RNA-Seq Differential Expression Analysis for Non-programmers Practice Key

---

June 05, 2025

Jason W. Bacon

Contents

June 05, 2025

0.1 Using this key

This key to the practice problems is provided to allow students to immediately *check their own work*. Do not look at this answer key before writing down your own answers. In doing so, you would only cheat yourself out of an opportunity to learn and prepare for the quizzes and exams. Writing things in your own words vastly improves your understanding.

0.2 Practice Problem Instructions

- Practice problems are designed to help you think about and verbalize the topic, starting from basic concepts and progressing through real problem solving.
- Use the latest version of this document.
- Read one section of this document and corresponding materials if applicable.
- Try to answer the questions from that section. If you do not remember the answer, review the section to find it.
- Do the practice problems *on your own*. Do not discuss them with other students. If you want to help each other, discuss *concepts* and illustrate with different examples if necessary. Coming up with the correct answer on your own is the only way to be sure you understand the material. If you do the practice problems on your own, you will succeed in the subject. If you don't, you won't.

  If you're still not clear after doing the practice problems, wait a while and do them again. This is how athletes perfect their game. The same strategy works for any skill.
- Write the answer in your own words. Do not copy and paste. Verbalizing answers in your own words helps your memory and understanding. Copying does not, and it demonstrates a lack of interest in learning.

  Answer questions completely, but *in as few words as possible*. Remove all words that don't add value to the explanation. Brevity and clarity are the most important aspects of good communication. Unnecessarily lengthy answers are often an attempt to obscure a lack of understanding and may lead to reduced grades. "If you can't explain it simply, you don't understand it well enough." -- Albert Einstein
- Check the answer key to make sure your answer is correct and complete.

  DO NOT LOOK AT THE ANSWER KEY BEFORE ANSWERING QUESTIONS TO THE BEST OF YOUR ABILITY. In doing so, you only cheat yourself out of an opportunity to learn and prepare for the quizzes and exams.
- ALWAYS explain your answer. No exceptions. E.g., justify all yes/no or other short answers, show your work or indicate by other means how you derived your answer for any question that involves a process, no matter how trivial it may seem, draw a diagram to illustrate if necessary. This will improve your understanding and ensure full credit for the homework.
- Verify your own results by testing all code written, and double checking short answers and computations. In the working world, no one will check your work for you. It will be entirely up to you to ensure that it is done right the first time.
- Start as early as possible to get your mind chewing on the questions, and do a little at a time. Using this approach, many answers will come to you seemingly without effort, while you're showering, walking the dog, etc.
- For programming questions, adhere to all coding standards as defined in the text, e.g. descriptive variable names, consistent indentation, etc.

Chapter 1

Bioinformatics analysis background

1.1 The status quo

1.1.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. Why has RNA-Seq differential expression analysis traditionally been difficult for the average biologist?

   Installing some of the software requires I.T. skills and using some of the software requires computer programming skills.

2. What is "abandonware"?

   Abandonware is software that is no longer supported, and likely difficult to install.

3. What is "paperware"?

   Paperware is a special type of abandonware that was developed in order to publish a paper, and immediately abandoned.

4. Why don't biologists just find a collaborator with bioinformatics experience to help with computer analyses?

   People with the required skills are in short supply, so they either don't have time, or cost too much under the economic law of supply and demand.

5. What is needed in the bioinformatics community to make analyses easier?

   More high-quality application development.

1.2 A much easier approach

1.2.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What are three ways the approach in this text brings RNA-Seq differential expression analysis within reach of the average biologist?

   - It makes software installation trivial.
   - It eliminates the need to write computer programs just to use certain tools.

- It provides software written in a sustainable way, so that minimal maintenance is required and it will continue to function long into the future.

2. What is the minimum computer knowledge required to perform an RNA-Seq differential expression analysis?

   Only a basic understanding of the Unix command-line environment is required.

3. What computer knowledge is required to become a professional bioinformatician?

   Bioinformaticians need to be adept at multiple programming languages.

4. What kind of computer hardware is necessary for a typical differential expression analysis, and how long will it take to complete?

   A typical differential expression analysis can be done in a few days on a typical PC or Mac.

5. How does the approach used by this book facilitate software installation?

   All of the necessary software is installed via a package manager, simply by installing a single meta-package.

6. Describe four ways the use of a unified package manager helps scientists be more productive.

   - It makes software installation trivial, even for complex programs.
   - It eliminates the confusion of "context switching" between multiple installation methods, which is tiresome and leads to mistakes.
   - It prevents conflicts between software packages.
   - It simplifies the user's shell environment, since all of the software is installed in one place.

7. Why do the new tools created for this analysis approach require less maintenance than most?

   They are written in stable, portable languages.

1.3 How you can help

1.3.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is the easiest way for users of scientific software to help improve software quality for everyone?

   The can communicate with the developers when they encounter problems, so that the developers are motivated to improve their software.

1.4 What is a pipeline?

1.4.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is a pipeline?

   A pipeline is a procedure that involves multiple analysis stages, one after another, where the output of one stage is the input to the next.

2. What is the main problem with automated pipelines?

   Automated pipelines often waste time and resources, because the continue to run subsequent stages even if the quality of an earlier stage is inadequate.

1.5 FASTQ files

1.5.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. How many lines represent each read in a FASTQ file? Explain.

   At least four, but possibly more, as sequences and quality scores can be split across multiple lines.

2. Describe the four fields of a read in a FASTQ file.

   (a) A line containing the sequence identifier and optional additional text to describe the sequence.
   (b) One or more lines of sequence data.
   (c) A separator line starting with '+', marking the end of the sequence data.
   (d) One or more lines of PHRED scores indicating the likelihood of a read error for each base in the sequence.

