Techniques & Applications

Becoming data-savvy in a big-data world

Meng Xu and Seung Yon Rhee

Carnegie Institution for Science, Department of Plant Biology, 260 Panama Street, Stanford CA 94305, USA

Plant biology is becoming a data-driven science. High-throughput technologies generate data quickly from molecular to ecosystem levels. Statistical and computational approaches enable describing and interpreting data quantitatively. We highlight the purpose, common problems, and general principles in data analysis. We use RNA sequencing (RNAseq) analysis to illustrate the rationale behind some of the choices made in statistical data analysis. Finally, we provide a list of free online resources that emphasize intuition behind quantitative data analysis.

Big data analysis

Advances in digital, information, and communications technologies generate enormous amounts of data – coined big data – in all sectors of our society [The Economist (2010) http://www.economist.com/node/15557443]. To gain insight from big data, we need to aggregate, manipulate, manage, analyze, find patterns, and visualize the data. By 2018, demand for the needed talent to capitalize on big data is estimated to exceed the supply by 50–60% (http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation). The needed talent will be educated in science, technology, engineering, and mathematics (STEM), and well versed in statistics, computer science, and applied mathematics.

Big data has infiltrated most scientific fields ranging from climate science, astronomy, neuroscience, biomedicine, and genomics to plant biology. Big data refers to datasets that are: (i) too large to be stored and manipulated through a relational database on a single computer; or (ii) too complex to be analyzed and understood easily [1]. Without the necessary mindset and tools to find patterns in the data holistically, we cannot capitalize on the copious amounts of data that are readily accessible to us today.

Data analysis is an approach that combines statistics and programming to explore, make inferences from, and make predictions about data. It transforms data into statistical evidence to support or reject hypotheses. This may sound daunting for many plant biologists who have not been trained formally in data analysis concepts and techniques. Fortunately, the ubiquity of big data has created many online resources and tools to suit people at all levels of training and to help scientists acquire the necessary data analysis skills. In this Techniques and Applications article, we introduce the basic concepts of data analysis and provide useful free online books and courses that can help lower the barrier for thinking quantitatively and performing data analysis with ease.

Common problems in data analysis

Many problems in data analysis arise from limited understanding of fundamental concepts in data analysis. In this section, we highlight a few problems that occur frequently in plant biology publications, and address them in more detail in the next section.

How to compare two or more variables is a good example of misuse of data analysis concepts and methods. For example, to ask whether a gene's mRNA level is significantly changed after a treatment, we measure the gene's mRNA level multiple times before and after the treatment. We want to compare the results using a statistical test, but how do we choose the correct test to use? Also, after we get a significance level in the form of a P value, how do we interpret it? Specifically, does a P value of 0.05 mean we have a 5% chance of making a wrong conclusion?

Another common problem scientists encounter is how to find correlations among variables. For example, highly correlated mRNA levels of two genes across multiple treatments or developmental stages could imply potential gene function association. In order to detect correlation, Pearson's and Spearman's correlation coefficients are widely used. What are the strengths and weaknesses of each choice, which one should be used, and when? If we obtained a correlation coefficient between two genes that is significantly higher than the average level using randomly chosen gene pairs or the entire sample space, does it mean that the two genes are highly correlated in their expression?

We need to make many choices and interpretations when analyzing data quantitatively. In order to address these problems appropriately, we need to understand the fundamental principles behind data analysis methods.

Key concepts in data analysis

Understanding key principles of data analysis can help analyze data appropriately and efficiently. In this section, we discuss a few general concepts in data analysis that are useful in common tasks, including selecting statistical tests, interpreting the results, and conducting multiple tests.

Distribution patterns of data

It is important to understand how the data is distributed before deciding how to analyze it quantitatively. Distribution refers to all possible values that a measurement could take, and the probabilities of each possible value. It describes the range and shape of the data quantitatively. Many statistical methods make assumptions about data

Corresponding author: Rhee, S.Y. (srhee@carnegiescience.edu).

Keywords: statistics; data analysis; big data; RNAseq analysis.

1360-1385/

http://dx.doi.org/10.1016/j.tplants.2014.08.003

Box 1. RNAseq as a case study for applying key concepts in data analysis

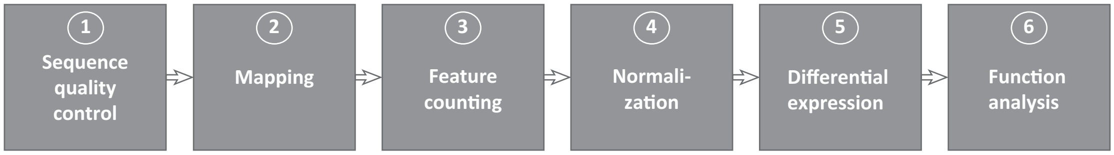
RNAseq is a powerful method for measuring transcript abundance of all genes in a sample. How to analyze RNAseq data is a good example of applying key data analysis concepts discussed in this article to real data. In this section, we briefly describe the main strategy used in RNAseq data analysis, and highlight the rationale behind the common statistical methods that are used.

