RNA-Seq Differential Analysis Pipeline Summary

Contents

List of Figures

This is a brief summary of the pipeline and tools I developed during my time in the MS program at UW -- Milwaukee.

I have 40 years of programming experience, 30 years in Unix systems management, and 20 years in scientific computing support, but my direct involvement in biology research is minimal, so I would love to hear from others about their experiences working with sequence data.

1 Required Skills

A major goal of my research is to make DE analysis accessible to a broader population of research biologists. Most open source bioinformatics software is designed to run on Unix-compatible systems, and most only offers and command-line interface (as opposed to a graphical user interface (GUI) or other menu-based interface). Hence, bioinformaticians must know how to use the Unix command-line at minimum.

Some software requires the user to develop and maintain other skills as well, such as R (statistical language) programming. For this pipeline and the tools that comprise it, only basic Unix command-line skills are necessary. No Perl, Python, or R programming is required to use this pipeline or adapt it to other studies.

2 Computing Resources

2.1 Do I Need a Cluster?

Use of a cluster is not necessary for most differential expression analyses. A powerful workstation with several cores and enough RAM can complete a a typical analysis in a reasonable amount of time. In fact, researchers are likely to spend more time learning HPC and waiting for help from support staff than it would take to do the analysis on their own on a decent workstation. There are some activities that are impractical without a cluster, such as de novo genome assembly, but most bioinformatics can be done on modern laptops and workstations.

This repository contains both SLURM scheduler scripts for running on a cluster and scripts for running on a standalone Unix machine using **xargs** to run jobs in parallel. The **xargs** scripts are tested primarily on a Mac Mini m1.

2.2 Mac

Apple's operating system, macOS, has been Unix compatible since the release of macOS X (macOS 10) in 2001. MacOS 10 is derived largely from FreeBSD and some earlier BSD systems, which have ties to the original Unix system created by AT&T. A Mac is the most realistic option for many biologists to gain access to a Unix environment, for several reasons:

- A Mac is by far the easiest Unix-compatible system for the average (non I.T.) person to maintain.
- Mac is the Unix system most likely to be well-supported by central I.T. departments. If your computer will be owned and managed by your organization, you will probably get better support for a Mac than for any other Unix-like system.
- Mac is the only Unix system supported by a wide variety of common commercial software products, such as MS Office, etc. Hence, using a Mac means you won't need two computers or virtual machines to run both commercial software and open source bioinformatics software.
- "Apple Silicon", Apple's ARM processor based systems are extremely fast and power-efficient compared to most PC hardware. An x86 (Intel or AMD) base PC comparable to a high-end Mac Studio would be very noisy due to the need for powerful cooling fans. You wouldn't want one in your office while it's running a large analysis.

Apple silicon (Apple's ARM-based system with unified memory are remarkably fast and power-efficient. A Mac Studio with 64 GiB of RAM will suffice for most analyses. 128 GiB will make it even less likely that you'll have to occasionally go elsewhere. Note that the RAM cannot be upgraded on Apple silicon systems, so don't skimp in order to save a few hundred dollars. Most bioinformatics software only utilizes CPUs, to don't worry too much about the GPU and Neural Engine specs. You will also likely need an external disk to accommodate all your data. External disks are risky, since they can be unplugged while in-use, but there is no better option for compact Mac systems.

Note You will want to disable the screensaver under System Settings -> Lock Screen. Some macOS screensavers use enormous amounts of memory and CPU, competing with scientific analysis software. If your system runs out of memory due to competition from the screensaver, an analysis that should take hours could end up taking days as the system "thrashes", frantically moving data in and out of swap space on the disk.

2.3 Remote Access to Macs

For remote access to a Mac via SSH, enable "Remote Login" under "Sharing" in System Preferences.

To prevent disruptive disconnections (even when jobs are running!), Add the following to `/etc/ssh/sshd_config`:

```
ClientAliveInterval 30
ClientAliveCountMax 0
```

You may prefer to enable remote access via "Screen Sharing". Enable VNC under the Screen Sharing options to enable access from non-Macs. Unlike SSH, programs run under Screen Sharing will not be terminated if you are disconnected for any reason.

