IMPLEMENTING VALIDATION PROCEDURES TO STUDY THE PROPERTIES OF WIDELY USED STATISTICAL ANALYSIS METHODS OF RNA SEQUENCING EXPERIMENTS

By

Pablo Daniel Reeb

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Fisheries and Wildlife-Doctor of Philosophy

2015

ABSTRACT

IMPLEMENTING VALIDATION PROCEDURES TO STUDY THE PROPERTIES OF WIDELY USED STATISTICAL ANALYSIS METHODS OF RNA SEQUENCING EXPERIMENTS

By

Pablo Daniel Reeb

RNA sequencing (RNA-seq) technology is being rapidly adopted as the platform of choice for transcriptomic studies. Although its major focus has been gene expression profiling, other interests, such as single nucleotide profiling, are emerging as the technology evolves. In addition, applications are being rapidly expanding in model and nonmodel organisms. The overall objective of this dissertation was to propose and implement validation procedures based on experimental data to estimate the properties of widely used statistical analysis methods of RNA-seq experiments.

The first study evaluated differential expression methods based on count data distribution and Gaussian transformed models. Parametric simulations and plasmode datasets derived from RNA-seq experiments were generated to compare the statistical models in terms of type I error rate, power and null p-value distribution. Overall, Gaussian models presented p-values closer to nominal significance levels and a p-value distribution closer to the expected uniform distribution. Researchers using models with these properties will have less false positives when inferring differentially expresses transcripts. Additionally, the use of Gaussian transformations enables the applications of all the well-known theory of linear models for instance to account for complex experimental designs.

The second study assessed the properties of dissimilarity measures for agglomerative hierarchical cluster analysis. The validation comprised dissimilarity measures based on Euclidean distance, correlation-based dissimilarities and count data-based dissimilarities. I used plasmode

datasets generated from two RNA-seq experiments with different sample structures and simulated scenarios based on informative and non informative transcripts. In addition, I proposed two measures, agreement and consistency, for comparing dendrograms. Dissimilarity measures based on non-transformed data resulted in dendrograms that did not resemble the expected sample structure, whereas dissimilarities calculated with appropriate transformations for count data were consistent in reproducing the expected dendrograms under different scenarios.

The third study compared variant calling programs that used reference genotypes obtained from a SNP chip. The evaluation included multiple samples and multiple tissue datasets and considered the effect of per base read depth. Sensitivity and false discovery rates were computed separately for heterozygous and homozygous sites in order to provide information for potentially different applications such as allele-specific expression or RNA-editing. Additionally, I explored the use of SNP called from RNA-seq to compute relationship matrices in population studies. Heterozygous sites with more than 10 reads per base and per sample were called with high sensitivity and low false discovery rates. Homozygous sites were called with higher sensitivity than heterozygous irrespective of depth but presented higher false discovery rates. A relationship matrix based on accurate genotypes obtained with RNA-seq presented a high correlation with a relation matrix based on genotypes from a SNP chip.

In conclusion, using synthetic and reference datasets, I compared statistical models to perform differential expression analysis, sampled-base hierarchical cluster analysis, and variant calling and genotyping. This validation framework can be extended to evaluate other methods of RNA-seq analysis as well as to evaluate the periodic publication of new and updated analysis methods. Choosing the most appropriate software can help researchers to obtained better results and to achieve the goals of their investigations.

To Romina, Sophia and Olivia.

ACKNOWLEDGMENTS

I dedicate this dissertation to my lovely family: Romina, Sophia and Olivia. Saying thank you will never be sufficient. Romina and I simultaneously embraced the challenges of building a family and moving abroad to continue my education. Both ventures have been tough but both were worth the efforts. We can say that one challenge has arrived to the end after writing five chapters. The remaining challenge has two little sprouts, Sophia and Olivia, and I hope we will never finish writing that story. I love you!

An especial recognition goes to my advisor Dr. Juan P. Steibel for being extremely patient and supporting throughout the whole program. Certainly, he has been more than a mentor while giving guidelines and advice, and offering friendship in the right moments.

I would like to say thank you to the members of my guidance committee Drs. James Bence, C. Titus Brown, Weiming Li, and Robert Tempelman. They all helped me to build a comprehensive academic background from their specific disciplines.

I shared a productive and friendly academic environment at Michigan State University, especially at the Animal Breeding and Genetics Group. I am taking with me an enriched experience of collaborative work with graduate students, professors, visiting scholars and post-docs.

Finally, I want to thank my large family and friends. All your short or long messages, phone calls, and hugs given virtually from distant places or in person in Michigan were essential for me and my family.

TABLE OF CONTENTS

LIST OF TABLES....ix

LIST OF FIGURES x

Chapter 1 Introduction.... 1

Chapter 2 Evaluating statistical analysis models for RNA sequencing experiments 11
- 2.1 Introduction....12
- 2.2 Material and Methods 15
  - 2.2.1 Simulations 15
  - 2.2.2 Plasmodes 18
    - 2.2.2.1 Null dataset (Cheung)....18
    - 2.2.2.2 Differentially expressed dataset (Bottomly) 20
  - 2.2.3 Comparison of alternative analysis tools for evaluating differential expression....22
    - 2.2.3.1 Filtering and normalization....22
    - 2.2.3.2 Differential expression analysis 23
    - 2.2.3.3 Multiple comparisons....23
  - 2.2.4 Evaluating and comparing results from alternative analysis packages 24
- 2.3 Results....26
  - 2.3.1 Simulations 26
  - 2.3.2 Plasmodes 28
    - 2.3.2.1 Null dataset (Cheung)....28
    - 2.3.2.2 DE dataset (Bottomly)....29
- 2.4 Discussion and conclusion....31
- APPENDIX 35

Chapter 3 Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using plasmode datasets....41
- 3.1 Introduction....42
- 3.2 Material and Methods 44
  - 3.2.1 Datasets....44
  - 3.2.2 Plasmodes 44
    - 3.2.2.1 Plasmodes from Bottomly dataset 45
    - 3.2.2.2 Plasmodes from MSUPRP dataset 47
  - 3.2.3 Clustering....49
  - 3.2.4 Cluster validation using results from plasmodes....50
- 3.3 Results....51
  - 3.3.1 Bottomly....51
  - 3.3.2 MSUPRP 57
- 3.4 Discussion 60
- 3.5 Conclusion....66

APPENDIX 68

Chapter 4 Assessing genotype call accuracy from RNA sequencing data 75
4.1 Introduction 76
4.2 Material and Methods 78
4.2.1 Resource population 78
4.2.2 Variant calling 79
4.2.3 Assessing genotype call accuracy 80
4.2.4 Equivalence between relationship matrices and correlation with SNP chip data 82
4.3 Results 83
4.3.1 Analysis across variant calling programs for 24 animals 83
4.3.2 Comparison of genotype calls from DNA-seq and RNA-seq from multiple tissues 85
4.3.3 Equivalence between relationship matrices 87
4.4 Discussion 89

Chapter 5 General discussion 94
5.1 Introduction 95
5.2 Objectives revisited 95
5.3 Future research directions 101

REFERENCES 103

LIST OF TABLES

Table 1 Classification rule to compute false and true positive rates..25

Table 2 Consistency for each dissimilarity measure..56

Table 3 Reference similarity matrix (S) for MSUPRP plasmodes 72

Table 4 Reference dissimilarity matrix (D) for MSUPRP plasmodes...72

Table 5 Cophenetic matrix of the reference dendrogram for MSUPRP plasmodes. 73

Table 6 Correlation between dissimilarity measures and reference dissimilarity.......................74

Table 7 Contingency table with classification rule used to compute sensitivity and error rate of genotype calling..81

LIST OF FIGURES

Figure 1 Algorithm used to simulate counts from existing estimates of model parameters 17

Figure 2 Multidimensional scaling analysis of experimental datasets 19

Figure 3 Algorithm used to generate plasmode datasets with no differentially expressed transcripts under a model with one classification variable 20

Figure 4 Algorithm used to generate plasmode datasets with differentially expressed transcripts under a model with two classification variables (block + treatment)..21

Figure 5 Simulation results from a scenario with three biological replicates 27

Figure 6 Plasmode results from Cheung dataset:..28

Figure 7 Plasmode results from Bottomly dataset: 29

Figure 8 Comparison of MAANOVA's p-value results...30

Figure 9 Steps used to process reads to obtain matrix of counts 36

Figure 10 P-value distribution of differential expression analysis performed with edgeR for Bottomly data using a model with block and treatment fixed effects..37

Figure 11 P-value distribution of non differentially expressed transcripts for a simulated scenario with three biological replicates 38

Figure 12 P-value distribution of non differentially expressed transcripts for plasmodes generated from Cheung dataset..39

Figure 13 P-value distribution of non differentially expressed transcripts for plasmodes generated from Bottomly dataset 40

Figure 14 Algorithm used to generate plasmodes from Bottomly dataset 47

Figure 15 Multidimensional scaling analysis of MSUPRP dataset 48

Figure 16 Typical dendrograms obtained for plasmode datasets from Bottomly experimental data with two dissimilarity measures under three scenarios 53

Figure 17 Agreement between dissimilarity measures using Bottomly plasmode datasets........54

Figure 18 Typical dendrograms obtained for plasmode datasets from MSUPRP experimental data with two dissimilarity measures 58

Figure 19 Agreement between dissimilarity measures using MSUPRP plasmode datasets 59

Figure 20 Sequencing layout of 24 barcoded samples...69

Figure 21 Plasmode generation for MSUPRP dataset..71

Figure 22 Reference dendrogram for MSUPRP plasmodes 73

Figure 23 Assessing SNP calling accuracy of RNA-seq samples of *longisimus dorsi*................84

Figure 24 Assessing SNP calling accuracy of DNA and RNA-seq samples of three tissues from the same animal...87

Figure 25 Scatterplot of correlation between relationship matrices for the 24 F2 females 89

Chapter 1

Introduction

Transcriptome analysis through next generation sequencing technologies (Metzker, 2010) is known as RNA sequencing (RNA-seq) (Wang et al., 2009). This technology has revolutionized transcriptomics since its first introduction in 2004 (Bennett, 2004) and has been rapidly applied to a variety of studies and species (Ozsolak and Milos, 2011; Marguerat et al., 2010; Wickramasinghe et al., 2014). Gene expression analysis has been the main objective in RNA-seq experiments (Oshlack et al., 2010) but other quantitative and qualitative aspects of transcript biology have been increasing in importance as the technology evolves. These other objectives include studies for detecting allele-specific expression (Quinn et al., 2014; Pirinen et al., 2015; Steibel et al., 2015), RNA editing (Lee et al., 2013; Ramaswami et al., 2013), alternative splicing (Griffith et al., 2010; Alamancos et al., 2014), novel transcripts (Roberts et al., 2011; Trapnell et al., 2010), and nucleotide variations (Quinn et al., 2013; Cánovas et al., 2010).

