Chapter 1 in: **Sampling Strategies for Marker Analysis.** Final Compendium of the Research Project **Development, optimisation and validation of molecular tools for assessment of biodiversity in forest trees** in the European Union DGXII Biotechnology FW IV Research Programme **Molecular Tools for Biodiversity**. Gillet, E.M. (ed.). 1999. URL http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm

Minimum sample sizes for sampling genetic marker distributions

Elizabeth M. Gillet[1,2]

[1] Institut für Forstgenetik und Forstpflanzenzüchtung, Universität Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

[2] Institut für Forstgenetik und Forstpflanzenzüchtung, Bundesforschungsanstalt für Forst- und Holzwirtschaft, Sieker Landstrasse 2, 22927 Grosshansdorf, Germany

Email: egillet@gwdg.de

Table of contents

TABLE 1: Minimum sample size for detecting all alleles present at frequencies not less than a given frequency in a deme

How many different alleles are present among the diploid individuals of a deme at a single gene locus? A **deme** is defined as a set of individuals, such as a family, stand, subpopulation, species,

etc. The only way to be sure that every allele is detected is to study every individual in the deme, hoping of course that any recessive alleles are present in homozygous form. Since it is seldom feasible to investigate the entire deme, the solution is to choose a sampling strategy that yields a high probability of detecting all alleles. If the true frequencies of the genotypes in the deme are known, Gregorius (1980) derived an exact formula for the detection probability for a sample of given size. By varying the sample size in this formula, the smallest size can be determined for which the probability of detecting all alleles reaches a desired value.

Usually, however, the true genotype frequencies are not known. Often, not even the number of different alleles in the deme is known. In such cases, the problem can be reformulated as: What is the minimum sample size for detecting all alleles that are not too rare? More precisely, for desired **threshold allele frequency** *alpha* and **detection probability** *epsilon*, what is the minimum sample size such that the probability is greater than or equal to *epsilon* that all alleles will be included in the sample that have frequencies not less than *alpha* in the deme.

The detection probability depends on how the alleles are associated to make up the diploid genotypes. For a given number of alleles at given frequencies, the detection probability assumes its minimum under complete homozygosity (Gregorius 1980). Reasoning intuitively, at most one new allele can be found per homozygous individual, as opposed to two in a heterozygous individual. The consequence is that the minimum sample size for detecting all alleles is greater under complete homozygosity than for all other forms of allelic association. A formula for the detection probability under complete homozygosity for alleles of known frequency and given sample size was derived by Gregorius (1980, Corollary 1).

If information is available neither about the number of alleles (rare or not) nor about their frequencies, calculations must proceed from the "worst case" single-locus genotype frequency distribution for which the smallest allele frequency is not less than *alpha*. This is the case of complete homozygosity where the number of alleles equals $n=[1/alpha]$ (*i.e.*, n equals the largest natural number less than or equal to $1/alpha$), of which $n-1$ alleles have the frequency *alpha* and the nth allele has frequency $1-(n-1)\,alpha$ (Gregorius 1980, Corollary 3). A formula for the detection probability for this "worst-case" situation is also given by Gregorius (1980, Corollary 3) and applied to calculate minimum sample sizes for different values of *alpha* and *epsilon* (his Table 1). This table is reproduced below.

Sampling strategy: For random sampling with replacement (or without replacement, if the deme is very large) among the diploid individuals of a deme, the following table lists the minimum sample size that ensures with probability *epsilon* that all alleles at a locus will be detected that are present at relative frequencies not less than *alpha*. Calculations proceed from the "worst case" of no prior knowledge about the number of alleles nor their frequencies and under the assumption of complete homozygosity.

Threshold allele frequency *alpha*	Detection probability *epsilon*		
	0.95	0.99	0.999
0.500	6	8	11
0.400	7	10	14
0.300	11	15	22
0.200	21	28	39
0.100	51	66	88
0.090	57	74	99
0.080	65	84	112
0.070	77	99	131
0.060	92	119	156
0.050	117	149	194
0.040	152	192	249
0.030	212	265	341
0.020	341	422	536
0.010	754	916	1146
0.009	850	1030	1285
0.008	972	1174	1462

Table 1: In order to detect all alleles of a locus that are present in a deme at frequencies not less than *alpha*, the minimum sample size is given such that probability of detection is always greater than *epsilon*, regardless of the actual number of alleles and the allele and genotype frequencies in the deme. (Reproduced from Gregorius 1980, Table 1)

Detecting haplotypes, cytotypes, and phenotypes

Since the probability of detecting all alleles when they are associated in complete homozygosity equals the probability of detecting all alleles in haploid individuals, a useful consequence is that the above table also lists the minimum sample sizes for detecting all alleles in haploid individuals. This carries over to the detection of all uniparentally inherited marker variants, such as chloroplast or mitochondrial DNA markers. In fact, the minimum sample sizes in the above table apply to any trait for which it holds that each individual exhibits exactly one trait state (*e.g.* phenotype, cytotype, haplotype).

