Draft: Investigation of High Heterozygosity Variants in Freeze2

AAChueva

July 14, 2016

Files that were used in the analysis:
1. File with TOPMed InDel Annotations (courtesy of Xiaoming Liu) 2. Feeeze 2 GDS GT only (includes all chromosomes)
3. HWE results for each ancestry (courtesy of Stephanie G.)

Pre-work that was done:
* Created a dataframe with variant.id, chr, pos, ref, alt, MAP20 and MAP35 fields.
* Extracted variants only for chromosome 22.

**Explore where SNPs tend to cluster on chromosome by looking at distance from SNP to telomere and centromere.

To do: Need to include code that I used to generate below file (at least a reference to which code i used to generate it)

Extracted coordinates for telomeres from this file downloaded by Cathy L from USCS on March 2011

/projects/users/cclaurie/genomics/hg19.Feb2009

```
##   chr     pos
## 1  21 9411500
## 2  21 9411785
## 3  21 9412269
## 4  21 9412658
## 5  21 9412808
## 6  21 9412886
```

```
##   chr     pos tel1_start_pos tel1_stop_pos tel2_start_pos tel2_stop_pos
## 1  21 9411500              0         10000       48119895      48129895
## 2  21 9411785              0         10000       48119895      48129895
## 3  21 9412269              0         10000       48119895      48129895
## 4  21 9412658              0         10000       48119895      48129895
## 5  21 9412808              0         10000       48119895      48129895
## 6  21 9412886              0         10000       48119895      48129895
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
```

```
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##    chrom left.base right.base
## 1      1 121535434  124535434
## 2      2  92326171   95326171
## 3      3  90504854   93504854
## 4      4  49660117   52660117
## 5      5  46405641   49405641
## 6      6  58830166   61830166
## 7      7  58054331   61054331
## 8      8  43838887   46838887
## 9      9  47367679   50367679
## 10    10  39254935   42254935
## 11    11  51644205   54644205
## 12    12  34856694   37856694
## 13    13  16000000   19000000
## 14    14  16000000   19000000
## 15    15  17000000   20000000
## 16    16  35335801   38335801
## 17    17  22263006   25263006
## 18    18  15460898   18460898
## 19    19  24681782   27681782
## 20    20  26369569   29369569
## 21    21  11288129   14288129
## 22    22  13000000   16000000
## X      X  58632012   61632012
## Y      Y  10104553   13104553
```