A guide to software packages for data analysis in molecular ecology

Andrew SCHNABEL[1], Peter BEERLI[2], Arnaud ESTOUP[3], David HILLIS[4]

[1]*Department of Biological Sciences, Indiana University South Bend, South Bend, Indiana 46615 USA*

[2]*Department of Genetics, University of Washington, Seattle, Washington 98195 USA*

[3] *Laboratoire de Modélisation et de Biologie Evolutive, 488 rue croix Lavit, URBL-INRA, 34090 Montpellier, France*

[4]*Department of Zoology and Institute of Cellular and Molecular Biology, University of Texas, Austin, Texas 78712 USA*

Abstract. We briefly discuss software packages for the analysis of molecular ecological data, focusing on three levels of analysis: parentage and relatedness, population genetic structure, and phylogeny reconstruction. For the first two levels of analysis, we have gathered lists of some of the packages that we consider to be the most useful and user-friendly. For each package, we provide information on names of authors, date of latest update, compatible operating systems, types of data handled and analyses supported, availability, and literature citations. For software packages dealing with phylogeny reconstruction, we refer the reader to specific literature and website sources where this information has already been compiled.

1. Introduction

Molecular ecologists use protein or DNA markers to address questions about interactions between organisms and their biotic and abiotic environments. These studies often result in the generation of large and complex molecular data sets, and one of the challenges facing many workers is how to analyze those data properly. In this chapter, we present summary information on several of the numerous computer software packages for the analysis of genetic relationships among individuals, populations, and species. We do not claim that this information is complete, because or perfectly up-to-date, because new programs and updates of older programs are appearing almost monthly. Although some overlap will inevitably exist between different levels of analysis, we have chosen to divide the summary into three areas: parentage and relatedness, population genetic structure and gene flow, and phylogeny reconstruction.

2. Relationships among individuals: parentage and relatedness

Our understanding of social systems, mating behaviours, correlates of reproductive success, and dispersal patterns in natural populations depends on the possibility of genetically differentiating individuals, assigning both male and female parentage to individual progeny, and estimating with sufficiently high precision the genetic relatedness between groups or pairs of interacting individuals (Queller & Goodnight 1989; Cruzan 1998; Parker *et al*. 1998; Estoup, this volume). Studies of plant populations often use parentage analyses to address questions of outcrossing rates,

genetic relatedness per se (Schnabel, this volume). On the other hand, in animal studies, relatedness and parentage are linked through studies of altruistic behavior, social and genetic mating systems, and kin selection (Hughes 1998; Rico, this volume). Although polymorphic genetic markers have been used for a long time in cases when pedigree information must be ascertained, as in animal breeding selection programs or in human paternity analysis, the advent of molecular markers with high levels of polymorphism has opened new perspectives for studies of parentage and relatedness in natural populations (Queller *et al*. 1993; Avise 1994; Estoup *et al*. 1994; Morin *et al*. 1994; Blouin *et al*. 1996; Taylor *et al*. 1997; Aldrich and Hamrick 1998; Hughes 1998; Parker *et al*. 1998; Prodöhl *et al*. 1998).

Compared with the number of software packages available for higher levels of analyses (see below), very few programs are available for the analysis of parentage or for the estimation of genetic relatedness (Appendix 1). Written specifically for plants, the set of programs by Ritland (1990) is the most widely used package for the analysis of outcrossing rates. More detailed parentage analyses are possible with PollenFlow (JD Nason, unpublished), which combines paternity exclusion analyses with the fractional paternity model of Devlin *et al*. (1988) and the maximum-likelihood models of Roeder *et al*. (1989) and Devlin & Ellstrand (1990), such that the user is able to obtain estimates both of pollen gene flow into the study population and relative fertilities of all possible male parents (Schnabel, this volume). A similar approach to parentage inference is taken in CERVUS (Marshall *et al*. 1998), which implements the likelihood models of Thompson (1975, 1976) and Meagher (1986). A very simple approach is taken by Danzmann (1997) in PROBMAX, which calculates probabilities that individuals are the offspring of specific parental pairs. The probability values indicate the number of loci sampled for each progeny that conform to Mendelian expectations for the pair of parents being tested. Finally, two well-developed programs are available for the estimation of relatedness based on the models of Queller and Goodnight (1989). The package Kinship tests hypotheses of pedigree relationships between pairs of individuals, and Relatedness uses a regression technique to measure relatedness between groups of individuals.

3. Relationships among populations: population genetic structure and gene flow

Many questions asked by molecular ecologists require that they conduct a survey of genetic diversity across several populations of a species. Such surveys estimate how much genetic variation a particular species maintains within its populations for a particular set of molecular markers (e.g., allozymes) and how that variation is partitioned among populations. Based on these data, inferences can be made about effective population sizes, natural selection, patterns of mating and dispersal, gene flow, and biogeographical history of the populations (e.g., Gentile & Sbordoni 1998; Godt & Hamrick 1998; Gonzalez *et al*. 1998; Xu *et al*. 1998). Hundreds of surveys of population genetic structure can be found in the literature, most of which prior to 1990 used either allozymes or mitochondrial DNA restriction sites (Hamrick & Godt 1989; Avise 1994). In the past decade, however, a large and rapidly growing number of studies have used a wider variety of molecular markers, such as DNA sequences, RAPDs, AFLPs, and microsatellites (e.g., Arden & Lambert 1997; Fischer & Matthies 1998; Paetkau *et al*. 1998; Winfield *et al*. 1998).

Given this great abundance of studies, it is not surprising that the number of available software packages for the analysis of population genetic structure is also large. Most of the early programs were written with allozymes in mind, and several of the currently available packages are still limited in the types of data that can be handled. In contrast to software for phylogenetic reconstruction (see below), there is no single source either in print or as a website that brings all of this information together. In collecting this diverse array of programs, we found that many of the more user-friendly programs (e.g.,

analyses they perform (Appendix 2). First, several packages calculate basic statistics of genetic variation, such as the proportion of polymorphic loci, the average number of alleles per locus, and heterozygosity. Those programs that handle a wider variety of data types also calculate statistics such as nucleotide diversity. Second, many packages will conduct tests for Hardy-Weinberg equilibrium. Third, most of the programs we report will estimate patterns of genetic stucturing using the hierarchical approach of Wright and/or Cockerham and Weir. A smaller proportion of the programs also include methods for analyzing microsatellite data using R_{ST} (e.g., Arlequin, Fstat, GENEPOP, RSTCALC) or Analysis of Molecular Variance (AMOVA in Arlequin). Fourth, a several of the programs will calculate one or more pairwise genetic distance measures (e.g., Nei's distance, Rogers distance), and will analyze those distances using some sort of clustering algorithm (e.g., UPGMA or neighbor joining). Last, several of the programs will estimate the level of linkage disequilibrium between loci. Although a number of home-grown programs for Macintosh computers must certainly exist, the large majority of packages we found were written for either a DOS or Windows platform. On a final note, the best available program for analyzing genetic structure within hybrid zones appears to be Analyse by Barton & Baird (1998).

All of the programs in Appendix 2 work within a traditional Wrightian framework of geographic structuring and estimation of gene diversity statistics based on allele frequencies. The introduction of coalesence theory by Kingman (1982) created new methods for analyzing population data (Hudson 1990; Beerli, this volume). Coalescence theory focuses on the sampled gene copies and looks backward in time to calulate the probability that two randomly chosen gene copies in the sample have a common ancestor *t* time units in the past. This process is driven only by the effective population size, N_e, and mutation rate. Kingman (1982) showed that the time when all lineages coalesced for a sample of 2, 4, and infinite gene copies is $2N_e$, $3N_e$, and $4N_e$, respectively. This Kingman coalescence process can be easily extended to incorporate other population parameters like population size, growth rate, recombination rate, and migration rates (Hudson 1990). Beerli (this volume) and Wakeley (1998) have shown that approaches based on coalescence theory are superior to approaches based on allele frequencies.