Reference

Gregorius H-R (1980) The probability of losing an allele when diploid genotypes are sampled. *Biometrics* 36: 632-652.

[Table of Contents]

TABLE 2: Minimum sample size for detecting all alleles present at frequencies not less than a given frequency in a deme showing Hardy-Weinberg-Proportions

Consider again the situation of Table 1, but suppose that the mode of gene action at the locus is codominance and that the genotypes are known to show **Hardy-Weinberg-Proportions (HWP)** within the deme. This means that if p_i is the relative frequency of the allele A_i in the deme ($p_i>0$ for all i and $Sum_i\ p_i=1$), and if P_{ij} is the relative frequency of individuals of genotype A_iA_j in the deme ($Sum_{\ i.le.j}\ P_{ij}=1$, where "*.le.*" stands for "*less than or equal to*"), then the relative genotype frequencies equal

$$P_{ii}=p_i^2 \quad \text{for all } i \quad \text{and} \quad P_{ij}= 2p_i\, p_j \quad \text{for all } i,j \text{ with } i < j \qquad \text{(HWP)}$$

Fulfillment of HWP signifies that the alleles are randomly associated in the genotypes. This information about the organization of the alleles can be used to refine the sampling strategy set forth in Table 1.

Again, a sampling strategy is sought that yields a high probability of detecting all alleles that are not too rare. The precise formulation remains the same: For desired **threshold allele frequency** *alpha* and **detection probability** *epsilon*, the probability should be greater than or equal to *epsilon* that all alleles will be included in the sample that have frequencies not less than *alpha* in the deme.

If the allele frequencies in the deme are known, Gregorius (1980, text following Corollary 1) showed that the minimum sample size when the genotype frequencies fulfill HWP is equal to one-half the minimum sample size for complete homozygosity (see Table 1). If information is available neither about the number of alleles nor about their frequencies, calculations must proceed from the "worst-case" single-locus genotype frequency distribution showing HWP for which the smallest allele frequency is not less than *alpha*. Again, this occurs when the number of alleles equals $n=[1/alpha]$, $n-1$ of which with frequency *alpha* and the *nth* allele with frequency $1-(n-1)\,alpha$. The result is that for HWP, the minimum sample size for detection of all alleles is one-half that for complete homozygosity. The table below thus results by halving the minimum sample sizes of Table 1.

Sampling strategy: For random sampling with replacement (or without replacement, if the deme is very large) among the diploid individuals of a deme that shows HWP at a locus that shows codominance of gene action, the table lists the minimum sample size that ensures with probability *epsilon* that all alleles at the locus will be detected that are present at relative frequencies not less than *alpha*. Calculations proceed from the "worst case" of no prior knowledge about the number of alleles nor their frequencies and under the assumption that the genotype frequencies show HWP.

Threshold allele frequency *alpha*	Detection probability *epsilon*		
	0.95	0.99	0.999
0.500	3	4	6
0.400	4	5	7
0.300	6	8	11
0.200	11	14	20
0.100	26	33	44
0.090	29	37	50
0.080	33	42	56
0.070	39	50	66
0.060	46	60	78
0.050	59	75	97
0.040	76	96	125
0.030	106	133	171
0.020	171	211	268
0.010	377	458	573
0.009	425	515	643
0.008	486	587	731

Table 2: Assume that the genotype frequencies in a deme show HWP at a locus that shows codominance of gene action. In order to detect all alleles that are present at this locus at frequencies not less than *alpha*, the minimum sample size is given such that the probability of detection is greater than *epsilon*, regardless of the actual number of alleles and their frequencies. (After Gregorius 1980, by halving the values in his Table 1)

Reference

Gregorius H-R (1980) The probability of losing an allele when diploid genotypes are sampled. *Biometrics* 36: 632-652.

