Detecting Heterogeneity in Population Structure Across Chromosomes: the CAnD Package

Caitlin McHugh[1*]

[1] Department of Biostatistics, University of Washington

[*] mchughc (at) uw.edu

April 24, 2017

Contents

1 Introduction

With the advent of dense, accurate and inexpensive genomic data, researchers are able to perform analyses that estimate ancestry across the entire genome. In particular, ancestry can be inferred across regions of the genome that are interesting for a disease trait, or can be inferred chromosome-wide to identify regions that have been passed down by an ancestral population.

The *CAnD* package provides functionality for the method that compares proportion ancestry in a sample set across chromosomes or chromosomal regions [1]. We calculate p-values for the observed difference in ancestry across chromosomes, properly accounting for multiple testing. An overall CAnD statistic and p-value are stored for each analysis.

This vignette describes a typical analysis workflow and includes some details regarding the statistical theory behind *CAnD*. For more technical details, please see reference [1].

2 Example workflow for CAnD

2.1 Preparing a data set for analysis

For our example, we will use a set of simulated data, the `ancestries` data set from the *CAnD* package. We begin by loading relevant libraries, subsetting the data, and producing summary statistics.

```
> library(CAnD)
> data(ancestries)
> dim(ancestries)
```

```
[1] 50 70
```

We initially can look at the columns of our `ancestries` object that correspond to the estimated proportion ancestries of chromosome one.

```
> ancestries[1:2,c(1,2,25,48)]
```

```
  IID    Euro_1      Afr_1   Asian_1
1   1 0.1536889 0.07994151 0.7663696
2   2 0.1108866 0.01604743 0.8730660
```

The `ancestries` data.frame holds simulated proportions for a set of 50 samples. Every row corresponds to a sample and each sample has a unique id, stored as `IID`. We imagine the proportions displayed in `ancestries` were estimated from a program such as FRAPPE [2], ADMIXTURE [3] or RFMix [4]. In this particular example, three ancestral subpopulations were assumed, namely `Euro`, `Afr` and `Asian`. The proportions can be locus-specific ancestry averaged across chromosomes, or could be any other sort of ancestral estimate for a portion of the genome. Furthermore, there can be any number of ancestral populations. Of course, the results are only interesting with two or more ancestries. In our sample data set, every sample has a column corresponding to the ancestral proportion for each of the three

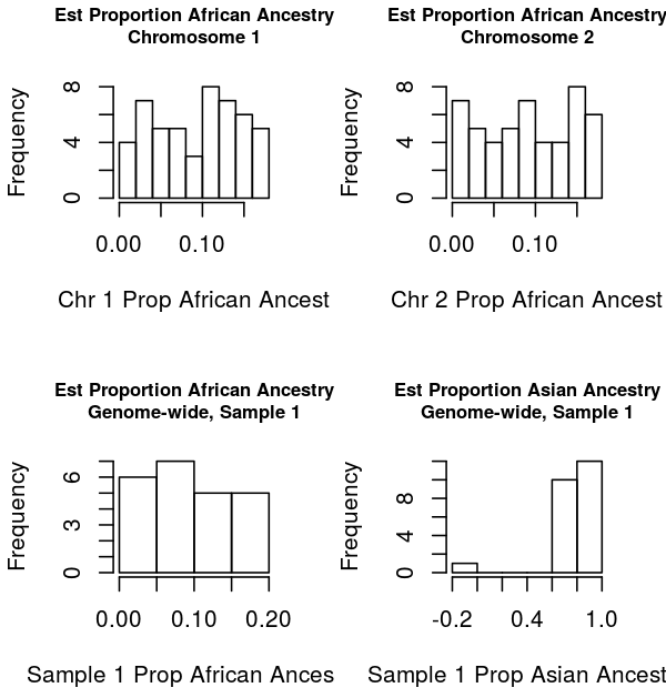

Figure 1: Histograms of estimated proportion African ancestry for all samples on chromosome 1 and 2, and estimated proportion African and Asian ancestry, genome-wide, for Sample 1.

ancestries for all autosomal chromosomes 1-22 and the X chromosome. The three proportions should sum to one for each chromosome within a sample.

