More notes on HWE

March 5, 2016

1 Preliminaries

Load utility R code; do setup:

```
source('../../../R/wlr.R') # load util code; path relative this folder or sibling in scripts/larrys

## [1] "SVN Id, I miss you.  $Id: wlr.R  2016-03-05 or later ruzzo $"

setup.my.wd('hwe')          # set working dir; UPDATE if this file moves, or if COPY/PASTE to new file
setup.my.knitr()
```

Some of the following should be copied into the methods supplement.

2 The Reference Genome is a "Hardy-Weinberg Haplotype"

It is useful to understand how often the reference sequence reflects rare alleles at polymorphic sites.

When a haploid "reference genome" is constructed from a single diploid individual, homozygous positions are, of course, recorded in the reference, while at heterozygous positions, one of the two alleles is selected to be the reference nucleotide essentially at random, e.g., based on which variant accumulated more reads in the sequencing run. Casting this as a more explicitly probabilistic process (and ignoring read errors and potentially biased coverage), to construct the reference, draw two samples from the HWE population at each site; if the two samples differ, choosing the one with more reads is equivalent to choosing the first draw, since they are equiprobable; if the two samples agree, it is again equivalent to choosing the first draw. So, the reference sequence is equivalent to drawing *one* sample from HWE at each position. I.e., the reference genome is what we might call a "Hardy-Weinberg haplotype"—exactly equivalent to a randomly selected haplotype drawn from the HWE population.

The above analysis would not apply to a "reference genome" constructed from DNA pooled from more than one individual. However, the CCMP 1335 reference sequence[1] was derived from an isogenic cell culture: the sequencing project isolated a single cell from the CCMP 1335 culture, then allowed it to divide to produce enough DNA for sequencing. Reproduction in culture is believed to have been exclusively mitotic, and so the reference construction model presented above is as appropriate as it would be for a more typical sequencing project based on a sample from a single multicellular organism. Unobserved sexual reproduction in culture might increase the variance in observed read counts at heterozygous sites, but with rare exceptions (e.g., homozygous lethality), should not alter the mean 50-50 mixture of the two alleles.

Keep the following algebraic derivation of the same result "just in case," but commented out in methods, since I think the version above is more readable and no less rigorous.

... Thus (ignoring read errors and assuming unbiased coverage), the probability that the minor allele is enshrined as the reference at that position is

$$q^2 + \frac{1}{2}(2pq) = q(q+p) = q.$$

(In the founder cell from which the sequencing culture was derived, this position was homozygous for the minor allele or it was heterozygous and the minor allele happened to have more reads than the equally-likely alternative.)

While irrelevant for our immediate purposes, this model extends to tri- and quad-allelic positions, too. If the four allele frequencies are p, q, r, s, then the probability, that, say, the q allele appears in the reference sequence is:

$$q^2 + \frac{1}{2}(2pq + 2rq + 2sq) = q(q+p+r+s) = q,$$

and similarly for the others by symmetry.

3 Heterozygous Sites Outnumber Homozygous Non-Reference Sites 2 to 1

In both H-isolates, we see a roughly 2:1 ratio between numbers of heterozygous (het) versus homozygous non-reference (homnr) positions. In principle, that ratio depends on both the distribution of allele frequencies in the sampled populations and on the reference genome, so it is natural to ask whether the 2:1 ratio is "special" in some way. For example, that is the ratio HWE would predict if all non-reference alleles had 0.5 frequency. In contrast, homozygous non-reference positions would be quite rare if the reference genome exclusively recorded major alleles (the allele with highest frequency at a given site, typically ≥ 0.5), but—as predicted by neutral theory—most variants were rare (frequency $\ll 0.5$).

We show below that the 2:1 ratio is expected, *independent* of allele frequencies, given the way the reference genome was constructed. Specifically, since the reference sequence is effectively a random haploid genome, it will record the major allele at most polymorphic positions, but will sometimes record a rare allele (equally rarely). Resequenced individuals will only rarely appear to be homozygous non-reference relative to the former class of reference positions, but will commonly be homozygous non-reference with respect to the later class. These effects counterbalance to yield the observed 2:1 ratio—exactly 2:1 when only biallelic positions are considered, and (slightly) greater than 2:1 when (typically rarer) multi-allelic positions are considered.

