contiBAIT: Improving Genome Assemblies Using Strand-seq Data

Kieran O'Neill, Mark Hills and Mike Gottlieb

December 16, 2015

koneill@bcgsc.ca

Contents

1 Licensing

Under the Two-Clause BSD License, you are free to use and redistribute this software.

2 Introduction

Strand-seq is a method for determining template strand inheritance in single cells. When strand-seq data are collected for many cells from the same organism, spatially close genomic regions show similar patterns of template strand inheritance. ContiBAIT allows users to leverage this property to carry out three tasks to improve draft genomes. Firstly, in assemblies made up entirely of contigs or scaffolds not yet assigned to chromosomes, these contigs can be clustered into chromosomes. Secondly, in assemblies wherein scaffolds have been assigned to chromosomes, but not yet placed on those chromosomes, those scaffolds can be placed in order relative to each other. Thirdly, for assemblies at the chromosome stage, where scaffolds are ordered and separated by many unbridged sequence gaps, the orientation of these sequence gaps can be found.

All three of these tasks can be run in parallel, taking contig-stage assemblies and ordering all fragments first to chromosomes, then within chromosomes while simultaneously determining the relative orientation of each fragment.

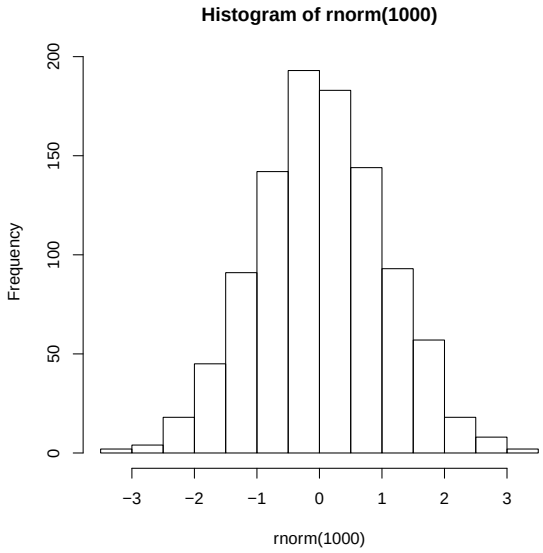
3 Input

ContiBAIT requires input in BAM format. Multiple BAM files are required for analysis, so ContiBAIT specifically calls for users to identify a BAM directory in which to analyse. BAM files should be sorted prior to analysis. To read in BAM files into ContiBAIT, create a strandFreqTable instance by calling strandSeqFreq

■strandSeqFreqTable.R■

4 Assigning contigs or scaffolds to chromosomes

```
> #Loading
> library(contiBAIT)
> plot(hist(rnorm(1000)))
```



5 Ordering contigs or scaffolds within chromosomes

First we load our artificial murine chromosome data. Some words about the meaning of linkage.group, animal.tab, and reorientedTable or whatever it's called.

```
> #   data(contigOrderingExample)
> #   reorientedTable <-  vignetteTestData[[1]]
> #   animal.tab <- vignetteTestData[[2]][[1]]
> #   linkage.group <- vignetteTestData[[3]][[1]]
> #   set.seed(666)
```

Next we sort the artificial mouse contigs using contiBAIT.

```
>   ordering <- 'ordering'
```

The resulting ordering can then be plotted.

6 Writing out to a BED file

This file can be passed to bedtools along with the original (draft) reference genome to create a new FASTA file containing the assembled genome.