A Test Document

Paul M. Magwene

August 9, 2012

1 ‘The Crisis’ by Thomas Paine

These are the times that try men’s souls. The summer soldier and the sunshine patriot will, in this crisis, shrink from the service of their country; but he that stands by it now, deserves the love and thanks of man and woman. Tyranny, like hell, is not easily conquered; yet we have this consolation with us, that the harder the conflict, the more glorious the triumph. What we obtain too cheap, we esteem too lightly: it is dearness only that gives every thing its value....

2 Math

Here’s a simple equation: $\cos^2\varphi + \sin^2\varphi = -e^{i\pi}$ and here is a an equation with a sum: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$. We follow these up with a ‘display equation’ as shown below:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Let’s do some typesetting of numbers, for example: 11, 12 and 17. Now, here’s how those look in math mode: 11 and 17. Are they any different? If so, which looks better?

3 Programming

And here is some verbatim:

```
import math
x = math.cos(2*math.pi)
l = 1.1 * tan(x)
print "Can you tell the difference between O and 0?"
print "How about 1 and l?"
```

4 Author Summary

Quantitative or complex phenotypes are traits that are under the control of multiple genes and environmental factors. Identifying the parts of the genome that contribute to variation in complex traits (Quantitative Trait Loci or QTLs), and ultimately the genes and alleles that are mechanistically responsible for trait variation, is a primary challenge in animal and plant breeding, population studies of human health and disease, and evolutionary genetics. In this study we describe an analytical framework that allows investigators to marry a QTL mapping approach called "bulk segregant analysis" (BSA) with high-throughput genome sequencing methodologies in order to map traits quickly, efficiently, and in a relatively inexpensive manner. This framework provides a statistical basis for analyzing BSA experiments that use next-generation sequencing and will help to accelerate the identification of QTLs in both model and non-model organisms.

5 Discussion

The use of a test based on the G-statistic provides a straightforward framework for analyzing BSA-sequencing data. The G-statistic has several advantages over the use of allele frequency differences as the basis for QTL estimation. For example, as shown in the supporting information (Text S2), G is expected to decrease much more rapidly around the causal site than bias in allele frequencies, implying narrower intervals of support around QTLs. Also in contrast to statistics based on the divergence of allele frequencies, G takes into account the strength of evidence related to sample size. This feature of the G-statistic can also potentially complicate analyses, as variance in read depth contributes to variance in G over relatively small spatial scales. However, as we show above, weighted averaging of G effectively smooths out 'high frequency' noise associated with sequencing variation.

5.1 Bulk Size and Sequencing Considerations

Our simulations suggest that for the experimental design considered here using bulk sizes as large as 15-20% of the phenotyped segregant population increases power to detect causal QTLs despite the fact that this means relatively smaller allele frequency differences between bulks. This is due to tradeoffs between bulk-size, selection intensity, and the variance of allele frequencies under the hierarchical sampling. Consider, for example, a single locus with alleles A_0 and A_1, where the effect of A_1 is additive and the two homozygotes differ by $2a$ units on average. Assuming no segregation distortion, and an F_2 population generated from inbred lines, the change in the allele frequency of A_1 in the high bulk after truncation selection is approximately $\Delta q = \frac{1}{8} i \frac{2a}{\sigma_p}$ where i is the intensity of selection, and $\frac{2a}{\sigma_p}$ is the 'standardized effect of the locus' (these quantities can be related to the selection coefficient, s, by $s \approx i \frac{2a}{\sigma_p}$). Given truncation selection on a normal distribution, the intensity of selection is given by $i = z/p$ where p is the proportion of selected individuals and z is the probability density function at the truncation point. Since the intensity of selection increases at a rate much less than $1/p$ (e.g. see Falconer and Mackay 1996, Fig. 11.3), an n-fold decrease in p results in a much less than n-fold change in the intensity of selection. For example, let $\frac{2a}{\sigma_p} = 0.2$ and consider truncation on the upper 20%, 10%, and 1%, of the phenotypic distribution. The increase in the frequency of A_1 in the high bulk given these truncation points is approximately 3.5%, 4.4%, and 6.7% respectively (translating to allele frequency differences of 7%, 8.8%, and 13.4% in the two-bulk case). On the other hand, the variance of the realized frequencies of the alleles in each bulk is inversely proportional to bulk size. Thus, a twenty-fold decrease in bulk size translates to less than a two-fold increase in allele frequency divergence, but a twenty-fold increase in the variance of allele frequencies. As long as average coverage, C, is moderate to large, the benefit of increasing n_s offsets the relatively smaller penalty resulting from a decrease in selection intensity. However, there is little benefit to increasing sequencing coverage beyond the size of the bulks.

Sequencing can introduce complications such as biases toward particular nucleotide calls; however in general this should effect both segregant bulks in the same direction. Due to the averaging affect of G', unless such biased sites are common over very large map distances they are unlikely to have substantial affects on results derived under our proposed framework. Similarly, a low percentage of mismapped reads or miscalled SNP calling are unlikely to be problematic for our framework, again because of the averaging affect of G'. However caution should be exercised in genomic regions that are particularly problematic in this regard, such as repeat rich regions.

5.2 Other Experimental Designs

In this paper we have focused on QTL mapping with an F_2 experimental design, but clearly our framework can be extended to other designs. Common alternatives include mapping populations produced by imposing one or more generations of inbreeding on an F_2, such as Recombinant Inbred Lines (RILs). The increased homozygosity of such populations should also be taken into consideration, as it increases the expected change in allele frequency due to selection but it also decreases the number of independent chromosomes that are sampled for a given number of selected individuals. Chromosomes in such RILs experience as much as twice the number of crossovers as do F_2 populations so the physical size of the smoothing window W should be reduced to take this reduced linkage disequilibrium into account. Even greater reductions of linkage disequilibrium can be accomplished by an alternative design that imposes additional generations of random mating, rather than inbreeding, on an F_2, resulting in more precise localization of QTLs. Additional generations of outcrossing (beyond the F_2) will likely magnify deviations of the null allele frequency from 0.5 owing to segregation distortion and/or inadvertent selection. This can be accommodated by application of formulas in Text S1 with q estimated from all sites within a genomic window.

Other experimental designs, such as backcrosses, will not have allele frequencies of 0.5. For these situations the null expected distributions of G and G' can be approximated using the equations presented in Text S1, although in this case it will be necessary to know the parental origin of the SNP alleles. Similarly, since G can be generalized to an arbitrary number of classes, one-tailed scenarios involving comparison to either a theoeretical population or a random sampling of segregants can be addressed in this framework.