Copy-number variants in a wild population of maize: Paper outline

Simon Renny-Byfield and Jeffrey Ross-Ibarra

2015-03-31

Introduction

Analysis of resequencing data has provided an in-depth look at species-wide variability in maize and teosinte, but we still have little understanding of how evolutionary processes shape diversity within individual populations. In particular, previous efforts have shown that the vast majority of the maize genome is affected by presence-absence or copy number variation, but we do not yet understand how this genomic flux impacts diversity. Here, we investigate within-population processes and the role of presence-absence variation by resequencing the genomes of 20 teosinte from a single natural population near Palmar Chico, Mexico. We scan the genome for evidence of recent positive selection, and identify thousands of presence-absence variants across euchromatic portions of the genome. We show that patterns of diversity in these regions deviate meaningfully from standard population genetic expectations, and argue that population genetic analysis in complex plant genomes must take into account the effect of such polymorphisms on genome-wide patterns of diversity.

In addition we will examine patter of of nucleotide diversity, Tajima's D, patterns of selection and the prevalence of CNV calls in the two sub-genomes of maize.

The role of the paper is to address several unanswered but interesting questions regarding CNV in natural populations:

1. How prevalent are CNVs in a single wild population of maize.
2. At what frequency are these CNVs segregating.
3. Are CNVs of any "selective significance"

and somewhat separately..

4. Are CNVs more prevalent in maize1 or maize2?
5. Are the patterns of diversity over CNVs variable between the two sub-genomes.

Below I will give a brief outline of the paper, in terms of methods, results so far and future directions/analyses to do.

Methods

We will use two main methods to examine CNV in the Palmar Chico pop:

1. A custom CNV calling pipeline, using the distribution of GC corrected coverage (normalized to the reference B73) across all genes on a sample by sample basis.
2. A publicly available program called cn.mops is used to estimate copy-number over genomic regions looking at sample wide coverage (this cannot use the reference B73 as a standard).

In principle both of these analyses can be presented together, hopefully each supporting the conclusions of the other. In the following section I will describe the two methods in a little more detail.

1. Custom CNV calling

The first method of calling CNVs is one designed by myself. It is relatively simple compared to some of the advanced methods available, but at the time we pursued this because we wanted to totally understand what was going on. As such the paper will require detailed description of exactly what we did to achieve the copy-number estimation.I provide here an outline of the processes with a few key figures, and the precise code and detailed description can be found at in this GitHub repo.

1.1 Coverage over genes

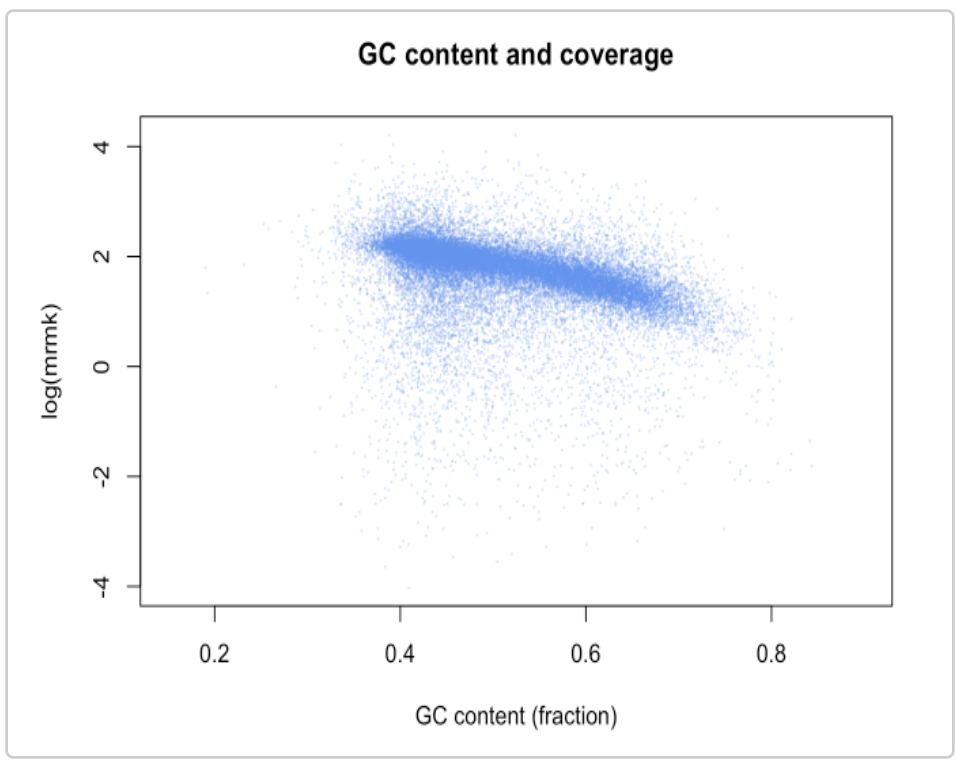
The first step is to estimate coverage over the primary transcripts and use that coverage information, normalized by GC content and later GC by relation to coverage in B73, to call CNV across the maize genome. Thus, I will estimate coverage over:

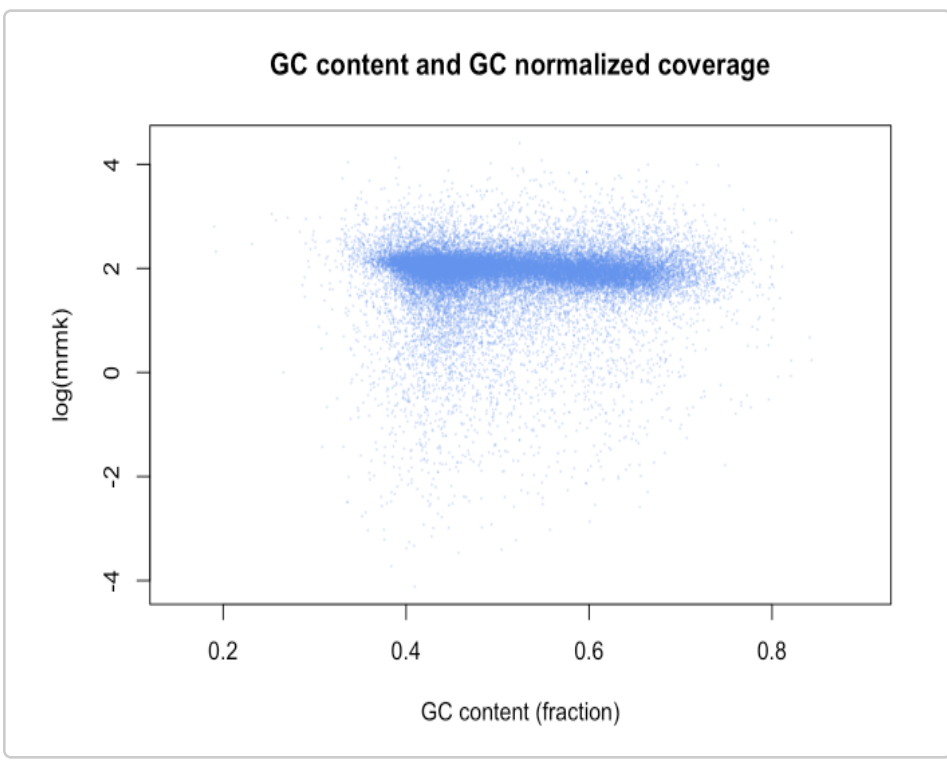
- exons
- cds
- 5' UTR
- and 3'UTR

of primary transcripts. Thus the first step will be estimating coverage using `bedtools multicov` as detailed in the above mentioned GitHub repo.

1.2 Normalization

The data are normalized by reads per kb per million (RPKM) using the script gene_dpeth.R and are then subsequently GC normalized using the `withinLaneNormalization` of `EDASeq`.

