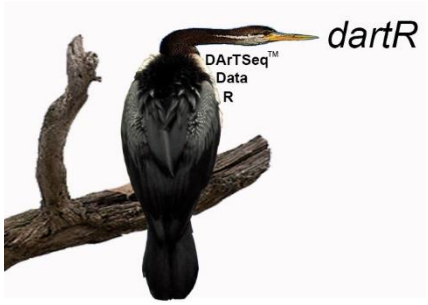
UNIVERSITY OF CANBERRA

SNP Analysis using dartR



TechNote: Fixed Difference Analysis

Version 2.0

Copies of these workshop notes are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

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

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

Fixed Difference Analysis

Introduction

The objective of a fixed difference analysis is to identify genetically diagnostic units in studies of species delimitation and phylogeography. First applied in species delimitation in grasses (Davis, Manos, & Davis, 1991; Davis & Nixon, 1992), the approach was subsequently applied in allozyme studies of animals (Georges & Adams, 1996; Georges, Adams, & McCord, 2002) and recently to SNP datasets (Georges *et al.*, 2018; Unmack *et al.*, 2019).

Typically, populations of a species or species complex do not mate randomly across the landscape. Barriers to dispersal of individuals may impede or prevent gene flow, and because of genetic drift and mutation these barriers result in population differentiation and hence 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 remarkably few individuals per generation is enough to prevent divergence of allelic profiles between randomly mating populations (Wright, 1931). The accumulation of fixed differences can result from both long-standing reproductive isolation or long-standing geographic isolation, and fixed differences cannot, on their own, distinguish between the two. Fixed differences are thus necessary but not sufficient to demonstrate reproductive isolation.

DartR provides a number of functions to conduct fixed difference analysis. In these Analysis Guidelines, we introduce fixed difference analysis as a means of identifying aggregations of populations into diagnosable lineages or Operational Taxonomic Units that can be considered further as either representing species or diagnostic lineages within species.

Fixed differences are Diagnostic

Fixed differences 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 (Moritz, Schneider, & Wake, 1992).

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.

Sympatry versus Allopatry

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 have in the past been regarded as different sympatric species, until examined using genetic tools (e.g. the butterflies Caeruleuptychia helios and Magneuptychia keltoumae; Nakahara *et al.*, 2017).

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 these 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 spatially, 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) or both, 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. It eliminates from consideration as species populations that may be distinctive based on their allelic profiles, but which lack fixed allelic differences. As such the fixed difference analysis is more conservative than pure phylogenetic approaches, and so works against taxonomic inflation.

Subjective considerations, taking into account all available evidence, will be required to decide which of the 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 however desirable this may be, is not possible in allopatry.

This will become clearer as we work though an analysis.

Paraphyletic Species

When a population diverges to the point of reproductive isolation (speciation) it may leave behind a series of populations, to which it previously belonged, in a paraphyletic relationship. One or more of the residual populations are sister to the new species, but belong to the same taxon as the remaining residual populations.

Such 'paraspecies' (Crisp & Chandler, 1996) occur when a diagnostic entity regarded as a species is nested within an aggregation of lineages that are not diagnosable or diagnosable only at the level of lineages within a species. Evidence for such paraspecies will become available when the OTUs arising from the fixed difference analysis are mapped against the phylogeny. Paraspecies are difficult to discover from the phylogeny alone, for both conceptual and practical reasons.

How do Fixed Differences Arise?

Fixed differences in allopatry can arise through drift whereby one isolated population loses a particular allelic state at a locus and the other isolate loses the alternate allele at that locus. This process clearly depends on the population sizes. Two large populations are less likely to accumulate fixed differences in allopatry than two small populations. The acquisition of a fixed difference via drift between a large population and a small

population is also unlikely. The small population may shed alleles via drift, but if the large population retains its diversity at that locus, then no fixed difference will emerge. Fixed differences that arise through drift, where one or both of the populations is large, would normally involve alleles that are already rare in the large population(s).

Populations that fluctuate in size through time require particular attention. If two large populations have historically experienced a bottleneck, then fixed differences may arise through drift at that time when population sizes are small. A similar case occurs with one large population that has experienced an historical bottleneck and the second population that is and has been small.

