Optimal DNA Pooling for the Detection of Single Nucleotide Polymorphisms

David M. Ramsey[1], Andreas Futschik[2]

[1] Department of Mathematics & Statistics, University of Limerick, Limerick, IRELAND, e-mail: david.ramsey@ul.ie
[2] Department of Statistics, University of Vienna

Abstract: We consider the optimal pooling of DNA to detect single nucleotide polymorphisms (SNPs), sites along the genome at which a population shows variation. The focus is on the detection of low frequency variants. Pooling individuals increases the probability that a rare variant appears in the sample. However, as the pool size increases, the mean number of reads from an individual decreases, making it harder to distinguish reads of a rare variant from errors. A hypothesis test for the detection of SNPs is defined. On the basis of this test, we determine the asymptotically optimal pool size given the parameters of the genome sequencer used, the number of lanes available and a specified significance level.

Keywords: genome sequencing; optimal pooling; single nucleotide polymorphisms.

1 Introduction

The genome consists of sequences made of 4 nucleotides (bases). At a majority of the sites in these sequences, each individual in a population has the same base. A site where there is variation is called a single nucleotide polymorphism (SNP). At such sites, in general, just two of the four bases appear. These variants are called alleles, the most common (rare) is termed the major allele (minor allele, respectively). We treat chromosomes, rather than members of a species, as individuals. However, our analysis can be generalized.

Since any reasonable test detects alleles of relatively large frequency with power close to 1, we concentrate on the detection of low frequency alleles. Following Futschik and Schlötterer (2010), one may use the following test: accept that there is a minor allele if in any lane the number of reads for a non-major allele exceeds a given threshold. We develop their work by specifying this threshold given the parameters of the sequencer and significance level required. An estimate of the power of this test is derived, which is used to find the optimal pool size for detecting low frequency alleles. For more on the practical issues involved in gene pooling see Kenny *et al.* (2010).

2 Description of the Problem and a Simplified Model

Genome sequencers read DNA from a pool (of m individuals) placed in a lane. Suppose we have k independent pools, i.e. the sample size is $n = km$. Consider a given site. Each lane gives a random number of reads for that site. If the same (large) amount of genetic material is taken from each individual, we may assume that the number of reads from an individual given that there are r reads in a lane has a binomial distribution with parameters r and $1/m$. Assume that each read is incorrect with a small probability ϵ, independently of other reads. Also, suppose that only two alleles are possible, the major allele and the putative minor allele.
Let $\mathbf{R} = (R_1, R_2, \ldots, R_k)$, where R_i is the total number of reads for that site in lane i. It is assumed that the R_i are i.i.d. from the Poisson(λ) distribution. In addition, suppose good estimates of λ and ϵ are available for the gene sequencer used.
The major allele is inferred to be the one with the largest number of reads in the whole sample. As we are interested in detecting low frequency alleles, we may assume that for reasonable sample sizes the major allele is correctly identified with probability 1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, where X_i is the number of reads of the putative minor allele in lane i.
Denote the minor allele frequency at a given locus by p. We wish to define an optimal pooling procedure (maximizing power) while controlling the type I error rate for a test of the following hypotheses.
H_0: The locus is not a SNP, i.e. $p = 0$.
H_A: $p = p_0$, where p_0 is some small positive value.

3 A Test for the Presence of a Minor Allele

Consider the test statistic $U = \max_{1 \le i \le k} X_i$, i.e. U is the maximum number of reads of a putative minor allele in a lane. Hence, under H_0, U is the maximum of independent observations from the Poisson($\lambda\epsilon$) distribution. The critical value for the test, u_k, is the smallest integer satisfying

$$P(U \le u_k | H_0) \ge 1-\alpha \Rightarrow P(X_i \le u_k | H_0)^k \ge 1-\alpha \Rightarrow P(X_i \le u_k | H_0) \ge \sqrt[k]{1-\alpha}.$$

Thus we can take the $\sqrt[k]{1-\alpha}$ quantile of the Poisson($\lambda\epsilon$) distribution as the critical value. We reject H_0 if and only if $U > u_k$. Note that this procedure takes into account the fact that we essentially have a multiple testing problem based on k test statistics $X_1, X_2, \ldots, X_k$. The critical value used in the test can be approximated using the Bonferroni procedure. However, this test does not take into account that such a procedure is repeated for each site. Hence, the value of α chosen should reflect this.
Under H_A, the number of minor alleles in the sample has a Bin(n, p_0) distribution. This can be approximated by the Poisson(np_0) distribution.

