Amrit Baweja, Kevin Elaba, Rajee Ganesan & Ethan Gaskin

Joel McManus

03713: Bioinformatics Data Integration Practicum

May 4 2023

Course Final Report

Introduction

Most RNA-sequencing studies revolve around short-read sequencing, where RNA molecules are isolated and enriched (generally for mRNA), and then reverse transcribed into complementary DNA[1]. These cDNA sequences are then fragmented, randomly primed and amplified using PCR to yield an RNA-seq cDNA library processed by the sequencing instrument[1]. This then outputs millions of "short" reads which can be assembled into the sequences they originated from, a process known as transcriptome assembly[1].

In this process, the reads of the organism can be mapped to a reference genome to determine which genes the reads originated from, and to reconstruct the corresponding transcripts[1]. In the event there is no common reference, these reads can also be mapped de novo, where there is no external information available to guide the reconstruction and is done from inference based on the sequence and assembly gaps[1]. De novo assembled sequences are generally uninformative on their own, and must later be assigned human-readable identifiers and have their functional and evolutionary properties characterized to understand their biological relevance[1]. This is known as transcriptome annotation, to understand the functional connection of the sequences to phenotype expression.

For many recently sequenced organisms, there is no reference genome to map to, and many are sequenced using de novo transcriptome assembly. For this reason, many do not have widely used transcriptome annotations.

In this project, we plan on using public data to make a transcriptome annotation of the budgerigars (Melopsittacus undulatus), better known as the common parakeet, which are sister taxon to songbirds and one of the three order of birds that possess vocal learning abilities[1]. The species is thought to originate from the endemic Australian genus, Melopsittacus, and was once proposed to be a link between the Neophema and Pezoporus based on the barred plumage[2]. However, based on DNA analysis, the budgerigar has been placed closer to the lories and fig parrots[2]. The authors utilized RNA extracted from parrot brain and ileum tissue (59 separate samples) which were collected from LPS time scale experiment, and performed RNA-sequencing utilizing the Illumina HiSeq 2500 instrument (PRJNA879979).

We use this RNA-sequencing data as well as the reference genome to develop an annotation and search for proteins and regulatory elements, and provide insight into the function and biological process of transcripts and proteins they encode. This involved the measurement of almost all transcripts, including all mRNA and sequences within a tissue region. This, in turn, creates a global picture of cell function.

Materials

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

Methods

There are various tools that can be used for transcriptome functional annotation: the steps generally follow[1] homology transfer and identity assignment via sequence search, sequence feature annotation and then gene ontology (GO) and biochemical pathway annotation.

In order to initially process the raw RNA reads, the initial pipeline process performs quality control with alignment with Bowtie2 and produces a gene expression matrix (following analysis with samtools and bedtools/genome coverage). From there, we utilized GFFRead and BBMap to convert files and then ran BLAST and Augustus to create a gene prediction based solely on the sequence.

