Analysing Sequencing Data within Populations

Ben Blackburne

Deletion Data

Illustrated in figure are PCO plots of the deletion distances calculated by four different metrics. These are: Euclidean (equation 1), Manhattan (2) (i.e. taxicab-space), and two others chosen as representative of the 19 distance metrics tried: Kulczynski (3) and Bray-Curtis (4). Their respective forumulae are given below:

$$d_{jk} = \sqrt{\sum_i (x_{ij} - x_{ik})^2} \tag{1}$$

$$d_{jk} = \sum_i |x_{ij} - x_{ik}| \tag{2}$$

$$d_{jk} = 1 - 0.5\left(\frac{min(x_{ij} - x_{ik})}{\sum_i x_{ik}} + \frac{min(x_{ij} - x_{ik})}{\sum_i x_{ij}}\right) \tag{3}$$

$$d_{jk} = \frac{\sum_i |x_{ij} - x_{ik}|}{\sum_i (x_{ij} + x_{ik})} \tag{4}$$

The deletion plots illustrate how different distance metrics perform differently in separating populations. In each case, the two eigenvalues of the two principal components do not drastically outweigh the others, indicating that the majority of the separation cannot be plotted on two dimensions. However, clearly CEU, YRB and JPT/CHB are separable in two dimensions (and for Bray-Curtis, the just first factor is sufficient). The rest of the distance may be accounting for variations within rather than between populations.

Insertion Data

For the insertion data (figure 2) a similar pattern is observed, with the most successful separation given by Bray-Curtis.

