SESSION 2

GENE FLOW – MEASUREMENT AND IMPLICATIONS

Chairman & Session Organiser: Professor A Gray
Institute of Terrestrial Ecology, Furzebrook

Defining and measuring gene flow

A F Raybould, R T Clarke
ITE Furzebrook Research Station, Wareham, Dorset, BH20 5AS, UK

ABSTRACT

We describe developments of regression-based methods for estimating the relationship between gene flow and distance using allelic variation at marker loci. These new methods allow the correct testing of the difference between regression slopes, and the assignment of confidence intervals to estimates from the regression equation, when the regression is between matrices of pairwise data, rather than between independent values. We show that there are significant differences in gene flow estimates from RFLPs and isozymes in sea beet. Estimates of the distance at which populations exchange one migrant per generation are lower for wild cabbage than for sea beet.

INTRODUCTION

A full assessment of the environmental impact of GM crops will consider the movement of transgenes into and among natural populations of crop relatives. Regardless of variation in sexual compatibility and selection, variation in gene flow among crop relative populations might cause some transgenes to spread and become common very quickly, whereas others might remain rare and be confined to areas very close to the source crop. Because transgenes are inherited in a mendelian way, gene flow among non-transgenic natural populations is an 'appropriate biological model' (Gliddon, 1994) for predicting the movement of transgenes within and among populations of crop relatives.

Gene flow is the movement of genetic information among individuals, populations or taxa. In plants, "potential" gene flow is the movement of seed and pollen as a function of distance (i.e. dispersal). "Actual" gene flow is the amount of fertilisation (in the case of pollen) and establishment of reproductive individuals (in the case of seeds) as a function of distance from a source (Levin & Kerster, 1974). Clearly, not all pollen will effect fertilisation, and not all seeds will establish reproductive plants, therefore actual gene flow can be much lower than potential gene flow. Also, the shapes of the pollen and seed dispersal curves may not predict the rate of change of gene flow with distance because the probability of fertilisation/establishment by a pollen grain/seed may vary with distance (e.g. Levin, 1981).

Gene flow is measured in two ways: by 'direct and' 'indirect' methods. The most common direct method for plants is the observation of seed and pollen movement, which gives an estimate of potential gene flow (dispersal). Other direct methods use genetic markers to estimate actual gene flow. A simple method is to introduce or identify a plant in a population with a unique genetic marker (e.g. an isozyme allele) and to follow the appearance of the marker in the next generation (e.g. Latta *et al.*, 1998). A more sophisticated approach uses markers to identify the fathers of half-sib families. If the markers are highly variable (e.g. microsatellites) and the number of potential fathers is relatively small, the father of each seed can be identified unambiguously (e.g. Dow & Ashley, 1998).

Indirect methods use the distribution of genetic variation to infer actual amounts of gene flow. The most powerful type of data is allele frequencies at one or more discrete loci. These data can be treated in a number of ways, but essentially high variation in allele frequency between populations or patches of plants indicates that gene flow between the populations is low, whereas similar allele frequencies across populations imply high gene flow.

Direct methods, whether estimating potential or actual gene flow, only measure gene flow at the time of the observations. Indirect genetic methods, on the other hand, measure average amounts of actual gene flow, by reflecting the cumulative effects of temporal variation in the spatial distribution of dispersal and establishment over preceding years, including rare, unpredictable events (e.g. Slatkin, 1985). If rare long-distance dispersal events (i.e. founder effects) have shaped the genetic structure of a species, direct methods may give lower estimates of gene flow than indirect methods (e.g. Campbell & Dooley, 1992). On the other hand, direct estimates can be higher than indirect estimates, for example where genetic drift has removed immigrant alleles from populations (e.g. Rasmussen & Brødsgaard, 1992). If possible, it is desirable to use both types of method.

In this paper, we describe indirect methods for estimating gene flow in two crop relatives: *Beta vulgaris* ssp. *maritima* (sea beet, the wild relative of sugar beet) and *Brassica oleracea* ssp. *oleracea* (wild cabbage, a wild relative of cultivated cabbage and oilseed rape). Our aims are to derive robust estimates of gene flow for these species, and to draw attention to the assumptions that underlie these estimates.

ESTIMATION OF GENE FLOW FROM ALLELE FREQUENCY DATA

The starting point for indirect estimates is data on the distribution of a genetic polymorphism. Neutral, co-dominant allelic variation at discrete loci gives the most powerful type of information. Dominant markers can be used, but very large sample sizes are needed to achieve the same statistical power (Lynch & Milligan, 1994). Since the mid-1960s, isozymes have provided plant population geneticists with a ready source of co-dominant markers. Microsatellites are now becoming the markers of choice because the number of polymorphic loci and the number of alleles per locus tend to be higher than isozymes. However, the mutation mechanisms at microsatellite loci are poorly understood, which presents problems in deciding the most appropriate estimator of genetic structure (and hence gene flow) at these loci.

The most common method of inferring gene flow from allele frequency data uses the 'Infinite Island' model of Wright (1931), in which an infinite number of finite populations ('islands') produce and receive migrants. The importance of an infinite number of islands is that allele frequencies in the system as a whole do not change. In each generation, islands have an effective population size of N, of which a proportion, m, are migrants. Wright showed that the product Nm is related to a parameter F_{ST} (see below), such that

$$F_{ST} \approx 1/(4\mathrm{Nm} + 1)$$

Nm is the amount of gene flow - the number of migrants per population per generation averaged over all islands.

