

---

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

---

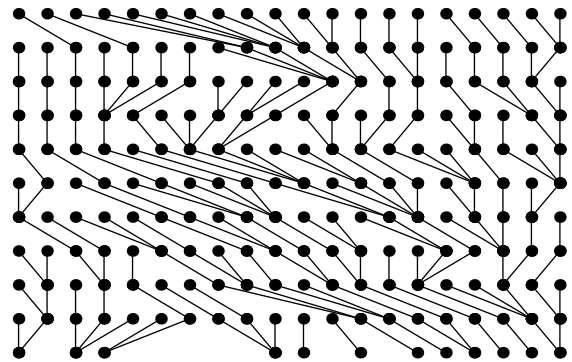
Peter Beerli  
Department of Genetics, Box 357360,  
University of Washington, Seattle WA 98195-7360,  
Email: beerli@genetics.washington.edu

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 variability 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 bootstrap 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 Most of these methods will be superseded by either maximum likelihood concepts in the context of gene frequency data, or methods taking the genealogy of the sample into account [second lecture].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 heterozygosity 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 population 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

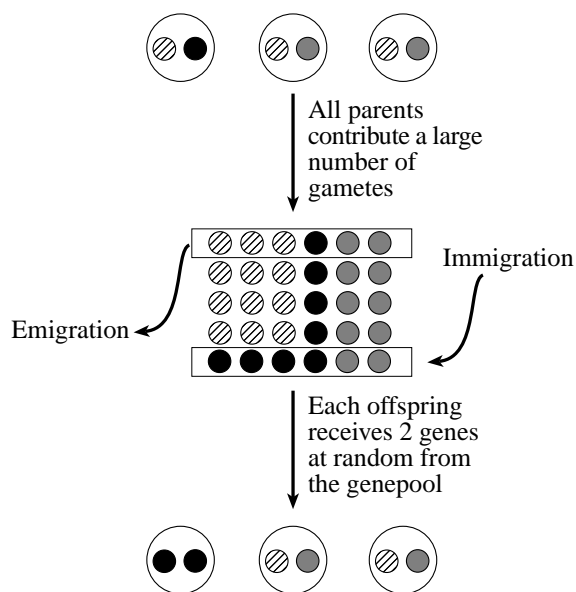


Figure 2: Sequence of events in a migration model

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).

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

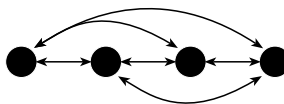


Figure 3: Four populations

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.

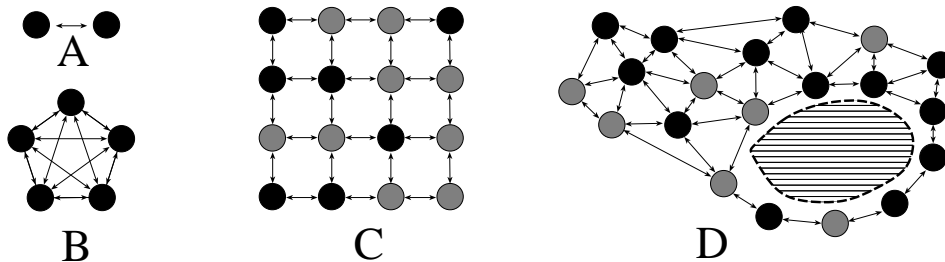


Figure 4: Migration models: A, B:  $n$ -island model, C: Stepping stone model (2-dimensional), D: arbitrary migration matrix model. Black disks are sampled subpopulations, gray disks are unsampled subpopulations

## 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.

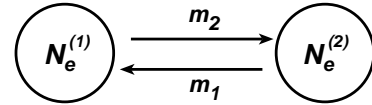
## Estimates of migration rate

---

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  $\Theta$  and  $m/\mu$  with  $\mathcal{M}$



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

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

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  $\mu$ ; 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  $\Theta_1$  and  $\Theta_2$  values and one symmetric migration rate  $\mathcal{M}$ .

<sup>1</sup>WEIR (1996) called this  $\theta$ , but we will use  $\Theta$  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 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  $F_{st}$  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.  $F_{ST}$ , 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](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>