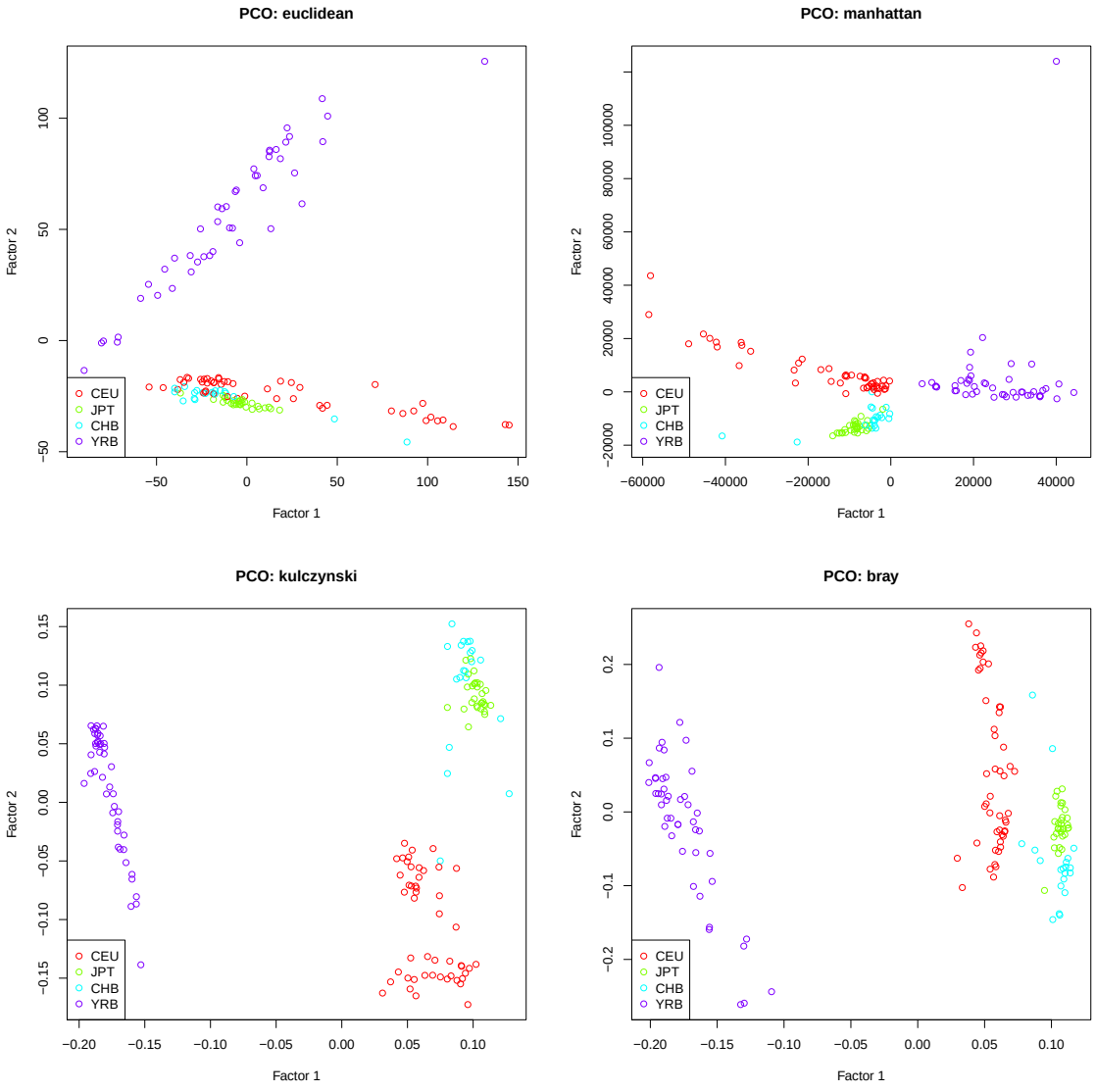


Figure 1: PCO analyses for Deletion Data

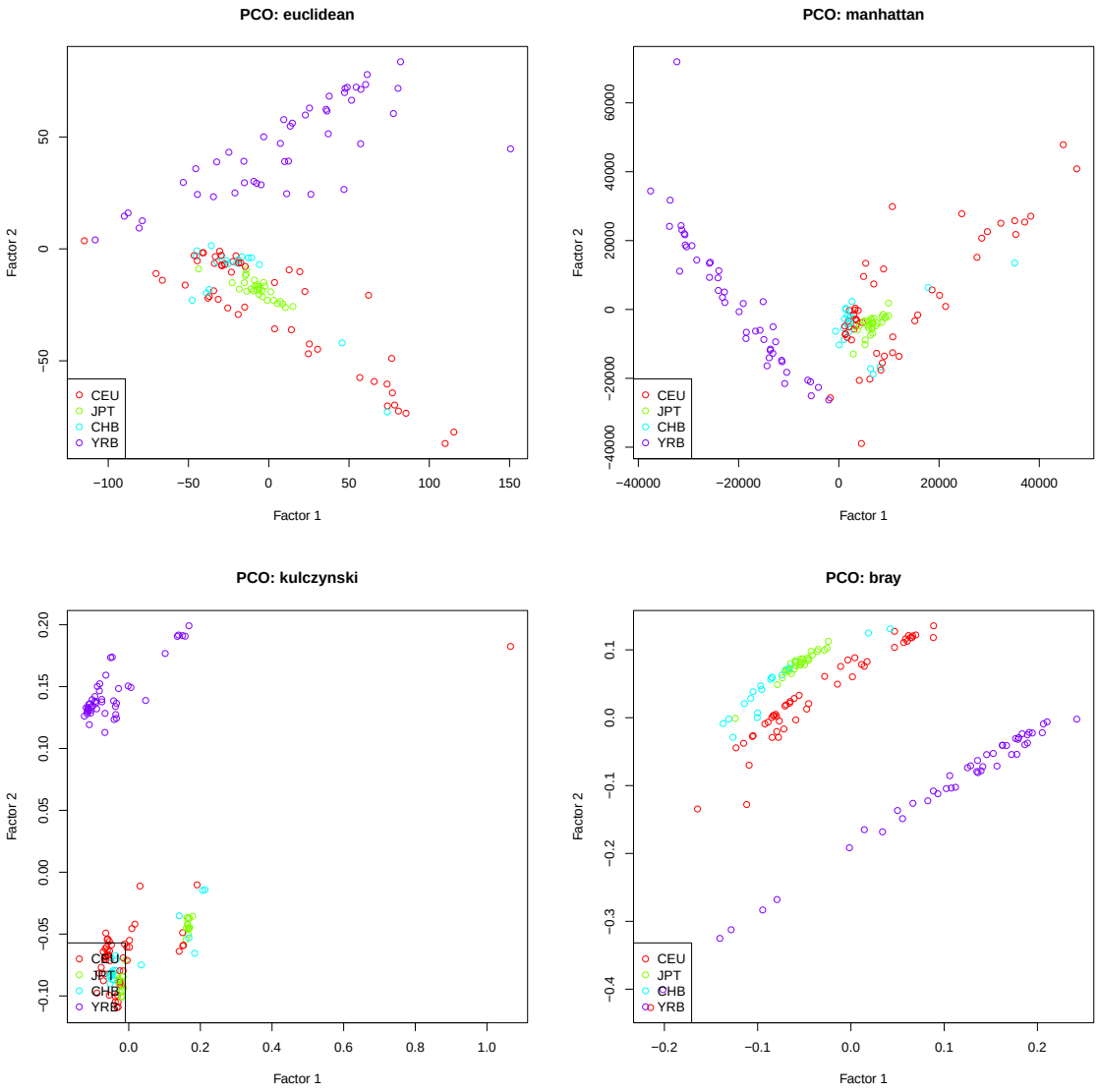


Figure 2: PCO analyses for insertion data

Combined Data

Combining the data together (and removing the samples that only have deletion or insertion entries) produces the result in figure 3. These data are similar to the deletion-only plots, and the eigenvalues (not shown) show a similar pattern too.

Effect of indel size

By restricting the data to deletions of length greater than 10 (according to the IndelSize column), we reduce the dataset to 50 302 deletions. We perform the same analysis on these data in figure 4. For comparison, we also analyse a random sample of 50 302 deletions (regardless of size) in 5.

It is not totally clear that there is any advantage to restricting to large deletions. But I can make some observations. The JPT-CHB separation seems to be much greater in the Bray-Curtis plot (although there are some anomolous points amongst CHB). Also, there are clearly two separate clusters in CEU, which was hinted at in earlier Kulczynski plots. Whether the overall clustering of the plots is any better than the random control is not totally clear, however.

Another potential method along these lines could lie in weighting indels (i.e. large indels produce larger distances than small indels).

Removal of lowest coverage samples

In figure 6 we successively remove samples with less than a given threshold coverage (1,2,3,4,6 and 8X) and reanalyse the data. Removal of samples does not yield tighter clusters, and once samples of <3X are removed the data are noticeable less separable.

Single chromosome analysis

Figure 7 shows the Bray-Curtis PCO analysis for chromosomes 1,6 and X on the combined indel datasets; in comparison with all chromosomes. Clearly, single chromosomes are sufficient to give good separation for the populations. This might be important if we don't have rights to all the data.