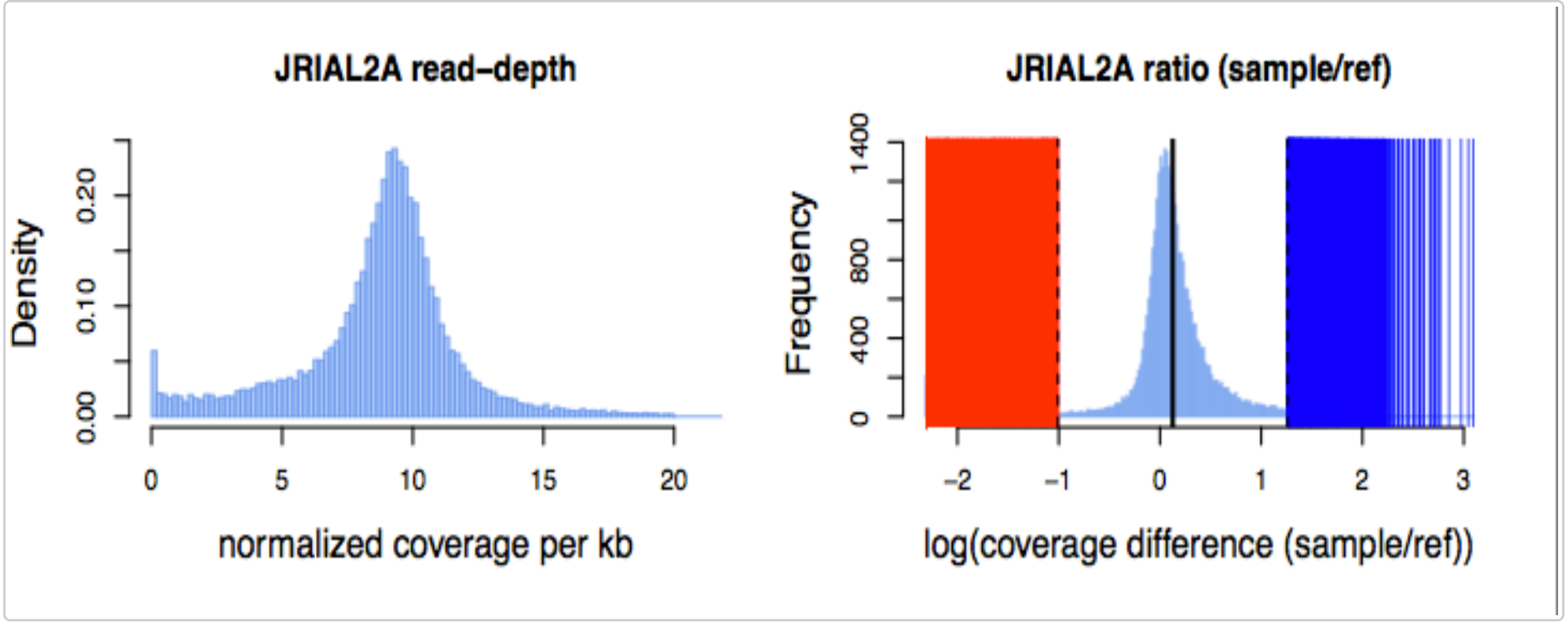




1.3 Comparing to B73 reference

The coverage over intervals within then normalized by dividing by coverage over the same region in the reference sample B73.

CNVs are scored by being 2 (or 3) Standard deviation away from the mean coverage over all intervals. This is performed separately for each sample and an example is given below. The left panel is distribution of coverage over all intervals in sample JRIAL2A, and the right panel is the ratio of coverage in the sample and B73. The vertical lines (dashed) indicate the standard deviation limit and the coloured lines indicate the CNV calls for that sample.



I doubt we will need to show this in the main paper but we will have to add to supplementary info, I reckon.

