HaploDistScan

Mats E. Pettersson

May 31, 2017

1 Introduction

The HaploDistScan package is intended to provide a set of function for performing a genome-wide scan for regions containing haplotypes that are substantially more different from each other than is the case for most parts of the analysed genome. Thus, it is typically best to use a data set with as little population structure as possible, as its presence will increase background variation, and weaken power to detect regions of interest.

These regions can have different provenance and means of maintenance - such as inversions repressing recombination, counter-selection against recombinants, an ongoing or partial selective sweep, or an introgression event - but the common feature is that there is at least one group of haplotypes that is much more similar within itself than compared to the rest of the haplotypes in the population. This situation will cause the distribution of all the distances to be, at least, bimodal, as the short within-group distances are separate from the out-of-group, or across-group if there are two district haplotypes, ones.

The HaploDistScan procedure calculates all pair-wise distance within each genomic region, and extracts a set of metrics from the resulting distribution. The metrics are the following:

- Hartigan's diptest statistic. This statistical test measures deviation from unimodality. We use the statistic rather than the p-value, as we want to estimate significance from the genomic distribution rather than a case with no deviation, in order to compensate for the level of structure present in the analysed data set.
- The standard deviation divided by the total number of SNPs in each region.
- The range (difference between largest and smallest distance) divided by the total number of SNPs in each region.

2 Workflow

The workflow of a complete analysis, starting with a phased genotype file, is as follows: First, we process our genotype file and generate a corresponding data frame and list of samples.

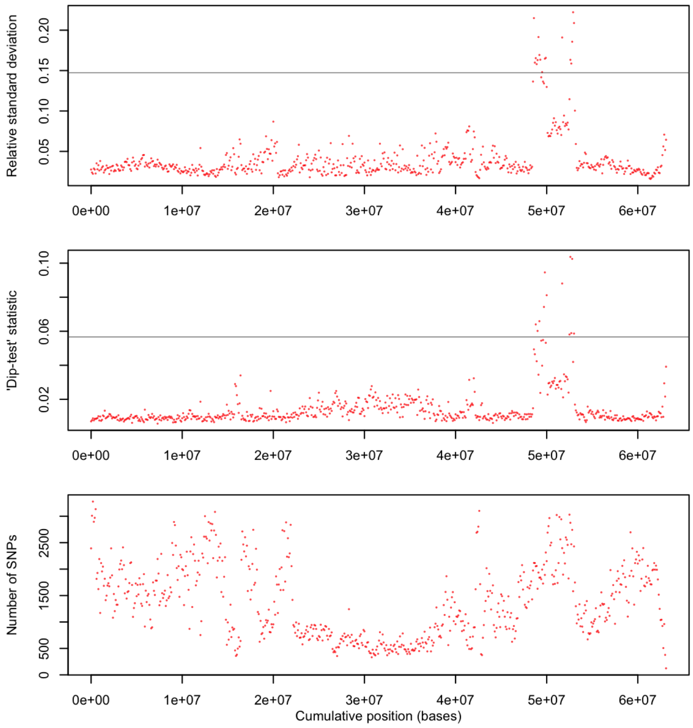
```
> ruff_chr6 <- generate_geno_df("ruff_chr6_GT.txt")
```

Then, we prepare a data frame with the location of the non-overlapping windows, and perform the scan:

```
> ruff_chr6_w <- construct_window_annotation(ruff_chr6$geno)
> ruff_chr6_s <- calculate_haplotype_distances(ruff_chr6$geno, ruff_chr6_w)
```

Finally, we visualise the results:

```
> ruff_chr6_top_pos <- plot_hap_dist_scan(ruff_chr6_s, ruff_chr6_w, 50)
```


Relative standard deviation
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
'Dip-test' statistic
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Number of SNPs
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Cumulative position (bases)