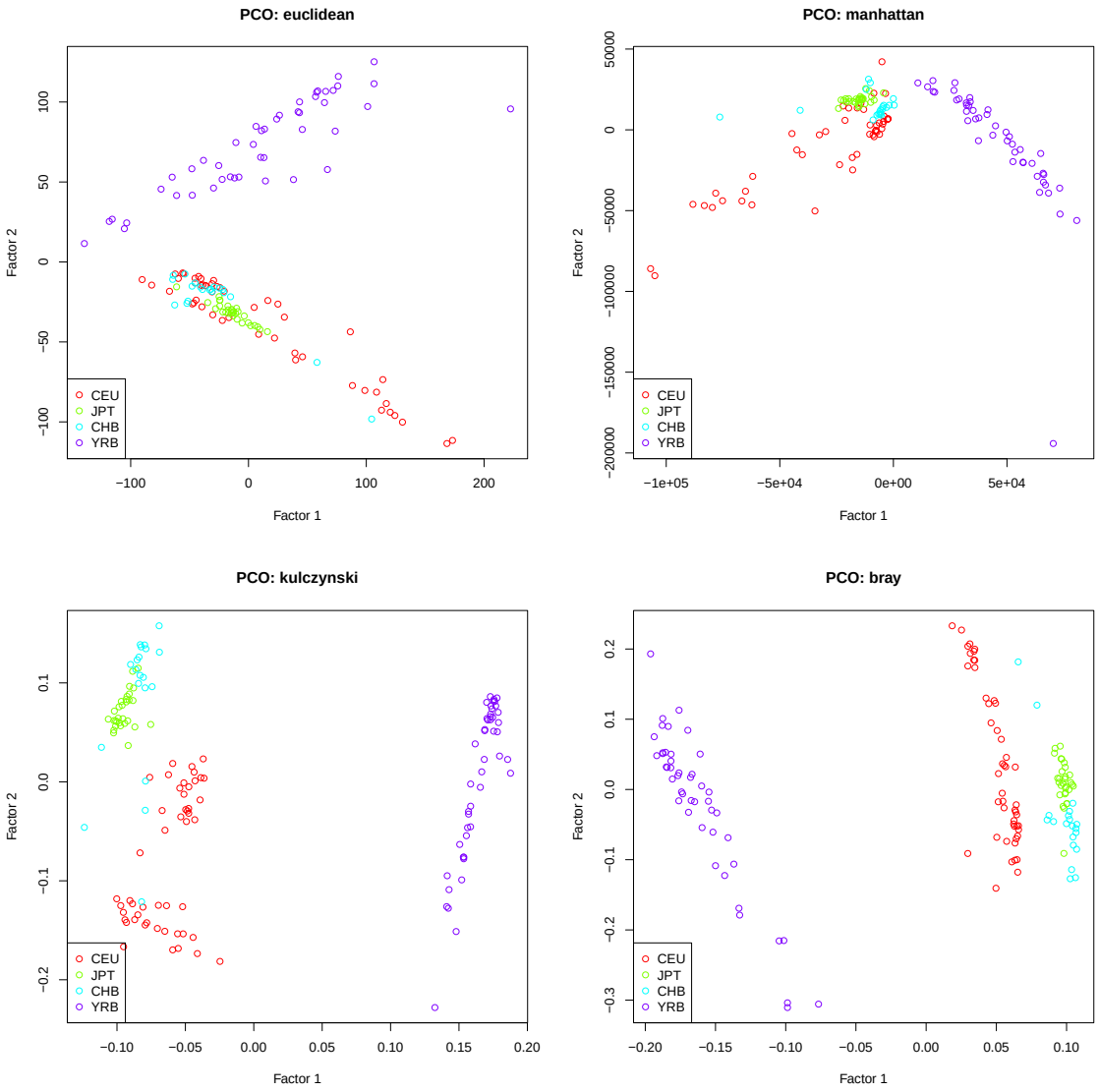


Figure 3: PCO analyses for both insertion and deletion data