2. cn.mops

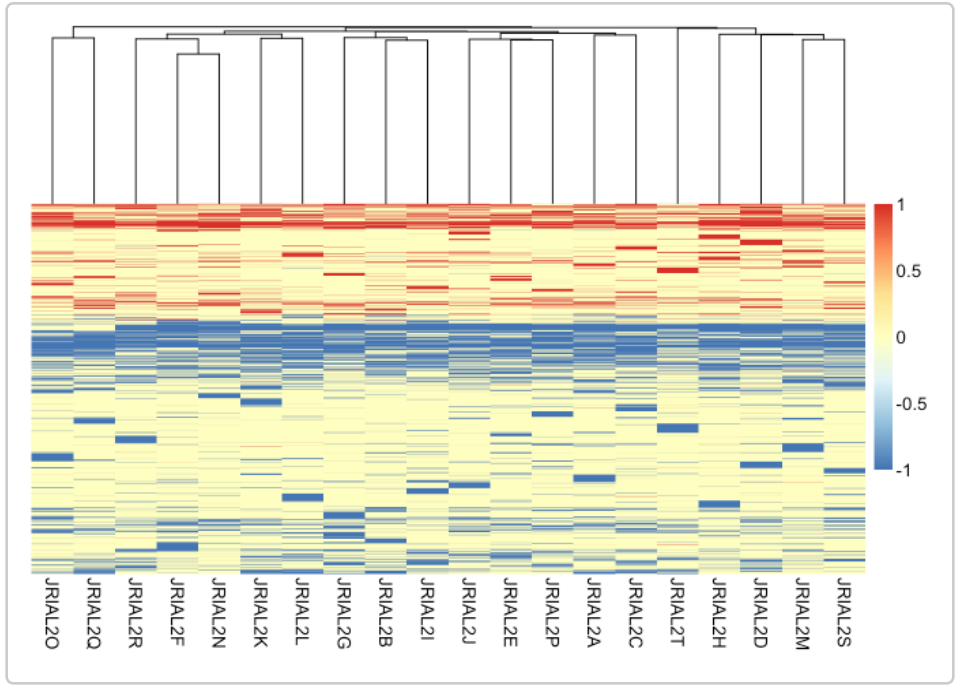
This is much easier in terms of methods, we just cite the paper and declare our parameters. I think maybe only a brief sentence or two about what `cn.mops` does is needed.

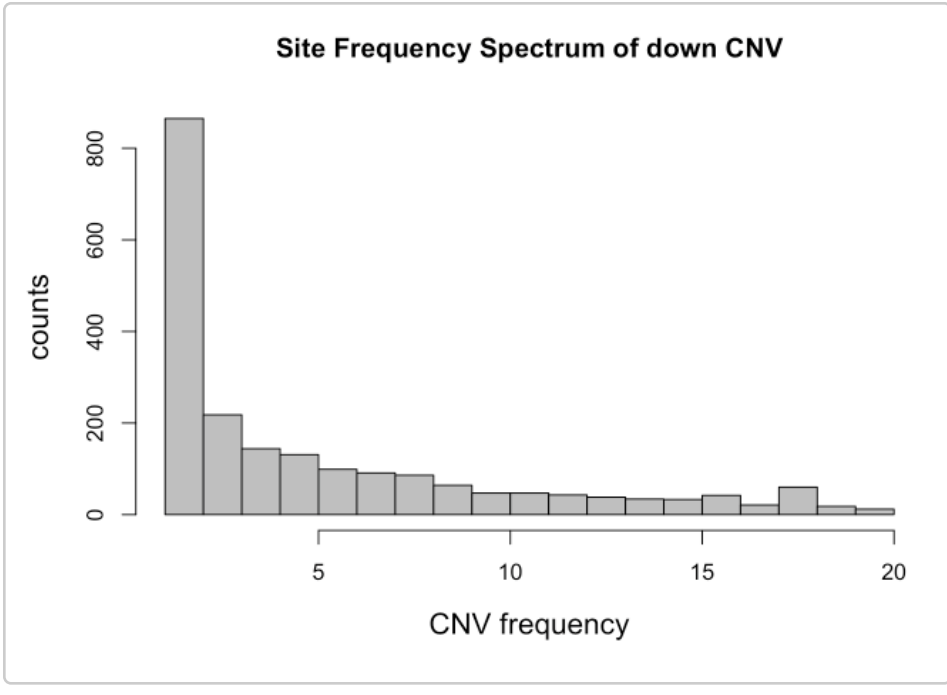
The problem will be how to integrate the two approaches…

