User Manual
Team1: Amrit Baweja, Kevin Elaba, Rajee Ganesan & Ethan Gaskin

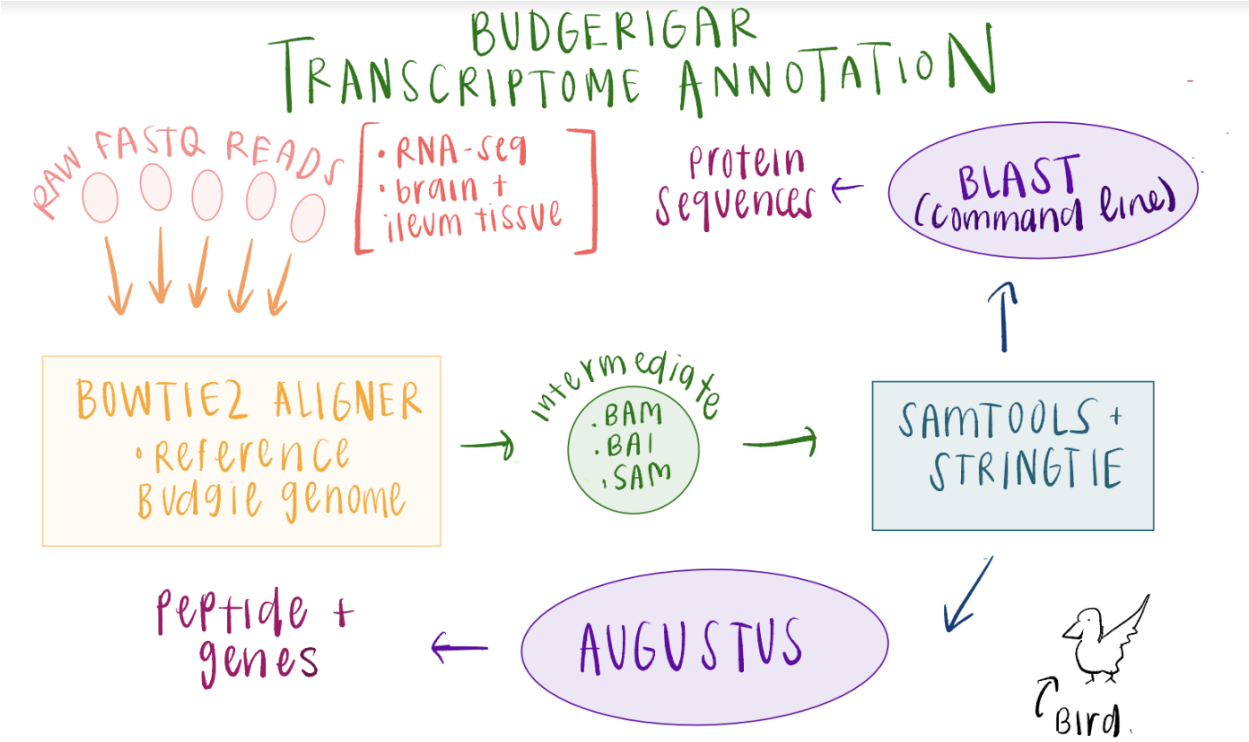
Table of Contents:

Introduction

About this Pipeline

This pipeline aims to make a transcriptome annotation of the common parakeet, budgerigar (Melopsittacus undulatus). Here, we utilize RNA-seq data as well as the reference genome to develop an annotation and search for proteins and regulatory elements, as well as evaluate the genetic connection of the budgerigar to other birds in similar clades or evolutionary regions.

Workflow



Briefly, we integrate multiple bioinformatics tools to generate a transcriptome annotation. The above figure showcases our workflow, where we use BowTie2, Samtools, StringTie, GFFRead, BBMap, BLAST, and Augustus.

Materials

You can download the budgerigar annotation and sample data with the following links:

- Budgerigar reference annotation: GCA_012275295.1
- RNA-seq of budgerigar brain tissue (Sample 1): SRX17566013
- RNA-seq of budgerigar brain tissue (Sample 2): SRX17566012
- RNA-seq of budgerigar brain tissue (Sample 3): SRX17566011
- RNA-seq of budgerigar brain tissue (Sample 4): SRX17566010
- RNA-seq of budgerigar brain tissue (Sample 5): SRX17566009

Packages

Our pipeline integrates the following packages. All version information is consolidated within a Anaconda environment, which can be reproduced with the ‘environment.yml’ file.