Two main groups of programs exist for coalescence analysis (Appendix 3), those that use segregating sites in the sample and those that integrate over all possible genealogies. In the sofware package SITES (Hey & Wakeley 1997), the coalescent is used to generate expectations for the number of segregating sites in a sample of sequences, and these expecetations are subsequently used to estimate population parameters. Most other programs listed in Appendix 3 integrate over all possible genealogies (e.g., MIGRATE). These programs are very computer intensive, but they use all possible information in the data, such as the history of mutation events. They also can be applied to several different types of molecular data other than sequence data. The general approach is to find the maximum likelihood of the population parameters, where the likelihood function is defined as the sum of probabilities over all possible genealogies. For each of these genealogies, one calculates the probability given the parameters and given the sampled data (Beerli, this volume).

4. Relationships among species: phylogeny reconstruction

The field of molecular systematics has become an increasingly important part of ecological studies during the past two decades. During that time, the number of computer programs for data preparation (e.g., entering and aligning DNA sequences), phylogenetic inference, tree comparisons, and other associated analyses has mushroomed to the point of being beyond the scope of any paper or website. For those readers considering phylogenetic analysis for the first time, we recommend reading Hillis (this volume) and the volume, *Molecular Systematics* (Hillis *et al*. 1996), within

analysis of phylogenetic and population genetic data. Because that publication is now nearly 3 years old, some of the information may be out of date, but nonetheless it represents a good starting point. Alternatively, we advise visiting the website of the J. Felsenstein lab at the University of Washington (http://evolution.genetics.washington.edu). At that website, one can find descriptions of approximately 120 phylogeny packages that are arranged by (i) method of phylogenetic inference; (ii) computer systems on which they work; (iii) most recent listings; and (iv) those most recently updated.

5. Acknowledgments

We thank all those people who sent us information about their programs and all those who maintain websites with information about their own and other people's software.

References

Aldrich PR, Hamrick JL (1998) Reproductive dominance of pasture trees in a fragmented tropical forest mosaic. *Science*, **281**, 103-105.

Ardern SL, Lambert DM (1997) Is the black robin in genetic peril? *Molecular Ecology*, **6**, 21-28.

Avise, JC (1994) *Molecular Markers, Natural History and Evolution*, Chapman & Hall, London UK.

Bahlo M, Griffiths RC (1998) Inference from gene trees in a subdivided population. *Theoretical Population Biology*.

Barton NH, Baird SJE (1998) Analyse 2.0. Edinburgh: http://helios.bto.ed.ac.uk/ evolgen/index.html.

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F, (1998) GENETIX, logiciel sous Windows™ pour la génétique des populations. Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier (France)

Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*, **5**, 393-401.

Cornuet JM, Luikart G (1997) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001-2014.

Cruzan MB (1998) Genetic markers in plant evolutionary ecology. *Ecology*, **79**, 400-412.

Danzmann RG (1997) PROBMAX: A computer program for assigning unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. *Journal of Heredity*, **88**, 333.

Devlin B, Ellstrand NC (1990) The development and application of a refined method for estimating gene flow from angiosperm paternity analysis. *Evolution*, **44**, 248-259.

Devlin B, Roeder K, Ellstrand NC (1988) Fractional paternity assignment, theoretical development and comparison to other methods. *Theoretical and Applied Genetics*, **76**, 369-380.

Estoup A, Solignac M, Cornuet JM (1994). Precise assessment of the number of patrilines and of genetic relatedness in honey bee colonies. *Proceedings of the Royal Society of London: Biological Sciences*, **258**, 1-7.

Fischer M, Matthies D (1998) RAPD variation in relation to population size and plant fitness in the rare *Gentianella germanica* (Gentianaceae). *American Journal of Botany*, **85**, 811-820.

Garnier-Gere P, Dillmann C (1992) A computer program for testing pairwise linkage disequilibrium in subdivided populations. *Journal of Heredity*, **83**, 239.

Gentile G, Sbordoni V (1998) Indirect methods to estimate gene flow in cave and surface populations of *Androniscus dentiger* (Isopoda: Oniscidea). *Evolution*, **52**, 432-442.

Godt MJW, Hamrick JL (1998) Allozyme diversity in the endangered pitcher plant *Sarracenia rubra* ssp. *alabamensis* (Sarraceniaceae) and its close relative *S. rubra* ssp. *rubra*. *American Journal of Botany*, **85**, 802-810.

Gonzalez S, Maldonado JE, Leonard JA, Vila C, Barbanti Duarte JM, Merino M, Brum-Zorrilla N, Wayne RK (1998) Conservation genetics of the endangered Pampas deer (*Ozotoceros bezoarticus*). *Molecular Ecology*, **7**, 47-56.

Goodman SJ (1997) Rst Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology*, **6**, 881-885.

Goudet J (1995) Fstat version 1.2: a computer program to calculate Fstatistics. *Journal of Heredity*, *86*, 485-486.

Griffiths RC, Tavaré S (1996) Computational methods for the coalescent. In: *Progress in Population Genetics and Human Evolution* (eds. Donnelly P, Tavaré S). IMA Volumes in Mathematics and its Applications. Springer Verlag, Berlin.

Hamrick JL, Godt MJW (1989) Allozyme diversity in plant species. In: *Plant PopulationGenetics, Breeding and Genetic Resources* (eds. Brown AHD, Clegg MT, Kahler AL, Weir BS), pp. 43-63. Sinauer, Sunderland, MA.

Hey J, Wakeley J (1997) A coalescent estimator of the population recombination rate. *Genetics*, **145**, 833-846.

Hillis DM, Moritz C, Mable BK (1996) *Molecular Systematics*. Sinauer, Sunderland, MA.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1-44.

Hughes C (1998) Integrating molecular techniques with field methods in studies of social behavior: a revolution results. *Ecology*, **79**, 383-399.

Kingman J (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235-248.

Kuhner MK, Yamato J, Felsenstein, J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* , **140**, 421-430.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429-439.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations**.** ***Molecular Ecology*****, 7,** 639-655**.**

Meagher, TR (1986) Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely male parents**.** ***The American Naturalist*****, 128,** 199-215**.**

Morin PA, Wallis J, Moore JJ, Woodruff DS (1994) Paternity exclusion in a community of wild chimpanzees using hypervariable simple sequence repeats. *Molecular Ecology*, **5**, 469-478.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Vyse E, Ward R, Strobeck C (1998) Variation in genetic diversity across the range of North American brown bears. *Conservation Biology*, **12**, 418-429.

Parker PG, Snow AA, Schug MD, Booton GC, Fuerst PA (1998) What molecules can tell us about populations: choosing and using a molecular marker. *Ecology*, **79**, 361-382.

Prodöhl P A, Loughry WJ, McDonough CM, Nelson WS, Thompson EA, Avise JC (1998) Genetic maternity and paternity in a local population of armadillo assessed by microsatellite DNA markers and field data. *The American Naturalist*, **151**, 7-19.

Queller CR, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258-259.

Queller CR, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in Evolution and Ecology*, **8**, 285-288.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research*, **67**, 147-158.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA*, **94**, 9197-9201.

Raymond M, Rousset F (1995a) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Raymond M, Rousset F (1995b) An exact test for population differentiation. *Evolution*, **49**, 1280-1283.

Ritland K (1990) A series of FORTRAN computer programs for estimating plant mating systems. *Journal of Heredity*, **81**, 235-237.

Roeder K, Devlin B, Lindsay BG (1989) Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, **45**, 363-379.

Rozas J, Rozas R (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Computer Applic. Biosci*, **11**, 621-625.

Rozas J, Rozas R (1997) DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. *Comput. Applic. Biosci.*, **13**, 307-311.

Schneider S, Kueffer JM, Roessli D, Excoffier L (1997) Arlequin ver. 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Sork VL, Campbell D, Dyer R, Fernandez J, Nason J, Petit R, Smouse P, Steinberg E (1998) Proceedings from a Workshop on Gene Flow in Fragmented, Managed, and Continuous Populations. National Center for Ecological Analysis and Synthesis, Santa Barbara, California. Research Paper No. 3. Available at http://www.nceas.ucsb.edu/papers/geneflow/

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: *Molecular Systematics* (eds. Hillis DM, Moritz C, Mable BK), pp.407-514. Sinauer, Sunderland, MA.

Taylor AC, Horsup A, Johnson CN, Sunnucks P, Sherwin B (1997) Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Molecular Ecology*, **6**, 9-19.

