Review

October 15th, 2014

General Comments

Pan *et al.* constructed 12 segregating populations, analyzed the genetic recombination landscape of these populations, and correlated the recombination hotspots with various genomic features. The authors present nice data sets that could be useful for future genetic studies. However, I am not convinced by the interpretation of the data, especially the methods used for hotspots and genomic features studies. To make the work acceptable, more work need to be done to either improve/clarify their methods and distinguish their results from those of Bauer *et al.*[1]. Below I include comments on specific aspects of the paper and hope they may be useful in revising the manuscript. There are a number of places with typo (e.g. "non-colineraity" should be "non-colinearity" on page 9) or grammatical error (e.g. "averaged 1879.3 cM" might be better to replaced by "an average of 1,879.3 cM"), careful revision of the text with this in mind would improve legibility.

[1] Bauer E, *et al*: Intraspecific variation of recombination rate in maize. Genome Biol 2013, 14:R103.

Results

- The number of recombination bins and the length of genetic maps were calculated using the same data to measure recombination events in two ways, it is not suprising to me that the two results should be correlated. I do not think it is necessary to present this result as a conclusion unless you have a point to make.
- I have concerns about the statement that "thereby improving the reference genome". Because B73 is not a founder line for most of the 12 populations, given the large number of genomic variations of maize[2], is it possible that the non-colinearity regions you observed are population specific inversions but not reference genome errors? Or the non-colinearity could be caused by genotyping errors or mis-placement of markers on the genetic map. Your have to rule out these possibilities.
- You should backup this statement "As expected, the longer the chomosome is the more recombination events occur", it is beyond my expectation.
- My major concern about this paper is that it may be not statistically legitimate to compare genomic features of hotspot regions

[2] Springer NM, *et al*: Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 2009, 5:e1000734.

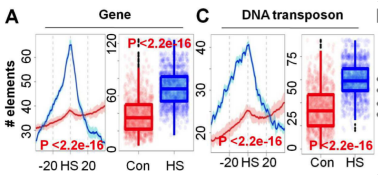

Figure 1: Snapshot of Figure 2 from the paper

with randomly selected genomic regions given SNPs on the SNP50 array may enriched in genic regions. And most of the *P* values lower than 2.2e-16 also seem worrisome (see right figure). Instead of choosing random genomic regions, non-hotspot regions of same size and SNP density were suggested to be selected for testing as described in Myers *et al.*[3]. In addition, SNP array normally suffers from some degree of ascertainment bias, it is not clear how much it will affect this analysis, but address this concern here or in discussion seems warranted.

[3] Myers S, *et al*: A fine-scale map of recombination rates and hotspots across the human genome. Science 2005, 310:321-324.

- You proposed a hypothesis about the function of genes[4] with intragenic recombination and conducted GO term enrichment to test it. However, in the conclusion, you neither rejected nor accepted your hypothesis, instead you proposed another possibility. I think there are two much speculations in this part, it might belongs to discussion.

[4] "It is possible that many of the genes with intragenic recombination belong to non-functional genes or pseduo-genes ..."

- The relationship of hotspots and gene expression should also using regions of same size and similar SNP density as control rather than randomly selected regions.
- A citation should be given for the `"genome-wide significance level"` for determining the threshold of GWAS. Or why not just use well accepted `FDR` or `bonferroni` method to control multiple test problem? And the population structure seems not well controlled in your GWAS (see right)?
- In the text, you should check whether **Fig. 4A** pointed to the right figure.
- In the last part of result section, the header is "intragenic recombination is significantly associated with gene expression and phenotypic variation in maize". However, I could not find any evidence in that section to support this statement, except some case studies.

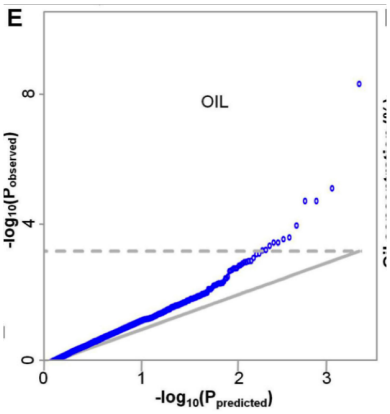

Figure 2: Snapshot of Figure 4E from the paper

Discussion

- You may want to talk more about how to use the 65 recombination hotspots in marker assisted selection.
- I am confused about this sentence[5], does the beginning and end be part of the genomic region *per se*.

[5] "Recombination is more likely to occur at the beginning and end of the genomic elements and not the genomic element regions `per se`, ..."

Methods

- To demonstrate the in-house Perl scripts works the same as (or better than) the estabilished methods, direct comparing their results with a test dataset would do the work. Technical details

(e.g. command `"flips"`) may belong to the software manual (or README file) on the script sharing website.

- The permutation procedure of hotspot identification need to be clarified, e.g., which value was permutated (or random shuffled), what test statistic was used and how to derive the threshold. Or citation.
- The terms in the formula (see right) should be defined.

$$4N_e r / kb$$

Tables and Figures

- Figure 3A: legend of `"# Genes with >= 1 SNP"` should be `"# Genes with <= 1 SNP"`?

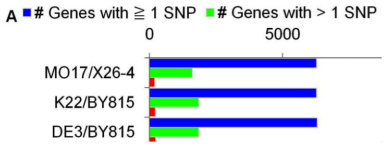

Figure 3: Snapshot part of Figure 3A

- Figure S2: For a better comparison of recombination events in 12 populations, x-axis and band width of the histograms in **Figure S2** should keep the same.
- Figure S4: 1Mb sliding bin might be too large for a genome wide distribution of DNA transposons.
- Figure S8: it seems to me the negative correlation was driven by only two dots.

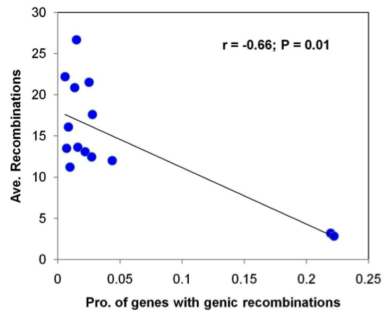

Figure 4: Snapshot of Figure S8

- Figure S11: From these qq-plots, some of the population structures were not completely controlled for the GWAS.