The parameter F_{ST} has been variously formulated. Perhaps the most widely adopted is the analysis of variance approach of Weir & Cockerham (1984). Consider a single locus with two alleles A and a with overall frequencies p and (1-p) respectively. Arbitrarily assign values of X=1 for A and X=0 for a. If X_{ijk} denotes the value (0 or 1) for the *i*th allele in the *j*th individual in the *k*th population, then the total variance of X can be partitioned as follows:

$$X_{ijk} = p + a_k + b_{jk} + w_{ijk}$$

where a_k represents the difference (from overall p) in the frequency of A in population k (with variance of the set of a_k equal to σ_a^2), the b_{jk} denote differences between individuals within populations (variance σ_b^2), and the w_{ijk} represent differences between alleles within individuals (variance σ_w^2). The total variance of X = p(1-p) $= \sigma_T^2 = \sigma_a^2 + \sigma_b^2 + \sigma_w^2$. F_{ST} is the proportion of the variance that is due to differences in the mean allele frequency among populations. In other words

$$F_{ST} = \sigma_a^2 / \sigma_T^2$$

There are different ways of treating loci with more than two alleles and for combining data from several loci. Weir & Cockerham's (1984) suggested approach, which is widely used, is to estimate the overall F_{ST} as

$$F_{ST} = \sum_r \sum_s \sigma_{a(rs)}^2 / \sum_r \sum_s \sigma_{T(rs)}^2$$

which is effectively an average of the individual F_{ST} values ($F_{ST(rs)} = \sigma_{a(rs)}^2 / \sigma_{T(rs)}^2$) for each allele s of each locus r, weighted by respective total variances ($\sigma_{T(rs)}^2$).

It is relatively straightforward to estimate F_{ST} for a group of populations, and so derive the average amount of gene flow among them. However, one may be more interested in whether gene flow declines with increasing geographic distance between populations, a phenomenon termed "isolation by distance" (Wright, 1943). Using computer simulations, Slatkin (1993) showed that under a variety of demographic models, isolation by distance was represented by a linear relationship between log Nm and log distance (D) estimated between all pairs of populations. Normal tests of significance of a correlation or regression of log Nm and log D are not valid because the number of points (pairs) is higher than the number of independent pieces of information (populations). If there is no isolation by distance present in the study (null hypothesis), then each of the n populations of plants could have come from any of the n geographic positions. Therefore the significance is obtained from a Mantel randomisation test (Mantel, 1967), which effectively repeatedly randomises the positions of the n populations. The test significance is the proportion of correlations between Log Nm and Log D from say 10000 randomised data sets that are equal to or less than the observed correlation (Nm and distance are negatively correlated when there is isolation by distance).

A regression relationship between log Nm and log D allows us to derive an indirect estimate of the actual gene flow between a pair of populations a given distance apart. It also enables us to estimate the distance (which we term 'isolation distance') at which gene flow is reduced to some critical level. However, in both cases, without confidence intervals such estimates are

of little value. The pairwise nature of the data means that valid confidence intervals cannot be derived from the formula for standard errors of parameter estimates and predictions available for normal linear regression of independent observations. We are developing a more appropriate regression model that incorporates terms for the variability between populations in their levels of gene flow for a given distance. Maximum likelihood estimates of model parameters and their standard errors are used to derive confidence limits for the regression line and for the 'isolation distance'. We call this approach the 'Maximum likelihood population-effects' (MLPE) method.

GENE FLOW AMONG SEA BEET POPULATIONS

We estimated gene flow among ten sea beet populations in Dorset (see Raybould *et al.*, 1996b; 1997). Fifty plants per population were analysed for variation at 7 isozyme and 6 restriction fragment length polymorphism (RFLP) marker loci. The data were used to estimate all pairwise F_{ST} values separately for the isozyme and RFLP data. F_{ST} estimates were converted to log Nm values and regressed on log D. The results are displayed in Figure 1 and Table 1.

Table 1. Details of regressions (log Nm = a + b log distance) among 10 populations of sea beet in Dorset. *P* = Mantel one-sided test probability (b ‹ 0).

Marker Type	b	*P*	R^2
RFLP	-0.479	0.0007	33.2%
Isozyme	-0.068	0.4910	1.0%

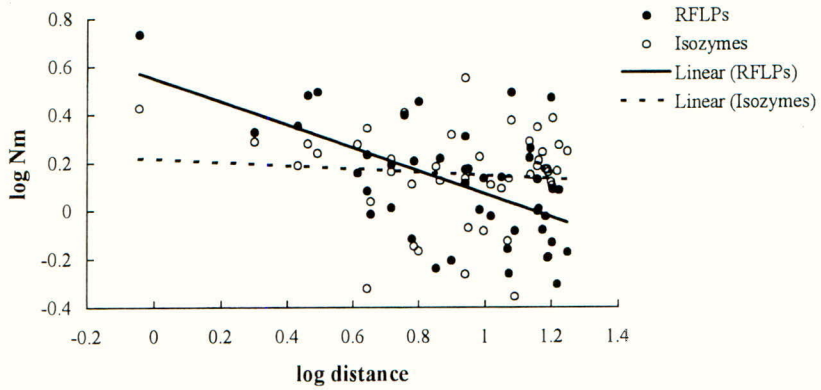

Figure 1. Regressions between log Nm and log distance among 10 populations of sea beet in Dorset. Nm estimated separately from isozymes and RFLPs.