A second way in which fixed differences can arise is where a locus is monomorphic across all populations, and then is subject to mutation in one of the populations. If the mutation is advantageous or linked to an allele that is advantageous, then it can sweep through the population and ultimately come to fixation. In the absence of gene flow with other populations, this mutated SNP locus will be established as a fixed difference.

Either way, the acquisition of a fixed allelic difference is a significant biological event.

How are Fixed Differences Obliterated?

Fixed differences are maintained by very low levels of exchange between populations if established by drift, and no exchange between populations in the case of a new advantageous mutation or a mutation that hitchhikes as part of a linkage group that contains an advantageous allele. Significant geneflow between populations, perhaps as low as one individual migrating per generation (Wright, 1931), either contemporary gene flow or recent historical gene flow will obliterate fixed differences. The gene flow could be episodic.

Convergent mutation at a SNP locus could also conceivably obliterate a fixed difference at that locus, but two mutations at the one SNP site are likely to be rare. If they involve a different transition or transversion than the initial mutation, this will result in more than two alleles at that locus; for diploid species, such loci will have been screened out by the DArT pipelines.

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. False positive fixed differences arising by chance will introduce to the set of entities under consideration as species those arising from sampling error.

A Fundamental Asymmetry

There is an important 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 simple observation. 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 observed at all loci thus allow 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 confidently aggregated into a single OTU on the basis of lack of fixed allelic differences regardless of the sample size, eliminating them from consideration as distinct species (strictly, there is insufficient evidence to reject the null hypothesis of the two populations belonging to the one species).

In contrast, 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, even if the probability any one particular locus is very small.

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 percent 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 without genotyping every individual in each population.

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 for both the allopatric and sympatric case.

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.

The Sympatric Case

The above analysis assumes that the populations being compared pairwise are allopatric. The simulation draws samples from the two population allele frequency profiles separately. In the sympatric case, the null proposition is that the two putative taxa are one, that is, that they share allele frequency profiles in panmixia.

Clearly, a count of fixed differences in sympatry is much stronger evidence of actual diagnosability that is the same count of fixed differences in allopatry.

The approach to this is to recalculate the false positive rate and associated statistics for populations that are being compared in sympatry.

The test can be done with

```
result <- gl.fdsim(gl, poppair=c("popname1","popname2"),
     sympatric=TRUE, verbose=3)
```

False negatives

We have argued that the absence of fixed differences or corroborated fixed differences is unambiguous. That is, if there are no fixed differences between two samples taken from two population – they share alleles at all loci – then no amount of additional sampling will yield a fixed difference. The observed absence of fixed differences between two populations is sufficient to conclude that the populations from which they are drawn have no fixed differences.

All would be well if the SNPs were called with 100% accuracy, but they are not. Some level of error will sneak in. This can lead to a single spurious call in one population of the SNP state in the other leading to the fixed difference at that locus to be undetected.

There are two ways to manage this. The first is to filter on read depth to remove those loci that have been called on the basis sequence with low read depth. This will reduce the rate of miscalls. A suitable lower threshold for read depth might be 10x.

```
gl <- gl.filter.rdepth(gl,lower=10,verbose=3)
```

The second way to manage this is by simple imputation, that is for example, to set to 0, any single SNP called as 1 in a population that is otherwise homogeneous 0.

```
gl <- gl.impute(gl,method="simple",nthreshold=1,nmin=10,verbose=3)
```

This approach should only be used if sample sizes are adequate, say n $\geq$ 10.

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.

Note: A single F1 hybrid in the data will be sufficient to cause the amalgamation of the two parent populations into a single OTU, so careful consideration is required.

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.

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, if the number of fixed differences does not exceed the false positive rate).

Note: Because a population with a small sample size may fail to be significantly different from many other populations, some subjective judgement is required in deciding with which population it should be amalgamated. Geographic proximity might be a consideration, or the amalgamating with the nonsignificant population with the largest sample size might be the option. Some thought is required. If all your populations have samples $\geq$ 10, these considerations rarely arise.

In some cases, cycling between Test and Reiterate might be required.

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). Having managed the incidence of false positives, 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.