The main strategy of analyzing RNAseq data is to count the number of sequence reads for each gene and infer differentially expressed genes across experimental groups [8,9]. We divide the data analysis process into six steps, as shown in the flow chart (Figure I).

The first four steps aim to count the number of reads for each gene [9,10]. First, sequence quality control analysis eliminates sequencing errors. High-quality reads are then mapped to a reference genome by sequence alignments. The number of mapped reads is counted. The gene counts should be normalized to account for sample or gene specific biases caused by factors such as different sequencing depth across samples or different gene lengths.

Differential gene expression analysis identifies genes with significantly changed expression levels using statistical tests. As described in the main text, a correct assumption of data distribution is critical for reliable parametric statistical tests. The negative binomial (NB) distribution has been shown to be a good model for count data [10]. Several popular tools like edgeR [11] and DESeq [12] are based on NB distribution. Nonparametric approaches such as SAMseq [13] do not assume a specific distribution and is appropriate only with moderate or large sample size (at least 12 replicates). Multiple hypothesis testing is needed because we often test differential expression of multiple genes. Therefore, it is important to control the false discovery rate (FDR). Currently, no single differential expression analysis method outperforms all others, and pros and cons of the methods should be considered when choosing a method [14]. An RNA-seq analysis exercise using Galaxy [15] may be useful to get a hands-on experience (https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise).

Gene function analysis assigns biological significance to genes identified from differential gene expression analysis. The strategy is to label genes with annotations and test for over- (or under-) representation of annotations compared to an appropriate background. For example, Gene Ontology (GO) enrichment analysis is commonly used to detect enrichment or depletion of functions in the gene groups of interest [16].



TRENDS in Plant Science

Figure I. Flow chart of data analysis process in RNAseq analysis.

distribution. Results from statistical analyses could only make sense when all the assumptions are (at least approximately) true. For example, Pearson's correlation coefficient assumes normally distributed variables and no significant outliers. Spearman's correlation coefficient can be used when these assumptions are violated at the cost of lower sensitivity. It transforms data points into rank order, and thus produces relatively robust results for different distributions. An appropriate distribution assumption is also important in RNAseq analysis (Box 1). Visualizing the data by graphing them (e.g., scatter plot, bar chart, box plot, etc.) can help explore data distributions as well as find potential patterns and outliers.

Statistical hypothesis testing

A statistical test allows decisions to be made about the data. A biological hypothesis is transformed into null and alternative hypotheses, in order to distinguish the two possibilities statistically. A null hypothesis (H_0) often refers to an assumption that the experimental observations (e.g., difference, correlation, enrichment, etc.) are obtained purely by chance. An alternative hypothesis (H_1) is an assumption of non-random effects in the observations compared to the background or control. These assumptions may or may not be true. Statistical tests are procedures to determine only whether we should reject the null hypothesis. Statistical hypothesis tests are widely used to compare large-scale data; for example, to find differentially expressed genes from RNAseq data and enriched biological functions in groups of differentially expressed genes (Box 1).

Test statistic and null distribution

A statistical test starts with choosing a test statistic and deriving its null distribution. The test statistic is a variable defined to quantify the population difference. The choice of the test statistic depends on the type of population difference to be tested. For example, for normally distributed numerical data (such as length or weight), we could quantify the mean difference between populations from which samples are drawn using the *t*-statistic. For categorical data (such as positive or negative response to treatment), we could quantify the frequency difference between groups using the chi-square statistic. The distribution of the test statistic under the null hypothesis is called the null distribution, which represents the background level of the population difference. If the chance of obtaining the observed data's test statistic value (or even more extreme values) from the null distribution is rare, the null hypothesis is likely to be rejected.

Assumptions in a statistical test

Many statistical tests make assumptions about data distribution and sample size because the null distribution depends on these parameters [2]. If the assumptions are not satisfied, the null distribution may not represent the correct background level of the test statistic, and the statistical test may produce biased results.

What if we are not sure what the data distribution is? The answer depends on the sample size. When the sample size is large, we can assume the null distribution to be normal because of the central limit theorem [3]. When the sample size is small (as is the case for many biological experiments with 3–6 replicates), statistical tests assuming normal distribution may generate misleading results.

Table 1. Useful online books and MOOCs on basic statistics

	Title	URL	Refs
Book	Handbooks of Biological Statistics (Introduction to choosing and applying appropriate statistical tests for particular experiments)	http://www.biostathandbook.com/	[2]
Book	OpenIntro Statistics (Introduction to data structures and statistical tools for data analysis)	http://www.openintro.org/stat/textbook.php	[17]
Book	The Elements of Statistical Learning (A comprehensive textbook on data mining and machine learning)	http://statweb.stanford.edu/~tibs/ElemStatLearn/download.html	[18]
Book	Advanced Data Analysis from an Elementary Point of View (Data analysis methods intended for advanced undergraduate students)	http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/	[19]
MOOC	Statistics One (Introductory statistics concepts assuming very little prior knowledge)	https://www.coursera.org/course/stats1	
MOOC	Case-Based Introduction to Biostatistics (Introduction to popular statistical methods in biochemical studies)	https://www.coursera.org/course/casebasedbiostat	
MOOC	Data Analysis (Introduction to applying statistical techniques to real data using R programming language, interpreting the results, and diagnosing potential problems in data analysis)	https://www.coursera.org/course/dataanalysis	
MOOC	Data Analysis and Statistical Inference (Introduction to statistical inference and linear regression)	https://www.coursera.org/course/statistics	