The RFLP data suggest a strong isolation by distance effect (significant negative slope and high R^2), whereas the isozyme data show no relationship between gene flow and distance (non-significant negative slope and low R^2). Because of the non-independence of the points in

the regression analyses, a formal test of the significance of the difference between the RFLP and isozyme slopes cannot be made using a conventional t-test. However, if there is no significant difference between the slopes, there will be no correlation (and regression relationship) between the difference in log Nm values for the two types of marker (i.e. log $Nm_{(RFLPs)}$ – log $Nm_{(isozymes)}$) and log distance. The regression slope is significantly different from zero (Mantel two-sided test $P = 0.0064$), showing that there is a statistically significant difference between the isozyme and RFLP regression slopes (Figure 2).

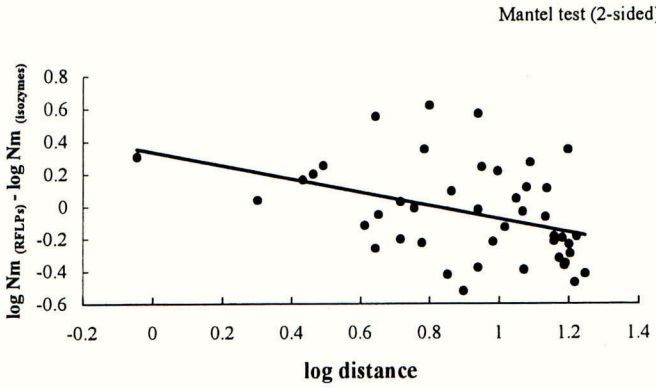

Figure 2. Regression between (log $Nm_{(RFLPs)}$ – log $Nm_{(isozymes)}$) and log distance among 10 populations of sea beet in Dorset.

The reason for the difference is uncertain at present. However, it is possible that the difference is due to the effects of a small subset of loci. If each RFLP locus is excluded in turn from the analysis, the correlation between log Nm and log distance for the 5 remaining loci becomes slightly less negative in each case, apart from *R1* which makes the correlation slightly more negative (Table 2). In all cases the correlation is significant. The results suggest a consistent pattern of isolation by distance at each RFLP locus.

In general, when an isozyme locus is excluded, the remaining 6 loci give a slightly more positive correlation compared with the full data set (Table 2). However, when either locus *got-3* or *pgi* is excluded, the correlation becomes negative, though non-significant. When both loci are excluded, the correlation becomes negative and significant ($P = 0.0077$) and the regression slope for log Nm based on the remaining five isozyme markers is not significantly different from that using the six RFLPs (Mantel two-sided test $P = 0.1085$).

The analysis is preliminary and needs to be confirmed by other tests. For example we need to obtain correlations for single loci. Single locus correlations are a problem at present because often a number of pairwise F_{ST} values are undefined because the two populations are fixed for the same allele and our current procedures for Mantel tests do not handle missing data. Nevertheless, we can speculate that the differences between the RFLP and isozyme data are mainly due to the effects of two loci, *got-3* and *pgi*. The case of *pgi* is particularly interesting

because a large number of studies have suggested that variation in PGI enzymes is adaptive (Riddoch, 1993).

Table 2. Changes in the correlation between log Nm and log distance when single loci are excluded from the analysis.

Isozymes			RFLPs		
Locus excluded	r	P (1-sided)	Locus excluded	r	P (1-sided)
None	0.003	0.4913	None	-0.631	0.0007
Acph	0.085	0.7410	*L3*	-0.614	0.0006
Est	0.030	0.5674	*L9*	-0.564	0.0013
Got-3	-0.080	0.2398	*R1*	-0.658	0.0004
Got-4	0.030	0.5758	*R4*	-0.611	0.0004
Per	0.009	0.5049	*R7*	-0.598	0.0009
6-pgdh	0.013	0.5324	*R13*	-0.567	0.0007
Pgi	-0.168	0.0927			
Got-3 & pgi	-0.318	0.0077			

For risk assessment purposes, it is useful to know the precision of our gene flow estimates. In conservation genetics, one migrant per generation (i.e. Nm = 1) is a useful rule of thumb for the minimum amount of gene flow necessary to prevent populations becoming fixed for different alleles through the effects of genetic drift (e.g. Mills & Allendorf, 1996). Therefore the distance at which Nm = 1 might be a useful way to compare the pattern of gene flow among species, and serve as a very rough measure of 'isolation distance' among natural populations. Again, because of the pairwise nature of the data, confidence intervals for estimates of the 'isolation distance' cannot be obtained by ordinary least-squares (OLS) regression techniques. We use our 'maximum likelihood population-effects' (MLPE) method to overcome this problem. For the beet RFLP data, an ordinary regression assuming independent points, gives Nm = 1 at 13.9 km with under-estimated 95% confidence intervals of 10.3-24.4 km. The MLPE method gives Nm = 1 at 14.5 km with wider, but in this case more accurate 95% confidence intervals of 6.7-39.3 km.

We have also carried out a study of seven *Brassica oleracea* (wild cabbage) populations using microsatellites (Raybould *et al.*, 1999), for which the estimated distance at which Nm = 1 was 6.1 km with best-estimate 95% confidence limits of 1.5-10.4 km. The lower 'isolation distance' was perhaps expected as wild cabbage pollen is insect dispersed, whereas beet pollen is dispersed by wind. In addition, beet seed may be dispersed by tides from some populations.