Results

1. Calling CNVs

At the moment the custom approach reveals between ~3,000 to 5,000 (depending on the stdev cut-off) in the single wild Palmar Chico population. A plot like the one below should probably be included in the paper.





These are of varying frequency, as you can see from the following sfs.. Lots of low frequency variants with very few fixed across the population.

This will of course change over in the future, we have decided to quantify the actual copy-number rather than the number of individuals exhibiting down CNVs.

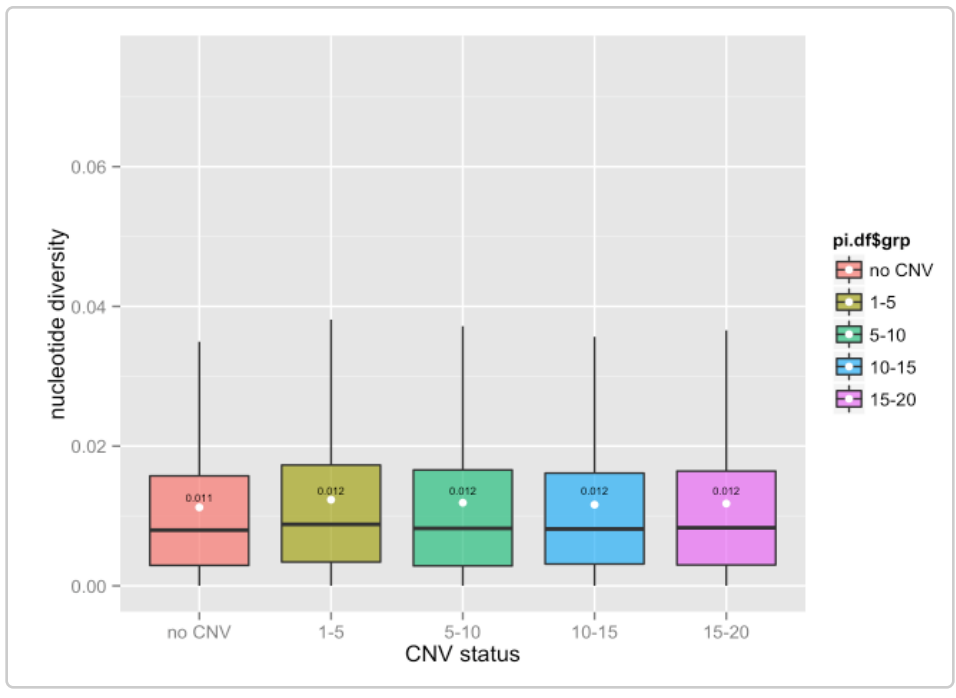
Also, we will want to look to see if the majority of CNVs are in *HWE*. THwy should be, and this will add weight to them being real.