3. What is the *numeric* PHRED score and the probability of a read error for the fifth base in the following read, assuming PHRED-33 scoring. Reminder: As for all practice questions, show your work.

```
@generand1
CAGTGCCGCTCAGGTGACTAAAATG
GAGCCAGCTCTAGCCCGCTCAGCCC
+generand1
@@@@@@@@@@@@@@@@@@@@@@@@@FG
GGGGGGGGGGGGGGGFFFFFF25<:
```

   '@' is character 64, so the number score is 64 - 33 = 31. P(error) = 10 ^ -(31/10) = .00079432823472428150.

4. Are quality scores typically consistent for a given read? Why or why not?

   No, they are typically lower near the ends due to the chemistry of the sequencing process being out of equilibrium.

5. What is generally considered a high quality score? An acceptable score?

   28 or higher is considered high quality, while 20 to 27 is considered acceptable.

1.6 The sequencing procedure

1.6.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. Describe the three major steps in the RNA-Seq lab protocol.

   (a) Extract RNA from a sample
   (b) Fragment it
   (c) Use a sequencing machine to read full or partial sequences

2. What is an adapter?

   An adapter is an artificial sequence added at each end of a DNA/RNA fragment to enable the sequencing machine to sequence it.

3. What is a tag?

   A tag is an artificial sequence, part of an adapter, that is unique to each sample. It enables us to determine which sample each read came from when samples are pooled for sequencing.

4. What is an insert?

   An insert is the natural DNA/RNA sequence portion of a read, i.e. the portion between the 5' and 3' adapters.

5. Do sequencers typically sequence entire mRNAs?

   No, DNA/RNA is fragmented to an average of a few hundred bases using restriction enzymes before being sequenced.

6. What is a typical read length for RNA-Seq?

   Typical read length is between 50 and 150 bases.

7. Are all the bases in our fragments represented in typical FASTQ files? Why or why not?

   No, reads are typically limited in length and fragments longer than the read length will not be read entirely.

1.7 Single vs paired-end sequencing

1.7.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is paired-end sequencing?

   Paired-end sequencing is a method of reading both the 5' end and the 3' end of each DNA/RNA fragment.

2. Describe two advantages of paired-end sequencing.

   - More of each fragment is represented in our FASTQ files.
   - If the 5' (forward) and 3' (reverse) reads are properly mated/paired, then we can determine the length of the original fragment.

3. How are paired-end reads represented in FASTQ files? What complication does this present for analysis?

   Forward and reverse reads are stored in separate files, with the mates at the same location within each file. Analysis programs must be careful to keep the forward and reverse files in sync. I.e., if a read is removed from one file, the mate must be removed from the same position in the other, since the position is the only way of knowing which reads are mates.

Chapter 2

RNA-Seq pipeline overview

2.1 The pipeline

2.1.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. How accurate are the RNA abundances we compare in an RNA-Seq differential expression analysis?

   The abundances are not very accurate, but accurate enough to draw some conclusions, especially where the change in transcription rate is many-fold.

2. Is RNA abundance used because it is the best measure of gene activity?

   RNA abundance is generally not the best measure of gene activity. We use it because it is the most convenient reflection of gene activity that we can currently quantify.

3. What conclusions can we draw from the results of an RNA-Seq differential expression analysis?

   We cannot draw any conclusions from such an analysis. The results are only suggestive and must be verified by other means.

2.2 Know your data

2.2.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. Why is it important to know our data?

   Knowing our data helps us understand exactly what it happening during an analysis. This kind of meticulous work is the difference between science and faith.

2.3 Pre-trim quality check with FastQC

2.3.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What does the yellow box indicate in a FastQC quality score report?

   The yellow box shows the interquartile range, meaning that the middle half of the quality scores for all reads are in this range.

2. What do we do if bases 46 and 47 have low quality scores, assuming the reads are 50 bases?

   In this case, we would remove bases 46 through 50. Just removing bases 46 and 47, while keeping 48 through 50, would alter the sequence of the read.

2.4 Adapter trimming

2.4.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is the benefit of trimming our raw data?

   Trimming removes junk sequences that affect the performance and quality of subsequent stages such as read mapping.

2. What are four things trimming aims to remove from our raw sequence data?

   - Adapters (especially 3' adapters) that were not automatically removed at the sequencing center.
   - Poly-A tails.
   - Ends of reads with low PHRED scores.
   - Reads that are very short and likely to map to more than one location.

3. Show the trimmed version of the following read, assuming the 3' adapter is AGATCGGAAGAG.

```
@
GGGTGATTACACACACCCATTAGCTTTGGGATCGTACGTACTGAAGATCGGAAGAGCTATGCAT
+
:1124?????????????????????????????????????????????????????>>???
```

```
ATTACACACACCCATTAGCTTTGGGATCGTACGTACTGA
```

4. What must trimming software do specially for paired-end data?

   When trimming paired-end data, the software must keep the forward and reverse files in sync. If it removes a read from one file, it must also remove the mate from the other.

2.5 Post-trim quality check

2.5.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. Describe two improvements we should see in the post-trim quality check as compared to pre-trim.
   - There should be very little remaining adapter contamination indicated.
   - Any questionable PHRED scores we saw in the raw data should be eliminated in the post-trim checks.
2. Should the post-trim quality check show zero adapter contamination?

   No, some adapter content will still be present since FastQC and MultiQC show all known adapters, not just the one used for our reads, and similar sequences sometimes occur naturally.

2.6 Read mapping (Alignment)

2.6.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is read mapping?

   Read mapping is aligning the reads in a library to a reference genome or transcriptome, in order to determine the location in the genome from which they were derived.
2. What is a reference genome or transcriptome?

   A references is sequences representing a hypothetical individual, possibly compiled from many individuals and using the most common (consensus) bases where there are differences among them.
3. Do reads always match the reference genome or transcriptome perfectly?