ASSUMPTIONS INVOLVED IN THE F_{ST}-BASED APPROACH

The F_{ST} based approach to estimating gene flow outlined above is very simple, but has many potential problems with both the interpretation of indirect estimates and with the population genetic models used to make these estimates (Bossart & Prowell, 1998). An essential assumption of the approach is that the observed distribution of genetic variation is the result of equilibrium between gene flow and genetic drift. This assumption can be violated for many reasons, such as recent colonisation, selection acting on markers (e.g. through

environmental effects) and high mutation rates. To some extent these problems can be overcome by using a wide variety of markers. A perhaps more serious problem is that the infinite island model is probably a gross oversimplification in most natural populations. Some models consider a finite number of islands, whereas others assume migration between neighbouring populations only (stepping stone models). A 'general' model would consider a matrix of migration rates between all populations (considered together, rather than as separate 2-population systems), however analysis based on such a model would be too complex to provide useful estimates of gene flow (e.g. see Nagel, 1997).

It is now possible to analyse data on population gene frequencies using migration models that are more realistic than the island model. For example, Tufto *et al.* (1998) present an analysis of data on 21 sub-populations of sea beet spread over two transects along the shore of Furzey Island in Poole Harbour, on the South Coast of England (Raybould *et al.*, 1996a). In the models of Tufto *et al.* (1996, 1998), explicit distance distribution functions are specified for pollen dispersal (assumed to be two-dimensional) and seed dispersal (assumed to be one-dimensional along the shore) based on perceived underlying physical movement processes. These functions are used with the assumed, known or estimated effective population sizes to derive a matrix of equations for the migration transition probabilities between each pair of populations. This matrix can be used to repeatedly calculate the probability distribution (i.e. multivariate normal variance-covariance matrix) for the observed population gene frequencies using a range of parameter values for the dispersal functions until the likelihood is maximised. This maximum likelihood approach can also be used to distinguish between alternative models and dispersal functions. Tufto *et al.* (1998) found that the Furzey Island sea beet data was best fitted by a model assuming the sub-populations formed an isolated meta-population with the standard deviation for both pollen and seed dispersal distances equal to 75m. This type of method therefore provides estimates of potential gene flow.

CONCLUSIONS

New markers and statistical approaches to the indirect estimation of gene flow are being developed continuously. The traditional approach based on F_{ST} estimates and the infinite island model may eventually be replaced by methods that are more realistic and as easy to use. However, we agree with Bohonak *et al.* (1998) who 'believe that the limitations of traditional approaches are generally understood and that they still provide a valuable first approximation in many cases'. Comparisons of gene flow among species have previously been based on differences in mean F_{ST} or Nm values regardless of the spatial scale of the study (e.g. Hamrick & Godt, 1996). Comparisons based on the above regression and modelling approaches are more informative, especially for GMO risk assessment, in that they incorporate the effects of scale and inter-population distances.

REFERENCES

Bohonak AJ; Davies N; Roderick GK; Villablanca FX (1998). Is population genetics mired in the past? *Trends in Ecology and Evolution* **13,** 360.

Bossart JL; Prowell DP (1998). Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology and Evolution* **13,** 202-206.

Campbell DR; Dooley JL (1992). The spatial scale of genetic differentiation in a hummingbird-pollinated plant – comparison of models of isolation by distance. *American Naturalist* **139,** 735-748.

Dow BD;Ashley MV (1998). High levels of gene flow in bur oak revealed by paternity analysis using microsatellites. *Journal of Heredity* **89,** 62-70.

Gliddon C (1994). The impact of hybrids between genetically modified crop plants and their related species: biological models and theoretical perspectives. *Molecular Ecology* **3,** 41-44.

Hamrick JL; Godt MJW (1996). Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society B* **351,** 1291-1298.

Latta RG; Linhart YB; Fleck D; Elliot M (1998). Direct and indirect estimates of seed versus pollen movement within a population of ponderosa pine. *Evolution* **52,** 61-67.

Levin DA (1981). Dispersal versus gene flow in plants. *Annals of the Missouri Botanical Garden* **68,** 233-253.

Levin DA; Kerster HW (1974). Gene flow in seed plants. *Evolutionary Biology* **7,** 139-220.

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

Mills LS; Allendorf FW (1996). The one-migrant-per-generation rule in conservation management. *Conservation Biology* **10,** 1509-1518.

Nagel JE (1997). A comparison of alternative strategies for estimating gene flow from genetic markers. *Annual Review of Ecology and Systematics* **28,** 105-128.

Rasmussen IR; Brødsgaard B (1992). Gene flow inferred from seed dispersal and pollinator behavior compared to DNA analysis of restriction site variation in a patchy population of *Lotus corniculatus* L. *Oecologia* **89,** 277-283.

Raybould AF; Goudet J; Mogg RJ; Gliddon CJ; Gray AJ (1996a). Genetic structure of a linear population of sea beet revealed by isozyme and RFLP analysis. *Heredity* **76,** 111-117.

Raybould AF; Mogg RJ; Clarke RT (1996b). The genetic structure of *Beta vulgaris* ssp. *maritima* (sea beet) populations: RFLPs and isozymes show different patterns of gene flow. *Heredity* **77,** 245-250.

Raybould AF; Mogg RJ; Gliddon CJ (1997). The genetic structure of *Beta vulgaris* ssp. *maritima* (sea beet) populations. II. Differences in gene flow estimated from RFLP and isozyme loci are habitat-specific. *Heredity* **78,** 532-538.

Raybould AF; Mogg RJ; Clarke RT; Gliddon CJ; Gray AJ (1999). Variation and population structure at microsatellite and isozyme loci in wild cabbage (*Brassica oleracea* L.) in Dorset (UK). *Genetic Resources and Crop Evolution* (in press).