2.4 Open Source Unix

There are multiple BSD and Linux operating systems suitable for bioinformatics analysis. If you are tech-savvy enough to manage your own Unix system, you can get probably get by for most analyses with a used PC for as little as $500. You'll want something with at least 8 cores and 128 GiB RAM, as mentioned in the Mac section above.

A PC running a POSIX OS such as BSD or Linux will be cheaper and more upgradable than a Mac, and can house more internal disk storage. My personal choice is FreeBSD with at least 2 disks configured as a ZFS RAID. The FreeBSD installer makes it trivial to configure ZFS RAIDs and boot from them directly. The desktop-installer (https://acadix.biz/desktop-installer.php) package makes it easy to set up a graphical desktop environment of your choice. Also, on FreeBSD, all the software needed for this pipeline can be installed by running **pkg install rna-seq**. See Section 3 for more information.

GhostBSD is a FreeBSD derivative with a fully graphical installer and management tools, similar to Debian and Ubuntu Linux. GhostBSD uses FreeBSD ports and packages, but from its own repository which lags behind FreeBSD by a month or more. If you don't need the very latest packages (most people do not), then GhostBSD is a good choice for those who are less tech-savvy.

Any Linux distribution can be used along with pkgsrc (see Section 3) to install the tools needed for this pipeline. Most of them may also be available in the native Linux package manager, though they will need to be installed individually, as the rna-seq meta-package only exists in FreeBSD ports and pkgsrc. Debian Linux, and its derivatives such as Ubuntu, are popular and easy to install and manage. Redhat Enterprise Linux (RHEL) and its derivatives are used on most HPC clusters. They are more stable and support some commercial software, but use older Linux components which may not work with the latest open source software.

2.5 Virtual Machines including Cloud Services

Virtual machines may be more convenient for some users, depending on available I.T. support, funding model, etc. Pay-as-you-go services like cloud computing can be problematic in academic research, where funding is fixed and must be known ahead of time. Using cloud services also requires some special skills. Discuss this with your I.T. staff to find out whether this is a good option for you.

Note that virtualized network interfaces and storage, including those used by Amazon EC2, Google Cloud Platform, Microsoft Azure, and other cloud-based computing platforms, will not perform nearly as well as "bare metal". Only the CPU and memory should be expected to match the performance of real hardware, because these are the only components that typically don't use any software emulation on top of the actual hardware. However, this performance hit may be an acceptable price to pay to be free from the need to maintain your own hardware. Look for a platform that is fast enough to complete your analysis, not necessarily the fastest possible.

3 Software Installation

All commands needed for this analysis are assumed to be in your PATH, meaning you can just run the commands without doing any special preparation like loading virtual environments, containers, etc. If programs are installed in non-standard directories, you will need to update your PATH (e.g. by loading an environment module if they are used at your site).

All programs necessary for this pipeline can be installed on FreeBSD using **pkg install rna-seq**.

On any POSIX platform (BSD, Linux, macOS, SunOS, etc.), they can be installed using the **pkgsrc** rna-seq package (http://pkgsrc.org/-). Pkgsrc does not require administrator privileges, so it can be used on computers managed by your organization without submitting help requests.

The FreeBSD port and pkgsrc package install all tools to the same directory structure, following the filesystem hierarchy standard. Hence, you need only add 1 directory to your PATH at most to gain access to all of the software. The FreeBSD port installs under `/usr/local`, which is already part of the standard PATH, so no adjustments are necessary on FreeBSD.

Pkgsrc can be set up in about 10 minutes using **auto-pkgsrc-setup** (http://netbsd.org/~bacon/). Binary packages are available for some platforms (NetBSD, macOS, RHEL Linux). In this case, we can use **pkgin install rna-seq**. On most platforms, pkgsrc users install from source using **cd (prefix)/biology/rna-seq; bmake install**.

This repository includes a script that installs the software using FreeBSD ports or pkgsrc.

Programs used in this pipeline (and included in the rna-seq meta-packages) are described in the sections that follow.

4 Recompression

Recompressing raw data from gzip format to xz format generally reduces disk usage by about 40%. This reduced xenopus raw reads from 143 GB to 92 GB.

Compression with **xz** at standard compression levels is time-consuming, so I only do this for long-term data, not for intermediate results. Decompression with **unxz** is very fast, so reading xz-compressed files does not slow down the analysis.