[Table of Contents]

TABLE 3: Minimum sample size for detecting all haplotypes among the haploid gametophytes produced by a single tree that is heterozygous at a given number of loci

If a tree is heterozygous at *m* gene loci, then it can produce haploid gametophytes (ovules whose haplotypes are inferable *e.g.* from the primary endosperm of conifer seeds, or pollen) that have any of 2^m different multilocus haplotypes. Assuming regular segregation and absence of linkage

between the loci, each of the 2^m haplotypes has equal chances of being formed, and thus the haplotypes should be uniformly distributed among the gametophytes. How should sampling be performed so that all haplotypes will be detected in a single sample? A sampling strategy is needed that yields a chosen high probability *epsilon* of including all haplotypes in the sample.

Sampling strategy: For random sampling with replacement (or without replacement, if the base set of gametophytes is very large) among the gametophytes, the table lists the minimum sample size that ensures with probability *epsilon* that all haplotypes will be detected. These sample sizes were calculated using a formula of Gregorius (1980) based on the respective number of uniformly distributed haplotypes.

No. *m* of heterozygous loci	No. of haplotypes = 2^m	Frequency of each haplotype $=1/2^m$	Detection probability *epsilon* of all haplotypes		
			0.95	0.99	0.999
1	2	0.500000	6	8	11
2	4	0.250000	16	21	29
3	8	0.125000	38	51	68
4	16	0.062500	90	115	150
5	32	0.031250	203	255	327
6	64	0.015625	453	557	703

Table 3: If a tree is heterozygous at *m* unlinked loci, each showing codominance of gene action and regular segregation, its gametophytes should have 2^m different haplotypes, each with relative frequency $1/2^m$. The table lists the minimum sample size such that probability of detection of all 2^m haplotypes is greater than *epsilon*. (Reproduced from Gillet 1998)

References

Gillet EM (1998) HAPLOGEN - User's Manual : Qualitative inheritance analysis of zymograms and DNA electropherograms in haploid gametophytes. URL http://www.uni-forst.gwdg.de/forst/fg/index.htm

Gregorius H-R (1980) The probability of losing an allele when diploid genotypes are sampled. *Biometrics* 36: 632-652.

[Table of Contents]

TABLE 4: Minimum sample size for detecting all multilocus genotypes in a progeny from self-fertilization of an individual

If a tree that is heterozygous at *m* diploid gene loci showing codominance of gene action can reproduce by self-fertilization, then it can produce offspring with any of 3^m different multilocus genotypes over the *m* loci. Assuming regular segregation at each locus, and absence of linkage

between loci, the relative frequencies of the different genotypes will differ. However, the "rarest" types - the complete homozygotes - should have frequencies of approximately $(1/4)^m$. How should sampling be performed so that all multilocus genotypes will be detected in a single sample? Once again, a sampling strategy is needed that yields a chosen high probability *epsilon* of including all genotypes in the sample.

Sampling strategy: For random sampling with replacement (or without replacement, if the base set of offspring is very large) among the gametophytes, the table lists the minimum sample size that ensures with probability *epsilon* that all genotypes will be detected. These sample sizes were calculated using a formula of Gregorius (1980) based on the number of genotypes and the frequency of the "rarest" genotype.

No. *m* of heterozygous loci	No. of offspring genotypes = 3^m	Frequency of "rarest" genotype = $(1/4)^m$	Detection probability *epsilon* of all genotypes		
			0.80	0.90	0.95
1	3	0.500000	9	13	19
2	9	0.062500	57	79	104
3	27	0.015625	304	396	500
4	81	0.003906	1504	1879	2297

Table 4: If a tree is heterozygous at *m* loci, then the number of different genotypes among the offspring equals 3^m, and the relative frequency of the rarest genotype equals $(1/4)^m$. The minimum sample size is listed such that probability of detecting all genotypes is greater than *epsilon*. (Reproduced from Gillet 1998)

References

Gillet EM (1998) DIPLOGEN - User's Manual : Qualitative inheritance analysis of zymograms and DNA electropherograms in diploid individuals. URL http://www.uni-forst.gwdg.de/forst/fg/index.htm

Gregorius H-R (1980) The probability of losing an allele when diploid genotypes are sampled. *Biometrics* 36: 632-652.

[Table of Contents]

TABLE 5: Minimum sample sizes for qualified estimation of the frequency distribution of types within a single deme

How can the relative frequencies of the types (genotypes, phenotypes) present in a deme be estimated? A sampling strategy is needed such that the probability will be large that the sample frequencies will not deviate by more than a given amount from the true frequencies. The following considerations are based on the system analytic approach developed by Gregorius (1998), which is explained below.

