Pandaseq Tutorial Documentation

Release 0.0

Adina Howe

Aug 17, 2017

Contents

Contents:

CHAPTER 1

Merging paired-end Illumina reads with pandaseq

Introduction

This is a tutorial for working with overlapping reads in a metagenome and making the most of this library preparation.

As a note, in general, I do not suggest sequencing with overlapping read (unless you are sufficiently concerned about sequencing errors, e.g., 16S rRNA genes with low sampling). In particular, for assembly, overlapping paired-end reads are using your sampling depth to get at redundant information. You'd be better served for assembly to increase insert sizes and sequence more unique information (in my opinion). That being said, overlapping paired ends seem to be most of what we deal with, so I merge them to get the longest reads.

I use Josh Neufield's PandaSeq for two reasons: 1) I've met Josh and he is a capable and nice guy, with good documentation for this open-source tool 2) RDP's Jim Cole and Qiong Wang have recommended this tool to me from their testing

See the github repo:

https://github.com/neufeld/pandaseq

And these articles:: #. Andre P Masella, Andrea K Bartram, Jakub M Truszkowski, Daniel G Brown and Josh D Neufeld. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 2012, 13:31. http://www.biomedcentral.com/1471-2105/13/31 #. Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl. Acids Res. 41(Database issue):D633-D642; doi: 10.1093/nar/gkt1244 [PMID: 24288368]

The Tutorial

Installation

This tutorial is setup to run after running Trimmomatic, see this tutorial. After you do read trimming, you'll have 2 sets of files: se1, se2, pe1, and pe2. For merging paired end reads, you should use the "quality" trimmed paired read files (pe1 and pe2).

To install the latest and greatest:

```
git clone https://github.com/neufeld/pandaseq.git
```

And follow the directions on the repo:

```
sudo apt-get install build-essential libtool automake zlib1g-dev libbz2-dev pkg-config
```

And then make the command call globally accessible:

```
cp pandaseq /usr/local/bin/
```

Merging reads

To merge reads, we are going to want to run some version of:

```
pandaseq -f forward.fastq -r reverse.fastq
```

But I also like to take care of a few options (-F = fastq output, -d = logging options, and to save a log file):

```
pandaseq -F -f s1_pe -r s2_pe -d rbfkms -u unmerged_pandaseq.fa 2> pandastat.txt 1>␣
↪merged_pandaseq.fastq
```

And that’s it! The output will be a merged PE file of any overlapping P1 and P2 reads and the unmerged reads. You can concatentate these together (note that that the unmerged is FASTA not FASTQ, maybe that will be an option that is fixed later). If you’re assembling, I would go through your unmerged reads and separate them out into P1 and P2 reads again (see my hack). This way you can feed them into the assembler as pairs at least.

CHAPTER 2

Indices and tables

- genindex
- modindex
- search