For intermediate output files, I generally use **zstd**. This tool has pretty much obsoleted **gzip**, since it is both faster and produces better compression ratios.

All of the tools I developed can directly read and write files compressed with **gzip**, **bzip2**, **xz**, **lz4**, and **zstd**.

5 Raw File Aliases

Sequencing centers may require ridiculously compact filenames for the sequence files (are the running MS-DOS??). To avoid confusion and mistakes caused by cryptic filenames, I first create symbolic links with more descriptive names that are both easier for people to read and easier for shell scripts to parse.

For this study, I also reduced the time points to ranks 1 through 5. The axolotl and xenopus data have the same number of samples and time points, to standardizing the filenames allows the exact same analysis scripts to be used for both data sets. This saves a lot of time on bug fixes and documentation.

```
Symbolic link                       Raw filename from sequencing center
-------------------------------------------------------------------------
sample01-time1-rep1-R1.fastq.xz@ -> ../../../Raw/X1NaA/X1NaA_1.fq.xz
sample01-time1-rep1-R2.fastq.xz@ -> ../../../Raw/X1NaA/X1NaA_2.fq.xz
sample02-time1-rep2-R1.fastq.xz@ -> ../../../Raw/X2NaB/X2NaB_1.fq.xz
sample02-time1-rep2-R2.fastq.xz@ -> ../../../Raw/X2NaB/X2NaB_2.fq.xz
sample03-time1-rep3-R1.fastq.xz@ -> ../../../Raw/X3NaC/X3NaC_1.fq.xz
sample03-time1-rep3-R2.fastq.xz@ -> ../../../Raw/X3NaC/X3NaC_2.fq.xz
sample04-time2-rep1-R1.fastq.xz@ -> ../../../Raw/X4d7A/X4d7A_1.fq.xz
sample04-time2-rep1-R2.fastq.xz@ -> ../../../Raw/X4d7A/X4d7A_2.fq.xz
sample05-time2-rep2-R1.fastq.xz@ -> ../../../Raw/X5d7B/X5d7B_1.fq.xz
sample05-time2-rep2-R2.fastq.xz@ -> ../../../Raw/X5d7B/X5d7B_2.fq.xz
sample06-time2-rep3-R1.fastq.xz@ -> ../../../Raw/X6d7C/X6d7C_1.fq.xz
```

```
sample06-time2-rep3-R2.fastq.xz@ -> ../../../Raw/X6d7C/X6d7C_2.fq.xz
sample07-time3-rep1-R1.fastq.xz@ -> ../../../Raw/X7d12A/X7d12A_1.fq.xz
sample07-time3-rep1-R2.fastq.xz@ -> ../../../Raw/X7d12A/X7d12A_2.fq.xz
sample08-time3-rep2-R1.fastq.xz@ -> ../../../Raw/X8d12B/X8d12B_1.fq.xz
sample08-time3-rep2-R2.fastq.xz@ -> ../../../Raw/X8d12B/X8d12B_2.fq.xz
sample09-time3-rep3-R1.fastq.xz@ -> ../../../Raw/X9d12C/X9d12C_1.fq.xz
sample09-time3-rep3-R2.fastq.xz@ -> ../../../Raw/X9d12C/X9d12C_2.fq.xz
sample10-time4-rep1-R1.fastq.xz@ -> ../../../Raw/X10d18A/X10d18A_1.fq.xz
sample10-time4-rep1-R2.fastq.xz@ -> ../../../Raw/X10d18A/X10d18A_2.fq.xz
sample11-time4-rep2-R1.fastq.xz@ -> ../../../Raw/X11d18B/X11d18B_1.fq.xz
sample11-time4-rep2-R2.fastq.xz@ -> ../../../Raw/X11d18B/X11d18B_2.fq.xz
sample12-time4-rep3-R1.fastq.xz@ -> ../../../Raw/X12d18C/X12d18C_1.fq.xz
sample12-time4-rep3-R2.fastq.xz@ -> ../../../Raw/X12d18C/X12d18C_2.fq.xz
sample13-time5-rep1-R1.fastq.xz@ -> ../../../Raw/X13d27A/X13d27A_1.fq.xz
sample13-time5-rep1-R2.fastq.xz@ -> ../../../Raw/X13d27A/X13d27A_2.fq.xz
sample14-time5-rep2-R1.fastq.xz@ -> ../../../Raw/X14d27B/X14d27B_1.fq.xz
sample14-time5-rep2-R2.fastq.xz@ -> ../../../Raw/X14d27B/X14d27B_2.fq.xz
sample15-time5-rep3-R1.fastq.xz@ -> ../../../Raw/X15d27C/X15d27C_1.fq.xz
sample15-time5-rep3-R2.fastq.xz@ -> ../../../Raw/X15d27C/X15d27C_2.fq.xz
```