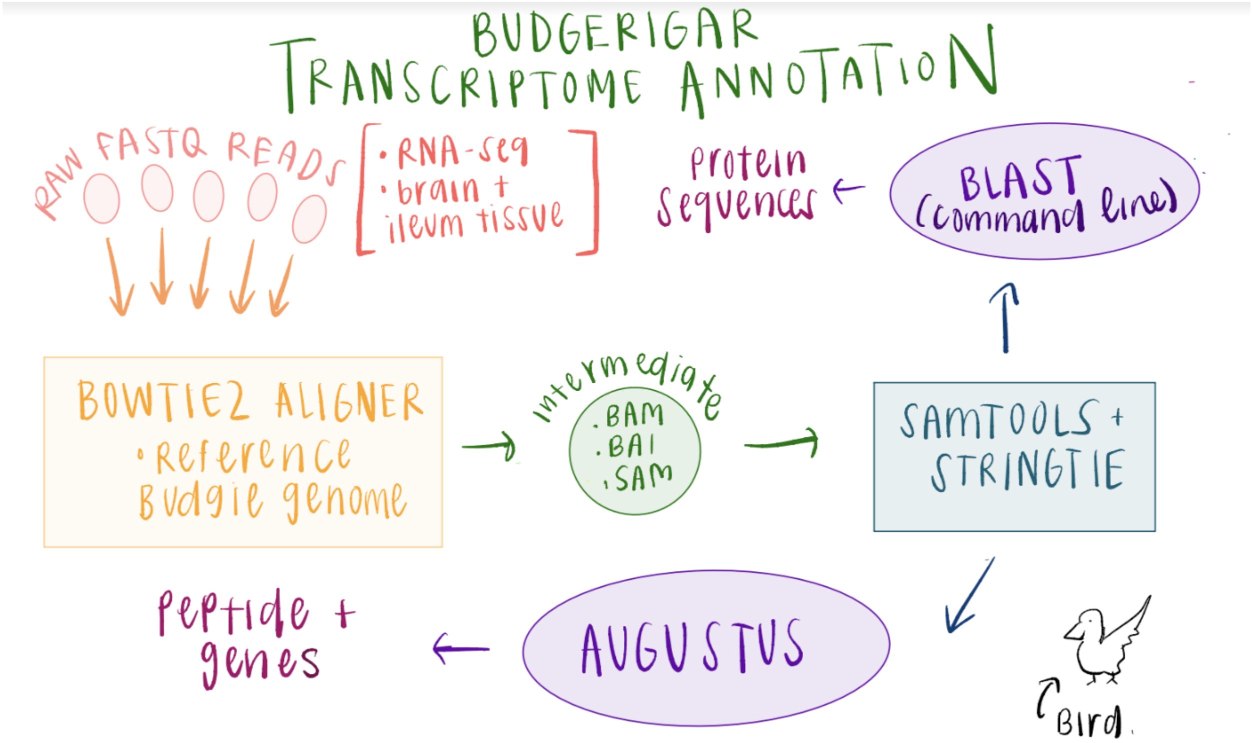


Figure 1. Sample flowchart of the pipeline's steps and goals.

Using the Pipeline

Our pipeline integrates the following packages. All version information is consolidated within an 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. Anaconda was utilized 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 is 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+ & Local refseq_rna Database

BLAST+ is a command line version of NCBI's BLAST software (including BLASTn, BLASTp, etc.). BLAST+ allows submission of FASTA files from the command line to the NCBI web server or to a locally installed BLAST database. To facilitate pipeline speed, we use a locally installed version of NCBI's refseq_rna database, limit queries to an e-value of 0.01 (our default), and use the csv output format. Note: downloading refseq_rna database uses ~45 GB of space and will take approximately 10 minutes.

