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 16, 2015

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 two methods that compare proportion ancestry in a sample set across chromosomes or chromosomal regions [1]. Both of the methods 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

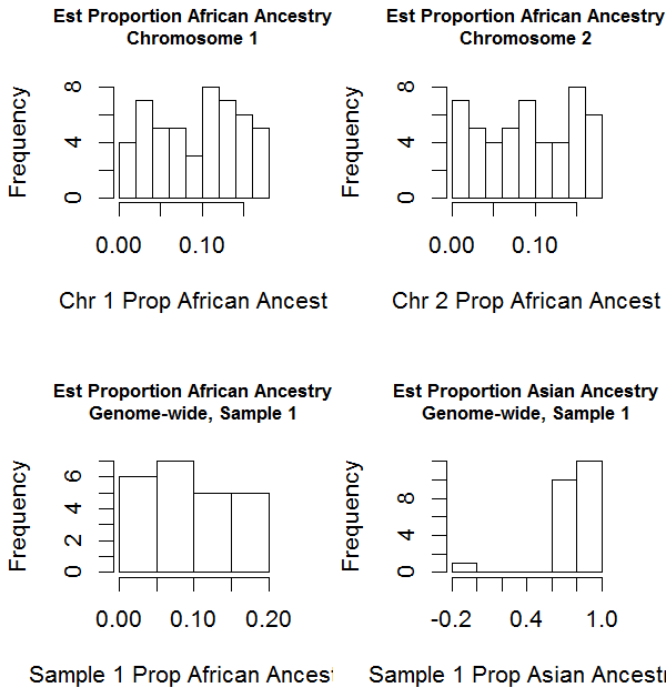

Figure 1: Histograms of estimated proportion African ancestry for all samples on chromosome 1 and 2, and estimated proportion African and Asian ancestry, genome-wide, for Sample 1.

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 = 1
 p-values = 2.92e-65
 p-values = 1.5e-07
 p-values = 3.24e-05
 p-values = 6.84e-08
 p-values = 0.000101
 p-values = 0.00275
```

```
 p-values = 0.00442
 p-values = 0.098
 p-values = 0.0966
 p-values = 8.69e-05
 p-values = 0.172
 p-values = 9.28e-06
 p-values = 0.00233
 p-values = 0.000421
 p-values = 0.000373
 p-values = 0.00433
 p-values = 3.63e-06
 p-values = 7.35e-05
 p-values = 0.000467
 p-values = 1.58e-07
 p-values = 0.00234
 p-values = 6.01e-55
observed CAnD statistic = 1030
calculated CAnD p-value = 0
```

```
> test(param_cRes)
[1] "parametric"
> overallpValue(param_cRes)
[1] 0
> overallStatistic(param_cRes)
[1] 1025.047
> 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
1.000000e+00 2.920087e-65 1.501976e-07 3.243111e-05 6.843154e-08 1.005373e-04
      Euro_7       Euro_8       Euro_9      Euro_10      Euro_11      Euro_12
2.753122e-03 4.422774e-03 9.796305e-02 9.656372e-02 8.686459e-05 1.718562e-01
     Euro_13      Euro_14      Euro_15      Euro_16      Euro_17      Euro_18
9.281243e-06 2.329372e-03 4.209857e-04 3.727050e-04 4.327478e-03 3.631504e-06
     Euro_19      Euro_20      Euro_21      Euro_22       Euro_X
7.352685e-05 4.674701e-04 1.580841e-07 2.342359e-03 6.009352e-55
```

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.

If we run the CAnD test with a set of 20 samples or smaller, we will receive a warning that perhaps the sample size is too small for this method. As an alternative, we can run the non-parametric CAnD test.

2.3 Running the Non-Parametric CAnD Test

The non-parametric CAnD test is an alternative to the CAnD test introduced in Section 2.2 when there is a small number of samples. The non-parametric CAnD test calculates S_k, the number of samples for which the difference between the mean ancestry of all chromosomes excluding chromosome c and chromosome c is greater than zero. Under the null hypothesis that there is no ancestry difference between chromosome c and a pool of the other chromosomes, S_k follows a Binomial$(n, 0.5)$ distribution, where n is the number of samples. For more details, see Section 3.2 and reference [1].

To run the non-parametric CAnD test, we first must subset `ancestries` to the proportion of ancestry for one subpopulation. For this vignette, we will consider the estimated proportion of Asian ancestry across all chromosomes for each sample.

