THESIS

A COMPREHENSIVE COMPENDIUM OF ARABIDOPSIS RNA-SEQ DATA

Submitted by

Gareth A. Halladay

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2020

Master's Committee:

Advisor: Asa Ben-Hur

Hamidreza Chitsaz
Anireddy Reddy

ABSTRACT

A COMPREHENSIVE COMPENDIUM OF ARABIDOPSIS RNA-SEQ DATA

In the last fifteen years, the amount of publicly available genomic sequencing data has doubled every few months [1–3]. Analyzing large collections of RNA-seq datasets can provide insights that are not available when analyzing data from single experiments. There are barriers towards such analyses: combining processed data is challenging because varying methods for processing data make it difficult to compare data across studies; combining data in raw form is challenging because of the resources needed to process the data. Multiple RNA-seq compendiums, which are curated sets of RNA-seq data that have been pre-processed in a uniform fashion, exist; however, there is no such resource in plants.

We created a comprehensive compendium for *Arabidopsis thaliana* using a pipeline based on Snakemake. We downloaded over 80 Arabidopsis studies from the Sequence Read Archive. Through a strict set of criteria, we chose 35 studies containing a total of 700 biological replicates, with a focus on the response of different Arabidopsis tissues to a variety of stresses. In order to make the studies comparable, we hand-curated the metadata, pre-processed and analyzed each sample using our pipeline.

We performed exploratory analysis on the samples in our compendium for quality control, and to identify biologically distinct subgroups, using PCA and t-SNE. We discuss the differences between these two methods and show that the data separates primarily by tissue type, and to a lesser extent, by the type of stress. We identified treatment conditions for each study and generated three lists: differentially expressed genes, differentially expressed introns, and genes that were differentially expressed under multiple conditions. We then visually analyzed these groups, looking for overarching patterns within the data, finding around a thousand genes that participate in stress response across tissues and stresses.

ACKNOWLEDGEMENTS

I want to thank Dr. Asa Ben-Hur for encouraging me to find my direction with research, and giving me time while I navigated through my project. He provided many opportunities to grow as a data scientist, and I learned a great deal from his extensive knowledge in the field of Bioinformatics. I also want to thank my committee members, Dr. Hamidreza Chitsaz and Dr. Anireddy Reddy for supporting my research and agreeing to be on my committee.

I am genuinely appreciative of Chris Wilcox, Russ Wakefield, and Dave Matthews for letting me TA for them and providing me with mentorship, always listening to my troubles, and giving me invaluable advice. I also want to extend my thanks to Michael Hamilton, Swapnil Sneham, Basir Shariat, and the various people within the graduate school department who always encouraged me and pushed me to move forward.

My friends have been a constant source of support and have stuck with me through this entire journey. I am truly lucky to have them in my life. I would like to thank Cole Frederick and his family. They gave me a second home in Fort Collins and have always been welcoming when I needed a reprieve from my work. My family encouraged me to go back to school, and to figure out where I will find my passion. They have always been there for me, and I am exceptionally lucky to have their support. Last, but not least, thank you Dr. Howe. You pushed me towards grad school and encouraged me in a thousand small ways that I have continued to reflected on over the past few years.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iii
LIST OF TABLES vi
LIST OF FIGURES vii

Chapter 1 Biological Background 1
1.1 Motivation 1
1.2 Nucleic Acids 2
1.2.1 Deoxyribonucleic Acid 2
1.2.2 Ribonucleic Acid 3
1.3 The Flow of Genetic Information 4
1.3.1 Transcription 4
1.3.2 Translation 5
1.4 Splicing 6
1.4.1 Alternative Splicing 7
1.5 Differential Alternative Splicing and Differential Gene Expression 8
1.5.1 RNA Sequencing Methods 8

Chapter 2 Biological Reproducibility, Replicability, and Data Reusability 11
2.0.1 Incomplete Metadata 12
2.0.2 Inconsistent Terminology for Identification of Metadata 14
2.0.3 Current Solutions 17

Chapter 3 Computational Reproducibility 21
3.1 Version Control 22
3.2 Workflow Management Tools 25
3.3 Package Management and Continuous Analysis 27
3.4 Data Storage and Existing Compendiums 30

Chapter 4 Data Processing and Results 33
4.1 Data Acquisition 33
4.2 Sequence Alignment 36
4.3 Filtering 36
4.4 Data Visualization 39
4.5 Principal Component Analysis 40
4.6 t-Distributed Stochastic Neighbor Embedding 40
4.7 Comparison of PCA and t-SNE using Expression Data 43
4.8 Differential Gene Expression 47
4.9 Differential Intron Retention 47

4.10 Differential Gene Expression and Differential Intron Retention 48
4.11 Identification of Genes that Undergo Differential Expression Under Multiple Stresses 49

Chapter 5 Conclusion and Future Work 56

Bibliography 58

LIST OF TABLES

2.1 Sequence Read Archive Completeness Information 15
2.2 Variations in how data is described 17

4.1 Sequence Read Archive Experiments 39
4.2 Major differences between PCA and t-SNE 43
4.3 Treatment stresses used to find differentially expressed genes 53

LIST OF FIGURES

1.1 Structure of DNA . . . 3
1.2 Information Flow and the Sequence Hypothesis . . . 4
1.3 RNA Splicing . . . 7
1.4 Types of Alternative Splicing . . . 9

2.1 Inadequate annotation of studies in the SRA for *A. thaliana* . . . 13
2.2 SRA RunTable Examples . . . 16

3.1 The number of papers from Bioinformatics between 2009 and 2017 that reference a version controlled repository name in the title or the abstract of the paper . . . 23
3.2 Bioinformatics Tools Poll . . . 28

4.1 Steps for gathering, processing, and analyzing sequence data . . . 34
4.2 Categorization of Treatment Conditions . . . 37
4.3 Pie Charts that describe the compendium . . . 38
4.4 Clustering methods in expression data . . . 45
4.5 Clustering methods by Tissue Type . . . 46
4.6 DEG and DIR Genes . . . 50
4.7 DEG and DIR Genes . . . 51
4.8 DEG and DIR Genes . . . 52
4.9 Histogram of Multiple Stress Frequency . . . 54
4.10 Comparison to genes we identified as MST genes vs previous published results . . . 55

Chapter 1

Biological Background

1.1 Motivation

Analyzing large collections of RNA-seq datasets can provide insights that are not available when analyzing data from single experiments. Through aggregating datasets, new trends can be quickly identified, as well as areas that have not been widely explored. There are barriers towards aggregating large datasets:

- Combining processed data is challenging because varying methods for processing data make it difficult to compare data across studies.
- Combining data in raw form is challenging because processing it is resource intensive.

Multiple RNA-seq compendiums, which are curated sets of RNA-seq data that have been pre-processed in a uniform fashion, exist [4–8]; however, there is no such resource in plants. These compendiums have improved the usability of the data and have helped facilitate the development of new bioinformatic and statistical methods [6]. Datasets that may have only been used for one publication suddenly provide additional value, and lead to opportunities for future publications within the domain.

We created a comprehensive compendium for *Arabidopsis thaliana* using a pipeline based on Snakemake. Over a three year period, starting in 2016, we downloaded over 80 Arabidopsis studies from the Sequence Read Archive. Through a strict set of criteria, we chose 35 studies containing a total of 700 biological replicates, with a focus on the response of different Arabidopsis tissues to a variety of stresses. In order to make the studies comparable, we hand-curated the metadata, pre-processed and analyzed each sample using our pipeline.

We performed exploratory analysis on our compendium using PCA and t-SNE. We discuss the differences between these two methods and show that the data separates primarily by tissue type,

and to a lesser extent, by the type of stress. We identified treatment conditions for each study and generated three lists: differentially expressed genes, differentially expressed introns, and genes that were differentially expressed under multiple conditions. We then visually analyzed these groups, looking for overarching patterns within the data, finding around a thousand genes that participate in stress response across tissues and stresses.

1.2 Nucleic Acids

Nucleic acids are found in abundance in all living things. They transmit and express the information to the interior operations of the cell. Eventually this information is expressed through the different levels of proteins in the cell. Enormous effort has gone into determining these sequences and how the cell adapts to survive.

1.2.1 Deoxyribonucleic Acid

The first nucleic acid we will describe is deoxyribonucleic acid, DNA. DNA encodes hereditary information and provides instructions for the development and functioning of all organisms. A DNA molecule forms a double helix structure; this structure is comprised of two strands of nucleotides that connect like a twisted ladder [9].

The nucleotide is the basic building block that forms the nucleic acid. The nucleotide is comprised of three parts: a sugar (deoxyribose), a phosphate group, and a nitrogenous base. There are four types of nucleotides in DNA: Adenine (A), Cytosine (C) Guanine (G) and Thymine (T). These types are differentiated by the structure of the nitrogenous base and can be further categorized as pyrimidines and purines. Cytosine and Thymine are pyrimidines and their molecular structure is comprised of a single carbon nitrogen ring. Adenine and Guanine are purines and their molecular structure is comprised of two carbon nitrogen rings. This double ring makes them larger in size.

A single strand of DNA is composed of a sequence of these nucleotides. Two complementary strands of DNA are joined together through hydrogen bonds. Cytosine always bonds with Guanine, and Adenine always bonds with Thymine. The two strands of DNA run in opposite directions of

each other. Each strand has a 5' end and a 3' end. These names come from the carbon numbers within the sugar group that forms the bonds between nucleotides within the same strand.

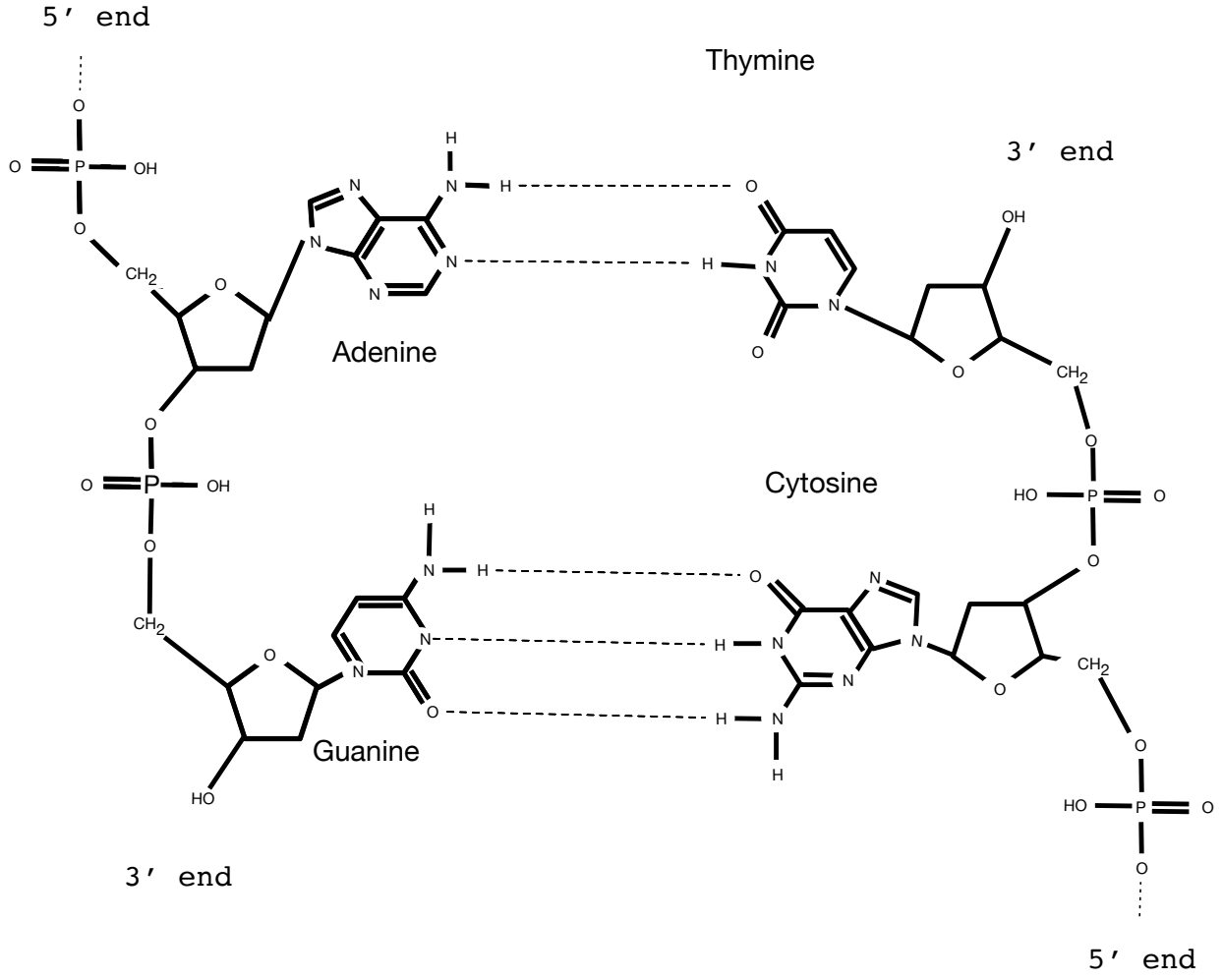

Figure 1.1: The structure of DNA. The four different types of bases can be grouped into pyrimidines and purines. The pyrimidines (Cytosine (C) and Thymine (T)) are bases containing one carbon nitrogen ring. The purines (Adenine (A) and Guanine (G)) contain two carbon nitrogen rings. These four bases pair together through hydrogen bonds; Cytosine always bonds with Guanine, and Adenine always bonds with Thymine.

1.2.2 Ribonucleic Acid

The second nucleic acid is ribonucleic acid, RNA, and it is usually single-stranded. RNA has a very similar structure to DNA, but instead of having a deoxyribose, it has a a ribose sugar. RNAs

nitrogenous bases are Cytosine (C), Uracil (U), Adenine (A), and Guanine (G). Unlike DNA, the Thymine is replaced with Uracil but Uracil pairs with Adenine [10]. RNA carries the same sequence information as the corresponding non-transcribed strand of DNA. RNA molecules play many roles within the cell from catalyzing biological reactions, to controlling gene expression. There are four major types of RNA: messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and small nuclear RNA (snRNA). mRNA acts as a template from the DNA to the protein, rRNA links the amino acids together to form proteins, tRNA delivers amino acids to the ribosome during translation, and snRNA is confined to the nucleus and plays an important role in the maturation of mRNAs.

1.3 The Flow of Genetic Information

The Central Dogma, proposed by Francis Crick, describes the flow of genetic information. Genetic information, the sequence of bases, passes from nucleic acid to nucleic acid until the unidirectional transfer to protein. Commonly, this process includes two key steps in expression, transcription (DNA to RNA) and translation (RNA to protein) [11].

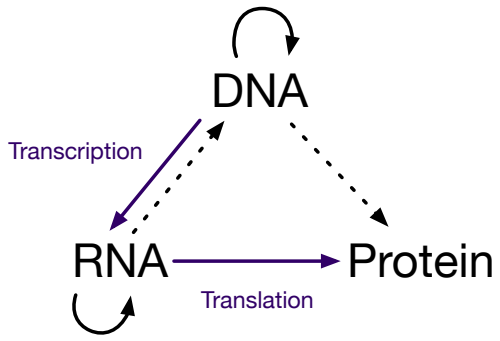

Figure 1.2: The Central Dogma of Biology.

1.3.1 Transcription

Transcription is the first step in this process where the information from the DNA transcribed into mRNA. The stretch of DNA transcribed is called a transcription unit. Transcription starts with

an RNA polymerase and other transcription factors binding to the promoter region of the DNA. The double helix in the DNA is separated by breaking the hydrogen bonds between the complementary nucleotides. Ribonucleotide triphosphates (NTPs) form base pairs with the complementary strand of DNA (the anti-sense strand). These ribonucleotides then join together through an enzyme called RNA polymerase to form the pre-messenger RNA.

In eukaryotes the pre-mRNA undergoes further co-transcriptional modifications to become a mature messenger RNA, mRNA. These modifications include the addition of a 5' cap at the beginning of the mRNA, the addition of a poly-A tail (a string of A nucleotides) at the end of the mRNA, and a process called splicing, where segments of the mRNA are removed. At the end of these modifications, the mRNA molecule is transported out of the nucleus before being transcribed into a polypeptide.

1.3.2 Translation

Translation is where the mRNA is used as a template to build a protein. The strand of mRNA is translated one codon at a time. A codon is comprised of a group of three nucleotides. There are 61 codons, most of these codons translate to a particular amino acid. Some codons translate to the same amino acid, because there are fewer amino acids than there are codons (20 common ones). One codon acts as a starting signal (AUG), and three act as termination signals (UAA, UAG, or UGA). The cytoplasm contains two other molecules that play key roles in translation: ribosomal RNA (rRNA) and transfer RNA (tRNA). Each tRNA molecule carries a single amino acid. At the other end of the tRNA molecule, there is a trinucleotide sequence, called the anticodon, which is complementary to the corresponding codon. The anticodon binds to a codon, and the tRNA molecule deposits its amino acid with the help of the ribosome. The ribosome is divided into a small and a large subunit, during translation these subunits join together on the mRNA strand to help with the formation of the polypeptide.

1.4 Splicing

The mRNA transcript undergoes several modifications. The transcript consists of coding regions (exons) and non-coding regions (introns). One modification that the transcript goes through is called splicing. Splicing is where the non-coding regions are removed, and the coding regions are joined together, forming the final sequence. There are four loosely conserved core sequence regions that contribute to accurate splicing in both plants and animals.

- A 5' splice site at the beginning of the intron, called the donor splice site, which contains a conserved GT.
- A 3' splice site end of the intron, called the acceptor splice site, which contains a conserved AG
- A branch point, with a conserved A located in the range of 18 to 40 nucleotides upstream of the 3' splice site.
- A polypyrimidine tract ('C' and 'U') following the branch point.

While there have been extensive studies into the spliceosomes for yeast and humans, the plant spliceosome has yet to be isolated [12]. The first report on plant in vitro splicing, which is a technique that helped characterize the assembly and composition of the spliceosome in mammals and yeast, was published in 2018 [13].

Since this area is still in development for plants, we will describe the composition and process for the major U2 spliceosome type in metazoans. Splicing occurs over several steps and is catalyzed by small nuclear ribonucleoproteins (snRNPs): U1, U2, U4, U5 and U6. During the first step U1 binds to a complementary sequence at the 5' end of the intron and the pre-mRNA is cleaved in this region. Through the pairing of adenine and guanine, the cleaved end attaches to a conserved branching point downstream and forms a lariat. Next, the snRNPs U2, U4, and U6 help position the 5' end of the intron and U5 helps position the 3' end of the intron so that the 3' end can be cut and joined to the 5' end. The adjacent exons are covalently bound together and the lariat is released.

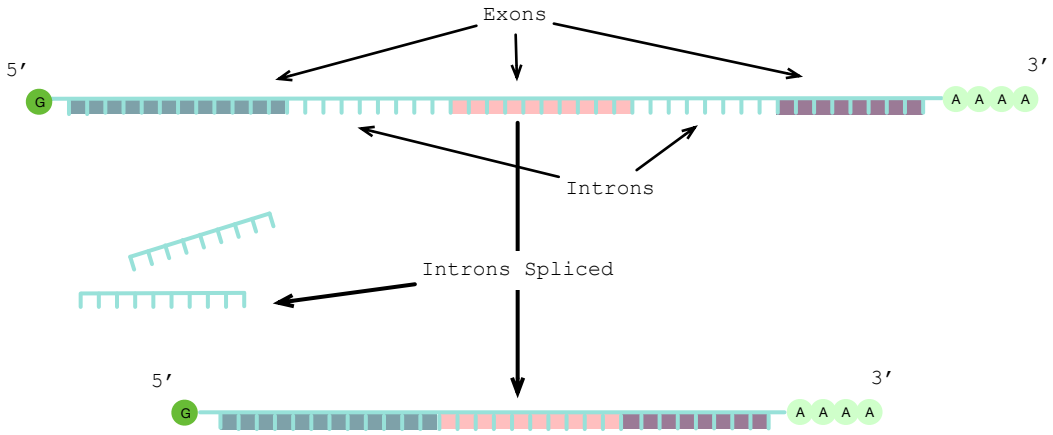

Figure 1.3: How pre-mRNA is spliced during transcription

1.4.1 Alternative Splicing

Shortly after RNA splicing was discovered, scientists learned that if the primary transcript had multiple introns, the introns could be spliced in different ways to form multiple transcripts from a single gene. This means that a single gene could result in multiple protein isoforms which increases the functional capacity of a gene. These splicing patterns change in different tissues and different stresses.