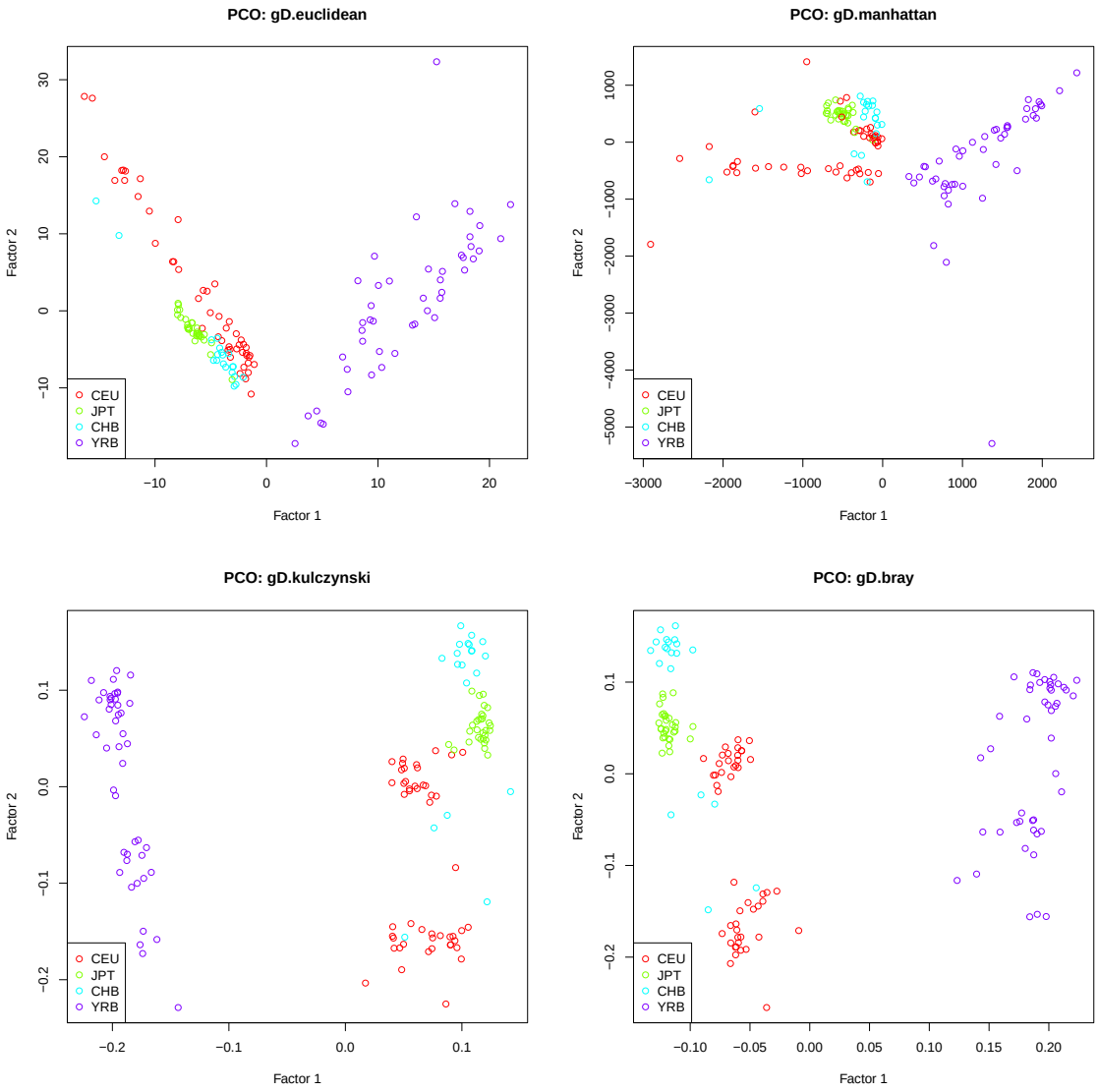


Figure 4: PCO analyses for the 50 302 large (>10) deletions

