Reid

Oct-1-14

A few hints for installing a non-model organism genome in BSgenomes for running CRISPseek with a non-model genome.

These instructions are for command line Linux, and should more-or-less work for Mac, and if you're clever enough, you should be able to get it to work on Windows.

Requirements:

PERL

R

libraries: CRISPRseek, BSgenome

1. To install CRISPRseek, start R as a superuser (sudo R) and enter:

source(http://bioconductor.org/biocLite.R)

biocLite("BSgenome")

biocLite("BiocGenerics")

biocLite("IRanges")

biocLite("GenomicRanges")

biocLite("Biostrings")

For the last four packages, they may be automatically pulled in when you install BSgenome, it won't hurt to reinstall them.

2. Download the tarball for CRISPRseek. Although it says that CRISPRseek can work on R >=3.0.1, I could not get it to run on R 3.0.2 (Frisbee Sailing). But it is simple enough to install as the tarball. Go to:

www.bioconductor.org/packages/release/bioc/html/CRISPRseek.html

Scroll down to the appropriate package. For Linux, take the Package Source <<CRISPRseek_1.0.3.tar.gz>

Move the tarball into an appropriate folder (doesn't really matter where) and start R as a superuser. Then type:

install.packages("CRISPRseek_1.0.3.tar.gz", type="source")

Exit R and continue on with building the non-model reference genome.

2. Download your genome from an appropriate resource. If you're not sure what you're looking for, select one that says "SCAFFOLDS" or "SUPERCONTIGS". It should be a large file, not one that says "CONTIG2SCAFFOLD" or "CONTIG2SUPERCONTIG", those should be smaller. Be sure to take the most recent version, which should be indicated by a higher number somewhere in the file. For example, I downloaded <<Aedes-aegypti-Liverpool_SCAFFOLDS_AaegL3.fa.gz>> from Vectorbase, where "SCAFFOLDS" represents the best sequence compression for this species and AaegL3 is the third version of the Liverpool strain of *Ae. aegypti* by Vectorbase.

3. Now that you've downloaded your genome, you need to also "copy link location". You can do this by right mouse clicking on the hyperlink and selecting "copy link location". Paste it into gedit, or notepad, or anywhere that you'll remember where it is.

4. Make a folder somewhere in your upper home directory to keep your databases convenient. I put all of mine in ~/.dbases/bioconductor. That way, I know where they are, but never have to see them.

5. Make a new directory in ~/.dbases/bioconductor called seqs_srcdir, and move your genome into this directory and unpackage it. For example:

mkdir ~/.dbases/bioconductor/seqs_srcdir && mv Aedes-aegypti-Liverpool_SCAFFOLDS_AaegL3.fa.gz ~/.dbases/bioconductor/seqs_srcdir/

cd ~/.dbases/bioconductor/seqs_srcdir && gunzip Aedes-aegypti-Liverpool_SCAFFOLDS_AaegL3.fa.gz

6. Now we have to *somewhat* reformat the fasta file. BSgenome seems like it will not build a genome with too many scaffolds (it is really meant for chromosomal data), but we can make a pseudo molecule separated by a string of "NNNNNNNNs" to compress to a single sequence so that R won't freak out. To do this, we first convert to a one sequence per line FASTA file. I have a short PERL script for this which makes it simple to run:

```
#!/usr/bin/perl –w

# This PERL script will convert a FASTA file to have only single line breaks for each sequence. Run as:

# fa2oneline.pl filein.fasta > fileout.fasta

use strict;

my $input_fasta=$ARGV[0];

open(IN,"<$input_fasta") || die ("Did not find FASTA file $input_fasta $!);
```

```
my $line = <IN>;

print $line;

while ($line = <IN>)

{

chomp $line;

if ($line=~m/^>/) {print "\n",$line,"\n";}

else { print $line; }

}

print "\n";
```

To run this PERL script, you can copy and paste it into your current directory (or a directory in your $PATH) and save as <<fa2oneline.pl>>.

Then:

```
chmod +x fa2oneline.pl
```

and to run do:

```
./fa2oneline.pl filein.fasta > fileout.fasta
```

or if you put it in a directory in your $PATH, use:

```
fa2oneline.pl filein.fasta > fileout.fasta
```

Now, let’s switch out the headers for NNNNNNNNNNs:

```
grep –v “>” fileout.fasta > seqs
```

```
awk ‘{print
“NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN”$1}’ seqs | sed
‘:a;N;$1ba:s/\n//g’ > oneline
```

Now we just have to add on a new header:

```
echo “>aedes” > newhead
```

```
cat newhead oneline >> aedes.fa
```

Notice that the extension is now “fa” rather than “fasta”. This seemed to be important.

And we can clean up a bit:

rm newhead oneline fileout.fasta seqs

9. Now, we can make the seed file:

Package: BSgenome.Aaegypti.VB.aa3

Title: Full genome sequences for Aedes aegypti version aa3 (VB version 3)

Description: Full genome sequences for Aedes aegypti (mosquito) as downloaded from Vectorbase (aa1, Oct. 2014)

Version: 3

organism: Aedes aegypt i

species: mosquito

provider: VB

provider_version: aa3

release_date: Oct. 2014

release_name: Vectorbase v3

source_url: https://www.vectorbase.org/download/aedes-aegypti-liverpoolscaffoldsaaegl3fagz

organism_biocview: Aedes_aegypti

BSgenomeObjname: Aaegypti

seqnames: paste ("aedes”)

PkgExamples: genome$chr1

seqs_srcdir:/home/reid/.dbases/bioconductor/seqs_srcdir

mseqnames: character(0)

nmask_per_seq: 0

Note that the field “seqnames” just says “aedes”, not “aedes.fa”. BSgenome will automatically add in the fa

When you prepare your file, follow this format exactly. For future builds, all I do is to exactly replace “Aaegypti” with “Astephensi” for example, and “aa3” with “as1”. Don’t get fancy. Also for some reason,

it seems like Bioconductor really only wants the first initial for the generic epithet, so just use one initial for it.

Save the file as:

BSgenome.Aaegypti.VB.aa3-seed

Where: BSgenome is always BSgenome, Aaegypti is your species, VB is your provider, and aa3 is your genome version.

10. Let's try to forge the genome for BSgenome now (taken from H. Pages "How to forge a BSgenome data package"):

sudo R

> library(BSgenome)
> forgeBSgenomeDataPkg("./BSgenome.Aaegypti.VB.aa3-seed")

This step will take a while to build. Once it is built, we have to go into check the DESCRIPTION file, and tweek it just a bit if it is dodgy.

cd ./BSgenome.Aaegypti.VB.aa3/
nano DESCRIPTION

make sure it says Version: 1.0.0 for the version. If it says "1", change it to "1.0.0".

11. Exit R and build the package:

sudo R CMD build BSgenome.Aaegypti.VB.aa3

It should build fine

11. Run a check on the package

sudo R CMD check BSgenome.Aaegypti.VB.aa3_1.0.0.tar.gz

It should go through and five a small error at the end, but seemed to be fine for downstream.

12. Install the genome on your computer for future usage:

sudo R

install.packages("./BSgenome.Aaegypti.VB.aa3_1.0.0.tar.gz")

13. Check your installation of your genome:

library(BSgenome)

installed.genomes()

This should list all of your locally-installed genomes:

“BSgenome.Aaegypti.VB.aa3” “BSgenome.Astephensi.VB.as1” etc….

You can also check for genomes that are available through BSgenomes with:

available.genomes()

These genomes are going to be better reconstructed than for locally-installed ones, but having a genome, even if it is in the form of thousands of supercontigs, is still pretty nice.

Good luck!