Population Genetic Analysis of ISSR Data

Theresa M. Culley

Dominant markers, such as RAPDs and ISSRs, have become popular in molecular studies in recent years. As an alternative to isozymes, these types of markers are especially attractive given their hypervariable nature, the vast numbers of loci that can be examined, and the small amount of fresh or dried material used per sample. However, one significant limitation is that in population genetic studies, analysis of dominant data is not as straightforward as for codominant markers such as isozymes. Consequently, the accuracy of estimations using dominant markers is usually reduced relative to codominant markers (Lynch and Milligan 1994).

Aside from assumptions of band homology (i.e., marker alleles/bands shared by two individuals descend from a common ancestor) and diallelic loci, the main reason that ISSRs and RAPDs are difficult to analyze statistically is because of their dominant nature. This makes it difficult to calculate allele frequencies that are used to generate many common genetic statistics (e.g., H_e, F_{st}, θ, G_{st}). With dominant data, the presence of a band can denote *either* a dominant homozygote (band present/present) or a heterozygote (band present/absent). This makes it impossible to calculate the frequency of the dominant allele directly from gels (as you would with codominant data because you can see both dominant and recessive alleles). The traditional technique is to determine allele frequencies from the number of band absences, assuming Hardy-Weinberg equilibrium. For example, the frequency of the absent or null allele (q) is first calculated from the observed number of individuals without a band (q^2). The frequency of the dominant allele (p) is calculated as (1-q). Both allele frequencies (p and q) are then used as they would be in the analysis of codominant data. If F_{is} (Wright's inbreeding coefficient within populations) is known from a previous study with codominant markers, an alternative approach is to calculate allele frequencies at a locus using the equation:

$$P_{aa} = q^2(1-F_{is}) + qF_{is}$$

where P_{aa} is the expected frequency of the homozygous recessive genotype (i.e., denoted by band absence) (see Lynch and Milligan 1994; B. Weir, personal communication). This method does not involve assumptions of Hardy-Weinberg equilibrium because F_{is} is used (F_{is} can be thought of as a measure of deviation from H-W; if $F_{is} = 0$, then the equation reduces to H-W conditions). One computer program that can use this method is POPGENE.

There are several problems with these approaches for dominant data such as ISSRs. First, Hardy-Weinberg equilibrium cannot be measured directly and could easily be violated in ISSR loci (as sometimes happens in isozyme loci). Second, shared band absences should not be used because they may not always indicate a homozygous recessive genotype. Band absence may be caused by a number of different factors, such as loss of a primer annealing site (because of nucleotide sequence differences), insertions or deletions in the fragment between the two primer sites, or experimental error (although this can be minimized by running replicate gels). Consequently, it may be unlikely that the absence of bands in two individuals arose from an identical ancestral mutation (i.e., that recessive alleles are homologous). The use of shared band absences may overestimate the relatedness among individuals, and they should be used with caution in population analyses and should never be used above the species level (Black 1997).

Band absences can also overwhelm the signal coming from positive band matches (D. Crawford, pers. comm.); this would be particularly undesirable if absences did not derive from identical ancestral mutations.

There are at least four methods that address these problems:

1. **Lynch and Milligan** (1994)
   - this method incorporates the sampling variance of the frequency of null homozygotes into calculations of allele frequencies, providing an estimate of Wright's F_{st}, as well as genetic distance. However, this method *still* uses band absences, assumes that genotype frequencies are in Hardy-Weinberg equilibrium, and assumes that mating is random ($F_{is} = 0$). If this method is to be used, consult Lynch and Milligan (1994) for recommendations on loci pruning and sample numbers.

   Software: RAPDFst will give Lynch and Milligan's F_{st} along with Weir and Cockerham's θ, and Wright's F_{st}.

The next three methods avoid calculating allele frequencies (assuming H-W) altogether:

2. **Band matching similarity coefficients**
   - This method incorporates only bands that are present. These statistics, like the coefficients of Jaccard (1908) and Nei and Li (1979) [same as Dice (1945) and Sørensen (1948)], compare the number of bands shared between individuals or populations:

		Individual A 1	0
Individual B	1	a	b
	0	c	d

1 = band present
0 = band absent

Jaccard's coefficient: $J = \frac{a}{a + b + c}$

Dice / Sørensen / Nei and Li's coefficient: $D/S = \frac{2a}{2a + b + c}$ $N = \frac{2a}{(a + b) + (a + c)}$

Example What is the similarity coefficient for two individuals with the following data matrix?