Impact of Low Sampling Intensity

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 can greatly distort or complicate the decisions on which OTUs represent species and which represent diagnosable lineages within species. The fixed difference

analysis is of greatest utility when sampling is comprehensive across the geographic range of the taxon/taxa under study.

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 or 1000s.

(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. A strategy might be to consider sample sizes and amalgamate populations with small sample sizes each with one with a large sample size, where the difference between the two does not significantly exceed the false positive rate. Then repeat the analysis.

Worked Example

As an example, let us consider a SNP data generated for a freshwater turtle from range of sites across northern Australia.

```
gl_nth <-
gl.load(file="Tutorial_dartR_fixed_difference_analysis.Rdata")
```

In `gl_nth`, we have the genotypes for individuals assigned to populations (sampling sites). These are labelled with their current putative assignment to species based on the species defined by Georges and Thomson (2010). The data have already been filtered, as follows:

```
gl_nth <- gl.filter.secondaries(gl_nth)
gl_nth <- gl.filter.callrate(gl_nth,threshold=0.95,v=3)
gl_nth <- gl.filter.reproducibility(gl_nth,threshold=0.995,v=3)
gl_nth <-
gl.filter.callrate(gl_nth,method="ind",threshold=0.8,v=3)
gl_nth <- gl.filter.monomorphs(gl_nth,v=3)
```

after having examined the data using the corresponding report functions to determine appropriate thresholds.

In preparation for the fixed difference analysis, sample sites at the boundary of two regions that show evidence of contemporary admixture have been removed. For example, examination of a PCA plot provided evidence of contemporary admixture between populations of *Emydura tanybaraga* from the Mitchell River in Queensland, west of the Great Dividing Range and *Emydura macquarii* from the Barron and Russell-Mulgrave rivers east of the range. Populations from these drainages were omitted from the fixed difference analysis. The Daly River of the Northern Territory has all three currently described species *Emydura victoriae*, *E. tanybaraga* and *E. subglobosa worrelli*, and evidence of admixture. The Daly River populations were excluded from the fixed difference analysis, to be examined separately later.

Sample sites with a low number of individuals have (n=1, n=2), where possible, been amalgamated with other sample sites within the same drainage basin.

This leaves us with 34 sites for the northern *Emydura*.

```
table(pop(gl_nth))
```

Emsub_Bamu	Emsub_Bensbach	Emsub_Fly	Emsub_Jardine	Emsub_Kikori	Emsub_Morehead
26	4	55	16	5	7
Emsub_Purari	Emsub_Vailala	Emsub_Vanapa	Emtan_Archer	Emtan_Blyth	Emtan_Darwin
3	23	8	4	12	11
Emtan_Holroyd	Emtan_Mitchell(Q)	Emtan_Pascoe	Emtan_Staaten	Emtan_Wenlock	Emvic_Carson
10	13	9	6	10	10
Emvic_Drysdale	Emvic_Dunham	Emvic_Fitzmaurice	Emvic_Fitzroy(WA)	Emvic_Isdell	Emvic_Mitchell(WA)
10	5	8	11	12	4
Emvic_Ord	Emvic_Pentecost	Emvic_Victoria	Emwor_Calvert	Emwor_Leichhardt	Emwor_Limmen
15	5	16	10	10	10
Emwor_Liverpool	Emwor_Macarthur	Emwor_Nicholson	Emwor_Roper		
10	8	33	19		

Note that most of the populations have respectable sample sizes.

Explore

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

```
D <- gl.dist.pop(gl_nth,v=3)
```