   No, there are differences in the DNA among individuals, so the reference sequences will not be identical to the reads in our library.
4. Are read mapping and alignment the same thing?

   Read mapping is one kind of alignment, as is read trimming and BLAST searches.
5. Is it easier to align RNA reads to a genome or two a transcriptome?

   Aligning to a transcriptome is easier, because the transcriptome is much smaller (less territory to search for a match) and the sequences in the transcriptome are processed (introns are removed), so they will match our reads more closely.
6. What is a properly paired (mated) alignment?

   A properly paired alignment is one where the 5' and 3' mated reads in paired end sequencing align to the same general location in a genome or the same feature in a transcriptome.
7. What is a primary alignment?

   A primary alignment is an alignment that matches a read better (has a higher alignment score) than other locations in the genome/transcriptome, in the event that a read has more than one possible match.
8. Why do most people use BAM instead of SAM file format?

   BAM files are much smaller than SAM files, so it saves space and speeds up reading and writing of the files.

9. What is pseudoalignment and why is it used?

   Pseudoalignment is an alternative to traditional alignment that does not attempt to match entire sequences, but uses statistical methods combined with partial matches. It is generally faster than traditional read mapping.

10. How is a genome browser used to investigate mapped reads?

    A browser such as IGV allows us to visualize the mapping of our reads to a reference.

2.7 Quantification (Abundance estimation)

2.7.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is quantification?

   Quantification is essentially counting the reads aligned to each feature of interest (usually a gene or transcript).

2. What inputs are necessary to perform quantification?

   - An alignment map, usually a BAM or SAM file, which indicates the location(s) to which each read mapped.
   - A feature file (usually GTF or GFF), which indicates the location of each feature (e.g. gene) in the genome.

3. Where can we obtain a feature file?

   Feature files should generally be obtained from the same site as the genome or transcriptome used.

4. What is a pileup?

   A pileup is a grouping of reads aligned to the same feature.

5. Why do pileups generally look incomplete and messy?

   Reads are only portions of fragments, which are portions of transcripts. Hence, many pieces of the original transcript are not included in the set of reads that form a pileup.

2.8 Normalization

2.8.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. Why do we need to normalize abundance estimates obtained in the quantification stage?

   Different samples have different amounts of total DNA/RNA, so the abundance estimates will vary by sample. Normalization allows us to compare abundances across samples.

2. What is the best way to ensure accurate normalization across samples? Why don't we always use this approach?

   Spike-in controls, which add a known concentration of known sequences provides the most accurate normalization. This approach entails extra cost, however. Also, if spike-in controls were not added before sequencing, we have no choice but to use other (i.e. statistical) methods instead.

2.9 Differential expression analysis

2.9.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is a fold-change?

   Fold-change is the ratio of abundance estimates across two conditions, such as wild-type and mutant.

2. Why do we often use log(FC) instead of FC?

   The log(FC) makes it easier to compare the magnitude of changes in up- and down-regulated genes, because the scale of log(FC) is the same. E.g. FCs of 0.4 and 2.5 are the same magnitude, which is not obvious, while log(0.4) = -0.916 and log(2.4) = +0.916.

3. How reliable are fold-changes for determining a biologically significant event?

   Fold-changes are not very reliable, unless the ratio is large. They are based on read counts, which are very crude estimates of biological activity for many reasons.

4. What is technical variation?

   Technical variation is variation in samples not due to biological differences. It is due to other factors such as lab technique and some uncontrollable factors in sequencing, etc.

5. Does a fold-change of 3.6 indicate a significant biological change?

   Not necessarily: The FC value of 3.6 is only a crude estimate, while the real FC in the cells may be smaller. Also, the impact of a given FC depends heavily on the function of the transcript. A FC of 3.6 in one gene product may cause drastic changes in the organism, while the same FC for another gene product may have very little effect. Some changes may be due to causes not of interest to our study as well.

6. What is a P-value?

   A P-value is the probability that a change at least this extreme would occur at random.

7. Does a low P-value indicate a biologically significant event?

   Absolutely not. First, P-values are based only on the numbers used to compute them, and do not incorporate any knowledge of biology. Second, the read counts on which the P-values are based are not very accurate, as we have discussed.

8. If P-values are not reliable, why do we use them?

   We use P-values because we don't have anything better to guide us at this stage, and they do provide *some* guidance about where to focus our efforts. Using P-values to guide us is far better than picking genes to investigate at random.

2.10 Functional analysis

2.10.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. No practice questions for this section.

Chapter 3

Installing the RNA-Seq software

3.1 The status quo: A perilous obstacle course

3.1.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. How hard should it be to install an open source software package?

   Open source installation can and should be trivial in all cases, due to the existence of numerous package managers.

2. Why is it dangerous to download binaries from individual developer websites?

   Binaries could be infected with malware such as viruses, Trojan horses, etc. Downloading binaries from numerous different websites means trusting *all* of those website to maintain strong security, which is not a realistic expectation.

3. What limitations will we encounter when using binaries from developer websites?

   Developer-provided binaries generally only support a few specific CPU types and operating systems, which may not even be the same for all the software we need.

4. What is a major issue with software installed within the R environment using `install.packages()`?

   Some of the software on which the R package depends may have been installed outside of R. R cannot ensure that a compatible version is installed or that the build parameters are correct. When this software is upgraded or removed, the R package will cease to function.

3.2 A secure and reliable approach: Package managers

3.2.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is the major advantage of the dreckly package manager?

   Dreckly can be used to install the exact same software on almost any Unix-like operating system, such as GNU/Linux systems and macOS.

2. What is the quickest way for people who only have an MS Windows machine to gain access to Unix and all the software needed for an RNA-Seq analysis?