```
> head(ancestries[,asianCols[6:10]],2)
```

```
    Asian_6   Asian_7   Asian_8   Asian_9  Asian_10
1 0.7033519 0.8624828 0.8008158 0.7110126 0.7712863
2 0.6553125 0.8405822 0.8683530 0.7424734 0.9043579
```

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

```
 [1] "Asian_1"  "Asian_2"  "Asian_3"  "Asian_4"  "Asian_5"  "Asian_6"  "Asian_7"
 [8] "Asian_8"  "Asian_9"  "Asian_10" "Asian_11" "Asian_12" "Asian_13" "Asian_14"
[15] "Asian_15" "Asian_16" "Asian_17" "Asian_18" "Asian_19" "Asian_20" "Asian_21"
[22] "Asian_22" "Asian_X"
```

```
> asianEsts <- ancestries[,asianCols]
> dim(asianEsts)
```

```
[1] 50 23
```

We now have the proper input to call `nonParam_CAnD` for the Asian ancestral subpopulation. We can consider all ancestral subpopulations in turn, or simply consider the Asian ancestry. Furthermore, we can consider all chromosomes genome-wide, or perhaps consider just the autosomes by subsetting the columns further in `ancestries`. Here, we will perform the analysis on the Asian ancestry for all chromosomes genome-wide.

By default, the non-parametric CAnD test will correct the p-values calculated for multiple testing using Bonferroni multiple testing correction, where the number of tests corresponds to the number of chromosomes or chromosomal regions under examination. Genome-wide this is 23 tests, one for each of the autosomes and one for the X chromosome. Accessor functions for the resulting `CAnDResult` allow us to view the results.

```
> cRes <- nonParam_CAnD(asianEsts)
> cRes
CAnD results for non-parametric test
Bonferroni correction was used
p-values = 1
 p-values = 1
 p-values = 0.353
 p-values = 0.152
 p-values = 0.000549
 p-values = 1
 p-values = 0.0598
 p-values = 1
 p-values = 1
 p-values = 1
 p-values = 0.0215
 p-values = 1
 p-values = 0.152
 p-values = 0.755
 p-values = 0.0215
 p-values = 0.152
 p-values = 0.755
 p-values = 0.152
 p-values = 0.0598
 p-values = 0.755
 p-values = 0.0215
 p-values = 0.0598
 p-values = 4.09e-14
observed CAnD statistic = 135
calculated CAnD p-value = 9.55e-11
> summary(pValues(cRes))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.05985 0.15180 0.45530 1.00000 1.00000
> test(cRes)
[1] "non-parametric"
> overallpValue(cRes)
[1] 9.553791e-11
> overallStatistic(cRes)
[1] 135.445
> BonfCorr(cRes)
```

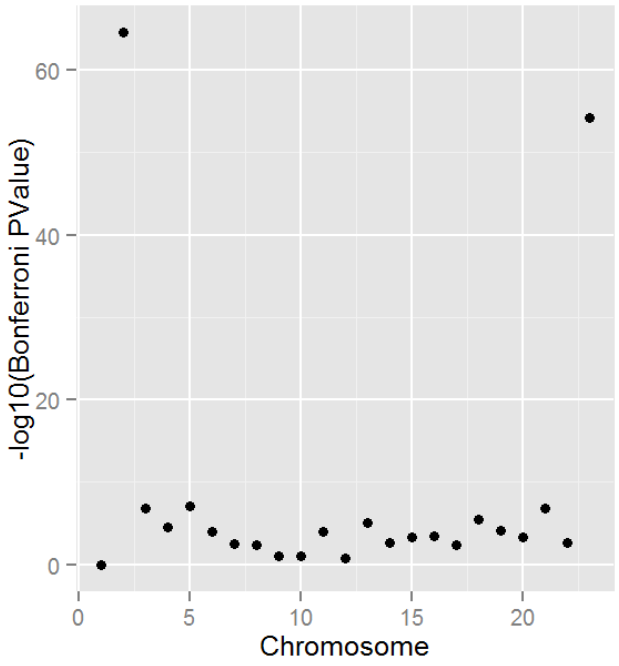

Figure 2: The calculated p-value from the parametric CAnD method to detect heterogeneity in proportion European ancestry by chromosome.