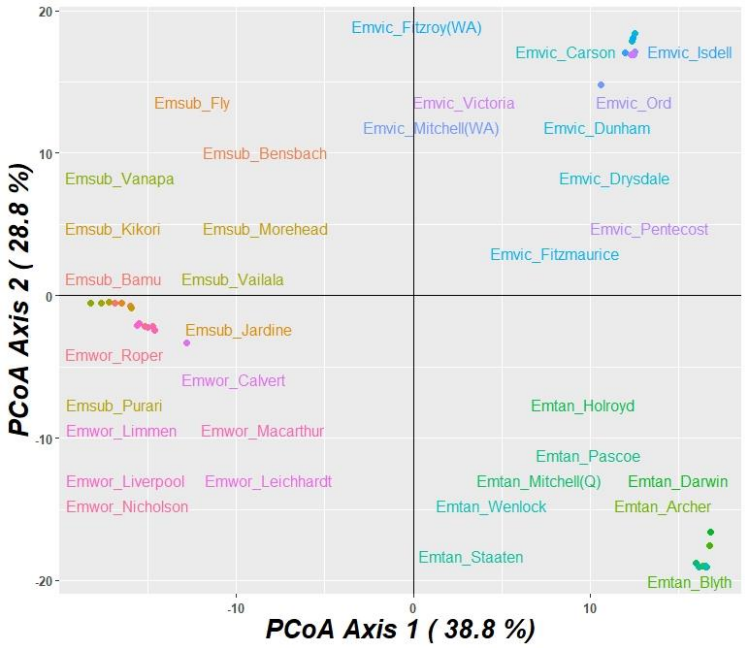
```
  Reporting inter-population distances
  Distance measure: euclidean
    No. of populations = 34
```

```
    Average no. of individuals per population = 12.29412
    No. of loci = 12334
    Miniumum Distance:  7.87
    Maximum Distance:  45.13
    Average Distance:  33.14
```

```
pcoa <- gl.pcoa(D,v=3)
```

```
   Performing a PCoA, individuals as entities, no correction applied
    Ordination yielded 4 informative dimensions from 34 original dimensions
    PCoA Axis 1 explains 38.8 % of the total variance
    PCoA Axis 1 and 2 combined explain 67.6 % of the total variance
    PCoA Axis 1-3 combined explain 80.9 % of the total variance
```

```
gl.pcoa.plot(pcoa,D,xaxis=1,yaxis=2)
```



Bit messy with the site labels, but there is considerable structure among sample sites evident in the top two dimensions of the ordination, and three major groupings corresponding to *Emydura tanybaraga*, *E. victoriae* and *E. subglobosa*. The question is, how distinct are these sample sites and is there structure within each. How many diagnosable taxa/lineages are there?

Compare

A first step in the fixed difference analysis is to calculate a matrix of fixed differences between the sample sites taken pairwise.

```
D <- gl.fixed.diff(gl_nth,v=4)
```

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

```
   D[[1]]$gl - the input genlight object;
   D[[2]]$fd - raw fixed differences (dist object);
   D[[3]]$pcfd - percent fixed differences (dist object);
   D[[4]]$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.

We can examine the fixed difference matrix