   Windows users can install a Unix-like system such as GhostBSD in a virtual machine on their Windows PC. With GhostBSD, installing the RNA-Seq tools is a trivial effort and very fast.

3.3 Dreckly package manager

3.3.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What are three advantages of the dreckly package manager.
   - Dreckly can be used on almost any modern Unix-like operating system (and in fact on some not-so-modern systems).
   - Dreckly does not require administrative rights, so anyone can use it, even on a computer managed by their I.T. department.
   - Dreckly can install and use its own modern compilers instead of those provided by the operating system. This is a major advantage on systems with older compilers that cannot build modern software.
2. What is necessary to use the dreckly system after it has been installed?

   Users simply need to ensure that dreckly is in their PATH, which can be done by sourcing the startup scripts that auto-dreckly-setup installs.
3. How can a dreckly user update all of the software installed by dreckly?

   Updating can be done by simply running **auto-dreckly-update --defaults**, assuming the sysutils/auto-admin package is installed.

3.4 GhostBSD

3.4.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. Who can run GhostBSD in a virtual machine?

   Anyone with a modern operating system, including Windows or Unix, and a graphical display, can run GhostBSD in a VM.
2. Briefly describe four advantages of GhostBSD for the RNA-Seq analysis in this text.
   - GhostBSD is easy to install, even for those who don't know Unix.
   - GhostBSD provides the easiest way to install all the necessary RNA-Seq software.
   - GhostBSD uses the ZFS filesystem, which compresses all files, saving precious space that may be needed for bioinformatics work.
   - GhostBSD is very fast and reliable.
3. What are three major advantages of VirtualBox over other VMMs?
   - VirtualBox is free and open source.
   - VirtualBox has strong support for running graphical operating systems.
   - VirtualBox runs on most popular operating systems, and almost any operating system can run under VirtualBox.
4. How is the process of installing GhostBSD different under VirtualBox, as opposed to installing on real hardware?

   There is no difference: The process is exactly the same.
5. What is Software Station?

   Software Station is GhostBSD's graphical user interface to the FreeBSD ports collection.

Chapter 4

RNA-Seq: A detailed example

4.1 A much easier approach

4.1.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is the major hurdle for biologists who want to compute fold-changes and P-values using most existing software?

   Most existing software requires the user to understand R programming and data structures just to use the tool.

2. What is a hurdle for doing scientific computing in general?

   Many of the scientific analysis programs are difficult to install and use.

3. What are the minimum skills needed to do most kinds of bioinformatics?

   Some knowledge of the Unix command-line (and scripting, which is really the same thing) is necessary in order to use many of the software tools.

4.2 A programming-free work flow

4.2.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. How strictly should you adhere to the tools listed in the flowchart of this section?

   The tools listed here are only suggestions. In analyzing RNA-Seq data, we should try alternatives wherever they are available.

4.3 Saving terminal output

4.3.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is device independence?

   Device independence is a feature of Unix that makes all I/O the same, whether it is going to/from a file or any I/O device.

2. What are the three standard streams open by default in every Unix process, and to what file or device are they normally connected?

   - Standard input (stdin) is normally connected to the terminal keyboard.
   - Standard output (stdout) is normally connected to the terminal screen.
   - Standard error (stderr) is normally connected to the terminal screen.

3. What is redirection?

   Redirection is a Unix feature that allows us to disconnect each of the standard streams from the terminal and connect it to a file or other device.

4. What are pipes?

   Pipes are a feature of Unix that allows us to disconnect each of the standard streams from the terminal and connect it to another *process*, rather than a file or device.

5. Show how to list the files in the current directory, save the output to "listing.txt", and see it on the screen at the same time.

```
ls -l | tee listing.txt
```

4.4 Obtaining raw reads

4.4.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What makes saccharomyces cerevisiae a good organism for practicing RNA-Seq analysis?

   - The genome is small, so analysis steps will complete quickly.
   - It is a well-studied model organism, to the reference genome and transcriptome are fairly complete.
   - Note: It would be incorrect to include biological replicates and technical replicates in this answer. These are features of the particular data set, not the organism. If you included these in your answer, this is an indication that you need to slow down and think about the question in more depth, rather than hastily copy seemingly relevant information from the book.

2. What is the SRA?

   The SRA, Sequence Read Archive, is a repository containing most sequence data from previously published studies.

3. What is sra-tools?

   The sra-tools suite is a command-line tool set used to perform efficient, selective downloads of SRA data.

4.5 Recompression

4.5.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is the best compression tool for sequence data that will be stored long-term?

   Data that will be stored long-term, such as raw reads, are generally worth recompressing with **xz**, which provides the best compression ratio by far.

2. What is the best compression tool for temporary files containing sequence data?

   Temporary files should generally be compressed using a fast compression tool, to avoid slowing down the analysis. The best in terms of speed vs compression ratio is **zstd**. However, we should also consider which compression tools are supported by the analysis tools that must read or write the data. There are a few tools that make it difficult to use anything other than **gzip**.

3. Some filesystems, such as ZFS, can automatically compress all of our files. Does this render compression tools useless?

   No, because the LZ4 compression used by ZFS does not save as much space as other tools, and sometimes we can speed things up by reducing the amount of disk I/O in exchange for some added CPU load used by a compression tool like **zstd**.

4. If a compression tool is slowing down an analysis by using too much CPU time, does this mean we should switch to a different tool?

   Not necessarily: We can reduce CPU use of any tool by lowering the compression level. This will use more disk space, but it may be a good trade.

5. How can we easily write a program that can read or write files compressed with any of the common tools?

   Use the `xt_fopen()` function from libxtend.

4.6 Descriptive naming to avoid calamities

4.6.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. Describe two reasons we should give our raw data files clear and descriptive names.