Thompson, EA (1975) The estimation of pairwise relationships**.** ***Annals of Human Genetics*****, 39,** 173-188**.**

Thompson, EA (1976) Inference of genealogical structure. ***Social Science Information*****, 15,** 477-526**.**

Tufto J,Engen S, Hindar K (1996) Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics*, **144**, 1911-1921.

Yeh FC, Yang RC, Boyle TJB, Ye ZH, Mao JX (1997) POPGENE, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta,

Yeh FC, Boyle TJB (1997) Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian Journal of Botany*, **129**, 157.

Wakeley J (1998) Segregating sites in Wright's island model. *Journal of Theoretical Population Biology*, **53**, 166-174.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847-855.

Weir, B (1996) *Genetic Data Analysis*. Sinauer, Sunderland, MA.

Winfield MO, Arnold GM, Cooper F, Le Ray M, White J, Karp A, Edwards KJ (1998) A study of genetic diversity in *Populus nigra* ssp. *betulifolia* in the Upper Severn Area of the UK using AFLP markers. *Molecular Ecology*, **7**, 3-11.

Xu J, Kerrigan RW, Sonnenberg AS, Callac P, Horgen PA, Anderson JB (1998) Mitochondrial DNA variation in natural populations of the mushroom *Agaricus bisporus*. *Molecular Ecology*, **7**, 19-34.

Appendix 1: List of software packages for the study of parentage and relatedness using molecular markers

Package name, lastest update (author)	Operating system	*Types of data handled*	Analyses supported	Availability, literature citation
CERVUS v. 1.0, 17/6/98 (T Marshall)	Windows95	Diploid, codominant markers	Uses a most-likely approach to parentage inference and estimates confidence in parentage of most likely parents. Can be used to calculate allele frequencies, run simulations to determine critical values of likelihood ratios and analyse parentage in populations of animals and plants. A simulation system can estimate the resolving power of a series of single-locus marker systems for parentage inference.	Freeware from http://helios.bto.ed.ac.uk/evolgen/index.html Marshall *et al*. (1998)
Kinship v. 1.2, (KF Goodnight, DC Queller, T Posnansky)	MacOS (PowerPC and 68K)	Diploid, codominant markers	Performs maximum likelihood tests of pedigree relationships between pairs of individuals in a population. The user enters two hypothetical pedigree relationships, a primary hypothesis and a null hypothesis, and the program calculates likelihood ratios comparing the two hypotheses for all possible pairs in the data set. Includes a simulation procedure to determine the statistical significance of results. Also calculates pairwise relatedness statistics.	Freeware from http://www.bioc.rice.edu/~kfg/GSoft.html Queller & Goodnight (1989)
MLT (K Ritland)	DOS	Diploid or tetraploid, codominant markers	A set of programs that finds maximum-likelihood estimates of outcrossing rates for plant populations. Also estimates parental gene frequencies and inbreeding coefficients. Special programs within the package can handle autotetraploids and ferns.	Freeware by contacting the author at ritland@unixg.ubc.ca Ritland (1990)

Appendix 1: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
PollenFlow v. 1.0, 26/3/98 (JD Nason)	MacOS (PowerPC and 68K)	Diploid, codominant markers	Implements two different models. First, a paternity exclusion-based model estimates total rate of pollen immigration from a single external source into a defined local population. Second, a likelihood-based model estimates relative male fertility within a population as well as pollen immigration from one or more external sources. Male fertility estimates are adjusted to eliminate biases due to cryptic gene flow.	Freeware by contacting the author at john-nason@uiowa.edu Sork *et al.* (1998)
PROBMAX, 17/11/97 (RG Danzmann)	DOS	Codominant and dominant/ recessive diploid markers	Ascertains the parentage of individuals when genotypic data on both parents and progeny are available. Also includes PROBMAXG, which generates possible progeny genotypes from the parental mixtures to test whether a given set of genetic markers will be able to discriminate all progeny back to parental sets, and PROBMAXN, which allows testing of possible parent/progeny assignments if null alleles segregating at some markers are suspected.	Freeware by anonymous ftp to 131.104.50.2 (password = danzmann) or contact the author at rdanzman@uoguelph.ca Danzmann (1997)
Relatedness v. 5.0.4, 29/6/1998 (KF Goodnight, DC Queller	MacOS (PowerPC and 68K	Diploid, codominant markers	Estimates genetic relatedness between demographically-defined groups of individuals using a regression measure of relatedness. Calculates symmetrical and asymmetrical relatedness and jackknife standard errors. Allows up to 32 demographic variables in defining those individuals to be used in calculating the relatedness	Freeware from http://www.bioc.rice.edu/~kfg/GSoft.html Queller & Goodnight (1989)

Appendix 2: List of software packages that will analyze geographically structured populations using traditional estimators based on gene frequencies.

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
Analyse v. 2.0, 5/98 (SJE Baird, NH Barton)	MacOS (PowerPC)	Diploid and haploid genetic markers, quantitative trait values, spatial coordinates (1 and 2 dimensions), environmental variables	Likelihood analysis of data from hybrid zones. Performs three types of analyses: general data handling (e.g., selecting subsets of the data satisfying particular criteria), analysis of random fluctuations in genotype frequency (e.g., estimating Fst, Fis, and standardized linkage disequilibrium), and analysis of a set of multilocus clines (e.g., estimating variation between clines).	Freeware from http://helios.bto.ed .ac.uk/evolgen/inde x.html Barton & Baird (1998)
Arlequin v. 1.1, 17/12/97 (S Schneider, JM Kueffer, D Roessli, L Excoffier)	Windows 3.1 or later	RFLPs, microsatellites, allozymes, RAPDs, AFLPs, allele frequencies, DNA sequences	Calculates gene and nucleotide diversity, mismatch distribution, haplotype frequencies, linkage disequilibrium, tests of Hardy-Weinberg equilibrium, neutrality tests, pairwise genetic distances, analyses of molecular variance (AMOVA).	Freeware from http://anthropolo gie.unige.ch/ arlequin Schneider *et al*. (1997)
DnaSP v. 2.52, 9/97 (J Rozas, R Rozas); v. 2.9 is available as a beta version	Windows 3.1 or later	DNA sequences	Estimates several measures of DNA sequence variation within and between populations (in noncoding, synonymous or nonsynonymous sites), and also linkage disequilibrium, recombination, gene flow, and gene conversion parameters. Also can conduct several tests of neutrality.	Freeware from http://www.bio .ub.es/~julio/Dn aSP.html Rozas & Rozas (1995, 1997)
Fstat v. 1.2, 12/95; Fstat for windows v. 2.3, is available as beta upon request (J Goudet)	DOS; new version will be Windows compatible	Allozymes, microsatellites, mtDNA RFLPs	Calculates gene diversity statistics of Weir and Cockerham (Weir, 1996). Computes jackknife and bootstrap confidence intervals of the statistics or can test gene diversity statistics using a permutation algorithm.	Freeware by writing to J. Goudet at jerome.goudet@ izea.unil.ch Goudet (1995)

Appendix 2: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
GDA, 11/7/97 (PO Lewis, D Zaykin)	Windows 3.1 or later	Allozymes, microsatellites	Calculates standard gene diversity measures, Wright's F-statistics using the method of Weir and Cockerham (Weir, 1996), genetic distance matrices, UPGMA and neighbor-joining dendrograms, exact tests for disequilibrium	Freeware; http://chee.unm.edu/gda Designed to accompany *Genetic Data Analysis* (Weir, 1996).
GENEPOP v. 3.1b, 12/97 (M. Raymond, F. Rousset)	DOS	Allozyme, microsatellites	Calculates exact tests for Hardy-Weinberg equilibrium, population differentiation, and genotypic disequilibrium among pairs of loci. Computes estimates of classical population parameters, such as allele frequencies, Fst, and other correlations. Includes Linkdos (Garnier-Gere and Dillmann, 1992), which is a program for testing pairwise linkage disequilibrium.	Freeware from 3 ftp sites: ftp://ftp.cefe.cnrs-mop.fr/genepop/ ftp://ftp2.cefe.cnrs-mop.fr/pub/pc/msdos/genepop/ ftp://isem.isem.univ-montp2.fr/pub/pc/genepop/ Raymond & Rousset (1995a, b)
GENETIX v. 3.3, 14/05/98 (K Belkhir, P Borsa, L Chikhi, J Goudet, F Bonhomme)	Windows95/NT	Allozymes, microsatellites	Calculates estimates of classical parameters (e.g., genetic distances, variability parameters, Wright's fixation indices, linkage disequilibrium) and tests their departure from null expectations through permutation techniques. The interface is not user-friendly for everyone, because it is currently only in French.	Freeware from http://www.univ-montp2.fr/~genetix/genetix.htm Belkhir *et al.* (1998)
Immanc, 17/10/97 (JL Mountain)	Windows 3.1 or later, MacOS (PowerPC), NeXT HP-RISC, Sun UltraSPARC	Allozymes, microsatellites, RFLPs	Tests whether or not an individual is an immigrant or is of recent immigrant ancestry. The program uses Monte Carlo simulations to determine the power and significance of the test.	Freeware from http://mw511.biol.berkeley.edu/software.html Rannala & Mountain (1997)