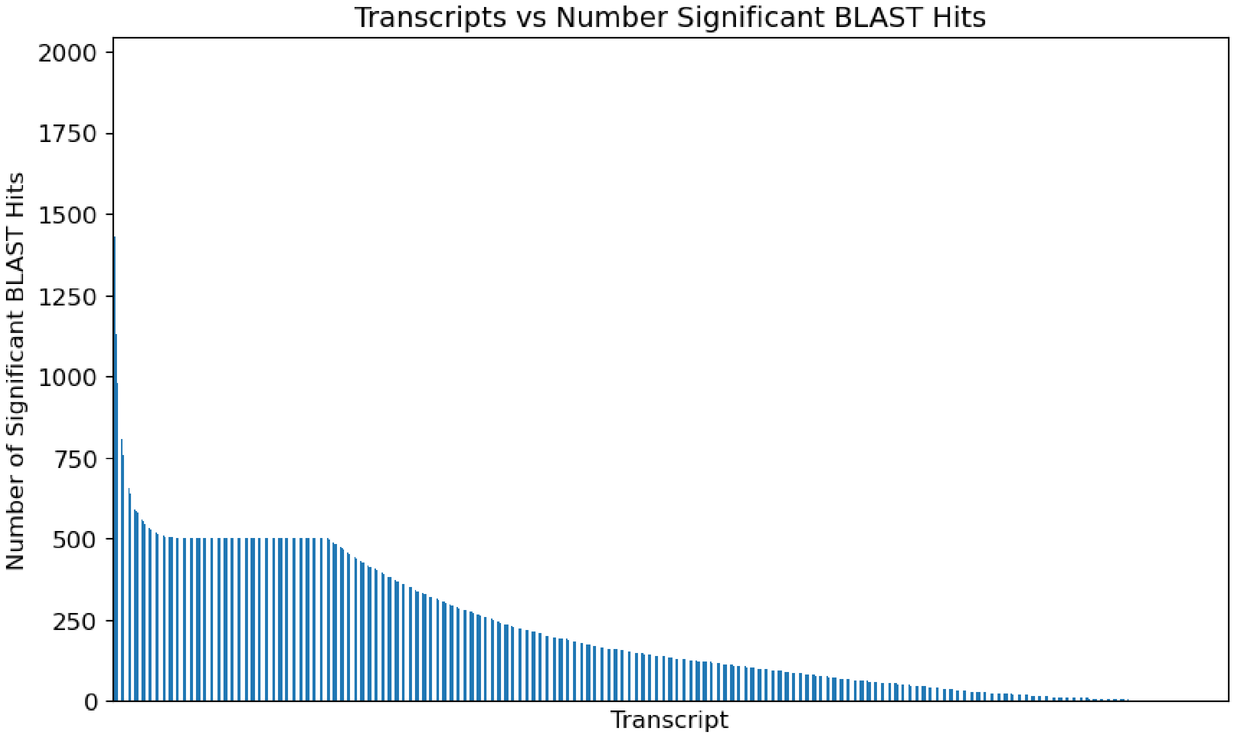
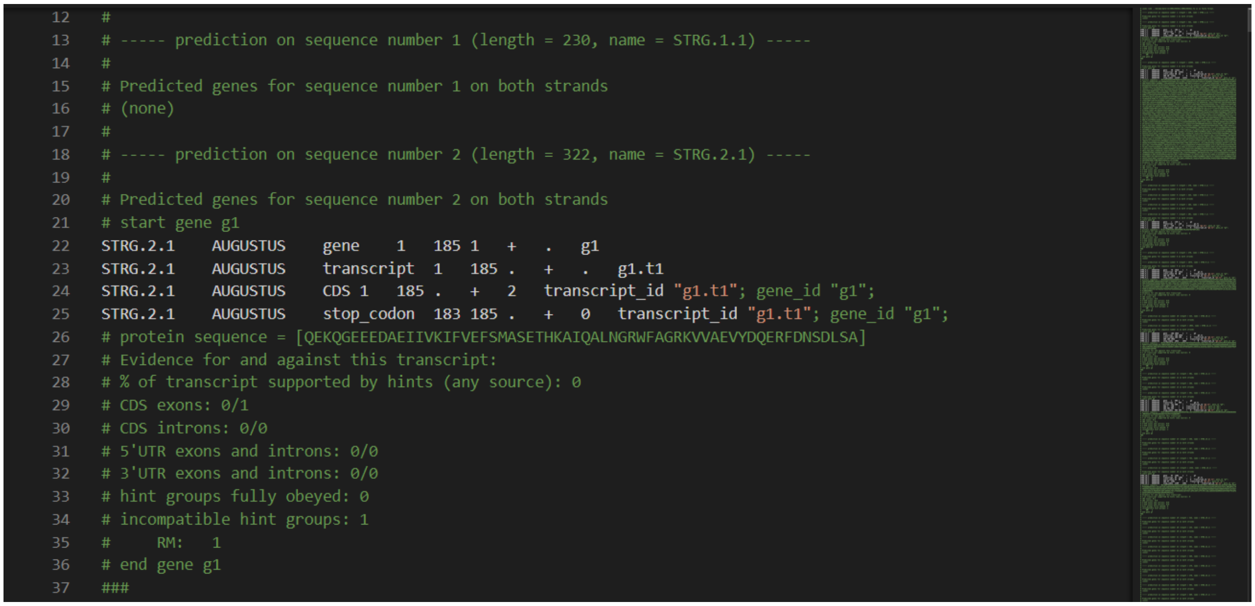


Figure 2. Results from our initial analysis identifying the transcripts with significant BLAST hits for the first sample (SRR21563621). Note: we omit transcript STRG.3.1, which had 13170 hits and was significantly skewing the plot.

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 sources like blast results and RNASeq data.

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



```
#
# ----- prediction on sequence number 1 (length = 230, name = STRG.1.1) -----
#
# Predicted genes for sequence number 1 on both strands
# (none)
#
# ----- prediction on sequence number 2 (length = 322, name = STRG.2.1) -----
#
# Predicted genes for sequence number 2 on both strands
# start gene g1
STRG.2.1    AUGUSTUS    gene    1   185 1   +   .   g1
STRG.2.1    AUGUSTUS    transcript  1   185 .   +   .   g1.t1
STRG.2.1    AUGUSTUS    CDS 1   185 .   +   2   transcript_id "g1.t1"; gene_id "g1";
STRG.2.1    AUGUSTUS    stop_codon  183 185 .   +   0   transcript_id "g1.t1"; gene_id "g1";
# protein sequence = [QEKQGEEEDAEIIVKIFVEFSMASETHKAIQALNGRWFAGRKVVAEVYDQERFDNSDLSA]
# Evidence for and against this transcript:
# % of transcript supported by hints (any source): 0
# CDS exons: 0/1
# CDS introns: 0/0
# 5'UTR exons and introns: 0/0
# 3'UTR exons and introns: 0/0
# hint groups fully obeyed: 0
# incompatible hint groups: 1
#     RM:   1
# end gene g1
###
```