A typical RNA-seq experiment includes the following steps (Marguerat et al., 2010; Oshlack et al., 2010). First, a sample is extracted from the tissue of interest, then RNA is purified and submitted to a series of processes known as library preparation. This process comprises poly-A RNA isolation, RNA fragmentation, reverse transcription to cDNA, adapter ligation, size selection, and PCR enrichment. The library preparation converts the input RNA into small fragments of cDNA that are ready to be sequenced. Second, the cDNA libraries are placed in the sequencing machine and the fragments are sequenced in parallel in a predefined number of cycles. At each cycle, fluorescent labeled nucleotides are added and the signals emitted are recorded and converted to base-calls. As a result, the sequencing machine provides a number of short reads with length equal to the number of cycles and with read abundance proportional to the fragment abundance. Third, the sequenced reads are filtered (for quality) and mapped to a reference genome or transcriptome (either pre-existing or assembled from the sequenced reads themselves).

After reads have been aligned, qualitative and quantitative information can be extracted from them in order to be applied for different types of analysis. Qualitative information refers to the study of the sequences of bases *per se* and can be used, for example, for discovering and calling single nucleotide polymorphisms (SNPs). Quantitative information, on the other hand, is based on the characteristic that the number of reads aligning to any given region of the genome is proportional to the abundance of fragments for that region within the sample (Mortazavi et al., 2008). These read counts can be summarized and aggregated over biologically meaningful units, such as exons, transcripts, or genes in order to provide a direct measure of expression (Oshlack et al., 2010). The most common use of this expression profiling is the analysis of differential expression (DE) (Oshlack et al., 2010). Differential expression analysis pursues the identification of genes whose transcripts show substantial changes in abundance across experimental conditions, such as differences between strains, tissues, or treated versus untreated individuals. Finally, both qualitative and quantitative information can be combined in studies like allele-specific expression (ASE). In ASE experiments, the goal is to identify genes for which the two alleles in an individual are expressed at different levels. ASE studies, first call for SNP and then identify heterozygous sites. For each of the heterozygous sites the number of reads supporting each of the alleles is counted and a statistical test is performed to evaluate whether the alleles are expressed in balance.

RNA-seq has expanded the field of transcriptomics as no other previous technology (e.g. microarray), not only for model but also for nonmodel organisms in general (Ekblom and Galindo, 2011). Some of the reasons that contribute to this unprecedented expansion in nonmodel organisms, and for fish and wildlife species (Qian et al., 2014) specifically, are the cost-effectiveness and the independence of prior genomic knowledge of the species of interest (Ekblom and Galindo, 2011; Vijay et al., 2013). For instance, Salem et al. (2012) used RNA-seq analysis to identify SNPs markers associated with growth-rate in rainbow trout. The authors compared two full-sib families, one selected for improved growth. Putative SNPs identified by RNA-seq as associated with growth trait were further estimated by genotyping individuals from 40 families using a designed array platform. Smith et al. (2013) studied the ability of crimson spotted rainbowfish to adapt to temperature stress. The authors tested for differential expression across two groups of fish exposed to different conditions using a *de novo* assembled transcriptome. The genes identified as differentially expressed were related to critical metabolic pathways for temperature tolerance and provided candidates to extend investigations of population adaptations to climate change. Babbit et al. (2012) used RNA-seq to compare gene expression patterns in the ovarian tissue of juvenile and adult female baboons and were able to link the differentially expressed genes to selection occurring in human and chimpanzee. This information is valuable to study the evolution of gene regulation in humans, specifically providing insight into the loci that contributed to shifts in developmental timing and physiology during human evolution. All the applications of RNA-seq aimed at making population-wide inferences (e.g: differential expression, allele-specific expression), including the cited ones, use a sample from the population in question, and thus, they rely on statistical analysis to make inferences. Consequently, the validity of the conclusions depend directly on the validity of the statistical methods used in the studies.

Even though a number of statistical analysis methods and tools have been developed for various applications of RNA-seq experiments (Chen et al., 2011), limitations and challenges still remain in determining their statistical validity due to potential variation introduced in each of the steps of the design and execution of an RNA-seq study (Ozsolak and Milos, 2011; Marguerat et al., 2010). A characteristic aspect of RNA-seq is that the measure of expression, total number of reads mapping to a certain region in the reference, is of discrete nature (Anders and Huber, 2010) in contrast to continuous intensity measures obtained in microarray analysis. These count data require statistical models that can account for the discrete nature of the data when testing for differential expression or when summarizing massively parallel multivariate expression records through clustering analysis. A simple Poisson model seems appropriate for this type of data when the experiment includes only technical replicates (Marioni et al., 2008). However, due to extra sources of variation at the biological level, read counts are always overdispersed and other models have been proposed, for example, based on the negative binomial distribution (Robinson et al., 2010; Anders and Huber, 2010; Hardcastle and Kelly, 2010; Di et al., 2011). Other characteristics of RNA-seq experiments such as the small sample size and large number of responses have limited the direct application of generalized linear mixed models developed for count data in other disciplines. Current models based on discrete distributions (i.e. Poisson, negative binomial) are limited to simple experimental designs such as pairwise or multiple fixed factors. Although those designs are extremely useful in most of lab experiments, they cannot account for random effects and hierarchical structures which are commonly present in ecological experiments and studies. Thus, it would be valuable to evaluate the use of transformation in linear mixed models using realistic datasets. The evaluation should include properties of interest when inferring results such as the rate of false positives and statistical power. Specifically, some authors (Langmead et al., 2010; Law et al., 2014) have proposed the use of data transformation to fit simpler Gaussian linear model analysis, which allow a straightforward implementation of multilevel models. Thus, there are multiple recommended ways of analyzing RNA-seq for differential expression and clustering whose statistical properties need to be evaluated.

Hierarchical cluster analysis has been the preferred statistical method to represent results of samples and genes after differential expression analysis (Liu and Si, 2014). Hierarchical clustering is also a good unsupervised technique to discover subpopulation structure with respect to a set of multivariate responses (Legendre and Legendre, 2012). Thus, hierarchical clustering of samples using gene expression data from RNA-seq experiments holds great promise for the study of natural variation in relation to gene expression in ecological experiments. Hierarchical clustering analysis relies on a measure of similarity/dissimilarity among the features (i.e. genes, transcripts or samples) to build a dendrogram. But dissimilarity measures may be greatly affected by the distribution of the measured variables. Dissimilarity measures for RNA-seq derived gene expression count data are mainly based on logarithmic transformation of raw counts. Additionally, a specific Poisson dissimilarity measure has been proposed for RNA-seq data (Witten, 2011). However, only parametric simulations and exemplar data have been used to compare the performance of these measures for sample clustering. Due to the high dispersion of count data derived from RNA-seq, a comparison based on experimental datasets and using a good measure of consistency of the resulting dendrogram representation could contribute to decide on the best dissimilarity measures for hierarchical clustering implementations. Another relevant question in this area is: are transformations used for differential expression analysis also appropriate for clustering analysis? Finding a dissimilarity measure that can reliably represent the sample structure before fitting any model could be of interest to conduct *a priori* exploratory data analysis to be followed up by inferential analyses.

Finally, another inferential problem that deserves study is the use of algorithms for calling SNP genotyping from RNA-seq data. SNP genotyping from RNA-seq is an application of RNA-seq experiments that has proven to be useful in some studies and we can expect that its use will increase in the future (Wickramasinghe et al., 2014; Seeb et al., 2011; Schunter et al., 2013; Narum et al., 2013). Currently, some questions remain about the equivalence of using this technology versus more traditional genotyping chips and DNA-seq. For instance, the effect of per base read depth when calling and genotyping variants should be studied in terms of its impact in sensitivity and specificity for calling SNP from RNA-seq data. Moreover, the statistical properties of called genotypes should be studied separately for heterozygous and homozygous genotypes because they are likely to be used for different purposes. Another important property of genotypes derived from RNA-seq data is how well they can represent the genetic structure of a population, when compared to more traditional measures of relatedness based on SNP chip genotypes and on genealogies.

In order to compare and decide among the available statistical methods in each of the mentioned applications of RNA-seq (e.g. differential expression analysis, clustering, SNP genotyping) a validation framework should be established for each case. Consequently, in this dissertation I followed the epistemological guidelines proposed by Mehta et al. (2004) for high-dimensional biology. Assessing statistical validity requires an explanation of what a method is supposed to do or what properties an estimate or test is supposed to exhibit. Although there is subjectivity in selecting the desired properties, once the criteria are chosen, methods can be evaluated objectively (Mehta et al., 2004). For instance, a desired property to compare among differential expression models is that the type I error rate should match the test's nominal error rate. Once we choose this property, we can compare the models by applying a validation, such as simulating data or other validating procedures.