The four splicing features mentioned above are not enough for splice site recognition. There are also *cis*-acting regulatory sequences and bind corresponding *trans*-acting factors. The *cis*-acting regulatory sequences are anywhere between 4 to 18 nucleotides long and can be categorized as splicing enhancers or as splicing silencers. Correspondingly the *trans*-acting factors that interact with these sequences are regulate the recruitment of splice site proteins. Splicing activator proteins bind to the splicing enhancer sites, increasing the probability of a nearby site to be part of a splice junction while the repressor proteins bind to splicing silencer sites and reduce the probability.

During alternative splicing, *cis*-acting regulatory elements in the mRNA sequence determine which portions of the transcript are retained and which are spliced out. These *cis*-acting regulatory elements alter splicing by binding to different *trans*-acting protein factors, such as SR (Serine-Arginine rich) proteins that function as splicing facilitators and heterogeneous nuclear ribonuleo-proteins (hnRNPs) that suppress splicing. The introduction of high-throughput technologies has

made the study of the alternative splicing more viable. The five forms of alternative splicing are shown in Figure 1.4.

1.5 Differential Alternative Splicing and Differential Gene Expression

After we start understanding the different processes that regulate the cell, the question becomes, how can we examine what is happening in the cell with changes in the environment? One common way is to measure the differences in alternative splicing and gene expression.

Differential alternative splicing is when alternative splicing occurs in response to a difference in environment. Several studies have looked at differential alternative splicing due to various biotic and abiotic stresses [14–17]. Some of these conditions include drought [18], heat [19], cold [20], exposure to chemicals [21], nutrient deficiencies [22], and viral and bacterial pathogens.

Alternative splicing can influence gene expression in several ways. On way is by generating transcript isoforms that are targeted by the nonsense mediated decay pathway, thereby down-regulating the gene expression. A second way is by generating protein variants with altered functions, thus allowing for a more varied proteome [14, 23].

Similar to differential alternative splicing, differential gene expression is where a gene is up- or down-regulated between two different biological conditions [24]. If, in two conditions, the difference between the number of sequence fragments that map back to a gene is significantly different (greater than the natural variation), then the gene is differentially expressed [25].

Studying these differences can help elucidate what is happening in the cell. We explore these differences with various conditions to better understand how the cell adapts to stressful environments and survives.

1.5.1 RNA Sequencing Methods

The main goals of RNA-seq are to identify the sequence, structure, and abundance within a sample. Sequencing techniques have been refined over time. Earlier transcriptomic methods,

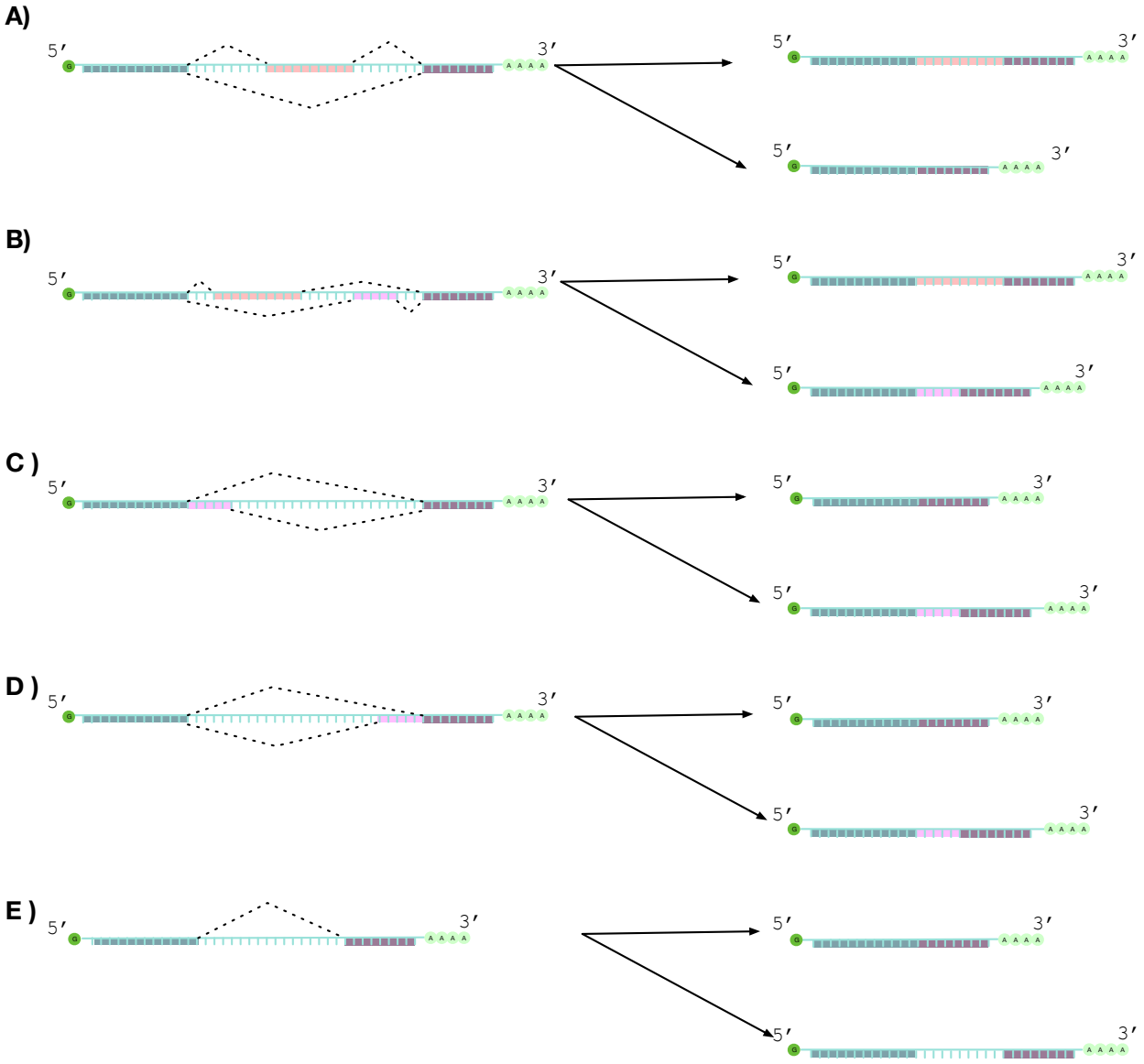

Figure 1.4: Forms of Alternative Splicing. **A) Exon Skipping** is where an exon is spliced out instead of being left in the final transcript. **B) Mutually Exclusive Exons** is where one of two consecutive exons is included in the final transcript. **C) Alternative Donor Sites** is where an alternative 5' splice site is selected, changing the 3' boundary of the upstream exon. This is the most prevalent form of AS in mammals and the most extensively studied. **D) Alternative Acceptor Sites** is where an alternative 3' splice site is selected, changing the 5' boundary of the downstream exon. **E) Intron Retention** is where an intron is retained instead of being spliced out of the mature mRNA. This is the most frequent form in plants but less frequent in mammals.

including Sanger sequencing and microarray expression analysis, provided a strong basis for deciphering genetic structure. High-throughput next generation sequencing, NGS, methods have transformed the field of transcriptomics. Unlike microarrays, RNA-sequencing (RNA-seq) does not depend on prior sequence knowledge. It provides a direct measure of RNA abundance, has more success in quantifying transcripts found in low or high abundance, and is more cost effective [26]. The cost of sequencing the genome went from 100 million dollars in 2001, to less than 10,000 dollars in 2014 [27]. This influx of data has since transformed the field of genomics, allowing a more comprehensive study of the genetic differences and similarities, within and across species [28]. One of the fields that NGS has transformed is transcriptomics. The transcriptome refers to the sum total of all the mRNA molecules expressed from the genes of an organism. Transcriptomics aims to identify and quantify expression levels of individual transcripts. Knowledge of the structure and abundance of the transcriptome helps researchers interpret expression in different tissues under normal conditions, and how this changes under different stresses.

RNA-Seq is widely used to compare gene expression between different experimental conditions, characterize alternative splicing, to look at mutations and to build co-expression networks. This data can also be used to discover novel exon-intron boundaries and verify current gene annotations.

Since RNA-seq opens up the possibility of studying gene expression in more depth, many research techniques are transitioning towards using RNA-Seq for understanding alternative splicing and differential expression.

Chapter 2

Biological Reproducibility, Replicability, and Data Reusability

Reproducibility, replicability and data reusability have recently gained a lot of attention in the research community [29–35]. This attention is a result of changes in how we handle and present data. One major change is that computers are more commonly used to perform scientific analyses and have allowed for the analysis of increasingly large amounts of data. The second major change is a positive shift within the scientific community towards sharing resources. The National Institutes of Heath (NIH) states that “rapid and unrestricted sharing of data and research resources is essential for advancing research on human heath and infectious diseases" [36]. In 2016, the NIH mandated that data be shared from all trials that use NIH funding [37].

Several foundations have also started to require resource sharing [38]. The Gordon and Betty Moor Foundation requires that all foundation-funded projects make data widely available [39]. Pediatric cancer foundations such as St. Baldrick’s Foundation and Alex’s Lemonade Stand Foundation for Childhood Cancer have incorporated data sharing policies into their grant processes [40]. The Bill and Melinda Gates foundation has adopted an open access policy on data that the foundation financially supports. This policy requires that associated data from peer reviewed publications be accessible to the public [41]. Because of policies like these and advances in technology, research groups have put more effort into making data easily accessible and developing frameworks to communicate the methods used to process and analyze this data. Even with these efforts, technical and descriptive barriers are still a major hurdle; we will introduce a few issues we encountered during this research.

In 2016, Nature conducted an online survey about reproducibility in research. Out of 1,576 participants, 52% believed there is a significant crisis of reproducibility in research, 70% of those surveyed had failed to reproduce results from another lab, and more than half admitted to failing

to reproduce results from experiments within their own labs [34]. In 2015, a paper published in PLOS Biology analyzed past studies and estimated that published irreproducible clinical research exceeds 50%, resulting in roughly $28 billion dollars spent per year in the USA on irreproducible research [35].

To address this issue, many journals, institutions and individual researchers are creating guidelines and recommendations for how to achieve reproducibility [30, 35]. First, we should define what these terms mean:

- **Reproducibility** refers to the ability for fellow researchers to duplicate the findings of a study using the same methods and the same data [29–31, 33].

- **Replicability** refers to the ability for fellow researchers to duplicate (within a measure of standard error) the findings of a study using the same methods and **new data** [30, 33].

Several academic libraries, including Elsevier, Frontiers, Nature Research, and Springer Nature, have signed the Transparency and Openness Promotion (TOP) Guidelines [42]. These guidelines were established in 2013 by the Center for Open Science to increase openness, integrity and reproducibility in research [43].

Increasing the standards for reproducibility and replicability also helps encourage data reuse. For results to be replicable, enough information must be provided for different labs to thoroughly understand the sequence of steps used to prepare the data.

There are many factors that affect reevaluation and reuse of data and scientific results, including: the initial design of experiments, careful recording of lab procedures, statistically sound analyses, and data reporting and deposition. We will focus on the area of data reporting because this greatly impacted our work.

2.0.1 Incomplete Metadata

For our research project, we downloaded and processed the raw data from over 80 studies, ultimately using 35 of these studies. Please refer to the methods section of this paper to see how

we chose this data. Even in the studies that we used, the most prevalent issue we encountered was incomplete or poorly annotated metadata from the Sequence Read Archive (SRA). Well-annotated metadata is essential for replication, reproducibility and integration of the data into other research questions.

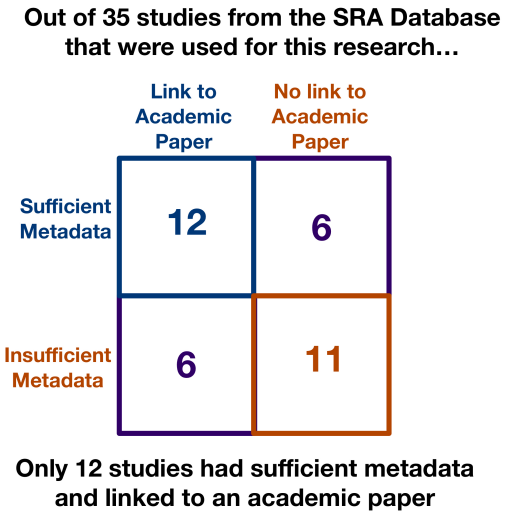

Figure 2.1: Inadequate annotation of studies in the SRA for *A. thaliana*

As illustrated in Figure 2.1, 18 out of the 35 studies did not reference an academic paper and 11 out of those 18 studies did not provide enough details within the metadata to adequately describe the growth conditions and/or treatment of the plants. For further information, refer to Table 2.1. Groups sharing data are doing so to benefit the scientific community and fulfill requirements for funding and publishing. Because the global reuse of data is still a new concept, research groups are unaware of the problems that other researchers will encounter when attempting to use their data. Some of the most common fields missing in the metadata included:

- The age of the plants
- The growth protocols for the plants (e.g. soil composition, temperature, light, etc)
- The quantity and duration of treatments applied to the plants

Sometimes, important sections of the data were missing and we did not have a complete picture of the experiment until we read the corresponding paper. Often there was no link to the corresponding paper. For example, one study applied multiple conditions to the plants. Because of this, there were multiple control samples, which were labeled as control treatments or mock treatments. It was impossible to pair these control samples with the treatment samples without a visual aid from the paper that used the same language [44, 45]. Another study used time points instead of hours to specify the progression of the experiment. The submitter specified that the time points were 3 hours apart but did not clarify whether the first time point was taken 0 or 3 hours into the experiment. This was clarified by a visual aid provided in the associated paper [14]. Without this information, it is impossible to accurately identify and incorporate hidden sources of variability.

To compensate for incomplete metadata and missing links to academic papers, we spent a substantial amount of time searching for the associated papers for each study and filling in the metadata from the methods and materials sections.

Finding these papers was not a straight forward task. We searched for an associated academic paper using the following information, when provided:

- Terminology provided by the abstract and from existing data
- The associated institution
- The contributor's name/s

We then confirmed the paper by matching details such as authors, institution, year published, and methods. We were able to find all but 2 of the 18 academic papers that were missing from the associated studies. While not all data has published results, when possible, it is extremely important to link the paper to promote replication and data reuse within the scientific community.

2.0.2 Inconsistent Terminology for Identification of Metadata

The second issue we faced was the identification of biological characteristics for each replicate (or run) in the SRA database. When looking at a study or experiment in the SRA database, a

Table 2.1: Additional information detailing if SRA experiments had sufficient metadata and if the data linked to an associated academic paper. Column one provides the accession. Column two provides a true or false value for whether there was sufficient data. Sufficient data includes details about growth conditions, age of plants, tissue type, and the amount and length of the treatment stress. Column three provides a true or false value for whether there was a link to an associated academic paper.

* indicates that an academic paper was linked to the study after research was finished.

SRA Accession	Sufficient Data	Associated Paper
DRP003686	F	T
DRP004486	F	F
ERP022071	F	F
ERP111840	T	T
SRP011128	F	F
SRP026260	F	T
SRP058527	F	T
SRP060410	F	T
SRP073212	F	F
SRP073711	T	T
SRP074485	T	T
SRP074890	F	T
SRP076862	T	T
SRP081056	T	T
SRP082177	F	T
SRP090416	T	T
SRP091014	F	F
SRP091628	T	F
SRP096554	T	T
SRP097690	T	F *
SRP101403	T	T
SRP102893	F	F
SRP107981	T	T
SRP108611	T	T
SRP124769	T	T
SRP134263	T	F
SRP136536	T	F
SRP145580	T	T
SRP148881	T	F
SRP155798	T	F
SRP156748	F	F
SRP161785	F	F
SRP162472	F	F
SRP177951	F	F *
SRP187477	F	F

researcher can download the RunTable which gives the researcher metadata associated with each replicate.

12 Runs found

Run	BioSample	Sample name	Experiment	MBases	MBytes	genotype	source name	treatment
SRR1909030	SAMN03397403	GSM1631206	SRX950685	3,778	2,418	rbm25-1	rbm25-1 ABA_3	0.1 mM ABA, 6 hours
SRR1909029	SAMN03397400	GSM1631205	SRX950684	3,736	2,411	rbm25-1	rbm25-1 ABA_2	0.1 mM ABA, 6 hours

20 Runs found

Run	BioSample	Sample name	Experiment	MBases	MBytes	genotype variation	infection	source name	time
SRR1257408	SAMN02730145	GSM1371370	SRX522636	1,032	701	T474D	infected with CaLCuV	T474D_infected20	20 dpi
SRR1257407	SAMN02730149	GSM1371369	SRX522635	1,729	1,176	T474D	infected with CaLCuV	T474D_infected20	20 dpi

11 Runs found

Run	BioSample	Sample name	Library name	Alias	Experiment	MBases	MBytes	SRA accession	Title	genotype
ERR2869074	SAMEA5066807	SAMEA5066807 ERS2878010	Col-0 H2O rep_a_p	E-MTAB-7374:Col-0 H2O rep_a	ERX2874949	9,462	3,302	ERS2878010	Col-0 H2O rep_a	wild type genotype
ERR2869075	SAMEA5066808	SAMEA5066808 ERS2878011	Col-0 H2O rep_b_p	E-MTAB-7374:Col-0 H2O rep_b	ERX2874950	6,481	2,275	ERS2878011	Col-0 H2O rep_b	wild type genotype

Figure 2.2: Examples of RunTable identifiers from the Sequence Read Archive

Several examples of RunTables from the SRA database are shown in Figure 2.2. While exploring different studies stored in the SRA database we noticed that the identifiers used to describe biological and technical replicates could be split into two categories. The first set of identifiers appear to act as primary keys to the database and uniquely identify each biological and technical replicate, examples of which include Run, BioSample, and Experiment. The second set of identifiers are manually entered by the contributor of the data and describe the biological characteristics of the replicate. Examples of these identifiers include genotype, ecotype, treatment, etc. This second set of identifiers do not adhere to a controlled vocabulary. Instead the contributors of the data use semi-structured textual descriptions to describe growth conditions, biological attributes, and treatment protocols. Identifiers that are not globally unique and persistent create a source of unnecessary complexity and ambiguity [46–50].

Inconsistent identifiers lead to several major issues:

- Inability to create reusable scripts to incorporate essential details from the experiment
- Difficulty in identifying relevant data to a particular experimental protocol or outcome
- Higher occurrence of incomplete metadata
- Increased chance of the metadata being misinterpreted

Furthermore, this leads to an inability to replicate past experiments and causes uncertainty with data reuse. Table 2.2 gives an example of database keys from the SRA metadata of several studies that were used during this research.

Table 2.2: List of terms used by different research groups as identifiers to describe the metadata for *A. Thaliana* in the SRA RunTable. Each column begins with the identifier we chose followed by a list of terms that we encountered to represent the same element.

ecotype	**genotype**	**age**	**tissue**	**treatment**	**time**
accession	genotype	age	body_site	co2_condition	duration_of_treatment
cultivar	genotype_type	dev_stage	organ	exposure_time	exposure_time
ecotype	genome_variation	developmental_stage	source_name	growth_condition	time
ecotype_background	source_name	Stage	tissue	growth_conditions	time_point
isolate				infection	
strain				label	
source_name				light_treatment	
				sample_name	
				Sample_name	
				source_name	
				stress	
				treated_with	
				treatment	
				Title	

2.0.3 Current Solutions

The ability to store and share data in public repositories is still a new part of the academic landscape. We have highlighted that incomplete and non-standardized metadata is a prevalent issue within the SRA database. This issue also touches other popular data archives, such as GEO, GenBank, and ArrayExpress. [46, 51–54]. Researchers face challenges with sharing data such as additional costs and time, risks that errors within published work will be exposed and that data the research group produced will be used by other labs, causing a loss of future opportunities or credit [38]. We believe that the benefits of sharing data in a reusable way outweigh these risks. Sharing data makes previous work easier to evaluate through reproduction and strengthens scientific conclusions through replication [30]. We will explore some of the current approaches used to correct metadata.

The first approach is educating researchers publishing data about data management plans and best practices. Journals, foundations, and governmental agencies are requiring data management and stewardship plans for data generated during research to help maximize investments [55]. In 2016, the FAIR Guiding Principles for scientific data management was published in Scientific Data and has been endorsed by the G7, the European Commission, and the NIH [56]. FAIR stands for Findability, Accessibility, Interoperability, and Reusability and gives a set of guidelines promoting 'good data management' [55]. The goal of this initiative is to create an environment where data is easy to discover through computational methods. While FAIR guidelines are being accepted by funding agencies worldwide, they have been slower to gain traction with individual labs [57, 58]. Over the last year there has been huge "organizational (International Data Week), technological (Google), and policy driven strides (GO FAIR)" that are believed to further clarify the guidelines and boost education [57].

