not be determined.

Where have we come?

The above Session was designed to give you a very brief overview to the Diversity Arrays Technology pipelines for producing SNP and associated data. Having completed this Session, you should now be familiar the following concepts.

- The concept of a SNP marker and how they are generated.
- The distinction between DArTSeq and SilicoDArT datasets.
- The coding used for SNP genotypes – 0 for homozygous reference, 2 for homozygous alternate, 1 for heterozygous, and NA for 'missing'.
- The coding used for SilicoDArT genotypes – 0 for absent, 1 for present, and NA for missing.

Further reading

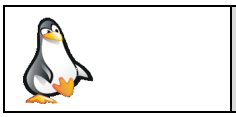
Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., & Kilian, A. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. BMC Proceedings 5(Suppl 7), P54. doi:10.1186/1753-6561-5-S7-P54.

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., … Uszynski, G. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods in Molecular Biology 888*:67–89.

Georges, A., Gruber, B., Pauly, G.B., Adams. M., White, D., Young, M.J., Kilian, A., Zhang, X., Shaffer, H.B. and Unmack, P.J. 2018. Genome-wide SNP markers breathe new life into phylogeography and species delimitation for the problematic short-necked turtles (Chelidae: *Emydura*) of eastern Australia. Molecular Ecology 27:5195-5213. [**Note:** This paper gives some detail on the DArT pipelines for generating the SNP datasets]

Georges, A., Unmack, P.J., Kilian, A., Zhang, X. and Dissanayake, D.S.B. (2025). Lineages as species or lineages within species – using diagnosability to better inform species delimitation (Chelidae: *Emydura*). Molecular Phylogenetics and Evolution, in review (**http://ssrn.com/abstract=5226604**). [**Note:** This paper gives some detail on the DArT pipelines for generating the SilicoDArT datasets]

Session 2: Getting data into dartR



If you are coming back to us, create or load a project, and do not forget to set the default directory for files using `gl.set.wd()`.

A sensible workflow

Let us begin by jumping the gun and defining a sensible pipeline for entering your data, as a context for the material in this and subsequent Sessions.

1. Examine the data provided by Diversity Arrays Technology in Excel to confirm that it conforms to expectations of the dartR package.

   For the SNP data, there needs to be a `AlleleID` column, and the start and end columns of the locus metadata needs to be defined by the columns headed by asterisks. The row with the locus metadata labels needs to be the same row that holds the individual (= specimen or sample labels). This is usually the case, but some older datasets may need a little modification.

   For the SilicoDArt data, there needs to be a `CloneID` column and the locus metadata needs to be defined by the columns headed by asterisks.

2. Prepare the metadata associated with each individual. This dataset, stored in csv format, contains at a minimum the individual/specimen labels in a column headed `id`, and a population column that assigns individuals to groups or populations in a column headed `pop`. Other columns are optional, but might include latitude, longitude of capture, sex, or other possible groupings of the individuals.

3. Read the data into dartR

We elaborate on this workflow in the sections that follow.

How dartR stores SNP data

The package dartR relies on the SNP data being stored in a compact form using a bit-level coding scheme. SNP data coded in this way are held in a genlight object that is defined in R package adegenet (Jombart, 2008; Jombart and Ahmed, 2011). Refer to the tutorial prepared by Jombart and Collinson (2015), called *Analysing genome-wide SNP data using adegenet 2.0.0*, if you require further information.

The complex storage arrangement of genlight objects is hidden from the user because it is accompanied by a number of “accessors”. These allow the data to be accessed in a way similar to the manipulation of standard objects in R, such as lists, vectors and matrices.

A genlight object can be considered to be a matrix containing the SNP data encoded in a particular way. The matrix entities (rows) are the individuals, and the attributes (columns) are the SNP loci. In the body of this individual x locus matrix are the SNP data, coded as 0 for homozygous reference state, 1 for heterozygous, and 2 for homozygous alternate (or SNP) state.

Note: This coding is quite different from that used by Diversity Arrays Technology in their 1-Row and 2-Row csv files provided as part of their report.