6 Pre-trim Quality Checks

The first analysis step is to run FastQC on the raw files, and MultiQC to combine the results for all samples into one interactive HTML report. A couple of example plots from the MultiQC report are shown here.

The read quality plot (axolotl data) shows that almost all of the reads are very high quality (A phred score of 20 means a 1/100 chance of a read error, 30 means a 1/1000 chance of error).

Figure 1: Read Quality

The adapter content plot shows that up to 2.5% of reads in some samples have adapter contamination near the 3' end.

Figure 2: Adapter Content

We can also run tools such as **blt fastx-stats** (from biolibc-tools (https://github.com/auerlab/biolibc-tools), which is installed by the rna-seq meta-packages. Biolibc-tools is a collection of simple tools I created to provide permanent solutions to numerous simple problems that often present a nuisance to researchers. It can be thought of as "putty" to fill in small gaps in typical analyses.

```
# blt fastx-stats Results/01-organize/Raw-renamed/sample01-day00-rep1-R1.fastq.xz
Filename:           Results/01-organize/Raw-renamed/sample01-day00-rep1-R1.fastq.gz
Sequences:          57943764
Bases:              8691564600
Mean-length:        150.00
Standard-deviation: 0.00
Min-length:         150
Max-length:         150
A:                  2478136460 (28.51%)
C:                  1873856390 (21.56%)
G:                  1887781386 (21.72%)
T:                  2451663620 (28.21%)
N:                  126744 (0.00%)
GC:                 3761637776 (43.28%)
```

7 Trimming

Trimming is not really necessary for RNA-Seq differential analysis, since the aligners will deal with adapter contamination, poly-A tails, etc. (Liao, 2020, doi: 10.1093/nargab/lqaa068). However, trimming beforehand reduces the workload for subsequent analysis stages, and allows them to produce cleaner results.

While I was perfectly happy with **cutadapt**, I wrote my own tool for this, called **fastq-trim**, (https://github.com/auerlab/fastq-trim) just to see how much improvement was possible over a quality existing tool. As shown on the website, **fastq-trim** is about 2.5 times as fast as cutadapt and 4 times as fast as Trimmomatic, while also using far less memory than either of them.

It uses algorithms very similar to cutadapt and a more efficient implementation in C. Testing on a sample of 100,000 reads produced only 6 trimmed sequences that differed slightly from the cutadapt output.

Trimming the xenopus data took 1.5 hours using 15 cores (5 jobs at a time, 3 cores per job) on my personal cluster (10-year-old Dell PowerEdge R415 servers). Average time to trim an individual file was about 30 minutes. It also ran in about 1.5 hours on my Mac Mini M1, using 4 jobs at a time under xargs.

Output from one sample of xenopus data is below. Results for all other samples were similar.

```
*** FASTQ TRIM ***

  Minimum match:     3
  Minimum quality:   24
  Minimum length:    30
  Phred base:        33
  Adapter matching:  Smart
  Maximum mismatch:  10%
  Filename:          Results/01-organize/Raw-renamed/sample02-time1-rep2-R1.fastq.xz
  Filename:          Results/01-organize/Raw-renamed/sample02-time1-rep2-R2.fastq.xz
  Mode:              Paired
  Adapters:          AGATCGGAAGAG AGATCGGAAGAG



Reads:                             107099288
Reads with adapters:                 5150891 (4%)
Reads with Poly-As:                   349808 (0%)
Bases with Q < 24:                  60688727 (0%)
Reads with low Q bases removed:      5173865 (4%)
Reads < 30 bases after trimming:      124850 (0%)
Mean adapter position:                   134
Mean read length:                        150

Run time: 1337.33 real      1621.45 user       137.50 sys
```

8 Post-trim Quality Check

Running FastQC and MultiQC on the trimmed data should show improvements in many metrics. The plot below shows that most adapter content has been removed. FastQC may flag some sequences as adapters that the trimming tool wasn't even looking for, so don't expect to find zero contamination after trimming. And again, removing adapter contamination isn't even necessary for RNA-Seq differential analyses. We trim merely to get rid of most contamination in order to improve the efficiency of subsequent stages.