The second approach involves manual curation efforts after the studies and data have been published [59]. One example of manual curation is through crowd-sourcing. This is where groups of people volunteer or are paid to quality assess and fix metadata. One example of crowd-sourcing metadata includes a web portal called CRowd Extracted Expression of Differential signatures (CREEDS) and Metacrowd. CREEDS provides annotated differential expression signatures and was initiated via an online class through Coursera where 70 participants annotated 2460 single-gene perturbation signatures, 839 disease signatures, and 906 drug perturbation signatures from GEO [48]. Another example of crowd-sourcing project is called MetaCrowd. Metacrowd uses a crowd-sourcing platform that connects companies with a distributed group of people who can annotate datasets, called CrowdFlower (Now called Figure-Eight). These annotators isolate metadata from datasets in the GEO database. As an initial experiment, the annotators were provided six of the most frequently occurring keys (age, cell line, disease, strain, tissue, and treatment) along with variants of these keys, totalling up to 355 terms. These terms were separated into lexical, value, and concept similarity groupings. These terms were also grouped computationally and then both the crowd-sourcing and the computational groupings were compared to an answer key. The project

ultimately hopes to use a mix of crowd-sourcing and computational algorithms to help assess and fix metadata. [60].

Another example of manual curation is the use of biocurators. Biocurators consist of people with expertise in both biology and computational skills. Biocurators extract biological information from scientific literature, integrate it into databases, and communicate with researchers to ensure accuracy of metadata and promote data exchange [59, 61]. Standardized labelling systems through ontologies are gaining more attention as open data initiatives become more popular [62]. Some examples include the Plant Ontology (PO), the Plant Trait Ontology (TO), and the Plant Experimental Conditions Ontology (PECO) [63]. These classification systems have helped allow parts of the curation to automated, but there is still a need for manual curation to help draw more complex connections between studies [63–65].

Similar to manual curation, multi-experiment compendiums exist where different research labs have gathered and processed raw data using the same computational pipeline. This creates a centralized location where other researchers can access ready to analyze summaries [5, 6, 66, 67]. Because these datasets are often limited to a small set of species, there is more of an opportunity to standardize the metadata. For example the Expression Atlas is a multi-species compendium in which there is a group of biocurators who extract information from the literature to ensure accuracy and enrich the annotations [68]. Examples of these packages will be described in the next chapter.

Several computational packages have been developed that attempt to improve the curation of metadata. These methods can be categorized into two approaches: automated natural language processing (NLP) and inferring metadata from gene expression profiles [48]. The first method, automated natural language processing uses computational algorithms to extract information from existing metadata and related academic journals. Examples of this include: GEOMMTX [69], MetaSRA [54] and GeoBoost [51]. The second method uses machine learning models to predict metadata, such as tissue type, gender, and library type, via the analysis of expression data [48]. Examples include:

- Tissue and cell type prediction using SHARQ [70], URSA [71], CIBERSORT [72], and xCell [73]
- Gender prediction using MassiR [74]
- Age, gender and tissue type prediction using phenopredict [75] and the automated label extraction [76]

Predicting more complex attributes such as growth protocols and treatment remain elusive and still have to be manually curated [48].

As described earlier, we manually curated our metadata. This was extremely time intensive and difficult due to lack of standardization between labs. With the amount of data being generated and the initiatives towards data sharing, there is a growing need for scalable solutions.

While this section covered the biological reproducibility there are other aspects of reproducible that we were able to control. The next chapter will cover technological concerns of reproducibility like data processing and go into more depth about compendiums that already exist.

Chapter 3

Computational Reproducibility

In the previous chapter, we highlighted the idea of a "reproducibility crisis". One reason research has been difficult to reproduce and replicate is due to incomplete and inconsistent experimental descriptions [77]. Another facet of the crisis involves the technical aspects of experiments. This relates to the incomplete or inaccurate descriptions of how the raw data was processed, and an inability to download, run, or find corresponding software packages [3, 29, 78–80]. These issues are being critically examined in what has recently been called a "software crisis" [81, 82].

In the last fifteen years, the amount of publicly available genomic sequencing data has doubled every few months [1–3]. With more data being produced, many individualized pipelines are being created to analyze different scientific questions. These pipelines are often poorly described and irreproducible [83].

Computational reproducibility requires the following information:

- The code used to process the raw data
- The order of execution to run the code and packages
- A computing environment that includes software packages and dependencies

Each part has a corresponding tool or set of tools that aids the developer by compartmentalization and organization. This includes version control systems, workflow management tools, and package management systems. We will go over what these are, how they support reproducibility and replicability and examples that have gained favor in the bioinformatics community. At the end of the chapter, we will summarize current online compendiums and the pipelines used to re-analyze groups of expression data.

3.1 Version Control

Version control helps the developer keep track of code being used in the research project. There are many benefits to including a version control system in your research. The first benefit is that these tools enable the researcher to document the evolution of the computational research project. Version control keeps track of changes made to files and directories. This is similar to a laboratory notebook for scientific computing, in that it keeps a lasting record of events [84]. Originally it was developed to help programmers coordinate software projects, but these tools have spread to the scientific domain for the following reasons:

- Changes made to source code, documentation, and academic papers are checked in and time stamped allowing anyone to easily navigate to previous versions
- Promotes development by providing a backup of the project and the ability to coordinate over multiple computers
- Allows multiple people to concurrently work on the same project

The second benefit of using version control is being able to easily communicate new research with the scientific community. Version control helps scientists around the world collaborate together. One way version control encourages collaboration is by giving scientists who are working in remote locations, without access to the internet, a way to work asynchronously and integrate their work into the project at a later date [85]. Version control can also help new contributors understand the progression of a research project. Because of these features, several groups have suggested that version control facilitates scientific reproducibility and advances research transparency [84, 86, 87].

We will discuss advantages of using one of the most widely used version control tools, Git [84, 85]. Git is a version control system that was created in 2005 by Linus Torvalds [88]. Since its introduction, Git has been widely adopted [86]. One reason for the success is that this tool allows each researcher to keep an individual copy of the project on their local machine. New updates can be pulled into the project (keeping the local copy up to date) and new features or corrections can

be shared (allowing other researchers to benefit from each others' work). This is called distributed version control, and it ensures that there is no single point of failure [85].

Another advantage is that Git is the backbone of GitHub. GitHub is a Git repository hosting service that provides a website using a graphical interface. This is a space where the community can contribute to a shared repository. GitHub also provides additional features like a linked wiki for documentation and task management tools to enhance productivity [89].

As shown in figure 3.1 GitHub has been increasingly used and cited in academic papers.

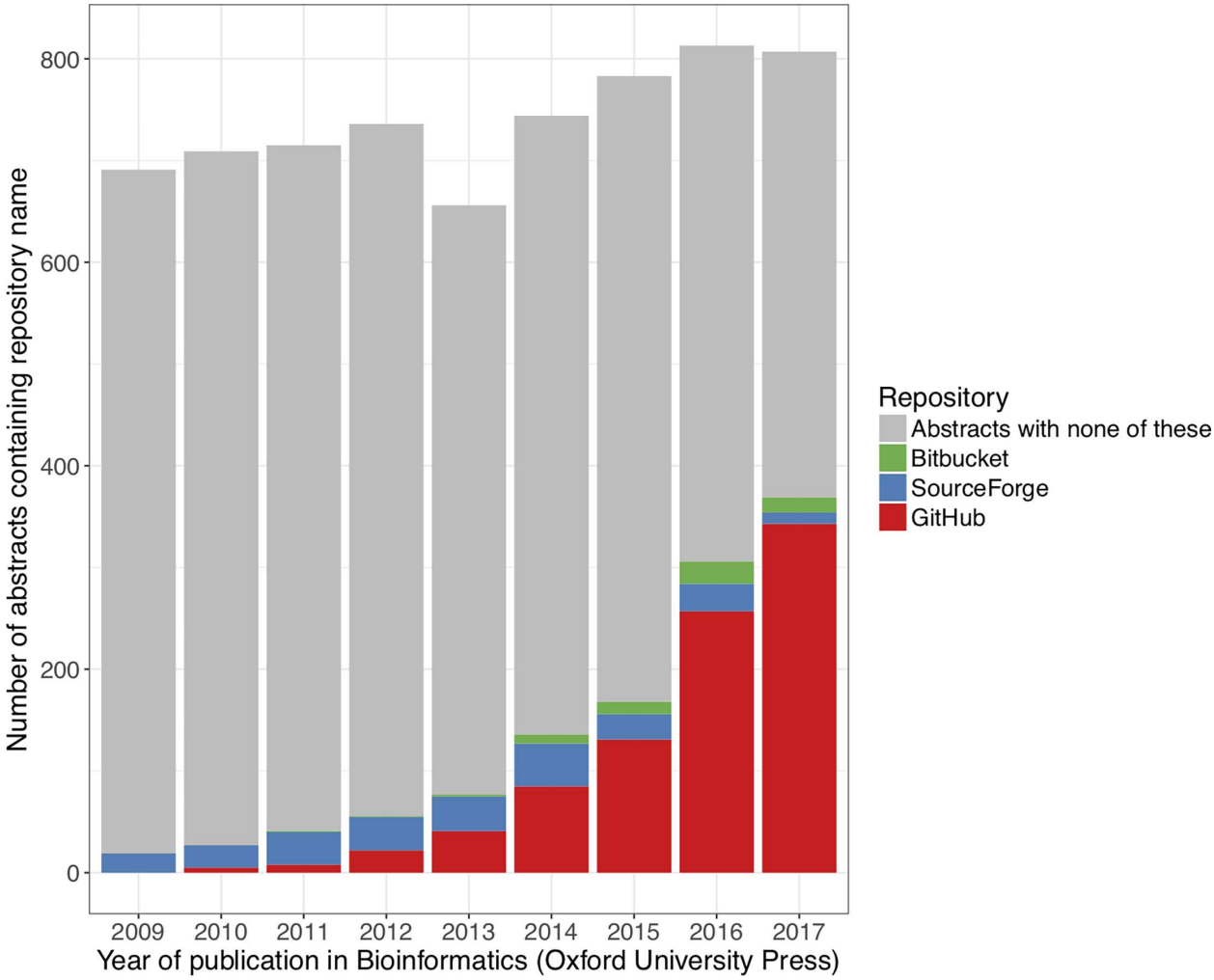

Figure 3.1: The number of papers from Bioinformatics between 2009 and 2017 that reference a version controlled repository name in the title or the abstract of the paper, from Russell et al [79].

Another serious issue concerning scientific reproducibility is finding relevant software that will be embraced and maintained within the scientific community. Finding relevant software can be hindered by the rate of publication, dubbed software lag. Even when the paper is published and the software is released, it is hard to identify software packages that will gain popularity within the field [82, 84]. If the software is not maintained, then it goes through a process coined software collapse. Software collapse is the idea that software, if not maintained, will get to the point that it is no longer usable [90]. This makes reproducing the results of any published paper that depends on that software onerous.

The question then becomes how to identify software that will remain relevant and not undergo software collapse. It has been shown that depending solely on a journal's impact factor is a poor predictor for a software tool's longevity [91, 92]. Alternative metrics have been proposed as an additional way to predict this [91–93]. Examples of these metrics include quantifying mentions from forums like BioStars, social media platforms like Twitter, and community lists curated by experts in the field [82, 91, 94]. Because GitHub has been so widely adopted, it provides metrics about software popularity [82]. Here are examples of two sets of alternative metrics provided by GitHub:

- The amount of stars, forks, and watchers [91]. The higher these metrics are, the more interest the tool has elicited [82].
- The amount of commits, contributors and how recently the project has been worked on. This can help identify how actively a tool is maintained [79].

Once a potential software package has been identified in GitHub, the interface provides additional insights. Through the issues section, the researcher is able to see types of issues encountered and if the contributors actively deal with bugs and questions from the community. The insights section presents graphs that give the researcher a better feeling about who has been contributing and what has been changing within the project. This is important because a more accepted and actively developed tool correlates to an easier installation and a larger community provides more

support through documentation and forums. It is important to note that any metric that relies on community support can be skewed. An example of this includes more popular researchers being cited more frequently, telling us more about the researchers than the software. Because of this researchers can target the metrics maximizing exposure and inflating the prestige of the research or software [92].

As we discussed in section 2.0.3, journals and funding agencies are requiring that raw sequencing data be made publicly available. Similarly, more academic journals are requiring that software be made publicly available. Two journals that currently include GitHub in their peer review process are the Journal of Open Source Software and ReScience [95]. The Journal of Open Source software was created to provide software developers a cost effective, streamlined platform to introduce new software to the community [95]. ReScience is a peer-reviewed journal, launched in 2015, that focuses on replication of computational research [96]. ReScience uses and publishes via GitHub and harbors the philosophy that if software can be replicated (which in this situation means a newly implemented algorithm achieving the same results) then there is enough information between the two publications to confidently be able to replicate that algorithm in the future [82, 96].

eLife and Nature also require or strongly recommend the use of GitHub for making software available. eLife launched a repository in 2017 and requires all papers introducing novel software to copy the software to that account [97, 98]. Nature Research recommends depositing code on a community repository like GitHub [99]. We believe that as software collapse gains more attention, repositories like GitHub will become a standard for ensuring reproducibility within the scientific community.

3.2 Workflow Management Tools

Part of advancing research transparency includes keeping precise records of each software package used to process the data and what parameters were chosen. The sequence of programmatic steps used to process the raw sequencing data is called a pipeline or a workflow [100]. Historically, scientists have depended on collecting or creating multiple programmatic scripts that need to be

run in a specific order to process the data [100]. Each of these steps can be done manually but important details needed to retrace the analysis are often accidentally left out [101–103]. Workflow management tools help record details needed to reproduce the analysis and quickly process new sets of data.

It is desirable for these tools to accommodate other features like:

- The ability to parallelize tasks for efficient performance [104]
- The ability to easily integrate new tools [105]
- The ability to easily share workflows with the scientific community [104]
- Fault tolerance, enabling automatic or manual restarts if a workflow fails [106]

There are several styles of workflow management tools. Some systems such as: BioPipe [101], Galaxy [102], GenePattern [107], GeneProf [108], Mobyle [109], PegaSys [110], and Taverna [111] are graphical interfaces that use visual programming to coordinate pre-configured tools [100]. Visual programming is an approach where the developer drags and drops programmatic elements onto a graphical interface creating a workflow [112]. These tools tend to be less flexible because it is harder to integrate new tools or custom scripts but are easier to use for scientists that have limited software engineering experience [102].

Other systems like Bpipe [113], GXP Make [114], Pwrake [115], Ruffus [116], and Snakemake [117] use a text based workflow. Text based workflows are often more flexible and can be easier for developers to collaborate through version control, but are more technically involved [100, 117].

We chose Snakemake for our workflow manager for the following reasons. The first reason has to do with a strong community. Snakemake has been actively developed since 2011 with over 100 contributors. Johannes Köster, the creator, has been a part of the project since the beginning, which has provided a clear direction. Snakemake has kept current with newer technologies like the ability to modularize workflows for reuse, integration of the Common Workflow Language (CWL), and integration with Docker and BioContainers [104]. It is extremely important that the software

making the workflow tool does not break. It is inconvenient if one of the tools used in the analyses undergoes software collapse but easier to trade out for a similar tool. If the software used for the workflow breaks, the entire pipeline has to be rebuilt, which is a much larger investment in time.

The second reason has to do with how well it integrates in with our lab. Snakemake uses Python, which is one of the primary languages our lab uses, so workflows can be easily shared and extended. Snakemake is semantically familiar because it is based off of GNU Make. Make is a build automation tool that allows the developer to create executables based on rules whose execution is triggered by the absence or modification of dependencies [118].

In the last section, we discussed how alternative metrics can indicate if a tool has been accepted within the scientific community. When we started our research we searched on the BioStar forums for popular pipelines and found that Snakemake had a lot of positive feedback. Figure 3.2 provides another example of alternative metrics. This graph shows two metrics that were aggregated by Albert Vilella at the end of 2018. the first alternative metric quantifies a tools popularity on GitHub. The score is calculated using the following formula $watch + (star/5) + (fork * 10)$. Dr. Vilella amplified or decreased the value of each metric because they show a different level of commitment to that repository. Watch is the number of people who have chosen to receive notifications about the repository, star indicates the number of people who are interested in the repository, and fork represents the number of people who have copied the repository. The second metric is a poll shared on Twitter by Albert Vilella. The poll drew the participation of around 500 people. We believe this is another indicator that Snakemake has remained popular since its release. Ultimately, the decision in workflow management tools depends on the requirements of the research project and existing skills within the lab. All of these tools strive to ensure reproducibility by automating data processing and each tool has its own strengths and syntax.

3.3 Package Management and Continuous Analysis

Once you have created a workflow, the next step is to make sure that the workflow is reproducible and shareable. As mentioned in the previous section, a developer can take measures

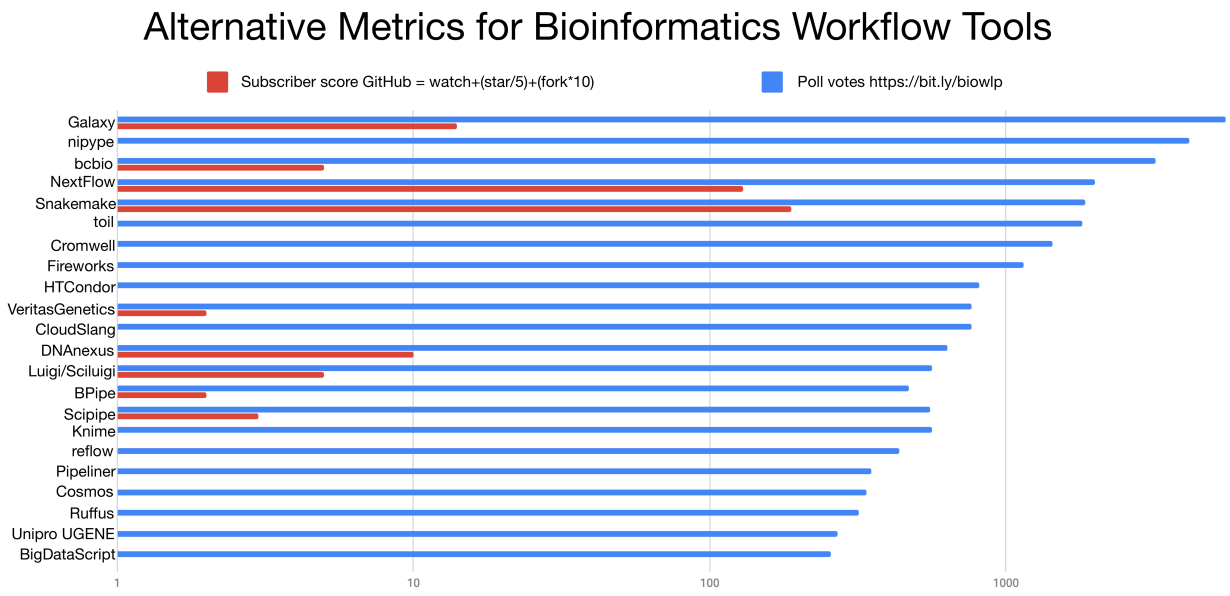

Figure 3.2: This chart shows alternative metrics aggregated by Albert Vilella: 1) Quantifies popularity on GitHub* by scaling the number of users who watch, fork, and star the repository 2) Results from an online poll shared on Twitter in December 2018 asking participants about preferences in Bioinformatics tools. Figure obtained from https://bit.ly/biowl

* Snakemake is not on GitHub, it is hosted by similar service, BitBucket. This is important to take into consideration because BitBucket is not as widely used and might influence the metrics.

towards using packages that will remain updated, however there is no guarantee that the software will be maintained.

Processing sequence data is complex and often depends on the use of multiple packages. Because of the complexity, there are several challenges to reproducing and sharing these workflows. Keeping the versions of these packages and their dependencies is essential towards transparency and reproducibility [94]. Two ways this can be accomplished are package management systems and containers.

Package management systems keep track of software versions and dependencies of tools used in a workflow [119]. Conda is a package management system that also serves as an environment management system, and has been widely accepted in the Bioinformatics community. Conda allows the user to define and build a group of software packages in a local environment. These isolated environments prevent conflicts when different analyses require different versions of the

same package. BioConda is a software repository for Conda that provides and maintains over 6,000 recipes for downloading and running bioinformatics packages [120, 121]. This has been a valuable tool in alleviating some of the pain that comes from trying and managing new software.