Note also that a genlight object used by dartR differs in some important respects from the default genlight object of adgenet (a dartR genelight object is a superset of an adegenet genlight object). By this we mean that all functions in the adegenet package work on dartR genlight objects, but dartR genlight objects have other essential components. So creating a genlight object to hold your data manually from a vcf or csv format requires a few steps in addition to importing the data to an adegenet genlight object, as outlined later in this tutorial.

Genlight objects not only have the SNP data, but also allow for attachment of locus metadata to the loci, and attachment of individual metadata to the individuals/samples. This is represented diagrammatically below.

Locus metadata

The locus metadata included in the genlight object are those provided as part of your Diversity Arrays Technology report. These metadata are obtained from the Diversity Arrays Technology csv file when it is read in to the genlight object. The locus metadata are held in an R data.frame that is associated with the SNP data as part of the genlight object.

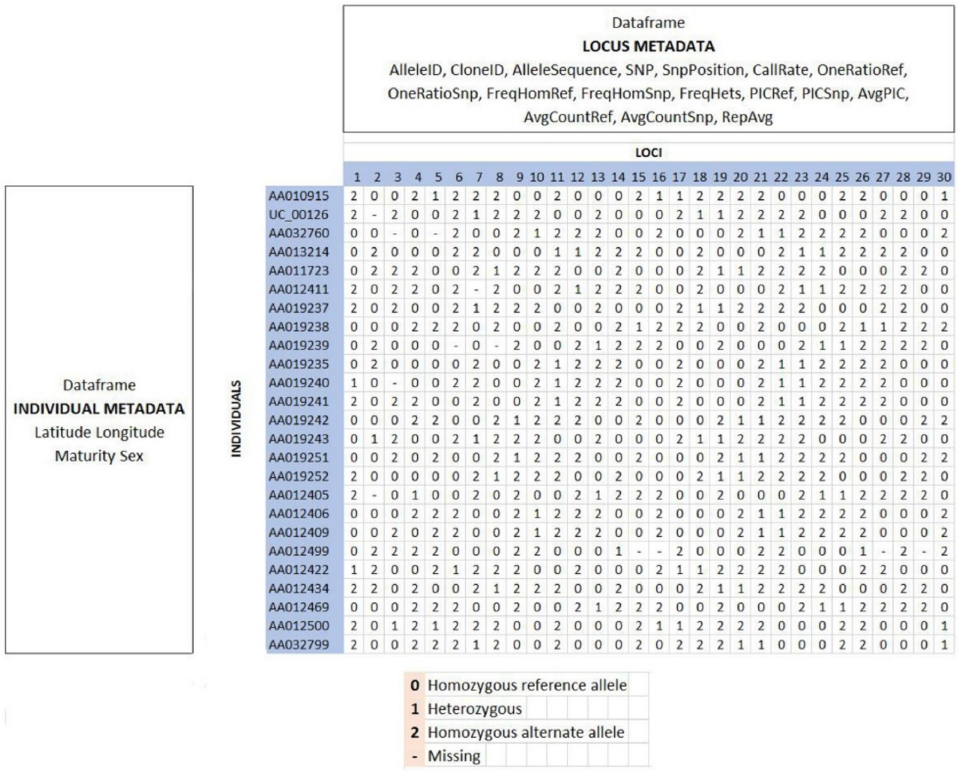


Figure 3. A diagrammatic representation to illustrate the arrangment for storing data in a genlight object. The data are genotypes in locus by individual matrix. Note that the coding of the genotypes changes from that used by Diversity Arrays Technology to the `0,1,2,NA` coding of dartR. Dataframes containing the metadata for loci and for individuals are associated with the genotypes. Dataframes containing the metadata for loci and for individuals are associated with the genotypes. The data and metadata are handled sensibly by package {adegenet} accessors [e.g. `nLoc()`, `nInd()`, `pop()`].

