Digest genome

Lucas Nell

2017-03-21

Note: Installing `SimRAD` requires the following code:

```
source('https://bioconductor.org/biocLite.R')
biocLite('Biostrings')
biocLite('ShortRead')
biocLite('zlibbioc')
install.packages('SimRAD')
```

```
set.seed(699)
genome_seq <- ref.DNAseq('aphid_genome.fa.gz', prop.contigs = 0.1)
paste(substr(genome_seq, 1, 10), '...',
      substr(genome_seq, nchar(genome_seq) - 9, nchar(genome_seq)))
```

```
## [1] "GGTAGATCGC ... GGGATGTCAT"
```

```
re_df <- data_frame(enzyme = c('ApeKI', 'SbfI', 'PstI', 'EcoT22I', 'BstBI', 'AscI',
                               'BspEI', 'AclI', 'FspI', 'MluI-HF', 'NruI-HF'),
                    sites = list(c('G', 'CAGC', 'G', 'CTGC'), c('CCTGCA', 'GG'),
                                 c('CTGCA', 'G'), c('ATGCA', 'T'), c('TT', 'CGAA'),
                                 c('GG', 'CGCGCC'), c('T', 'CCGGA'), c('AA', 'CGTT'),
                                 c('TGC', 'GCA'), c('A', 'CGCGT'), c('TCG','CGA'))) %>%
    filter(!enzyme %in% c('SbfI', 'PstI', 'EcoT22I', 'AscI', 'BspEI', 'FspI'))
```

The restriction enzymes below were filtered (reasons follow):

- *SbfI*: Not blocked by CpG methylase
- *PstI*: Not blocked by CpG methylase
- *EcoT22I*: "... not sensitive to dam, dcm, or CG methylation" link
- *AscI*: "AscI is strongly inhibited by NaCl and ammonium acetate"
- *BspEI*: Only impaired by CpG methylase
- *FspI*: "Ligation is 25% -75%."

This function runs an in silico digestion on an enzyme, given its sites as a character vector:

```
digest_enzyme <- function(enzyme_sites, dna_seq = genome_seq) {
    if (is.list(enzyme_sites)) {
        enzyme_sites <- enzyme_sites[[1]]
    }
    names(enzyme_sites) <-
        c('cut_site_5prime1', 'cut_site_3prime1',
          'cut_site_5prime2', 'cut_site_3prime2',
          'cut_site_5prime3', 'cut_site_3prime3',
          'cut_site_5prime4', 'cut_site_3prime4')[1:length(enzyme_sites)]

    call_list <- as.list(c(DNAseq = dna_seq, verbose = FALSE, enzyme_sites))

    dig <- do.call(insilico.digest, call_list)

    return(dig)
}
```

Running that on all digestion enzymes in `re_df` (takes ~30 seconds):

```
re_df <- re_df %>%
    mutate(digest = lapply(sites, digest_enzyme))
```

> From six sequencing lanes, we identified 809,651 sequence tags (at least five times) from one or both flanks of 654,998 of the 2.1 million ApeKI cut sites lying within the single copy genomic fraction.

(Elshire et al. 2011, p 5)

From above, I've created below an object storing the proportion of cut sites that I'll assume get sequenced:

```
seq_p <- 654998 / 2.1e6
```

Printing summary of each digestion, where all numbers assume `seq_p` proportion of sites get sequenced:

```
## ---    ApeKI    ----
## Loci per Mbp = 181.06
## Total loci = 98,144
##
## ---    BstBI    ----
## Loci per Mbp = 78.82
## Total loci = 42,725
##
## ---    AclI    ----
## Loci per Mbp = 93.09
## Total loci = 50,463
##
## ---    MluI-HF    ----
## Loci per Mbp = 49.92
## Total loci = 27,061
##
## ---    NruI-HF    ----
## Loci per Mbp = 23.32
## Total loci = 12,638
```

References

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE* **6**:e19379.