Riddoch BJ (1993). The adaptive significance of electrophoretic mobility in phosphoglucose isomerase. *Biological Journal of the Linnean Society* **50,** 1-17.

Slatkin M (1985). Gene flow in natural populations. *Annual Review of Ecology and Systematics* **16,** 393-430.

Slatkin M (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47,** 264-279.

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

Tufto J; Raybould AF; Hindar K; Engen S (1998). Analysis of genetic structure and dispersal patterns in a population of sea beet. *Genetics* **149,** 1975-1985.

Weir BS; Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358-1370.

Wright S (1931). Evolution in Mendelian populations. *Genetics* **16,** 97-159.

Gene flow and risk assessment

C J Gliddon
School of Biological Sciences, University of Wales, Bangor, Gwynedd, LL57 2UW, UK

ABSTRACT

The primary aim in assessing the amount of gene flow from higher plants, in the context of risk assessment of the release of GMO's, is to obtain an estimate of the distribution of the dispersal stages (e.g. pollen and seeds). A knowledge of this distribution will then permit estimation of the probability of the dispersal stage travelling further than a specified distance from its source. Such cumulative dispersal probabilities can then be used to design both any appropriate confinement measures for a release and monitoring procedures to ensure the efficacy of the confinement.

A standard statistical method for analysing gene-flow data is described and applied to data obtained from crop plants or their relatives, previously published in the scientific literature. The concept of relative risk is developed and applied to this data. Attention is drawn to the inherent variability of both gene flow and its estimates in these experiments and to the consequent problems in attempting to quantify risk to the environment.

INTRODUCTION

In Europe, we are approaching a possible transition between the confined use in the environment of genetically modified organisms (GMO's), primarily in small-scale field trials for the purpose of research and development, and a scaling-up of releases (farm-scale trials), prior to moving on to their unconfined commercial release. While there is still no harmonisation of regulations relating to the release of GMO's on a global scale, there is a large degree of conformity in the aims of the regulations, namely prevention or minimisation of environmental damage. To this end, the assessment of risk associated with the release of GMO's into the environment is required during the research and development stage with the intention that salient information obtained from monitoring those small and medium scale releases will be used to improve the assessment of risk associated with commercial (unconfined) releases.

The spread of genes due to spatial dispersal of pollen has been of great interest to agronomists, population geneticists and environmentalists. Contamination of seed crops by an undesirable source of pollen - such as a different cultivar, wild or weedy relative - led to many field studies in order to evaluate isolation distances required to maintain a satisfactory level of genetic purity of seeds. Results of such trials have often been used in order to establish, modify or contest isolation requirements for Certified Seed production (e.g. Stringam, 1978). Agronomists have also been looking for the best strategies to minimise isolation distances: barren zones, removal of border lines or pollinators.

More recent concerns relating to pollen flow have arisen from the development of GM crops. The nature of these concerns often varies, depending on the stage of development of the GM variety. At early stages of development there is often a perceived need to prevent, or at least minimise, the escape of 'novel' genes from a field trial into wild or feral relatives (Ellstrand & Hoffman, 1990; Dale, 1992; Raybould & Gray, 1993). This need represents, in essence, the converse of the problems faced in traditional seed multiplication, in that contamination of other plants is to be prevented rather than contamination of the trial by them. At later stages of development, it must be realised that escape of genes from the GM crop in medium to large scale trial is inevitable when sexually compatible relatives exist in the same environment. Nevertheless, it would be useful to be able to quantify likely rates of spread into the environment to enable the design of appropriate monitoring protocols and/or mitigation procedures should the risk assessment necessitate them. A further need for quantification of levels of gene flow arises in late stages of development where a single crop species has been modified for a variety of end-uses. For example, if oil seed rape varieties are being developed for use as human food, in industrial processes and for production of pharmaceuticals, it is clearly desirable to have a knowledge of the magnitude of gene flow which may occur, to permit the appropriate segregation in agriculture of the different forms of the crop.

RISK ASSESSMENT

Mackenzie & Henry (1990) described the present paradigm in evaluating risk as follows:

$$risk = exposure \times hazard$$

In the application of this conceptual relationship to GMO's released into the environment, *exposure* is a measure of the organism's (insert's) ability to escape from its place of deliberate introduction. The subsequent fate of any escaped organisms (inserts) then needs to be quantified in terms of the likelihood of persistence, increase and spread in the environment. The third term in the equation, *hazard*, refers to the impact of the escaped organism on the existing ecosystem.

In order for information to be useful in addressing the problem of quantification of *exposure* for use in risk assessment, it is necessary that it produce data allowing the estimation of the probability of escape as a function of distance from the intentional introduction, together with some estimate of the effect of size of the introduction and some measure of the likelihood of persistence.

The quantification of *hazard* or impact of the escaped organism on the existing ecosystem has rarely been addressed by monitoring of field trials, not least because field trials have been designed primarily to preclude the possibility of escape and, hence, of any *hazard* occurring. Furthermore, the quantification of *hazard* should involve both biological and socio-economic criteria since, for example, the local extinction of a species certainly has an impact on the local ecosystem but the question of whether this extinction constitutes damage rather than benign change requires a value judgement for its answer.