Figure 3. Results from gene prediction in .gtf format for one of the samples (SRR21563621)

Required Input Files

All reference files were 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

Getting Started with 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 from the alignment result. **StringTie** would generate a .gtf file from the alignment results provided in .bam format, given a reference genome fasta file. **GFFRead** would then convert the .gtf file into a .fasta format, and then use **BBMap's** 'reformat.sh' script to convert the .fasta file into .fastq format. **BLAST** was then used to identify transcripts (outputted in .csv file using BLAST outfmt 10) for further analysis of proteins, and **Augustus** will generate a .gtf file containing predicted genes as well as their corresponding protein sequences.

Discussion

The final result of our pipeline identifies a final list of proteins, each with a corresponding score to the sequence transcripts inputted into the program, as well as a file identifying genes correlated with the same sequences.

There were over 9,700 identified transcript calls (following the StringTie application) for each sample, around 70,000 transcripts identified in total. For each of the samples, there were 9565, 17262, 16709, 13259 and 12860 transcripts with identified RNA-seq transcript matches and 877, 6048, 4029, 2821, and 2219 transcripts with no BLAST results, respectively. Parallelization allowed us to BLASTn all ~70,000 transcripts vs 52,197,929 refseq_rna sequences[9] (3.6 trillion alignments) in ~25 min. Figure 2 details the significant transcripts for the first sample. We also ran Augustus to identify the genes correlated with the transcripts, explored the annotations in gtf format (Figure 3), and we attempted to visualize the expression levels and changes utilizing the Integrated Genome Viewer (but had file format issues).

It would be interesting to make the pipeline more amenable to switching species (as of current, we don't have code to directly pull reference genomes for species other than the common parrot). Future studies, outside of developing visualizations of the results from our pipeline, may include running the pipeline on close relatives of the budgerigar, and evaluating the connection of the genomic makeup of the budgerigar to other birds in similar clades or evolutionary regions. Further analysis might look into gene ontology and pathway annotation, using the InterProScan tool to annotate GO terms, as well as utilizing the Kyoto Encyclopedia of Genes and Genomes database to evaluate sequences and their connections to biochemical pathways. One improvement for the BLAST portion of the code would be to help automate the visualizations. Currently, it's difficult to process the many hits (which we tried to limit with a strict threshold e-value of 0.01) in an informed manner given there are millions of hits for the set of transcripts associated with each sample. Another step we could take is running the transcripts with few/no hits on other softwares to help researchers explore novel reads which will appear in different birds or tissues within the same bird.

References

1. Venket Raghavan, Louis Kraft, Fantin Mesny, Linda Rigerte, A simple guide to de novo transcriptome assembly and annotation, Briefings in Bioinformatics, Volume 23, Issue 2, March 2022, bbab563, https://doi.org/10.1093/bib/bbab563
2. Schweizer M, Seehausen O, Güntert M, Hertwig ST. The evolutionary diversification of parrots supports a taxon pulse model with multiple trans-oceanic dispersal events and local radiations. Mol Phylogenet Evol. 2010 Mar;54(3):984-94. doi: 10.1016/j.ympev.2009.08.021. Epub 2009 Aug 21. PMID: 19699808.
3. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W20-5. doi: 10.1093/nar/gkh435. PMID: 15215342; PMCID: PMC441573.
4. Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, Kanae Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D353–D361, https://doi.org/10.1093/nar/gkw1092
5. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9. [abstract | full text]
6. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.
7. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. Nature 587, 240–245 (2020). https://doi.org/10.1038/s41586-020-2876-6
8. Harshil Patel, Phil Ewels, Alexander Peltzer, Olga Botvinnik, Gregor Sturm, Denis Moreno, Pranathi Vemuri, silviamorins, Maxime U Garcia, Lorena Pantano, Mahesh Binzer-Panchal, nf-core bot, Matthias Zepper, Robert Syme, Gavin Kelly, Friederike Hanssen, James A. Fellows Yates, Chris Cheshire, rfenouil, … Sven F. (2023). nf-core/rnaseq: nf-core/rnaseq v3.11.1 - Plastered Radium Rhino (3.11.1). Zenodo. https://doi.org/10.5281/zenodo.7789554
9. NCBI Database: RNA-RefSeq. https://ftp.ncbi.nlm.nih.gov/blast/db/refseq_rna-nucl-metadata.json https://www.ncbi.nlm.nih.gov/refseq/statistics/