A method of investigation - or sampling strategy - includes specification of sample size. If a hypothesis about the true distribution *P* of the *n* types in the deme is given, then the method is defined to be **qualified** for *P*, if the sample size is large enough such that

$$Prob_P\,(d_0\,(S,U)\ .ge.\ lambda)\ \ .le.\ epsilon$$

holds, where "*.ge.*" stands for "*greater than or equal to*", "*.le.*" stands for "*less than or equal to*", and where

- $S = (s_i)_{i=1,...,n}$ is the random variable representing the vector of relative frequencies of the types in samples drawn from a deme containing individuals of *n* different types, where the sampling strategy specifies the sample size and sampling is done with replacement (or without replacement if the deme is very large);
- $d_0\,(S,P)$ is the genetic distance between *S* and the hypothetical distribution $P = (p_i)_{i=1,...,n}$ where $d_0\,(S,U) = 1/2 \times Sum_{i=1,...,n}\,|s_i-p_i|$ (Gregorius 1974a,b)
- $Prob_P$ denotes the multinomial sampling distribution under *P* for the given sample size;
- The left side of the above inequality, $Prob_P\,(d_0\,(S,P)\ .ge.\ lambda)$, is termed the **critical probability** corresponding to the **critical d_0-discrepancy** *lambda*.
- The value of *epsilon* specifies the **level of significance** and lies between 0 and 1.
- The values of *lambda* and *epsilon* are termed the **levels of qualification** of the method of investigation.

In words, qualification means that the relative frequency distribution observed in a sample of size specified by the method of investigation will have a given large probability of not deviating from the true frequency distribution by more than a given threshold amount. The objective is to choose a method of investigation that is qualified for the hypothesis *P*. If this is fulfilled, then the hypothesis *P* is qualifiedly rejected if $d_0\,(S,P)\ .ge.\ lambda$ holds and qualifiedly accepted if $d_0\,(S,P) < lambda$ holds.

If no information is available about the true distribution *P*, then it is necessary to choose a method of investigation that is qualified for all possible distributions of *n* types. It can be shown that if the method of investigation is qualified for the uniform distribution $U = (u_i)_{i=1,...,n}$ where $u_i=1/n$, then it is qualified for all distributions of *n* types. *U* can thus be termed the "worst-case" distribution of *n* types.

Sampling strategy: For sampling with replacement within a single deme containing a given number *n* of types, the following table gives the minimum sample size that ensures qualification (Gregorius 1998) on given levels for the "worst-case" frequency structure *U*. For up to 4 types, sample sizes were calculated using exact probabilities; for larger numbers of types, sample sizes were estimated by Monte-Carlo methods.

No. *n* of different types	Critical d_0-discrepancy *lambda*					
	0.05	0.06	0.07	0.08	0.09	0.10
2	401*	274*	208*	156*	128*	106*
3	499*	349*	256*	199*	154*	130*
4	610	424	315*	237*	189*	156*
5	721	496	371	280	221	181
6	817	567	420	325	253	211

Table 5: For a deme that contains individuals of *n* different types, the minimum sample size for qualified estimation of the frequency distribution at level of significance *epsilon*=0.05 is given for five different critical d_0-discrepancies *lambda.* An asterisk (*) signifies that the exact critical probability was calculated. Otherwise, minimum sample sizes numbers were estimated using Monte-Carlo methods.

References

Gregorius H-R (1998) The system analytical approach to the study of hypotheses. URL http://www.uni-forst.gwdg.de/forst/fg/index.htm

Gregorius H-R (1974a) On the concept of genetic distance between populations based on gene frequencies. Proceedings, Joint IUFRO Meeting S.02-04, pp. 17-22.

Gregorius H-R (1974b) Genetischer Abstand zwischen Populationen. I. Zur Konzeption der genetischen Abstandsmessung. *Silvae Genetica* 23: 22-27.

[Table of Contents]

TABLE 6: Minimum samples sizes for qualified joint estimation of the frequency distributions of types within several demes for estimation of multiple-deme parameters

How can the relative frequencies of the types (*e.g.* genotypes, phenotypes) present in each of several demes be estimated jointly? Joint frequency estimates are necessary for the estimation of multiple-deme parameters. A sampling strategy is needed such that the probability will be large that the joint deviation of the sample frequencies from the true frequencies in the demes will not exceed a chosen amount. The following considerations are based on the system analytic approach developed by Gregorius (1998) (also see explanations for Table 5).