E.g., suppose there are 100 biallelic loci, each with a 0.1 minor allele frequency in a population in HWE. In expectation, at each locus, $0.9^2 = 81\%$ of individuals are expected to be homozygous for the major allele, $2 \cdot 0.9 \cdot 0.1 = 18\%$ heterozygous, and only $0.1^2 = 1\%$ homozygous for the minor allele. In one individual, the number of homozygous non-reference positions will depend on the reference, of course. If the reference reflects the major allele at each locus, then only 1% of these loci will be homozygous non-reference (vs 18% het, an 18:1 ratio), but if the reference records the minor allele at 10% of loci (as expected in a random haplotype from this population), then the number of homozygous non-reference loci is expected to be $0.81 * 10 + 0.01 * 90 = 9.0$, so the het:homnr ratio is $2{:}1$.

More generally, consider a diploid population in Hardy-Weinberg equilibrium. Focus on a specific biallelic position having minor allele frequency $0 \leq q \leq 1/2$ and $p = 1 - q$. When re-sequencing another individual drawn from the same population, determining whether this position is heterozygous versus homozygous non-reference can be visualized as drawing three independent samples from the HWE population—the first draw determines the reference haplotype, and the other two define the genotype of the new individual. If all three are the same, that site is homozygous for the reference allele. If the three are not all the same, then only three distinct possibilities are relevant: Letting "a" denote the allele that was observed only once, and "b" the allele seen twice, the three draws yield abb, bab, or bba. Since the first letter defines the reference, outcome abb is the homozygous non-reference case, and the other two outcomes are heterozygous. These three outcomes are equally likely (with all three probabilities equal to p^2q or or all equal to q^2p, depending on whether "b" is the major or minor allele, resp.), so the heterozygous to homozygous non-reference ratio is 2:1. Inclusion of (a small number of) 3- and 4-state positions in the population will raise the proportion of heterozygous positions in a resequenced individual (by a similarly small amount).

Thanks to Joe Felsenstein for suggesting the above proof. My older version follows. Again, let's keep it, but I think the above is simpler.

Note: 3-/4-state sites add other cases to the "1 a, 2 b" draws where both are minor, but they maintain the 2:1 ratio. They also add an "abc" draw as a possibility, necessarily a het site, with a probability that includes the product of 2 or 3 minor alleles, hence will increase the 2:1 ratio by, typically, a small amount.

For the general case, consider a diploid population in Hardy-Weinberg equilibrium with n bi-allelic positions, the i^{th} having minor allele frequency $0 \leq q_i \leq 1/2$ and $p_i = 1 - q_i, i = 1, \ldots, n$. In a random diploid individual, the probability that position i is heterozygous is, of course, $2p_iq_i$ (independent of the reference sequence), and so the expected number of such positions is

$$N_{het} = 2\sum_{i=1}^{n} p_i q_i.$$

Again in a random diploid individual, we find the probability that position i is homozygous for the non-reference allele (with respect to the fixed reference derived as above) as follows. Let Z_i be the event that position i is homozygous non-reference, let $M_i(M_i^C)$ be the event that the reference genome shows the minor (major, resp.) allele at position i, and define $p(M_i) = r_i$. Then

$$\begin{aligned} P(Z_i) &= P(Z_i \mid M_i)P(M_i) + P(Z_i \mid M_i^C)P(M_i^C) = p_i^2 r_i + q_i^2(1 - r_i) \\ &= p_i^2 q_i + q_i^2 p_i = p_i q_i (p_i + q_i) = p_i q_i, \end{aligned} \tag{1}$$

where the last line follows from assuming $r_i = q_i$. Given this, the expected number of homozygous non-reference positions is

$$N_{homnr} = \sum_{i=1}^{n} p_i q_i,$$

exactly half of N_{het}.

4 Our CCMP Re-Sequencing Cultures Were Isogenic

As noted, the CCMP 1335 reference sequence[1] was derived from an isogenic cell culture—it was grown from a single isolated cell. In contrast, each of our "re-sequencing cultures" was grown from an estimated 5–10 cells isolated by flow cytometry (Methods **??**) from the relevant CCMP culture. Genetic diversity in the re-sequencing culture could potentially mask genomic signals of interest. E.g., a site that is homozygous non-reference in some but not all cells might be indistinguishable from a uniformly heterozygous site. Consequently, we considered whether each CCMP culture was isogenic.

Perhaps the simplest way to establish a single-species culture from an asexual or homothallic unicellular organism is to establish it from a single isolated cell. (CCMP 1335 is reputed[1] to have been established in this way.) Additionally, in a culture established decades ago from a few cells, it is plausible that one genotype may have grown to dominance. Thus, it is very plausible that all seven CCMP isolates are isogenic, but to be conservative, we looked to our data for direct confirmation.

