BMI245 Project

```r
#setwd("~/Documents/stanford/classes/spring17/bmi245/project/")
supp2 <- read.csv("colonna2014_suppl2.csv", header =TRUE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Regions that appear more than once in pairwise DDAF
newd <-  supp2 %>% group_by(POS) %>% filter(n()>1)
```

Genomic regions with expected high inter-population variation

Classic sweep examples: accepted examples of classic sweeps (that is, DARC, HERC2, MCM6, SLC24A5, SLC45A2

rs2789823 in VAV2 has 185 haplotypes in AFR population alone

SLC24A5

- Involved in skin color, pigmentation
- 15:48426484
- High DDAF

DARC (Colonna et al 2014, Fig 1)

- DDAF = .943 for AFR-EUR
- 1:159174683

rs6014096 in DOCK5 (AFR-ASN) rs1596930 in EXOC6B (AFR-ASN) rs12903208 in PRTG (AFR-ASN)

GSTCD

- DAF 0.85 in ASN 0.03 in AFR and 0.58 in EUR

ABCA12

- ATP-binding cassette, sub-family A member 12
- Suspected to be important in skin pigmentation adaptation in out-of-Africa expansion
- SNP: rs10180970 - DAF of 0.96 in Asia and 0.91 in Europe, compared with 0.13 in Africa

ZNF804A - SNPs + indels

- Regulation of TX factor related to schizophrenia
- intronic polymorphism (rs1344706)
- rs4667001 in the fourth exon changes mRNA levels
- Exomic INDEL, a novel variant; chromosome 2, position 185802211
- post-TL modifications
- present at high frequency in the ASN population
- DAF=0.83 versus 0.58 and 0.07 in EUR and AFR, respectively
- Phosphorylation of other proteins (for example, the deubiquitinating enzyme OTUB1) has been demonstrated to regulate susceptibility to pathogens of the Yersinia family [48], of which some members probably evolved in China [49], and thus we speculate that the insertion may have been selected in relation to pathogen resistance.

Notes

Mapping to the standard reference (the PGF MHC haplotype) results in large fluctuations in coverage and many poorly mapped reads (Fig. 1a). However, when the reference is augmented with an additional haplotype, identified by comparing the classical HLA genotypes of the sample with those of the eight reference haplotypes and noting that one of the eight haplotypes is a close match, read coverage and alignment are greatly improved (Fig. 1b,c).

1. Variance-aware reference structure: Pop VCF 1000 genomes.
2. Algorithms to match HTS seq data to (1) + Detect novel variation (new path?)
3. Project variation-aware ref onto primary sequence
4. Benchmarks

No genomic region/SNP/variant with high DDAF appears more than once in Colonna et al 2014 supplementary table 2.

References

Population variation

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-6-r88

Graph based representation of genomes

https://www.nature.com/ng/journal/v47/n6/full/ng.3257.html

Sankey Diagrams

https://www.r-bloggers.com/experimenting-with-sankey-diagrams-in-r-and-python/ https://www.r-bloggers.com/creating-custom-sankey-diagrams-using-r/

Google Drive Project

https://drive.google.com/open?id=0B-JZp8cuqCwMdURYNXo5a2NIM28

Github repo

https://github.com/nmferraro5/graph_genome_GENE245

DAF

https://www.biostars.org/p/128266/