Validating procedures for RNA-seq analysis methods have followed the same strategies that have been previously implemented in technologies such as microarray. Basically, the most used validating procedures can be grouped into the following categories: a) theoretical demonstration, b) exemplar dataset, c) simulation, d) gold-standard reference, and e) plasmode datasets. Theoretical derivations in the context of high-dimensional biology often lacks of a mathematical demonstration to support their validity (Mehta et al., 2006). The use of an exemplar dataset offer the advantage of accounting for the whole complexity of the transcriptome such as having a variance-covariance matrix that reflect real interactions among features. However, this approach must be considered only as an illustration in a particular dataset and not as evidence to support a method. Computer simulations have been the most commonly used procedure (Anders and Huber, 2010; McCarthy et al., 2012) due to ease in creating datasets under hypothetical scenarios by assuming a parametric data generation model. However, although gene expression data is easy to simulate, mimicking a realistic gene expression dataset is quite challenging. Gold-standard references obtained with a different technology have been used to validate results (Fang and Cui, 2011). Typical references for gene expression results from RNA-seq are qPCR data (Bullard et al., 2010; Rapaport et al., 2013). However, analysis models for qPCR data should themselves be validated (Steibel et al., 2009). Another example of gold standard is the use of mask-and-impute for estimating imputation accuracy (Badke et al., 2013; Gualdrón Duarte et al., 2013) and the use of genotyping chips to validate genotype calls from genome sequencing (Kumar et al., 2014). Finally, the use of plasmodes is another appropriate procedure that can be applied to validate a statistical method. This approach aims at generating datasets that preserve the characteristics of experimental data with the benefit of knowing the true status as it happens with simulated data.

Plasmodes are synthetic datasets generated from experimental data (Mehta et al., 2004). The term plasmode was introduced in 1967 by Cattell et al. (1967) as a tool for validating techniques in multivariate data. Unlike parametric simulations, data distributions and correlations generated in plasmodes can be more realistic because they are taken directly from experimental data thus no assumptions are required *a priori* to create the datasets. Plasmodes have to be generated according to the validation objectives and the available experimental data (Mehta et al., 2006) and one challenge is precisely how to create a plasmode for differential expression analysis versus a plasmode to validate results for sample- or even gene- based clustering.

In this dissertation, I show how to validate statistical analysis methods for RNA-seq data by creating plasmode datasets. I also use parametric simulations and a gold-standard reference to complement the validation analysis. Plasmodes were created in three different ways. First, I created plasmodes by applying resampling-based methods consistent with the null hypothesis (no differential expression). Second, I created differentially expressed plasmodes with a known proportion of differentially expressed genes. These two methods of creating plasmodes are particularly useful to evaluate statistical properties of differential expression analysis models and, for evaluating dissimilarity metrics for clustering. Third, I generated synthetic individuals by combining known proportions of transcript-specific read counts from two paternal individuals, which is more relevant for comparing dissimilarity metrics for hierarchical clustering. Finally, I used a gold-standard reference comparison to assess properties of variant calling methods.

Given the rapid advances in analysis tools designed for RNA-seq experiments, a better understanding of the properties of results obtained by commonly used statistical analysis methods would provide valuable information for researchers who either develop analysis models and tools or utilize analysis tools in their investigations. A systematic comparison through a valid framework, as provided by plasmodes, could supply reference datasets to be used as benchmarks for comparing analytic approaches. Furthermore, researchers interested in specific experimental data could use plasmodes generated by similar experimental conditions to conduct pilot *in silico* studies and to choose the most suitable analysis procedure. For instance, a researcher may evaluate whether to use a generalized linear model for analyzing the read counts of a differential expression experiment or to apply an appropriate transformation in order to use a more flexible general linear mixed model. As the range of objectives of RNA-seq studies expands (i.e. focusing not only on differential expression analysis) it will be of interest to contribute with analogous validation framework in such specific applications. In accordance with this demand, we provide metrics to evaluate dissimilarity matrices to represent sample relationships using hierarchical cluster analysis, and metrics to compare variant calling software. The evaluation of dissimilarity measures are of direct benefit to summarize and represent biological/technical variability in experimental design, or relationships among individuals in population studies. The comparison of variant calling software provides information on genotyping accuracy that could improve other related analysis such as allele-specific expression or RNA editing.

Thus, the overarching goal of this dissertation was to propose and implement validation procedures based on experimental data to estimate the properties of widely used statistical analysis methods of RNA-seq experiments.

The specific objectives were:

1. To evaluate statistical models for differential expression analysis in RNA-seq experiments
2. To assess the properties of dissimilarity measures for agglomerative hierarchical clustering of RNA-seq data
3. To compare sensitivity and false discovery rates for SNP genotyping in variant calling programs

Chapter 2

Evaluating statistical analysis models for RNA sequencing experiments

Reeb, P. D., and Steibel, J. P. (2013). Evaluating statistical analysis models for RNA sequencing experiments. *Front. Genet.* 4, 1-9. doi:10.3389/fgene.2013.00178.

2.1 Introduction

RNA sequencing (RNA-seq) technology is being rapidly adopted as the platform of choice for high-throughput gene expression analysis (Ozsolak and Milos, 2011). Many methods have been proposed to model relative transcript abundances obtained in RNA-seq experiments but it is still difficult to evaluate whether they provide accurate estimations and inferences.

Sound statistical analysis of RNA-seq data should consider not only the factors of any basic experimental design, but also the characteristics of "omic" studies (genomic, proteomic, transcriptomic, etc). An RNA-seq experimental design must consider treatment and block structures, and combine them according to the principles of a well-planned design: randomization, blocking and replication (Auer and Doerge, 2010). Typically, fixed or random effects such as library multiplexing, sequencing lane, flow cell, individual sample, tissue, or time can be crossed or nested with treatments or other experimental conditions. Such a design is used to model thousands of correlated variables (i.e transcripts), usually, in a context of small number of biological replicates. Although the development of reliable models that account for all these factors is challenging, it is even more difficult to assess the validity of a particular analysis model (Pachter, 2011).

Validity of statistical models for differential expression analyses has been evaluated by (i) applying the model to a novel dataset, (ii) deriving analytical proofs, (iii) using simulations, (iv) comparing to a gold-standard measure, or (v) constructing plasmodes. In (i) the true status of nature is unknown, therefore this method must only be accepted as an illustration and not as evidence to support a model. However, any of the last four options, or a combination of them, could be used to demonstrate adequacy of a model. Obtaining a mathematical demonstration (ii), may be impossible for some models (Gadbury et al., 2008). Most of the models rely on assumptions that are difficult to verify and the consequences of departures from assumptions may not be clear. Computer simulation (iii) has been the most commonly used procedure (Anders and Huber, 2010; McCarthy et al., 2012) . This preference is due to ease in creating datasets under diverse scenarios by controlling the set of parameters used in the simulation. Nevertheless, such generated data depend on the parameterization selected and the assumptions of the simulation model. Moreover, these dataset may constitute a partial representation of reality as the complexity of RNA-seq data is hard to mimic. Typical gold-standard (iv) for gene expression are qPCR data (Bullard et al., 2010; Rapaport et al., 2013). However, analysis models for qPCR data should themselves be validated (Steibel et al., 2009). The use of plasmodes (v) is another appropriate procedure that can be applied to validate a statistical method. This approach aims at generating datasets that preserve the characteristics of experimental data with the benefit of knowing the true status as it happens with simulated data.

A plasmode is a dataset obtained from experimental data but for which some truth is known (Mehta et al., 2004). Plasmodes have been applied in microarrays (Gadbury et al., 2008), admixture estimation methodologies (Vaughan et al., 2009) and qPCR (Steibel et al., 2009). This procedure has not been extensively applied in RNA-seq since it requires large sets of raw data with an accurate description of the experimental conditions under which they were obtained. This information is essential to accurately develop plasmodes under null and alternative hypotheses. Only recently, an initiative has provided a repository with ready-to-use databases from RNA-seq studies (Frazee et al., 2011).

Processed data obtained from RNA-seq experiments are essentially counts that in the simplest model represent total number of reads mapping to a region in a reference genome or transcriptome. A comprehensive comparison of stochastic models that have been proposed is presented in Pachter (2011). Although different discrete distributions such as binomial, multinomial, beta-binomial, Poisson, and negative binomial, have been proposed to model RNA-seq data, Poisson and negative binomial are the most implemented ones in RNA-seq analysis software. A simple Poisson model seems appropriate when the experiment includes only technical replicates from a single source of RNA (Marioni et al., 2008). In practice, however, due to extra sources of variation, the observed dispersion is larger than the expected for a simple Poisson distribution and to correctly account for over-dispersion, generalized Poisson (GPseq) (Srivastava and Chen, 2010), mixed Poisson (TSPM) (Auer and Doerge, 2011), Poisson log-linear (PoissonSeq) (Li et al., 2012) and negative binomial (edgeR, DESeq, baySeq, NBPSeq) (Robinson et al., 2010; Anders and Huber, 2010; Hardcastle and Kelly, 2010; Di et al., 2011) are used instead. Regardless of the model, calculating dispersion parameters requires special statistical and numerical approaches due to the small sample sizes and large number of responses used in RNA-seq studies. In particular, borrowing information across transcripts when estimating model parameters, as used in microarrays (Smyth, 2004; Cui et al., 2005), has been also proposed for RNA-seq (Robinson and Smyth, 2008; Anders and Huber, 2010; Zhou et al., 2011). Another challenging issue for these statistical analysis models, is the ability to handle different experimental sources of variation. Most of the models allow fitting simple effect models and pair-wise comparison between treatments but only a few allow multiple factors (McCarthy et al., 2012). Currently, to the best of our knowledge, there is only one available model that can fit random effects (Van De Wiel et al., 2013). Methods that can accommodate complex hierarchical designs and provide more powerful tests to detect differentially expressed transcripts are under active research. On the other hand, microarray analysis models and software usually assume a Gaussian distribution for response variables, but they accommodate fixed and random effects in a straightforward manner (Rosa et al., 2005; Cui et al., 2005). Consequently, an alternative to model counts in RNA-seq experiments is to transform counts and use Gaussian models (Langmead et al., 2010; Smyth et al., 2012).