Appendix 2: Continued

Package name, lastest update (author)	**Operating system**			**Availability, literature citations**
			Analyses supported	
		Types of data handled		
Migrlib v. 1.0 (J Tufto)	Unix (available as a collection of S-Plus functions and some C code)	Allele frequencies	Estimates the pattern of migration in a subdivided population from genetic differences generated by local genetic drift. Functions are also provided for carrying out likelihood ratio tests between alternative models such as the island model and the stepping stone model.	Freeware from http://www.math.ntnu.no/~jarlet/migration Tufto *et al*. (1996)
PMLE12 v. 1.2, 4/3/96 (B Rannala)	Windows 3.1 or later, MacOS (PowerPC or 68K), NeXTStep	Allozymes, mtDNA RFLPs	Estimates the gene flow parameter theta for a collection of two or more semi-isolated populations by (pseudo) maximum likelihood. For discrete-generation island model, theta=2Nm. For a continuous-generation island model, theta is the ratio of the immigration rate phi to the individual birth rate lambda.	Freeware from http://mw511.biol.berkeley.edu/bruce/exec.html Rannala & Hartigan (1996)
POPGENE v. 1.21, 22/12/97 (F Yeh, RC Yang, T Boyle)	Windows 3.1 or later	Co-dominant or dominant markers using haploid or diploid data.	Calculates standard genetic diversity measures, tests of Hardy-Weinberg Equilibrium, Wright's F-statistics, genetic distances, UPGMA dendrogram, neutrality tests, linkage disequilibrium	Freeware from http://www.ualberta.ca/~fyeh/index.htm Yeh & Boyle (1997); Yeh *et al*. (1997)
RSTCALC v. 2.2, 6/10/97 (SJ Goodman)	DOS, Windows 3.1 or later	Microsatellites	Performs analyses of population structure, genetic differentiation, and gene flow. Calculates estimates of Rst, tests for significance and calculates 95% CI.	Freeware from http://helios.bto.ed.ac.uk/evolgen Goodman (1997)
TFPGA (Tools for Population Genetic Analyses), 12/5/98 (MP Miller)	Windows 3.1or later	Codominant (allozyme) and dominant (RAPD, AFLP) genotypes	Calculates descriptive statistics, genetic distances, and F-statistics. Performs tests for Hardy-Weinberg equilibrium, exact tests for genetic differentiation, Mantel tests, and UPGMA cluster analyses.	Freeware from http://herb.bio.nau.edu/~miller No citation available

Appendix 3: List of software packages that will analyze geographically structured populations using estimators based on coalescence.

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
Bottleneck v. 1.1.03, 27/11/97 (JM Cornuet, G Luikart, S Piry)	Windows95	Allele frequencies	Detects recent reductions in effective population size from allele frequency data. Tests whether a set of loci shows a significant excess of heterozygosity (i.e., the observed heterozygosity is larger than the heterozygosity expected at mutation-drift equilibrium and assuming a given mutation model).	Freeware from http://www.ensam.inra.fr/~piry Cornuet & Luikart (1997)
Fluctuate v. 1.50B, 6/2/98 (M Kuhner, J Yamato)	Windows 95/NT, MacOS (PowerMac); UNIX; available also as C source code	DNA sequences	Estimates the effective population size and an exponential growth rate of a single population using maximum likelihood and Metropolis-Hastings importance sampling of coalescent genealogies.	Freeware from http://evolution.genetics.washington.edu/lamarc.html Kuhner *et al.* (1995, 1998)
Genetree, 9/6/98 (M Bahlo, RC Griffiths)	Windows 95/NT, Dec Alpha; available also as C source code	DNA sequences	Finds maximum likelihood estimates of population sizes, exponential growth rates, migration matrices, and time to the most recent common ancestor.	Freeware from http://www.maths.monash.edu.au/~mbahlo/mpg/gtree.html Griffiths & Tavaré (1996) Bahlo & Griffiths (1998)
Migrate-0.4 v. 0.4.3, 25/5/98 (P Beerli)	Windows 95/98/NT, MacOS (PowerMac), Dec Alpha, LINUX/Intel, NeXTStep; available also as C source code	Allozymes, microsatellites, DNA sequences	Menu driven, character-based program that finds 4+1 maximum-likelihood estimates of population parameters for a two-population model: effective population sizes for subpopulation1 and subpopulation 2, migration rates between the two subpopulations, and for multilocus data, a shape parameter for the distribution of the mutation rate.	Freeware from http://evolution.genetics.washington.edu/lamarc.html Beerli, this volume

Appendix 3: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
Migrate-n v. Alpha-3, 25/5/98 (P Beerli)	Windows 95/98/NT, MacOS (PowerMac), Dec Alpha, LINUX/Intel, NeXTStep; available also as C source code	Allozymes, microsatellites, DNA sequences	Menu driven, character-based program that finds n*n maximum-likelihood estimates of population parameters for n-population model: effective population sizes for each subpopulation, migration rates between the n subpopulations, and for multilocus data, a shape parameter for the distribution of the mutation rate.	Freeware from http://evolution.genetics.washington.edu/lamarc.html Beerli, this volume
Recombine v. 1.0, 17/6/98 (MK Kuhner, J Yamato, J Felsenstein)	MacOS (PowerMac), Windows95/NT; available as C source code that will compile on DEC ULTRIX, DEC alpha, INTEL machines, NeXT, SGI, but needs gcc to compile on Suns	DNA or RNA sequences, single nucleotide polymorphisms	Fits a model which has a single population of constant size with a single recombination rate across all sites. It estimates 4Nu and r, where N is the effective population size, u is the neutral mutation rate per site, and r is the ratio of the per-site recombination rate to the per-site mutation rate.	Freeware from http://evolution.genetics.washington.edu/lamarc.html No citation available
SITES v. 1.1, 21/4/98 (J Hey)	DOS, MacOS; also available as ANSI C source code	DNA sequences	Generates tables of polymorphic sites, indels, codon usage. Computes numbers of synonymous and replacement base positions, pairwise sequence differences, and GC content. Performs group comparisons and polymorphism analyses and estimates historical population parameters. Primarily intended for data sets with multiple closely related sequences.	Freeware from http://heylab.rutgers.edu/index.html#software Hey & Wakeley (1997) Wakeley & Hey (1997)

Analysis of geographically structured populations: (Traditional) estimators based on gene frequencies

Peter Beerli
Department of Genetics, Box 357360,
University of Washington, Seattle WA 98195-7360,
Email: beerli@genetics.washington.edu

This is an introduction and overview of the currently used methods for the analysis of population subdivision and estimation of migration rates. We will discuss theoretical population models such as the group of single migration parameter models with two or n islands, stepping stone models, and multi-parameter models such as the migration matrix model. In this lecture I will concentrate on approaches using gene frequencies, and will neglect complicating evolutionary forces such as selection and age structured populations. Sewall Wright introduced 1922 the fixation index F and the term F statistic. This summary statistic is based on the avariability in and between subpopulations. For different data types (e.g. enzyme electrophoretic markers, microsatellite markers, sequence data) different coefficients are in use (e.g. F_{ST}, R_{ST}). These different methods take into account that the variability generating process, mutation, is different for different types of data. Most of these F_{ST} based estimators were developed for symmetrical population models. I will discuss an extension which is able to cope with asymmetrical population models, compare these different methods, and analyze their performance. Confidence limits of F_{ST} of population parameters can be found using the boostrap over loci, or a maximum likelihood ratio test if we are working in a maximum likelihood framework. Most of these methods will be superseded by either maximum likelihood concepts in the context of gene frequency data, or methods taking the genealogy of the sample into account [second lecture].