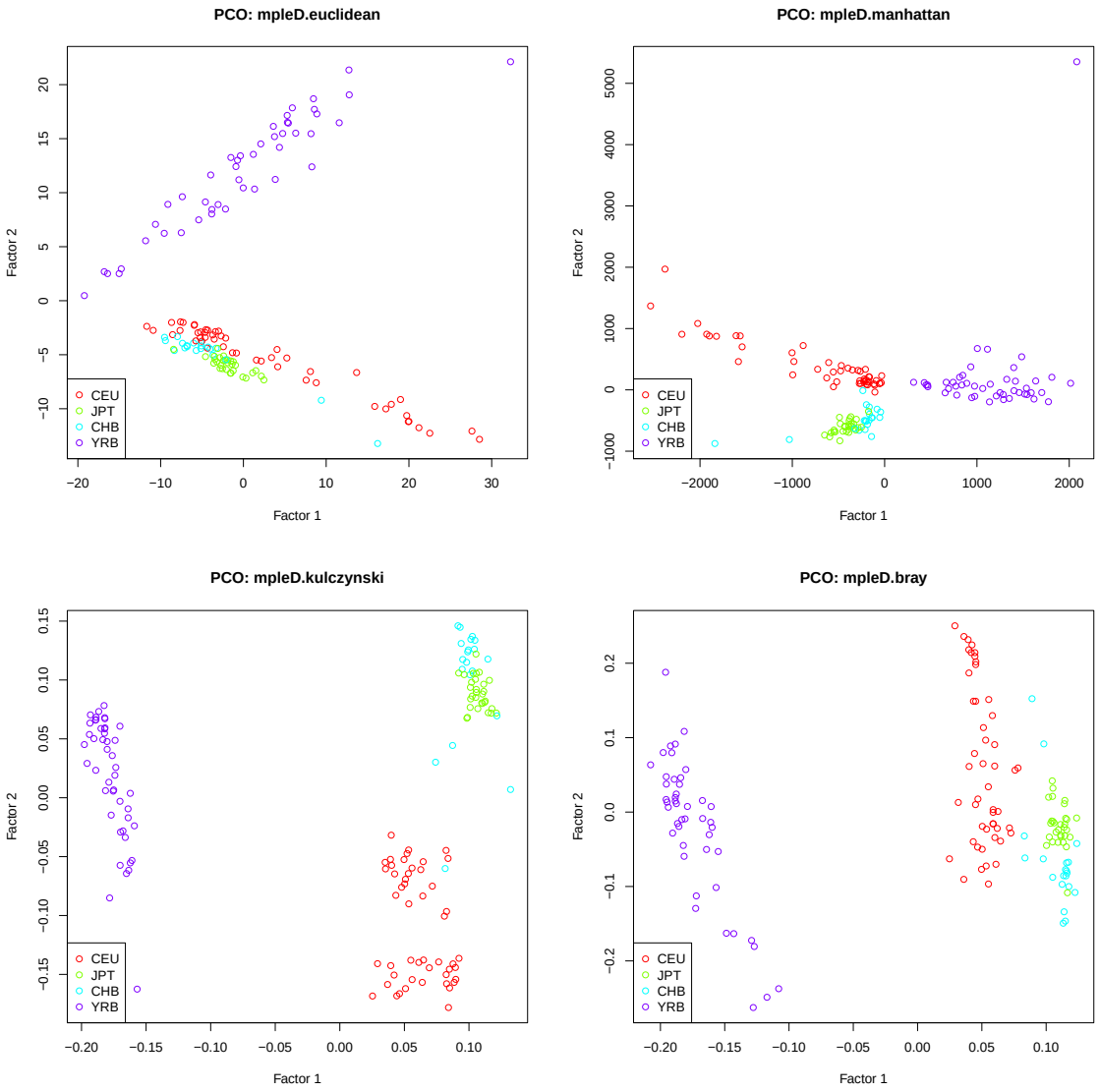


Figure 5: PCO analyses for a random set of 50 