In any case, given the multitude of available statistical models and the complexity of experimental design of many gene expression studies, researchers often find themselves having to decide between competing models and analysis program. In other cases, although a researcher may have an a priori designated software and model for RNA-seq data analysis, the question is if the fitted model produces sound inferences.

In this paper, we present and apply a methodology for evaluating statistical methods for RNA-seq experiments by combining results from computer simulations and plasmodes. We follow the epistemological guidelines stated in Mehta et al. (2006) for high-dimensional biology and provide a general framework that can be adapted to different experimental conditions.

2.2 Material and Methods

2.2.1 Simulations

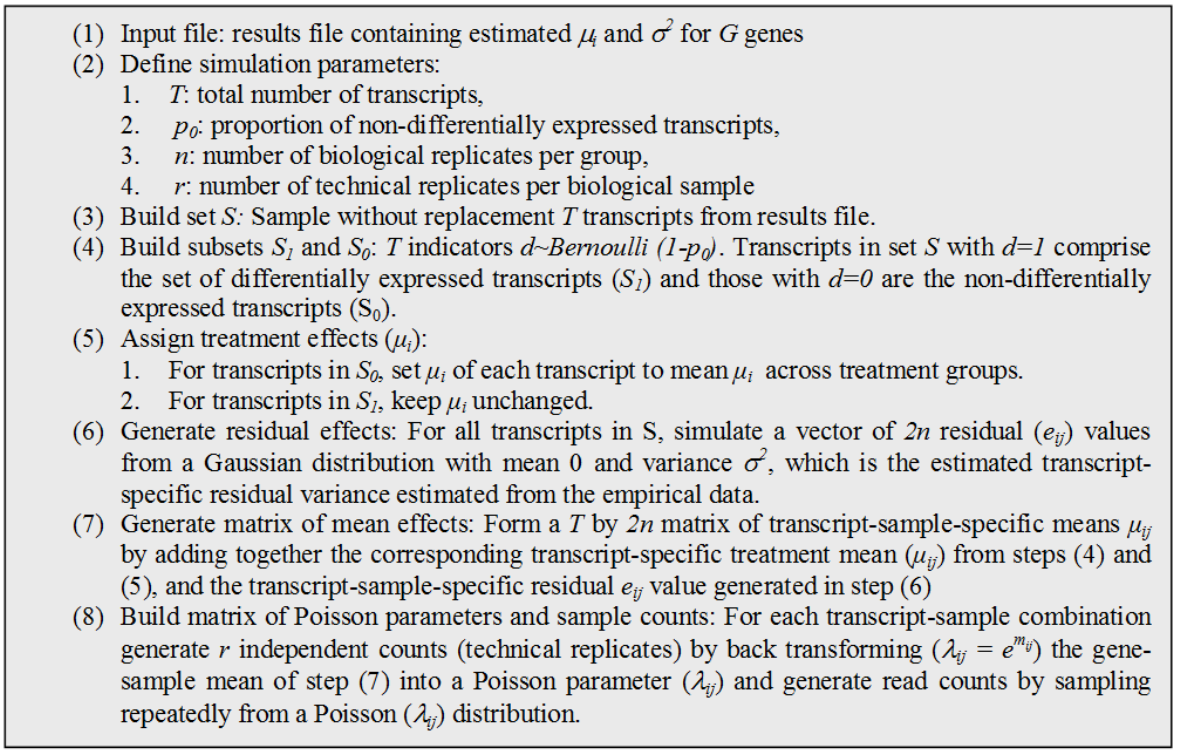
Simulated datasets were created conditional on estimated parameter values and results that had been previously obtained (Ernst et al., 2011). The data consisted of read counts from an RNA-seq experiment based on a developmental expression study (Sollero et al., 2011). Experimental and alignment protocols are described in the supplemental material (Figure 9). Estimations for parameters μ_i and σ^2 were obtained by fitting generalized linear Poisson models with log-library size as an offset variable using function lmer (Bates et al., 2013) from R (R Development Core Team, 2014).

Equation [1] represents the generalized linear model used to generate the simulated datasets:

$$\begin{cases} y_{ij} \sim Poisson(\lambda_{ij}) \\ log(\lambda_{ij}) = O_{ij} + \ \mu_i + e_{ij} \\ e_{ij} \sim N(0, \sigma^2) \end{cases} \qquad [1]$$

where y_{ij} is the read count for a particular transcript in treatment *i* and sample *j*, O_{ij} is a known off-set value (in this case the total library size), μ_i is the group mean, e_{ij} is a sample-specific residual. The transcript sub-index (g) was omitted for convenience.

Given estimates of parameters from equation [1] for transcripts, we simulated read counts by following the algorithm described in Figure 1. The output from this procedure consisted of a matrix of counts of size *T* by *2nr* with a known proportion (p_0) of differentially expressed transcripts and known group effects ($\boldsymbol{\mu_i}$). Treatment is represented in this matrix by *nr* columns, but with only *n* independent (biological) replicates. While this simulation is not based on the negative binomial distribution, it continues to be an over-dispersed Poisson process commonly used to simulate RNA-seq counts (Blekhman et al., 2010; Auer and Doerge, 2011; Hu et al., 2011). The resulting over-dispersed Poisson counts will have means, variances, and treatment effects sampled from those estimated from experimental data. The procedure can be repeated K times to produce several simulated datasets.



(1) Input file: results file containing estimated μ_i and σ^2 for G genes

(2) Define simulation parameters:

  1. T: total number of transcripts,
  2. p_0: proportion of non-differentially expressed transcripts,
  3. n: number of biological replicates per group,
  4. r: number of technical replicates per biological sample

(3) Build set S: Sample without replacement T transcripts from results file.

(4) Build subsets S_1 and S_0: T indicators d~$Bernoulli\ (1\text{-}p_0)$. Transcripts in set S with $d=1$ comprise the set of differentially expressed transcripts (S_1) and those with $d=0$ are the non-differentially expressed transcripts (S_0).

(5) Assign treatment effects (μ_i):

  1. For transcripts in S_0, set μ_i of each transcript to mean μ_i across treatment groups.
  2. For transcripts in S_1, keep μ_i unchanged.

(6) Generate residual effects: For all transcripts in S, simulate a vector of $2n$ residual (e_{ij}) values from a Gaussian distribution with mean 0 and variance σ^2, which is the estimated transcript-specific residual variance estimated from the empirical data.

(7) Generate matrix of mean effects: Form a T by $2n$ matrix of transcript-sample-specific means μ_{ij} by adding together the corresponding transcript-specific treatment mean (μ_{ij}) from steps (4) and (5), and the transcript-sample-specific residual e_{ij} value generated in step (6)

(8) Build matrix of Poisson parameters and sample counts: For each transcript-sample combination generate r independent counts (technical replicates) by back transforming ($\lambda_{ij} = e^{m_{ij}}$) the gene-sample mean of step (7) into a Poisson parameter (λ_{ij}) and generate read counts by sampling repeatedly from a Poisson (λ_{ij}) distribution.

Figure 1 Algorithm used to simulate counts from existing estimates of model parameters

We set *K=1000* and *T=5000*, producing 1000 simulated datasets with 5000 transcripts each. Noteworthy, when sampling transcripts in S, it is assumed that all transcripts are differentially expressed (no significance testing is performed). But subsequently, the mean treatment differences (in the log-scale) are zeroed out if the transcripts are assigned to S_0. For transcripts assigned to S_1, mean differences are kept unchanged; consequently S_1 includes a whole distribution of treatment effects from very small to large according to the distribution estimated from the experimental data.

We simulated nine scenarios by combining three levels of biological replication (n=3, 5, 10) and three levels of technical replication (*r*=1, 3, 5). The proportion of differentially expressed transcripts was set to 0.1.

2.2.2 Plasmodes

In contrast to simulation datasets based on equation [1], we generated plasmode datasets not based on any model. Plasmodes were generated using data available in the online resource ReCount (Frazee et al., 2011). From the whole collection of analysis-ready datasets, we chose to work with two RNA-seq experiments to illustrate the generation of 1) a null dataset, where there are no obvious systematic effects that explain variance in gene expression and, 2) a dataset with treatment and block effects.

2.2.2.1 Null dataset (Cheung)

The data originated in a study of immortalized B-cells from 41 (17 females and 24 males) unrelated CEPH (Centre d' Etudes du Polimorphisme Humain) grandparents (Cheung et al., 2010). The samples were sequenced using the Illumina Genome Analyzer. To generate a plasmode dataset, we selected the 21 samples from male individuals that were represented with only one technical replicate. The resulting gene expression data exhibits extensive variation that cannot be attributed to any systematic factor (Figure 2a). Any random partition of the dataset into two (or more) categories should shield a null dataset where no differential expression is expected beyond the normal sample-to-sample variation. Consequently this dataset lends itself to create plasmodes to evaluate statistical properties of analysis models under the null hypothesis.