The locus metadata would typically include:

SNP	the mutational change and its position in the sequence tag, referenced from zero
SnpPosition	position (zero is position 1) in the sequence tag of the defined SNP variant base
TrimmedSequence	(optional)The sequence containing the SNP or SNPs (the sequence tag), trimmed of adaptors.
CallRate	proportion of samples for which the genotype call is non-missing (that is, not "-")
OneRatioRef	proportion of samples for which the genotype score is 0
OneRatioSnp	proportion of samples for which the genotype score is 2
FreqHomRef	proportion of samples homozygous for the Reference allele
FreqHomSnp	proportion of samples homozygous for the Alternate (SNP) allele
FreqHets	proportion of samples which score as heterozygous, that is, scored as 1
PICRef	polymorphism information content (PIC) for the Reference allele
PICSnp	polymorphism information content (PIC) for the SNP
AvgPIC	average of the polymorphism information content (PIC) of the Reference and SNP alleles
AvgCountRef	sum of the tag read counts for all samples, divided by the number of samples with non-zero tag read counts, for the Reference allele row
AvgCountSnp	sum of the tag read counts for all samples, divided by the number of samples with non-zero tag read counts, for the Alternate (SNP) allele row
RepAvg	proportion of technical replicate assay pairs for which the marker score is consistent

In addition, dartR calculates the minor allele frequency and an estimate of read depth, and stores it in the locus metadata.

These metadata variables are held in the genlight object as part of a data.frame called loc.metrics, which can be accessed in the following form:

```
# Make a genlight object to work with
  gl <- testset.gl

# Only entries for the first 10 individuals are shown
  gl@other$loc.metrics$RepAvg[1:10]
```

```
## [1] 1.000000 1.000000 1.000000 1.000000 0.989950 1.000000 0.993274
## [8] 1.000000 1.000000 0.980000
```

You can check the names of all available loc.metrics via:

```
names(gl@other$loc.metrics)
```

```
## [1] "AlleleID" "CloneID" "AlleleSequence" "SNP"
## [5] "SnpPosition" "CallRate" "OneRatioRef" "OneRatioSnp"
## [9] "FreqHomRef" "FreqHomSnp" "FreqHets" "PICRef"
## [13] "PICSnp" "AvgPIC" "AvgCountRef" "AvgCountSnp"
## [17] "RepAvg" "clone" "uid" "rdepth" "maf"
```

Examine the first 10 values of RepAvg, CallRate and some other listed locus metadata in `testset.gl` and your own dataset.

Depending on the report from Diversity Arrays Technology you may have additional, fewer or different `loc.metrics` (e.g. `TrimmedSequence` may be available only on request).

These metadata are used by the dartR package for various purposes, so if any are missing from your dataset, then there will be some analyses that will not be possible. For example, `TrimmedSequence` is used to generate output for subsequent phylogenetic analyses that require estimates of base frequencies and transition and transversion ratios.

`AlleleID` is essential (with its very special format), and dartR scripts for loading your data sets will terminate with an error message if this is not present.

Individual metadata

Individual (=specimen/sample) metadata are user specified, and do not come from DArT. Individual metadata are held in a second dataframe associated with the SNP data in the genlight object. See the figure above.

Two special individual metrics are:

`id`	Unique identifier for the individual or specimen that links back to the physical sample
`pop`	A label for the biological population from which the individual was drawn

Individual metrics are supplied by the user by way of a metafile, provided at the time of inputting the SNP data to the genlight object. A metafile is a comma-delimited file, usually named ind_metrics.csv or similar, that contains labelled columns. The file must have a column headed `id`, which contains the individual (=specimen or sample labels) and a column headed `pop`, which contains the populations to which individuals are assigned.

These special metrics can be accessed using:

```
pop(gl)
popNames(gl)
indNames(gl)
```

Try these for yourself to see the output they produce.

A number of other user-defined metrics can be included in the metadata file. Examples of user-defined metadata for individuals include:

sex	Sex of the individual (Male, Female)
maturity	Maturity of the individual (Adult, Subadult, juvenile)
lat	Latitude of the location of collection
long	Longitude of the location of collection

These optional data are provided by the user in the same metafile used to assign id labels and assign individuals to populations at the time of reading in the data.

The individual metadata are held in the genlight object as a dataframe named `ind.metrics`. You can check the names of all available `ind.metrics` via:

```
names(gl@other$ind.metrics)
```

```
[1] "id"      "pop"     "lat"     "lon"     "sex"     "maturity"  "collector" "location"  "basin"    "drainage"
```

and can be accessed using the following form:

```
# Only first 10 entries shown
  gl@other$ind.metrics$sex[1:10]
```

```
[1] Male    Male    Male    Male    Unknown Male    Female  Female  Male    Female
Levels: Female Male Unknown
```

Try these for yourself to see the output they produce.