Suppose one of the CCMP cultures had several, say, f, founder cells, independently drawn from the HWE population. Extending the analysis from Section 2, at a biallelic position having minor allele frequency $q = 1 - p \leq p$, the probability that the $2f$ chromosomes of the f founder cells hold exactly $j = 0, \ldots, 2f$ copies of the non-reference allele is:

$$B(j, f) = p\binom{2f}{j}p^{2f-j}q^{j} + q\binom{2f}{j}p^{j}q^{2f-j}.$$

This is the probability of exactly j "successes" when performing $2f$ trials in a weighted *mixture* of two binomial distributions, one with weight p and success probability q, and the other with weight q and success probability p. Graphically, the probability mass function for this system will place all mass at the discrete points $j/(2f), j = 0, \ldots, 2f$. According to neutral theory, we should expect many positions to exhibit small minor allele frequencies q. Intuitively, when q is sufficiently small, the most likely scenario is that the major allele is the reference nucleotide. In this case, the most likely number of copies of the nonreference allele captured among the f founders is $j = 0$, with $j = 1$ being next most likely, and $j = 2, 3, \ldots$ being increasingly unlikely (the first term in the formula above). However, if the minor allele is the reference nucleotide (which happens with probability q), then the most likely outcome is that $j = 2f$ nonreference alleles (i.e., only major alleles) are seen, with $j = 2f-1, 2f-2, \ldots$ being increasingly unlikely (the second term in the formula). These two series cross at $j \approx f$, and their sum in minimized when $j = f + 1$, with the net result that $B(j, f)$, as a function of j, is convex ("U-shaped"), with most probability mass *away* from the middle (except when $f = 1$, when the $j = 1$ case *is* the middle). The top row of Figure H2-1 gives three examples of this, showing the probability mass function $B(j, f)$ versus non-reference fraction $j/(2f)$ corresponding to a minor allele frequency of $q = 0.04$ and various numbers of founders f.

Above intuition is probably adequate, but here's a proof. Flip convex switch to false to hide in supplement but keep the text, in case...

We formally verify for each $f \geq 1$ and $q \leq 1/(4f + 1)$ that $B(j, f)$ is a convex function of j as follows. Let $c_i = \binom{2f}{i}$, let $r = q/p$ and note that $r \leq \frac{1/(4f+1)}{(1-1/(4f+1))} = \frac{1}{4f}$. Then

$$B(j, f) = p\binom{2f}{j}p^{2f-j}q^{j} + q\binom{2f}{j}p^{j}q^{2f-j} = p^{2f+1}c_j r^j \left[1 + r^{2f-2j+1}\right],$$

and, for $0 < j < 2f$,

$$\begin{aligned} B(j-1, f) - 2B(j, f) + B(j+1, f) &= p^{2f+1}c_j r^j \left[\frac{c_{j-1}}{c_j}r^{-1}(1 + r^{2f-2j+3}) - 2(1 + r^{2f-2j+1}) + \frac{c_{j+1}}{c_j}r^{1}(1 + r^{2f-2j-1})\right] \\ &= p^{2f+1}c_j r^j \left[\left(\frac{c_{j-1}}{c_j}r^{-1} - 2 + \frac{c_{j+1}}{c_j}r\right) + \left(\frac{c_{j-1}}{c_j}r - 2 + \frac{c_{j+1}}{c_j}r^{-1}\right) r^{2f-2j+1}\right] \\ &> 0, \end{aligned}$$

where the final inequality holds since

$$\min_{0<j<2f}\left(\frac{c_{j-1}}{c_j}, \frac{c_{j+1}}{c_j}\right) r^{-1} = \frac{r^{-1}}{2f} \geq 2.$$

Thus, $B(j, f)$ is convex. Since sums of convex functions are themselves convex, any *mixture* of sites with minor allele frequencies below $1/(4f+1)$ will also have its minimum near the middle, except when $f = 1$.