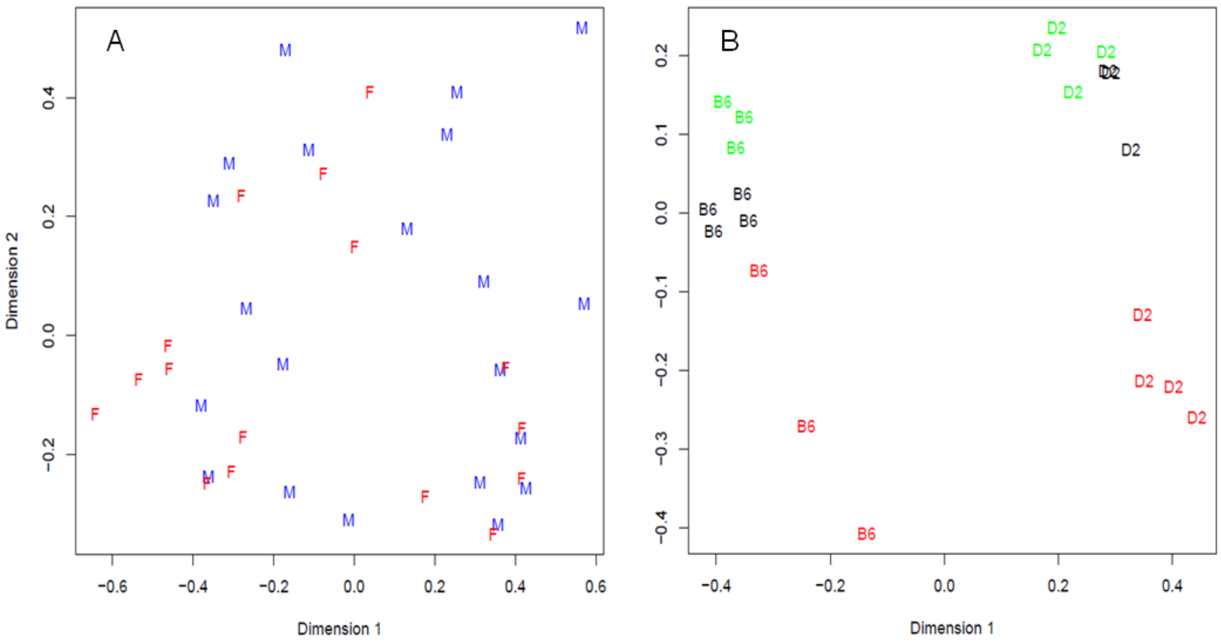
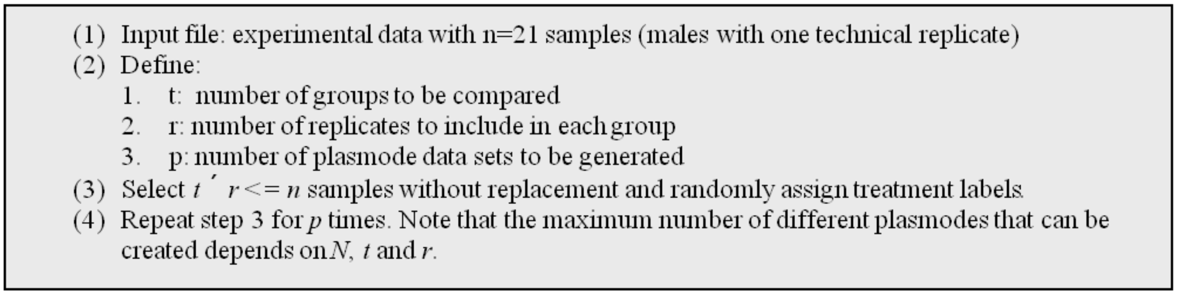


Figure 2 Multidimensional scaling analysis of experimental datasets

(A) Cheung samples: F=Females and M=males; **(B)** Bottomly samples: labels correspond to strain (treatment) B6=C57BL/6J, D2=DBA/2J, and colors to flowcell number (block): red=4, black=6 and green=7. In Cheung dataset there is not clear distinction between females and males while in Bottomly samples are first grouped in two large groups corresponding to strain B6 and D2 and then in subgroups consistent with flowcell number

To generate null datasets, we proceeded as explained in Figure 3. Using *n=21* samples from males, we generated *p=10* plasmodes each with *t=2* groups and *r=10* biological replicates in each group.

Notice that no parametric model is used at any time. We constructed plasmodes by reshuffling data and assigning an arbitrary treatment label. In this way overall distribution and gene-to-gene correlations remain unchanged with respect to real data.



(1) Input file: experimental data with n=21 samples (males with one technical replicate)
(2) Define:
  1. t: number of groups to be compared
  2. r: number of replicates to include in each group
  3. p: number of plasmode data sets to be generated
(3) Select *t* ´ *r* <= *n* samples without replacement and randomly assign treatment labels.
(4) Repeat step 3 for *p* times. Note that the maximum number of different plasmodes that can be created depends on *N*, *t* and *r*.

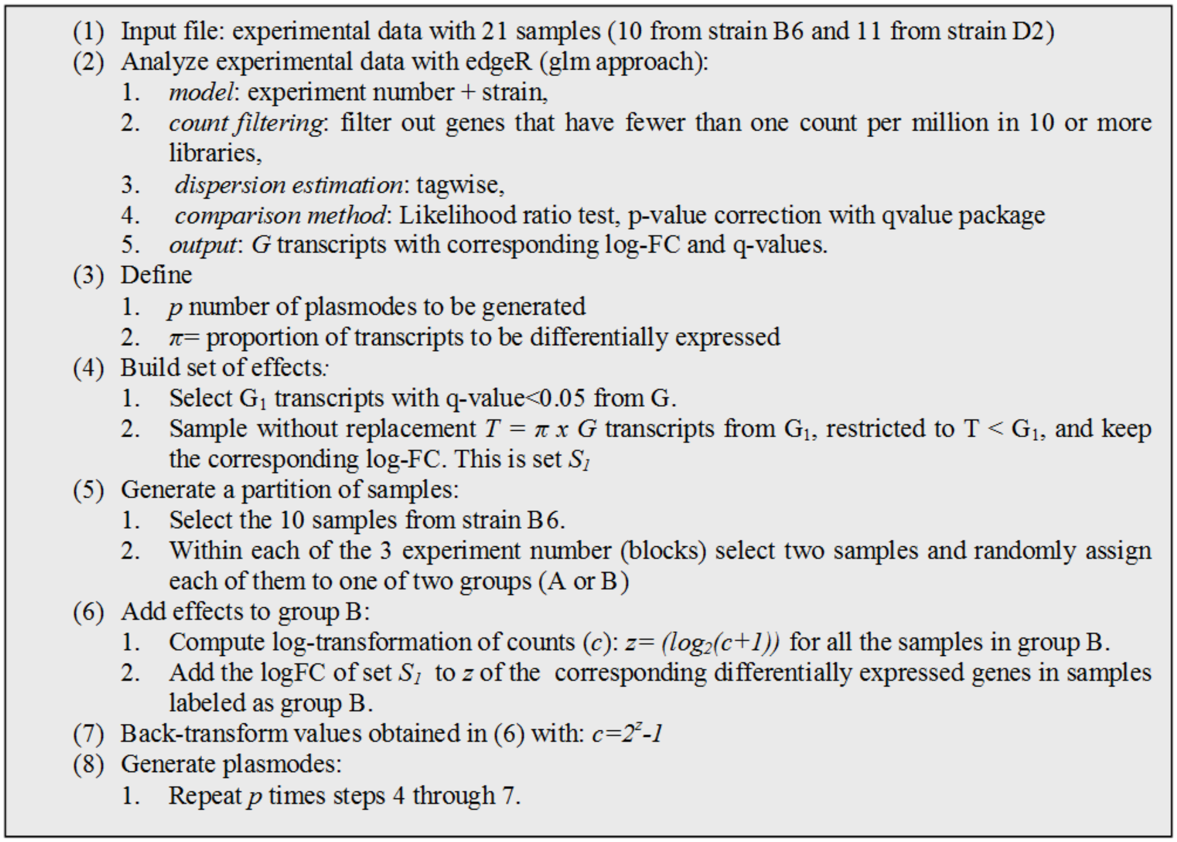
Figure 3 Algorithm used to generate plasmode datasets with no differentially expressed transcripts under a model with one classification variable

2.2.2.2 Differentially expressed dataset (Bottomly)

In Bottomly et al. (2011), the authors arranged 21 samples from two inbred mouse strains (B6 and D2; n for B6=10, n for D2=11) on 21 lanes of three Illumina GAIIx flowcells and they analyzed the RNA-seq reads with a simple one-way classification (strain) model. After performing descriptive analysis of gene expression data, we found that not only strain but also the experiment number (flowcell) explained a large amount of the variation (Figure 2b). For example, the first principal dimension clearly divides samples from each strain, but the second principal dimension shows substantial variation between flowcells, especially flowcell 4 (red) versus the other two.

Consequently, we blocked by experiment and used edgeR to fit a model with strain and experiment as fixed effect, resulting in a large number of putatively differentially expressed genes (Figure 10). Due to a strong experiment effect, we decided to conduct randomization for plasmode construction within experiment number as detailed in Figure 4.