```
D$fd
```

		Emydura subglobosa									Emydura tanybaraga								Emydura victoriae										Emydura s. worrelli					
		Bamu	Bensbach	Fly	Jardine	Kikori	Morehead	Purari	Vailala	Vanapa	Archer	Blyth	Darwin	Holroyd	Mitchell(Q)	Pascoe	Staaten	Wenlock	Carson	Drysdale	Dunham	Fitzmaurice	Fitzroy(WA)	Isdell	Mitchell(WA)	Ord	Pentecost	Victoria	Calvert	Leichhardt	Limmen	Liverpool	Macarthur	Nicholson
E. subglobosa	Bensbach	0																															
	Fly	0	0																														
	Jardine	2	13	1																													
	Kikori	0	4	0	6																												
	Morehead	0	0	0	2	1																											
	Purari	0	5	0	17	4	4																										
	Vailala	0	1	0	6	0	0	0																									
	Vanapa	11	66	15	84	35	52	55	0																								
E. tanybaraga	Archer	737	825	619	682	929	748	997	859	1235																							
	Blyth	777	864	650	717	975	788	1046	902	1298	7																						
	Darwin	667	756	554	635	857	676	930	792	1162	4	0																					
	Holroyd	728	798	613	662	897	727	970	839	1198	2	7	5																				
	Mitchell(Q)	606	667	510	563	751	603	815	701	1038	1	5	3	0																			
	Pascoe	763	839	642	690	939	752	1009	873	1249	3	10	3	1	0																		
	Staaten	741	813	626	670	914	734	980	851	1221	2	10	5	0	0	0																	
	Wenlock	711	782	597	646	874	713	946	810	1179	2	9	2	1	1	1	1																
E. victoriae	Carson	639	738	552	720	820	675	885	747	1138	902	895	701	863	783	907	901	868															
	Drysdale	679	782	589	765	860	720	925	783	1184	950	940	743	906	822	952	946	909	1														
	Dunham	611	711	525	692	791	649	859	717	1110	849	841	662	815	735	862	844	821	26	42													
	Fitzmaurice	631	734	547	714	820	674	880	743	1128	863	841	674	824	746	867	862	826	44	72	1												
	Fitzroy(WA)	617	719	528	696	796	655	859	721	1115	882	871	676	842	753	885	878	847	2	1	26	41											
	Isdell	606	703	523	679	771	639	838	702	1090	870	865	674	832	746	874	869	831	17	21	31	47	3										
	Mitchell(WA)	653	758	562	733	832	695	902	760	1152	923	915	731	884	795	927	912	883	56	80	154	170	63	92									
	Ord	575	677	495	659	751	616	817	682	1061	811	805	632	780	705	825	819	789	16	27	0	0	20	22	130								
	Pentecost	646	752	561	733	834	690	902	760	1156	879	871	688	842	769	891	883	847	26	47	0	1	25	38	157	0							
	Victoria	593	696	513	678	775	638	839	706	1087	808	802	635	778	701	819	812	781	33	56	0	0	25	37	154	0	0						
E.s. worrelli	Calvert	101	157	60	190	207	112	265	178	417	550	563	503	517	448	540	533	515	670	711	636	657	652	634	688	606	673	615					
	Leichhardt	134	199	77	234	242	140	310	217	491	907	968	866	870	744	915	881	871	889	929	844	865	861	841	903	807	886	825	6				
	Limmen	130	205	80	247	250	149	318	219	490	920	985	884	876	766	943	904	883	901	942	861	879	872	854	912	820	904	839	1	6			
	Liverpool	126	198	78	243	252	144	315	215	483	941	999	888	908	776	958	928	898	897	937	866	887	876	856	917	827	905	843	4	7	2		
	Macarthur	110	172	71	201	217	124	276	194	429	756	819	715	737	642	784	760	751	794	834	751	774	759	743	804	712	789	731	0	7	0	3	
	Nicholson	89	134	54	162	168	94	209	146	366	654	708	618	629	546	667	651	637	696	731	648	673	673	656	705	620	690	636	0	0	0	4	0
	Roper	98	156	60	188	191	110	245	165	401	754	794	703	726	628	768	744	726	760	796	720	738	729	710	774	686	757	699	0	2	0	1	0

Note that there are quite a few comparisons with zero fixed differences or only one (uncorroborated) fixed differences. These are candidates to aggregate on the basis of presenting no diagnostic allelic differences.

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,tpop=1,verbose=3)
```

```
New population groups

Group:Emsub_Bamu+

[1] "Emsub_Bamu" "Emsub_Bensbach" "Emsub_Fly" "Emsub_Jardine" "Emsub_Kikori"
"Emsub_Morehead" "Emsub_Purari" "Emsub_Vailala" "Emsub_Vanapa"


Group:Emtan_Blyth+

[1] "Emtan_Blyth" "Emtan_Darwin" "Emtan_Archer" "Emtan_Mitchell(Q)"
"Emtan_Staaten" "Emtan_Holroyd" "Emtan_Pascoe" "Emtan_Wenlock"


Group:Emvic_Carson+

[1] "Emvic_Carson" "Emvic_Drysdale" "Emvic_Fitzroy(WA)" "Emvic_Isdell"
```

```
Group:Emvic_Dunham+
[1] "Emvic_Dunham" "Emvic_Ord" "Emvic_Pentecost" "Emvic_Victoria"
"Emvic_Fitzmaurice"


Group:Emwor_Calvert+
[1] "Emwor_Calvert" "Emwor_Limmen" "Emwor_Macarthur" "Emwor_Nicholson"
"Emwor_Roper" "Emwor_Leichhardt" "Emwor_Liverpool"
```