   - Descriptive names make it easier for people to tell what the file contains just by looking at the name.
   - Descriptive names make it easier for our analysis scripts to parse the filenames, leading to fewer mistakes.

2. What can we do if the sequencing center gives us files with cryptic names, without losing any information about the original names, and without using additional disk space?

   We can create links to the original files with more descriptive names. A link is just another name for the same file, so it does not duplicate the contents.

4.7 Pre-trim quality check

4.7.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. How can we easily interpret the quality scores we see in the raw data files?

   We can interpret the quality scores using the **phred-decode** script provided in the DIY-RNA-Seq-DE repository.

2. Show a command that prints summary statistics about the file `sample01-cond1-rep3.fq.xz`. How do we install this command?

```
shell-prompt: blt fastx-stats sample01-cond1-rep3.fq.xz
```

This command is part of the biolibc-tools package, which is installed with the rna-seq meta-package.

3. Show a command that produces summary quality information about the file `sample01-cond1-rep3.fq.xz`, placing results in `Results/Raw-QC`. Hint: You may need to review Section 4.5.

   shell-prompt: xzcat sample01-cond1-rep3.fq.xz \| fastqc stdin:sample01-cond1-rep3 -o Results/Raw-QC

4. Show a command for viewing FastQC results in the directory `Results/Raw-QC`.

   shell-prompt: firefox Results/Raw-QC

5. What does a red flag in the FastQC left panel indicate?

   A red flag indicates suspected low-quality data, though it does not always indicate a real problem, as FastQC has no way of knowing the specifics about the study. We have to know what to expect from our own data in order to interpret FastQC results correctly.

6. What does the thin red line indicate in the per-base quality graph?

   The thin red line indicates the median PHRED score across all reads at that position.

7. What do the whiskers indicate in the per-base quality graph?

   The whiskers indicate the 10th to 90th percentile range of PHRED scores at this position, so 80% of all reads have a PHRED score within the whiskers at that position.

8. What does the per-tile quality graph tell us?

   If the data came from an Illumina sequencer, this gives us a graphical depiction of quality by location on the slide.

9. What does the per sequence quality scores graph tell us?

   This graph shows the statistical distribution of *average* quality scores for each read.

10. What does the per base sequence content graph tell us? What causes the bias often seen at the 5' end of each read?

    This graph shows the percent of each base (A, C, G, T) at each position. The bias is due to the fact that DNA/RNA was fragmented using restriction enzymes, which bind to certain motifs. Hence the first few bases of each read are not random.

11. What does the per sequence GC content graph tell us?

    This graph shows the distribution of GC content across all reads. The mean of this distribution should match the known GC content of the species.

12. What does the per base N content graph tell us?

    This graph indicates the number of unknown bases at each position, which should be very low.

13. What does the sequence length distribution graph tell us? What do we expect to see here for typical raw RNA-Seq data?

    This graph tells us the distribution of sequence lengths? Since most RNA-Seq data comes from fixed-length short-read sequencers, the lengths here should all be the same.

14. What does the sequence length distribution level graph tell us? What do we expect to see here for RNA-Seq data?

    This graph indicates how many sequences are found N times for any given value of N >= 2. Duplicated sequences are normal for RNA-Seq data, since there are many transcripts from some genes.

15. What does the overrepresented sequences graph tell us?

    This graph indicates whether there were sequences that occurred more often than they should by chance.

16. What does the adapter content graph tell us? What do we expect to see here and why?

    This graph indicates the number of known adapter sequences spotted starting at each possible position in the reads. It is normal to see some adapter content in raw reads, because some fragments are shorter than the read length, so that the 3' adapter is included in the read. Adapters become increasingly likely toward the 3' end of the reads.

17. What does MultiQC do?

    MultiQC combines the FastQC graphs for all FASTQ files into a single summary graph with interactive features.

4.8 Adapter trimming

4.8.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is the goal of read trimming in an RNA-Seq analysis?

   Read trimming aims to remove adapters, bases with low PHRED scores, and poly-A tails, in order to improve the quality of the reads for subsequent stages.

2. What is the disadvantage of using a trimming tool that automatically detects which adapter sequences were used? What is the advantage?

   When using such a tool, we won't know our data as well. The advantage isn't much: It literally only saves us the few seconds required to run a tool like **fastq-scum** to find out what adapters our reads contain.

3. What compression tools can be used implicitly with **fastq-trim** input and output files?

   The **fastq-trim** command supports all popular compression tools implicitly. The choice of tools should be based on the need to save disk space and what tools are supported by the next stage of our pipeline.

4. Show a **fastq-trim** command that will trim reads in the paired-end files `sample01-cond2-rep05-R1.fastq.xz` and `sample01-cond2-rep05-R2.fastq.xz`, saving the output to `sample01-cond2-rep05-R1-trimmed.fastq.gz` and `sample01-cond2-rep05-R2-trimmed.fastq.gz`. The 3' adapter begins with AGATCGGAAGAG. Trim sequences that match at least the first four bases of the adapter with no more than 15% differences. Trim ends with PHRED scores below 25 and poly-A tails of length five or more. Discard reads less than 20 bases in length after trimming. Save a copy of all terminal output to `trim.out`, assuming you are using a Bourne shell derivative.

```
shell-prompt: fastq-trim --3p-adapter1 AGATCGGAAGAG --polya-min-length 5 \
    --min-match 4 --max-mismatch-percent 15 --min-qual 25 --min-length 20 \
    sample01-cond2-rep05-R1.fastq.xz sample01-cond2-rep05-R1-trimmed.fastq.gz \
    sample01-cond2-rep05-R2.fastq.xz sample01-cond2-rep05-R2-trimmed.fastq.gz \
    2>&1 | tee trim.out
```

4.9 Post-trim quality check

4.9.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What kind of changes should we expect to see in the FastQC basic statistics post-trim results, as compared with the pre-trim results, assuming fixed-length reads?