```
[1] TRUE
```

2.4 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 parametric CAnD test in Figure 2.

```
> plotPvals(param_cRes, main="Parametric CAnD P-values\nProportion European Ancestry Genc
```

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")
```

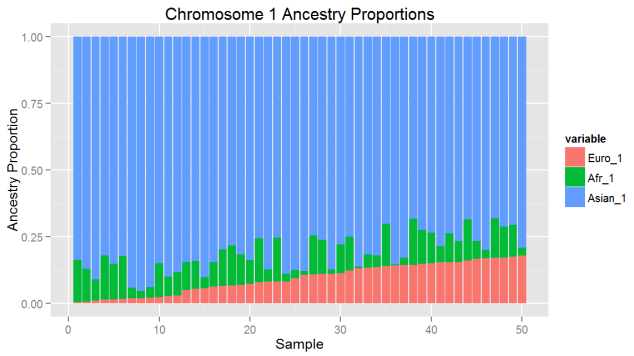

Figure 3: Barplot of chromosome 1 ancestry proportions, ordered by increasing proportion European ancestry.

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 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 $p_c, c \in \{1, \ldots, m\}$.

The overall CAnD statistic is calculated as

$$\chi^2_{CAnD} = -2\sum_{c=1}^{m} ln(p_c)$$

which is an implementation of Fisher's combined probability test. Under the null hypothesis, $\chi^2_{CAnD} \sim \chi^2_{2m}$ with the assumption that the p-values are independent and that $p_c \sim U(0,1)$.

3.2 Non-Parametric CAnD Methods

The non-parametric CAnD method is most beneficial with small sample sizes [1]. Let S_k be the non-parametric statistic for the number of individuals in N for which $D^c_{ik} > 0$,

$$S_k = \sum_{i \in N} \mathbb{I}\{D^c_{ik} > 0\}$$

S_k is equivalent to a sign statistic and under the null hypothesis of no ancestry difference between chromosome c and a pool of the other chromosomes, S_k follows a Binomial(n,0.5) distribution, where n is the number of individuals in the set N. We obtain S_k and the corresponding p-value, p_c, for all m chromosomes in turn.

To find the CAnD statistic, we apply Fisher's combined probability test to get

$$\chi^2_{CAnD} = -2\sum_{c=1}^{m} ln(p_c)$$

which assumes $p_c \sim U(0,1)$ and all p_c are independent. Under these assumptions, $\chi^2_{CAnD} \sim \chi^2_{2m}$.

4 Session Info

- R version 3.2.0 RC (2015-04-08 r68161), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: CAnD 1.0.0
- Loaded via a namespace (and not attached): BiocStyle 1.6.0, MASS 7.3-40, Rcpp 0.11.5, colorspace 1.2-6, digest 0.6.8, ggplot2 1.0.1, grid 3.2.0, gtable 0.1.2, labeling 0.3, munsell 0.4.2, plyr 1.8.1, proto 0.3-10, reshape 0.8.5, reshape2 1.4.1, scales 0.2.4, stringr 0.6.2, tools 3.2.0

5 References

1. McHugh, C., Brown, L., Thornton, T. Detecting heterogeneity in population structure across chromosomes in admixed populations. Manuscript in Preparation.

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.