There are 5 aggregations of sample sites, each aggregation comprising sites that, when compared pairwise, have no corroborated fixed allelic differences (`tpop=1`) 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
```

	Emsub_Bamu+	Emtan_Blyth+	Emvic_Carson+	Emvic_Dunham+	Emvic_Mitchell(WA)	
Emtan_Blyth+	152					
Emvic_Carson+	256	313				
Emvic_Dunham+	282	334	1			
Emvic_Mitchell(WA)	402	450	31	93		
Emwor_Calvert+	13	130	296	320	442	

There are two things of note here. The first is that, even though we aggregated sample sites on the basis of no corroborated fixed differences, the outcome has some pairs of aggregations that still have no corroborated fixed differences (e.g. the Emtan_Carson+ aggregation and the Emtan_Dunham+ aggregation 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 a few fixed differences (e.g. Emwor_Calbert+ vs Emsub_Bamu+). The question that arises is, are these false positives arising from the finite sample sizes? We can examine this later on.

We run the collapse process one more time.

```
D3 <- gl.collapse(D2,tpop=1,verbose=3)
New population groups
Group:Emvic_Carson++
[1] "Emvic Carson+" "Emvic Dunham+"


D3$fd
```

	Emsub_Bamu+	Emtan_Blyth+	Emvic_Carson++	Emvic_Mitchell(WA)
Emtan_Blyth+	152			
Emvic_Carson++	197	252		
Emvic_Mitchell(WA)	402	450	28	
Emwor_Calvert+	13	130	235	442

And that is as far as the collapsing of the fixed difference matrix can go, with **tpop** set to 1.

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 4.

```
Populations, aggregations and sample sizes
  Emsub_Bamu+ Emtan_Blyth+ Emvic_Carson++  Emvic_Mitchell(WA) Emwor_Calvert+
          147           75             92                   4             100
```

With such low sample sizes, and the number of loci being considered, it is possible that the 28 fixed differences observed between, say, Emvic_Carson++ and Emvic_Mitchell(WA) with n=4 arose by sampling error.

This concern can be accommodated by testing the observed differences for significance.

```
D4 <- gl.fixed.diff(D3,test=TRUE,v=3)
```

As this script will take a long time to run, you might like to add the parameter **nreps=100** for sake of illustration.

```
Comparing populations pairwise
Emvic_Carson++ vs Emvic_Mitchell(WA)   [p = 0.2434 ,ns]
```

So as we suspected, the 28 fixed allelic differences between Emvic_Carson++ and Emvic_Mitchell(WA) does not exceed the false positive rate, given the exceptionally low sample sizes. Note however that the 13 fixed allelic differences between Emsub_Bamu+ and Emwor_Calvert+ did significantly exceed the false positive rate based on sample sizes of 147 and 100 respectively.

Now all that remains is to amalgamate the non-significant Emvic_Carson++ and Emvic_Mitchell(WA) and prepare a final summary.

```
gl.final <-
gl.merge.pop(D4$gl,old=c("Emvic_Carson++","Emvic_Mitchell(WA)"),ne
w="Emvic_Carson+++")

D5 <- gl.fixed.diff(gl.final,v=3)

Populations, aggregations and sample sizes
    Emsub_Bamu+    Emtan_Blyth+ Emvic_Carson+++  Emwor_Calvert+
            147              75              96             100