The "R" distributions (analogous to those histogrammed, e.g., in Fig. **??** in the main paper) expected from the model outlined here, would reflect (a) a theoretical distribution similar to the dots shown in Fig H2-1 for alleles captured in the founder population, but (b) summed over many positions with varying minor allele frequencies, and (c) "blurred" by binomial sampling as the sequencer accumulates reads from both alleles at sites not rendered completely mono-allelic by (a) or drift. (We ignore in-culture selective pressure.) The gray bar graphs in Fig H2-1 reflect a simple

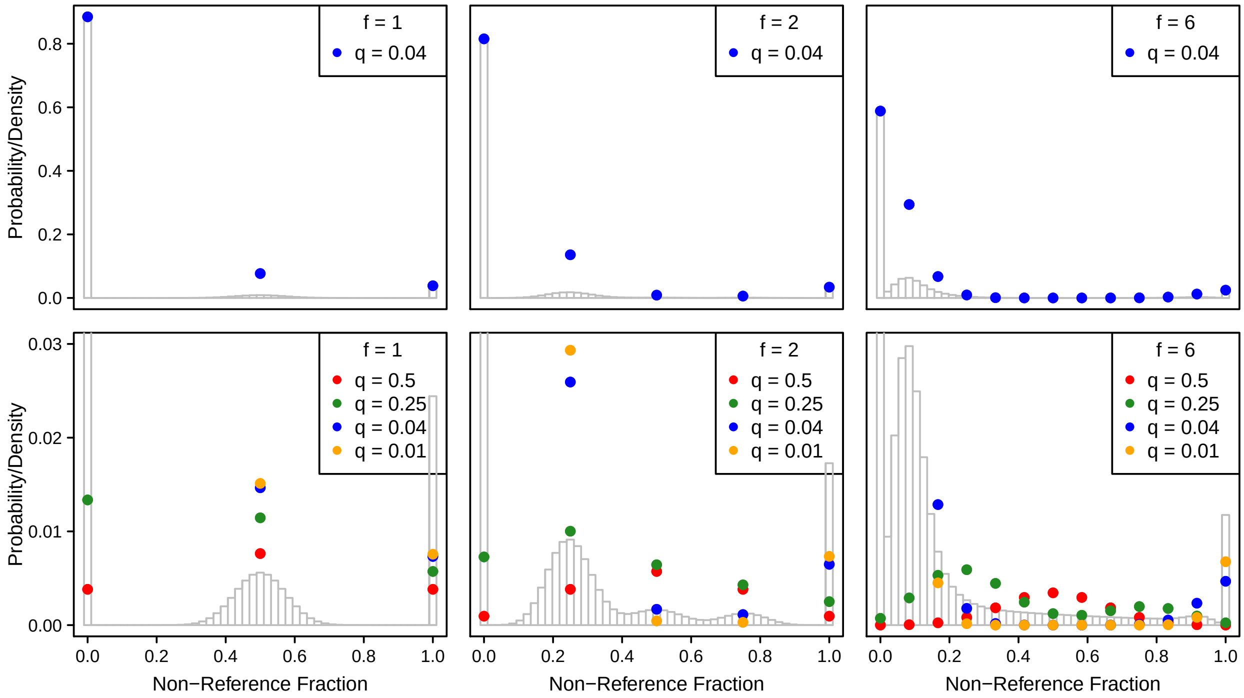

Figure H2-1: Non-Reference Modeling: top row shows simulation results assuming $f = 1$ (left), 2 (middle) or 6 (right) founder cells are sampled from a HWE population. At sites having minor allele frequency $q = 0.04$, the probability $B(j, f)$ (y-axis) that these f cells hold a specified fraction $j/(2f), j = 0, \ldots, 2f$ (x-axis) of the non-reference allele at that site is plotted (blue dots). The superimposed gray bar graphs show the effect of sampling reads during sequencing—e.g., in the upper left panel, 8% of biallelic sites having this minor allele frequency would be expected to be heterozygous (blue dot at $x = 0.5$), but binomial sampling of reads for both alleles will spread their apparent nonreference proportions as shown ("hump" in gray bar graph centered under 0.5). The theoretical model used assumes coverage 48 at all sites, with no errors or bias in sequencing or mapping. Bottom row: analogous simulations, assuming a weighted mixture of minor allele frequencies q (see legend), with weights inversely proportional to q. (Note the change in y scale in the lower row; the leftmost points and gray bars are clipped to expose more detail at small y values.)

simulation of this. Note that the "U" shaped scenario does not apply individually to larger q values (red and green points in the bottom panels of Figure H2-1), but does apply collectively to a mixture including many sites with small q, even when some sites with larger q are present, as seen in Figure H2-1. In aggregate, these effects add variability to the data, but do not alter the main features of our model, namely, presence of a fair number of positions with apparent non-reference frequency near 1.0, and, with one key exception, absence of a peak in the R-distribution near 0.5. The key exception is when $f = 1$: establishment of the culture from a single individual means that all heterozygous sites in that individual are retained at a 50-50 allele frequency in the descendant population (as fixed heterozygous sites in all offspring if only mitotic division happens in culture, and maintained on average if there is unobserved sexual reproduction in culture). In this case, the CCMP culture is isogenic. Two other exceptions that allow a peak near 0.5 are:

- When the estimated $k = 5$–10 cells isolated to start the sequencing culture are isogenic. This would happen if $k = 1$, or if $k \geq 2$, but all happened to be mitotic descendants of only one of the f CCMP founders. Neither event seems very likely, but neither is problematic for our analysis—this case is equivalent to the $f = 1$ case, in that the *sequenced* portion of the CCMP culture is isogenic, which is all that is needed for our subsequent analysis.
- When most variants in the wild population are assumed to have allele frequencies near 0.5. However, this strongly violates neutral theory, which predicts that rare variants will greatly outnumber common variants. In

the later situation, the dominant peaks will be at $0.0, 1.0$ and at $1/(2f)$ (rare alleles are more likely to be seen in one of the $2f$ chromosomes than in several), as seen in Fig H2-1.

Thus, the presence of the peak near 0.5 in the "R" histograms (e.g., Fig. **??**) for 6 isolates demonstrates that each re-sequencing culture was isogenic (having been established from, or eventually dominated by the descendants of, a single isolated cell). An important consequence is that the exact number of cells used to establish the re-sequencing culture (the "$k = 5$–10" estimate) is not relevant for our subsequent analysis—all k are genetically identical. (Again, interpretation of CCMP 1014 is hampered by lower data quality.)

5 The H-clade retains sexual reproduction but the L-clade is asexual

Based on Section 4 each of our CCMP strains had a single founder cell. Assuming each was drawn from a common population in HWE, the single founder cell of each (non-1335) strain would have had a heterozygous to homozygous non-reference ratio of at most 2:1 with respect to the CCMP 1335 reference, as shown in Section 3. Homozygous non-reference positions in the founder will appear exclusively non-reference in its descendants, no matter how many were included in our re-sequencing culture, even if recombination were occurring in culture, for the simple reason that no alternative nucleotide exists at that position in any cell in the entire culture. In consequence, the 2:1 ratio will be recapitulated when re-sequenced. Thus, the dearth of homozygous non-reference positions in the L-isolates in comparison to the ≈90k predicted by this analysis (1/2 of the ≈180k observed heterozygous positions reported in Table **??**) argues strongly against sex in the wild for all five.

In contrast, the ≈2:1 heterozygous to homozygous non-reference ratio observed in the H-isolates (with respect to the CCMP 1335-based reference) is consistent with HWE, and thus sexual reproduction in those populations, based on the assumption that the allele frequencies in the H-isolates have not drastically changed in the time since the L-clade founder emerged from the population that was the common ancestor to all isolates.

6 Sex in Culture?

At various points we have said things like: "argues against sex in culture for CCMP 1335 (at least during the NNN year interval between the original sequencing[1] and our re-sequencing)." I think this is in the current draft of the main paper too. At this point I suggest we *delete* it from the main paper, at least. I believe it is true, but will be unsurprising to most, and although it would be nice to present some concrete evidence bearing on the question, I think the evidence we have may raise debate that will detract from the main story.

Specifically, I think our evidence is this: the arguments above show that all CCMP cultures are monoclonal, and the 5 L-cultures started with essentially the same clone (no sex in the wild). From that point, sex in culture will not create new alleles, but will mix them. *IF* we had re-sequenced from single cells, homozygous non-reference counts would clearly show sex/nosex. But since we sequenced $k = 5$–10 cells, if we assume sex in culture, we are back to the scenario where mixing may hide the homnr positions. Conveniently, the original simulations I did for a different purpose fit this scenario very well: assume HWE populations with all allele frequencies = 0.5 (or zero). The conclusions I advocated before are still valid: small k leaves a footprint clearly distinct from the observed data; larger k, say 10, gives a bell-shaped curve (the "orange" curves in the simulations, I think) like what we see in the L-clade, with few apparent homnr positions, but a distinctly wider bell than observed. The latter situation is very implausible, IMHO.

But rather than invite debate on a less interesting topic, I propose we just remain silent on the question of sex in culture. Another option would be a short section in the supplement outlining this argument (e.g., an elaboration of the previous paragraph, with a few of the simulations shown as a figure). That would also be OK with me.

References

[1] Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* **306**, 79–86 (2004).