Mid point report

Muhammad Ilyas

2023-06-12

Contents

Research Questions

In this dissertation I will try to answer the following questions using transcriptomic data from Japanese quails (*Coturnix japonica*).

1. How does gene expression vary across different tissues?
2. How phenotypic differences are shaped by differentially expressed genes?

Differential gene expression

DNA is responsible to determine the functions and properties of each individual cell, across the living world. The blueprint for this is the functional units of DNA called genes. By selectively switching on and off a particular set of genes, cells can dynamically access and translate specific instructions through ‘gene expression’. The selected genes’ information is transcribed into RNA molecules, which can then be utilised to directly control gene expression or translated into proteins. This indicates that in a certain condition and time the set of RNAs transcribed reflects the current state of a cell and can reveal underlying mechanisms. What’s more interesting is that the study of differential gene expression enables the comparison of gene expression profiles from various tissues and situations in order to pinpoint genes that are crucial for phenotypic determination. For instance, studying the differences between healthy and diseased tissues can reveal new information about the genetic factors that influence pathology.

Relevancy to research

The transcriptome is the whole set of RNAs transcribed from the genes of a cell. As discussed above, their relative abundances reflect the level of expression of the corresponding genes, for a specific developmental stage or physiological condition. Although RNAs are not the final products of the transcription–translation process, the study of gene expression and differential gene expression can unveil important aspects about the cell states under investigation.

In past years, hybridization-based approaches such as microarrays, were the most used solutions for gene expression profiling and DE analysis, thanks to their high throughput and relatively low costs [3]. These technologies consist in an array of probes, whose sequences represent particular regions of the genes to be monitored. The sample under investigation is washed over the array, and RNAs are free to hybridize to the probes with a complementary sequence. A fluorescent is used to label the RNAs, so that image acquisition of the whole array enables the quantification of the expressed genes. Although widely used in quantitative transcriptomics, these techniques have several limitations [3, 4]:

reliance on prior knowledge about the genome for probe design;

possibility to monitor only some portions of the known genes and not the actual sequences of all transcribed RNAs;

high background levels due to cross-hybridization, i.e. imperfect hybridization between quasi-complementary sequences;

limited dynamic range due to background noise and signal saturation;

need for normalization to compare data from different arrays.

The advent of NGS has revolutionized transcriptomics and quickly established RNA-seq as the preferred methodology for the study of gene expression [3, 5]. The standard workflow of an RNA-seq experiment is described in the following. The RNAs in the sample of interest are initially fragmented and reverse-transcribed into complementary DNAs (cDNAs). The obtained cDNAs are then amplified and subjected to NGS. In principle, all NGS technologies can be used for RNA-seq, even though the Illumina sequencer (http://www.illumina.com) is now the most commonly used solution [6]. The millions of short reads generated can then be mapped on a reference genome and the number of reads aligned to each gene, called ‘counts’, gives a digital measure of gene expression levels in the sample under investigation.

Although RNA-Seq is still under active development, it is now widely used in place of microarrays to measure and compare gene transcription levels because it offers several key advantages over hybridization-based technologies [3–5, 7–9], such as:

reconstruction of known and novel transcripts at single-base level;

broad dynamic range, not limited by signal saturation;

high levels of reproducibility.

The flexibility enabled by single-base resolution probably represents the most powerful feature, as it allows the quantification and sequencing of all the transcripts present in a sample. Compared with microarrays, that can only assay portions of transcripts corresponding to probes, RNA-seq leverages on the sequencing framework to overcome the pure quantification task, enabling new applications, such as transcriptome profiling of non-model organisms [10, 11], novel transcripts discovery [12], investigation of RNA editing [13, 14] and quantification of allele-specific gene expression [15].

Despite all these newsworthy features and apparently easy scheme of data analysis, RNA-seq studies produce large and complex data sets, whose interpretation is not straightforward [16, 17]. Data analysis is further challenged by technical issues inherent to the specific NGS technology, such as sequencing errors in the output reads due to miscalled bases [2], or to biases introduced by the different steps of the RNA-seq protocol, such as amplification, fragmentation and reverse-transcription [18–20]. In particular, protocol-specific bias may under- or over-represent specific loci leading to biased results, thus necessitating careful data quality control and normalization. The latter issue is described in details in the 'Count bias and normalization' section. Nevertheless, if a well-annotated reference genome or transcriptome is available and if the aim of an RNA-seq study is the detection of DE genes, a basic data processing pipeline consists in the following steps: (i) read mapping, (ii) counts computation, (iii) counts normalization and (iv) detection of differentially expressed genes (Figure 1). More sophisticated pipelines can be tailored on the specific need by considering the addition of pre- and post-processing modules to be used before and after read mapping.