We generated 10 plasmodes executing step 4 to 7 with p=10 and π=0.20. Notice that in step 3, we used edgeR to obtain a list of DE genes (set G) to build a plasmode with some genes under alternative hypothesis. Any other statistical software, however, can be used with the only requirement of defining a sufficient small q-value threshold. After genes are selected no model is used at any time. Similar to the previous section plasmodes are constructed by reshuffling data, but in this case an effect estimated from real data is added to selected genes. Again, we expect that this procedure yields plasmodes with identical distribution to real data for non differentially expressed genes and with comparable effect sizes for differentially expressed genes.



```
(1) Input file: experimental data with 21 samples (10 from strain B6 and 11 from strain D2)
(2) Analyze experimental data with edgeR (glm approach):
    1. model: experiment number + strain,
    2. count filtering: filter out genes that have fewer than one count per million in 10 or more
       libraries,
    3. dispersion estimation: tagwise,
    4. comparison method: Likelihood ratio test, p-value correction with qvalue package
    5. output: $G$ transcripts with corresponding log-FC and q-values.
(3) Define
    1. $p$ number of plasmodes to be generated
    2. $\pi$= proportion of transcripts to be differentially expressed
(4) Build set of effects:
    1. Select $G_1$ transcripts with q-value<0.05 from G.
    2. Sample without replacement $T = \pi x G$ transcripts from $G_1$, restricted to $T < G_1$, and keep
       the corresponding log-FC. This is set $S_1$
(5) Generate a partition of samples:
    1. Select the 10 samples from strain B6.
    2. Within each of the 3 experiment number (blocks) select two samples and randomly assign
       each of them to one of two groups (A or B)
(6) Add effects to group B:
    1. Compute log-transformation of counts ($c$): $z= (log_2(c+1))$ for all the samples in group B.
    2. Add the logFC of set $S_1$ to $z$ of the corresponding differentially expressed genes in samples
       labeled as group B.
(7) Back-transform values obtained in (6) with: $c=2^z-1$
(8) Generate plasmodes:
    1. Repeat $p$ times steps 4 through 7.
```

Figure 4 Algorithm used to generate plasmode datasets with differentially expressed transcripts under a model with two classification variables (block + treatment)

2.2.3 Comparison of alternative analysis tools for evaluating differential expression

To illustrate the use of simulated datasets and plasmodes we compared three R (R Development Core Team, 2014) packages from Bioconductor (Gentleman et al., 2004). Two of them, edgeR and DESeq, were designed specifically for statistical analyses of RNA-seq experiments while the third one, MAANOVA (Cui et al., 2005), was originally conceived for analyzing microarray experiments. As mentioned before, MAANOVA has the ability of fitting hierarchical models that can better accommodate complex experimental design assumptions. However, such flexibility comes at the price of assuming a Gaussian distribution. Data transformation and use of permutation to set significance thresholds can help alleviate these limitations, but its performance may still be contingent upon sample size and total read counts per transcript. Consequently, we included MAANOVA in this study and compare it to two well established packages for RNA-seq analysis.

2.2.3.1 Filtering and normalization

A double filtering criterion was applied to all datasets previous to normalization and statistical analysis. Transcripts with 2 or more reads per million in at least as many libraries as number or biological replicates were kept in the analysis. In the simulation study, technical replicates were summed up before filtering. Consequently, the technical replicate level only represents increased sequencing depth.

Normalization aimed at accounting for differences in library size and composition not attributable to treatments. To conduct the analysis with edgeR, data were normalized using the scaling method proposed by Robinson and Oshlack (2010) and the logarithm of the resulting effective library size were used by default as offsets in the model.

Analyses with DESeq were performed on counts previously normalized by function estimateSizeFactors. According to Anders and Huber (2010), this normalization method is similar to the one proposed by Robinson and Oshlack (Robinson and Oshlack, 2010) in edgeR, and it is the recommended procedure by the authors of DESeq.

Normalized values to use in MAANOVA were obtained with function voom() of the limma package (Smyth, 2005). The process, analogous to the one proposed in (Smyth et al., 2012), included adjustment for compositional structure using function calcNormFactors() of edgeR and transformation to log2-counts per million.

2.2.3.2 Differential expression analysis

edgeR: Differential expression was tested by likelihood ratio tests using the generalized linear model functionality and estimating tagwise dispersions.

DESeq:To look for differentially expressed genes, function nbinomGLMTest was applied using the dispersion estimates generated by function estimateDispersions.

MAANOVA: In the linear model fit by MAANOVA lane was treated as a fixed array effect of a single-color microarray. Differential expression analysis was performed using both, moderated F-test (Fs) and transcript by transcript F-test (F1). Significance was assessed using 100 sample permutations (Yang and Churchill, 2007).

2.2.3.3 Multiple comparisons

It is recognized that correction of p-values when making multiple comparisons is essential in high throughput differential expression analyses (Storey and Tibshirani, 2003). The most common procedure used is the computation of the false discovery rate or FDR (Benjamini and Hochberg, 1995). Properties of methods to estimate FDR rely heavily on the distribution of p-values (Li et al., 2012). In this case we did not aim at selecting individual differentially expressed genes or gene sets but we focused at studying the properties of tests in terms of type I and type II error rates. Consequently, we concentrate on comparison of nominal and empirical type I and type II error rates without applying multiple correction and we discuss how departures of assumed values can further affect decisions when applying p-value corrections.

2.2.4 Evaluating and comparing results from alternative analysis packages

To compare performances of derived tests in terms of power and type I error rates, we generated receiver operator characteristic (ROC) curves by computing true positive rate (TPR) and false positive rate (FPR) at given significance thresholds. The TPR was calculated as the proportion of true positives (TP) over the total number of simulated differentially expressed transcripts (S_1). FPR, on the other hand, was calculated as the proportion of false positives (FP) over the total number of transcripts simulated with no differential expression (S_0). See table 1 for details.

Table 1 Classification rule to compute false and true positive rates

Analysis method status	Transcript simulation status		Total
	Not differentially expressed	Differentially expressed	
not declared significant	TN	FN	R_0
declared significant	FP	TP	R_1
Total	$\#S_0$	$\#S_1$	G

FP=number of false positives (transcripts in S_0 set declared differentially expressed), TP= number of true positives (transcripts in S_1 declared differentially expressed), FPR= false positive rate= FP/$\#S_0$, TPR=true positive rate=TP/$\#S_1$

Finally, distributions of p-values were compared by quantile-to-quantile plots and histograms.

Analyses were performed at the Michigan State University High Performance Computing Center facilities using R (version 2.15.1), edgeR (version 3.0.8.4.6), limma (version 3.14.4), DESeq (version 1.10.1) and MAANOVA (version 1.28.0).

2.3 Results

2.3.1 Simulations

Figure 5 shows results obtained for a simulation with 3 biological replicates and 1 technical replicate. Similar results were found in other simulated scenarios (data not shown).

The quantile-to-quantile plot in Figure 5 allows evaluation of the fit of observed p-values to the uniform (0,1) distribution expected under null hypothesis (Leek and Storey 2011). P-values corresponding to MAANOVA showed a more characteristic pattern whereas edgeR and DESeq presented significant departures from such distribution. Furthermore, the logarithmic scale allows to easily inspect the behavior of very small p-values. DESeq presented larger p-values than expected up to a cutoff of 0.001, while the opposite pattern occur for p-values smaller than 0.001. Both MAANOVA approaches presented a close to expected pattern with a small deviation for p-values smaller than 0.0001. To compute the logarithm, all p-values equal to zero were replaced by the minimum observed p-value and thus generated the plateau at the end of the distributions of MAANOVA results. In addition, quantile-to-quantile plots allowed us to select Fs and F1 tests computed with permutation against the tabulated approach (Figure 8a-b). An alternative representation of p-value distribution using histograms is presented in the supplemental material (Figure 11).

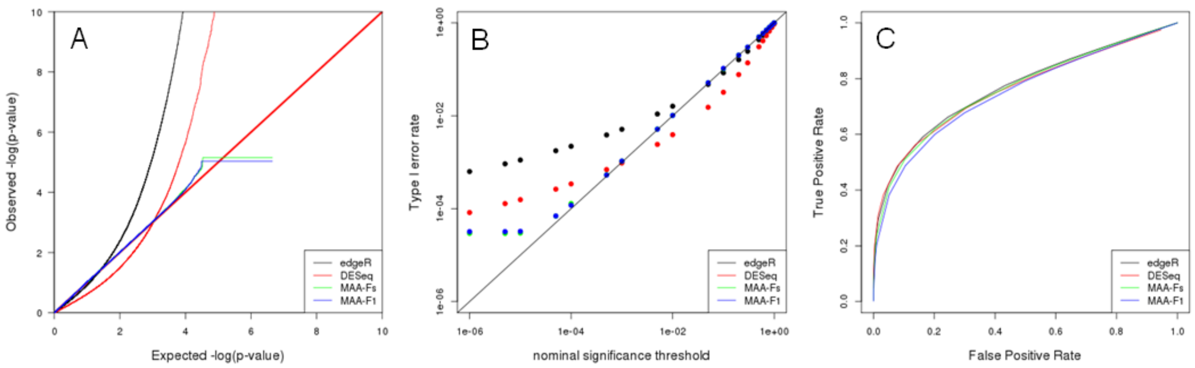


Figure 5 Simulation results from a scenario with three biological replicates

(A) Q-Q uniform plot of non differentially expressed transcripts, **(B)** type I error rate vs. nominal significance values, and **(C)** ROC curves. Models: i) edgeR (blue), ii) DESeq (red), iii) MAA-Fs: MAANOVA Fs moderated test using permutation (green), and iv) MAA-F1: MAANOVA F1 transcript by transcript test using permutation (blue)

In concordance with the observed p-value distributions, the realized type I error rates levels for DESeq and edgeR were much different than expected in comparison with MAANOVA approaches (Figure 5b). All the packages presented higher realized significance levels when evaluated at nominal values below 0.01, with edgeR being the most liberal, and MAANOVA the least deviated from nominal values.