302 deletions

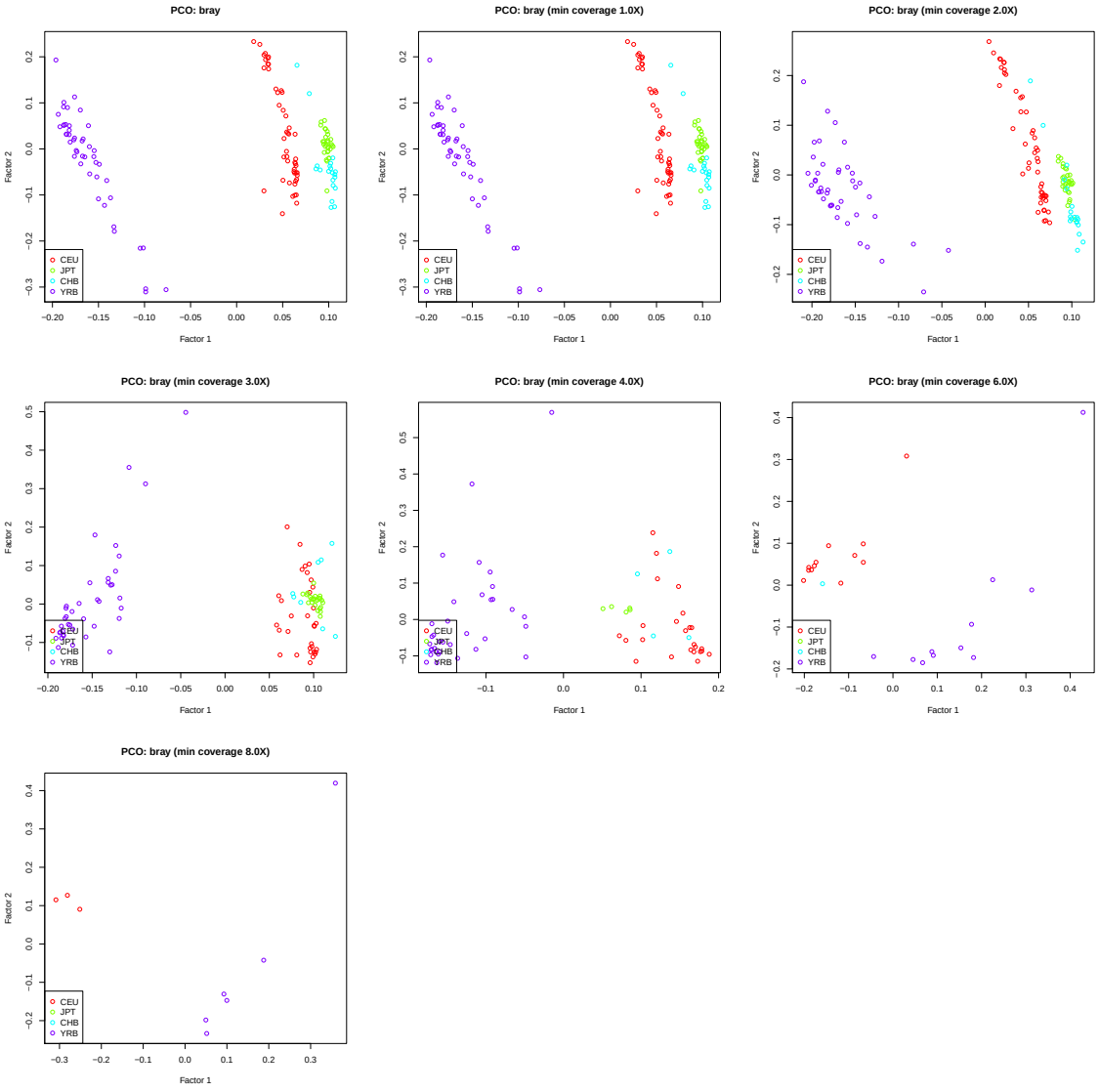
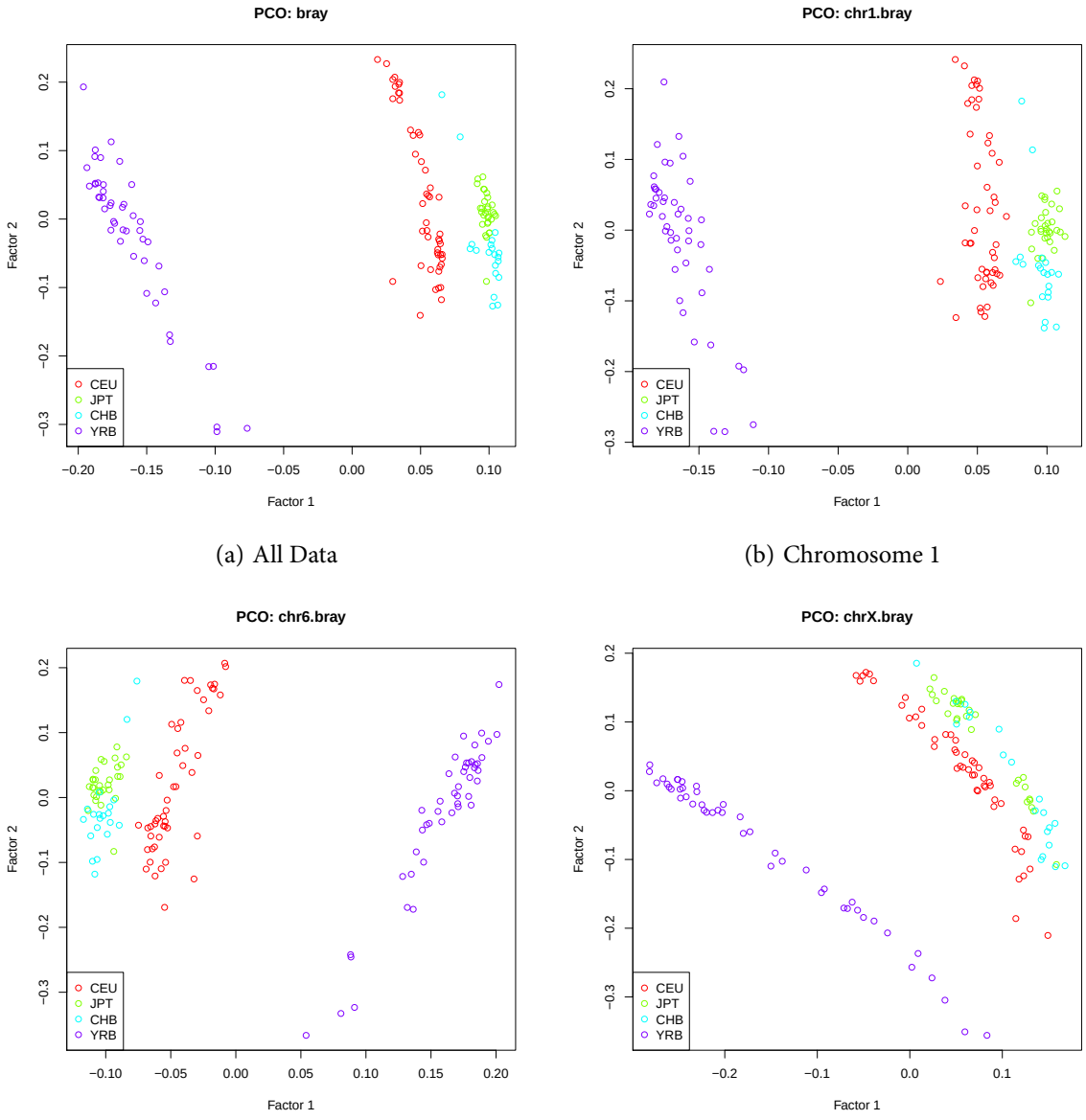