EXISTING ANALYSES

Virtually all of the sampling methods and monitoring protocols described in the literature fail to describe the minimum levels of detection which could be achieved using their particular protocol. This problem is exacerbated by the designs of the experiments - in the vast majority of cases using higher plants, the marked organisms are in a small minority of total organisms in the design. This results in the experimental design making it difficult to detect the spread of the marker in relation to the probability of recovering the non-marked gene. For example, if a marker is represented by 1% of the total organisms, even if its spread is uniform across the entire experimental area, it will only be recovered in 1% of samples. This fault of experimental design could well account for the very small distances that have been reported for the spread of GMO's (e.g. Scheffler *et al.*, 1993).

The vast majority of reported monitoring results fail to try to fit a distribution to the data. The results are usually presented as a simple histogram of raw data with either no attempt at further analysis or a simplistic analysis consisting of calculation of a mean dispersal distance (occasionally with errors which have been erroneously calculated assuming an underlying normal distribution). Furthermore, results are usually presented for marker genes at a given distance as a percentage of total genes sampled. This is inappropriate as it is scale dependent, the correct form being of marker genes at a given distance as a percentage of the total number of marker genes recovered. This method removes the dependence on size of the source of marker genes and correctly emphasises the rate of decrease of marker genes recovered with distance (see Kareiva *et al.*, 1994).

While undoubtedly there has been useful information collected from monitoring of the release of GMO's, the absence of appropriate analysis of the data makes it virtually useless, in its present form, for the purpose of risk assessment.

ALTERNATIVE APPROACHES

Gene flow may be studied by both indirect and direct methods. Indirect methods involve the use of techniques developed in population genetics theory to estimate rates of gene-flow in natural populations (Goudet *et al.*, 1994; Raybould *et al.*, 1996; Raybould *et al.*, 1997). Such indirect methods are problematic in that they need natural populations. However, in cases where the perceived risk involves spread of escaped genes through populations of crop relatives, these methods are ideal since they combine the effects on rate of spread of both gene-flow and population structure (i.e. connectivity). Direct methods involve the estimation of the parameters of dispersal distributions from actual field experiments. Traditionally, dispersal was assumed to follow a bivariate normal distribution (Wright, 1943; Haldane, 1948). However, dispersal distributions from plants have, in the main, been found to be strongly leptokurtic (e.g. Levin & Kerster, 1974) and this led to the suggested use of an exponential power function of the form e^{-ax^b} (Bateman, 1947; Kareiva *et al.*, 1994). Rather than use this essentially descriptive distribution, Lavigne *et al.* (1996) and Tufto *et al.* (1997) have proposed using methods based on a consideration of Brownian Motion in three dimensions to describe pollen deposition. It is clear that under some conditions, for example wind strength varying in direction during an experiment, this mechanistic method gives a superior fit compared to the descriptive exponential power function (Tufto *et al.*, 1997).

Nevertheless, this paper will apply the exponential power function to a range of published data on dispersal in crop plants.

The methodology used follows and extends that suggested by Kareiva *et al.* (1994), using a dispersal reliability function, *R(x)*, associated with an exponential power function for the density of exposure:

$$R(x) = e^{-ax^b} \qquad (1),$$

where a and b are constants which are estimated from the data, defining the rate of decay (shape) of the distribution. *R(x)* then gives the relative density of genes as a function of distance, x, from the source. Kareiva *et al.* (1994) suggest using a maximum likelihood approach in which the data can be realised as a binomial sampling process in which the binomial probabilities of obtaining a seed carrying a marker gene are proportional to *R(x)*. In order to estimate a and b using this method, a prior estimate of the contamination, c, at distance 0 from the source is required. In situations where the published data was not sufficient to allow the use of this maximum likelihood approach, the function *c.R(x)* was fitted using the method of least squares. The parameter c defines the level of contamination within the source of pollen (zero distance), b affects the convexity of the curve relative to simple exponential decay (the lower b, the greater the convexity) and a affects the rate of exponential decay. It should be noted that a, b and c are highly correlated, exacerbating problems of estimation.

For use in assessing risk, an expression describing the probability of a gene travelling further than a specified distance, x, is required. This is a cumulative density function (c.d.f.) with shape (rate of decay) described by the parameters a and b estimated above.

The appropriate c.d.f., derived from equation (1), is as follows:

$$\Gamma(\alpha,\beta) / \Gamma(\beta)$$

where α is $a.x^b$,
β is $2/b$ for dispersal in two dimensions (all directions),
$\Gamma(\bullet,\bullet)$ is the incomplete Gamma function and $\Gamma(\bullet)$ is the complete Gamma function

The c.d.f. above describes the decrease in exposure as a function of distance from the source, relative to the exposure with no isolation. Since, in any given release, the hazard is independent of isolation distance, the c.d.f. is also a measure of relative risk. That is, the probabilities obtained estimate the magnitude of risk, relative to the risk measured at a distance of zero from the release site.

RE-ANALYSIS OF EXISTING DATA

Data was obtained from the scientific literature by computer-searching bibliographies. From several hundred publications, those papers which provided suitable raw data on effective gene flow as a function of distance were selected and the data analysed in a consistent manner.

Source references for these papers and brief details of the experiments are given in Gliddon *et al.* (1999).

The estimates of parameters *a* and *b*, together with their 95% confidence intervals for a selected set of crops are given in Table 1. For completeness, the 'contamination rate' describing the percentage of marked genes recovered at distance zero is also given although, as described above, it plays no further part in the subsequent analysis. It can be seen from Table 1 that less than half of the experiments have estimates of *a* and *b* that do not include zero in their confidence interval (bold italics in the Table). This indicates that the experimental data for these experiments alone shows a SIGNIFICANT decrease in recovery of marked pollen as a function of distance. In the main, this failure to describe a decay in contamination with distance for many crops is regarded as an indication of the lack of appropriate data collected in the experiments rather than as a biological attribute of the crop.