Indiv. A 011000101011 where: a = 3
Indiv. B 010100100101 b = 2
c = 3

Jaccard's coefficient: J = 3 / (3 + 2 + 3) = 0.375
Dice's coefficient: S = 2(3) / [2(3) + 2 + 3] = 0.545
Nei and Li's coeff.: N = 2(3) / [(3+2) + (3 + 3)] = 0.545

Distances can be calculated from any of these coefficients as (1-similarity). Jaccard's coefficient is the most simple, but the coefficient of Nei and Li puts more weight on positive matches. These methods are advantageous because band absences can be excluded from analyses and there are no assumptions of Hardy-Weinberg equilibrium. However, there are several similarity coefficients that incorporate band absences (d) into their formulas (see Sneath and Sokal 1973, p. 131) and these should be used with caution. Examples are as follows:

Simple Matching coefficient: $SM = \frac{(a + d)}{(a+b +c + d)}$

Yule coefficient: $Y = \frac{(ad - bc)}{(ad + bc)}$

Baroni-Urbani Buser coefficient: $BUB = \frac{[\sqrt{(ad)} + a]}{[a + b + c + \sqrt{(ad)}]}$

Software:

- Vera Ford's programs - calculate Nei and Li/Dice/Sørensen's coefficient either among all pairs of individuals as a distance measure (using !WXDNL), or averaged across populations as a similarity measure (using !WAVSIML). For more information on the formulas used, see Crawford et al. (1998).
- MVSP - calculates many different distance/similarity measures, including coefficients of Jaccard, and Nei and Li.
- NTSYS - calculates many coefficients, including Jaccard and Dice
- RAPDPLOT – calculates the Nei and Li coefficient and allows the user to test for support of phenogram branches using bootstrapping.
- PAUP – calculates the *restriction site* Nei and Li coefficient, but note that this is *not* the same as the Dice coefficient!!!

3. AMOVA – Analysis of Molecular Variance

– this method partitions observed variation into within and among population components using genetic distances; the among population variance component is called ϕ_{st}, an analog to Fst and θ. The method was not originally designed to analyze purely dominant data (RAPD or AFLP), and presently the data must be treated as molecular haplotypes (i.e., data assumed to be made up of completely linked sites with no recombination), rather than RAPD data (data made up of independent sites with linkage equilibrium). As such, the resulting population genetic structure indices (ϕ_{st}) are not yet comparable with those obtained for codominant markers (θ; see Stewart and Excoffier 1996). Both programs listed below are presently being developed to adequately handle RAPD data.

Software:

- Arlequin
- AMOVA-PREP / AMOVA1.55 – the first program prepares data for entry into AMOVA1.55 (previously WINAMOVA) by first creating a distance matrix. Presently, you can choose among the following genetic distances: Euclidean distance, non-Euclidean distance, and the simple-matching coefficient (which

incorporates band absences). The non-Euclidean distance is really that of Nei and Li (see Huff et al. 1993):

$$D = 100\left(1 - \frac{2\,n_{xy}}{n_x + n_y}\right)$$

where n_x and n_y are the number of bands observed in individuals x and y, and n_{xy} is the number of bands shared by the two individuals.

AMOVA 1.55 then reads the distance matrix and completes the analysis. An option in this program is Bartlett's statistic, which is a measure of variance heterogeneity (e.g., for comparing how variable pair-wise genetic distances are within each of two populations).

4. Shannon-Weaver Diversity Index

- this method was adopted from community ecology and has been used to describe species richness. It does not assume H-W, but does presume that diversity estimates based on band phenotypes (bands present/absent) approximate genetic diversity (see Whitkus et al. 1998) – as do the band-matching similarity coefficients. Whitkus et al. (1998) use the Brillouin formula to eliminate bias associated with finite sample sizes. One caveat is that this technique does incorporate shared band absences.

Software: MSVP calculates the Shannon Index (also the Simpson and Brillouin's indices).

Common Questions:

Why can't I use common descriptive statistics like A, A_p, H_e, or H_o for my ISSR data?

Genetic variation statistics like the ones you mentioned should not be used for dominant data because they are based on observed heterozygosity or allele frequencies. As mentioned previously, heterozygotes are indistinguishable from dominant homozygotes because they are both represented by a band. In addition, it is not possible to calculate allele frequencies from dominant data, except by using band absences (not a good thing!). Presently, the only descriptive statistic that can be used is percent polymorphic loci (P), if it is calculated directly from the data (i.e., search thru the loci and count which ones contain both present and absent bands in a group of individuals). However, this may be an underestimate in cases when a band at a given locus is present in all individuals, because it is unknown if heterozygotes are present (which could make the locus polymorphic).

If I want to measure population differentiation with a dominant marker, which statistic is best to use?

At the present time, genetic distances using band-matching similarity coefficients (e.g., Jaccard, or Nei and Li) might be the best technique because it does not use band absences or assume Hardy-Weinberg equilibrium. For dominant data, G_{st}, F_{st}, and θ are not advisable because they are based on allele frequencies. Lynch and Milligan's F_{st} is a better estimate for dominant data, but it still uses band absences and assumes Hardy-Weinberg equilibrium. Other

genetic distance measures, such as Nei's genetic distance, are also built on allele frequencies. The Shannon-Weaver diversity index is relatively suitable for dominant data, but is not frequently used and there are few studies (as of yet) with which to compare your results. The AMOVA method (using non-Euclidean distances) is very promising, but at the present time, the results are not directly comparable with those obtained with codominant markers. This method should be monitored for future developments.