ROC curves had similar patterns for each of the nine simulated scenarios. Power improved at a given FPR as the number of technical and/or biological replicates increased. In the scenario with 3 biological replicates, the enhancement in power when adding technical replicates seems to be particularly greater than in a scenario with 5 or 10 biological replicates (data not shown). In the case with 3 biological replicates and 1 technical replicate (Figure 5c), edgeR and DEseq had similar power while the MAANOVA analyses reported less power.

2.3.2 Plasmodes

2.3.2.1 Null dataset (Cheung)

Q-Q plot in Figure 6a shows the adequacy of p-values to the uniform distribution for each of the plasmode datasets analyzed with the different models. All the models presented large dispersions with some cases being close to the expected values and some being far apart. In particular, edgeR results tend to be above the identity line which means that observed p-values are smaller than expected. On the contrary, DESeq and both MAANOVA tests tend to have a more conservative behavior as they presented larger observed p-values than expected. See also Figure 6b where edgeR presented inflated type I error rates for nominal significance threshold smaller than 0.01.

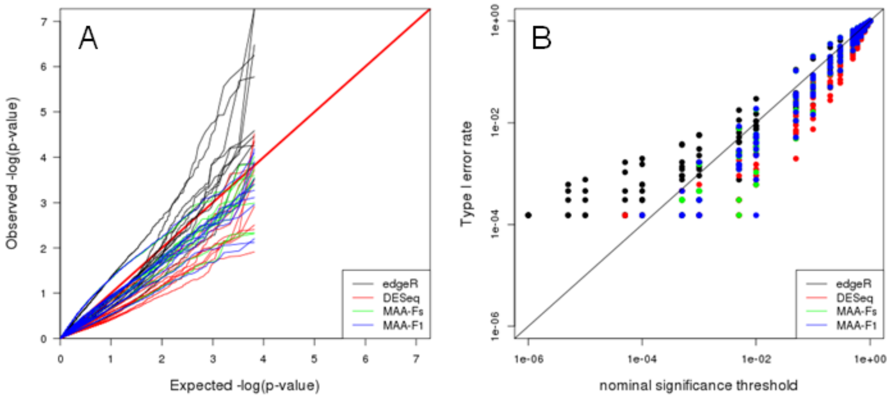


Figure 6 Plasmode results from Cheung dataset:

(A) Q-Q uniform plot of non differentially expressed transcripts, **(B)** type I error rate vs. nominal significance values. Models: i) edgeR (blue), ii) DESeq (red), iii) MAA-Fs: MAANOVA Fs moderated test using permutation (green), and iv) MAA-F1: MAANOVA F1 transcript by transcript test using permutation (blue)

2.3.2.2 DE dataset (Bottomly)

The p-value distributions (Figure 7a) presented similar dispersion patterns to the one observed in the plasmodes generated from Cheung dataset utilizing edgeR and DESeq. However, p-value distributions for MAANOVA tests were more homogeneous across datasets with the p-values from F1 test tabulated approach being closer to the expected values under uniform distribution. ROC curves for DESeq and edgeR were analogous after adjusting for type I error rates. Besides, both programs reported higher power than analysis performed with MAANOVA (Figure 7c). Interestingly, and opposite to previous datasets, the best F test to apply when using MAANOVA was F1 with tabulated F values-compare the proximity to the red line in Figure 8e in contrast to the pattern in Figure 8f.

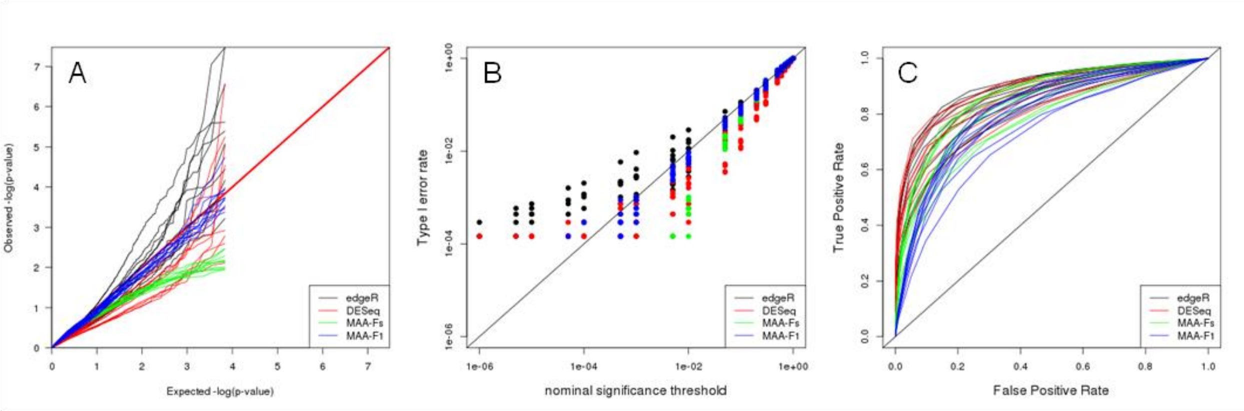


Figure 7 Plasmode results from Bottomly dataset:

Q-Q uniform plot of non differentially expressed transcripts, **(B)** type I error rate vs. nominal significance values, and **(C)** ROC curves. Models: i) edgeR (blue), ii) DESeq (red), iii) MAA-Fs: MAANOVA Fs moderated test with permutation (green), and iv) MAA-F1: MAANOVA F1 transcript by transcript test tabulated (blue)

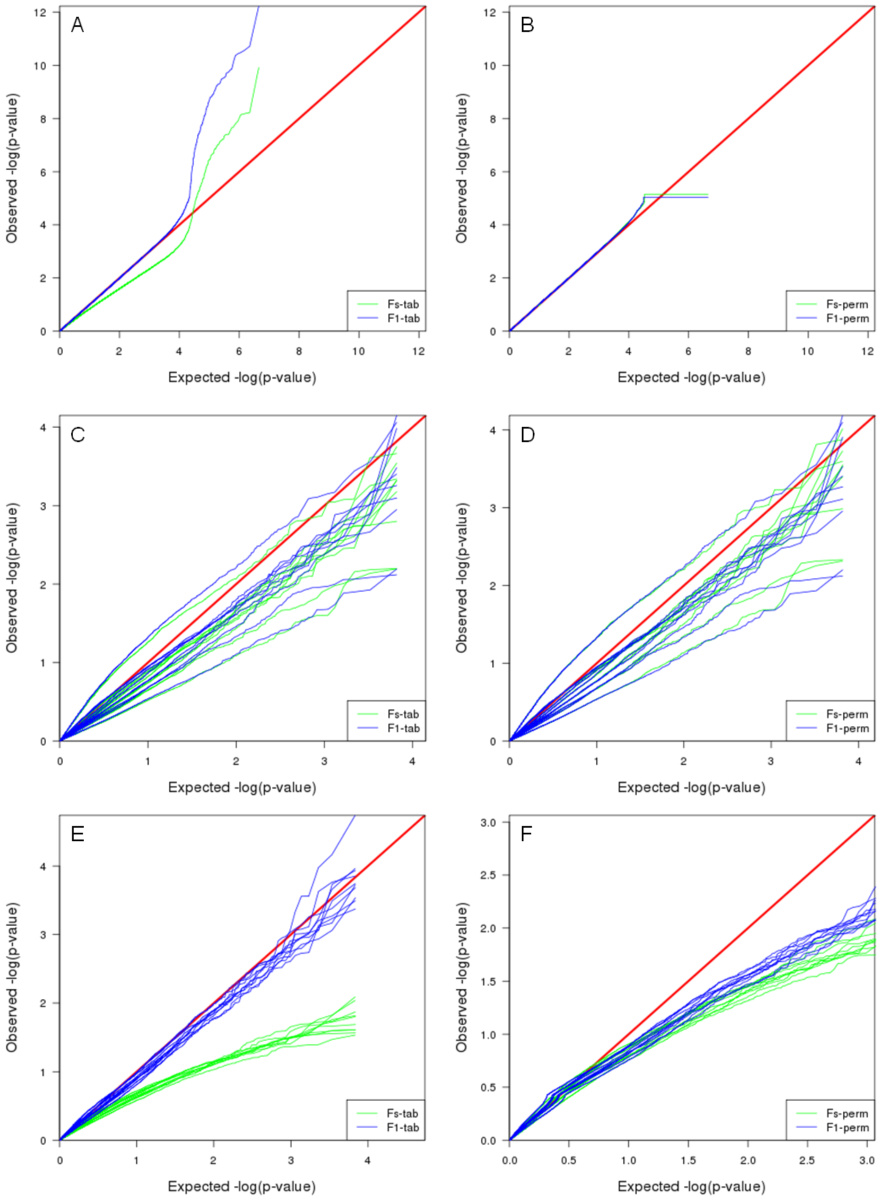


Figure 8 Comparison of MAANOVA's p-value results

Comparison of MAANOVA's p-value results of non differentially expressed transcripts using Fs moderated test and F1 transcript by transcript test, with a tabulated (right) or permutation (left) approach. In the simulated dataset **(A-B)** the permutation approach presented a more characteristic uniform distribution, the plateau at the end is caused by the replacement of zeroes by the minimum observed p-value when computing logarithm. Plasmodes generated from Cheung dataset, presented similar patterns either using a tabulated or a permutation approach **(C-D)**. Plasmodes generated from Bottomly presented better patterns for Fs with tabulated and F1 with permutation approach **(E-F)**

2.4 Discussion and conclusion

Validating and comparing methods to analyze RNA-seq data is essential for providing powerful statistical packages that can detect differentially expressed genes in downstream analyses (Robles et al., 2012). In this paper we illustrate how to utilize plasmode datasets in combination with simulations to evaluate analysis methods more comprehensively.

