Create structural variants from reference genome

Lucas Nell

24 March 2017

Contents

Updated 27 March 2017

Loading packages:

```
suppressPackageStartupMessages({
    library(magrittr)
    library(ggplot2)
    library(purrr)
    library(dplyr)
    library(ShortRead)
})
```

1 Data to on which to base simulations

Reference:

Bickel, R. D., J. P. Dunham, and J. A. Brisson. 2013. Widespread selection across coding and noncoding DNA in the pea aphid genome. *G3: Genes|Genomes|Genetics* **3**:993–1001. Available from http://www.g3journal.org/content/3/6/993

The main points are below (all quotes are from p 996):

- "We sequenced 21 genetically distinct lines of pea aphids. . . "
- ". . . we calculated F_{st} levels across the genome, comparing 11 pea aphid lines from the Northeast US (New York and Massachusetts) and 10 from California. We observed no structure, with an overall F_{st} value of -0.021. We conclude that pea aphid populations in the United States function as a single, panmictic population."
- "[θ_w and θ_π] for all sites across the genome were 0.0050 and 0.0045, respectively"

2 Initial information

From the paper's information above, we have. . .

```
theta_w <- 0.0050
theta_pi <- 0.0045
```

I'm going to simulate a sample size of 10. (The period is prepended to avoid conflicts with other object names.)

```
.n <- 10
```

3 Calculating parameters from paper data

The two main pieces of information I want for the calculations are (1) the proportion of segregating sites and (2) some measure of how different individuals are at segregating sites.

3.1 Segregating sites

Watterson's (1975) estimator (θ_w) is as follows:

$$\theta_w = \frac{K}{a_n}$$

where K is the proportion of segregating sites. Variable a_n is below:

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

where n is the number of individuals sampled. Thus the the proportion of segregating sites is simply $K = \theta_w a_n$.

For simplicity, I'm going to first make a function to compute a_n for a given value or values of n.

```
a_n <- function(n) {
    # Inner function to get a single "harmonic number"
    harm_n <- function(inner_n) {
        harm_n_vec <- 1 / 1:inner_n
        return(sum(harm_n_vec))
    }
    if (any((n %% 1) != 0)) stop("n must be entirely integers")
    sapply(n - 1, harm_n)
}
```

Now to compute the proportion of segregating sites for my chosen sample size of 10, I simply multiply θ_w and a_{10}.

```
(seg_sites <- theta_w * a_n(10))
```

```
## [1] 0.01414484
```

3.2 Diversity at segregating sites

Nei and Li's (1979) measure of nucleotide diversity, θ_π, is calculated using the following equation:

$$\theta_\pi = \sum_{ij} x_i x_j \pi_{ij}$$

Here, x_i and x_j represent the frequencies of the ith and jth unique sequences respectively and π_{ij} represents the proportion of divergent sequence between the ith and jth unique sequences.

If I assume that all lines will be unique sequences—a safe assumption if whole genomes are considered—then the above equation can be expressed as follows:

$$\theta_\pi = \sum_{ij} \frac{1}{n^2} \pi_{ij}$$

Then, since the number of total pairwise combinations between n sequences can be simplified to $\binom{n}{2}$...

$$\theta_\pi = \binom{n}{2} \frac{1}{n^2} \bar{\pi}$$

where $\bar{\pi}$ is the mean proportional sequence divergence between any two sequences. Solving for $\bar{\pi}$ yields the following:

$$\bar{\pi} = \frac{\theta_\pi n^2}{\binom{n}{2}}$$

Since I've already calculated the proportion of segregated sites, I want the mean divergence at segregated sites only. (This improves computational and coding efficiency because I only have to worry about segregating sites.) To do that, I divide θ_π by the proportion of segregated sites. This leaves me with the proportional nucleotide divergence between two sequences at segregating sites.

```
(seg_div <- (theta_pi / seg_sites) * .n^2 / choose(.n, 2))
```

```
## [1] 0.7069715
```

(See the `README.md` file for why I'm including `./genome_data/` in file paths.)