Flags

The genlight object used by dartR has some additional information not normally accessed by the user. If these data are not in the genlight object, various functions may throw an error.

To ensure your manually generated genlight object (say converted from a vcf file) is compliant, be sure to use

```
gl <- gl.compliance.check(gl)
```

History

A history of manipulations is also stored in the genlight object. This is convenient should you wish to interrogate (or indeed repeat) the process that created the current version of the genlight object.

Display the history of a genlight object using

```
gl.report.history(gl)
```

How dartR stores SilicoDArT data

dartR also stores SilicoDArT presence/absence data in a genlight object, but distinguishes the data from SNP data by setting ploidy=1.

The locus metadata would typically include:

AlleleSequence	Sequence of the tag which is present in samples with genotype score "1"
TrimmedSequence	Same as the full sequence, but with removed adapters in short marker tags
AvgReadDepth	Sum of the tag read counts for all samples, divided by the number of samples with non-zero tag read counts.
StDevReadDepth	Standard deviation of the number of tag reads for all samples with non-zero tag read counts
CallRate	Proportion of samples for which the genotype call is either "1" or "0", rather than "-"
CloneID	Unique identifier of the sequence tag
OneRatio	Proportion of samples for which the genotype score is "1"
PIC	Polymorphism Information Content
Qpmr	Average of the normalized non-zero tag read counts divided by the standard deviation of the normalized non-zero tag read counts (If standard deviation is zero or near zero, the Qpmr value will be 100).
Reproducibility	Proportion of technical replicate assay pairs for which the marker score is consistent

Note that the locus metadata supplied by Diversity Arrays Technology may vary from service to service. The SilicoDarT data and associated metadata can be accessed in the same way as for SNP data, as described above.

Reading DArT files into a genlight object

SNP genotypes

SNP data can be read into a genlight object using `gl.read.dart()`. This function intelligently interrogates the input csv file to determine

- if the file is a 1-row or 2-row format, as supplied by Diversity Arrays Technology Pty Ltd.
- the number of locus metadata columns to be input before reading the the SNP data themselves.
- the number of lines to skip at the top of the csv file before reading the specimen IDs and then the SNP data themselves.
- if there are any errors in the data.

An example of the function used to input data is as follows:

```
gl <- gl.read.dart(
  filename="sample_data_2Row.csv",
  ind.metafile="sample_metadata.csv")
```

The filename specifies the csv file provided by Diversity Arrays Technology, and the `ind.metafile` parameter specifies the csv file which contains metrics associated with each individual (id, pop, sex, environmental data, etc).