Instead, empirically derived data distributions or nonparametric (distribution-free) alternative tests should be used (Box 1). Even if the data distribution is known, careful examination of outliers is needed. Data outliers may dramatically shift the sample means, especially when sample size is small.

Meaning of a P value

A statistical test returns a P value that refers to the probability of the data's test statistic being observed from the null distribution when the null hypothesis is true. A P value is widely used as an indicator to make decisions about statistical significance. When the P value is below a certain threshold, usually 0.05 or 0.01 for historical reasons [4], we reject the null hypothesis and accept the alternative hypothesis. Recent research suggests that this conventional threshold is too optimistic, and a 0.005 or 0.001 level of significance is more appropriate for conducting reproducible studies [5].

The P value says nothing about the probability of H_0. Therefore, a P value of 0.05 does not mean a 5% chance of having false positive results. Rather, it means the test statistic value (or even more extreme values) can be observed at a 5% probability purely by chance. The actual false positive rate could be much higher [6]. A P value can be used only to reject the H_0, but not to accept it. A high P value does not mean the H_0 is true, or that there is no difference between the samples; it simply means that the evidence is not strong enough to reject H_0 at the specified significance level.

Statistical significance represented by P values may not necessarily predicate biological importance [6]. Minor differences between samples can be deemed statistically significant given large enough sample size. For example, a Pearson's correlation coefficient that is significantly different from zero does not necessarily imply a strong correlation. The degree of correlation depends on the value of the correlation coefficient itself rather than the significance level.

Controlling for false positive rates

The number of statistical tests performed in an analysis affects the number potential false positives. When testing multiple hypotheses, the P value is no longer a good measure of significance because of high cumulative false positive results. Each test has a certain probability of getting a false positive result. Therefore, the number of total false positive results could be high if thousands of hypothesis tests are conducted, such as differential expression analysis using microarrays or RNAseq data (Box 1). Two approaches to address this problem are family-wise error rate control and false-positive rate control [7].

These fundamental concepts are just a few examples of ideas and approaches commonly used in data analysis. More theoretical and practical skills can be obtained from books, courses, and online resources.

Resources for gaining skills in data analysis

The Internet is the best source for getting one's feet wet into data analysis. Some structured learning about basic statistical concepts and methods could help. There are several online books and Massive Open Online Courses (MOOCs) that we found particularly helpful for providing a solid background in statistics (Table 1). However, these are just a few courses out of many others that we have not evaluated. Therefore, we encourage the readers to explore the MOOC space to find the next level courses or other introductory courses. If one comes across a problem that is not readily accessible from the resources suggested here, we encourage taking advantage of local help. For example, most universities have open office hours in the Statistics or Applied Mathematics departments.

Acknowledgments

We thank J. Guo, R. Zhang, and R. Rellán-Álvarez for their comments on the manuscript. This work was supported in part by grants from the National Science Foundation (IOS-1026003) and the Department of Energy (BER-65472).

References

1 Ward, J.S. and Barker, A. (2013) Undefined by data: a survey of big data definitions. *ArXiv:1309.5821.[cs.DB]*

2 McDonald, J.H. (2009) *Handbook of Biological Statistics*. (2nd edn), Sparky House Publishing

3 Lumley, T. *et al.* (2002) The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* 23, 151–169

4 Fisher, R.A. (1934) Chapter III: Distributions. In *Statistical Methods for Research Workers.* (5th edn) (Crew, F.A.E. and Cutler, D.W., eds), pp. 41–70, Oliver and Boyd

5 Johnson, V.E. (2013) Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19313–19317

6 Nuzzo, R. (2014) Scientific method: statistical errors. *Nature* 506, 150–152

7 Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445

8 Van Verk, M.C. *et al.* (2013) RNA-Seq: revelation of the messengers. *Trends Plant Sci.* 18, 175–179

9 Anders, S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* 8, 1765–1786

10 Bullard, J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94

11 Robinson, M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140

12 Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* 11, R106

13 Li, J. and Tibshirani, R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* 22, 519–536

14 Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14, 91

15 Blankenberg, D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* 10, 11–21 Chapter 19 Unit 19

16 Huang, D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13

17 Diez, D.M. *et al.* (2012) *OpenIntro Statistics: Second Edition*, CreateSpace Independent Publishing Platform

18 Hastie, T. *et al.* (2011) *The Elements of Statistical Learning*, Springer

19 Shalizi, C.R. (2013) *Advanced Data Analysis from an Elementary Point of View*, Cambridge University Press