   Total bases will reduced, read length will show a range now whereas it may have been constant before trimming, and total sequences will likely be reduced (if some reads were discarded because they were too short after trimming).

2. What kind of changes should we expect to see in the FastQC read length distribution post-trim results, as compared with the pre-trim results, assuming fixed-length reads?

   We should now see a curve representing a range of read lengths rather than the pyramid showing a constant read length of the raw reads.

3. What kind of changes should we expect to see in the FastQC per base read quality post-trim results, as compared with the pre-trim results?

   The mean, average, interquartile, and 10-90 percentile range indicators should all be improved where the scores were low for the raw reads. None of them should be in the red, indicating low PHRED scores, after trimming.

4.10 Reference genome and transcriptome

4.10.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is a reference genome?

   A reference genome is a representation of the DNA sequences of a hypothetical individual.

2. What is a reference transcriptome?

   A reference transcriptome is a representation of the coding sequences (genes, transcripts, exons) from a hypothetical individual.

3. Why don't we use the sequences from a real individual as a reference?

   Every individual has unique DNA, and a reference using consensus will better match most individuals on average.

4. What are the common types of differences in DNA sequences between two individuals?

   - SNPs (single nucleotide polymorphisms) are differences in a single base position.
   - Indels (insertions/deletions) are where one or more additional bases are inserted into the genome relative to another individual.

5. What file format usually contains reference genomes and transcriptomes? How is a sequence represented in a FASTA file?

   FASTA file format is usually used for references. A sequence, such as a chromosome in a genome or a transcript in a transcriptome, is represented as a comment line beginning with a '>' followed by the sequence ID and arbitrary text description. This line is followed by one or more lines of sequence data.

6. Why is the sort order of reference files important?

   Many bioinformatics programs rely on data being sorted in order to operate efficiently.

7. How accurate are the reference genomes and transcriptomes for most organisms? Why?

   References are not very accurate for most organisms, because they are very difficult to build. Only the most common model organisms have highly accurate references available.

8. Where can we find most reference genomes and transcriptomes? Which one should we use?

   The three major sites are Genbank in the US, Ensembl in Europe, and the DNA data bank of Japan. They share data with each other, so the choice is usually a matter of geographic proximity for speed, or personal preference for the website design.

9. Why would we choose not to simply download a genome reference as one file?

   We may not want to have, e.g., sex chromosomes, the mitochondrial chromosome, and/or unassembled sequences in our reference.

10. What is the advantage of using a command-line download tool such as **curl** instead of downloading reference file interactively with a web browser?

    Using a tool like curl, we can script the download, so that it is perfectly documented and easily reproducible.

11. Show a **curl** command for downloading chromosome 1 of Mus musculus (common mouse) from Ensemble.

```
shell-prompt: curl -O https://ftp.ensembl.org/pub/release-114/fasta/mus_musculus/dna/ ↩
    Mus_musculus.GRCm39.dna.chromosome.1.fa.gz
```

12. Show a command for viewing the content of the mouse chromosome 1 downloaded in the previous question, one screen at a time.

```
shell-prompt: gunzip -c Mus_musculus.GRCm39.dna.chromosome.1.fa.gz | more
```

13. What are checksums for? How to they work?

    Checksums allow us to quickly confirm the integrity of a file we downloaded. The checksum value is computed from the file content, and is almost never the same for two difference files.

14. What is a feature file?

    A feature file contains the name, location, size, etc. of all known features (genes, transcripts, exons, etc.) in a genome.

15. How do we obtain a reference transcriptome?

    There are two ways:

    - Download the cDNA reference from one of the web sites.
    - Construct a reference using a feature file (GTF or GFF), which lists the location of every known feature in the genome, and a reference genome.

16. What is a simple sanity check for a transcriptome reference built from a genome and feature file?

    We can count the features in the transcriptome FASTA and the same features in the feature file from which it was built. They should match closely.

4.11 Read mapping (Alignment)

4.11.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. What is the difference between "read mapping" and "alignment"?

   Alignment refers to any matching of sequences for any purpose, including adapter removal, etc. Read mapping uses alignment to determine the location in the genome from which reads originated.

2. What is an alignment map?

   An alignment map is a file containing reads, the location(s) in the reference to which they best align, and the alignment scores (estimate of how well they align to that location).

3. What is the purpose of an index in read mapping?

   An index greatly speeds up the search for good alignments for each read, vs searching the entire reference for each read.

4. Show the commands for generating a kallisto index called `kallisto-mouse.index` from a reference called `mouse-transc fasta`.

   ```
   shell-prompt: samtools faidx mouse-transcriptome.fasta
   shell-prompt: kallisto index --index=kallisto-mouse.index mouse-transcriptome.fasta
   ```

5. How do we determine the success of **kallisto index** to an extent that justifies moving on to the next step of performing pseudo-alignments and quantification?

   Carefully examine the terminal output from kallisto index to check for warnings and errors, and check that the index file size is comparable to the size of the transcriptome reference file.

6. What is multithreading?

   Multithreading is the use of multiple cooperating processes at same time to speed up completion of a task.

7. What is a named pipe, and when are they needed?

   A named pipe is a Unix pipe with a filename attached to it. We can use a named pipe to pipe data into programs that cannot read input from the standard input, but require an input filename.

8. What, in general, are the advantages of a named pipe (or an unnamed pipe for that matter) over using a regular file for intermediate results?

   - A regular file might occupy a large amount of disk space. A pipe uses no disk space, just a small amount of memory while in use.
   - A pipe is much faster than disk I/O, even if the disk is an SSD.
   - Using a regular file, the process creating the file must complete before the process reading it can start. With a pipe, they can run at the same time.

