# Notes on Schraiber et al., 2012 - *Genomic Tests on Variation in Inbreeding Among Individuals and Among Chromosomes*

Vince Buffalo

August 22, 2013

## 1 Levene's 1949 Model: The Probability of Heterozygotes with No Inbreeding

First, we consider equation (1) in Schraiber et al. (2012), the probability of seeing some number heterozygotes in a sample with no inbreeding derived from (Levene, 1949) (but I went back to Haldane (1954) and saw the same derivation there, earlier). The setup is that we have $n$ diploid individuals, giving us $n$ possible "bins" to put $2n$ balls into. If we only consider polymorphic sites (assuming biallelic), there are two alleles: ancestral and derived. For our $2n$ alleles in our sample, call $i$ the number of derived alleles. $i$ is bound between 1 and $2n-1$ (that is, $i \in \{1, 2, \ldots, 2n-1\}$) since if $i = 0$, we have no derived alleles and our site is not polymorphic, and if $i = 2n$, our site is entirely derived alleles, and again, our site is not polymorphic.

There are three additional constraints. First, if we define $h$ as the number of heterozygous individuals (that is, individuals with one of the $i$ derived alleles and one of the $2n-i$ ancestral alleles), then if $h$ is odd, $i$ must be odd, and second, if $h$ is even, then $i$ must be even. If this is unclear, consider the case where $h$ is even. If an additional derived allele is added, it either goes into a heterozygous individual, and makes it homozygous (so $h' = h - 1$, making $h$ odd) or it makes a homozygous individual heterozygous (so $h' = h + 1$, making $h$ odd). Finally, the last constraint is that $h \leq \min(i, 2n - i)$, which can be understood as the most heterozygotes you can have is the minimum number of derived or ancestral alleles. This is easy intuitively too: in the extreme case that if *every* minor allele ($i$ or $2n - i$) was in a heterozygous individual, then the most heterozygous genotypes possible is the number of minor alleles.

### 1.1 Combinatorics

Schraiber uses a classic colored ball and bin combinatorial approach equation borrowed on Levene (1949) and Haldane (1954) to look at the probability of heterozygotes given our seen

number of derived alleles, $i$, and our sample size, $n$. Combinatorial approaches are very powerful, so I will go into some depth in this derivation.

When deriving probabilities using combinatorial approaches, our goal is to enumerate all outcomes in the sample space to determine the denominator, and then derive an expression for the number of outcomes for a particular event as the numerator. So to derive the probability of $h$ heterozygotes given we see $i$ derived alleles in a sample of $n$, we first want to determine the denominator by asking how many total ways are there are to arrange $2n - i$ red balls (ancestral alleles) and $i$ white balls (derived alleles).

One way of thinking about this is to take all arrangements of $2n$ balls, treating each ball as distinguishable (there are $(2n)!$ of these). Since we don't care about the arrangement of indistinguishable white balls, we divide the number of arrangements of $i$ white balls ($i!$). We also don't care about the number of arrangements of $2n - i$ indistinguishable red balls, so we divide out $(2n - i)!$ red balls. This leaves us with:

$$\frac{(2n)!}{i!(2n-i)!} \tag{1}$$

arrangements of $i$ white balls and $2n - i$ red balls, the well-known binomial coefficient. In this case, our problem can be set up this way: given numbered (distinguishable) balls, how many ways are there to grab $i$ and paint them white? Note that certain arrangements of balls, i.e. for $n = 5$ individuals we will have arrangements that are lead to the same genotype: $(0,1)(1,0), (0,0)(1,1)(0,1)$ and $(1,0)(1,0), (0,0)(1,1)(0,1)$ are the same. We will account for this in a bit.

Now, let $n_2$, $n_1$, and $n_0$ be the number of bins with 2, 1, and 0 white balls. Immediately, notice that $n_1 = h$, the number of heterozygotes, and $n_0 + n_1 + n_2 = n$.

Since $i$ is the number of white balls (derived alleles), the number of $n_0$ bins (ancestral homozygous, bins with only red balls) is $n_0 = (2n - i - h)/2$, or intuitively, the number of ancestral (red) balls $(2n - i)$, minus the number in heterozygotes $(h)$, leaving the number of ancestral balls in homozygous bins. Since we want a count of the number of *bins*, and not *balls*, and these are by definition in homozygous state, we divide by 2.