D5$fd
```

	Emsub_Bamu+	Emtan_Blyth+	Emvic_Carson+++
Emtan_Blyth+	152		
Emvic_Carson+++	166	223	
Emwor_Calvert+	13	130	198

Summary

A fixed difference analysis was applied to 34 populations of freshwater turtle in northern Australia (*Emydura*) to see if there was any evidence to challenge the null hypothesis that they comprised a single species. Four taxa are currently recognised – *Emydura subglobosa subglobosa*, *Emydura subglobosa worrelli*, *Emydura tanybaraga* and *Emydura victoriae*. Is there evidence of cryptic taxa within these currently recognised species and subspecies?

Five diagnosable OTUs were identified on the basis of corroborated fixed allelic differences. The distinction between two of these (*Emydura victoriae* [Carson River aggregation] and *Emydura victoriae* [Mitchell River (WA)] did not significantly exceed the false positive rate (p = 0.2434) and were amalgamated to yield four significantly diagnosable OTUs.

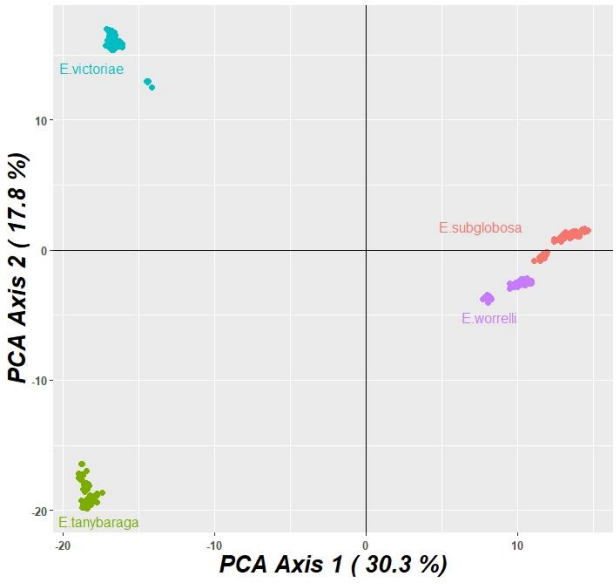
There was no evidence to reject the null hypothesis represented by the existing taxonomy of the three northern species of *Emydura*. The subspecies of *Emydura subglobosa* were diagnosable, but marginally with 13 fixed differences in comparison with the 130-223 fixed differences between the other taxa.

A graphical representation of the genetic divergence among the taxa is provided in the form of a PCoA.

```
popNames(gl.final) <-
c("E.subglobosa","E.tanybaraga","E.victoriae","E.worrelli")

pcoa_final <- gl.pcoa(D5)

gl.pcoa.plot(pcoa_final,gl.final)
```



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.

In the above example, we sidestepped this issue by asking if the fixed difference analysis challenged the currently accepted taxonomy. But what if you are approaching this *de novo.* All species are lineages, but not all lineages are species.

A first step in making the distinction between lineages that are to be regarded as species and lineages considered to represent diversity within species 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.

Choices can be made on how stringent to be in filtering. One non-standard filtering option might be to filter out loci that are not supported by a read depth of 10 or more. This will reduce the number of loci to work with, but increase the reliability of the SNP calling.

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 if aggregation is not to occur. 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., 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]

Other References

Crisp M & Chandler G. 1996. Paraphyletic species. *Telopea* 6: 813–844.

Davis JI, Manos PS & Davis I. 1991. Isozyme variation and species delimitation in the Puccinellia nuttalliana complex (Poaceae): an application of the phylogenetic species concept. *Systematic Botany* 16: 431–445.

Davis JI & Nixon KC. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology* 41: 421–435.

Georges A, Gruber B, Pauly GB, White D, Adams M, Young MJ, Kilian A, Zhang X, Shaffer HB & Unmack PJ. 2018. Genomewide 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.

Georges A & Adams M. 1996. Electrophoretic delineation of species boundaries within the short-necked freshwater turtles of Australia (Testudines: Chelidae). *Zoological Journal of the Linnean Society* 118: 241–260.

Georges A, Adams M & McCord W. 2002. Electrophoretic delineation of species boundaries within the genus Chelodina (Testudines: Chelidae) of Australia, New Guinea and Indonesia. *Zoological Journal of the Linnean Society* 134: 401–421.

Georges A & Thomson S. 2010. Diversity of Australasian freshwater turtles, with an

annotated synonymy and keys to species. *Zootaxa* 37: 1–37.

Moritz C, Schneider CJ & Wake DB. 1992. Evolutionary relationships within the ensatina eschscholtzii complex confirm the ring species interpretation. *Systematic Biology* 41: 273–291.

Nakahara S, Zacca T, Huertas B, Neild AFE, Hall JPW, Lamas G, Holian LA, Espeland M & Willmott KR. 2017. 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.

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

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97–159.

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 true 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 reference 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 reference 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

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

so for one OR the other

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

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

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

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

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

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

$$\mu = f_A \quad \text{..... (14)}$$

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

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 (v1.9.1 and later) 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 non-significant pairs can be amalgamated manually with the function gl.merge.pop, which requires a decision on which pair to amalgamate when one population with a small sample size does not differ significantly from a number of other populations.

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), and the null proposition is adopted.

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 ∂ = 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).