Containers are another way to keep track of software versions and tool dependencies. Each package added to Bioconda also has a corresponding Docker BioContainer automatically uploaded to an image registry site called Quay [121]. Developers can create an images which represent a software environment. An image is a self-contained, read-only snapshot of a group of applications, packages, and whatever the software needs to run (operating system, dependencies, etc) [122]. Then anyone else can download that image and run it using a container program. There are several container programs available, including Docker [122], Singularity [123], and Shifter [124]. We will discuss Docker because it is one of the most widely adopted container technologies within the scientific community [125]. Docker images can be several gigabytes in size but only have to be downloaded once and can be started quickly and with minimal overhead [125]. This technology integrates with cloud computing platforms such as Google, Amazon and Microsoft [125].

Docker has been suggested as a way to reproduce research by removing dependency management and limiting the effects of software collapse [122, 125]. Similar to Git, Docker images can be tagged so that older versions that correspond to academic papers can be easily accessed and reproduced.

Biocontainers is a community based project that provides infrastructure and guidelines to create, manage and distribute bioinformatics containers [121]. All Biocontainers are Docker based and both Conda and Docker interface with Snakemake [126].

These tools all support reproducibility and replicability but the responsibility still remains with the researcher. There are several other recommendations for enhancing reproducibility. The first involves producing logs for each step in the analysis. Logging each step provides additional information for further assessment and way to help debug the workflow if an error is encountered. The second recommendation is to include a file that logs the specific versions of packages used along with the results. The last recommendation is include some type of automatic regression

testing whenever new tool versions are available. Continuous analysis involves rerunning all code in a pipeline on a known dataset when there are changes to the code, new updates for existing packages, or additions to the pipeline [125]. When paired with logging, the researcher is able to quickly check and make sure that the additions did not cause anything to break and that there were no significant changes to the results.

We used Git as a version control tool and Conda as both a package manager and a virtual environment in our project. Git was used to backup our scripts and enabled us to easily work from our home computers and the research computers at the university. Conda was used to download bioinformatics packages and to create two virtual environments in order to accommodate tools that used Python 2 versus tools that used Python 3.

3.4 Data Storage and Existing Compendiums

The rise of computational science with new technology, methodological advances, and increased computing power has dramatically increased our ability to collect more complex and higher dimensional data [127]. The Sequence Read Archive (SRA) and the Gene Expression Omnibus (GEO) were established in response to a need of a public repository for high-throughput gene expression data [128, 129].

The Gene Expression Omnibus (GEO) was established in 2000 as a worldwide resource for gene expression studies [130]. GEO contains the gene expression profiles for microarray and high throughput studies, with written descriptions of experimental design and associated metadata. GEO requires the researcher to provide any raw data that was generated during the experiment, and collaborates with the Sequence Read Archive to store these files [131]. GEO strives to make the data accessible to the research community, allowing researchers who do not have the resources or skills to analyze their own genomics experiments ready to use data [130]. While GEO is a valuable resource, each study is processed distinctly which, as mentioned above, makes it difficult to compare.

The Sequence Read Archive (SRA) is the National Institutes of Health (NIH) primary archive for high-throughput sequencing data and is a part of the International Nucleotide Sequence Database Collaboration (INSDC). This collaboration includes: the Sequence Read Archive (SRA), the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ). Data submitted to any of these three organizations is shared. Submission of raw data into the SRA is mandated by most funding agencies and open access journals.

While massive amounts of of data are submitted to the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA), it is not processed using a uniform workflow. There has been a focus on creating computational pipelines that connect various software tools and give a standardized way to store versions and parameters. Focusing on reproducibility improves the quality of published research, makes it easier to analyze multiple datasets from different labs, and speeds up progress by promoting the reuse and repurposing of data. Because of these issues, several research groups have collated and processed a subset of this raw data using a uniform pipeline. We will mention existing pipelines and what they provide.

Gemma was established in 2012 by the Pavlidis Lab at the University of British Columbia. Gemma provides a set of tools for meta-analysis of genomics data and currently hosts coexpression and differential expression results from 10,542 studies [4]. Data can be submitted by registered users or has been obtained from the GEO database. Each dataset undergoes automated [69] and manual curation of the metadata using established ontologies to standardize the vocabulary and make it easier to explore the datasets. After the metadata is curated, the data is analyzed by performing sequence analysis and gene assignment with a current annotation to enable comparisons across platforms [4].

The Expression Atlas is hosted by the European Molecular Biology Laboratory in the European Bioinformatics Institute. Experiments are categorized as differential or baseline depending on what purpose the data was generated for. The expression data is processed by a pipeline named iRAP [132]. Currently there are 3,564 datasets across 43 genomes. From these, 809 of the datasets are RNA-Seq based. As stated on the website, each dataset is manually curated by PhD

biologists who extract and structure information from the literature to accurately represent each experiment [5, 7, 68].

The Recount Project processes publicly available human RNA-Seq data into expression data for genes, exons, exon-exon splice junctions and base-level coverage using a pipeline named Rail-RNA [66]. The data is available for bulk download through an R package as RangedSummarized-Experiment objects or via the website [6, 133].

ARCHS4 is a web resource that is produced by the Ma'ayan Lab at the Icahn School of Medicine at Mount Sinai. ARCHS4 gives access to the gene and transcript levels from published studies on human and mouse RNA-seq data. The raw data is processed through a pipeline using a web-service called Elysium [67] via AWS. ARCHS4 focuses on providing a pipeline that is publicly available through AWS, and on cost effective processing of the data, to help accommodate such a rapidly growing field [8].

These compendiums serve an important role in scientific discovery. They promote data reuse and because multiple experiments are processed in the same way, the pipelines help ensure technical reproducibility. There are hardly any compendiums available that focus on plants, in comparison to mammals. In order to look at larger patterns in deferentially expressed genes and deferentially retained introns, we created a pipeline and collected public sequencing data for Arabidopsis thaliana.

Chapter 4

Data Processing and Results

In this chapter we break down each step used to process the sequencing data. As mentioned in Chapter 3, we used Snakemake to manage the acquisition and processing of each dataset. Figure 4.1 outlines the steps for gathering, processing and analyzing data.

4.1 Data Acquisition

We chose to collect data for *Arabidopsis thaliana* because it is one of the most widely studied plants and consequently there is access to a large amount of high quality public sequencing data. *Arabidopsis* is a small flowering plant in the mustard family and is a standard reference plant for biology [134, 135]. Though, at first glace, it is not obvious why *Arabidopsis* was chosen as a reference, it has several advantageous features. These include a short life cycle, a wide breadth of natural variation in physiological traits and a compact genome [135, 136]. The *Arabidopsis* genome is composed of 5 chromosomes and is around 135 mega base pairs. At a meeting held by the Arabidopsis Genome Project in 1989 at the NSF, a goal was set to completely sequence the *Arabidopsis* genome by the year 2000. This meeting brought together researchers from three continents and led to *Arabidopsis* being the first plant and the third multi-cellular organism to be completely sequenced [134–137].

We searched for *Arabidopsis* experiments on the SRA. As mentioned in Chapter 2, the metadata is not standardized so we used a combination of keywords to maximize the amount of data we collected. These keywords included: *Arabidopsis thaliana*, *A. thaliana*, *Arabidopsis*, Col-0, WT (and these keywords paired with different treatment stresses). Once a dataset was identified, the RunInfo table was downloaded. The RunInfo table is a file provided by the SRA that holds the treatment conditions and replicate structure of all of the samples in a specific study. We extracted the metadata using a custom Python script and then manually standardized our results using techniques discussed in Chapter 2. This script was manually adapted per experiment to correspond

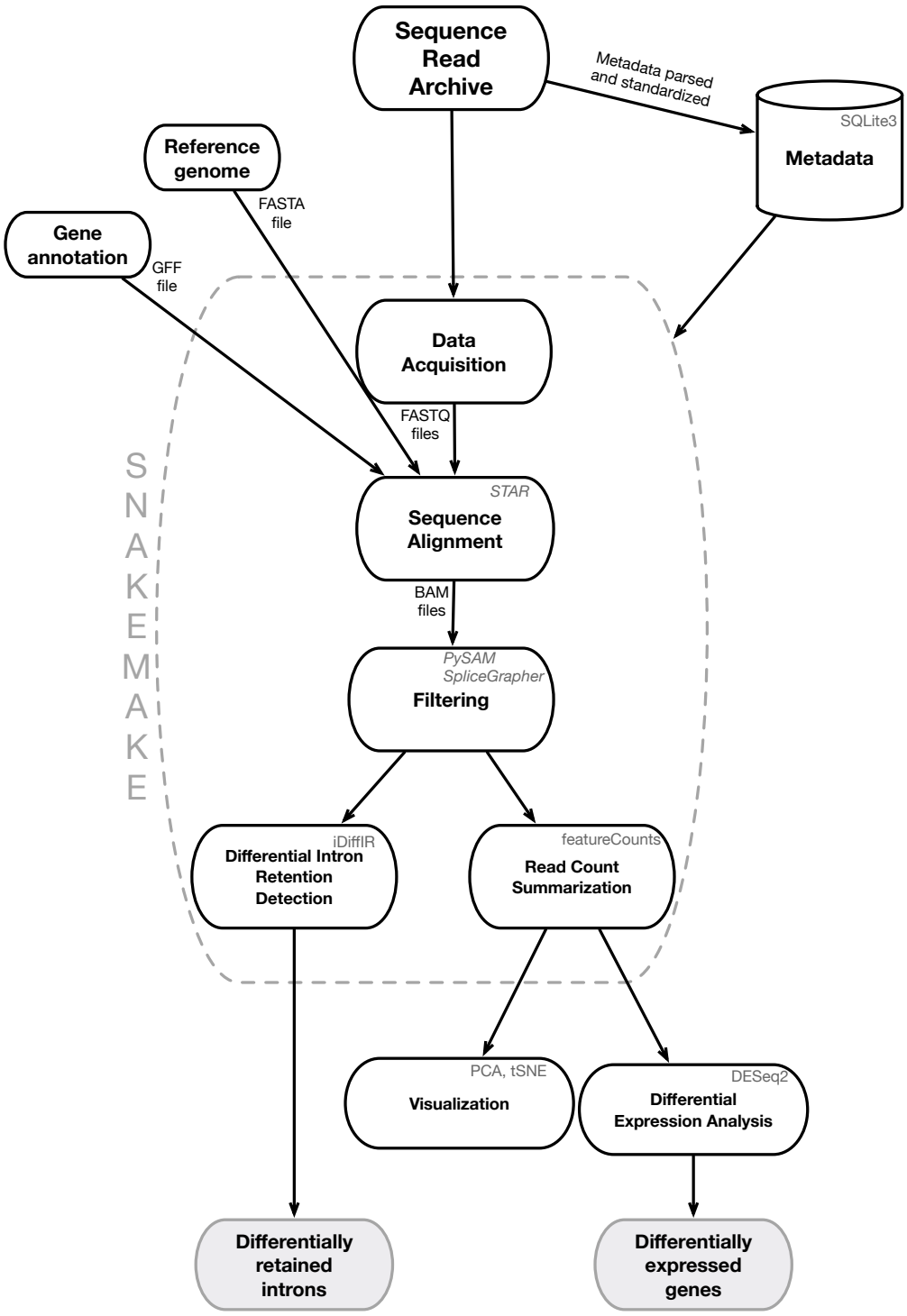

Figure 4.1: An outline of the major steps used to gather, process and analyze RNA-seq data. Chapter 4 shows technical details, software used, and decisions made for each step in the pipeline.

with the specific language used by the research team that uploaded the data. Sequencing data was then downloaded using the SRA toolkit, v2.9.6 [138].

Eighty-eight studies were downloaded, processed, and checked for quality and relevancy to the project. Out of the 88 studies, 35 studies met our quality standards. We chose our data using the following criteria:

- Organism: *Arabidopsis thaliana*, ecotype: Columbia (Col-0) with no genetic mutations.
- At least two biological replicates per condition.
- All biological replicates minimally required:
  - $\geq$ 1 Gbases;
  - read length $\geq$ 75 base pairs;
  - 75% or more of the reads needed to align specifically to one location in the reference genome, using the RNA-seq aligner, STAR.

Table 4.1 provides information about the datasets downloaded.

In our final set of studies, the most common tissue types were: leaf, seedling, and root; which composed 89% of the biological replicates, as shown in Figure 4.3 B. The most common treatment conditions were: abscisic acid (ABA), cold, drought, heat, and mineral deficiencies; which composed 80% of the biological replicates, as shown in Figure 4.3 A. Other information that describes experimental setup includes the type of platform used to process the samples, whether the data was sequenced as single-end or paired-end reads, and the read length.

The following next-generation platforms were used in our data: Illumina Genome Analyzer II, Illumina HiSeq 2000, Illumina HiSeq 2500, Illumina HiSeq 3000, Illumina HiSeq 4000, and NextSeq 500. Figure 4.3 C shows the majority of the experiments used Illumina HiSeq 2500 or Illumina HiSeq 2000 to sequence the samples. These platforms produce files that consist of millions of short sequencing reads that represent fragments from the original RNA-molecules [139]. Each

of these platforms are Illumina based. Illumina has dominated the sequencing industry and set the standard for massively parallel sequencing [140].

RNA-seq samples can either be sequenced as:

- single-end, where the sequencer reads a fragment from one end to the other, generating the sequence of bases.
- paired-end, where the fragment is sequenced from both ends.

Paired-end reads give slightly better alignments rates, however, if studying well-annotated organism, single-end reads are sufficient for gene expression analysis [141, 142]. Figure 4.3 D shows that the majority of studies chose to sequence data using a paired-end protocol.

Longer sequencing reads can provide more reliable information by increasing the chance that the read will be mapped uniquely to the genome. By filtering out studies with read lengths < 75 bp, we ended up with read lengths ranging between 75-300 bp, depending on the study. Once the samples have been processed, they are stored in a FASTQ file [143], and this file is what is uploaded to the SRA database [144].

4.2 Sequence Alignment

Once the sequence samples are downloaded from the SRA database, the technical replicates are combined, and the biological replicates are mapped to a reference genome using the Spliced Transcripts Alignment to a Reference (STAR) package, v2.7.0d. [174]. We used the TAIR10 reference genome and gene annotations from The Arabidopsis Information Resource (TAIR) database [175, 176]. STAR outputs Sequence Alignment Map (SAM) files, or the binary equivalent, called a BAM file [177].

4.3 Filtering

After STAR aligns the reads, the output is filtered for false-positive splice junctions using a custom script that utilizes features from SpliceGrapher [178], 0.2.6v, and PySAM [179], v0.15.2.

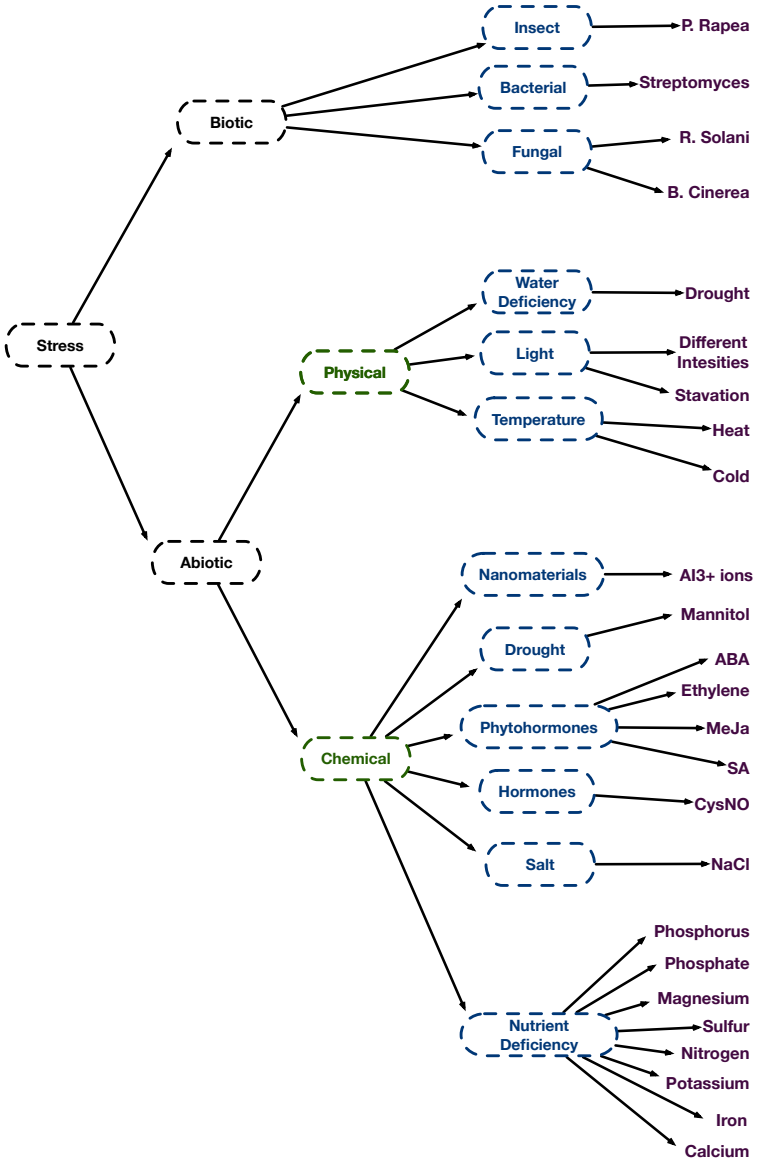

Figure 4.2: Categorization of treatment conditions. To obtain these categorizations, terminology was gathered from associated academic papers and each treatment was looked up in the Plant Experimental Conditions Ontology (PECO). A mixture of these two sources were combined to ensure that the treatment conditions accurately represented what the original authors were biologically investigating.

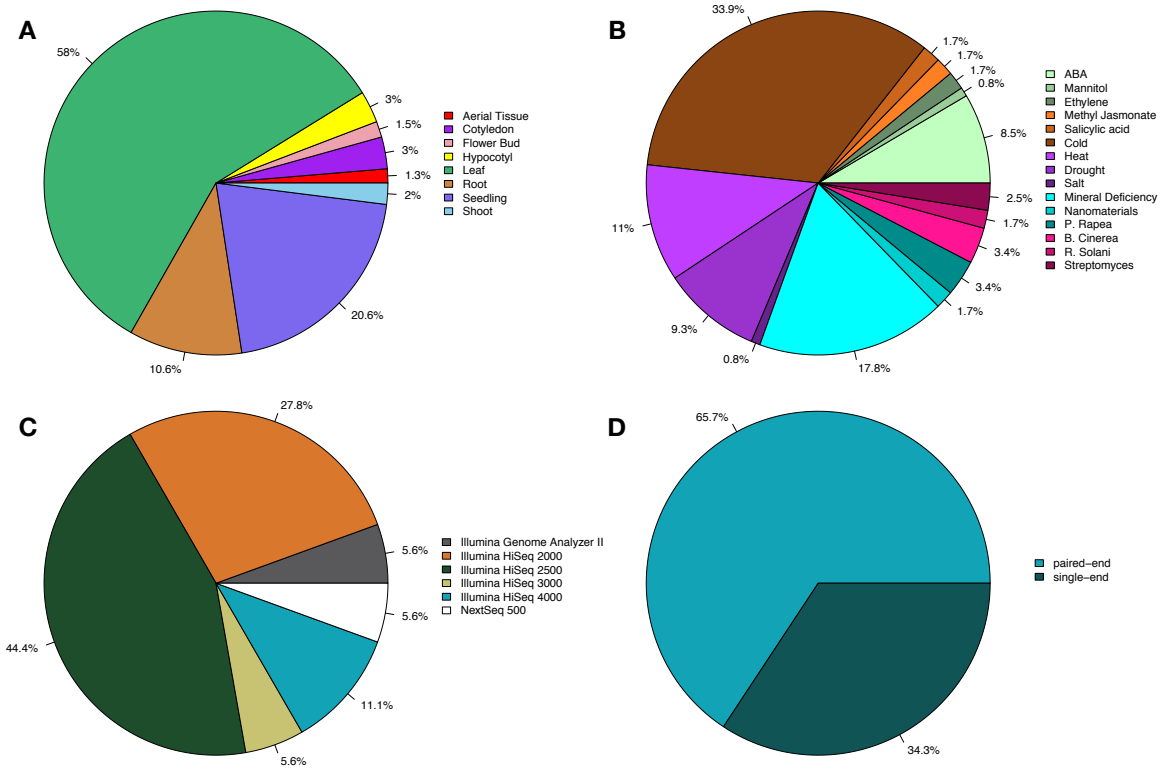

Figure 4.3: Graph A describes the distribution of tissue types within biological replicates. Graph B describes the distribution of treatments groups within biological replicates. Graph C describes the distribution of Illumina sequencing platforms within studies. Graph D describes the distribution of Sequencing protocols within studies.

