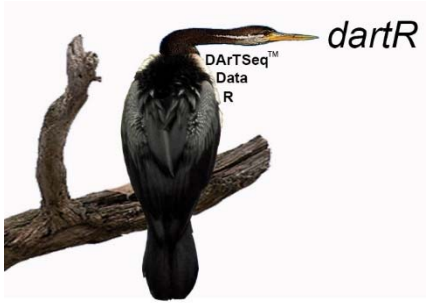
SNP Analysis using dartR



Fixed Difference Analysis

Copies of these workshop notes are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: **arthur.georges@canberra.edu.au**

Contents

Fixed Difference Analysis

Introduction

The objective of a fixed difference analysis is to identify genetically diagnostic units in studies of species delimitation or phylogeography.

Typically, populations of a species or species complex do not freely connect across the landscape. Barriers to dispersal of individuals may impede or prevent gene flow, leading to population structure across the landscape.

As two isolated populations diverge, they will accumulate allele frequency differences through drift, selection and mutation. At some point, allele frequencies at a particular locus may come to fixation for one state in one population (say homozygous reference or 0) and to fixation for the other state in the other population (homozygous alternate or 2). These two populations will have acquired a fixed allelic difference.

Allele frequencies may ebb and flow, but once a locus becomes fixed for an allele or suite of alleles, there is no return, in the absence of convergent mutations (rare for SNPs) or gene flow. The acquisition of a fixed difference between two diverging populations is thus considered to be a significant biological event.

Accumulation of fixed differences between two populations is a robust indication of lack of gene flow, because exchange of only one individual per generation is enough to prevent divergence of allelic profiles. Long-standing reproductive isolation or long-standing geographic isolation can both allow the accumulation of fixed differences.

Fixed differences are Diagnostic

Fixed differences also have an important property. They allow unambiguous assignment of an individual to its source population (or species). Unlike loci for which alleles are present in both populations in different frequencies, a locus for which its alleles are fixed and different between two populations is a diagnostic character. The allelic state of an individual at that locus unambiguously assigns that individual to one or the other populations. This diagnosability property is of particular value in studies of species delimitation.

Fixed Differences are not Transitive

Fixed allelic differences between populations, taken pairwise, are not transitive. Populations A and B can exhibit no fixed differences, and populations B and C can exhibit no fixed differences, but populations A and C can have accumulated fixed allelic differences. For example, the percent frequencies of the alternate allele at a given locus might be

	Locus 01
Pop A	0
Pop B	25
Pop C	100

in which case a fixed difference occurs between population A and C, but not between populations A and B or B and C. In practice, this occurs when you have a geographical cline, whereby adjacent populations experience some level of geneflow that prevents or episodically removes any fixed differences, but fixed differences nevertheless accumulate through isolation by distance. In the extreme case of a ring species, individuals at a particular location may segregate into two populations based on fixed differences in sympatry, only to be linked by a series of intermediate populations that form a ring around a mountain range or insular coastline for example.

The non-transitive nature of fixed differences is accommodated in the fixed difference analysis implemented in dartR. Clinal variation is accommodated in the designation of diagnostic operational taxonomic units (OTUs), by iteratively amalgamating populations that are not differentiated by fixed differences. In this way, even a ring species will ultimately be regarded as a single OTU.

Allopatry versus Sympatry

Sympatry

Fixed allelic differences between two putative taxa that have been in sympatry long enough to have cross-bred were this possible, is unambiguous evidence of reproductive isolation and their status as distinct species. Such evidence is more definitive than morphological evidence which may admit the possibility of phena, that is, morphological variants arising through differing developmental histories, or as polyphenisms. On occasion, even the two sexes of a species were regarded as different sympatric species, until examination using genetic tools (e.g. the butterflies *Caeruleuptychia helios* and *Magneuptychia keltoumae*; Nakahara et al., 2018).

How one comes to suspect that two taxa exist in sympatry varies with the circumstances. It might be that the two taxa are each widespread and distinctive, and have recently been found in sympatry. Two different phenotypes, unremarkable in the context of a cline, might be found in microsympatry raising suspicions that there are two species rather than a single polytypic species. In the case of truly cryptic species, suspicions may be aroused because the location in which both are found throws a strong deviation from Hardy-Weinberg Equilibrium, and a STRUCTURE analysis yields two genetically distinctive groupings. Whatever the case may be, a fixed difference analysis would begin with the putative sympatric taxa identified and separated *a priori* as putative taxa.

Allopatry

Cases in allopatry are simpler in the sense that the populations subject to study are clearly defined, but are more complex in that it is not possible to objectively decide if diagnosable aggregations of populations are species, or if they represent structure within a species. This is because the diagnosability can arise from either reproductive isolation (characteristic of species) or geographic isolation (characteristic of lineages within species), and it is difficult to distinguish the two.

The fixed difference analysis cannot resolve this conundrum, but rests upon the premise that diagnosability is a necessary but not sufficient criterion for assigning species status. In that sense, the fixed difference analysis identifies a set of diagnosable aggregations of populations that are candidates for consideration as species, taking into account also a phylogeny. Subjective considerations, taking into account all available evidence, will be

required to decide which diagnosable aggregations of populations should be regarded as species, and which should be regarded as structure within species. Unlike the sympatric case, an objective decision is not possible in allopatry.

This will become clearer as we work though an analysis.

False Positives

One of the limitations of fixed difference analysis is the possibility of false positives arising because finite samples of individuals are typically collected from the sampling sites.

A Fundamental Asymmetry

There is an asymmetry in fixed difference analysis analogous to the asymmetry in hypothesis testing.

In hypothesis testing, a significant difference can be accepted with a measurable level of uncertainty (usually < 0.05), but a non-significant difference is ambiguous. When a result is non-significant, the test might have failed because there is no difference, or because the sample sizes were insufficient to detect a difference when it existed. Interpretation of a non-significant difference is thus ambiguous, and requires an accompanying power analysis.

In the case of fixed allelic differences, the asymmetry lies in the fact that if two sets of individuals are drawn from two populations and found to share alleles at all of the loci examined, then no amount of additional sampling will uncover a fixed difference. Shared alleles at all loci allows a definitive conclusion that the two populations from which the individuals were drawn have not accumulated fixed differences. The result is definitive.

The presence of fixed differences in the sample set, on the other hand, is ambiguous. They might represent true fixed differences between the two populations, or they might have arisen simply by chance (false positives), given the sample sizes. To interpret an observed count of fixed differences between two populations, we need an estimate of the accompanying false positive rate.

By this reasoning, two populations can be aggregated into a single OTU on the basis of lack of fixed allelic differences regardless of the sample size, but a decision to regard two populations as distinct relies on sample sizes that are adequate for distinguishing real fixed differences from false positives (sampling error). This has important consequences for interpretation of fixed differences in support of identifying diagnosable OTUs.

Compounding Error

When considering a single locus, relatively few individuals per population are required to practically eliminate a false positive, for all but extreme differences in allele frequencies between the two populations. For example, if the allele frequencies of the focal SNP locus are 50:50, and the sample sizes are 5 individuals from population A and 5 individuals from population B, it does not require explicit calculation to realise that the probability of a false positive is vanishingly small. It is the probability of getting all 5 individuals in one population as homozygous reference and all 5 individuals in the other population as homozygous alternate, by chance, given p=q=0.5.

A number of issues complicate these calculations. The first is that, although the probability of a false positive at one locus might be vanishingly low, the calculations are typically conducted over very many loci, and the errors compound. The probability of finding a false fixed difference across 60,000 SNP loci can be substantial, regardless of how small this probability is for one particular locus.