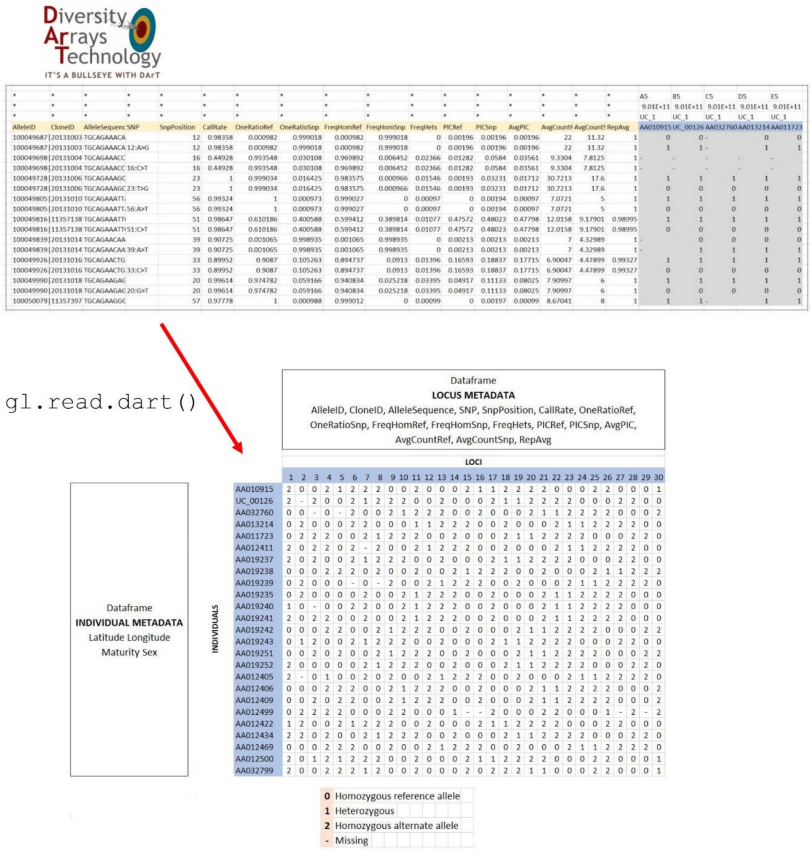


Figure 3. A diagrammatic representation to illustrate the process for reading Diversity Arrays Technology data in a genlight object. The data are genotypes in locus by individual matrix. Note that the coding of the genotypes changes from that used by Diversity Arrays Technology to the 0,1,2,NA coding of dartR. Dataframes containing the metadata for loci and for individuals are associated with the genotypes. The data and metadata are handled sensibly by package {adegenet} accessors [e.g. nLoc(), nInd(), pop())].

The resultant genlight object contains the SNP genotypes, the individual metadata and the locus metadata.

If you have saved the sample files provided into your working directory, read `sample_data_2Row.csv` into a genlight object using `DFwt16-sample_metadata.csv` to assign individual metics.

Verify that the genlight object contains the right number of loci, individuals and populations.

SilicoDArT genotypes

SNP data can be read into a genlight object using `gl.read.silicodart()`. This function intelligently interrogates the input csv file to determine

- the number of locus metadata columns to be input (the first typically being `CloneID` and the last `Reproducibility`).

- the number of lines to skip at the top of the csv file before reading the specimen IDs and then the SNP data themselves.
- if there are any errors in the data.

An example of the function used to input data is as follows:

```
gl <- gl.read.silicodart(
  filename="sample_data_silicodart.csv",
  ind.metafile="sample_metadata.csv")
```

The filename specifies the csv file provided by Diversity Arrays Technology, and the `ind.metafile` parameter specifies the csv file which contains metrics associated with each individual (id, pop, sex, environmental data, etc).

The resultant genlight object contains the SilicoDArT presence/absence genotypes, the individual metadata and the locus metadata.

	If you have saved the sample files provided into your working directory, read `sample_data_silicodart.csv` into a genlight object using `sample_metadata.csv` to assign individual metrics. Verify that the genlight object contains the right number of loci, individuals and populations.

Reading non-DArT files into a dartR genlight object

If you are working with data that have not been prepared by Diversity Arrays Technology, you can still input the data to dartR provided you can get it into the appropriate format.

The way to do this is covered in a separate Tutorial.

Saving a genlight object

Reading the data in from an Excel spreadsheet and converting to a genlight object takes a lot of computation, and so time. You will also have done some tidying up of the data. It is sensible to save your genlight object in binary form using

```
gl.set.wd(getwd())
gl.save(gl,file="tmp.Rdata")
```

and then read it in again with

```
gl.new <- gl.load("tmp.Rdata")
```

	Try saving `gl` or your own genlight object to your workspace, and verify that it has been saved to the appropriate directory. Then restore it to a new genlight object.

Tidy up the workspace

We have created files that we will not use again, so they should be removed from the workspace.

```
rm(gl.new, gl.1row, gl.2row)
```

Where have we come?

In this Session, we have covered a range of topics on data entry, the storage of data and some preliminary approaches to examining those data. Having completed the Session, you should understand

- What is a sensible pipeline for preliminary handling of your SNP data.
- How a genlight object is organised in terms of the SNP scores (which are different from the scores used by Diversity Arrays Technology) and how locus and sample metadata are associated with the genotypes.
- The different types of locus metadata generated by Diversity Arrays Technology, and how to look up what each metric means.
- How to read data from Diversity Arrays Technology into a genlight object.
- How to interrogate the locus and individual (specimen/sample) metadata.