Table 4.1: Experiments from the Sequence Read Archive that fit our quality standards and additional requirements. Column one provides the SRA accession and a citation to the corresponding academic paper. Column two provides the number of treatment conditions associated with the experiment. Column three provides the number of biological replicates associated with each treatment condition. Column four provides the tissue type of the biological replicates. Column five provides the age range when the biological replicates were sampled. Column six categorizes the type of treatment exposed to the plant.

SRA Accession	# Conditions	# Replicates	Tissue	Age (days)	Treatment
DRP003686 [145]	13	2	Root	28	Mineral Deficiency
DRP004486 [146]	3	3	Leaf	44	Mineral Deficiency
ERP022071 [14]	26	3	Leaf	35-39	Cold
ERP111840 [147]	3	3	Leaf	33	SA
SRP011128 [148]	2	3	Root	13	Mineral Deficiency
SRP026260 [149]	3	2	Leaf	unknown	Drought
SRP058527 [150]	3	2	Leaf	21	Heat
SRP060410 [15]	11	2	Seedling	6	Light
SRP073212 [44, 45]	56	2, 3	Leaf	36	Drought, B. Cinerea, P. Rapea
SRP073711 [151]	14	2	Seedling	3	ABA
SRP074485 [152]	18	2	Cotyledon, Hypocotyl	5	Light
SRP074890 [153]	2	3	Leaf	28	CysNO
SRP076862 [154]	4	2	Root, Shoot	3	Ethylene
SRP081056 [16]	4	3	Leaf	35	Heat
SRP082177 [155]	2	3	Leaf	35	Heat
SRP090416 [156]	4	3	Leaf	10	Carbon Starvation
SRP091014 [157]	2	2	Leaf, Root	13	Mineral Deficiency
SRP091628 [158]	3	2, 3	Root	7	Mineral Deficiency
SRP096554 [159]	6	2	Leaf	28	Drought, Light Deprivation
SRP097690 [160]	2	3	Flower Bud, Leaf	stages 1-12	Heat
SRP101403 [161]	3	3	Seedling	10	MeJA
SRP102893 [162]	3	3	Root	30	Al3+ ion, nAl2O3
SRP107981 [163]	2	2	Root	10	Heat
SRP108611 [164]	2	3	Seedling	7	ABA
SRP124769 [165]	2	3	Seedling	12	ABA
SRP134263 [166]	2	3	Seedling	7	Heat
SRP136536 [167]	2	3	Seedling	11	Drought
SRP145580 [168]	3	3	Seedling	9	ABA, Mannitol
SRP148881 [169]	2	4	Aerial	28	Drought
SRP155798 [170]	7	3	Leaf	14-28	Drought
SRP156748	3	3	Leaf	30	Heat
SRP161785	6	2, 3	Seedling	unknown	Streptomyces AGN23
SRP162472 [171]	4	4	Root, Shoot	14	R. solani
SRP177951 [172]	3	3	Seedling	14	NaCl
SRP187477 [173]	3	3	Seedling	10	Cold

A splice junction is filtered out if it is not in the gene annotation, or if it does not show up in the other biological replicates. Once filtered, the sequencing data is passed to the next steps in our analysis, which includes: calculating gene expression, differential gene expression, and differential intron retention [139].

4.4 Data Visualization

Identifying patterns present across multiple datasets remains challenging. The goal of dimensionality reduction, or embedding, is to build a 2D or 3D map which represents the similarities

in the higher dimensional dataset. Visualizing the data is valuable because it allows us to detect dominant patterns in the genetic expression. It is important to note that the variation in data from high-throughput sequencing technologies might not be due to biological differences. Batch effects are differences in the data due to how the data was prepared and sequenced, instead of biological reasons [180]. These include differences in laboratory conditions, equipment, and technicians [181].

Dimensionality reduction methods assume that the data has been normalized. We used DeSeq2 along with the variance-stabilizing transformation (VST) to normalize the data prior to analysis. The goal of VST is to factor out the dependence of the variance on the mean (over dispersion) [182].

4.5 Principal Component Analysis

Principal component analysis (PCA) is a mathematical procedure takes a set of possibly correlated variables and produces a set of directions, such that those directions are orthogonal (linearly independent) and ranked according to the variance along those directions. The number of principal components produced is less than or equal to the number of original variables. Because PCA creates a ranked list according to variance, it has been used as a linear feature extraction technique and was one of the original dimensionality reduction approaches [183]. Often a large fraction of the variance is captured within the first few components. These components can be plotted using a scatter plot, making it possible to observe similarities, differences and common groupings [184].

A DESeq2 package, plotPCA was used to generate the images in Figures 4.4 and 4.5. We will go into more detail about these images in Section 4.7.

4.6 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is one of the newest and most commonly used dimensionality reduction techniques [185]. t-SNE was developed on the premise that keeping similar data points close together is more important than preserving the distance of dissimilar data

points. The technique was introduced by Laurents van der Maaten and Geoffrey Hinton in 2008. t-SNE builds off Stochastic Neighbor Embedding (SNE) which was introduced by Geoffrey Hinton and Sam Roweis in 2002 [186].

Because this is a newer method, we will go into its mathematical details. The algorithm begins by calculating the similarity between two points in a higher dimensional space using Equation (4.1):

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{j' \neq i'} exp(-||x_{i'} - x_{j'}||^2/2\sigma_i^2)}. \tag{4.1}$$

$p_{j|i}$ is calculated by centering a Gaussian over x_i, then measuring the density of all other points under this Gaussian and re-normalizing according to these points. This gives us a set of probabilities $p_{j|i}$, which measure the similarity between a pair of points i and j. If $p_{j|i}$ is large, it implies that the points are close, or very similar, and if $p_{j|i}$ is small, it implies that the points are dissimilar. The bandwidth of the Gaussian, with a given variance σ_i^2, is set for each point such that there is a fixed number of other points taken into consideration. This variance is different for every point, allowing a fixed number of points to be considered within the Gaussian, even if the density of the dataset varies. This fixed number is termed the perplexity, and can be thought of as the number of nearest neighbors considered when calculating the probabilities of the high and low dimensional representations for each point.

The final set of similarities in the high dimensional space is calculated by taking the joint probabilities between a pair of points, as shown in Equation (4.2):

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{4.2}$$

This averages the probability that point j will choose point i and the probability that point i will pick point j. N is the number of input data points to the algorithm. Next, a corresponding set of points is laid out in low dimensional space at random, and then the joint probabilities are calculated in the lower dimensional space using Equation (4.3):

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + ||y_k - y_l||^2)^{-1}}, \tag{4.3}$$

Instead of using a Gaussian distribution, a Student-t distribution with 1 degree of freedom is used: $(1+||y_k-y_l||^2)^{-1}$. This distribution is used to increase optimization and to compensate for an issue called 'the crowding problem'. Due to the fact that the size or distance between points scales down exponentially as the dimensions are decreased, the points end up being compressed or 'crowded' together when the data is embedded into the lower dimension. The Student t-distribution is heavier tailed than the Gaussian distribution, which allows dissimilar points to be modelled farther apart.

Once these probabilities have been calculated, if the mapping points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j, the conditional probabilities between corresponding points in the higher and lower dimension will be the same. The discrepancy between p_{ij} and q_{ij} is measured using the Kullback-Leibler divergence, shown in Equation (4.4):

$$KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}. \tag{4.4}$$

P represents the joint probability distribution in the high dimensional space and Q represents the joint probability distribution in the low dimensional space. The KL cost function preserves local structure by assigning a large penalty to points that are close in higher dimensional space (p_{ij}) but modelled far away in the low dimension space (q_{ij}). If the p_{ij} is small (the points x_i and x_j are very dissimilar in high dimensional space), hence preserving the distance in the low dimensional space is not as important to the cost function.

t-SNE minimizes the KL divergence by using gradient descent. There is a performance limiting factor in using the KL cost function in the gradient descent algorithm. All pair-wise interactions within the dataset have to be considered in order to move a point to lower the KL divergence measure. This does not scale well for genomics research because there is a major slowdown for datasets that are larger than five to ten thousand points. The Barnes-Hut approximation combines

similar points into a single interaction and uses this interaction as an approximation for that set of similar points [187].

Rtsne, v0.15, was used to generate the images in Figures 4.4 and 4.5. Default parameters were used for learning rate and number of iterations. The perplexity was manually adjusted depending on the size of the data set. Table 4.2 compares major differences between PCA and tSNE. In the next section we will discuss these comparisons in regards to our data.

4.7 Comparison of PCA and t-SNE using Expression Data

Table 4.2: Major differences between PCA and t-SNE

	PCA	**t-SNE**
Type of algorithm	deterministic	stochastic
Projection onto a low dimensional space	linear	non-linear
Global or local approach	global	local and global
Unique solution	yes	no

We ran PCA and t-SNE on multiple subsets of the data. We started by looking at samples that contained the three most sampled tissues: leaf, seedling and root. Out of all of our samples, 89% were classified as one of these three tissues. A breakdown of the tissue types and the treatment conditions can be seen in Figure 4.3.

We found that both t-SNE and PCA primarily grouped the data by tissue type. Which supported conclusions previously published in, "Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis" [46]. The researchers in this study collected and re-annotated 6,057 microarray expression profiles of *Arabidopsis* from GEO and used PCA to show that samples of the same tissue tended to cluster together, even if the plants were exposed to different treatment conditions. We similarly demonstrate this in Figure 4.4, plots A and B.

PCA was not originally created for the purposes of dimensionality reduction and visualization. Instead, it is focused on finding orthogonal projections that contain the highest variance. Because of this property, it has a hard time modelling data that is not linearly correlated, e.g. a spiral. t-SNE

is better able to model shapes because it emphasizes preserving local structure. Interestingly, in Figure 4.4 plot C and D we use the same data points that are in plots A and B, but color these points to show treatment conditions rather than tissue types. t-SNE was able to further cluster some of the samples by condition, while PCA did not separate the conditions out as well.

Since tissue type predominated, we further analyzed the separate tissues in Figure 4.5. Plots A and B show clustering with leaf samples. While PCA does show trends, the following conditions: cold, drought, *P. Rapea* (a small butterfly), and *B. Cinerea* (a fungus) all cluster around the same location. t-SNE does a much better job showing distinct clusters. On reason for this is because t-SNE uses a probability distribution; it naturally expands dense clusters and contracts sparse ones, to even out the cluster densities. This allows t-SNE to show each cluster more distinctly, but in turn, the density of the clusters can not be used as part of the interpretation.

Plots C and D cluster seedling samples. It is interesting that both plots seem to separate out light (the seedlings were exposed to red, blue, and white lights) and ABA samples, while the other treatment conditions do not show distinct differences. It is also interesting to note that the location of the clusters are not similar between PCA and t-SNE. PCA is deterministic: every time you run the algorithm on the same dataset, you get the same plot. t-SNE is stochastic and cares more about preserving the probability distribution between points than where the points are projected onto the lower dimension. As a result, it is good practice to run the t-SNE multiple times and with multiple parameters to understand the relationships between points in the dataset.

A final observation corresponds to plots E and F which cluster root samples. t-SNE seems to work better with larger sample sizes. When the sample sizes were too small, like the number of samples for roots (45), there were not enough samples to form interesting relationships between the data points.

After observing patterns within the data, we analyzed each treatment condition for differentially expressed genes and differentially expressed introns. We will briefly describe the packages and process used to generate this data, and observe differences in expression over the different treatment conditions.

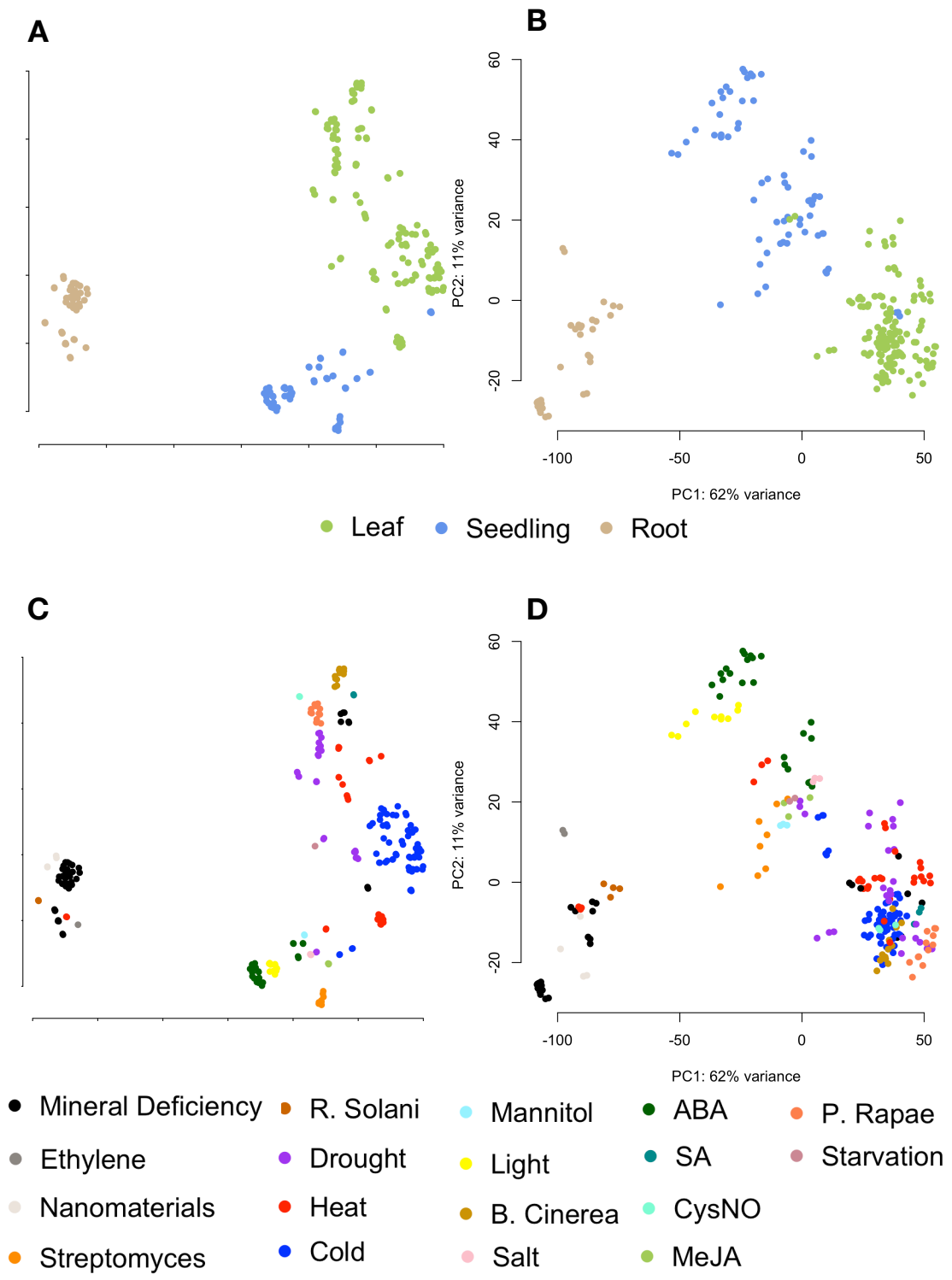

Figure 4.4: Figures A and C represent t-SNE plots created using an R package, Rtsne, with a perplexity of 13. Figures B and D are PCA plots created using a DeSeq2 package called plotPCA. Figures A and B correspond to tissue types for each biological sample. Figures C and D correspond to condition types for each biological sample.

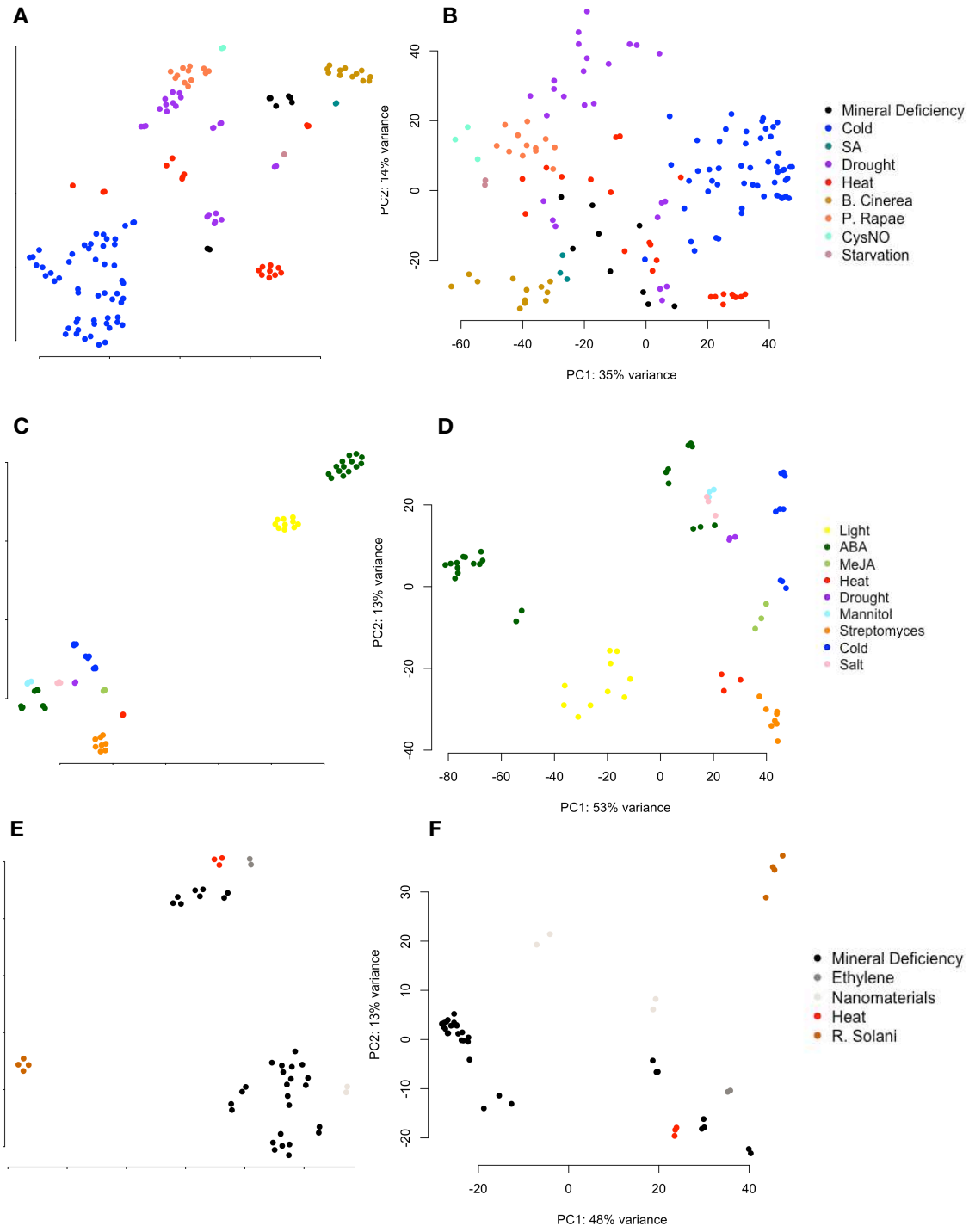

Figure 4.5: Figures A, C, and E represent t-SNE plots created using an R package, Rtsne, with the corresponding perplexity parameters: 10, 4, and 13. Figures B, D, and F are PCA plots created using a DeSeq2 package called plotPCA. Figure A and B are leaf samples, Figure C and D are seeding samples, and Figures E and F are root samples.

4.8 Differential Gene Expression

Read counts per gene were quantified using featureCounts, v1.6.4 [188], using the TAIR10 reference genome and gene annotations. Multi-mapping reads and reads with ambiguous assignments were excluded.

We used the R package DESeq2 [189, 190], v1.24.0, to identify deferentially expressed genes. As shown in Table 4.1, the majority of publicly available genomic data for plants contains a low number of biological replicates, $(n \leq 3)$. DESeq, edgeR [191], and limma [192] are packages that have methods that compensate for lower numbers of biological replicates. While normalization methods focus on normalizing library size, DESeq and edgeR also adjust for differences in library composition. Adjusting for library composition is modelled by assuming that genes with huge differences will generally be rare and should be given less influence. These packages incorporate information about a gene's expression across each biological replicate to shrink the variance, giving more influence to moderate differences (or "house-keeping" genes) [191, 193–195].