NOTE: Many studies of genetic variation (usually with isozymes) have used both F_{st} (or θ) and Nei's genetic distance. It is debatable whether both measures should be calculated on the same data set because of the different assumptions in their underlying models. For example, the derivation of θ assumes that genetic divergence between two groups is due to drift only, while Nei's genetic distance assumes that both drift and mutation are important. The former statistic has been called a short-term genetic distance (Reynolds, Weir, and Cockerham 1983) because it measures divergence on a short evolutionary time scale, while Nei's genetic distance addresses longer evolutionary time.

Common Statistics of Genetic Variation

	Statistic	Calculated from: allele frequencies.	Calculated from: observed heterozygotes	Calculated from: band-matching	Appropriate to Use With: Isozymes (codominant)	Appropriate to Use With: ISSRs (dominant)	Comments
Descriptive Statistics	A	X			Yes	No	
	A_p	X			Yes	No	
	P	X		X	Yes (allele freq.)	Yes (band-matching)	
	H_o		X		Yes	No	
	H_e	X			Yes	No	sometimes estimated for dominant markers
Population Differentiation	F_{st} (F_{is}, F_{it})	X	X		Yes	No	
	θ	X			Yes	No	sometimes estimated for dominant markers (Lynch & Milligan)
	ϕ_{st}	X		X	Not used	Yes (band-matching)	calculated from genetic distances
	G_{st}	X			Yes	No	
	Genetic Distance	X		X	Yes (allele freq.)	Yes (band-matching.)	sometimes estimated for dominant markers (Lynch & Milligan)
	Shannon Index			X	Not used	Yes	

A = number of alleles per locus
A_p = number of alleles per polymorphic locus
P = percentage of polymorphic loci
H_o = proportion of observed heterozygotes
H_e = proportion of expected heterozygotes (under H-W equilibrium; equals 2pq in two allele-case)
F_{st} = proportion of the total diversity that is partitioned among populations (Wright, 1951)
θ = analog of F_{st}, incorporates effects of small and unequal sample/population sizes (Weir and Cockerham, 1984)
ϕ_{st} = analog of F_{st}, generated from an AMOVA (based on genetic distances)
G_{st} = analog of F_{st}, but calculated differently – assumes H-W (see Nei, 1973)

For a description of most of these statistics, see Berg and Hamrick (1997).

SOFTWARE INFORMATION:

Program	Platform	Availability
Arlequin	WIN	http://anthropologie.unige.ch/arlequin
AMOVA1.55	WIN	FTP at: http://anthropologie.unige.ch/ftp/comp/win/amova
AMOVA-PREP	WIN	http://herb.bio.nau.edu/~miller/amovaprep.htm
MVSP	DOS/WIN	Exeter Software at: http://www.ExeterSoftware.com/cat/mvsp.html $140 (demo version available)
NTSYS	WIN	Exeter Software at: http://www.ExeterSoftware.com/cat/ntsyspc.html $275 (educational discounts available)
PAUP	MAC/WIN	Sinauer Associates, Inc. at: http://www.lms.si.edu/PAUP/ $85 for WIN/DOS, $100 for MAC
POPGENE	WIN	http://www.ualberta.ca/~fyeh/index.htm
RAPDFst	DOS	FTP at lamar.colostate.edu/pub/wcb4
RAPDPLOT	DOS	FTP at lamar.colostate.edu/pub/wcb4
V. Ford's programs	DOS	anonymous FTP at: 140.254.12.151 in folder incoming/ISSR

LITERATURE CITED:

Berg EE, Hamrick JL (1997) Quantification of genetic diversity at allozyme loci. *Canadian Journal of Forest Research* 27: 415-424.

Black W (1997) Explanation of RAPDPLOT 3.0. Documentation file distributed with the RAPDPLOT program via ftp at lamar.colostate.edu/pub/wcb4.

Crawford DJ, Esselman EJ, Windus JL, Pabin CS (1998) Genetic variation in running buffalo clover (*Trifolium stoloniferum*: Fabaceae) using random amplified polymorphic DNA markers (RAPDs). *Annals of the Missouri Botanical Garden* 85: 81-89.

Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology* 26: 297-302.

Huff DR, Peakall R, Smouse PR (1993) RAPD variation within and among natural populations of outcrossing buffalograss [*Buchloë dactyloides* (Nutt.) Engelm.]. *Theoretical and Applied Genetics* 86: 927-934.

Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles* 44: 223-270.

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. Molecular Ecology 3: 91-99.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Science USA* 76: 5269-5273.
Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105: 767-779.
Sneath PHA, Sokal RR (1973) *Numerical Taxonomy*. WH Freeman and Company, San Francisco, California, USA.
Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skr.* 5: 1-34.
Stewart CN, Excoffier L (1996) Assessing population genetic structure and variability with RAPD data: application to *Vaccinium macrocarpon* (Amercian Cranberry). *J. Evol. Biol.* 8: 153-171.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.
Whitkus R, de la Cruz M, Mota-Bravo L, Gomez-Pompa A (1998) Genetic diversity and relationships of cacao (*Theobroma cacao* L.) in southern Mexico. *Theoretical and Applied Genetics* 96: 621-627.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics* 15: 323-354.