First we can examine the estimated proportions, both by sample and by ancestry. We create histograms of these values, seen in Figure 1.

```
> par(mfrow=c(2,2))
> hist(ancestries$Afr_1,main="Est Proportion African Ancestry\nChromosome 1",
+      xlab="Chr 1 Prop African Ancest",cex.main=0.8)
> hist(ancestries$Afr_2,main="Est Proportion African Ancestry\nChromosome 2",
+      xlab="Chr 2 Prop African Ancest",cex.main=0.8)
> afrCols <- seq(from=25,to=(25+22))
> asianCols <- seq(from=(25+22+1),to=ncol(ancestries))
> hist(as.numeric(ancestries[1,afrCols]),main="Est Proportion African Ancestry\nGenome-wi
+      xlab="Sample 1 Prop African Ancestry",cex.main=0.8)
> hist(as.numeric(ancestries[1,asianCols]),main="Est Proportion Asian Ancestry\nGenome-wi
+      xlab="Sample 1 Prop Asian Ancestry",cex.main=0.8)
```

The `data.frame` is the only input file required to run the CAnD tests. For each test, we will subset the columns to the particular ancestry of interest.

2.2 Running the CAnD Test

The CAnD test detects heterogeneity in population structure patterns across chromosomes. CAnD uses local ancestry estimated from SNP genotype data to identify significant differences in ancestral contributions to chromosomes in samples from admixed populations. Statistically, CAnD compares a chromosome c with a pool of all other chromosomes. The null hypothesis is that the mean difference between ancestry proportion on chromosome c and the mean ancestry proportion across all other chromosomes is zero. For more details, see Section 3.1 and reference [1].

We will perform the CAnD test on the estimated proportions of European ancestry. In order to do this, we first subset `ancestries` to the columns of interest.

```
> euroCols <- seq(from=2,to=(2+22))
> head(ancestries[,euroCols[20:23]],2)
```

```
     Euro_20   Euro_21    Euro_22      Euro_X
1 0.03326729 0.1663629 0.09374132 0.003506331
2 0.09929861 0.1444173 0.04413520 0.009132646
```

```
> colnames(ancestries[euroCols])
```

```
 [1] "Euro_1"  "Euro_2"  "Euro_3"  "Euro_4"  "Euro_5"  "Euro_6"  "Euro_7"  "Euro_8"
 [9] "Euro_9"  "Euro_10" "Euro_11" "Euro_12" "Euro_13" "Euro_14" "Euro_15" "Euro_16"
[17] "Euro_17" "Euro_18" "Euro_19" "Euro_20" "Euro_21" "Euro_22" "Euro_X"
```

```
> euroEsts <- ancestries[,euroCols]
> dim(euroEsts)
```

```
[1] 50 23
```

```
> head(euroEsts[,1:5],2)
```

```
     Euro_1    Euro_2     Euro_3     Euro_4     Euro_5
1 0.1536889 0.9195305 0.12707480 0.01081360 0.11025599
2 0.1108866 0.9758581 0.05505619 0.09160169 0.07248881
```

Then, we can simply run the CAnD test across all chromosomes for the estimated European ancestry in our 50 samples.

```
> param_cRes <- CAnD(euroEsts)
> param_cRes
```

```
CAnD results for parametric test
Bonferroni correction was used
p-values = 0.00971
 p-values = 7.16e-65
 p-values = 4.97e-07
 p-values = 8.59e-05
 p-values = 1.68e-07
 p-values = 0.000251
 p-values = 0.00421
```

```
 p-values = 0.00801
 p-values = 0.145
 p-values = 0.163
 p-values = 0.000133
 p-values = 0.336
 p-values = 1.88e-05
 p-values = 0.00491
 p-values = 0.000826
 p-values = 0.000624
 p-values = 0.011
 p-values = 1.05e-05
 p-values = 0.000151
 p-values = 0.00138
 p-values = 3.79e-07
 p-values = 0.0188
 p-values = 6.01e-55
observed CAnD statistic = 11300
calculated CAnD p-value = 0
```

```
> test(param_cRes)
[1] "parametric"
> overallpValue(param_cRes)
[1] 0
> overallStatistic(param_cRes)
[1] 11295.89
> BonfCorr(param_cRes)
[1] TRUE
```