Allele Frequency Profiles

The second issue is more insidious. The probability of a false positive at a locus depends critically on the allele frequencies in the populations at that locus. For example, the probability of a false positive fixed difference at a locus with allele frequencies PopA = 99.5:0.5 and PopB = 0.5:99.5 is going to be quite high. To calculate the probability of false positives across all loci will require knowledge of the allele frequencies in each population at each locus. Of course, this information is unavailable.

The next best option is to use the observed allele frequencies across loci in simulations to count the number of false positives that are expected to occur by chance – the False Positive Rate. This has been implemented in R package dartR.

Pragmatic Decision Required

To undertake these calculations, it is necessary to provide a practical definition of a false positive. If two populations with true allele frequencies of 99.95:0.05 and 0.05:99.95 throw a fixed difference in two finite samples of individuals, would we call this a false positive? Probably not. The populations are effectively fixed and different at that locus in the two populations.

In addition, a locus with allele frequencies of PopA = 99.5:0.5 and PopB = 0.5:99.5 is much more likely to come to fixation than it is to move in the opposite direction. So, the true difference might be considered fixed from a practical point of view, and scoring it as fixed based on the sample data is not of great consequence.

A second consideration, is that the two populations being compared may contain true fixed differences, such that true positives will be conflated with the false positives. The challenge for the simulation is to admit that the comparison is not between two allelic profiles that share all alleles at some non-zero frequency (a simple null model), but between two populations that may have fixed differences unknown in number.

Whatever way you look at this challenge, a threshold, delta (∂), needs to be set when generating the expected false positive rate. Delta is a threshold specifying how extreme the divergence between two populations (not samples) needs to be in order to score the difference as fixed. A value of ∂ = 0.02 might be appropriate.

With parameter ∂ set, and with simulations, we are able to generate an estimate of the number of false positives expected given the sample sizes. This false positive rate and its error in estimation serves as a basis for deciding if the observed number of fixed differences reflects the presence of real fixed differences between two populations or if they arose by chance alone.

Fixed Difference Analysis

We are now in a position to devise a fixed difference analysis to identify sets of our sampling sites for which, collectively, individuals are diagnosable by one or more fixed allelic differences.

Explore — The first step is to examine the data graphically to identify putative boundary zones exhibiting evidence of hybridization or introgression that may be taken out of the analysis and considered separately. Whether you do this will depend on your view of hybridization and its impact on species delimitation. Retaining sampling sites with some level of hybridization or introgression at the boundary of what would otherwise be distinct entities will result in the amalgamation of those entities into a single OTU. Maybe that is what you want; or maybe you are tolerant of some level of hybridization between good species at a zone of contact.

Compare — The second step is to consider the sampling sites as the fundamental entity for the analysis. We then compare each sampling site with each other sampling site to calculate the number of fixed allelic differences between them.

Amalgamate — The third step is to amalgamate the individuals from sampling sites for which there are no fixed differences, in the knowledge that ***absence of fixed differences in the sample set implies absence of fixed differences in the populations from which they were drawn***. This step provides a set of putative operational taxonomic units, or OTUs.

Because of the low but inherent error rate in calling SNPs and assigning states between heterozygous and homozygous at a locus, you might want to base this decision on the absence of corroborated fixed differences, that is, setting tpop=1, so that two fixed differences or more are required to prevent amalgamation of two populations into a single OTU.

Reiterate — The fourth step is to repeat the procedure until no further amalgamations are possible. This iterative procedure accommodates the non-transitivity of fixed differences. Clines will amalgamate into putative OTUs even though some populations within the OTU will have fixed differences in comparison with others. Populations along a cline will daisy-chain into putative OTUs by this procedure.

Test — The fifth step is to consider the statistical significance of the observed fixed differences between the putative OTUs derived above. The OTUs can then be further amalgamated on the basis of lack of significance (that is, the number of fixed differences does not exceed the false positive rate).

At the end of the analysis, we will have classified the sampling sites into OTUs each diagnosable by one or more fixed allele differences (two or more if tpop=1). We can be confident that these resultant OTUs are not subject to contemporary gene flow and have not been subject to such geneflow in the recent past.

The OTUs can be designated as Evolutionarily Significant Units (ESUs), subspecies or species, drawing upon all available evidence. If your approach is phylogenetic, then you can map the diagnosable OTUs against the tree to evaluate which clades should be regarded as candidate species.

Comprehensive Geographic Sampling

Fixed difference analysis relies on comprehensive sampling across the landscape so as to avoid interpreting sparsely sampled populations as diagnosable OTUs when in fact there exist intermediate populations with allelic profiles that would unite them. An excellent treatment of this issue is provided by Chambers and Hillis (2020, *Systematic Biology*, 69:184–193). Failure to achieve comprehensive coverage of the distribution of a species complex influences the decisions on which OTUs represent species and which represent diagnosable lineages within species, and reduces the objectivity of the value the fixed difference analysis adds to these decisions.

Adequate Sample Sizes

A second set of issues arise when the number of individuals per sampling locality is small. First of all, small sample size increases the false positive rate for fixed differences, and so the risk of identifying diagnosable OTUs arising through sampling error. This can be accommodated in part by testing the number of fixed differences between two populations statistically, but the statistical test incorporated into dartR relies on a reasonable estimate of the allele frequency profile for each population, and for this to be so, the sample sizes should be $\geq$ 10 (2n = 20).

The bottom line is that, if you want a robust fixed difference analysis, you need to sample comprehensively across the range of the suspected species complex you are working with and collect 10 or more individuals per sample site.

Practical Considerations

Some would argue that this is rarely achievable. Your options then are

(a) Where possible, manually amalgamate populations that are in sufficiently close proximity to warrant an assumption that they belong to the same diagnosable taxon. In the case of aquatic organisms, this manual amalgamation might be warranted for populations with low sample sizes within the single catchment.

(b) Consider increasing the level of corroboration of fixed differences required to prevent amalgamation. Here we have argued for corroborated fixed differences with npop=1, that is, for at least two fixed differences to preclude amalgamation. But when the sample sizes are low in some or many populations, then consideration should be given to increasing the level of corroboration. One way to select a threshold is to examine the fixed difference matrix or the average number of fixed differences between populations and pick a value that is clearly a low outlier in comparison with the "norm" between populations. A value of npop of 5, or 8 or even 20 might be justified if the mean number of fixed differences among populations is typically in the 100s.

(c) Manually apply the testing of fixed differences against the estimate of the false positive rate, taking particular care to note that these comparisons are not transitive. Fixed differences between a population with a low sample size and other populations might not exceed the false positive rate in a number of comparisons, rendering the decision on which populations to amalgamate challenging, and subjective.

The worst case scenario is that a great number of diagnosable OTUs will emerge from the analysis. As argued above, large aggregations do not come into question as diagnosable OTUs, but a plethora of smaller diagnosable OTUs arising from the analysis because of false positive fixed differences will be an annoyance when making the ultimate decision as which OTUs are species and which are diagnosable lineages within species. Definitive interpretation acceptable to reviewers will be challenging.

Worked Example

As an example, let us consider a SNP data generated for a freshwater turtle from range of sites across eastern Australia (Georges et al., 2018).

```
gl <- readRDS(file="Emydura_sth_for_workshops.Rdata")
```

In `gl`, we have the genotypes for individuals assigned to populations (sampling sites). The two marked in red below are outgroup taxa – *Emydura subglobosa* and *E. victoriae* respectively. The remainder belong to a species complex from southern and eastern Australia referred to collectively as the Southern *Emydura*.