For each experiment, we identified control and treatment conditions used within the publications. We ran differential gene expression analysis and obtained a list of genes with an absolute log fold change ≥ 2 and a false discovery rate (FDR) ≤ 0.01.

4.9 Differential Intron Retention

iDiffIR [196] is a package that specializes in identifying differential intron retention (DIR) events using RNA-Seq data. iDiffIR searches for DIR events by applying a log fold change statistic, shown in Equation (4.6), to every intron identified within a particular gene annotation (GTF or GFF) file. The mean read depth across an intron is quantified using Equation (4.5):

$$\mu_r(I) = \frac{1}{|I|} \sum_{i \in I} r(i). \tag{4.5}$$

Where $r(i)$ is the number of reads at a particular genomic index, i.

To account for the fact that introns have low expression levels, the read coverage is normalized separately for the introns and the exons. The read coverage is also normalized across experimental conditions; this helps avoid the false detection of differentially retained introns as a result of differential gene expression. Differential expression between introns in two conditions is quantified by the following log-fold change statistic:

$$log_{\hat{FC}}(I) = log_2(\frac{a + \hat{\mu}_r(I_1)}{a + \hat{\mu}_r(I_2)}). \tag{4.6}$$

Where $\hat{\mu}_r(I_1)$ and $\hat{\mu}_r(I_2)$ denote the adjusted mean read depth from the introns in conditions 1 and 2. a is a pseudo-count parameter that controls for large fold-change values that occur in regions of low expression and is necessary to avoid filtering genomic regions that are not expressed in one of the conditions. Genes that exhibit DIR events were detected and quantified for each condition in our studies, as shown in Figure 4.6 B. We further combined genes that exhibited DIR across different studies based on treatment conditions, as shown in Figure 4.7 B

4.10 Differential Gene Expression and Differential Intron Retention

The heat, drought, and carbon starvation stresses triggered significant up-regulation and down-regulation of differential gene expression. The chart shows that several of these studies show DEG counts that were thousands of genes higher than found in other studies. This was reflected in both Figure 4.6 A and Figure 4.7 A. Salt, nitric oxide (via CysNO), P. Rapea (a butterfly), and B. Cinerea (a fungus) stresses triggered a overwhelming up-regulation, in comparison to the number of down-regulated genes, in differential expression.

Mineral deficiencies included: phosphate, iron, calcium, molybdenum, nitrogen, phosphorus, sulfur, zinc, copper, manganese, potassium, magnesium, and boron. With exception of nitrogen, potassium and phosphate, these deficiencies did not trigger a large change in differential gene expression. It was surprising to see that while there was almost no differential gene expression,

there were increased levels of differential intron retention. These differences can be seen in Figure 4.6, A and B.

The cold treatment study was interesting for many reasons. The study looks at differential gene expression and genes that exhibited differential alternative splicing in response to a temperature change from $20°$C to $4°$C. Samples were taken on the first and third day, every three hours. The differentially expressed genes oscillate over time reflecting changes in the circadian rhythm. The authors called this process "gating" [44], where the magnitude of changes in gene expression depend on the time of day that the sample was taken, as seen in Figure 4.8. It would be interesting to collect other experiments that kept track of the time of day that the plant was sampled, in order to see how the circadian rhythm affects other conditions.

The cold treatment study exhibited the largest levels of differential intron retention. Each individual experiment averaged around 350 genes with differential intron expression events, totaling to around 1,200 unique genes. This can be seen in 4.6 B and Figure 4.7 B. However, the dramatic increase in genes that exhibited differential intron retention between the cold condition and all other conditions in Figure 4.7 might be due more to how much data this study produced, rather than the treatment condition.

One issue with this compendium is that while some treatment conditions are examined by numerous studies, producing a wealth of data, others do not have much publicly accessible data. Because of this difference, our compendium will be a great place to investigate current data and generate scientific queries.

4.11 Identification of Genes that Undergo Differential Expression Under Multiple Stresses

A set of genes that were differentially expressed under multiple stress treatments, MST genes, were identified. We calculated the score of each gene by summing the number of unique biotic or abiotic stresses that caused the gene to be differentially expressed. This was achieved by grouping

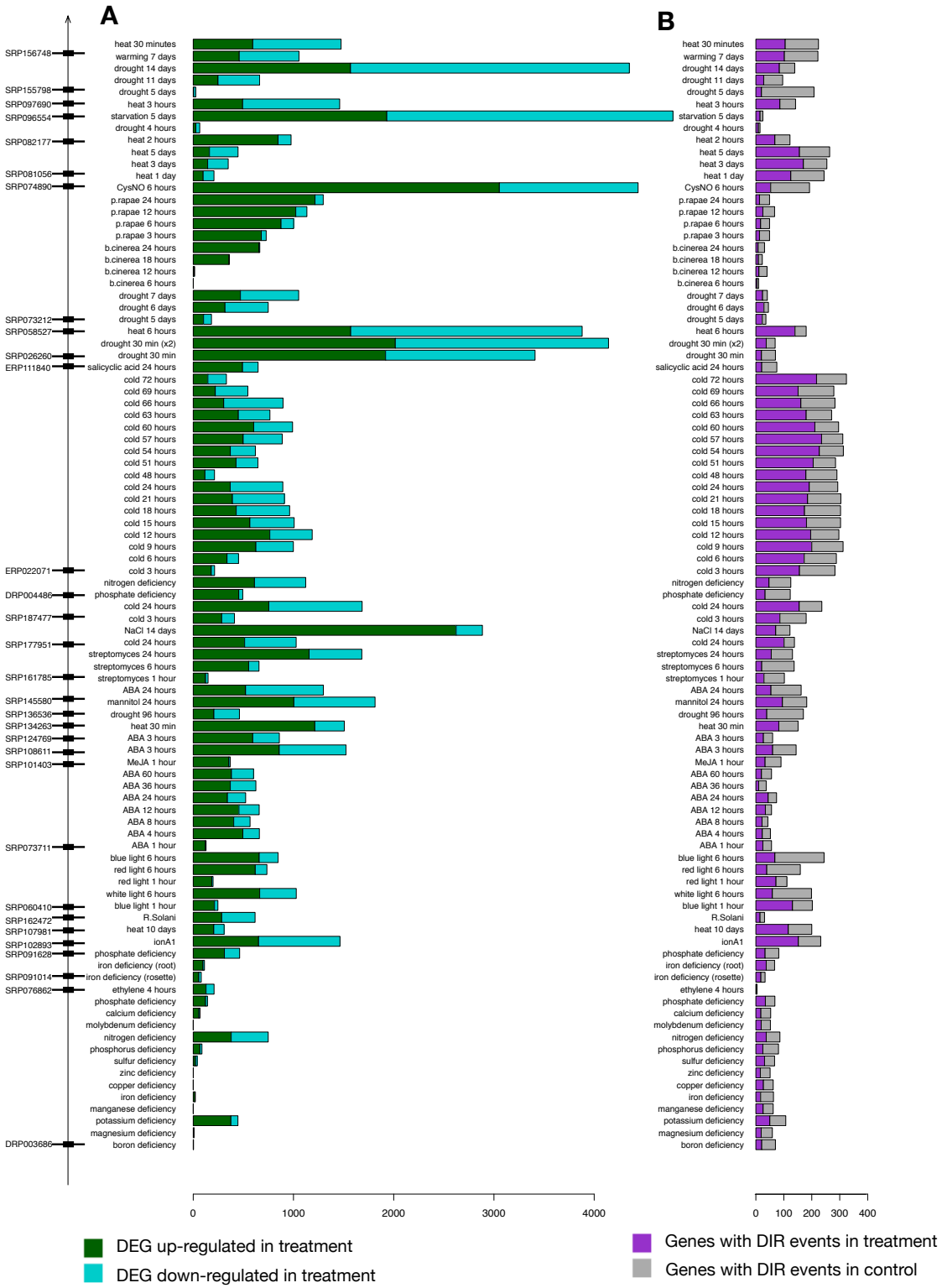

Figure 4.6: Barplot A illustrates the extent that genes are differentially up-regulated or down-regulated per treatment. Barplot B illustrates genes that contain an increase in differential intron retention events in the treatment condition vs an increase in differential intron retention events in the condition. Sequence Read Archive accessions are shown to the left to give frame of reference to how many treatment conditions are analyzed per condition.

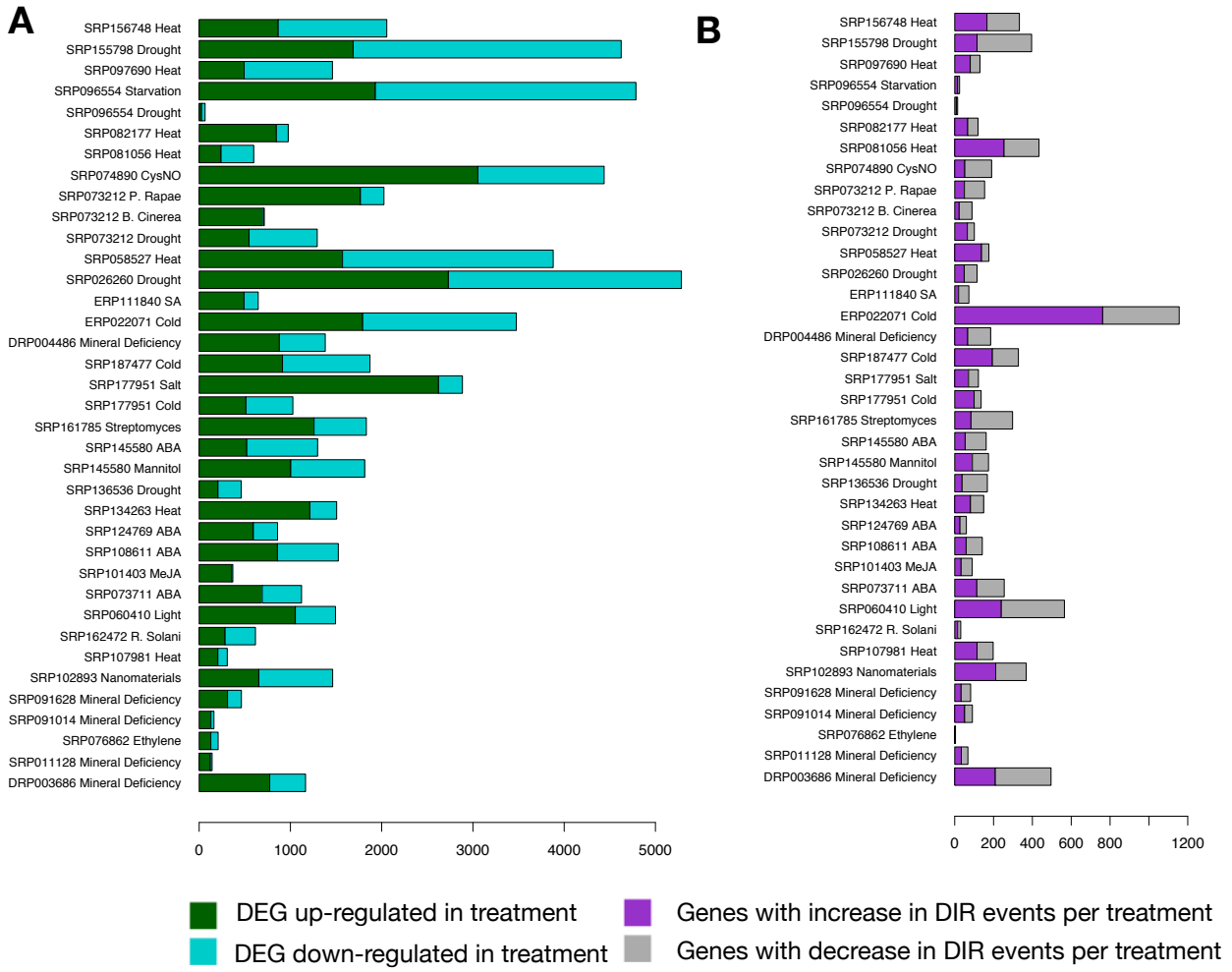

Figure 4.7: Barplot A quantifies up-regulated and down-regulated differentially expressed genes for each study using DESeq2 with an absolute log fold change ≥ 2 expression cut off and a FDR ≤ 0.01. Barplot B quantifies genes that contain up-regulated and down-regulated differentially retained introns using iDiffIR.

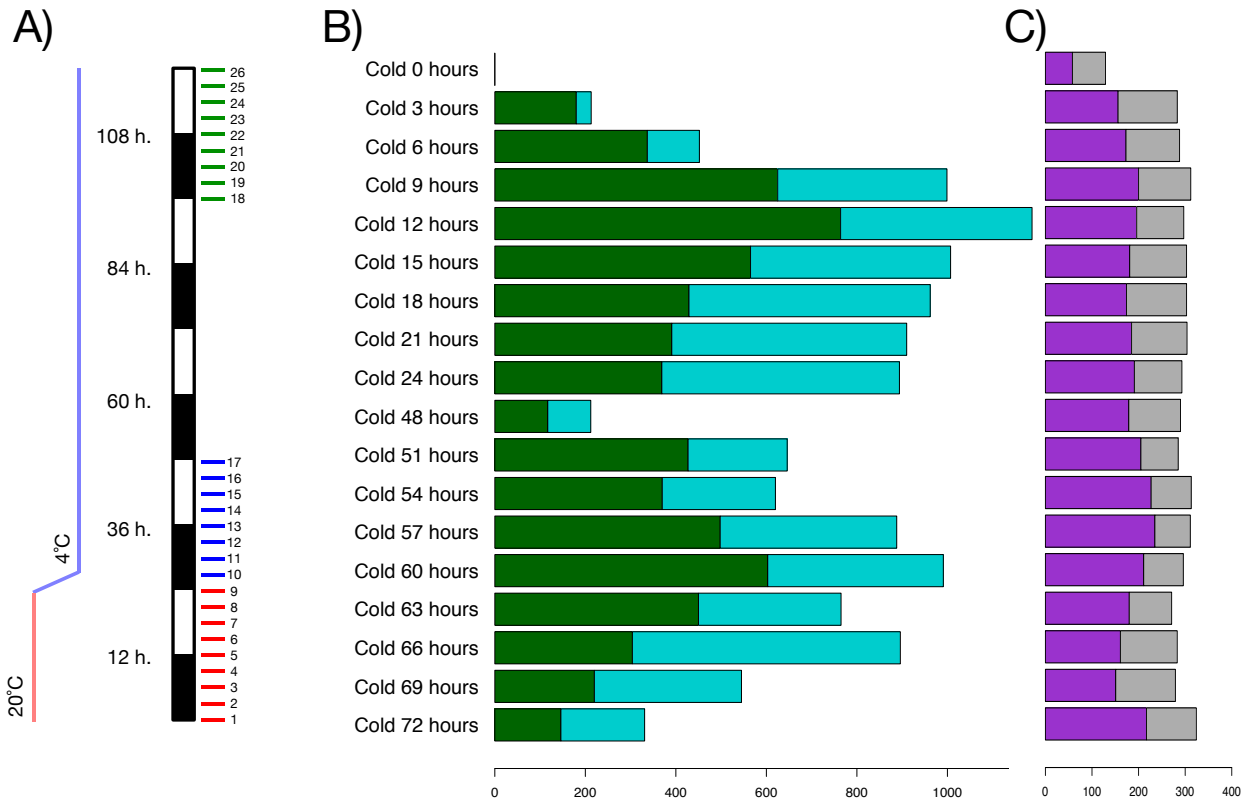

Figure 4.8: A) is an image adapted from Coolen et al. [44] that shows the sampling times and strategy for the experiment. B) Shows the differential gene expression for control vs. cold. C) shows genes with differential intron retention events for control vs. cold.

similar studies and creating a comprehensive list of genes that were differential expressed for each of these groups. The groups were defined as shown in Table 4.3:

Table 4.3: Groups defined for generating a comprehensive a list of differentially expressed genes. Column one shows the label used to generate the groups and plots. Column two gives a brief description of the conditions. Column three quantifies how many studies were in each group.

Group	Description	Number of Studies
ABA	abscisic acid	4
B. Cinerea	necrotrophic fungus	1
Cold		3
CysNO	nitric oxide donor	1
Drought		5
Ethylene	plant hormone	1
Heat		7
Light	exposure to red, blue and white light	1
Mannitol		1
Methyl jamonate (MeJA)	a plant growth regulator	1
Mineral Deficiency		5
Nanomaterials		1
P. Rapae	butterfly	1
R. Solani	pathogenic fungus	1
Salicylic acid (SA)		1
Salt		1
Starvation	carbon deprivation	1
Streptomyces	gram positive bacteria	1

As with data used to generate the PCA and t-SNE visualizations, we focused on the three primary tissue types: roots, leaves, and seedlings. We plotted this data on a set of histograms and found it interesting that over 1,000 genes were differentially expressed in 15 or more unique experiments, as shown in Figure 4.9 A. There were far fewer genes that had differentially retained introns, as shown in Figure 4.9 A.

We wanted to compare the results of our compendium with results acquired in a previous publication, “Functional-genomics-based identification of genes that regulate Arabidopsis responses to multiple abiotic stresses” [197]. In this study, the authors collected data from 17 micro-array experiments, including a heat experiment that they conducted for this research. The treatment con-

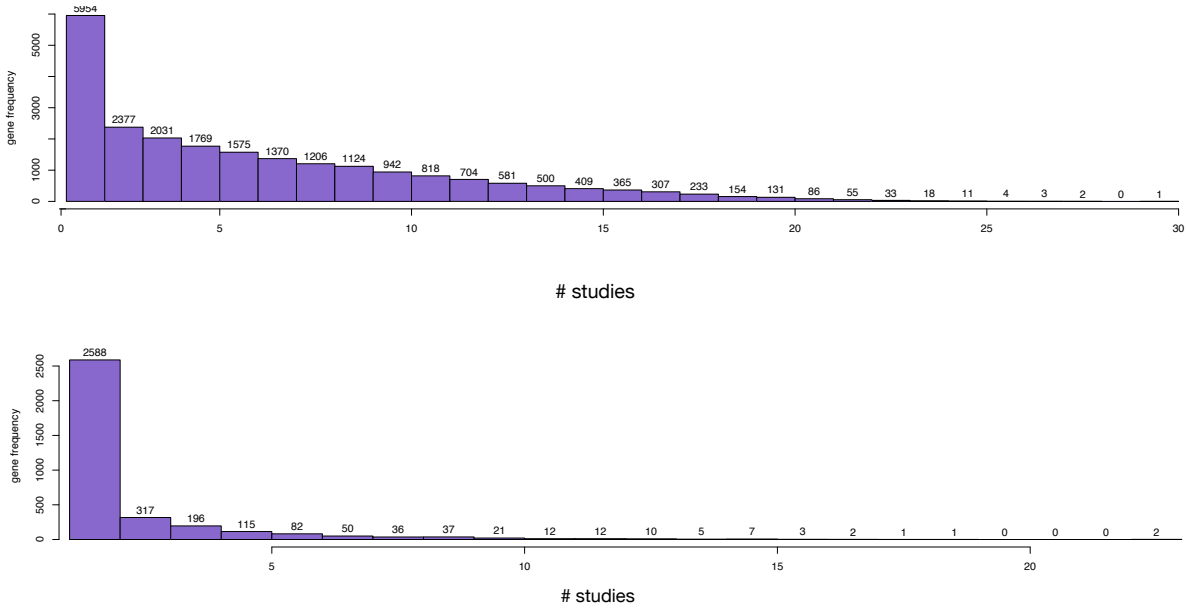

Figure 4.9: These histograms represent a combination of the three tissue types we focused on: leaf, root, and seedling. This included 32 studies from the Sequence Read Archive. Histogram A shows how many genes were differentially expressed exclusively in that number of studies. Histogram B shows the how many genes had exhibited differential intron retention in that number of studies.

ditions that they covered were: heat, drought, cold, salt, osmotic stress (similar to drought), high light, oxidative stress, and abscisic acid.

Surprisingly, there was only a 11% overlap between the list of genes from the previously published results, as illustrated in Figure 4.10 A. To explore this further, we wondered if it was because we were testing different sets of treatment conditions. We filtered our conditions in the following ways:

- We only used conditions that matched what the authors had specified in their paper
- We filtered out experiments that last more than 10 hours. The paper noted that stress responses were generally triggered upstream within the stress signal transduction pathways but some responses did peak at around 10 hours [197].

We then calculated the overlap between our new set of differentially expressed genes and the published results. Interestingly, we doubled the amount of genes that overlapped between the two groups, but that was still only a 16% overlap.