Figure 6: Increasing cutoff



(a) All Data

(b) Chromosome 1

(c) Chromosome 6

(d) Chromsome X

Figure 7: Bray-Curtis PCO analysis of all data and single chromosomes

How much coverage is needed?

Future work - since including even low coverage samples seems to be useful, how low can coverage get before this information is lost?

How do SNPs compare?

Todo.

Discussion

There are examples of previous work directly and indirectly related to this analysis [1] [2] [3] [4]. In fact, in one recent study [2] the point is made that in PCO analyses the majority of the variation is not contained in the first two or three factors, as observed here. Their solution is to perform PCO and cluster within the n-dimensional space using a monte-carlo method. It may be worth investigating this further.

Are there really two clusters within CEU? If so, then it can be demonstrated that the large indels are more effective at distinguishing between these. The analyses cited above might be useful in verifying these clusters.

References

[1] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, Dec 2006.

[2] Patrick A Reeves and Christopher M Richards. Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. *PLoS ONE*, 4(1):e4269, Jan 2009.

[3] John P Huelsenbeck and Peter Andolfatto. Inference of population structure under a dirichlet process model. *Genetics*, 175(4):1787–802, Apr 2007.

[4] Marc Bauchet, Brian McEvoy, Laurel N Pearson, Ellen E Quillen, Tamara Sarkisian, Kristine Hovhannesyan, Ranjan Deka, Daniel G Bradley, and Mark D Shriver. Measuring european population stratification with microarray genotype data. *Am J Hum Genet*, 80(5):948–56, May 2007.