In preparation for the fixed difference analysis, sample sites at the boundary of two regions and that show evidence of contemporary admixture have been removed. Sample sites with a low number of individuals have, where possible, been amalgamated with other sample sites within the same drainage basin.

This leaves us with 40 ingroup sites for the Southern *Emydura* and 2 populations for the outgroup taxa.

```
table(pop(gl))
```

NEQNormanby	10_NEQBarrCair	101_LEBCoopEulb	102_LEBCoopCull	11_NEQRussEube
11	4	9	10	10
MDB_Warrego	113_MDBWarrDart	114_MDBWarrSanf	115_MDBWarrBiny	117_MDBCondArch
3	10	10	10	10
12_NEQJohnWari	121_MDBBordGoon	123_MDBGwydBing	124_MDBMacqCudg	125_MDBLachForb
10	10	9	10	10
127_MDBMurrAlbu	131_MDBMurrMBri	14_CEQRossRoss	19_CEQBurdMist	29_CEQStyxStyx
10	10	10	10	5
35_CEQFitzFair	36_CEQFitzCarn	37_CEQFitzKorc	38_CEQFitzTaro	43_CEQFitzAlli
10	10	4	5	10
48_SEQKolaKola	53_SEQBurnBara	58_SEQMaryBoru	60_SEQMaryTiar	62_SEQFrasBowa
5	10	6	4	5
63_SEQFrasMack	65_SEQPineBuny	67_SEQBrisWive	70_NENTweeUki	73_NENRichCasi
5	3	10	10	10
NENClarence	NENBellinger	84_NENMaclGeor	NENHunter	96_LEBCoopAvin
10	7	10	10	10
EmsubRope	**EmvicVict**			
5	5			

The data were filtered as per Georges et al. (2018).

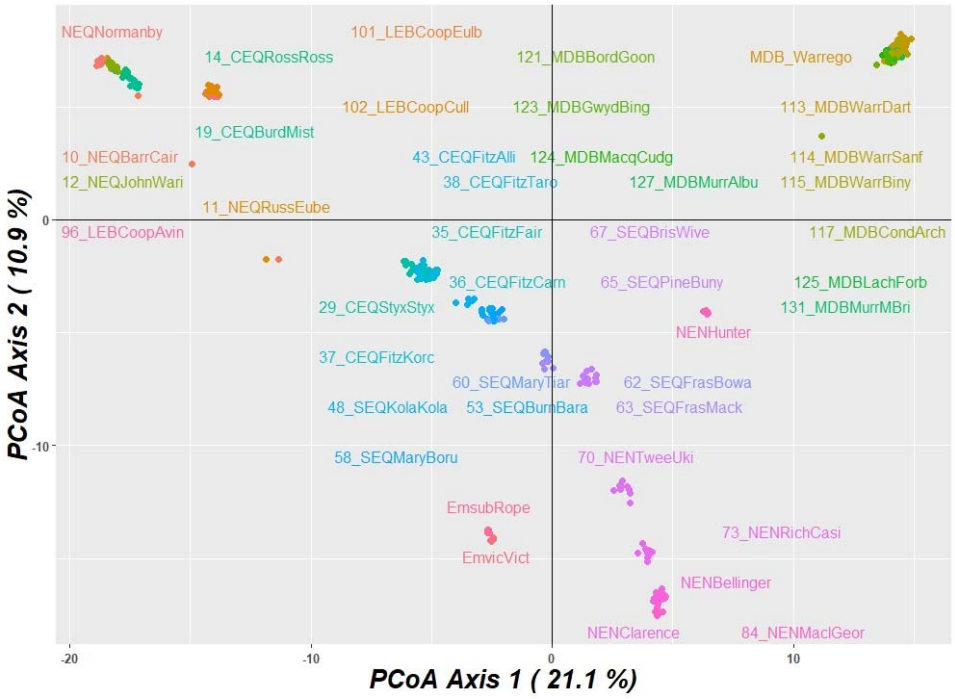
```
gl <- gl.filter.callrate(gl, threshold=0.95, verbose=3)
```

```
gl <- gl.filter.repavg(gl, threshold = 0.99, verbose=3)
gl <- gl.filter.secondaries(gl, verbose=3)
```

Explore

We can visualize the similarities using a PCoA applied to Euclidean distances calculated from the SNP genotypes.

```
pcoa <- gl.pcoa(gl)
gl.pcoa.plot(pcoa,gl)
```



Bit messy with the site labels, but there is considerable structure among the ingroup sample sites evident in the top two dimensions of the ordination. The question is, how distinct are these sample sites. Are they each diagnosable?

Compare

A first step in the fixed difference analysis is to calculate a matrix of fixed differences.

```
D <- gl.fixed.diff(gl)
```

Object `D` is a list containing the revised `gl` object and square matrices, as follows

D[[1]]$gl – the input genlight object;

D[[2]]$fd – raw fixed differences;

D[[3]]$pcfd – percent fixed differences;

D[[4]]$nobs – mean no. of individuals used in each comparison;

D[[5]]$nloc – total number of loci used in each comparison;

D[[6]]$expfpos – if test=TRUE, the expected count of false positives for each comparison [by simulation];

D[[7]]$sdfpos – if test=TRUE, the standard deviation of the count of false positives for each comparison [by simulation];

D[[8]]$prob – if test=TRUE, the significance of the count of fixed differences [by simulation].

Note that the D[[6]] to D[[8]] are populated with NAs unless the test parameter is set to TRUE.

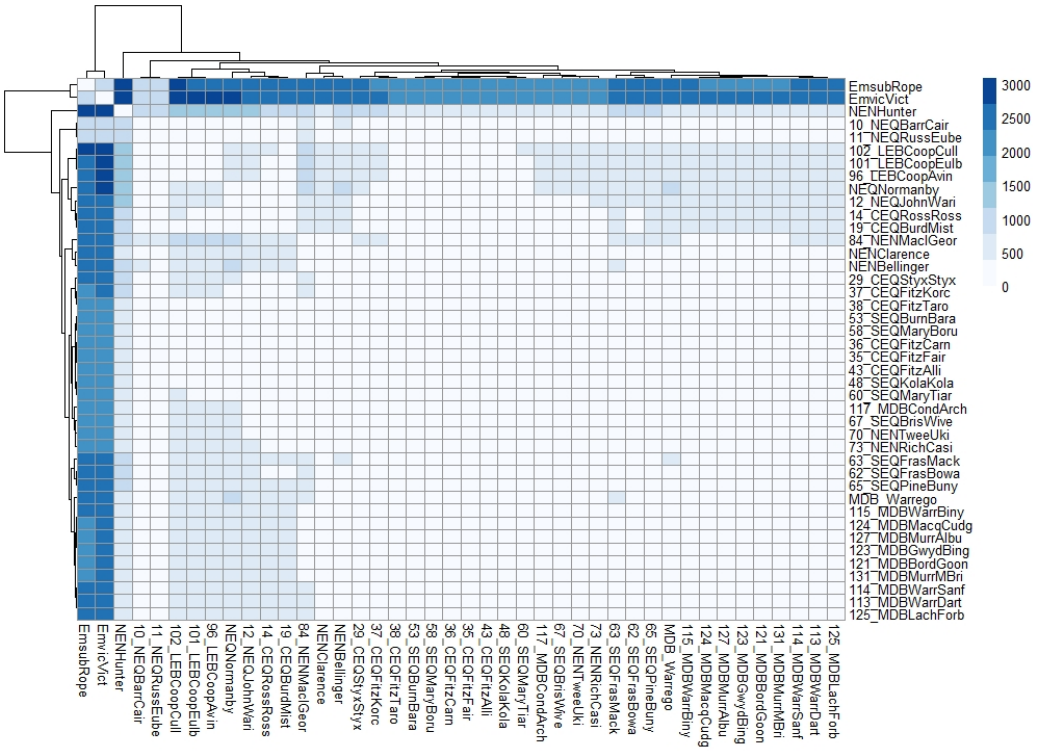
The matrices are too big to show here but you can peruse then on the screen.

```
D
```

Alternatively, the fixed difference matrix can be visualized as a heatmap.

```
gl.plot.heatmap(D)
```

There is not a lot of structure in the data, apart from differences between the outgroup taxa (*Emydura subglobosa and E. victoriae*) and the ingroup taxa.