Relevant Literature

Studies on differential gene expression

Studies on Japanese quail

Data

Quail transcriptome data

Provided by Dr. Barbara's Lab from Centre of Ecology and Conservation. The experiment was designed to investigate the effect of reproductive investment on animal physiology. Two lines with high and low investment based on egg size were selected to find up-regulation or down-regulation of genes. Both the lines were provided enough food to control for food related stress. The tissues investigated were brain, follicles and liver.

Data Wrangling

The raw data needs processing to before any analysis can be performed. A pipeline was designed to process the raw transcriptome data. The process involves quality check, trimming adapter sequences and discarding low quality reads, mapping the transcriptome against a reference genome and obtaining count tables. The count tables can be used for down stream analysis of the data to find out which genes are differentially expressed. Due to the large size of data the servers were used for carrying the process. Most of the analysis was performed in R using Bioconductor, except for the mapping and generating count tables which were done using STAR [1] and featureCounts [2] respectively. The fastp [3] program was accessed using the R wrapper Rfastp [4] to do the quality control and adapter trimming. The reference genome of Japanese quail obtained from [5].

```
cd dissertation_project/data
```

Next Steps

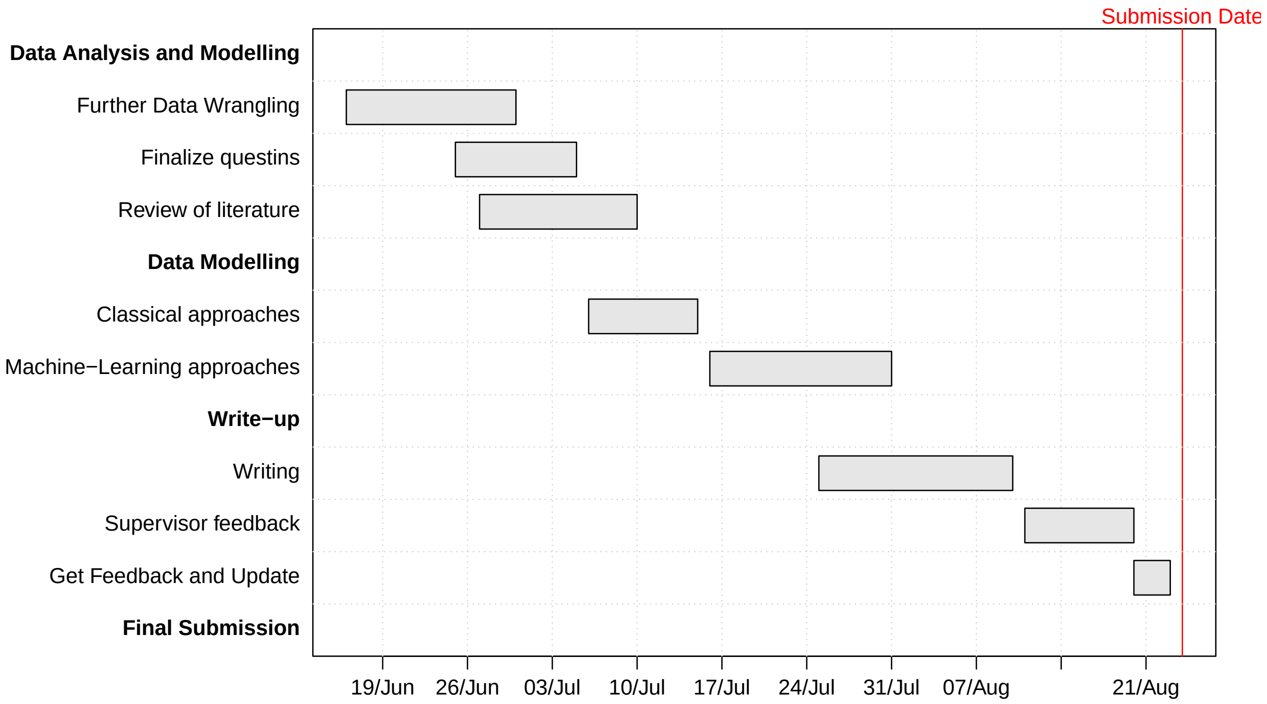

Figure 1: Gant chart for next steps

References

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29: 15–21. doi:10.1093/bioinformatics/bts635

2. Liao Y, Smyth GK, Shi W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30: 923–930. doi:10.1093/bioinformatics/btt656

3. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

4. Wang W, Carroll T. Rfastp: An ultra-fast and all-in-one fastq preprocessor (quality control, adapter, low quality and polyX trimming) and UMI sequence parsing). 2022. doi:10.18129/B9.bioc.Rfastp

5. Morris KM, Hindle MM, Boitard S, Burt DW, Danner AF, Eory L, et al. The quail genome: Insights into social behaviour, seasonal biology and infectious disease response. BMC Biology. 2020;18. doi:10.1186/s12915-020-0743-4