Table 1: Estimates of *a* and *b*

Crop	a	a min	a max	b	b min	b max	Contamination	Range (metres)	
ALFALFA	0.099	-0.330	0.528	0.501	-0.272	1.274	0.620	50	1610
ALFALFA	0.007	-0.036	0.049	1.191	-0.367	2.748	0.201	5	105
MAIZE	***0.522***	***0.337***	***0.707***	***0.958***	***0.624***	***1.291***	***0.838***	***0.5***	***24***
MAIZE	***1.413***	***1.058***	***1.768***	***0.974***	***0.160***	***1.788***	***0.046***	***1***	***34***
POTATO	0.523	-2.514	3.560	0.942	-1.619	3.504	0.031	0.75	21.5
POTATO	0.577	-15.549	16.703	0.647	-22.105	23.398	0.001	0.75	9.5
RADISH	***1.643***	***0.703***	***2.583***	***0.761***	***0.215***	***1.306***	***0.855***	***0.2***	***14***
RADISH	0.274	-0.064	0.612	0.613	0.294	0.932	1.000	6	180
RADISH	0.214	-0.074	0.501	0.555	0.221	0.889	0.522	2.5	150
RAPE	0.835	-2.400	4.071	0.773	-1.856	3.401	0.028	1	70
RAPE	***0.750***	***0.554***	***0.946***	***0.206***	***0.102***	***0.156***	***0.019***	***n/a***	***n/a***
RAPE	***0.832***	***0.693***	***0.971***	***0.304***	***0.236***	***0.374***	***0.035***	***n/a***	***n/a***
RAPE	***1.332***	***1.053***	***1.611***	***0.535***	***0.451***	***0.619***	***0.051***	***3***	***55***
RAPE	***0.752***	***0.526***	***0.978***	***0.652***	***0.541***	***0.763***	***0.026***	***3***	***55***
RYEGRASS	0.018	-0.021	0.057	0.816	0.262	1.371	0.060	3	76
RYEGRASS	***0.133***	***0.073***	***0.194***	***0.774***	***0.593***	***0.955***	***0.348***	***0.5***	***915***
RYEGRASS	0.003	-0.001	0.007	1.555	1.191	1.919	0.454	0	100
RYEGRASS	0.031	-0.104	0.165	1.169	-0.023	2.361	0.454	10	70
SUGAR BEET	***1.310***	***0.330***	***2.290***	***0.524***	***0.097***	***0.952***	***0.141***	***0.2***	***23***
WHEAT	***0.483***	***0.385***	***0.581***	***0.872***	***0.684***	***1.060***	***0.254***	***0.025***	***30***
WHEAT	0.038	-0.020	0.096	0.870	0.411	1.329	0.107	3	48

a min, b min: lower 95% confidence limit of *a* and *b* respectively
a max, b max: upper 95% confidence limit of *a* and *b* respectively
Range: The distances from the souce over which the experiment was carried out

DISCUSSION

Variability within crops

A cursory glance at Table 1 shows that there can be immense variability in rates of decay estimated from different experiments for particular crops. The relative risk for the five experiments carried out with oil seed rape is shown graphically in Figure 1. In part, this

variability may be ascribed to the lack of good data and, therefore, the lack of statistical power in the estimates. However, an alternative explanation is that this variability is an inherent feature of the pollination biology of certain crop species.

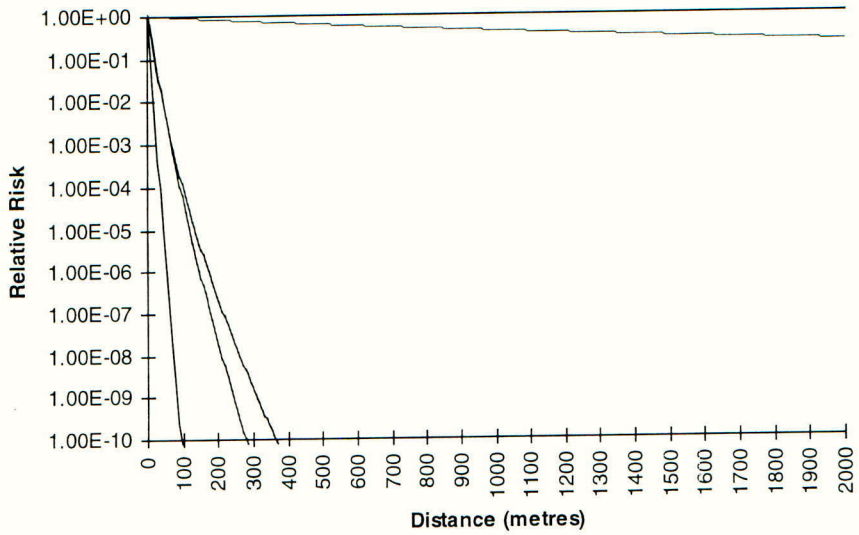

Figure 1. Relative risk in five different experiments using oil seed rape

Unfortunately, the existing data is not sufficient to discriminate between these two alternatives. The large body of literature on pollination biology certainly suggests that, at least for insect pollinated species, the crop species itself is unlikely to be a sufficient descriptor of expected pollen movement. Factors including the type and density of surrounding vegetation, flowering stage of other vegetation and meteorological conditions are likely to influence the distance that pollen is carried to a far greater extent than the species of pollen donor. It is obvious that such variability needs to be taken into account if carrying out a risk assessment as a precursor to the possible release of a GMO.