Aggregate

At this point we might consider aggregating sample sites pairwise where they have not accumulated any fixed differences. To explain the procedure, consider the dartR function

```
D2 <- gl.collapse(D,verbose=3)
```

```
Starting gl.collapse
Starting gl.collapse
  Comparing populations for absolute fixed differences
  Amalgamating populations with zero fixed differences

Initial Populations
 NEQNormanby 10_NEQBarrCair 101_LEBCoopEulb 102_LEBCoopCull 11_NEQRussEube MDB_War
rego 113_MDBWarrDart 114_MDBWarrSanf 115_MDBWarrBiny 117_MDBCondArch 12_NEQJohnWar
i 121_MDBBordGoon 123_MDBGwydBing 124_MDBMacqCudg 125_MDBLachForb 127_MDBMurrAlbu
131_MDBMurrMBri 14_CEQRossRoss 19_CEQBurdMist 29_CEQStyxStyx 35_CEQFitzFair 36_CEQ
FitzCarn 37_CEQFitzKorc 38_CEQFitzTaro 43_CEQFitzAlli 48_SEQKolaKola 53_SEQBurnBar
a 58_SEQMaryBoru 60_SEQMaryTiar 62_SEQFrasBowa 63_SEQFrasMack 65_SEQPineBuny 67_SE
QBrisWive 70_NENTweeUki 73_NENRichCasi NENClarence NENBellinger 84_NENMaclGeor NEN
Hunter 96_LEBCoopAvin EmsubRope EmvicVict
```

```
Group:10_NEQBarrCair+
[1] "10_NEQBarrCair" "11_NEQRussEube" "12_NEQJohnWari" "14_CEQRossRoss" "19_CEQBur
dMist"

Group:101_LEBCoopEulb+
[1] "101_LEBCoopEulb" "102_LEBCoopCull" "96_LEBCoopAvin"

Group:113_MDBWarrDart+
 [1] "113_MDBWarrDart" "114_MDBWarrSanf" "115_MDBWarrBiny" "117_MDBCondArch" "121_
MDBBordGoon" "123_MDBGwydBing" "124_MDBMacqCudg" "125_MDBLachForb" "127_MDBMurrAlb
u" "131_MDBMurrMBri" "MDB_Warrego"

Group:35_CEQFitzFair+
[1] "35_CEQFitzFair" "36_CEQFitzCarn" "37_CEQFitzKorc" "38_CEQFitzTaro" "43_CEQFit
zAlli" "58_SEQMaryBoru" "53_SEQBurnBara" "48_SEQKolaKola" "60_SEQMaryTiar"

Group:65_SEQPineBuny+
[1] "65_SEQPineBuny" "67_SEQBrisWive" "70_NENTweeUki"  "73_NENRichCasi" "NENClaren
ce" "84_NENMaclGeor" "NENBellinger"
```

There are 5 aggregations of sample sites, each aggregation comprising sites that, when compared pairwise, have no fixed allelic differences at any loci.

The output matrix can be examined by accessing the fd matrix in the class fd object that was produced by gl.collapse().

```
D2$fd
```

	10_NEQBarrCair+	101_LEBCoopEulb+	113_MDBWarrDart+	29_CEQStyxStyx	35_CEQFitzFair+	62_SEQFrasBowa	63_SEQFrasMack	65_SEQPineBuny+	EmsubRope	EmvicVict	NENHunter	NEQNormanby
10_NEQBarrCair+	0	204	76	74	8	160	229	38	718	694	676	1
101_LEBCoopEulb+	204	0	360	423	182	557	677	283	2608	2673	1276	525
113_MDBWarrDart+	76	360	0	120	10	115	155	8	1999	2068	456	417
29_CEQStyxStyx	74	423	120	0	0	189	276	61	2359	2436	893	387
35_CEQFitzFair+	8	182	10	0	0	10	18	1	1817	1886	407	117
62_SEQFrasBowa	160	557	115	189	10	0	3	16	2380	2454	818	580
63_SEQFrasMack	229	677	155	276	18	3	0	25	2498	2574	940	717
65_SEQPineBuny+	38	283	8	61	1	16	25	0	1917	1966	389	330
EmsubRope	718	2608	1999	2359	1817	2380	2498	1917	0	910	3022	2646
EmvicVict	694	2673	2068	2436	1886	2454	2574	1966	910	0	3089	2725
NENHunter	676	1276	456	893	407	818	940	389	3022	3089	0	1389
NEQNormanby	1	525	417	387	117	580	717	330	2646	2725	1389	0

There are two things of note here. The first is that, even though we aggregated sample sites on the basis of no fixed differences, the outcome has some pairs of aggregations that still have no fixed differences (e.g. 29_CEQStyxStyx and 35_CEQFitzFair+ in the revised fixed difference matrix). This is because of the non-transitive property of fixed differences, and is the reason the gl.collapse script needs to be run iteratively.

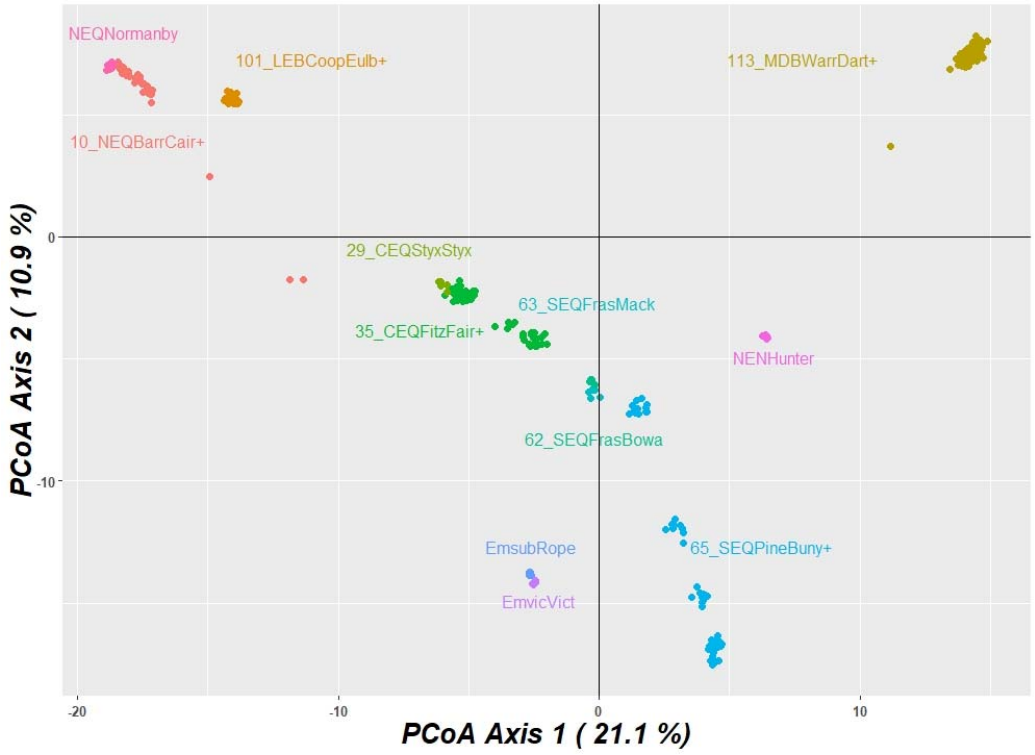
The second observation is that some sample sites/aggregations are supported by only one fixed difference. Some people like to work only with corroborated fixed differences by setting the parameter tpop=1. Leave that for another day.