Next, for $n_2$, the number of homozygous derived (bins with two white balls) we use the same logic. There are $i$ white balls, but we lose $h$ of these two homozygous bins. Thus, we are left with $(i - h)/2$ bins that are homozygous derived.

Next, $n_2$, $n_1$, and $n_0$ give us the counts for each of these events, as a function of $i$, $h$, and $n$. With these components, we can now get the number of distinguishable ways to balls to our three *classes*: $n_0$, $n_1$, and $n_2$. The number of ways of assigning $n$ objects to classes of sizes $n_0$, $n_1$, and $n_2$ is given by the multinomial coefficient:

$$\frac{n!}{n_2!n_1!n_0!} \tag{2}$$

However, note that this is off by a factor of $2^{n_1}$, because each of the $n_1$ heterozygotes can be arranged 2 ways, not just one. Thus we multiple by $2^{n_1}$, which now counts *all* the permutations of heterozygotes. This is quite important, as we are also counting all possible permutations of balls in heterozygotes in the binomial coefficient above too.

Finally, this leaves Equation 3, which is simply a rearranged **??** over Equation 1.

$$\mathrm{P}(h \mid i, n) = \frac{i!(2n-1)!}{(2n)!} \frac{2^h n!}{((i-h)/2)!h!((2n-i-h)/2)!} \tag{3}$$

## 2 Single Inbreeding Coefficient for a Sample

Next, Schraiber looks at a single inbreed coefficient for the sample, $F$. He uses $F$ as the probability of being identical by descent and $1-F$ as the complement of this. A probabilistic interpretation of $F$ requires $0 \leq F \leq 1$ (no excess of heterozygotes).

Schraiber assumes that the $j$ individuals at a polymorphic site that are i.b.d is binomially distributed with $p = F$. Thus:

$$\mathrm{P}(j \mid n', F) = \binom{n'}{j} F^j (1-F)^{n'-j} \tag{4}$$

Which is a standard binomial, with one minor caveat. Notice that we use $n'$ here instead of $n$. This is to handle a constraint: if $i$ (the number of derived alleles) is odd, then there has to be at least one heterozygous site. We define $n' = n$ if $i$ is even, $n = n-1$ if $i$ is odd.

Now, let $k$ be the number of individuals that are homozygous for the derived allele because they are i.b.d (given we know $j$). $k$ has a hypergeometric distribution (note that there was another typo in the original paper here):

$$\mathrm{P}(k \mid j, i', n') = \frac{\binom{i'/2}{k}\binom{n'-i'}{j-k}}{\binom{n'}{j}} \tag{5}$$

This is saying, in the context of a hypergeometric distribution, if I draw $j$ individuals (where $j$ is the number of i.b.d. individuals, either derived or ancestral alleles), what's the probability of seeing $k$ autozygous derived alleles given there are $i/2$ pairs to choose from in a population of size $n$ (note that the model assumes some model of reproduction such that autozygosity can't occur by drawing the same derived allele twice, hence we divide by two)? To me, it seemed strange that we worry about draws of size $j$, but our reasoning for this is is that this entire equation will become used as a conditional in a joint density, and summed across all possible $j$ since we are uncertain about it.

Given $j$ (the number of i.b.d. individuals) and $k$ (the number of derived i.b.d. alleles), the number of individuals heterozygous at a site has the distribution given by Equation 3 with $n = n - j$ and $i = i - 2k$ (the paper says $j = j - 2k$ but there's no $j$ in equation (1), so I believe this is another typo).

Finally, we want an expression for $P(h \mid i, n, F)$ (and we see $h$, $i$, and $n$ in our real data), so we take the marginal densities over the unknowns $j$ (number of i.b.d individuals) and $k$ (number of autozygous derived) in our joint likelihood built up from the conditionals:

$$\mathrm{P}(h \mid i, n, F) = \sum_{j=0}^{n'} \sum_{k=0}^{j} P(h, i-2k, n-j) P(j \mid F, n') P(k \mid j, i', n') \tag{6}$$

Thus, this serves as our likelihood model to estimate $F$ using MLE approaches.