Result. *When there are b individuals with the minor allele, the distribution of the test statistic stochastically dominates the distribution of this statistic when one individual with the minor allele appears in each of b lanes.*
Let D denote the event that H_A is accepted given that it is true. Let the number of individuals with the minor allele in the sample be B and $\mu = E[B] = mkp_0$. We obtain

$$P[D] = \sum_{b=0}^{\infty} P[D|B=b]P(B=b) \geq \sum_{b=1}^{\infty} P[D|B=b]P(B=b).$$

Since $P[D|B=0] \leq \alpha$, we can treat the resulting bound as a good approximation of $P[D]$. For $b \geq 1$,

$$P[D|B=b] = P(U > u_c|B=b) = 1 - P(U \leq u_k|B=b) \geq 1 - P(V_1 \leq u_k)^b,$$

where V_1 is the number of correct reads from one individual. If p_0 is small enough to neglect the possibility of two individuals with the minor allele being in a pool, it follows that $P[D|B=b] \approx 1 - q_k^b$, where

$$q_k = \sum_{j=0}^{u_k} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}.$$

Hence,

$$P[D] \approx \sum_{b=1}^{\infty} \frac{e^{-\mu}\mu^b[1-q_k^b]}{b!} = 1 - e^{-\mu(1-q_k)}.$$

Since the exponent in this expression is linear in p_0, the asymptotically (as $p_0 \to 0$) optimal pool size is independent of p_0.

4 Results from Simulations

Simulations were carried out for each of the following models:

1. Mistakes from reading the major allele always resulted in observing the same allele (the minor allele, if one was present). Mistakes in reading the minor allele always resulted in observing the major allele.

2. Mistakes from reading an allele always resulted in observing the same allele (neither the major allele nor the minor allele, if one was present).

3. Mistakes from reading any allele gave the other three possibilities with equal probability.

It should be noted that Model 1 corresponds to the model described in Section 2. Under Models 2 and 3, more than two alleles can be observed at a site. In these cases, as before, the major allele is assumed to be the

allele with the largest number of reads in the whole sample. The putative minor allele is taken to be the non-major allele with the largest number of reads from a single lane. Note that it is possible to correctly reject H_0, but incorrectly infer which base is the minor allele. For such an error to occur, it is necessary for the number of errors in a lane to exceed both the threshold and the number of reads of the real minor allele. Hence, the probability of such an error is less than α. Tables 1-3 give results based on 10,000 simulations in each case. In each case $p = 0.01$ and $\alpha = 0.001$. It can be seen that the optimal pool size and empirical power are robust to deviations from the assumptions of the model.

TABLE 1. Optimal pool sizes (derived by simulation), theoretical and estimated power under Model 1. The power estimated by simulation is given in brackets.

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
ϵ=0.01	4, 0.3752 (0.3864)	3, 0.6145 (0.6252)	3, 0.8514 (0.8489)	3, 0.9427 (0.9425)
ϵ=0.005	4, 0.3752 (0.3825)	4, 0.6915 (0.6973)	4, 0.9048 (0.9065)	4, 0.9706 (0.9728)
ϵ=0.002	6, 0.4628 (0.4733)	6, 0.7885 (0.7995)	7, 0.9553 (0.9583)	4, 0.9706 (0.9707)
ϵ=0.001	7, 0.4628 (0.4678)	7, 0.7885 (0.7946)	6, 0.9553 (0.9581)	6, 0.9905 (0.9917)

TABLE 2. Optimal pool sizes, estimated power and the probability of wrongly determining the minor allele (given in brackets) under Model 2.

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
ϵ=0.01	4, 0.3696 (0.0006)	3, 0.6167 (0.0002)	3, 0.8504 (0.0000)	3, 0.9388 (0.0000)
ϵ=0.005	4, 0.3729 (0.0001)	4, 0.6948 (0.0001)	4, 0.9060 (0.0001)	4, 0.9728 (0.0000)
ϵ=0.002	6, 0.4723 (0.0001)	6, 0.7917 (0.0001)	6, 0.9588 (0.0001)	4, 0.9712 (0.0000)
ϵ=0.001	6, 0.4699 (0.0000)	6, 0.7883 (0.0000)	5, 0.9535 (0.0001)	6, 0.9907 (0.0000)

TABLE 3. Optimal pool sizes, estimated power and the probability of wrongly determining the minor allele (given in brackets) under Model 3.

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
ϵ=0.01	4, 0.3752 (0.0000)	3, 0.6133 (0.0000)	3, 0.8549 (0.0000)	3, 0.9421 (0.0000)
ϵ=0.005	4, 0.3766 (0.0000)	4, 0.6993 (0.0000)	4, 0.9056 (0.0000)	4, 0.9703 (0.0000)
ϵ=0.002	6, 0.4694 (0.0001)	6, 0.7923 (0.0000)	7, 0.9575 (0.0000)	4, 0.9712 (0.0000)
ϵ=0.001	6, 0.4711 (0.0000)	6, 0.7942 (0.0000)	6, 0.9546 (0.0000)	6, 0.9922 (0.0000)

Acknowledgments: D. M. Ramsey is grateful for the support of Science Foundation Ireland under the BIO-SI project (no. 07MI012)

References

Futschik A. and Schlötterer C. (2010). Massively parallel sequencing of pooled DNA sample - the next generation of molecular markers *Genetics*, **186**, 207-218.

Kenny E. M., Cormican P., Gilks W. P., Gates A. S., O'Dushlaine C. T., Pinto C., Corvin A. P., Gill M., Morris D. W. (2010) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Research* , doi: 10.1093/dnares/dsq029