Introduction and context

In the early twenties Sewall Wright introduced the notation of the fixation index F to characterize the influence of mating systems on heterozygosity in inbred guinea pig lines. Such an inbred line looks like a "natural" population (Fig. 1) with very few individuals; genes are passed in a random fashion to offspring, who replace their parents. WRIGHT (1973) wrote: "It became evident that the same set of parameters, the F-statistics, which measure relative change of heterozygosis in an array of diverging inbred lines also measures the differentiation of their gene frequencies" and we can apply it to geographically structured populations. F-statistic itself gives us a summary statistic about isolation of subpopulations and their variability, but if we want to understand more clearly the underlying processes we want to know the population parameters such as population size and migration rate and perhaps be able to determine routes of gene flow between populations. A general overview on the problems of estimating effects of migration on gene frequencies can be found in FELSENSTEIN (1982).

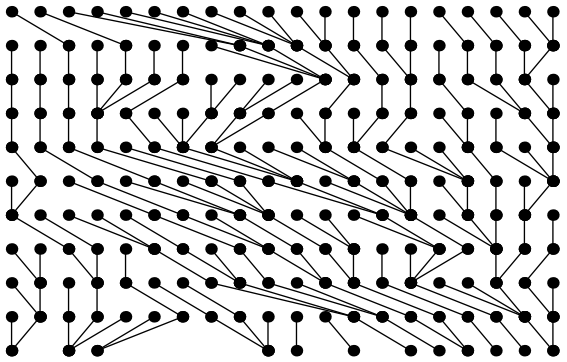

Figure 1: Wright-Fisher population model: idealized population with random mating. The genes are rearranged so that we can see the genealogy. Each line of dots is a generation, the number of individuals is 10 with 20 genes

Models of geographically structured populations

Most of the migration models have several very restrictive assumptions and assume a specific way of replacing individuals from one generation to the other (Fig. 2).

The n island model (Figure 4: A,B) **(WRIGHT, 1931):** All subpopulations have the same effective population size, $N_e^{(i)}$. Individuals migrate from one subpopulation to the other with the same rate m. The distances between subpopulations are not taken into account.

Stepping stone model (Figure 4: C) **(MALECOT, 1950; KIMURA, 1953):** All subpopulations have the same effective population size, $N_e^{(i)}$. The migration rate m is constant and defines the rate of exchange from one neighboring population to the other along the possible paths.

Continuum model (WRIGHT, 1940): in which a populaiton is spread out in geographical continuum. Unfortunately, these models have mathematical properties so that they are not able to define stable subpopulations at one location through time, although they come very close to our intuition about real populations.

Migration matrix model (Figure 4: B,D)**(BODMER and CAVALLI-SFORZA, 1968):** All subpopulations have the same effective population size, $N_e^{(i)}$. The migration rates between subpopulations

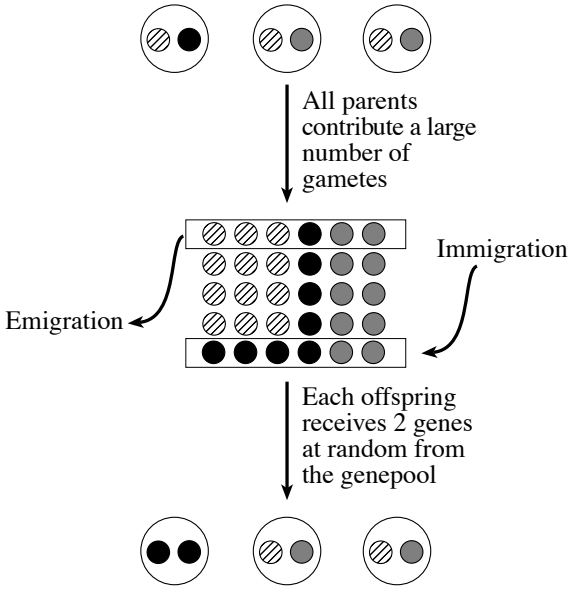

Figure 2: Sequence of events in a migration model

can be different and for four populations (Figure 3) one could have for example the following migration matrix (I chose the migration rates to reflect an isolation by distance model).

$$\begin{pmatrix} - & m & \frac{m}{2} & \frac{m}{4} \\ m & - & m & \frac{m}{2} \\ \frac{m}{2} & m & - & m \\ \frac{m}{4} & \frac{m}{2} & m & - \end{pmatrix}$$

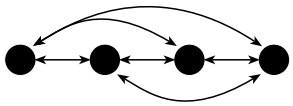

Figure 3: Four populations

In an arbitrary migration model some of the migration path can be disallowed (set to 0.0). A further extension of these models includes variable subpopulation size.

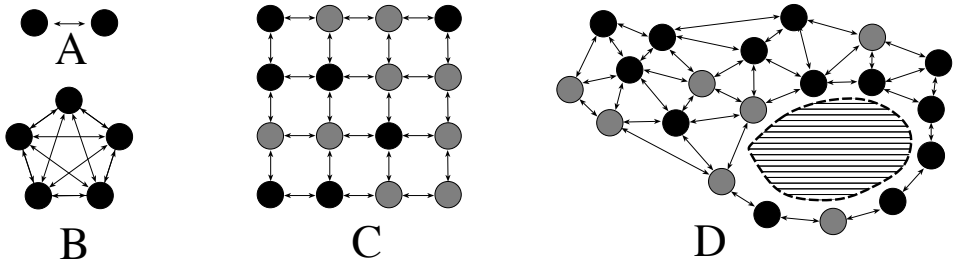

Figure 4: Migration models: A, B: *n*-island model, C: Stepping stone model (2-dimensional), D: arbitrary migration matrix model. Black disks are sampled subpopulations, gray disks are unsampled subpopulations

Transformation of variability into summary statistics

To develop a summary statistic we can use the variability in and between populations, but we need to consider the underlying model of evolution.

F_{ST}[1], G_{ST}, **Infinite allele model:** WEIR (1996), SLATKIN (1991)

R_{ST}, **Microsatellites:** SLATKIN (1993)

F_{ST}, **Sequences:** HUDSON *et al*. (1992b), NEI (1982), and LYNCH and CREASE (1990)

Assessments of confidence limits

Bootstrapping over loci is appropriate to generate confidence limits.

Estimates of migration rate

Wright's formula

$$F_{ST} = \frac{1}{1+4Nm}$$

to transform F_{ST} values into migration rates is still most commonly used. It assumes that the mutation rate is 0.0 and the number of subpopulations is very large. Also, we will not gain any information about the population sizes themselves, they are convoluted with the migration rates. Additionally, a mutation rate of 0.0 is perhaps appropriate for enzyme electrophoretic data, but it is not appropriate for microsatellites or intron-sequences. We can incorporate these relaxations of the assumptions. In a two population model (Fig. 5) we can solve the following equation system using the homozygosity within a population F_W and the homozygosity between populations F_B (NEI and FELDMAN, 1972) by replacing $4N\mu$ with Θ and m/μ with $\mathcal{M}$

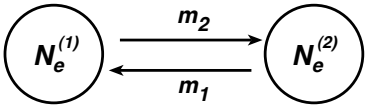

Figure 5: Two population model with population sizes $N_e^{(1)}$, $N_e^{(2)}$, and migration rates m_1, m_2.

$$\begin{aligned} F_W^{(1)} &= \frac{1}{2N_1} + \left(1 - 2\mu - 2m_1 - \frac{1}{2N_1}\right) F_W^{(1)} + 2m_1 F_B \\ F_W^{(2)} &= \frac{1}{2N_2} + \left(1 - 2\mu - 2m_2 - \frac{1}{2N_2}\right) F_W^{(2)} + 2m_2 F_B \\ F_B &= F_B\,(1 - \mu - m_1 - m_2) + m_1 F_W^{(1)} + m_2 F_W^{(2)} \end{aligned} \tag{1}$$

With one locus we can only solve for 3 parameters, either a constant $\Theta = 4N\mu$ ($4 \times$ effective population size $N_e \times$ mutation rate μ; because we do not know the mutation rate we include it into the estimate) and two migration rates $\mathcal{M}_1 = m_1/\mu$ and $\mathcal{M}_2 = m_2/\mu$ or for two different Θ_1 and Θ_2 values and one symmetric migration rate $\mathcal{M}$.