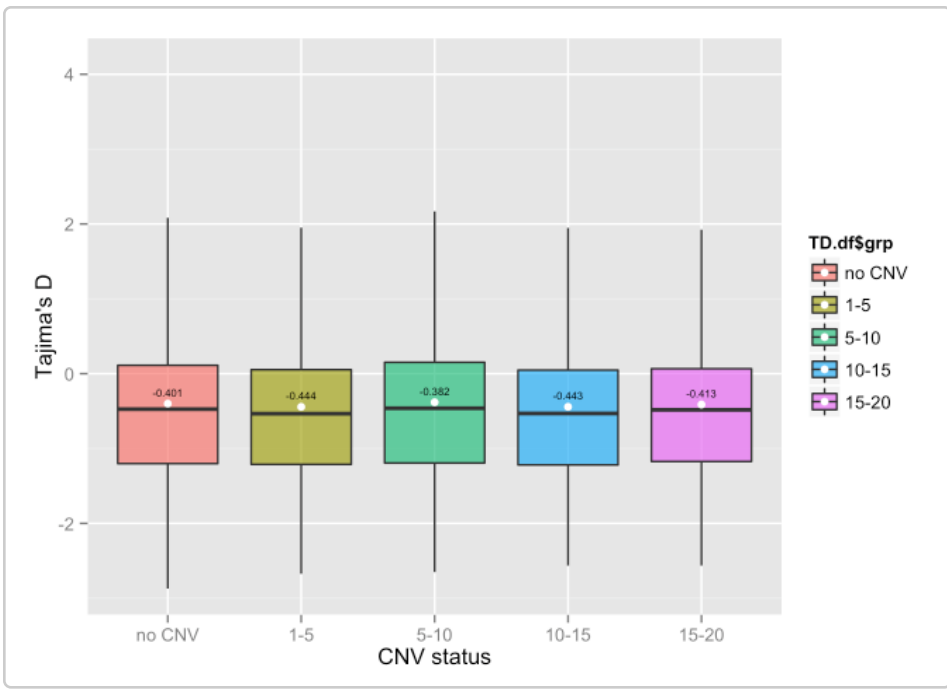
2. Patterns of diversity across CNVs

The next part of the paper should detail the patterns of diversity over CNV regions and compare these to regions of the genome that are “normal”. We can compare:

- pi
- singleton diversity and
- Tajima’s D.

At the moment we cannot find a particularly strong signal (if any). In response we have decided to examine these patterns over exon, cds, 5’ UTR and 3’UTR. Here is a place holder figure, but this will change for the paper.