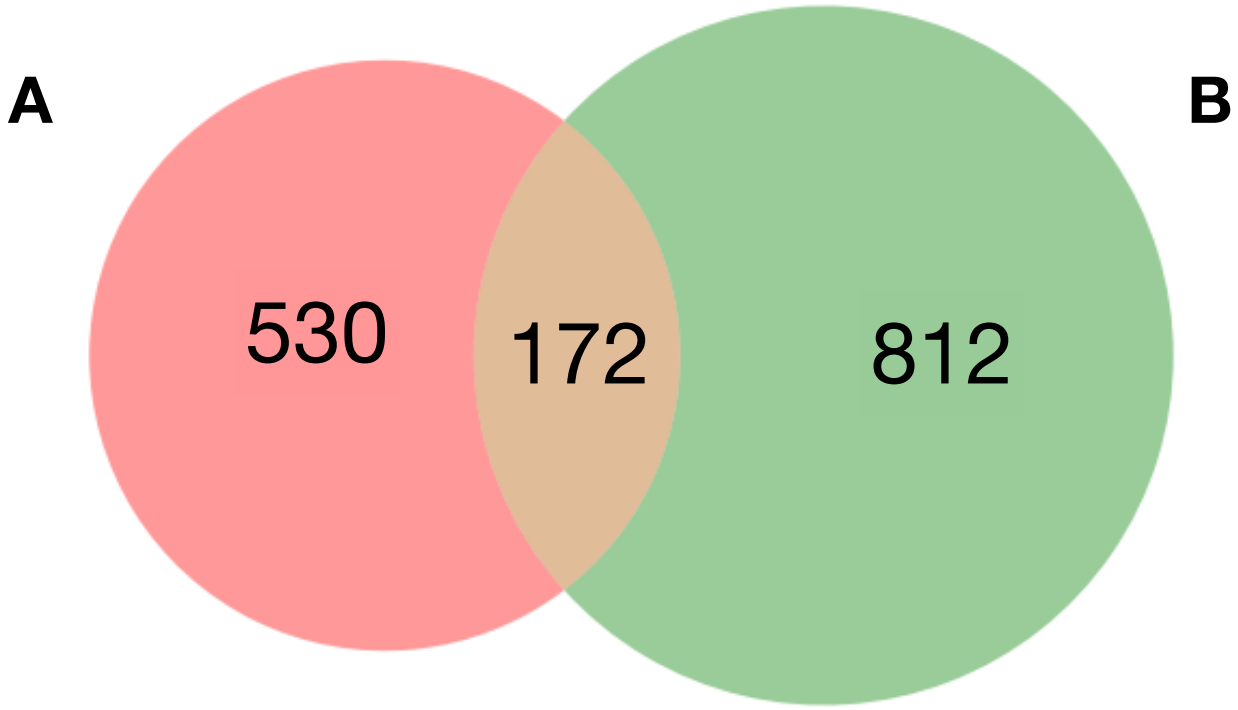

Figure 4.10: Group A represents the genes that we identified as MST genes. Group B represents the genes that were identified by Kant et al. [197]

Chapter 5

Conclusion and Future Work

There are multiple compendiums for mammal RNA-sequencing [4–8, 48, 49, 66, 67], there are not similar resources for plant data. These resources help advance science by providing large collections of data that can be combined to provide global insights into the 'omics fields. In our study, each sample was carefully curated, the metadata were comprehensively re-annotated, and the sequencing data was all processed using the same pipeline. This is a valuable resource for the plant community because there are currently no compendiums that focus on plant RNA-sequencing data. We also created a Snakemake sequencing pipeline to automate this process and will make it available for future use. This way our results can be reproduced and other research groups can use it to process their own data. There are some aspects of our pipeline that might be important to consider:

- Adding new studies to the compendium. Especially studies that investigate treatment conditions in which we have limited samples. The more studies we have, testing the same condition, the greater we can enhance data standardization techniques within our pipeline.

- Integrate other model plant organisms, like *Oryza sativa* and *sorghum bicolor*

- Investigate ways to automate the standardization and correctness of metadata. This is a bottleneck within the research process because of how long it takes to identify corresponding academic papers and which data is relevant.

- Updating software within the pipeline that has added new features, like Snakemake. As we mentioned in Chapter 3, software that is actively maintained has a decreased chance of undergoing software collapse. Since the induction of this project, Snakemake has added several desirable features that speed up processing and make it easier to organize the project. Actively integrating these features into the software will help prevent keep the pipeline relevant.

- Creating a web interface so that the data sets can be provided to the scientific community.

We provided a limited analysis to demonstrate the range of tissue types and treatment conditions we have acquired. We used t-SNE and PCA to cluster studies that used different types of treatment conditions and showed that the studies clustered by tissue type and secondarily treatment condition. An interesting set of observations could be made by clustering genes based on differential expression or differential intron retention patterns. This may provide some insight into genes that exhibit differential expression and intron retention under multiple stress factors, as well.

We gathered datasets for each study of differentially retained introns and differential gene expression. Other information we could gather in the future are other types of alternative splicing events and novel splice junctions. These would provide a more comprehensive view of how the expression is changing as the plants undergo each stress treatment. Lastly, we looked into genes that exhibit differential gene expression in multiple stresses, and compiled a list of these genes for future analysis. Each of these areas could be expanded in order to give insight into how plants adapt and evolve.

Bibliography

[1] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[2] Anna C Greene, Kristine A Giffin, Casey S Greene, and Jason H Moore. Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics*, 17(1):43–50, 2015.

[3] Serghei Mangul, Thiago Mosqueiro, Richard J Abdill, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Jared Littman, Benjamin Statz, et al. Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS biology*, 17(6):e3000333, 2019.

[4] Anton Zoubarev, Kelsey M Hamer, Kiran D Keshav, E Luke McCarthy, Joseph Roy C Santos, Thea Van Rossum, Cameron McDonald, Adam Hall, Xiang Wan, Raymond Lim, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, 28(17):2272–2273, 2012.

[5] Robert Petryszak, Maria Keays, Y Amy Tang, Nuno A Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic acids research*, 44(D1):D746–D752, 2015.

[6] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible rna-seq analysis using recount2. *Nature biotechnology*, 35(4):319, 2017.

[7] Irene Papatheodorou, Nuno A Fonseca, Maria Keays, Y Amy Tang, Elisabet Barrera, Wojciech Bazant, Melissa Burke, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Nancy

George, et al. Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic acids research*, 46(D1):D246–D251, 2017.

[8] Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, and Avi Ma'ayan. Massive mining of publicly available RNA-seq data from human and mouse. *Nature communications*, 9(1):1366, 2018.

[9] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[10] R.N. Shukla. *Analysis Of Chromosome*. Agrotech Press, 2014.

[11] Laurence A Moran. Basic Concepts: The Central Dogma of Molecular Biology, Jan 2007.

[12] Anireddy SN Reddy, Yamile Marquez, Maria Kalyna, and Andrea Barta. Complexity of the alternative splicing landscape in plants. *The Plant Cell*, 25(10):3657–3683, 2013.

[13] Mohammed Albaqami and Anireddy SN Reddy. Development of an in vitro pre-mRNA splicing assay using plant nuclear extract. *Plant methods*, 14(1):1, 2018.

[14] Cristiane PG Calixto, Wenbin Guo, Allan B James, Nikoleta A Tzioutziou, Juan Carlos Entizne, Paige E Panter, Heather Knight, Hugh G Nimmo, Runxuan Zhang, and John WS Brown. Rapid and dynamic alternative splicing impacts the Arabidopsis cold response transcriptome. *The Plant Cell*, 30(7):1424–1444, 2018.

[15] Lisa Hartmann, Philipp Drewe-Boß, Theresa Wießner, Gabriele Wagner, Sascha Geue, Hsin-Chieh Lee, Dominik M Obermüller, André Kahles, Jonas Behr, Fabian H Sinz, et al. Alternative splicing substantially diversifies the transcriptome during early photomorphogenesis and correlates with the energy availability in arabidopsis. *The Plant Cell*, 28(11):2715–2734, 2016.

[16] A Pajoro, E Severing, GC Angenent, and RGH Immink. Histone h3 lysine 36 methylation affects temperature-induced alternative splicing and flowering in plants. *Genome biology*, 18(1):102, 2017.

[17] Sergei A Filichkin, Michael Hamilton, Palitha D Dharmawardhana, Sunil K Singh, Christopher Sullivan, Asa Ben-Hur, Anireddy SN Reddy, and Pankaj Jaiswal. Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching. *Frontiers in plant science*, 9:5, 2018.

[18] Motoaki Seki, Taishi Umezawa, Kaoru Urano, and Kazuo Shinozaki. Regulatory metabolic networks in drought stress responses. *Current opinion in plant biology*, 10(3):296–302, 2007.

[19] Kang Yan, Peng Liu, Chang-Ai Wu, Guo-Dong Yang, Rui Xu, Qian-Huan Guo, Jin-Guang Huang, and Cheng-Chao Zheng. Stress-induced alternative splicing provides a mechanism for the regulation of microrna processing in arabidopsis thaliana. *Molecular cell*, 48(4):521–531, 2012.

[20] Jenna E Gallegos. Alternative splicing plays a major role in plant response to cold temperatures. *The Plant Cell*, pages tpc–00430, 2018.

[21] Yu Ling, Sahar Alshareef, Haroon Butt, Jorge Lozano-Juste, Lixin Li, Aya A Galal, Ahmed Moustafa, Afaque A Momin, Manal Tashkandi, Dale N Richardson, et al. Pre-mrna splicing repression triggers abiotic stress signaling in plants. *The Plant Journal*, 89(2):291–309, 2017.

[22] Wenfeng Li, Wen-Dar Lin, Prasun Ray, Ping Lan, and Wolfgang Schmidt. Genome-wide detection of condition-sensitive alternative splicing in arabidopsis roots. *Plant physiology*, 162(3):1750–1763, 2013.

[23] Anireddy SN Reddy. Alternative splicing of pre-messenger rnas in plants in the genomic era. *Annu. Rev. Plant Biol.*, 58:267–294, 2007.

[24] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome biology*, 14(9):3158, 2013.

[25] Ei-Wen Yang, Thomas Girke, and Tao Jiang. Differential gene expression analysis using coexpression and rna-seq data. *Bioinformatics*, 29(17):2153–2161, 2013.

[26] Andreas PM Weber. Discovering new biology through sequencing of RNA. *Plant physiology*, 169(3):1524–1531, 2015.

[27] Kris A Wetterstrand. DNA sequencing costs: data from the NHGRI genome sequencing program (GSP), 2013.

[28] Jerzy K Kulski. Next-generation sequencing — an overview of the history, tools, and "omic" applications. In *Next Generation Sequencing-Advances, Applications and Challenges*. IntechOpen, 2016.

[29] Evanthia Kaimaklioti Samota and Robert P Davey. Knowledge and attitudes among life scientists towards reproducibility within journal articles. *BioRxiv*, page 581033, 2019.

[30] Franklin Sayre and Amy Riegelman. The reproducibility crisis and academic libraries. *College & Research Libraries*, 79(1):2, 2018.

[31] Luca Alessandrì, Neha Kulkarni, Riccardo Panero, Martina Olivero, Maddalena Arigoni, Marco Beccuti, Francesca Cordero, and Raffaele A Calogero. RBP, a community for reproducible bioinformatics, 2018.

[32] Daniele Fanelli. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11):2628–2631, 2018.

[33] Yang-Min Kim, Jean-Baptiste Poline, and Guillaume Dumas. Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, 7(7):giy077, 2018.

[34] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.

[35] Leonard P Freedman, Iain M Cockburn, and Timothy S Simcoe. The economics of reproducibility in preclinical research. *PLoS biology*, 13(6):e1002165, 2015.

[36] Data Sharing and Release Guidelines. https://www.niaid.nih.gov/research/data-sharing-and-release-guidelines, Jun 2019.

[37] National Institutes of Health. NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources. https://grants.nih.gov/policy/sharing.htm, Feb 2018.

[38] YoSon Park and Casey S Greene. A parasite's perspective on data sharing. *GigaScience*, 7(11):giy129, 2018.

[39] Gordon and Betty Moore Foundation. Data Sharing Philosphy. https://www.moore.org/docs/default-source/Grantee-Resources/data-sharing-philosophy.pdf, Sep 2008.

[40] Olena Morozova Vaske and David Haussler. Data sharing for pediatric cancers. *Science*, 363, March 2019.

[41] Bill & Melinda Gates Foundation. Bill & Melinda Gates Foundation Open Access Policy. https://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy, 2019.

[42] Center for Open Science. TOP Guidelines. https://cos.io/top/, July 2019.

[43] BA Nosek, G Alter, GC Banks, D Borsboom, SD Bowman, SJ Breckler, S Buck, CD Chambers, G Chin, G Christensen, et al. Promoting an open research culture: The TOP guidelines for journals. *Science*, 348(6242):1422–1425, 2015.

[44] Silvia Coolen, Silvia Proietti, Richard Hickman, Nelson H Davila Olivas, Ping-Ping Huang, Marcel C Van Verk, Johan A Van Pelt, Alexander HJ Wittenberg, Martin De Vos, Marcel

Prins, et al. Transcriptome dynamics of Arabidopsis during sequential biotic and abiotic stresses. *The Plant Journal*, 86(3):249–267, 2016.

[45] Silvia Coolen, Johan A Van Pelt, Saskia CM Van Wees, and Corné MJ Pieterse. Mining the natural genetic variation in Arabidopsis thaliana for adaptation to sequential abiotic and biotic stresses. *Planta*, 249(4):1087–1105, 2019.

[46] Fei He, Shinjae Yoo, Daifeng Wang, Sunita Kumari, Mark Gerstein, Doreen Ware, and Sergei Maslov. Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *The Plant Journal*, 86(6):472–480, 2016.

[47] Priyanka Bhandary, Arun S Seetharam, Zebulun W Arendsee, Manhoi Hur, and Eve Syrkin Wurtele. Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. *Plant science*, 267:32–47, 2018.

[48] Zichen Wang, Alexander Lachmann, and Avi Ma'ayan. Mining data and metadata from the gene expression omnibus. *Biophysical reviews*, 11(1):103–110, 2019.

[49] Ashkaun Razmara, Shannon E Ellis, Dustin J Sokolowski, Sean Davis, Michael D Wilson, Jeffrey T Leek, Andrew E Jaffe, and Leonardo Collado-Torres. recount-brain: a curated repository of human brain rna-seq datasets metadata. *BioRxiv*, page 618025, 2019.

[50] Katrina Learned, Ann Durbin, Robert Currie, Ellen Towle Kephart, Holly C Beale, Lauren M Sanders, Jacob Pfeil, Theodore C Goldstein, Sofie R Salama, David Haussler, et al. Barriers to accessing public cancer genomic data. *Scientific Data*, 6(1):98, 2019.

[51] Tasnia Tahsin, Davy Weissenbacher, Karen O'Connor, Arjun Magge, Matthew Scotch, and Graciela Gonzalez-Hernandez. GeoBoost: accelerating research involving the geospatial metadata of virus Genbank records. *Bioinformatics*, 34(9):1606–1608, 2017.

[52] Oleksandr Lykhenko, Alina Frolova, and Maria Obolenska. Designing the database for microarray experiments metadata. In *2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF)*, pages 127–131. IEEE, 2017.

[53] Morgan GI Langille, Jacques Ravel, and W Florian Fricke. ”Available upon request“: not good enough for microbiome data!, 2018.

[54] Matthew N Bernstein, AnHai Doan, and Colin N Dewey. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, 33(18):2914–2923, 2017.

[55] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[56] John Wise, Alexandra Grebe de Barron, Andrea Splendiani, Beeta Balali-Mood, Drashtti Vasant, Eric Little, Gaspare Mellino, Ian Harrow, Ian Smith, Jan Taubert, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug discovery today*, 24(4):933–938, 2019.

[57] Digital Science, Mark Hahnel, Briony Fane, Jon Treadway, Grace Baynes, Ross Wilkinson, Barend Mons, Erik Schultes, Luiz Olavo Bonino da Silva Santos, Pavel Arefiev, and et al. The State of Open Data Report 2018, Oct 2018.

[58] John Brock. “A love letter to your future self”: What scientists need to know about FAIR data, Feb 2019.

[59] Philip E Bourne and Johanna McEntyre. Biocurators: contributors to the world of science, 2006.

[60] Amrapali Zaveri and Michel Dumontier. Metacrowd: Crowdsourcing biomedical metadata quality assessment. 2017.

[61] Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47, 2008.

[62] Sabina Leonelli, Robert P Davey, Elizabeth Arnaud, Geraint Parry, and Ruth Bastow. Data management and best practice for plant science. *Nature plants*, 3(6), 2017.

[63] Laurel Cooper, Austin Meier, Marie-Angélique Laporte, Justin L Elser, Chris Mungall, Brandon T Sinn, Dario Cavaliere, Seth Carbon, Nathan A Dunn, Barry Smith, et al. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic acids research*, 46(D1):D1168–D1180, 2017.

[64] Sarah G Odell, Gerard R Lazo, Margaret R Woodhouse, David L Hane, and Taner Z Sen. The art of curation at a biological database: principles and application. *Current Plant Biology*, 11:2–11, 2017.

[65] Laurel Cooper and Pankaj Jaiswal. The plant ontology: a tool for plant genomics. In *Plant Bioinformatics*, pages 89–114. Springer, 2016.

[66] Abhinav Nellore, Leonardo Collado-Torres, Andrew E Jaffe, José Alquicira-Hernández, Christopher Wilks, Jacob Pritt, James Morton, Jeffrey T Leek, and Ben Langmead. Rail-rna: scalable analysis of rna-seq splicing and coverage. *Bioinformatics*, 33(24):4033–4040, 2016.

[67] Alexander Lachmann, Zhuorui Xie, and Avi Ma'ayan. Elysium: Rna-seq alignment in the cloud. *bioRxiv*, page 382937, 2018.

[68] EBI Gene Expression Team. Expression atlas. https://www.ebi.ac.uk/gxa/about.html, 2019.

[69] Leon French, Suzanne Lane, Tamryn Law, Lydia Xu, and Paul Pavlidis. Application and evaluation of automated semantic annotation of gene expression experiments. *Bioinformatics*, 25(12):1543–1549, 2009.

[70] Darya Filippova. SHARQ: Search Human RNA-Seq Annotations. http://www.cs.cmu.edu/~ckingsf/sharq/, 2016.

[71] Young-suk Lee, Arjun Krishnan, Qian Zhu, and Olga G Troyanskaya. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23):3036–3044, 2013.

[72] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453, 2015.

[73] Dvir Aran, Zicheng Hu, and Atul J Butte. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18(1):220, 2017.

[74] Sam Buckberry, Stephen J Bent, Tina Bianco-Miotto, and Claire T Roberts. massiR: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics*, 30(14):2084–2085, 2014.

[75] Shannon E Ellis, Leonardo Collado-Torres, Andrew Jaffe, and Jeffrey T Leek. Improving the value of public rna-seq expression data by phenotype prediction. *Nucleic acids research*, 46(9):e54–e54, 2018.

[76] Cory B Giles, Chase A Brown, Michael Ripperger, Zane Dennis, Xiavan Roopnarinesingh, Hunter Porter, Aleksandra Perz, and Jonathan D Wren. ALE: automated label extraction from GEO metadata. *BMC bioinformatics*, 18(14):509, 2017.

[77] Sehrish Kanwal, Farah Zaib Khan, Andrew Lonie, and Richard O Sinnott. Investigating reproducibility and tracking provenance–a genomic workflow case study. *BMC bioinformatics*, 18(1):337, 2017.

[78] Florencio Pazos and Monica Chagoyen. Characteristics and evolution of the ecosystem of software tools supporting research in molecular biology. *Briefings in bioinformatics*, 2018.

[79] Pamela H Russell, Rachel L Johnson, Shreyas Ananthan, Benjamin Harnke, and Nichole E Carlson. A large-scale analysis of bioinformatics code on GitHub. *PloS one*, 13(10):e0205898, 2018.

[80] Neha Kulkarni, Luca Alessandrì, Riccardo Panero, Maddalena Arigoni, Martina Olivero, Giulio Ferrero, Francesca Cordero, Marco Beccuti, and Raffaele A Calogero. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC bioinformatics*, 19(10):211, 2018.

[81] Serghei Mangul, Thiago Mosqueiro, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Littman, Benjamin Statz, Angela Lam, et al. A comprehensive analysis of the usability and archival stability of omics computational tools and resources. *bioRxiv*, page 452532, 2018.

[82] Mikhail Dozmorov. GitHub statistics as a measure of the impact of open-source bioinformatics software. *Frontiers in bioengineering and biotechnology*, 6:198, 2018.

[83] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316, 2017.