Figure 3: Adapter Contamination after Trimming

9 Reference Transcriptome and/or Genome

I align reads to both a transcriptome (using **kallisto**) and a genome (using **hisat2**), and then compare the results for validation.

Aligning to the genome also enables discovery of novel transcripts.

For the first run on the axolotl reads, I used transcriptome `AmexT_v47_cds.fa` and genome `AmexG_v6.0-DD.fa` from http://www.axolotl-omics.org as references.

For well-annotated organisms, we can also create a transcriptome using **gffread**, a GTF or GFF3 file, and a genome FASTA. Generating from the GTF/GFF3 ensures that all annotated sequences present in GTF/GFF3 are also present in the transcriptome. This is not always the case for downloaded cDNA transcriptomes.

10 Alignment to a Transcriptome

I use **kallisto** for this step. It is extremely fast (if not generating pseudobams or bootstrap estimates, which are usually not necessary). Kallisto does not generate SAM (sequence alignment map, or BAM, a compressed equivalent) files, like most aligners to a genome. Creating pseudobams was an option prior to kallisto 0.50.0. Now, we need another aligner in order to generate SAM/BAM files.

Kallisto also performs quantification, producing raw read counts for each transcript (used by differential analysis tools) and TPM (transcripts per million) for human-consumption. A sample of **kallisto** output is shown below.

```
target_id               length  eff_length  est_counts  tpm
AMEX60DD201000002.1;    2001    1737.38     97.1699     0.798628
AMEX60DD201000003.1;    2508    2244.38     250.115     1.59129
AMEX60DD301000003.2;    1623    1359.38     0           0
AMEX60DD201000004.1;    1692    1428.38     90.0984     0.900701
AMEX60DD201000005.2;    357     101.384     19          2.67603
```

The "eff_length" and "est_counts" columns are used by DA tools such as DESeq2, EdgeR, FASDA, and Sleuth, to compute fold-changes and P-values. TPM, RPKM, and FPKM are reductions of read count information that not useful to DA tools, and are mainly used to give readers a rough idea about mRNA abundance.

11 Quantification

Quantification is the process of counting the reads aligned to each feature of interest (genes or transcripts for RNA-Seq) and estimating relative mRNA abundance from these data. Normally, this is a separate step, but as mentioned above, **kallisto** does this for us automatically during the alignment process.

12 Differential Analysis

This is the most problematic stage of the analysis. Most existing tools (DESeq2, EdgeR, Sleuth, etc.) require the user to write R scripts that read the alignment / quantification output (which differs for different aligners), manipulate it into complex R data structures (called data frames), and use R library functions to generate the fold-changes and P-values used to identify significant changes in expression.

Installing and maintaining R packages is also notoriously problematic. The installations using **install.packages()** and **BiocManager::install()** often fail, and will almost certainly cease to function a few months down the road due to routine updates to the underlying OS and packages installed outside of R.

In addition to technical issues with R, the results very greatly across different DA tools. Note that 3 samples are not nearly enough to achieve satisfactory statistical power in any setting. Many DA tools use sophisticated statistical tricks to increase power at the cost of higher false discovery rates. From Li, et al:

“ DESeq2 and edgeR had large discrepancies in the DEGs they identified on these datasets (Additional file 1: Fig. S1). In particular, 23.71% to 75% of the DEGs identified by DESeq2 were missed by edgeR. The most surprising result is from an immunotherapy dataset (including 51 pre-nivolumab and 58 on-nivolumab anti-PD-1 therapy patients) [8]: DESeq2 and edgeR had only an 8% overlap in the DEGs they identified (DESeq2 and edgeR identified 144 and 319 DEGs, respectively, with a union of 427 DEGs but only 36 DEGs in common). This phenomenon raised a critical question: did DESeq2 and edgeR reliably control their false discovery rates (FDRs) to the target 5% on this dataset? ”

Lastly, R is an interpreted language, which is roughly 100 times slower than fully compiled languages such as C, C++, or Fortran. This is not a problem for differential expression analyses using a small number of replicates, but it’s a big problem for population studies (Li, et al, 2022, https://doi.org/10.1186/s13059-022-02648-4).