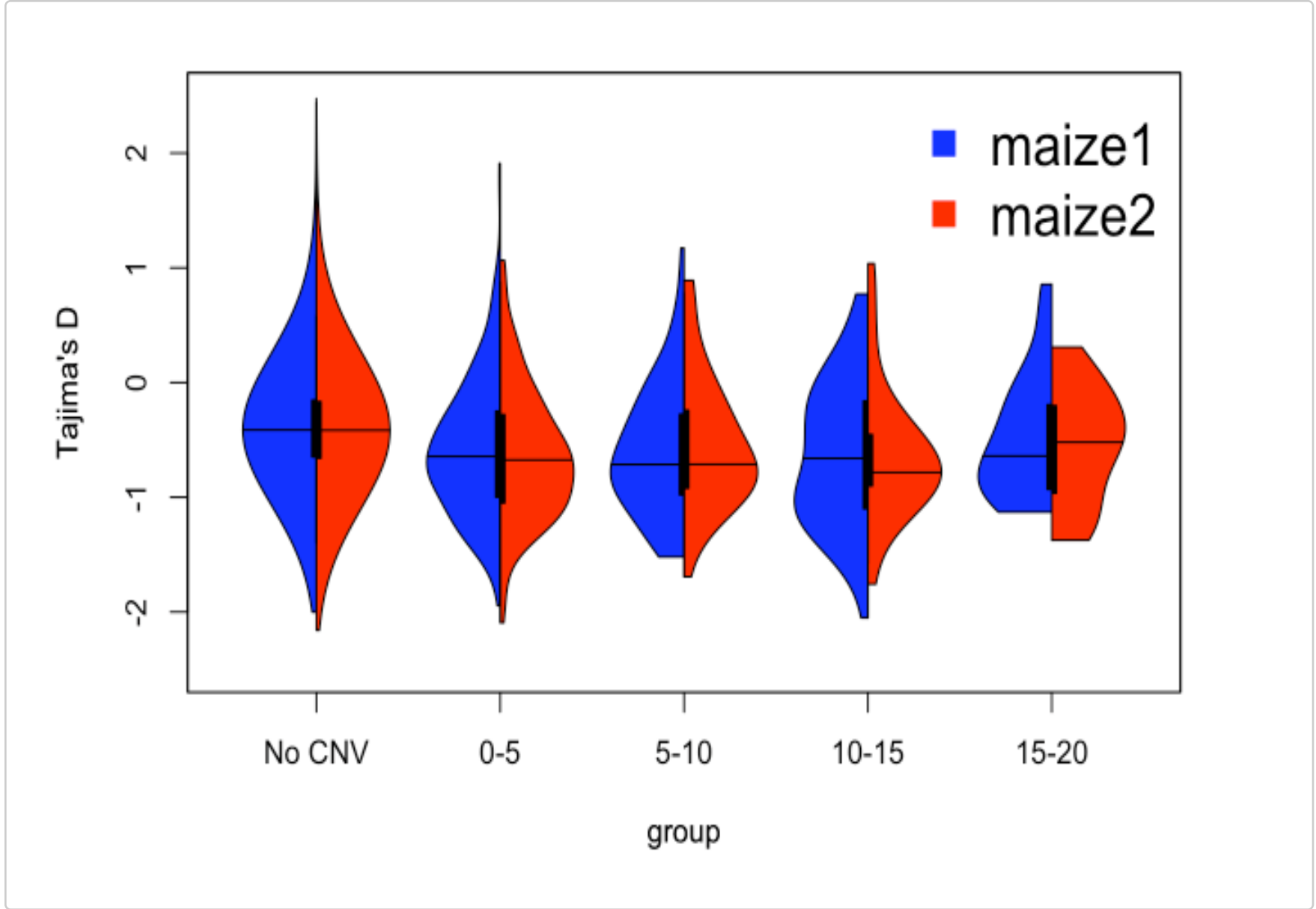


The idea is that we will see varying patterns of nucleotide diversity across CNVs at different frequencies. Perhaps more negative Tajima’s D in regions that show high copy-number CNVs….

3. Comparing maize1 and miaze2

Next, I think we should move on to examine maize1 and maize2 sub-genomes with respect to both the number of resident CNVs and the various estimates of Pi, singelton diversity and Tajima’s D. This should automatically include information on copy-number variants.

It looks as though there might be a an interesting pattern, at the moment we really need to wait for the more recent analysis to come through. Here is a place holder figure, we should probably have something like this:



We can also use *Variant Effect Predictor* to analyse and asses the impact of variants segregating within the population. Do maize1 and miaze2 seem to be effected equally by variants of different sorts? etc etc.

4. Patterns of selection across the maize Genome

Tim will, when has has a moment, perform genome-wide estimates of selection in this wild population of teosinte. This is nice and should dovetail with the previous two sections. For example we can:

1. Ask how much of the genome is affected by selection
2. Compare the impacts of selection on normal vs CNV regions (i.e. are CNVs selected against?).
3. Should researchers take inot account CNVs when looking for patterns of selection.
4. Compare patterns of selection in maize1 vs maize2
5. Link these to patterns of TajD, Pi and singleton diversity as well as outpout from `Variant Effect Predictor`.

Discussion and Conclusion

As the data are not yet fully developed I think it is pretty difficult to predict what we might be worthy of discussion. However, there were a few points that will need to be covered in the discussion section:

1. The remarkably large number of CNV calls within a single pop, there are almost certainly more if we sample more.
2. The merits of our own approach vs those of cn.mops and others.
3. What do the patterns of Pi, TajD and Singleton diversity tell us about segregating CNVs?
4. Does this impact the inference of selection across the genome.
5. Hopefully something about the difference betweem maize1 and maize2, and what we can infer about the ongoing process of fractionation.