The *m* demes are assumed to be given, with the relative size of deme *j* equal to c_j, where $c_j > 0$ and $Sum_j\, c_j = 1$. Denote $p_i\,(j)$ as the relative frequency of type *i* in deme *j*, where $Sum_i\, p_i\,(j) = 1$. These deme frequencies are translated into a weighted relative **joint frequency distribution** by multiplying each frequency by c_j, *i.e.*, $p_{ij}=c_j\, p_i\,(j)$. Then, $Sum_{i,j}\, p_{ij} = 1$ holds.

Assume that the *m* demes are of equal relative sizes, *i.e.*, $c_j=(1/m)$ for all *j*. If a hypothesis about the true distribution *P* of the *n* types in each of the *m* demes is given, then a method of investigation involving independent random samples from each deme is defined to be **qualified** for *P*, if the

sample sizes for all demes are large enough such that

$$Prob_P\,(d_0\,(S,P)\ \ .ge.\ lambda)\ .le.\ epsilon$$

holds, where "*.ge.*" stands for "*greater than or equal to*", "*.le.*" stands for "*less than or equal to*", and where

- $S = (s_{ij})_{\,i=1,...,n,\ j=1,...,m}$ is the random variable representing the vector of weighted relative joint frequencies of the types in independent random samples drawn from each of a set of demes containing individuals of *n* different types, where the sampling strategy specifies the sample size per deme and sampling is done with replacement (or without replacement if the demes are very large);
- $d_0\,(S,P)$ is the measure of discrepancy between *S* and the hypothetical distribution $P = (p_{ij})_{i=1...,n,\ j=1,...,m}$ where $d_0\,(S,P) = 1/2 \times Sum_{i=1,...,n,\ j=1,...,m}\ |s_{ij}-p_{ij}|$ (Gregorius 1974a,b);
- $Prob_P$ denotes the multinomial sampling distribution for *P* for the given sample sizes in the demes;
- The left side of the above inequality, $Prob_P\,(d_0\,(S,P)\ .ge.\ lambda)$, is termed the **critical probability** corresponding to the **critical d_0-discrepancy** *lambda*.
- The value of *epsilon* specifies the **level of significance** and lies between 0 and 1.

In words, qualification means that the relative joint frequency distribution observed in the deme samples will have a given large probability of not deviating from the true frequencies by more than a given threshold value. The objective is to choose a method of investigation that is qualified for the hypothesis *P*. If this is fulfilled, then the hypothesis *P* is qualifiedly rejected if $d_0\,(S,P)\ .ge.\ lambda$ holds and qualifiedly accepted if $d_0\,(S,P) < lambda$ holds.

If no information is available about the true distribution *P*, then it is necessary to choose a method of investigation that is qualified for all possible distributions of *n* types. It can be shown that if the method of investigation is qualified for the uniform distribution $U = (u_{ij})_{\,i=1,...,n,\ j=1,...,m}$ where $u_i = 1/(mn)$, then it is qualified for all distributions of *n* types in *m* demes. *U* can thus be termed the "worst-case" distribution of *n* types in *m* demes.

Sampling strategy: For independent random sampling with replacement within a given number *m* of demes of equal relative sizes that each contain individuals possessing any one of a given number *n* of types, the following table gives the minimum sample size that ensures qualification (Gregorius 1998) on given levels for the "worst-case" frequency structure *U*. The sample sizes were estimated using Monte-Carlo methods.

No. of demes of equal relative sizes	No. *n* of types present in the demes	Critical d_0-discrepancy *lambda*	
		0.05	0.10
2	2	240	60
2	3	350	87
2	4	440	110
2	5	544	138
3	2	207	52
3	3	300	75
3	4	391	101
3	5	481	121
4	2	175	45
4	3	271	68
4	4	357	90
4	5	445	110
100	2	81	20
100	3	152	38
100	4	220	56
100	5	289	73

Table 6: For given number of demes of equal relative sizes that contain individuals of *n* different types, the minimum sample size for each deme for qualified joint estimation of the demes' frequency distributions at level of significance *epsilon*=0.05 is given for two different values of the critical d_0-discrepancy *lambda*.

References

Gregorius H-R (1998) The system analytical approach to the study of hypotheses. URL http://www.uni-forst.gwdg.de/forst/fg/index.htm

Gregorius H-R (1974a) On the concept of genetic distance between populations based on gene frequencies. Proceedings, Joint IUFRO Meeting S.02-04, pp. 17-22.

Gregorius H-R (1974b) Genetischer Abstand zwischen Populationen. I. Zur Konzeption der genetischen Abstandsmessung. *Silvae Genetica* 23: 22-27.