We notice that the CAnD p-value is significant when considering the difference in chromosomal estimates of European ancestry genome-wide. To further investigate this, we can examine the p-values calculated for each chromosome.

```
> pValues(param_cRes)
      Euro_1       Euro_2       Euro_3       Euro_4       Euro_5       Euro_6
9.707144e-03 7.158090e-65 4.971329e-07 8.587877e-05 1.675722e-07 2.508206e-04
      Euro_7       Euro_8       Euro_9      Euro_10      Euro_11      Euro_12
4.209963e-03 8.006873e-03 1.452072e-01 1.629134e-01 1.332820e-04 3.356967e-01
     Euro_13      Euro_14      Euro_15      Euro_16      Euro_17      Euro_18
1.875395e-05 4.908740e-03 8.262300e-04 6.240482e-04 1.101079e-02 1.048361e-05
     Euro_19      Euro_20      Euro_21      Euro_22       Euro_X
1.506853e-04 1.375600e-03 3.793303e-07 1.878355e-02 6.009352e-55
```

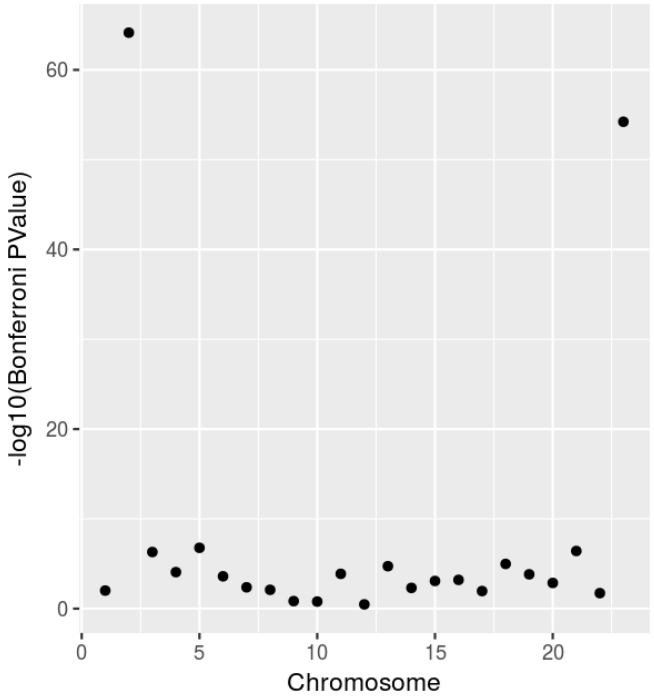

Figure 2: The calculated p-value from the CAnD method to detect heterogeneity in proportion European ancestry by chromosome.

The p-value comparing the X chromosome to the autosomes is highly significant, implying that the estimated European ancestry on the X chromosome is statistically significantly different from that on the autosomes.

2.3 Visualizing Results

There are two plotting functions available in *CAnD* to visualize results from the CAnD method.

The `plotPvals` function plots the calculated p-values against each chromosome/chromosomal region. We will show the results from the CAnD test in Figure 2.

```
> plotPvals(param_cRes, main="CAnD P-values\nProportion European Ancestry Genome-wide")
```

The `barPlotAncest` function plots the proportion ancestry for a given chromosome/chromosomal region for each sample. This visualization is an efficient way to compare the proportions ancestry across the entire sample. Note this is simply a summary plot and does not require running of the CAnD tests to produce. We see the results for our sample in Figure 3.

```
> chr1 <- ancestries[,c("Euro_1","Afr_1","Asian_1")]
> barPlotAncest(chr1,title="Chromosome 1 Ancestry Proportions")
```