We can examine a revised ordination, using the new collapsed matrix using

```
pca <- gl.pcoa(D2)
gl.pcoa.plot(pca,D2)
```

Note that the gl.pcoa and gl.plot.pcoa functions recognise D2 as a class 'fd' object and handle it appropriately.

The graph is starting to look more presentable, with less 'populations' to display because of the aggregations.



We could continue to run the aggregation script repeatedly until there are no populations that do not differ.

```
D3 <- gl.collapse(D2)
```

and would find that the Styx River sampling site now aggregates with the 35_CEQFitzFAir+ aggregation. Then run it again to be sure there is no further opportunity to collapse the matrix. However, there is a better way.

Recursive application of gl.collapse

Recursive application of gl.collapse is rather tedious. The best way to proceed is to use the script

```
D <- gl.collapse.recursive(gl,tpop=1,verbose=3)
```

This time let's examine corroborated fixed differences, with `tpop` set to 1.

The function first sets tpop to 0, then recursively collapses the matrix until no further improvement is made. It then sets tpop to 1, and repeats the procedure, until ultimately there is a final set of aggregations, each differing from the other by 2 or more fixed differences.

```
Setting tpop: 0
Iteration: 1

Initial Populations
 NEQNormanby 10_NEQBarrCair 101_LEBCoopEulb 102_LEBCoopCull 11_NEQRussEube MDB_War
rego 113_MDBWarrDart 114_MDBWarrSanf 115_MDBWarrBiny 117_MDBCondArch 12_NEQJohnWar
i 121_MDBBordGoon 123_MDBGwydBing 124_MDBMacqCudg 125_MDBLachForb 127_MDBMurrAlbu
131_MDBMurrMBri 14_CEQRossRoss 19_CEQBurdMist 29_CEQStyxStyx 35_CEQFitzFair 36_CEQ
FitzCarn 37_CEQFitzKorc 38_CEQFitzTaro 43_CEQFitzAlli 48_SEQKolaKola 53_SEQBurnBar
a 58_SEQMaryBoru 60_SEQMaryTiar 62_SEQFrasBowa 63_SEQFrasMack 65_SEQPineBuny 67_SE
```

```
QBrisWive 70_NENTweeUki 73_NENRichCasi NENClarence NENBellinger 84_NENMaclGeor NEN
Hunter 96_LEBCoopAvin EmsubRope EmvicVict

New population groups

Group:10_NEQBarrCair+
[1] "10_NEQBarrCair" "11_NEQRussEube" "12_NEQJohnWari" "14_CEQRossRoss" "19_CEQBur
dMist"

Group:101_LEBCoopEulb+
[1] "101_LEBCoopEulb" "102_LEBCoopCull" "96_LEBCoopAvin"

Group:113_MDBWarrDart+
 [1] "113_MDBWarrDart" "114_MDBWarrSanf" "115_MDBWarrBiny" "117_MDBCondArch" "121_
MDBBordGoon" "123_MDBGwydBing" "124_MDBMacqCudg" "125_MDBLachForb" "127_MDBMurrAlb
u" "131_MDBMurrMBri" "MDB_Warrego"

Group:35_CEQFitzFair+
[1] "35_CEQFitzFair" "36_CEQFitzCarn" "37_CEQFitzKorc" "38_CEQFitzTaro" "43_CEQFit
zAlli" "58_SEQMaryBoru" "53_SEQBurnBara" "48_SEQKolaKola" "60_SEQMaryTiar"

Group:65_SEQPineBuny+
[1] "65_SEQPineBuny" "67_SEQBrisWive" "70_NENTweeUki" "73_NENRichCasi" "NENClarenc
e" "84_NENMaclGeor" "NENBellinger"

Setting tpop: 0
Iteration: 2
Initial Populations
 NEQNormanby 10_NEQBarrCair+ 101_LEBCoopEulb+ 113_MDBWarrDart+ 29_CEQStyxStyx 35_C
EQFitzFair+ 62_SEQFrasBowa 63_SEQFrasMack 65_SEQPineBuny+ NENHunter EmsubRope Emvi
cVict

New population groups
Group:29_CEQStyxStyx+
[1] "29_CEQStyxStyx" "35_CEQFitzFair+"

Setting tpop: 0
Iteration: 3

No further amalgamation of populations at fd <= 0

Setting tpop: 1
Iteration: 1

Initial Populations
 NEQNormanby 10_NEQBarrCair+ 101_LEBCoopEulb+ 113_MDBWarrDart+ 29_CEQStyxStyx+ 62_
SEQFrasBowa 63_SEQFrasMack 65_SEQPineBuny+ NENHunter EmsubRope EmvicVict

New population groups
Group:10_NEQBarrCair++
[1] "10_NEQBarrCair+" "NEQNormanby"

Group:29_CEQStyxStyx++
[1] "29_CEQStyxStyx+" "65_SEQPineBuny+"

Setting tpop: 1
Iteration: 2

No further amalgamation of populations at fd <= 1
Completed: gl.collapse.recursive
```

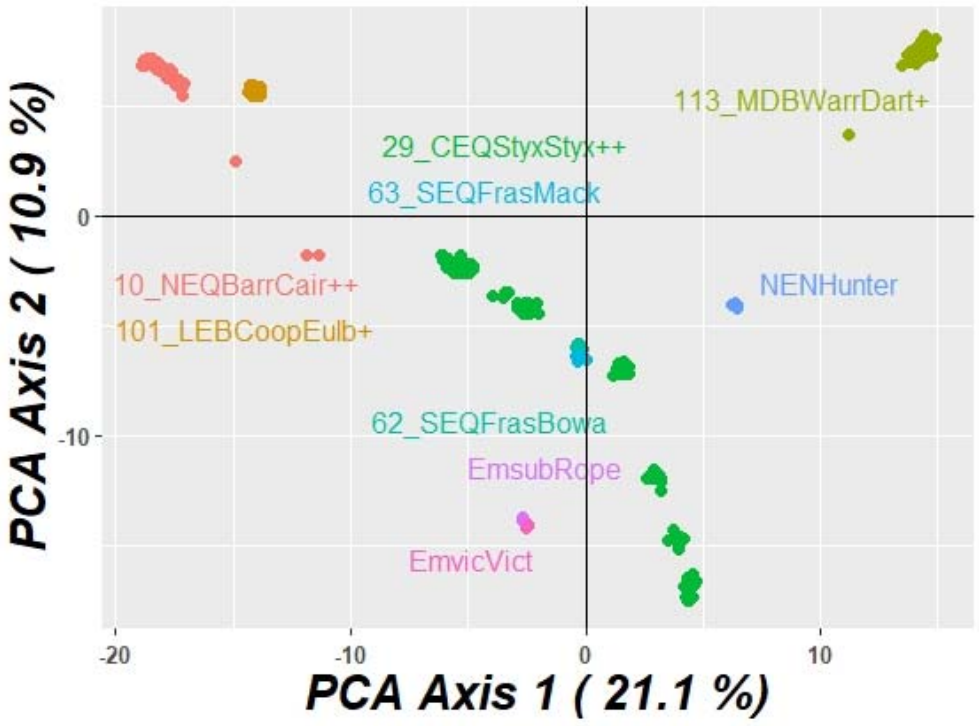
We just run the script again to be sure.

```
D <- gl.collapse(D)
```

```
No further amalgamation of populations at fd <= 0
  Recursive analysis complete
```