[1] WEIR (1996) called this θ, but we will use Θ for $4N_e\mu$ in approaches using coalescence theory

Problems with F-statistic approaches:

- Wright's formula is often inappropriate for real world situations.
- Rather complicated estimation procedure, when we consider more than two populations and want to estimate population sizes and migration rates.
- If for some subpopulations the F_W are smaller than the F_B the estimation procedure breaks down.
- Gene frequencies are considered to be the true gene frequencies of the sampled populations. This can produce wrong results with small sample sizes.
- Parameter estimates based on F_{ST} do not make full usage of the data [see second lecture].

Maximum likelihood estimators

- Estimation using PMLE of RANNALA and HARTIGAN (1996)
- Estimation using the approach of TUFTO *et al*. (1996)

Other approaches

- Distance measures (NEI and FELDMAN, 1972)
- Parsimony related (EXCOFFIER and SMOUSE, 1994)
- Rare allele approach (SLATKIN, 1985)

Summary

- We recognize several different migration models: n-island model, stepping stone model, and migration-matrix model. Their assumptions strongly influence the estimates of population parameters. Complications in computations of estimates can arise by relaxing assumptions such as equal population size or symmetric migrations.
- Quality of transformation of the variability in the data into summary statistics is dependent how well the underlying model for the estimator fits the data.
- Current F-statistic approaches assume symmetry of migrations and often equal population sizes.
- Allowing for unequal population sizes and unequal migration rates complicates migration rate estimation considerably. Also, in a F-statistics framework it is not possible to estimate all four parameters of a two population model with one locus (e.g. mtDNA).

- Maximum likelihood approaches, e.g. work by RANNALA and HARTIGAN (1996) and TUFTO *et al*. (1996), utilizing the distribution of gene frequencies promise to give good results, but some of this work is still in the beginning stages.
- For sequence data the current estimators based on F-statistics are less accurate than coalescence theory based estimators, because they do not not use information about the history of mutations.

Bibliography

BARTON, N. and SLATKIN, M., 1986 A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity (Edinburgh) **56 (Pt 3):** 409–15.

BODMER, W. F. and CAVALLI-SFORZA, L. L., 1968 A migration matrix model for the study of random genetic drift. Genetics **59:** 565–592.

EXCOFFIER, L. and SMOUSE, P., 1994 Using allele frequencies and geographic subdivision to reconstruct gene trees within species: Molecular variance parsimony. Genetics **136:** 343–359.

FELSENSTEIN, J., 1982 How can we infer geography and history from gene frequencies? Journal of Theoretical Biology **96:** 9–20.

HUDSON, R., BOOS, D., and KAPLAN, N., 1992a A statistical test for detecting geographic subdivision. Molecular Biology and Evolution **9:** 138–151.

HUDSON, R., SLATKIN, M., and MADDISON, W., 1992b Estimation of levels of gene flow from dna sequence data. Genetics **132:** 583–9.

KIMURA, M., 1953 "stepping-stone" model of population. Annual Report of the National Institute of Genetics, Japan **3:** 62–63.

LYNCH, M. and CREASE, T., 1990 The analysis of population survey data on DNA sequence variation. Molecular Biology and Evolution **7:** 377–394.

MALECOT, G., 1950 Some probability schemes for the variability of natural populations (french). Annales de l'Universite de Lyon, Sciences, Section A **13:** 37–60.

NEI, M., 1982 Evolution of human races at the gene level. In *Human Genetics, Part A: The Unfolding Genome*, edited by B. Bohhe-Tamir, P. Cohen, and R. Goodman, pp. 167–181, Alan R. Liss, New York.

NEI, M. and FELDMAN, M. W., 1972 Identity of genes by descent within and between populations under mutation and migration pressures. Theoretical Population Biology **3:** 460–465.

RANNALA, B. and HARTIGAN, J., 1996 Estimating gene flow in island populations. Genetical Research **67:** 147–158.

RANNALA, B. and MOUNTAIN, J., 1997 Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci **94:** 9197–9201.

ROUSSET, F. and RAYMOND, M., 1997 Statistical analyses of population genetic data: new tools, old concepts. Trends in Ecology and Evolution **12:** 313–317.

SLATKIN, M., 1985 Rare alleles as indicators of gene flow. Evolution **39:** 53–65.

SLATKIN, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–92.

SLATKIN, M., 1991 Inbreeding coefficients and coalescence times. Genetical Research **58:** 167–75.

SLATKIN, M., 1993 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

SLATKIN, M. and BARTON, N., 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.

SLATKIN, M. and MADDISON, W., 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

SLATKIN, M. and VOELM, L., 1991 Fst in a hierarchical island model. Genetics **127:** 627–629.

TUFTO, J., ENGEN, S., and HINDAR, K., 1996 Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics **144:** 1911–1921.

WEIR, BRUCE, S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.

WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. American Naturalist **74:** 232–248.

WRIGHT, S., 1973 The origin of the f-statistics for describing the genetic aspects of population structure. pp. 3-26 in Genetic Structure of Populations, ed. N. E. Morton. University Press of Hawaii, Honolulu .

Software, with emphasis on methods using gene frequencies

[this list is certainly not complete]

- ANALYSE An "easy-to-use" MacOS application for the analysis of hybrid zone data. Calculates several statistics: e.g. FST, and isolation by distance.
  Website through `http://helios.bto.ed.ac.uk/evolgen`

- ARLEQUIN is an exploratory population genetics software environment able to handle large samples of molecular data (RFLPs, DNA sequences, microsatellites), while retaining the capacity of analyzing conventional genetic data (standard multi-locus data or mere allele frequency data). A variety of population genetics methods have been implemented either at the intra-population or at the inter-population level.
  Website at `http://anthropologie.unige.ch/arlequin`

- DNASP computes (among lots of other things) different measures of the extent of DNA divergence between populations, and from these measures it computes the average level of gene flow, assuming the island model of population structure. DnaSP estimates the following measures: dST, gST and Nm, NST and Nm, FST and Nm (Rozas, J. and R. Rozas. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Applic. Biosci. 13: 307-311). Binary for Windows 3.1 and 95.
  Website at `http://www.bio.ub.es/~julio/DnaSP.html`

- GDA (Genetic Data Analysis) is a Microsoft Windows program for analyzing discrete genetic data based on WEIR (1996).
  Website at `http://chee.unm.edu/gda`

- GENEPOP is a population genetics software package for DOS and can be fetched by anonymous ftp from `ftp.cefe.cnrs-mop.fr` in the directory /PUB/PC/MSDOS/GENEPOP or can be used through a web interface at
  `http://www.curtin.edu.au/curtin/dept/biomed/teach/genepop/web_docs/gene_form.html`

- IMMANC is a program designed to test whether or not an individual is an immigrant or is of recent immigrant ancestry. The method is appropriate for use with allozyme, microsatellite, or restriction fragment length data. Loci are assumed to be in linkage equilibrium. The power of the test depends on the number of loci, the number of individuals sampled, and the extent of genetic differentiation between populations RANNALA and MOUNTAIN (1997). Binaries for Macintosh, Windows, and NEXTSTEP.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- MICROSAT estimates several indices using microsatellite data. C source code and binaries for DOS and Macintosh.
  Website at `http://lotka.stanford.edu/microsat.html`

- PMLE12 estimates the gene flow parameter theta for a collection of two or more semi-isolated populations by (pseudo) maximum likelihood using either allozyme or mtDNA RFLP data RANNALA and HARTIGAN (1996). C source code and binaries for Macintosh, Windows, and NEXTSTEP.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- POPGENE computes both comprehensive genetic statistics (e.g., allele frequency, gene diversity, genetic distance, G-statistics, F-statistics) and complex genetic statistics (e.g., gene flow, neutrality tests, linkage disequilibria, multi-locus structure). Binaries for Windows3.1, Windows95.
  Website at `http://www.ualberta.ca/ fyeh/index.htm`.