9. What are three simple commands to perform sanity checks for Kallisto outputs in the directory ./Kallisto-out?

   -
   ```
   shell-prompt: ls -l Kallisto-out/*.tsv
   ```
   -
   ```
   shell-prompt: wc -l Kallisto-out/*.tsv
   ```
   -
   ```
   shell-prompt: head Kallisto-out/*.tsv
   ```

10. What is one reason to map reads to a genome as well as to a transcriptome?

    Mapping to a transcriptome will only match reads to known genes. Mapping to a genome is the only way to discover new genes using the read data.

11. What is the main problem when mapping mRNA reads to a eukaryotic genome? What is the solution?

    Eukaryotic mRNA is processed before leaving the nucleus, most importantly by removing introns. This means that the RNA sequences of reads obtained from the cytoplasm will not match the genome, since the genome contains intron sequences and the reads do not. The solution is to use a "splice-aware" aligner, which can match reads with the introns removed to the original genome sequences.

12. Show the commands needed to build a Hisat2 index in ./Results/Hisat2 from the reference genome ./Results/Reference/genome.fa.

    ```
    shell-prompt: samtools faidx Results/Reference/genome.fa.
    shell-prompt: mkdir Results/Hisat2
    shell-prompt: cd Results/Hisat2
    shell-prompt: ln -s ../Reference/genome.fa .
    shell-prompt: ln -s ../Reference/genome.fa.fai .
    shell-prompt: hisat2-build genome.fa genome.fa
    ```

13. Show the commands needed to map reads from ./Results/Trimmed/sample01-cond1-rep01.fastq.gz to the reference described in the previous question, storing the alignment map in ./Results/Hisat2/sample01-cond1-rep01.bam.

    ```
    shell-prompt: mkdir Results/Hisat2
    shell-prompt: cd Results/Hisat2
    shell-prompt: hisat2 --threads 4 -x ./genome \
        -U ../Trimmed/sample01-cond1-rep01-trimmed.fastq.gz \
        | samtools sort > sample01-cond1-rep01.bam
    ```

14. What are two simple sanity checks we can perform on our Hisat2 alignment output?

    - Do a long listing (ls -l) of all the BAM files and look for any that are much smaller than the rest. This might indicate a failed alignment job.
    - Run **samtools quickcheck** on each of the BAM files to detect any obvious corruption of the files.

4.12 Differential expression analysis software

4.12.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is the biggest problem for the average biologist running differential expression (DE) software?

   Most popular DE software requires R programming skills just to use it. Unlike many tools, it's not just a matter of running a command with our data files as inputs.

2. What normalization method does FASDA use?

   FASDA uses the median ration normalization method, the same as DESeq2.

3. How consistent are the results across popular DE tools?

   Results are not very consistent at all. According to one study, EdgeR missed between 23% and 75% of SDE features reported by DESeq2, and in one dataset, there was only an 8% overlap between SDE features reported.

4. What is the advantage of using a non-parametric statistical method?

   Non-parametric methods make no assumptions about the distribution of the data, so they are stable. Parametric methods will not work well when the data do not fit the assumed distribution.

4.13 Quantification

4.13.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is quantification?

   Quantification is the process of determining how many reads or transcripts align to each feature of interest (e.g. gene or transcript).

2. Do we need to convert our read counts to transcript counts for the purpose of differential expression analysis?

   No, the ration of reads / transcripts is about the same, except for random variation, regardless of the biological condition, so converting to transcripts will not change our results.

3. What are the inputs to a quantification program?

   The inputs are an alignment map (SAM, BAM, or CRAM file), which indicates the location in the reference to which each read best aligns, and a feature file (GTF or GFF), which indicates the location, size, and other information about all known features for the organism. The locations in the alignment map are looked up in the feature file and the count for the matching feature is incremented.

4. Show a FASDA command to compute abundances for each transcript in the alignment map `./Results/Align/sample01-cond1-rep1.bam` using the feature file `./Results/Reference/t-rex.gff3`. The raw reads are 100 bases long.

   shell-prompt: fasda abundance 100 \ Results/Reference/t-rex.gff3 Results/Align/sample01-cond1-rep1.bam

4.14 Normalizing across samples

4.14.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. Why must we normalize counts across samples?

   The total RNA in each sample varies due to technical factors that have nothing to do with the rate of gene expression for each sample. Normalization attempts to eliminate these factors so that the read counts will be about the same where gene expression was about the same.

2. Does high technical variation indicate bad lab technique?

   Not necessarily. Lab technique is only one of many factors contributing to technical variation.

3. Why do we not use standard normalization methods such as TPM for differential analysis?

   Methods like TPM are used to normalize across features (e.g. genes) within a sample, adjusting for different feature lengths. What we need to do is normalize across samples for the same feature, effectively equalizing sample size.

4. Show a FASDA command to normalize counts in the files `Results/Abundance/sample01-cond1-rep01-abundance.tsv`, `Results/Abundance/sample02-cond1-rep02-abundance.tsv`, `Results/Abundance/sample03-cond2-rep01-abundance.tsv`, and `Results/Abundance/sample04-cond2-rep02-abundance.tsv`, placing the normalized counts in `./normalized.tsv`.

   shell-prompt: fasda normalize \ --output normalized.tsv \ Results/Abundance/sample01-cond1-rep01-trimmed-abundance.tsv \ Results/Abundance/sample02-cond1-rep02-trimmed-abundance.tsv \ Results/Abundance/sample03-cond2-rep01-trimmed-abundance.tsv \ Results/Abundance/sample04-cond2-rep02-trimmed-abundance.tsv

5. What is a quick sanity check to verify that the normalization command did not fail?

   We can use **wc** to count the lines in the normalized counts file and the lines in the individual counts files. They should all be identical, except for the one extra line in the individual files, which is a header line describing the columns.