We can now examine the final result with an ordination plot or heatmap.

```
pca <- gl.pcoa(D)
gl.pcoa.plot(pca,D)
```



Note the clinal series (green) extending up the coast from the Macleay River in the south to the Fitzroy River in the north. The initial collapse of populations had these as a group of diagnosable taxa that collapse on applying the technique recursively.

Testing for Significance

There is one last issue to consider, the possibility that distinctions between our final aggregations are based on false positives. Note that some of the populations have sample sizes of only 5. With such low sample sizes, and the number of loci being considered, it is possible that the 3 fixed differences observed between, say, Fraser Island (Bowarrady) and Fraser Island (MacKenzie) arose by sampling error.

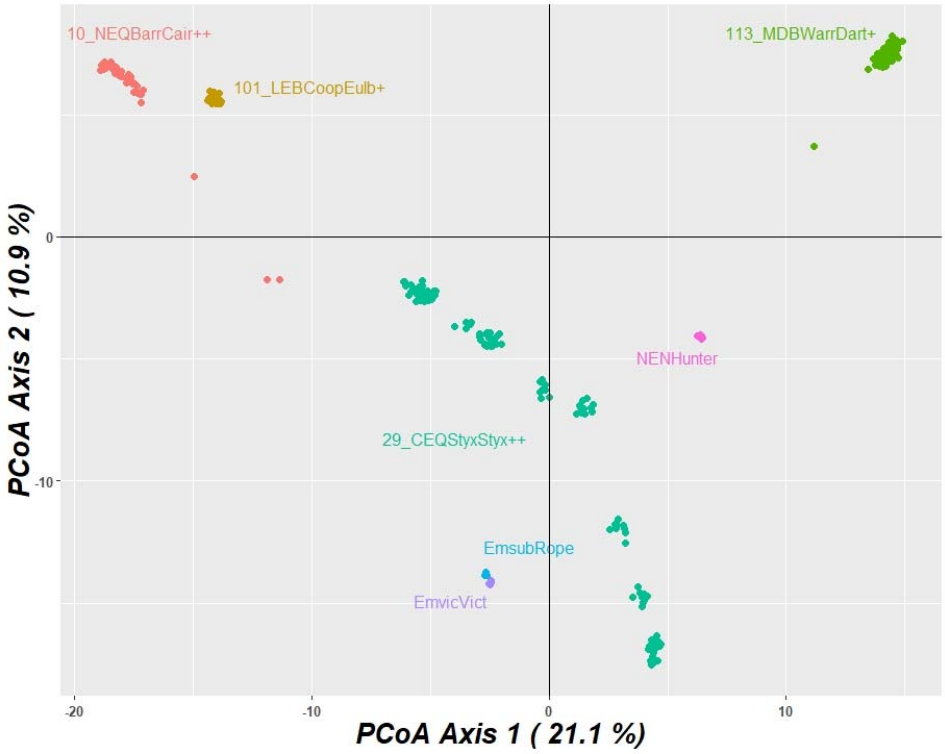
This concern can be accommodated by testing the observed differences for significance. For sake of illustration, we set the number of replicates in the simulations to only 100.

```
D <- gl.fixed.diff(D$gl,test=TRUE,reps=100)
```

	10_NEQBarrCair+	101_LEBCoopEulb+	113_MDBWarrDart+	29_CEQStyxStyx+	62_SEQFrasBowa	63_SEQFrasMack	65_SEQPineBuny+	EmsubRope	EmvicVict	NENHunter	NEQNormanby
10_NEQBarrCair+		0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.797
101_LEBCoopEulb+	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
113_MDBWarrDart+	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
29_CEQStyxStyx+	0.003	0.000	0.000		0.980	0.945	0.003	0.000	0.000	0.000	0.000
62_SEQFrasBowa	0.000	0.000	0.000	0.980		1.000	0.984	0.000	0.000	0.000	0.000
63_SEQFrasMack	0.000	0.000	0.000	0.945	1.000		0.999	0.000	0.000	0.000	0.000
65_SEQPineBuny+	0.000	0.000	0.000	0.003	0.984	0.999		0.000	0.000	0.000	0.000
EmsubRope	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000
EmvicVict	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000
NENHunter	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000
NEQNormanby	0.797	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

There are six comparisons that are not significant based on the simulations (four shown). For example, with p =1.000, the observed number of fixed differences between Fraser Island (Bowarrady, n=5) and Fraser Island (MacKenzie, n=5) did not exceed the number of false positives expected from sampling error.

We can further collapse the populations by amalgamating those that do not differ significantly, to yield a robust designation of operational taxonomic units (OTUs) well



supported by the evidence.

We now have our result. The original sampling sites have been aggregated on the basis of being insufficiently distinguishable using fixed differences. Our analysis first considered absolute fixed differences, then finished with an assessment on the basis of an estimate of false positive rate given the sample sizes.

Sample	sizes		
10_NEQBarrCair++	101_LEBCoopEulb+	113_MDBWarrDart+	29_CEQStyxStyx++
55	29	102	139
NENHunter	EmsubRope	EmvicVict	
10	5	5	

There are the two outgroup taxa as diagnosable OTUs, an OTU comprising the Burdekin River and those to the north along the coast, Cooper Creek, Murray-Darling Basin, the east coastal rivers from the Styx in the north to the Bellinger Coast in the south, and the Hunter River OTU. These entities can be diagnosed by fixed allelic differences, and are presumably free of contemporary or recent gene flow.

Note that the fixed difference analysis depends critically on adequate sampling across the range of the taxon concerned. This is necessary to capture clinal variation and not

interpret sparsely sampled sites on a cline as distinct OTUs. It was also necessary to detect any hybrid zones or zones of localized introgression. The sampling intensity needs to be high enough (say n=10) to both reduce the influence of false positives and to enable a robust assessment of false positives.

Fixed Differences in Species Delimitation

Taxonomically diagnosable units have been identified across the landscape, but are these diagnosable OTUs species? This is an age-old question that has no simple answer.

All species are lineages, but not all lineages are species. A first step in making this distinction is to insist that the lineages under consideration be diagnosable. This constraint alone greatly reduces the incorporation of lineages that have been subject to recent or contemporary allelic exchange, and so puts a constraint on taxonomic inflation. The fixed difference analysis provides a means of assessing lineages against the criterion of diagnosability.

In recent papers, we have outlined the steps for using SNPs in species delimitation. Ours is one view, and by no means universally accepted, but it presents a defensible approach that avoids over-splitting.