### 2.1 $\hat{F}_i$: $F$ per a Certain Number of Derived Alleles, Across Sites

Equation 6 can then be used as a likelihood for each derived allele $i$, by "replacing $n$ by $n_i$, the observed number of sites at which there are $i$ derived alleles, and $h$ by $h_i$, the number of sites that are heterozygous." The natural way to think about this is that we are working across sites, rather than across individuals.

Why would this be interesting? For a sample $n$, $i$ implies a genotype frequency $(i/2n)$. For certain sites that have the same number of derived alleles $i$ (same frequeny), and we want to estimate $F$, as if $i$ is low (i.e. derived mutations are recent, think neutral coalecent) a higher $F$ may indicate selection. In fact, Schraiber et al. (2012) use LRT to determine if a $F$ per $i$ fits better than a universal $F$ and the find it does. However, as the paper shows, even allowing for $F$ for each $i$, the data are not consistent with inbreeding coefficients being the same for each individual. This is why they move on to estimate individual-specific inbreeding coefficient.

## 3 Individual Specific Inbreeding Coefficients

Rather than just estimate $F_i$, one could estimate $F_j$, the inbreeding coefficient per individual $j \in \{1, \ldots, n\}$ (note this a different $j$ than before!). Then, we can find the set $(F_1, \ldots, F_n)$ given $i$ and $(h_1, \ldots, h_n)$ where $h_j = (h_{j1}, h_{j2}, \ldots, h_{jL})$ in which $h_{jl} = 1$ if individual $j$ is heterozygous at site $l$ and 0 otherwise (note: don't we need $i_l$?).

The exact likelihood is hard, so we compute the *composite likliehood*. Composite likelihoods are a type of pseudo-likelihood because they don't fully specify the *real* likelihood (probabilistic model). Borrowing an example similar to one in Pawitan 2013, assume you have a linear spatial distribution of something and you observe some array of data, $y_1, y_2, \ldots, y_n$. The joint probability we reach for would be: $p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2 \mid y_1) \ldots p(y_n \mid y_{n-1}, \ldots)$. However, this imposes a right to left model of our spatial distribution, and seems awkward. It would be more natural to think of first-order neighbor model $p(y_k \mid y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_n) = p(y_k \mid y_{k-1}, y_{k+1})$. But, the product of all the conditional probabilities is not the same as the likelihood, but instead a composite likelihood. Pawitan gives a nice example for the simple case where observe $(y_1, y_2)$:

$$p(y_1 \mid y_2)p(y_2 \mid y_1)$$

This is not a likelihood; the true likelihood is:

$$p(y_1)p(y_2 \mid y_1)$$

The former case is the composite liklihood, as is similar to the one used in Schraiber et

al.:

$$\mathrm{P}(h_1, h_2, \ldots, h_n \mid F_1, F_2, \ldots, F_n, i) = \prod_{j=1}^{n} \prod_{l=1}^{L} \mathrm{P}(h_{jl} = 1 \mid F_j, i_l)^{h_{jl}} \left(1 - \mathrm{P}(h_{jl} = 1 \mid F_j, i_l)\right)^{1-h_{jl}} \tag{7}$$

where $L$ is the number of polymorphic sites (note I added an $l$ subscript, as this is at a particular site and I think this is a typo in the original). Also:

$$\mathrm{P}(h_{jl} = 1 \mid F_j, i) = (1 - F_j)\frac{i(2n - i)}{n(2n - 1)} \tag{8}$$

Note that for some individual $j$'s $F_j$, this probability of being heterozygous has the same $1 - F$ factor as the frequency of a heterozygote under Generalized Hardy-Weinberg ($(1 - F)2p(1 - p)$). So we might naturally ask, what is $(i(2n - i))/(n(2n - 1))$? Well, the frequency of derived alleles is $i/(2n - 1)$ and the frequency of ancestral is $(2n - i)/2n$. I believe the frequency of $i$ has the denominator $2n - 1$ because the condition mentioned earlier.

# References


JBS Haldane. An exact test for randomness of mating. *Journal of Genetics*, 52(3):631–635, 1954.

Howard Levene. On a matching problem arising in genetics. *The Annals of Mathematical Statistics*, 20(1):91–94, 1949.

Joshua G Schraiber, Stephannie Shih, and Montgomery Slatkin. Genomic tests of variation in inbreeding among individuals and among chromosomes. *Genetics*, 192(4):1477–1482, November 2012.