It is clear that for insect pollinated species in particular, there can be a very strong effect of experimental conditions determining the amount of gene flow measured. One striking component of the experimental design that has often been shown to have a major effect is that of barren zones or borders separating potential pollen recipients from donors. Goplen *et al.* (1972) reported on the effects of isolation distance on contamination in sweetclover (*Melilotus alba*). They used a recessive gene marker, *cu*, for low coumarin content. In one of their experiments, a source plot of 0.4 ha was located 46 m and 389 m from two sink plots of a similar size. Both these plots showed a similar average level of contamination (0.14 and 0.13 respectively). However, within each of these plots, the contamination declined rapidly with distance from the source. An explanation for the above observations is that pollinators arrive in the recipient plots loaded with pollen from the donors. They then unload their pollen as

they move within the recipient plots. However, the likelihood of arriving at one or other recipient plots is not determined, to any major extent, by the distance from the source.

In conclusion, there is a need to be extremely careful in using published data on isolation with distance as an element of risk assessment. This due to two factors: first, the data in the literature have frequently not been analysed appropriately for the purposes of risk assessment; and, secondly, there is often great variability among experiments describing pollen movement in a single crop. In general, the variability within crops can be assumed to be greater in insect-pollinated species than in wind pollinated species. However, there is often insufficient data in the literature to allow reasonably confident predictions to be made about the levels of contamination to be expected in almost all crops.

The statistical method described above provides a basic standard for the analysis of experiments designed to measure the decay of contamination by pollen as a function of distance for use in risk assessment. It should be noted that although only three parameters need be estimated in this method, the majority of existing data-sets are insufficient to describe significant decay of contamination with distance. This is a clear indictment of the utility of many existing descriptions of pollen movement in crop plants in the context of risk assessment. The development of more complex methods of statistical description, taking into account, for example, the effects of wind strength and direction on pollen movement (e.g. Lavigne *et al.*, 1996; Tufto *et al.*, 1997), will not render existing data more useful as they will require estimation of more parameters from very sparse data-sets

REFERENCES

Bateman A J (1947). Contamination in seed crops. III. Relation with isolation distance. *Heredity*, **1**, 303-336.

Dale P J (1992). Spread of engineered genes to wild relatives. *Plant Physiology*, **100**, 13-15.

Ellstrand N C; Hoffman C A (1990). Hybridization as an avenue of escape for engineered genes. *BioScience*, **40**, 438-442.

Gliddon C; Boudry P; Walker D S (1999). Geneflow – a review of experimental evidence. In: *Environmental impact of genetically modified crops*, eds A J Gray, F Amijee & C J Gliddon. Genetically Modified Organisms Research Report 10, pp. 67-81. DETR: London

Goplen B P; Cooke D A; Pankiw P (1972). Effects of isolation distance on contamination in sweetclover. *Canadian Journal of Plant Science*, **52**, 517-524.

Goudet J; De Meeüs T; Day A J; Gliddon C (1994). The different levels of population structuring of the dogwhelk, *Nucella lapillus*, along the south Devon coast. In: *Genetics and evolution of aquatic organisms*, ed. A R Beaumont, pp. 81-96. Chapman & Hall: London.

Haldane J B S (1948). The theory of a cline. *Journal of Genetics*, **48**, 277-284.

Kareiva P; Morris W; Jacobi C M (1994). Studying and managing the risk of cross-fertilization between transgenic crops and wild relatives. *Molecular Ecology*, **3**, 15-21.

Lavigne C; Godelle B; Reboud X; Gouyon P H (1996). A method to determine the mean pollen dispersal of individual plants growing within a large pollen source. *Theoretical and Applied Genetics*, **93**, 1319-1326.

Levin D A; Kerster H (1974). Gene flow in seed plants. *Evolutionary Biology*, **7**, 139-220.

Mackenzie D R; Henry S C (1990). Towards a consensus. In: *The biosafety results of field tests of genetically modified plants and microorganisms*. Proceedings of the Kiawah Island Conference, eds D R MacKenzie & S C Henry, pp. 273-283. Agricultural Research Institute: Bethesda, Maryland, USA.

Raybould A F; Gray A J (1993). Genetically modified crops and hybridization with wild relatives: a U.K. perspective. *Journal of Applied Ecology*, **30**, 199-219.

Raybould A F; Goudet J; Mogg R J; Gliddon C J; Gray A J (1996). The genetic structure of a linear population of sea beet (*Beta vulgaris* ssp. *maritima*) revealed by isozyme and RFLP analysis. *Heredity*, **76**, 111-117.

Raybould A F; Mogg R J; Gliddon C J (1997). The genetic structure of *Beta vulgaris* ssp. *maritima* (sea beet) populations .2. Differences in gene flow estimated from RFLP and isozyme loci are habitat-specific. *Heredity*, **78**, 532-538.

Scheffler J A; Parkinson R; Dale P J (1993). Frequency and distance of pollen dispersal from transgenic oilseed rape (*Brassica napus*). *Transgenic Research*, **2**, 356-364

Stringam G R; Downey R K (1978). Effectiveness of isolation distance in turnip rape. *Canadian Journal of Plant Sciences*, **58**, 427-438.

Tufto J; Engen S; Hindar K (1997). Stochastic dispersal processes in plant populations. *Theoretical Population Biology*, **52**, 16-26.

Wright S (1943). Isolation by distance. *Genetics*, **28**, 114-138.