Our fundamental contention is that all species delimitation studies, whether traditional, genetic, or genome-based, should supplement any tree-based or network-based approach by cross-referencing with five additional tree-free analyses:

1. Construct ordination plots of the genetic affinities among individuals to identify both discrete and admixed genetic groups; separate out instances of contemporary hybridization and introgression for separate analysis;
2. Apply phylogenetic techniques to identify lineages;
3. Assess diagnosability of any lineages thus identified;
4. Explicitly consider the geographic relationships among all diagnosable lineages (sympatry, parapatry, allopatry);
5. Assess sampling intensity within sample sites and spatially; and
6. Incorporate knowledge for other comparative biological attributes of these lineages to inform decisions on taxonomic status – ESU, subspecies, species.

When dealing with SNP data, a fixed difference analysis is central to this six-step process, though the final sixth step will still require considerable subjective judgement when dealing with allopatric OTUs.

Tweaking the Analysis

The analysis can be adjusted to your tastes.

If you believe that the concept of absolute fixed differences is too stringent, then the parameter `tloc` can be set to something other than the default of zero. For example, setting `tloc=0.05` implies that allele frequencies at a locus of 95:5 vs 5:95 will be regarded as a fixed difference.

One reason for altering the value of `tloc` is to use the fixed difference analyses to examine structure across the landscape based on allele frequency variation, rather than the extreme of fixed differences. This provides and alternative to STRUCTURE.

If you think defining diagnostic OTUs on the basis of a single fixed difference is unwise, then setting `tpop=1` will require a fixed difference to be corroborated by another to prevent aggregation. Alternatively, you might look across the fixed distance matrix and set a higher value for `tpop`.

The default value of `delta` is set to 0.2. Delta is the threshold value for the minor allele frequency required to consider the true difference between two populations as operationally fixed. This can be adjusted.

The default value for the test of significance for fixed differences is set at 0.05. This can be adjusted to be more or less stringent using the `alpha` parameter.

Reading

Chambers, E.A. and Hillis, D.M. 2019. The multispecies coalescent over-splits species in the case of geographically widespread taxa. Systematic Biology 69:184–193.

Georges, A., Gruber, B., Pauly, G.B., 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.

Hillis, D.M. 2019. Species delimitation in herpetology. Journal of Herpetology 53:3-12.

Sukumaran J., and L. L. Knowles. 2017. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences USA 114:1607–1612.

Unmack, P.J., Young, M.J., Gruber, B., White, D., Kilian, A., Zhang, X. and Georges, A. 2019. Phylogeography and species delimitation of *Cherax destructor* (Decapoda: Parastacidae) using Genome-wide SNPs. Marine and Freshwater Research 70:857–869.

Unmack, P.J., Adams, M., Hammer, M.P., Johnson, J.B., Gruber, B., Gilles, A. and Georges, A. 2020. Plotting for change: an analytic framework to aid decisions on which lineages are candidate species in phylogenomic species discovery. Submitted [draft available on request]

References

Nakahara, S. , Zacca, T., Huertas, B., Neild, A.F.E., Hall, J.P.W, Lamas, G., Holian, L.A., Espeland, M., and Keith R. Willmott, K.R. 2018. Remarkable sexual dimorphism, rarity and cryptic species: a revision of the '*aegrota* species group' of the Neotropical butterfly genus *Caeruleuptychia* with the description of three new species (Lepidoptera, Nymphalidae, Satyrinae). Insect Systematics & Evolution 49:130-182.

Appendix: Accommodating False Positives

Introduction

A fixed difference at a biallelic SNP locus occurs between two populations (sampling sites) when all individuals in one population are fixed for the reference allele and all individuals in the other population are fixed for the alternate allele, or vice versa.

This simulation deals with the fact that a fixed difference between two samples taken from two populations A and B may represent a true fixed difference between those populations, or may represent a sampling error. How do we determine whether the observed count of fixed differences arising in comparison of two finite samples of individuals is sufficient to conclude that there are true fixed differences between the two populations from which they are drawn?

The simulation generates an expectation for the number of false positive fixed differences between two populations using the allele profiles for the samples and the sample sizes. The cases of sympatry and allopatry are considered separately. The false positive rate can be used to assess whether the observed count of fixed differences is real. Alternatively, the analysis is carried further to provide a test of significance (p value) for the observed fixed differences, taking into account the sample sizes.

Rationale

In the account that follows, $f_{Ai} \in [0,1]$ is the observed relative frequency of the reference allele at locus *i* of *k* loci scored for Population *A*, $p_{Ai} \in [0,1]$ is the true frequency of the reference allele at locus *i*, and n_A is the number of individuals sampled from Population *A*. The analysis applies only to biallelic data from unrelated individuals.

Allopatry

Consider a single locus. If p_A is the true relative frequency of the reference allele in population A from which a sample of n_A individuals is taken, and the individuals are independent (unrelated), then the probability that NONE of the $2n_A$ alleles will be the reference allele is

$$\Pr\{NONE\ A\ ref\} = (1 - p_A)^{2n_A} \quad \ldots\ldots (1)$$

If p_B is the true relative frequency of the reference allele in population B from which a sample of n_B individuals is taken, then the probability that ALL of the n_B alleles will be the reference allele is

$$\Pr\{ALL\ B\ ref\} = (p_B)^{2n_B} \quad \ldots\ldots (2)$$

with p_A and p_B varying independently.

The probability of a fixed difference arising in the samples of size n_A and n_B by chance is

$$\Pr\{Fixed\ Diff\ A\ ref\ B\ alt\} = (1 - p_A)^{2n_A}(p_B)^{2n_B} \quad \ldots\ldots (3)$$

where the alternate allele is fixed in population A and the reference allele is fixed in population B.

For the reverse

$$\Pr\{Fixed\ Diff\ A\ alt\ B\ ref\} = (p_A)^{2n_A}(1-p_B)^{2n_B} \quad \text{..... (4)}$$

so for one OR the other

$$\Pr\{Fixed\ Difference\} = (1-p_A)^{2n_A}(p_B)^{2n_B} + (p_A)^{2n_A}(1-p_B)^{2n_B} \quad \text{..... (5)}$$

The expected count of fixed differences between two populations A and B from which two samples of size n_A and n_B are drawn will be

$$fd = \sum_{i=1}^{i=k}(1-p_{Ai})^{2n_A}(p_{Bi})^{2n_B} + (p_{Ai})^{2n_A}(1-p_{Bi})^{2n_B} \quad \text{..... (6)}$$

for *i* = 1 to *k* loci, assuming that the loci are independent (i.e. not linked).

The true allele frequency distributions across individuals from each of population A and B are unknown, those frequencies will vary from locus to locus (that is, p_A and p_B are not constant), and the two populations may have loci exhibiting true fixed differences. Hence, equation 5 does not yield a practical solution to the problem of estimating the rate of false positives for given sample sizes.

Sympatry

The mathematics for the sympatric case is tractable. The null hypothesis is that the two samples representing putatively distinct taxa are from the same population.

Consider a single locus. If *p* is the <u>true</u> relative frequency of the reference allele in the population from which a sample of n_A individuals is taken, and the individuals are independent (unrelated), then the probability that NONE of the $2n_A$ alleles will be the <u>reference</u> allele is

$$\Pr\{NONE\ A\ ref\} = (1-p)^{2n_A} \quad \text{..... (7)}$$

If n_B individuals are taken because they assign to the second putative sympatric taxon, then the probability that ALL of the n_B alleles will be the <u>reference</u> allele is

$$\Pr\{ALL\ B\ ref\} = (p)^{2n_B} \quad \text{..... (8)}$$

The probability of a fixed difference arising in the samples of size n_A and n_B by chance is

$$\Pr\{Fixed\ Diff\ A\ ref\ B\ alt\} = (1-p)^{2n_A}(p)^{2n_B} \quad \text{..... (9)}$$

where the alternate allele is fixed in population A and the reference allele is fixed in population B.

