1 Introduction

This lecture is about detecting deletion polymorphisms from SNP genotype data. We first consider the "footprint" that a segregating deletion can leave in SNP genotype data on unrelated individuals and in mother–father–child trios. Next we extend the haplotype-based likelihood model (see lecture III) to accommodate deletions, and consider the implications of that model extension to the EM algorithm. Finally, we formulate the deletion detection problem as a model comparison task.

Generic methodology to be learned in this lecture (by examples): extending a model; challenges in model comparison; extending algorithms.

2 Deletion polymorphisms

Copy number variants (CNV) are junks of DNA whose copy number in the genome may vary between individual chromosomes. Thus, a certain variant may occur five times in one individual, but, say, just once in another individual. *Deletion polymorphisms* are special CNVs whose copy number is either 0 or 1. It is wildely believed that deletions may have a significant role both in the genetics of complex traits and in genome evolution.

3 Footprints of deletions in SNP genotype data

Deletions cannot be observed directly from SNP genotype data. Instead a *hemizygous* deletion (in a diploid individual) will be read as a run of homozygous genotypes at the SNPs spanned by the deletion. That is, the underlying pair AD, where "D" stands for a deletion, is read as the genotype AA, similarly aD as aa. A homozygous deletion is likely to cause a "no-call" in the genotyping measurement. Since no calls may occur also due to reasons other than deletions, one cannot detect even homozygous deletions directly. However, already an unusually run of consequtive markers with no-calls or homozygous genotypes in a single individual is a good infication of a possible deletion. But, since homozygous deletions (for no-calls) are very rare if the deletion is rare, and since only a very long region with all-homozygous markers in an individual could not be explained by long frequent haplotypes, in practice, deletions need to be inferred based on statistical signals in a sample of genotypes (or haplotypes). An example of such a signal is an excess of homozygous genotypes, or in other words, deviation from HWE.

In addition to the said signals, deletions can result in observed Mendelian inconsistencies in mother–father–child trios. For instance, suppose the mother has AD and the father has aa at a locus, and the deletion is transmitted to child. Then the child has the genotype Da, which is, however, observed as aa: now, as the mother's genotype is observed as AA, the observations contradict Mendelian inheritance (assuming no deletion or genotyping errors). However, such apparent inconsistencies may occur also due to reasons other than deletions, mainly genotyping errors. Thus, again, a single clue from an individual trio is not sufficient, but a bunch of such clues need to be aggregated in a larger sample of trios.

In summary, the following signals speak for the presence of a deletion polymoprhism if clustered into nearby markers in several individuals or trios:

- No-calls (sometimes called also "missing data" or "null genotypes"),
- deviation from Hardy–Weinberg equilibrium,
- apparent Mendelian inconsistencies in trios.

4 Haplotype-based likelihood approach

Suppose we are given a sample of genotypes (unrelated or trios) over a set of SNP markers. We are interested in deciding whether there is a deletion that spans the SNPs and occurs in some of the sampled genotypes. Our approach is to extend the haplotype inference methods described in previous lectures (Week III) to allow for a "deletion haplotype". We will only consider the more involved case of mother–father–child trios.

We extend the key notation and concepts as follows. We denote a deleted allele (at any locus) by D. The *deletion haplotype* is the sequence $\mathrm{DD}\cdots\mathrm{D}$ of length m, denoted d for short. Thus a haplotype is an element of the combined set $H_1 = H_0 \cup \{\mathrm{d}\}$, where $H_0 = \{0,1\}^m$ consists of the usual haplotypes over the m SNPs. A genotype is an element of $\{0,1,2,\mathrm{N}\}^m$, where N stands for a no-call. A haplotype pair, or "the true genotype", $\{h, h'\}$ is called a *homozygous deletion* if both the haplotype are the deletion haplotype $h = h' = \mathrm{d}$, and a *hemizygous deletion* if exactly on of the two haplotypes is the deltion haplotype, $h \in H_0$ and $h' = \mathrm{d}$.

It is practical to allow for probabilistic relations between the true and the observed genotypes. To this end we let $\tau(g; h, h')$ denote the probability of observing g when the true genotype is $\{h, h'\}$. We assume that these probabilities factorize over the markers:

$$\tau(g; h, h') = \prod_{s=1}^{m} \tau_s(g_s; h_s, h'_s)\,,$$

where $\tau_s(g_s; h_s, h'_s)$ is the probability of observing the genotype g_s at SNP s given that the underlying true allele pair is $\{h_s, h'_s\}$. Thus, we may think of τ_s as a 6×4 matrix. In these matrices the entries $(00, 0)$, $(01, 1)$, $(11, 2)$, $(0\mathrm{D}, 0)$, $(1\mathrm{D}, 2)$, and $(\mathrm{DD}, \mathrm{N})$ are typically set to close to 1, corresponding to the most likely observations given the true state of the genotype.

We may now write the probability of an observed trio genotype $\mathbf{g}$ given the haplotype frequency parameters p_h, $h \in H_1$, as

$$\Pr(\mathbf{g}; p) = L(p; \mathbf{g}) = \sum_{i,i',j,j' \in H_1} p_i p_{i'} p_j p_{j'} \tau(g^{\mathrm{m}}; i, i') \tau(g^{\mathrm{f}}; j, j') \tau(g^{\mathrm{c}}; i, j)\,;$$

$L(p; \mathbf{g})$, or sometimes $L(\mathbf{g})$ for short, is called the *likelihood of p* (for $\mathbf{g}$). For a sample of trio genotypes G, we assume independence: $\Pr(G; p) = \prod_{\mathbf{g} \in G} L(p; \mathbf{g}) = L(p; G)$.

