Pool-seq uncertainty analysis

Z. Forrest Elkins

Simple pairwise comparisons

Simulate a pairwise comparison between alleles at a single location without error.

```
# load dependencies

library(ggplot2)

library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --

v tibble  3.1.8      v dplyr   1.0.10

v tidyr   1.2.1      v stringr 1.4.1

v readr   2.1.3      v forcats 0.5.2

v purrr   0.3.4

-- Conflicts ------------------------------------------ tidyverse_conflicts() --

x dplyr::filter() masks stats::filter()

x dplyr::lag()    masks stats::lag()
```

```r
set.seed(172452)


# establish pairwise parameters


## allele frequencies for high and low populations
popH <- 0.3
popL <- 0.09
## absolute difference in allele frequencies
diff <- abs(popH - popL)
## scale data logarithmically
logdiff <- -log10(diff)
## print pairwise comparison value -- 'true' value
logdiff
```

```
[1] 0.6777807
```

Estimating 'effective coverage'

A simulated calculation of coverage using the math detailed in Tilk et al. 2019. In other words, my own miniature HAF-pipe functions. Our variable cvg_e is the 'theoretical coverage at which binomial sampling of reads would be expected to contain the observed amount of error from estimated frequencies.' Put another way, the estimated and theoretical root mean squared error (RMSE) rates will equal each other under 'effective coverage.'

```r
# modulate pairwise comparison with coverage as a source of error


# n is the number of sites, which in our case is just one
n <- 1


## estimated and true allele frequencies at one site
AFtrue <- rbinom(100,100,0.38) / 100
AFest <- rbinom(100,100,0.43) / 100


## effective coverage -- Tilk et al. 2019
cvg_e <- sum(AFtrue * (1 - AFtrue)) / sum((AFest - AFtrue) ^ 2)
cvg_e
```

```
[1] 30.42092
```

```r
## theoretical root mean squared error
RMSEthe <- sqrt( sum(AFtrue * (1 - AFtrue)) / (cvg_e * n) )
RMSEthe
```

```
[1] 0.8772115
```

```r
## estimated RMSE
RMSEest <- sqrt( sum((AFest - AFtrue) ^ 2) / n)
RMSEest
```

```
[1] 0.8772115
```

In this case, with the code above, error introduced by variation in coverage would be evaluated via the magnitude of difference between theoretical and estimated RMSE values.

Tilk, Susanne, Alan Bergland, Aaron Goodman, Paul Schmidt, Dmitri Petrov, and Sharon Greenblum. 2019. "Accurate Allele Frequencies from Ultra-Low Coverage Pool-Seq Samples in Evolve-and-Resequence Experiments." *G3* 9 (12): 4159–68.

Coverage and pairwise allele frequency comparisons

Simulate a pairwise comparison between alleles at a single location with coverage error.

Simple plot of random allele frequencies

Pull allele frequencies from a distribution for 'low' and 'high' poolseq populations and plot them.

```
## the following code is from my professional website
expl_freqs <- readRDS('expl-freqs.Rds')
arm <- expl_freqs %>%
  filter(chrom=='2L') %>%
  slice(1:800)


D <- (sample(400:600,800,replace = TRUE))/1000
N <- (sample(300:600,800, replace = TRUE))/1000
```

```
sampledata <- tibble(
  "pos" = arm$pos,
  "N" = N,
  "D" = D
)


plt <- ggplot(data=sampledata, aes(pos/16)) +
  geom_line(aes(y = N, colour = "non-explorers")) +
  geom_line(aes(y = D, colour = "explorers"))
plt + labs(x = "Position (Mb)")
```

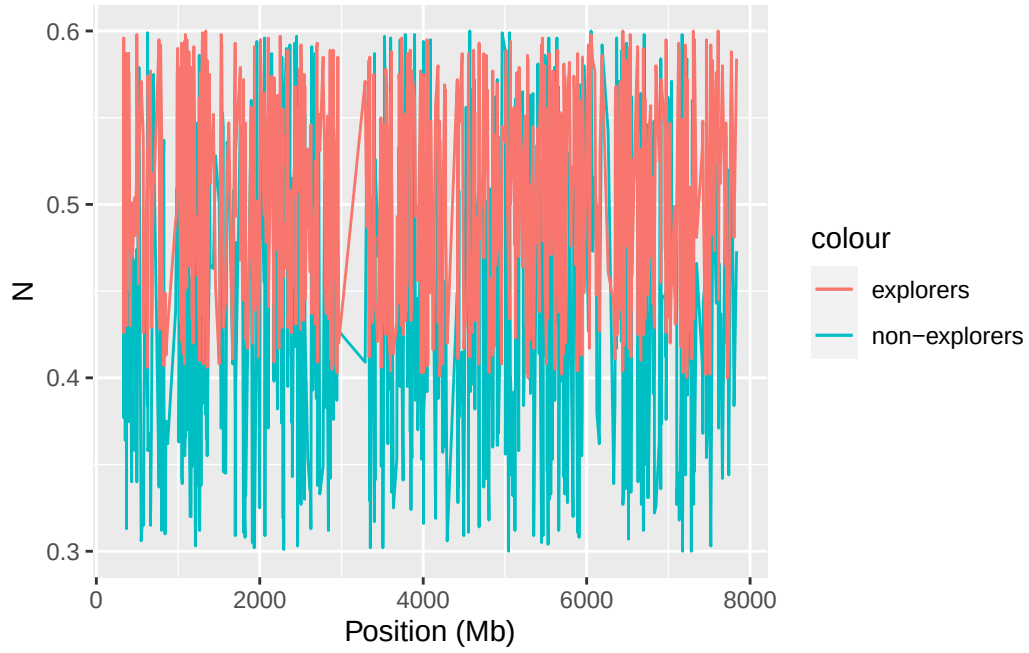

Figure 1: Allele differences between exploring and non-exploring flies.