Exercises

Exercise 1: 2-Row Format

- Open the file `sample_data_2Row.csv` in Excel. This is a set of SNP data for Emydura, a freshwater turtle, in 2-row format as would be supplied by Diversity Arrays Technology Pty Ltd.
- Refer to the documentation on the Diversity Arrays Technology web page to understand the scoring of SNPs in the 2-row format.
- Also refer to the MetaDataDefinition file provided by Diversity Arrays Technology as part of their report. In this case, a definition file is provided as `sample_metadata.xlsx`.
- Now examine the individual metadata in the file `sample_data_2Row.csv`. Note the two mandatory columns `id` and `pop`.
- Create a new script in the RStudio Editor Window and add the lines

```
# EXERCISE 1: 2-Row Format
# Input data from sample_data_2Row.csv, associate with
  sample_data_2Row.csv
```

- Add and execute a statement to read the SNP data in to dartR as a genlight object called `gl.2row`
- Add and execute a statement to examine a summary of the contents of `gl.2row`

- Use the `as.matrix()` function to display the genotypes for the first 5 individuals and the first 10 loci.

Exercise 2: 1-Row Format

- Open the file `sample_data_1Row.csv` in Excel. This is a set of SNP data for Emydura, a freshwater turtle, in 1-row format as would be supplied by Diversity Arrays Technology Pty Ltd.
- Refer to the documentation on the Diversity Arrays Technology web page to understand the scoring of SNPs in the 1-row format.
- Also refer to the MetaDataDefinition file provided by Diversity Arrays Technology as part of their report. In this case, a definition file is provided as `sample_metadata.xlsx`.
- Now examine the individual metadata in the file `sample_data_1Row.csv`. Note the two mandatory columns `id` and `pop`.
- Create a new script in the RStudio Editor Window and add the lines

```
# EXERCISE 2: 1-Row Format
# Input data from sample_data_1Row.csv, associate
  with sample_data_1Row.csv
```

- Add a statement to read the SNP data in to dartR as a genlight object called `gl.1row`
- Add a statement to examine a summary of the contents of `gl.1row`
- Use the `as.matrix()` function to display the genotypes for the first 5 individuals and the first 10 loci.

Exercise 3: SilicoDArT

- Open the file sample_data_silicodart.csv in Excel. This is a set of marker presence/absence data for Cherax destructor provided in SilicoDArT format by Diversity Arrays Technology Pty Ltd.
- Refer to the documentation on the Diversity Arrays Technology web page to understand the scoring of the data in the SilicoDArT format.
- Also refer to the MetaDataDefinition file provided by Diversity Arrays Technology as part of their report. In this case, a definition file is provided as `sample_metadata.xlsx`.
- Now examine the individual metadata in the file `sample_data_1Row.csv`. Note the two mandatory columns `id` and `pop`.
- Create a new script in the RStudio Editor Window and add the lines

```
# EXERCISE 3: SilicoDArT data
# Input data from sample_data_SilicoDArT.csv,
  associate with sample_data_SilicoDArT.csv
```

- Add a statement to read the SNP data in to dartR as a genlight object called `gs.`
- Add a statement to examine a summary of the contents of `gs.`

- Use the `as.matrix()` function to display the genotypes for the first 5 individuals and the first 10 loci.

References

Jombart T. and Caitlin Collins, C. (2015). Analysing genome-wide SNP data using adegenet 2.0.0. http://adegenet.r-forge.r-project.org/files/tutorial-genomics.pdf

Jombart T. and Ahmed, I. (2011). *adegenet 1.3-1*: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27: 3070–3071.

Jombart, T., Kamvar, Z.N., Collins, C., Lustrik, R., Beugin, M.P., Knaus, B.J., Solymos, P., Mikryukov, V., Schliep, K., Maié, T., Morkovsky, L., Ahmed, I., Cori, A., Calboli, F. and Ewing, R.J. (2018). Package 'adegenet'. Version 2.1.1. Exploratory Analysis of Genetic and Genomic Data. https://github.com/thibautjombart/adegenet

Ende