- RELATEDNESS 4.2 calculates average genetic relatedness among groups of individuals specified by up to three user-defined demographic variables. It also calculates F-statistics measuring inbreeding and genetic differences among sub-populations. Binary for Macintosh.
  Website at `http://www-bioc.rice.edu/~kfg/GSoft.html`
- RSTCALC is a program for performing analyses of population structure, genetic differentiation and gene flow using microsatellite data. Binary for Windows.
  Website through `http://helios.bto.ed.ac.uk/evolgen`

Analysis of geographically structured populations: Estimators based on coalescence

Peter Beerli
Department of Genetics, Box 357360,
University of Washington, Seattle WA 98195-7360,
Email: beerli@genetics.washington.edu

The rapid increase in the collection of population samples of molecular sequences, plus the great expansion of the use of microsatellite markers, makes it possible to investigate the patterns and rates of migration among geographically subdivided populations with much greater power than was previously possible. The difficulty with methods for analyzing these data has been that they do not allow the researcher to observe the genealogical tree of ancestry of the sampled sequences, but only make an estimate of it which has a great deal of uncertainty. Taking the uncertainty in our estimate of the genealogy into account is the major challenge for a proper statistical analysis of these data. The statistical approach of maximum likelihood is used to infer these rates and patterns, using the Markov Chain Monte Carlo (MCMC) method of computing the likelihoods. This method samples genealogies from the space of possible genealogies, using an acceptance-rejection method to concentrate the sampling in the regions which contribute most to the outcome. Even though the number of possible genealogies is vast, the MCMC sampling can avoid wasting computer time on possibilities that can have made little contribution to the observed outcome. This sampling of different genealogies in computing a likelihood for the parameters correctly accounts for our lack of knowledge of the true gene tree.

It can be shown that these ML-methods are superior to methods based on F_{ST}. Additionally, ML-methods can take into account variability in mutation rate and can estimate all relevant population parameters jointly and also analyze cases with different population sizes and migration rates. Comparison of different data types reveals that number of loci sampled is a key factor in reducing the variability of the parameter estimates.

The coalescent

Most current population genetics analyses are using theoretical findings of Sewall Wright and R. A. Fisher which were made in the early 20th century. Their work is based on a view which uses discrete generations of idealized individuals passing their genes to offspring in the next generation. This "looking forward" strategy implies that calculation of the probability of a given genotype is rather difficult. Kingman (1982a,b) formalized a "looking backward" strategy: the coalescent. Hudson (1990) and Donnelly and Tavaré (1997) give comprehensive reviews on the subject. Coalescence theory takes the relatedness of the sample into account, so it incorporates random genetic drift and mutation. This approach makes it very easy to calculate probabilities of a genealogy of a sample of individuals with a given effective population size, $P(g|\Theta)$. Hudson (1990) and others showed that we can extend this single population approach to multiple populations and estimate migration rates and also that we can include other forces such as growth, recombination, and selection.

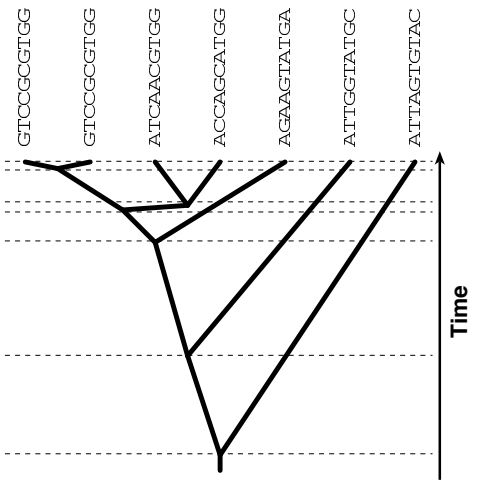

Figure 1: A coalescent tree with sampled sequences

Markov chain Monte Carlo (MCMC) integration

Construction of random genealogies (Simulation studies) is simple with the coalescent approach (e.g. the method of Slatkin and Maddison 1989). Inference of parameters is much harder, especially when we want not to lose any information in the data (Felsenstein 1992). In a likelihood framework we would like to simply integrate over all possible genealogies G and solve for the population parameters Θ at the maximum likelihood

$$L(\Theta) = \int_{g \in G} P(g|\Theta) P(D|g) dg, \tag{1}$$

where $P(D|g)$ is the likelihood of the genealogy with the sample data. This is not possible; there are too many different topologies with different branch lengths. But we can approximate by using a biased random walk through the genealogy space and then infer the parameters from the sampled genealogies correcting for the biased sampling:

$$L(\Theta) = \int_{g \sim P(g|\Theta_0)P(D|g)} \frac{P(g|\Theta)}{P(g|\Theta_0)} dg \tag{2}$$

(MCMC: Hammersley and Handscomb 1964, MCMC and coalescence: Kuhner et al. 1996)

Table 1: Simulation with unequal known parameters of 100 two-locus datasets with 25 individuals in each population and 500 base pairs (bp) per locus. Std. dev. is the standard deviation.

	Population 1		Population 2	
	$4N_e\mu$	$4N_em$	$4N_e\mu$	$4N_em$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

Two population exchange migrants

We will explore the details of the MCMC mechanism in a simple two population model with the parameters: $\Theta_1 = 4N_e^{(1)}\mu$, $\Theta_2 = 4N_e^{(2)}\mu$, $\mathcal{M}_1 = m_1/\mu$, $\mathcal{M}_2 = m_2/\mu$ (we need to scale by the unknown mutation rate μ of our data).

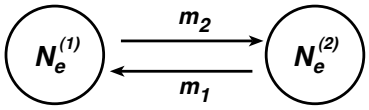

Figure 2: Two population model with population sizes $N_e^{(1)}, N_e^{(2)}$, and migration rates m_1, m_2.

- Assumptions: Population have constant size and exist forever, migration rate is constant through time, and the genetic markers are neutral.
- We can jointly estimate migration rates and population sizes
- Example of a simulation study (Table 1), where I generated 100 single locus data sets and then analyzed them with the program MIGRATE (Beerli 1997).
- Problems: perhaps not a natural situation; how long do we need to run the genealogy sampler?

Migration matrix model

- Assumptions: same as with 2 populations
- Simulation studies with (a) 4 sampled populations and (b) with 3 sampled population and one population where we don't have data.

Figure 3: Population structure used in simulations.

- Problems: how many genealogies to sample? Number of parameters increases quadratically.

Comparison with F_{ST}

Simulation studies can show that the ML-estimator delivers better result than F_{ST}, and results are still accurate when population sizes and/or migration rates are unequal (Table 1).

Hypothesis testing using likelihood ratios

The maximum likelihood framework makes it easy to test hypotheses. I expect that these tests will supersede standard test based on F_{ST}. I will show a few examples and hope that I am able to have a version of MIGRATE finished in March so that everybody can experiment with their own data in the "data section".

$$\mathrm{H}_0 : \hat{N}_e = N_e^{(x)}$$

$$\text{Test-statistic:} \qquad -2\log\left(\frac{\mathrm{L}(\Theta_x)}{\mathrm{L}(\hat{\Theta})}\right) \leq \chi^2_{df,\alpha}$$

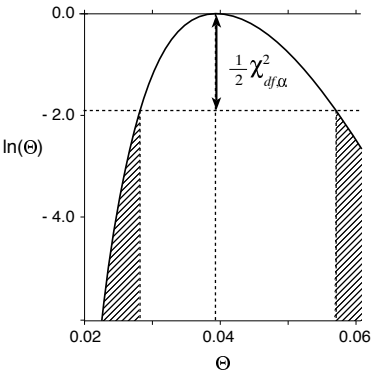

Figure 4: Likelihood ratio test: dashed areas are outside of the 95% confidence limit. Θ is $4N_e\mu$; $df = 1$, $\alpha = 0.05$

Data type and mutation rate

We have mutation models for infinite allele model, microsatellite stepwise mutation model (Valdez and Slatkin 1993, Di Rienzo et al. 1994), and finite sites sequence model (e.g. Swofford et al. 1996).