Li, et al also reported poor performance for large sample sizes (the performance issue is not relevant to our 3-sample study).

To alleviate many of these issues, I developed a new differential analysis tool called FASDA (https://github.com/auerlab/fasda), which is written entirely in C, is easily installed and updated using highly-evolved and reliable package managers, and is installed by rna-seq meta-packages mentioned above.

It processes kallisto output directly, so there is no need for R programming or knowledge of data frames. It computes *exact* P-values for low sample counts, and uses the Mann-Whitney U test (a.k.a. Wilcoxon rank-sum test) for sample counts of 8 or more. Both of these methods control false discovery.

Note Exact P-values are computed by literally generating all possible combinations of read counts and counting combinations with fold-changes as least as extreme as observed. Hence, such P-values are not estimates, like those produced by other tools.

Processing **kallisto** output with **fasda** requires only two commands, e.g.:

```
# Generate normalized counts from all samples under the same condition
# This uses the median-ratios normalization (MRN) method
fasda normalize --output normalized-counts-condition1.tsv \
    sample*-condition1-rep*/abundance.tsv

# Compute fold-changes and exact or Mann-Whitney P-values
fasda fold-change --output FCs-c1-c2.txt \
    normalized-counts-condition1.tsv \
    normalized-counts-condition2.tsv
```

Example output is shown below (MNC = mean normalized count, SD/C1 = standard deviation for MNC1 / MNC1, %Agr = how many samples agree on the direction of the change, FC = fold-change, P-val = exact P-values).

```
Feature                  MNC1    MNC2  SD/C1  SD/C2  %Agr  FC 1-2  P-val
AMEX60DD201000002.1;     99.1   102.2    0.2    0.3    67    1.03  0.79163
AMEX60DD201000003.1;    327.7   445.6    0.3    0.2    67    1.36  0.19409
AMEX60DD301000003.2;      0.0     0.0    0.0    0.0   100       *  1.00000
AMEX60DD201000004.1;     95.2    88.1    0.3    0.3    67    0.93  0.73842
AMEX60DD201000005.2;     23.5    16.3    0.6    0.3    67    0.69  0.49557
AMEX60DD201000006.1;    382.0   496.6    0.4    0.3    67    1.30  0.29852
```

13 Alignment to a Genome

I also align to the genome, mainly for comparison and validation of the kallisto analysis. This requires a splice-aware aligner, of which there are three in the mainstream:

- Hisat2, evolved from TopHat, is the fastest (Musich, 2021, doi: 10.3389/fpls.2021.657240) and most memory-efficient (except for indexing very large genomes, like axolotl, where it required over 70 GB). Aligning the reads has much more modest memory requirements.

- Salmon is notoriously difficult to build, so the only viable option for most researchers is to use the precompiled executables provided by the developers. Downloading and running executables this way is considered a very bad practice in I.T. due to the risk of viruses, Trojan horses, etc. It's better to install executables from a more trustworthy source, such as Debian packages, FreeBSD ports, pkgsrc, etc.

  Memory requirements are enormous, which means it will not run on a typical PC and cannot utilize all the cores on a typical HPC node, so the alignment will take much longer than necessary.

  The project currently has 290 open issues, which is not an indication of high code quality and active maintenance. https://github.com/-COMBINE-lab/salmon.

- STAR has the same issues with memory use and code quality as Salmon, but worse: 690 open issues at the time of this writing. https://github.com/alexdobin/STAR.

I experimented with all three of these aligners and now only use hisat2.

Unlike kallisto, these aligners do not perform quantification. FASDA performs quantification on the SAM/BAM/CRAM output of these aligners, generating an output file in the same format as kallisto's abundance.tsv, so that the normalization and DA stages are exactly the same regardless of the aligner used.

By default, FASDA uses **stringtie**, a highly-evolved abundance calculator often used on **hisat2** output. FASDA runs stringtie to compute abundances and then reformats the output into a kallisto-style abundance.tsv file.