Anaconda
Anaconda is a package management system commonly used in bioinformatics workflows. Users can customize their Anaconda by selecting the packages they want to install, creating new environments, and managing package versions and dependencies. We use Anaconda to consolidate all of the packages listed below into one environment, in order to create a reproducible pipeline.

Bowtie2
Bowtie2 is a tool used for aligning high-throughput sequencing reads to a reference genome.

Input:
- A set of RNA-seq reads in fastq format
- A reference genome in fasta format

Output:
- A SAM file that contains alignment information for each read
- A BAM file, a compressed version of the SAM file

Samtools
Samtools is a tool that takes .sam and .bam files as inputs for post processing alignment. We use samtools to sort the aligned files, which are required before assembly.

StringTie
StringTie is a tool which assembles RNA-Seq alignments into potential transcripts. It takes in the BAM files as input and a reference genome and produces a gtf file for each sample.

GFFRead
GFFRead is a tool used for converting GFF and GTF files to other formats. We use it to convert the sample.gtf files into a fasta format.

BBMap
BBMap is a tool used for mapping high-throughput sequencing reads to a reference genome. We use it to convert sample.fasta files into fastq format.

BLAST
BLAST is a widely used bioinformatics tool for identifying homologous sequences across different organisms. We input a fasta file for each sample and return a tabular text file containing a list of hits, which are sequences that match the query with a

significant score. The hits are ranked on their E-value, a measure of statistical significance of the match.

Augustus
AUGUSTUS is a gene prediction program which can be used for ab-initio predictions based solely on the sequence. AUGUSTUS can also incorporate hints on the gene structure coming from extrinsic source like blast results and RNASeq data.

- Input: Alignments in fasta format
- Output: Predicted genes and information in gtf format.

Package Installation

The only package you would need to install manually is Anaconda. Then run the 'install.sh' script to initialize an Anaconda environment and other libraries, and you're good to go!

Running The Pipeline

Required Input Files
All reference files can be downloaded from ncbi:

1. Reference genome annotation gtf file: serves as reference annotation file for StringTie.
2. Reference genome fasta file: serves as reference genome file for GFFRead
3. RNA-seq fasta file: experimental RNA profiling to build our own transcript

Required Directory Structure
For ease-of-use, each sample.fasta should have its own directory, named the same as the sample. The corresponding .gtf, .fa, and .fq files will be written to their respective sample folders.

Visually, the directory structure should look like the following:

```
Working_directory:
        samples:
                sample_1:
                        sample_1.fasta
                sample_2:
                        sample_2.fasta
        reference_genome.gtf
        reference_genome.fasta
        environment.yml
        install.sh
        wrapper.sh
```

Usage
Please make sure you have Anaconda installed, and install all necessary libraries by using the ‘install.sh’ script provided. Once you have the following directory structure set up, you are good to go!

Command Line + Parameters
After you’ve installed all necessary libraries, run ‘bash wrapper.sh’ with the 5 following parameters:

- sample_dir: the directory where all of your samples are located, separated within their own directories
- ref_annotation: path to .gtf genome annotation file
- ref_fa: path to reference genome .fasta file
- blast_db: path to local blast rna-seq db
- blast_algo: path to local blast algorithm

Your results will be produced within each sample folder and within the ‘augustus_output’ and ‘blast_output’ directories. If you were to use the ‘installation.sh’ script provided in the tar file, your execution would look like this:

bash wrapper.sh samples/ genomic.gtf genomic.fa refseq_rna ncbi-blast-2.13.0+/bin/blastn

Understanding the Pipeline

Bowtie2 will generate a .sam file from the RNA-seq data and reference genome in the working directory.

Samtools is then used to create a sorted .bam file and a bam.bai file from the alignment result.

StringTie would generate a .gtf file from the alignment results provided in sorted .bam format, given a reference genome fasta file.

GFFRead would then convert the .gtf file into a .fasta format. **BBMap** would then convert the .fasta file into a .fastq format (which we did not end up needing).

BLAST would use the input fasta files in order to identify statistically significant homologous protein hits, outputted in a tabular .txt file.

Augustus would also use the fasta files generated from GFFRead to generate a .gtf file containing predicted genes as well as their corresponding protein sequences.

Limitations
The current pipeline is fit with reference data for the budgerigar, so running this pipeline with any non-budgerigar RNA-seq data would not yield strong results. You are welcome to download and change the reference data given to fit your working needs.