The EM algorithm presented in Week III for finding maximum-likelihood haplotype frequencies extends to the present case in a straightforward manner. The update rule for the frequency of haplotype $h \in H_1$ becomes

$$p_h^{(t+1)} = \frac{1}{4n} \sum_{\mathbf{g} \in G} \big(I_h(\mathbf{g}) + I'_h(\mathbf{g}) + J_h(\mathbf{g}) + J'_h(\mathbf{g}) \big) / L(\mathbf{g})\,,$$

where the $I_h(\mathbf{g})$, $I'_h(\mathbf{g})$, $J_h(\mathbf{g})$, and $J'_h(\mathbf{g})$ are defined just as before, only replacing the haplotype set $\{0,1\}^m$ by the larger set H_1 and the error models $\tau_s(g_s; h_s + h'_s)$ by the slightly modified counterparts $\tau_s(g; h_s, h'_s)$ introduced above. (Computational implications of these modifications are examined in the exercises.)

5 On testing the hypothesis of no deletion

Suppose we are given a sample of n genotypes G (unrelated or trios) over a set of SNPs, and we are asked to estimate whether a deletion polymorphims covering the SNPs is present in the sample.

This task can be formulated as a parameter inference problem: estimating the frequency of the deletion haplotype, p_{d}, should yield a reasonable guess concerning the presence of a deletion. Namely, if p_{d} is substantially larger than 0, say $p_{\mathrm{d}} \geq 1/(2n)$ (corresponding to at least one occurrence in the sample), then the deletion probably is present; otherwise we would expect to get the estimate $p_{\mathrm{d}} = 0$ by maximum-likelihood inference. A shortcoming of this approach, however, is that it does not provide any quantification of the statistical significance associated with the claim that a deletion is present or absent.

To quantify significance, one may take an alternative approach: treat the issue as a formal hypothesis testing problem. The null hypothesis is that here is no deletion, and the alternative is that there is (or can be) a deletion. The former hypothesis amount to modelling the possible haplotypes by the set H_0, whereas the latter hypothesis amounts to modelling the possible haplotypes by the set H_1. Both are composite hypotheses, meaning that they define a class of probability models for the data (sampled genotypes); fixing the values of the parameters, the hapotype frequencies in our case, specifies a member of the class. Maxmimum-likelihood approaches to model class comparison or hypothesis testing compare the respective maximum likelihoods

$$\hat{L}_r(G) = \max\{L(p; G) : p \text{ is over } H_r\}\,, \qquad r = 0, 1\,.$$

Note that $\hat{L}_1(G)$ is at least as large as $\hat{L}_0(G)$, since the models are nested, that is, $H_0 \subseteq H_1$. Therefore, one, of course, should not always select the model that simply yields the larger maximum-likelihood.

The *likelihood ratio test* (LRT) is commonly applied in the case of nested models. The test statistic is

$$\lambda = \ln \frac{\hat{L}_1(G)}{\hat{L}_0(G)}\,.$$

Provided that certain conditions hold and the null hypothesis (the simpler model) is assumed to be true, 2λ follows asymptotically the $\chi^2(d)$ distribution, where $d > 0$ is the difference of the number of parameters of the models, $d = 1$ in our case. This, in principle, gives us a handy way to assign a p-value for the observations under the null hypothesis.

Practice has shown, however, that the asymptotics do not apply, at least with practical sample sizes. The reason is probably the fact that the maximum-likelihood parameter estimates p are not always interior points in the parameter space. In particular, the introduction of a deletion in the larger model may render the estimated frequency of some other haplotype to be zero, even though the maximum-likelihood estimate of that frequency is nonzero in the simpler model that does not allow for a deletion. Fortunately, the $\chi^2(1)$ distribution can be replaced either by a distribution obtained by simulating genotypes according to the null model or by an empirical distribution of the statistic obtained with data that is known or expected to contain no deletion polymorphims.

Exercises

V:1 (a) Extend the definition of the pure parsimony problem to allow for a deletion haplotype. (Hint: redefine $\{h, h'\}$ *explains* g.) (b) What modifications are needed in Gusfield's ILP model to make it valid for the extended pure parsimony problem?

V:2 Show that the values $I_i(\mathbf{g})$ for all $i \in H_1 = \{0,1\}^m \cup \{\mathrm{d}\}$ can be computed in time $O(m2^m)$, given the trio genotype $\mathbf{g}$ and the model parameters (haplotype frequencies p_h, $h \in H_1$ and the error models τ_s, $s = 1, 2, \ldots, m$) as input.

V:3 Suppose you are given the haplotype frequencies p_h, $h \in H_1$, and the error models τ_s, $s = 1, 2, \ldots, m$. For a random observed genotype $g \in \{0, 1, 2, \mathrm{N}\}^m$, write (in terms of the p_h and τ_s) the conditional probability that the underlying true genotype is (a) a homozygous deletion, (b) a hemizygous deletion, given the observed g. (c) How fast can you compute these probabilities given g as input?