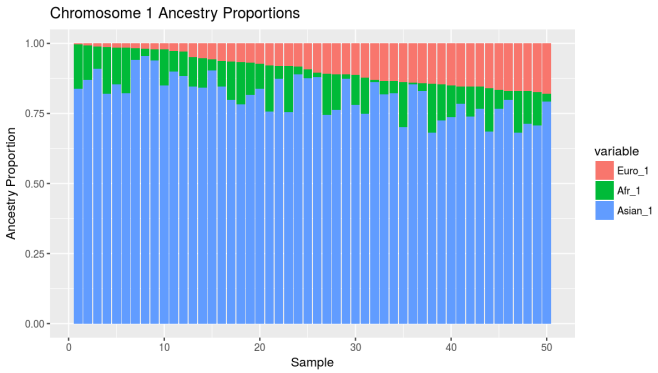

Figure 3: Barplot of chromosome 1 ancestry proportions, ordered by increasing proportion European ancestry.

3 Methods in brief

Define the proportion ancestry from subpopulation k for individual i to be a_{ik}, $i \in \{1, \ldots, N\}$. Let $G_{-c} = \{1, 2, \ldots, c-1, c+1, \ldots, 22, X\}$. For a given chromosome of interest c, we calculate the pooled mean of all chromosomes excluding c as

$$a_{ik}^{-c} = \frac{1}{22} \sum_{M \in G_{-c}} a_{ik}^{M}$$

The difference in ancestry between a given chromosome c and the average of all other chromosomes, in individual i and for a given ancestry subpopulation k, is

$$D_{ik}^{c} = a_{ik}^{-c} - a_{ik}^{c}$$

Denote the mean D_{ik}^{c} across all individuals i as $\overline{D_k^c}$.

3.1 CAnD Methods

The CAnD method tests for heterogeneity across m chromosomes [1]. We first define the t-statistic comparing differences in ancestry subpopulation k on chromosome c with a pool of the other chromosomes as

$$T_k = \overline{D_k^c} / \sqrt{v_k^2 / n}$$

where $v_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_{ik}^{c} - \overline{D_k^c})^2$ is the sample variance. Note this statistic takes into account the average ancestry difference between chromosome c and the mean ancestry of the other chromosomes across all individuals as well as within individuals. T_k has $n-1$ degrees of freedom and is a test statistic that tests the null hypothesis that the mean difference between the ancestry proportion on chromosome c and the ancestry proportion across all other chromosomes for subpopulation k is zero. We calculate

T_k for each chromosome c of interest, and obtain m p-values $p_c, c \in \{1, \ldots, m\}$. Then, we define

$$\widetilde{T}_k = \begin{pmatrix} T_k^1 \\ T_k^2 \\ \vdots \\ T_k^m \end{pmatrix} \sim \text{MVN}(0, \Sigma)$$

It follows that combined CAnD statistic, allowing for correlation between T_k^c and $T_k^{c'}$, is

$$\chi^2_{CAnD} = \widetilde{T}_k' \Sigma^{-1} \widetilde{T}_k \tag{1}$$

which follows a χ^2_m distribution under the null hypothesis.

4 Session Info

- R version 3.4.0 (2017-04-21), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Running under: `Ubuntu 16.04.2 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so`
- LAPACK: `/home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: CAnD 1.8.0
- Loaded via a namespace (and not attached): BiocStyle 2.4.0, Rcpp 0.12.10, backports 1.0.5, colorspace 1.3-2, compiler 3.4.0, digest 0.6.12, evaluate 0.10, ggplot2 2.2.1, grid 3.4.0, gtable 0.2.0, htmltools 0.3.5, knitr 1.15.1, labeling 0.3, lazyeval 0.2.0, magrittr 1.5, munsell 0.4.3, plyr 1.8.4, reshape 0.8.6, rmarkdown 1.4, rprojroot 1.2, scales 0.4.1, stringi 1.1.5, stringr 1.2.0, tibble 1.3.0, tools 3.4.0, yaml 2.1.14

5 References

1. McHugh, C., Brown, L., Thornton, T. Detecting heterogeneity in population structure across the genome in admixed populations. Genetics, 2016.

2. Tang, H., Peng, J., Wang, P., Risch, N.J. Estimation of individual admixture: Analytical and study design considerations. Genetic Epidemiology, 2005.

3. Alexander, D.H., Novembre, J., Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 2009.

4. Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. American Journal of Human Genetics, 2013.