For the reverse

$$\Pr\{Fixed\ Diff\ A\ alt\ B\ ref\} = (p)^{2n_A}(1-p)^{2n_B} \quad (10)$$

so for one OR the other

$$\Pr\{Fixed\ Difference\} = (1-p)^{2n_A}(p)^{2n_B} + (p)^{2n_A}(1-p)^{2n_B} \quad (11)$$

The expected count of fixed differences between two populations A and B from which two samples of size n_A and n_B are drawn will be

$$fd = \sum_{i=1}^{i=k}(1-p_i)^{2n_A}(p_i)^{2n_B} + (p_i)^{2n_A}(1-p_i)^{2n_B} \quad (12)$$

for *i* = 1 to *k* loci, assuming that the loci are independent (i.e. not linked).

The true allele frequency distributions across individuals is unknown, those frequencies will vary from locus to locus (that is, *p* is not constant across loci), and the two populations may have loci exhibiting true fixed differences. Hence, equation 5 does not yield an exact solution to the problem of estimating the rate of false positives for given sample sizes.

However, equation (10) achieves its maximum when *p* = 0.5, and so too will equation (11). An upper limit to the false positive fixed differences is thus given by

$$fd \leq 2k(0.5)^{2(n_A+n_B)} \quad (13)$$

which provides a convenient upper limit to the number of false positives to expect given the sample sizes.

Simulation

To resolve the allopatric case and provide a more refined estimate of the false positive rate in the case of sympatry, we turn to simulation.

Allopatry

In the allopatric case, we draw at random from the observed allele frequency distributions at a given locus for each of population A and B to derive an estimated sampling distribution for the true allele frequencies at that locus under binomial assumptions. For example, if f_A is the observed frequency of the reference allele at a given locus for population A, then appropriate estimates for the parameters of binomial distribution from which the sample frequencies are drawn are

$$\mu = f_A \quad (14)$$

$$\delta = \sqrt{\frac{f_A(1-f_A)}{2n_A}} \quad (15)$$

accurate when f_A is not too close to 0 or 1.

At a given locus for population A, we first sample a frequency f_A from the observed allele frequency distribution for that locus, then select a frequency p_A at random for the $2n_A$ alleles, where

$$p_A \sim B(2n_A, f_A)$$

Similarly, for population B,

$$p_B \sim B(2n_B, f_B)$$

Using the rbinom() function in the R {stat} package

$$p_A = rbinom(n = 1, size = 2n_A, prob = f_A)$$

$$p_B = rbinom(n = 1, size = 2n_B, prob = f_B) \quad \text{..... (16)}$$

These probabilities are combined using Equation 5 to yield an expected probability of a fixed difference at the focal locus. The calculations are then applied to all loci, and the probabilities summed (Equation 6) to obtain an estimate of the expected count of fixed differences between populations A and B.

The simulation is repeated for 1,000 iterations, or as many as necessary to constrain the precision of the expected count.

Sympatry

The simulations for the sympatric case are similar except that the null proposition is that the two putative taxa are drawn from the same population. That is

$$f = (f_A + f_B)/2$$

to replace f_A and f_B in the computations above.

A Pragmatic Decision

There remains the problem, in the allopatric case, of conflation of true fixed differences between the populations and false positives. This arises because the populations used in the simulations may have true fixed differences, each yielding a sample fixed difference, and these will be combined with false positives in count of expected fixed differences. It is not possible to infer from $f_A = 0$ that $p_A = 0$. For example, the upper 95% confidence limit for $f_A = 0$ is $p_A = 0.168$ for a sample size of 10 individuals (2n=20) (Clopper-Pearson estimate, refer to **http://epitools.ausvet.com.au/content.php?page=CIProportion**, accessed 5-Mar-18). Because it is not possible to infer from $f_A = 0$ that $p_A = 0$, true and false positives are conflated.

We deal with this contingency by setting a tolerance for the minor allele frequency (MAF < ∂) in the populations that will be accepted as contributing to a fixed difference. That is, a positive is a true positive if it arises where

$$p_A < \partial \text{ and } (1 - p_B) < \partial$$

or vice versa.

Setting a threshold ∂ serves two purposes. First, such extreme cases of 0 or 1 for allele frequencies are not well accommodated in the algorithms for sampling from a binomial distribution (e.g. rbinom() in R {stat}). The function rbinom() will consistently yield $p_A = 0$ for $f_A = 0$ when this is clearly not the case, and the poor approximation at extremes is accommodated by setting ∂ > 0. Second, true allele frequencies of 1:0 vs 0:1 is a true fixed difference and will always throws a positive in the sample set; we regard it as a true positive. But what of (1-∂):∂ vs 0:1 with ∂ vanishingly small? If this case throws a positive in the sample set, is it a false positive? In terms of indicating low levels of gene flow between populations A and B, an almost fixed difference (say, ∂ = 0.01) is arguably as informative as a strict fixed difference (∂ = 0). Any practical assessment would regard such a positive (with ∂ < 0.01) as a true positive.

Thus, to undertake the simulations, we need to make an operational decision on the value of ∂. This decision is likely to be controversial and case specific, so is left to individual researchers.

Implementation

These calculations have been implemented in dartR (v1.8) available in the CRAN repository. The function gl.fixed.diff now has the option to calculate p values for an observed fixed difference between two populations given the respective sample sizes and a decision on ∂. The function gl.collapse.pval will amalgamate populations or putative OTU for which the observed fixed differences are not significant at a specified level.

Examples

Example 1

Population A of the freshwater turtle *Emydura macquarii* from the Warrego River at Ambathella $(n_A = 2)$ in the northern basin of the Murray-Darling drainage (Australia) has 6 fixed differences in comparison with population B from the Lachlan River at Lake Forbes $(n_B = 10)$ in the southern basin. The comparison involved 2,025 polymorphic loci.

Applying the above calculations in a simulation with ∂ = 0.01 for 1,000 replications yielded an expected number of false fixed differences of 34 (±SD 3.4) and a probability that the observed count of fixed differences occurred by chance alone of *P* = 1.00. The observed fixed differences between population A (Ambathella) and population B (Lake Forbes) are not statistically significant. There is therefore no evidence that they belong to distinct operational taxonomic units (OTUs).

Example 2

Population A of the freshwater turtle *Emydura macquarii* from the Hunter River of SE coastal New South Wales $(n_A = 10)$ has 381 fixed differences in comparison with population B of SE coastal NSW and southern Queensland, extending from the Macleay River in the south to the Pine Rivers in the north $(n_B = 60)$. Population B arose from the

aggregation of sampling sites in the absence of fixed differences. The comparison involved 4,931 polymorphic loci.

Applying the above calculations in a simulation with ð = 0.01 for 1,000 replications yielded an expected number of false fixed differences of 18 (±SD 2.3) and a probability that the observed count of fixed differences occurred by chance alone of $P << 0.0001$. The observed fixed differences between population A (Hunter River) and population B (remaining SE Coast) are highly significant. There is therefore strong evidence to keep the Hunter River and the remaining coastal populations of SE coastal Australia as separate diagnosable operational taxonomic units (OTUs).