[84] Jeffrey Perkel. Democratic databases: science on GitHub. *Nature News*, 538(7623):127, 2016.

[85] Karthik Ram. Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine*, 8(1):7, 2013.

[86] Kevin Kskoglund. Understanding version control, Aug 2012.

[87] Jeffrey Perkel. TechBlog: Git: The reproducibility tool scientists love to hate, Jun 2018.

[88] Junio C Hamano. Git–a stupid content tracker. *Proc. Ottawa Linux Sympo*, 1:385–394, 2006.

[89] Klint Finley. What exactly is github anyway?, Jul 2012.

[90] Neil Chue Hong. To achieve the goals of e-science, we must change research culture globally. *Informatik-Spektrum*, 41(6):414–420, 2018.

[91] Heather Piwowar. Altmetrics: Value all research products. *Nature*, 493(7431):159, 2013.

[92] David Crotty. Altmetrics. *European heart journal*, 38(35):2647–2648, 2017.

[93] Grischa Fraumann. The values and limits of altmetrics. *New Directions for Institutional Research*, 2018(178):53–69, 2018.

[94] Philip A Ewels, Alexander Peltzer, Sven Fillinger, Johannes Alneberg, Harshil Patel, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. nf-core: Community curated bioinformatics pipelines. *bioRxiv*, page 610741, 2019.

[95] Jonathan P Tennant, Jonathan M Dugan, Daniel Graziotin, Damien C Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, Lauren B Collister, Christina K Pikas, Tom Crick, et al. A multi-disciplinary perspective on emergent and future innovations in peer review. *F1000Research*, 6, 2017.

[96] Nicolas P Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A Barba, Fabien CY Benureau, C Titus Brown, Pierre De Buyl, Ozan Caglayan, Andrew P Davison, et al. Sustainable computational science: the ReScience initiative. *PeerJ Computer Science*, 3:e142, 2017.

[97] Maria Guerreiro. Forking software used in eLife papers to GitHub, 2017.

[98] Serghei Mangul, Lana S Martin, Eleazar Eskin, and Ran Blekhman. Improving the usability and archival stability of bioinformatics software, 2019.

[99] Reporting standards and availability of data, materials, code and protocols.

[100] Jeremy Leipzig. A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*, 18(3):530–536, 2017.

[101] Shawn Hoon, Kiran Kumar Ratnapu, Jer-ming Chia, Balamurugan Kumarasamy, Xiao Juguang, Michele Clamp, Arne Stabenau, Simon Potter, Laura Clarke, and Elia Stupka.

Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Research*, 13(8):1904–1915, 2003.

[102] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, 2010.

[103] Stephen R Piccolo and Michael B Frampton. Tools and techniques for computational reproducibility. *GigaScience*, 5(1):30, 2016.

[104] Elise Larsonneur, Jonathan Mercier, Nicolas Wiart, Edith Le Floch, Olivier Delhomme, and Vincent Meyer. Evaluating Workflow Management Systems: A Bioinformatics Use Case. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2773–2775. IEEE, 2018.

[105] Jillian Rowe, Nizar Drou, Aymen Yousif, and Kristin Gunsalus. BioSAILs: versatile workflow management for high-throughput data analysis. *bioRxiv*, page 509455, 2019.

[106] Ola Spjuth, Erik Bongcam-Rudloff, Guillermo Carrasco Hernández, Lukas Forer, Mario Giovacchini, Roman Valls Guimera, Aleksi Kallio, Eija Korpelainen, Maciej M Kańduła, Milko Krachunov, et al. Experiences with workflows for automating data-intensive bioinformatics. *Biology direct*, 10(1):43, 2015.

[107] Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. GenePattern 2.0. *Nature genetics*, 38(5):500, 2006.

[108] Florian Halbritter, Harsh J Vaidya, and Simon R Tomlinson. GeneProf: analysis of high-throughput sequencing experiments. *Nature methods*, 9(1):7, 2012.

[109] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–3011, 2009.

[110] Sohrab P Shah, David YM He, Jessica N Sawkins, Jeffrey C Druce, Gerald Quon, Drew Lett, Grace XY Zheng, Tao Xu, and BF Francis Ouellette. Pegasys: software for executing and integrating analyses of biological sequences. *BMC bioinformatics*, 5(1):40, 2004.

[111] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.

[112] Janez Kranjc, Roman Orač, Vid Podpečan, Nada Lavrač, and Marko Robnik-Šikonja. ClowdFlows: Online workflows for distributed big data mining. *Future Generation Computer Systems*, 68:38–58, 2017.

[113] Simon P Sadedin, Bernard Pope, and Alicia Oshlack. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, 28(11):1525–1526, 2012.

[114] Kenjiro Taura, Takuya Matsuzaki, Makoto Miwa, Yoshikazu Kamoshida, Daisaku Yokoyama, Nan Dun, Takeshi Shibata, Choi Sung Jun, and JunâĂŹichi Tsujii. Design and implementation of GXP make–a workflow system based on make. *Future Generation Computer Systems*, 29(2):662–672, 2013.

[115] Masahiro Tanaka and Osamu Tatebe. Pwrake: A parallel and distributed flexible workflow management tool for wide-area data intensive computing. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pages 356–359. ACM, 2010.

[116] Leo Goodstadt. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics*, 26(21):2778–2779, 2010.

[117] Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[118] Francesco Strozzi, Roel Janssen, Ricardo Wurmus, Michael R Crusoe, George Githinji, Paolo Di Tommaso, Dominique Belhachemi, Steffen Möller, Geert Smant, Joep de Ligt, et al. Scalable workflows and reproducible data analysis for genomics. In *Evolutionary Genomics*, pages 723–745. Springer, 2019.

[119] Sebastian Schmeier. Towards reproducible computational biology: An introductory tutorial. 2018.

[120] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475, 2018.

[121] Amanda Cooksey. BioContainers Bonanza. https://www.youtube.com/watch?v=4vI2v8vYtzc, February 2019.

[122] Carl Boettiger. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1):71–79, 2015.

[123] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017.

[124] Lisa Gerhardt, Wahid Bhimji, Markus Fasel, Jeff Porter, Mustafa Mustafa, Doug Jacobsen, Vakho Tsulaia, and Shane Canon. Shifter: Containers for hpc. In *J. Phys. Conf. Ser.*, volume 898, page 082021, 2017.

[125] Brett K Beaulieu-Jones and Casey S Greene. Reproducibility of computational workflows is automated using continuous analysis. *Nature Biotechnology*, 35(4):342, 2017.

[126] Felipe da Veiga Leprevost, Björn A Grüning, Saulo Alves Aflitos, Hannes L Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, et al. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, 33(16):2580–2582, 2017.

[127] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[128] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[129] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[130] Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical Genomics*, pages 93–110. Springer, 2016.

[131] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 11 2012.

[132] Nuno A Fonseca, Robert Petryszak, John Marioni, and Alvis Brazma. iRAP-an integrated RNA-seq Analysis Pipeline. *bioRxiv*, page 005991, 2014.

[133] recount2: analysis-ready rna-seq gene and exon counts datasets.

[134] David W Meinke, J Michael Cherry, Caroline Dean, Steven D Rounsley, and Maarten Koornneef. Arabidopsis thaliana: a model plant for genome analysis. *Science*, 282(5389):662–682, 1998.

[135] Maarten Koornneef and David Meinke. The development of arabidopsis as a model plant. *The Plant Journal*, 61(6):909–921, 2010.

[136] Arabidopsis Genome Initiative et al. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *nature*, 408(6814):796, 2000.

[137] Josiane Meire Toloti Carneiro, Katherine Chacón Madrid, Bruna Caroline Miranda Maciel, and Marco Aurélio Zezzi Arruda. Arabidopsis thaliana and omics approaches: a review. *Journal of Integrated OMICS*, 5(1):1–16, 2015.

[138] Stephen Sherry and Chunlin Xiao. Ncbi sra toolkit technology for next generation sequence data. In *Plant and Animal Genome XX Conference (January 14-18, 2012). Plant and Animal Genome*, 2012.

[139] Alexander Dobin and Thomas R Gingeras. Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics*, 51(1):11–14, 2015.

[140] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1):341, 2012.

[141] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[142] R Hammerén. Comparison of PE and SE for RNA Seq. https://ngisweden.scilifelab.se/file/1540-1_Comparison_of_PE_and_SE_for_RNA-seq.pdf, 2016.

[143] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–1771, 2009.

[144] Takeru Nakazato, Tazro Ohta, and Hidemasa Bono. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*, 8(10):e77910, 2013.

[145] Sho Nishida, Yusuke Kakei, Yukihisa Shimada, and Toru Fujiwara. Genome-wide analysis of specific alterations in transcript structure and accumulation caused by nutrient deficiencies in Arabidopsis thaliana. *The Plant Journal*, 91(4):741–753, 2017.

[146] Tsuneaki Takami, Norikazu Ohnishi, Yuko Kurita, Shoko Iwamura, Miwa Ohnishi, Makoto Kusaba, Tetsuro Mimura, and Wataru Sakamoto. Organelle dna degradation contributes to the efficient use of phosphate in seed plants. *Nature plants*, 4(12):1044, 2018.

[147] James J Furniss, Heather Grey, Zhishuo Wang, Mika Nomoto, Lorna Jackson, Yasuomi Tada, and Steven H Spoel. Proteasome-associated hect-type ubiquitin ligase activity is required for plant immunity. *PLoS pathogens*, 14(11):e1007447, 2018.

[148] Ping Lan, Wenfeng Li, Tuan-Nan Wen, Jeng-Yuan Shiau, Yu-Ching Wu, Wendar Lin, and Wolfgang Schmidt. itraq protein profile analysis of arabidopsis roots reveals new aspects critical for iron homeostasis. *Plant physiology*, 155(2):821–834, 2011.

[149] Yong Ding, Ning Liu, Laetitia Virlouvet, Jean-Jack Riethoven, Michael Fromm, and Zoya Avramova. Four distinct types of dehydration stress memory genes in Arabidopsis thaliana. *BMC plant biology*, 13(1):229, 2013.

[150] Björn Pietzenuk, Catarine Markus, Hervé Gaubert, Navratan Bagwan, Aldo Merotto, Etienne Bucher, and Ales Pecinka. Recurrent evolution of heat-responsiveness in brassicaceae copia elements. *Genome biology*, 17(1):209, 2016.

[151] Liang Song, Shao-shan Carol Huang, Aaron Wise, Rosa Castanon, Joseph R Nery, Huaming Chen, Marina Watanabe, Jerushah Thomas, Ziv Bar-Joseph, and Joseph R Ecker. A transcription factor hierarchy defines an environmental stress response network. *Science*, 354(6312):aag1550, 2016.

[152] Markus V Kohnen, Emanuel Schmid-Siegert, Martine Trevisan, Laure Allenbach Petrolati, Fabien Sénéchal, Patricia Müller-Moulé, Julin Maloof, Ioannis Xenarios, and Christian Fankhauser. Neighbor detection induces organ-specific transcriptomes, revealing patterns underlying hypocotyl-specific growth. *The Plant Cell*, 28(12):2889–2904, 2016.

[153] Adil Hussain, Bong-Gyu Mun, Qari M Imran, Sang-Uk Lee, Teferi A Adamu, Muhammad Shahid, Kyung-Min Kim, and Byung-Wook Yun. Nitric oxide mediated transcriptome profiling reveals activation of multiple regulatory pathways in Arabidopsis thaliana. *Frontiers in plant science*, 7:975, 2016.

[154] Fan Zhang, Jae Yun Lim, Taewook Kim, Youngjae Pyo, Sibum Sung, Chanseok Shin, Hong Qiao, et al. Phosphorylation of cbp20 links microrna to root growth in the ethylene response. *PLoS genetics*, 12(11):e1006437, 2016.

[155] Waleed S Albihlal, Irabonosi Obomighie, Thomas Blein, Ramona Persad, Igor Chernukhin, Martin Crespi, Ulrike Bechtold, and Philip M Mullineaux. Arabidopsis HEAT SHOCK TRANSCRIPTION FACTORA1b regulates multiple developmental genes under benign and stress conditions. *Journal of experimental botany*, 69(11):2847–2862, 2018.

[156] Marion Eisenhut, Andrea Bräutigam, Stefan Timm, Alexandra Florian, Takayuki Tohge, Alisdair R Fernie, Hermann Bauwe, and Andreas PM Weber. Photorespiration is crucial for dynamic response of photosynthetic metabolism and stomatal movement to altered CO2 availability. *Molecular plant*, 10(1):47–61, 2017.

[157] Louis Grillet, Ping Lan, Wenfeng Li, Girish Mokkapati, and Wolfgang Schmidt. IRON MAN is a ubiquitous family of peptides that control iron transport in plants. *Nature plants*, 4(11):953, 2018.

[158] Shan Lu, Chenyi Li, Ye Zhang, Zai Zheng, and Dong Liu. Functional disruption of a chloroplast pseudouridine synthase desensitizes Arabidopsis plants to phosphate starvation. *Frontiers in plant science*, 8:1421, 2017.

[159] Trevor M Nolan, Benjamin Brennan, Mengran Yang, Jiani Chen, Mingcai Zhang, Zhaohu Li, Xuelu Wang, Diane C Bassham, Justin Walley, and Yanhai Yin. Selective autophagy of BES1 mediated by DSK2 balances plant growth and survival. *Developmental cell*, 41(1):33–46, 2017.

[160] Shuang-Shuang Zhang, Hongxing Yang, Lan Ding, Ze-Ting Song, Hong Ma, Fang Chang, and Jian-Xiang Liu. Tissue-specific transcriptomics reveals an important role of the unfolded protein response in maintaining fertility upon heat stress in Arabidopsis. *The Plant Cell*, 29(5):1007–1023, 2017.

[161] Chang Liu, Ying Xin, Le Xu, Zhaokui Cai, Yuanchao Xue, Yong Liu, Daoxin Xie, Yule Liu, and Yijun Qi. Arabidopsis ARGONAUTE 1 binds chromatin to promote gene transcription in response to hormones and stresses. *Developmental cell*, 44(3):348–361, 2018.

[162] Yujian Jin, Xiaoji Fan, Xingxing Li, Zhenyan Zhang, Liwei Sun, Zhengwei Fu, Michel Lavoie, Xiangliang Pan, and Haifeng Qian. Distinct physiological and molecular responses in Arabidopsis thaliana exposed to aluminum oxide nanoparticles and ionic aluminum. *Environmental pollution*, 228:517–527, 2017.

[163] Sara Martins, Alvaro Montiel-Jorda, Anne Cayrel, Stéphanie Huguet, Christine Paysant-Le Roux, Karin Ljung, and Grégory Vert. Brassinosteroid signaling-dependent root responses to prolonged elevated ambient temperature. *Nature communications*, 8(1):309, 2017.

[164] Yingfang Zhu, Bangshing Wang, Kai Tang, Chuan-Chih Hsu, Shaojun Xie, Hai Du, Yuting Yang, Weiguo Andy Tao, and Jian-Kang Zhu. An Arabidopsis Nucleoporin NUP85 modulates plant responses to ABA and salt stress. *PLoS genetics*, 13(12):e1007124, 2017.

[165] Paola Punzo, Alessandra Ruggiero, Marco Possenti, Roberta Nurcato, Antonello Costa, Giorgio Morelli, Stefania Grillo, and Giorgia Batelli. The PP 2A-interactor TIP 41 modulates ABA responses in Arabidopsis thaliana. *The Plant Journal*, 94(6):991–1009, 2018.

[166] Therese C Rytz, Marcus J Miller, Fionn McLoughlin, Robert C Augustine, Richard S Marshall, Yu-ting Juan, Yee-yung Charng, Mark Scalf, Lloyd M Smith, and Richard D Vierstra. Sumoylome profiling reveals a diverse array of nuclear targets modified by the sumo ligase siz1 during heat stress. *The Plant Cell*, 30(5):1077–1099, 2018.

[167] Min May Wong, Govinal Badiger Bhaskara, Tuan-Nan Wen, Wen-Dar Lin, Thao Thi Nguyen, Geeng Loo Chong, and Paul Verslues. Phosphoproteomics of Highly ABA-Induced1 identifies AT Hook Like10 phosphorylation required for growth regulation during stress. *bioRxiv*, page 413013, 2018.

[168] Yang Zhao, Zhengjing Zhang, Jinghui Gao, Pengcheng Wang, Tao Hu, Zegang Wang, Yueh-Ju Hou, Yizhen Wan, Wenshan Liu, Shaojun Xie, et al. Arabidopsis duodecuple mutant of PYL ABA receptors reveals PYL repression of ABA-independent SnRK2 activity. *Cell reports*, 23(11):3340–3351, 2018.

[169] Maroua Bouzid, Fei He, Gregor Schmitz, RE Häusler, APM Weber, T Mettler-Altmann, and Juliette De Meaux. Arabidopsis species deploy distinct strategies to cope with drought stress. *Annals of botany*, 124(1):27–40, 2019.

[170] Nora Marín-de la Rosa, Chung-Wen Lin, Yang Jae Kang, Stijn Dhondt, Nathalie Gonzalez, Dirk Inzé, and Pascal Falter-Braun. Drought resistance is mediated by divergent strategies in closely related Brassicaceae. *New Phytologist*, 2019.

[171] Viviane Cordovez, Liesje Mommer, Kay Moisan, Dani Lucas-Barbosa, Ronald Pierik, Roland Mumm, Victor J Carrion, and Jos M Raaijmakers. Plant phenotypic and transcriptional changes induced by volatiles from the fungal root pathogen rhizoctonia solani. *Frontiers in plant science*, 8:1262, 2017.

[172] Rafael Catalá, Cristian Carrasco-López, Carlos Perea-Resa, Tamara Hernandez-Verdeja, and Julio Salinas. Emerging roles of lsm complexes in posttranscriptional regulation of plant response to abiotic stress. *Frontiers in plant science*, 10, 2019.

[173] Chunzhao Zhao, Rong Yang, Yechun Hong, Zhizhong Ren, Kai Tang, Heng Zhang, and Jian-Kang Zhu. A role for pickle in the regulation of cold and salt stress tolerance in arabidopsis. *Frontiers in plant science*, 10:900, 2019.

[174] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[175] Seung Yon Rhee, William Beavis, Tanya Z Berardini, Guanghong Chen, David Dixon, Aisling Doyle, Margarita Garcia-Hernandez, Eva Huala, Gabriel Lander, Mary Montoya, et al. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic acids research*, 31(1):224–228, 2003.

[176] Tanya Z Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait, and Eva Huala. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*, 53(8):474–485, 2015.

[177] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[178] Mark F Rogers, Julie Thomas, Anireddy SN Reddy, and Asa Ben-Hur. SpliceGrapher: detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. *Genome biology*, 13(1):R4, 2012.

[179] Andreas Heger. Pysam. https://github.com/pysam-developers/pysam, 2019.

[180] Sonali Arora, Siobhan S Pattwell, Eric C Holland, and Hamid Bolouri. Uncertainty in RNA-seq gene expression data. *BioRxiv*, page 445601, 2018.

[181] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling

the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733, 2010.

[182] Michelle Badri, Zachary Kurtz, Christian Muller, and Richard Bonneau. Normalization methods for microbial abundance data strongly affect correlation estimates. *bioRxiv*, page 406264, 2018.

[183] Shuangge Ma and Ying Dai. Principal component analysis based methods in bioinformatics studies. *Briefings in bioinformatics*, 12(6):714–722, 2011.

[184] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303, 2008.

[185] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[186] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

[187] Laurens Van Der Maaten. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*, 2013.

[188] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[189] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[190] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[191] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[192] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[193] Benoît De Hertogh, Bertrand De Meulder, Fabrice Berger, Michael Pierre, Eric Bareke, Anthoula Gaigneaux, and Eric Depiereux. A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC bioinformatics*, 11(1):17, 2010.

[194] Nicholas J Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, Tom Owen-Hughes, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna*, 22(6):839–851, 2016.

[195] Josh Starmer. StatQuest: DESeq2, part 1, Library Normalization. https://www.youtube.com/watch?v=UFB993xufUU, 2017.

[196] Mike Hamilton, ASN Reddy, and Asa Ben-Hur. Predicting differential intron retention with iDiffIR. *Plant and Animal Genome*, 2016. http://combi.cs.colostate.edu/idiffir/introduction.html.

[197] Pragya Kant, Michal Gordon, Surya Kant, Gaston Zolla, Olga Davydov, Yair M Heimer, VERED CHALIFA-CASPI, Ruth Shaked, and Simon Barak. Functional-genomics-based identification of genes that regulate arabidopsis responses to multiple abiotic stresses. *Plant, Cell & Environment*, 31(6):697–714, 2008.