4.15 Computing fold-changes and P-values

4.15.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. The file `normalized.tsv` contains counts for condition 1 in columns 2, 4, and 6, and condition 2 in columns 3, 5, and 7. Column 1 contains the feature names. Show the Unix command to create the files `cond1.tsv` and `cond2.tsv`.

```
shell-prompt: cut -f 1,2,4,6 normalized.tsv > cond1.tsv
shell-prompt: cut -f 1,3,5,7 normalized.tsv > cond2.tsv
```

2. What is the first sanity check we should perform on **fasda fold-change** output?

   We should make sure that the fold-change output has the same number of features as the normalized counts inputs. This can be easily done using **wc -l**.

3. What is a fold-change?

   A fold-change is the ration of mean normalized counts across two conditions.

4. What is a P-value in terms of fold-changes?

   A P-value is the probability that a fold-change at least this extreme would have occurred at random, given the counts from which the fold-change was derived.

5. How reliable are P-values for determining which genes have seen a significant change in expression? What is one way to get a better idea about which DE features to focus on?

   P-value are not reliable at all, especially with low sample counts. One solution is to run multiple DE tools and compare closely examine the results where the produce very different P-values.

6. What typically happens when we sample RNA more than once under the exact same conditions?

   Typically, we see a fairly large amount of variation, even where the conditions are (allegedly) the same. RNA abundance fluctuates widely due to many factors, so we cannot expect a high degree of consistency across samples, or even from the same individual.

7. How do P-values help us cope with unreliable abundance estimates? Is the situation just hopeless?

   While the results are unreliable, there are typically thousands of DE features, and selecting based on P-values improves our odds of finding some that are truly interesting.

8. How to standard deviations in the counts help us understand the results of a DE analysis?

   The standard deviations help us understand why the P-value is what it is. Where counts vary a lot across samples, the P-values will be higher.

9. What do P-values tell us about the biological significance of a fold-change?

   P-values tell us absolutely nothing about biology. They are merely a mathematical function of a group of numbers, in this case, a set of read counts.

10. How is an exact P-value computed for a fold-change? Why don't we always use them?

    An exact P-value is computed by computing the fold-change of every possible arrangement of read counts, and dividing the fold-changes greater than or equal to the actual fold-change by the total number of possibilities. This is only feasible for a small number of samples, because the number of possible arrangements of counts grows rapidly with increasing sample size.

11. Why are read counts so unreliable? Keep your answer simple and general.

    There are numerous variables beyond the control of the lab biologists that affect read counts, such as individual differences among samples, and the precise timing of the sample extraction.

12. As a biologist, are we more concerned about false negatives (missing something significant) or false positives (exploring something that is actually not significant)?

    We're more concerned about false positives, because they can lead to significant wasted time and resources.

13. What are the typical effects of increasing the number of replicates?

    This generally leads to lower P-values and hence more SDE features flagged if using the same P-value cutoff.

4.16 Visualizing DE results

4.16.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. How are read counts represented in a heatmap?

   The color of each section of the heatmap represents the z-score of the counts for that feature. By using z-scores, we normalize the color scheme so that the color is the same for the mean of each feature, rather than having vastly different colors for features with different mean read counts.

2. What does the dendrogram next to the heatmap tell us?

   The dendrogram shows us how closely the changes in read counts correlate between two features. I.e., two features with a very similar fold-change will be closely linked by the dendrogram.

3. Why does a heatmap show read counts for all samples, rather than just the mean normalized count for each condition?

   All counts are shown for the same reason that FASDA shows standard deviation / mean count: It allows us to easily see how variable the read counts are for a given feature. It can also help us identify outliers among all the samples.

4. Which plotting library is the best?

   This is a silly question, as is any question asking "which is better" in general. The answer depends on the specific needs of each study, and to some extent on personal preference. Always be critical of the question rather than assuming it must have a correct answer, or trying to tell the inquirer what they want to hear.

5. How do we quickly and easily select the interesting SDE features from the list produces by **fasda fold0-change**?

   We can't do this quickly or easily. Selecting features of interest involves carefully examining the fold-changes, P-values, other statistics produced by **fasda fold-change**, and also considering the biology of each feature. E.g., we need to think about what minimum fold-change is actually significant for that specific gene or transcript.

4.17 Transcript or gene level expression analysis

4.17.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. Why is gene-level differential analysis problematic?

   The RNA-Seq protocol isolates transcripts, not genes, and a given gene may produce multiple different transcripts (isoforms) with different functions in the cell. The combined abundance of all isoforms from a given gene may be meaningless, and at the least contains less information than the separate abundances of each isoform.

2. How do we convert from transcript IDs to gene IDs (or names)?

   We use the parent feature listed in a feature file (GTF or GFF) to identify the gene from which a transcript was transcribed.

4.18 Multiple conditions

4.18.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

---

1. List the comparisons that must be performed for a differential analysis using three conditions, "wild-type", "mutant1", "mutant2", and "mutant3".

```
wild-type vs mutant1
wild-type vs mutant2
wild-type vs mutant3
mutant1 vs mutant2
mutant1 vs mutant3
mutant2 vs mutant3
```

4.19 Functional genomics

4.19.1 Practice

Note Be sure to thoroughly review the instructions in Section 0.2 before doing the practice problems below.

1. What is the goal of functional analysis following DE analysis?

   Functional analysis aimed to determine the biological impact of the significant fold-changes we uncovered. This may include examining the direct impact of changes in expression, as well as determining interactions with other genes.

2. Are there many computational tools to automatically perform a functional analysis?

   There are none. Functional analysis is largely a manual process that must be performed by a biologist with a good understanding of the genes / transcripts identified as SDE.