What's the effect of the data type to the estimate of migration rates? The data type is not that important, for the quality of the migration rate estimates, but the variance of the estimates is dependent on the number of unlinked loci (Fig. 5) having independent coalescent trees and the variability in the data, the more segregating sites or polymorphic loci are present the better the estimates of the migration rates.

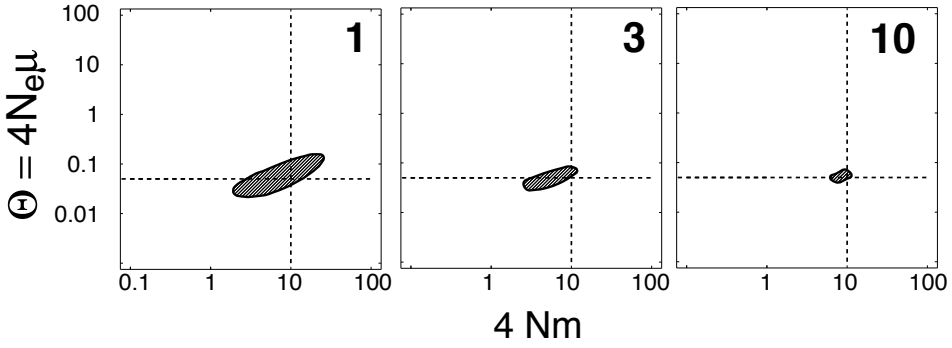

Figure 5: Variance of parameter estimates: the dashed area is the 95% confidence area, the numbers 1, 3, and 10 are the numbers of sampled loci

Mutation rate is not constant: incorporation of the variance of the mutation rate is possible by assuming that it follows a Gamma distribution (Fig. 6) and estimating the shape parameter α of this distribution jointly with the population parameters by integrating over all mutation rates x

$$L(\Theta, \mathcal{M}, \alpha) = \prod_l \int_0^\infty \frac{e^{-\alpha x/\Theta_l} x^{\alpha-1}}{\Gamma(\alpha)\left(\frac{\Theta_l}{\alpha}\right)^\alpha} L(x, \Theta_l, \mathcal{M}_l) dx,$$

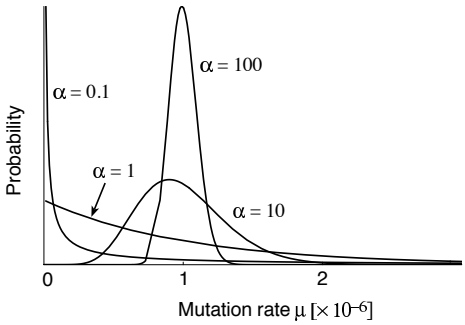

Figure 6: Gamma distributed mutation rates, with different shape parameter α and the same mean

Summary

- Coalescence theory enables us to estimate population parameters by including sample data and taking the possible histories of the populations into account.
- Expansion of the coalescence model to any migration model is possible.
- Maximum likelihood ratio test of arbitrary hypotheses.
- Multi-locus enzyme electrophoretic data and microsatellite markers delivers good migration rate estimates compared to mtDNA sequence data, because the quality of the result is dependent on the number of loci and the variability in the data.
- The assumption that the mutation rate over loci is constant is obviously wrong for electrophoretic markers and microsatellites and taking the variation of the mutation rate into account should improve the estimates of population parameters.

Bibliography

Citations with a ⋆ are recommended to read and/or introductory, citations with a ● are rather difficult.

BEERLI, P., 1997, MIGRATE DOCUMENTATION version 0.3. Distributed over the Internet: http://evolution.genetics.washington.edu/lamarc.html.

DI RIENZO A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN, and N. B . FREIMER, 1994, Mutational processes of simple-sequence repeat loci in human populations. Genetics **91** (8): 3166–3170.

⋆ DONNELLY, P. and S. TAVARÉ, 1997, *Progress in population genetics and human evolution*. IMA volumes in mathematics and its applications **87**, Springer, New York.

FELSENSTEIN, J., 1973, Maximum likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics **25**: 471–492.

FELSENSTEIN, J., 1988, Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics **22**: 521–565.

FELSENSTEIN, J., 1992, Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetics Research **59**: 139–147.

• GRIFFITHS, R. and S. TAVARÉ, 1994, Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci **344** (1310): 403–10, Department of Mathematics, Monash University, Clayton, Victoria, Australia.

HAMMERSLEY, J. and D. HANDSCOMB, 1964, *Monte Carlo Methods*. Methuen and Co., London.

⋆ HUDSON, R. R., 1990, Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44.

• KINGMAN, J., 1982a, The coalescent. Stochastic Processes and their Applications **13**: 235–248.

• KINGMAN, J., 1982b, On the genealogy of large populations. In *Essays in Statistical Science*, edited by J. Gani and E. Hannan, pp. 27–43, Applied Probability Trust, London.

⋆ KUHNER, M., J. YAMATO, and J. FELSENSTEIN, 1995, Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140** (4): 1421–30, Department of Genetics, University of Washington, Seattle 98195-7360, USA.

METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, and E. TELLER, 1953, Equation of state calculations by fast computing machines. Journal of Chemical Physics **21**: 1087–1092.

NATH, H. and R. GRIFFITHS, 1993, The coalescent in two colonies with symmetric migration. Journal of Mathematical Biology **31** (8): 841–51.

NATH, H. and R. GRIFFITHS, 1996, Estimation in an Island Model Using Simulation. Theoretical Population Biology **50**: 227–253.

• NOTOHARA, M., 1990, The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology **29** (1): 59–75.

SLATKIN, M., 1991, Inbreeding coefficients and coalescence times. Genetical Research **58** (2): 167–75, Department of Integrative Biology, University of California, Berkeley 94720.

⋆ SLATKIN, M. and W. MADDISON, 1989, A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123** (3): 603–613, Department of Zoology, University of California, Berkeley 94720.

⋆ SWOFFORD, D., G. OLSEN, P. WADDELL, and D. HILLIS, 1996, Phylogenetic Inference. In *Molecular Systematics*, edited by D. Hillis, C. Moritz, and B. Mable, pp. 407–514, Sinauer Associates, Sunderland, Massachusetts.

• TAKAHATA, N., 1988, The coalescent in two partially isolated diffusion populations. Genetical Research **52** (3): 213–22.

• TAKAHATA, N. and M. SLATKIN, 1990, Genealogy of neutral genes in two partially isolated populations. Theoretical Population Biology **38** (3): 331–50, National Institute of Genetics, Mishima, Japan.

VALDEZ A. M. and M. SLATKIN, 1993, Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133** (3): 737–749, Department of Zoology, University of California, Berkeley 94720.

WAKELEY, J. and J. HEY, 1997, Estimating ancestral population parameters. Genetics **145** (3): 847–855.

Software, with emphasis on using the coalescent

[this list is certainly not complete]

- LAMARC package [**L**ikelihood **A**nalysis with **M**etropolis **A**lgorithm using **R**andom **C**oalenscence. Three programs are currently available: COALESCE, FLUCTUATE, and MIGRATE. C-source code and binaries for Windows, Mac, LINUX, DUNIX, NEXTSTEP.
  Website at `evolution.genetics.washington.edu/lamarc.html`

- MISAT estimates the effective population size of a single population using microsatellite data and can also test if the one-step model or a multi-step model is appropriate. Binaries for Macintosh and Windows.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- SITES is a computer program for the analysis of comparative DNA sequence data (Hey and Wakeley, 1997. A coalescent estimator of the population recombination rate. Genetics 145: 833-846) . C source code and binaries for DOS and Macintosh.
  Website at `http://heylab.rutgers.edu`

- UPBLUE is a least square estimator for population size (Fu, Y. X., 1994. An phylogenetic estimator of effective population size or mutation rate. Genetics 136:685-692). Fortran program or use the website directly to calculate results
  `http://www.hgc.sph.uth.tmc.edu/fu/`

- Calculation of 4Nm using the method of SLATKIN and MADDISON (1989), you need to calculate the minimal number of migration events on the genealogy either by hand or using MacClade (Maddison and Maddison 1992, Sinauer). Pascal source code.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- Several programs for the estimation of population size, exponential growth, recombination rate, migration rate, time of the last common ancestor. Contact Bob Griffiths (email: ...) for more information.