a eZ idea of inversions, part II

Jeff Ross-Ibarra

July 13, 2016

In my previous post I explained my idea for why inversion polymorphisms may be common in complex plant genomes like maize. Here I'll describe what we're doing to start looking for them.

Inversions suppress recombination. The resulting lack of exchange means we expect divergence to build up between normal and inverted haplotypes, and such effects should be reflected in population genetic variability (e.g Fig. 1).

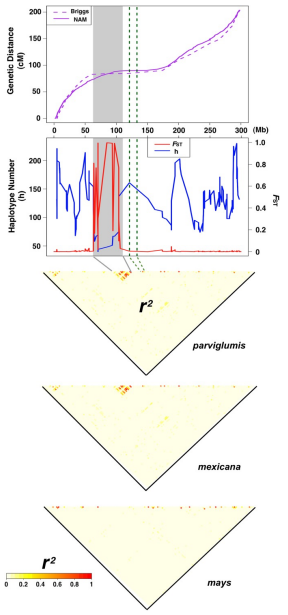

Figure 1: The impact of an inversion on diversity. Figure from Fang et al. 2012 showing reduced haplotype diversity, elevated F_{ST} between haplotypes, increased LD, and decreased rates of crossover inside a large inversion on maize chromosome 1.

I figured these effects should be reflected in a principle component analysis of diversity data from a sample segregating for an inversion. Indeed, when I look at GBS data from several thousand maize landraces from the SeeDs of Discovery project, I find my old friend Inv4m is readily identifiable (Fig. 2). I perform PCA on a sliding window along the chromosome, and in windows overlapping the

inversion, the first principle component cleanly divides all individuals into three clusters representing the three genotypes (two homozygotes and a heterozygote) for the inversion.

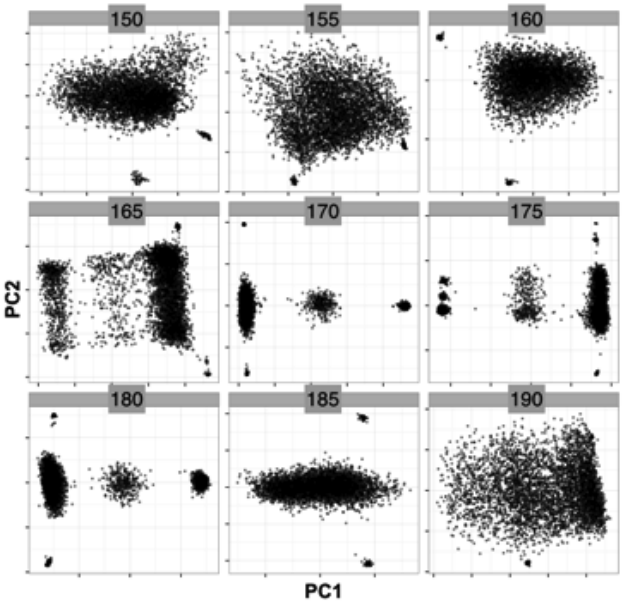

Figure 2: The first two principal components of a PCA on individuals from the maize SeeDs of diversity data, plotted in 5Mb windows for part of Chr. 4. You can see a high-res version at the 1Mb scale here.

So this is cool, as it hints at the possibility we may be able to use PCA to do a genome scan for putative inversions. I later discovered I was not alone in thinking about this possibility, but that Ma and Amos had suggested a similar approach years ago in humans. The sticking point, however, was how to identify windows with putative inversions. Our first realization was that scatterplots, while visually appealing, were not very useful for classification, and that most of the time putative inversions were represented as clusters on only one of the PC's (Fig. 3). This reduces the problem to one of identifying clusters of values in a single vector of data. I had recently learned to use Dirichlet Processes to cluster transposable elements based on their length, and this seemed a similar problem.

The DPP seems to work, and while we're still tweaking priors and testing results, it seems that using the posterior probability that there are three clusters,

$$\frac{P(k=3|\Theta)}{1-P(k=3|\Theta)} \div \frac{P(k=3|\Psi)}{1-P(k=3|\Psi)} \tag{1}$$

where Θ is the data and Ψ the prior, is a relatively simple means of identifying candidate regions that could be inversions. Of course even if these look real, we'll also need to convince ourselves these regions are real, and not population structure, a large indel, or some other feature of the data. But even a cursory glance at the data finds lots of convincing-looking regions (see Fig. 3A) and I suspect we will find a lot of new stuff!

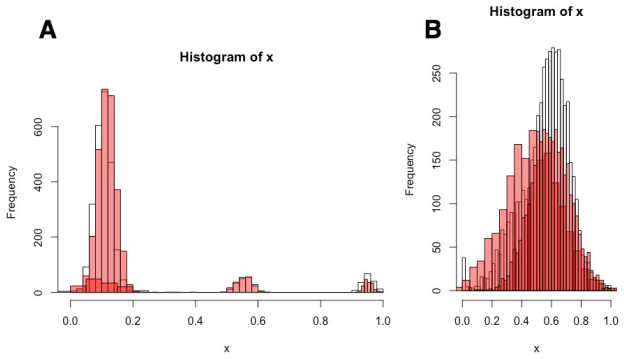

Figure 3: Histogram of scaled [0,1] eigenvalues along an individual principal component for a particular region on chromosome 10 in the SeeDs data. The region shown in (A) appears to be consistent with an inversion polymorphism forming 3 clusters representing the 3 genotypes at the inversion, while (B) appears a normal region of the genome lacking any evidence of an inversion.