Parametric simulations can benefit a particular model depending on the distribution and specifications used to generate the dataset. For example, it can be argued that in our simulation study, edgeR and DESeq resulted too liberal compared to MAANOVA due to the additive generalized Poisson model that was used to simulate the dataset. However, results from two independent plasmode datasets, generated without using specific parametric models, confirmed the same behavior (Figure 6b and Figure 7b). Moreover, a common problem of parametric simulations is that genes are simulated independently. Such misspecification is overcome in plasmode datasets where the residual correlation structure among genes after adjusting for systematic effects is preserved with respect to the original dataset.

Exploring the joint null distribution of p-values for a particular test helps to determine the adequacy of a model and to decide the best method to correct for multiple comparisons, and doing so requires generation of multiple accurate high-dimensional datasets (Leek and Storey, 2011). For example, we compared null p-value distribution obtained for the two types of MAANOVA F tests (Fs or F1) combined with two methods to compute the p-values (tabulated F or permutation). The choice of the best combination varies for each dataset: in the simulation study, either Fs or F1 using permutation provide a p-value distribution close to a uniform distribution while none of the F tests using tabulated values provide a reasonable distribution (Figure 8a-b). Plasmodes generated from Cheung dataset presented similar patterns for all the combinations (Figure 8c-d), then Fs and F1 using permutation were chosen as suggested by Cui et al. (Cui et al., 2005). Conversely, in the analysis of plasmodes generated from Bottomly dataset, F1 test using tabulated F values was the best approach (Figure 8e-f). According to Cui et al. (2005), the F1 test for a fixed effect model has a standard F distribution and critical values could be obtained from F tables. These results are important because typical correction by FDR as proposed by Benjamini and Hochber (1995) may not be appropriate if the underlying uniform distribution is not supported. Other strategies have been adapted from Storey (2002) to estimate FDR for RNA-seq data and which correction should be applied is a topic of reseach (Li et al., 2012). All in all, these results emphasize the need to validate methods under realistic conditions and to select a base dataset for a plasmode where total sample size and sequencing depth (magnitude of counts) are considered.

In addition to the base dataset used to build a plasmode, the specific algorithm for plasmode generation should vary according to the objective of the study. Gadbury et. al (2008) presented an algorithm that generates the partition of samples into two groups and repeatedly samples different effect sets to be added to that unique partition. In this work, we propose to make several partitions from the original set of samples and add a set of effect in each case (Figure 4). This approach constitutes a way to incorporate valuable information on biological variation. For example, one can easily study the dispersion of patterns in the Q-Q plots or ROC curves. Alternatively, both approaches, Gadbury et al.(2008) and the one presented in this paper, can be combined to study the influence of different sets of genes as well as sample variability.

Moreover, the construction of a plasmode must consider all the experimental conditions under which the base data were collected. Treatment and block effects may be easily identified from the experimental design but further restrictions in randomization (flowcell, lane, time) or technical issues (operator, use of technical replicates) may arise only from inspecting protocol details and applying explorative statistical analyses. For instance, descriptive analysis of the Cheung dataset and visualization of samples using multidimensional scaling analysis (Figure 2a) suggested that no specific effects were present in the data structure; therefore we used it as an example to build a null plasmode. However, the same procedure applied to the Bottomly dataset indicated that not only the main strain, but also a characteristic effect due to flowcell number was an important source of variation (Figure 2b). Consequently, strain and block (flowcell) were considered in two parts of the plasmode generation algorithm: firstly, when defining the model to select the effects (step 2 in Figure 4), and secondly, when partitioning samples within each flowcell (step 5 in Figure 4). These considerations allowed us to generate appropriate null and alternative datasets. A similar process should be followed with any new dataset plausible of being used as a base for plasmode generation.

We used the plasmodes and simulated data to illustrate the selection of optimal differential expression analysis strategies. To this end, we focused on comparing true and false positive rates of tests to assess type I error rates and power. While we did not intend to perform a comprehensive evaluation of analysis protocols for RNA-seq data analysis, we did want to include two broad types of methods: 1) those directly tailored to count data by using negative binomial distributions (DESeq, EdgeR) or 2) a Gaussian model after transformation (MAANOVA). We found that edgeR and DESeq incurred in inflated type I error rates for small significance levels (Figures Figure 5b, Figure 6b, and Figure 7b) while MAANOVA's p-values tend to be closer to the nominal significance levels. Admittedly, after adjusting for type I error rates, power was similar for edgeR and DESeq and higher than that from MAANOVA (Figure 7c). However, in a real data scenario, adjusting is not possible because the true status is unknown.

These results emphasize the fact that RNA-seq data are complex and to decide what method to use may be experiment-specific due to the unknown distributions of expression levels. Plasmodes may contribute to decisions on which method to choose by using a similar pre-existing dataset and comparing results. It is critical to select a dataset that has a complete description of the experimental design and detailed protocols of how the data were obtained. Using this information, it is possible to design proper null and alternative datasets. For example, it was easy to find a set of differentially expressed genes in the mouse dataset that studied two inbred lines. Contrarily, in the human dataset, it was not possible to explain the variation in expression only as a consequence of gender effects. The human subjects came from an outbred population and factors such as age, weight, or other characteristics could have explained differences in gene expression. Granted, any of the mentioned effects could have been included in the model if the information was available. The promising results obtained from this approach emphasize the need of promoting and improving systematic data sharing across the research community to facilitate plasmode building.

Finally, the flexibility of plasmode construction allows comparing model tuning selection for downstream analysis but also upstream analysis, as normalization procedures or alignment pipelines, could be contrasted. Future uses of plasmodes could be: comparison of alignment programs for a given statistical analysis model or even exploring interaction of statistical model and read processing protocols to find optimal combined pipelines for data processing from reads-to-p-values.

APPENDIX

RNA-seq data processing

In a previous experiment, we used the Pig oligoarray for transcriptional profiling of developing pig skeletal muscle, and results for a study comparing transcript profiles of *longissimus dorsi* muscle from fetuses at 40 and 70 d of gestation in two different breed (Sollero et al., 2011). For this study, we used the same RNA samples from one of the breed types profiled with the Pigoligoarray (n = 3 for each developmental age) with deep sequencing technology (Illumina GAIIx) to obtain 50nt paired end reads from 6 libraries (2 conditions, 3 bio-replicates each). The processing steps are described in the following Figure.

1. Filter passing read pairs were aligned to the S. scrofa reference genome (Sscrofa9, April 2009, Ensemble release 61) using the spliced RNA aware aligner, TopHat. Reads from each library were aligned separately. The reference gene annotation (same version as above) was provided to TopHat to provided information about predicted splice junctions, but TopHat also predicts novel splice junctions.
2. Novel splice junctions predicted from each of the 6 libraries were combined with the splice annotations from the reference to create a single, non-redundant set of predicted splice sites. TopHat is better able to map spliced reads if a list of potential junctions is provided as input.
3. Each library was aligned to the reference a second time, providing the non-redundant set of potential splice sites as input.
4. The alignments produced from each library were used as input to Cufflinks. Cufflinks was used to examine RNA-Seq alignments and to generate a set of predicted transcripts based on assembly of overlapping reads.
5. The transcript models generated by Cufflinks for each library were combined into a single, non-redundant set of transcript/gene models with Cuffcompare. The reference annotation was also provided to Cuffcompare to associate the predicted gene models with their most likely reference model.
6. The models generated by Cuffcompare are filtered to remove those models with little support from the underlying read alignments. Specifically, if aligned reads are observed in only one of the six libraries, that model is removed from the final set.
7. The alignments for each library produced during the second round of TopHat (step 3) and the curated set of gene models (step 6) were used as input to htseq-count. This program compared a set of alignments to an annotation file and reported the number of fragments uniquely aligned to each gene in the annotation. The models generated by Cuffcompare may have multiple transcripts modeled for a particular gene but htseq-count only reports the total fragments for a gene.

Figure 9 Steps used to process reads to obtain matrix of counts

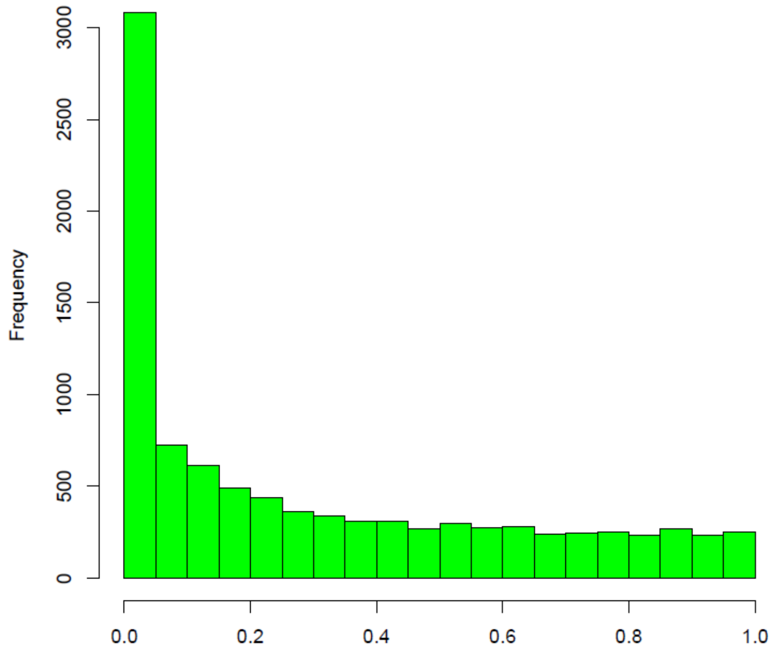


Figure 10 P-value distribution of differential expression analysis performed with edgeR for Bottomly data using a model with block and treatment fixed effects

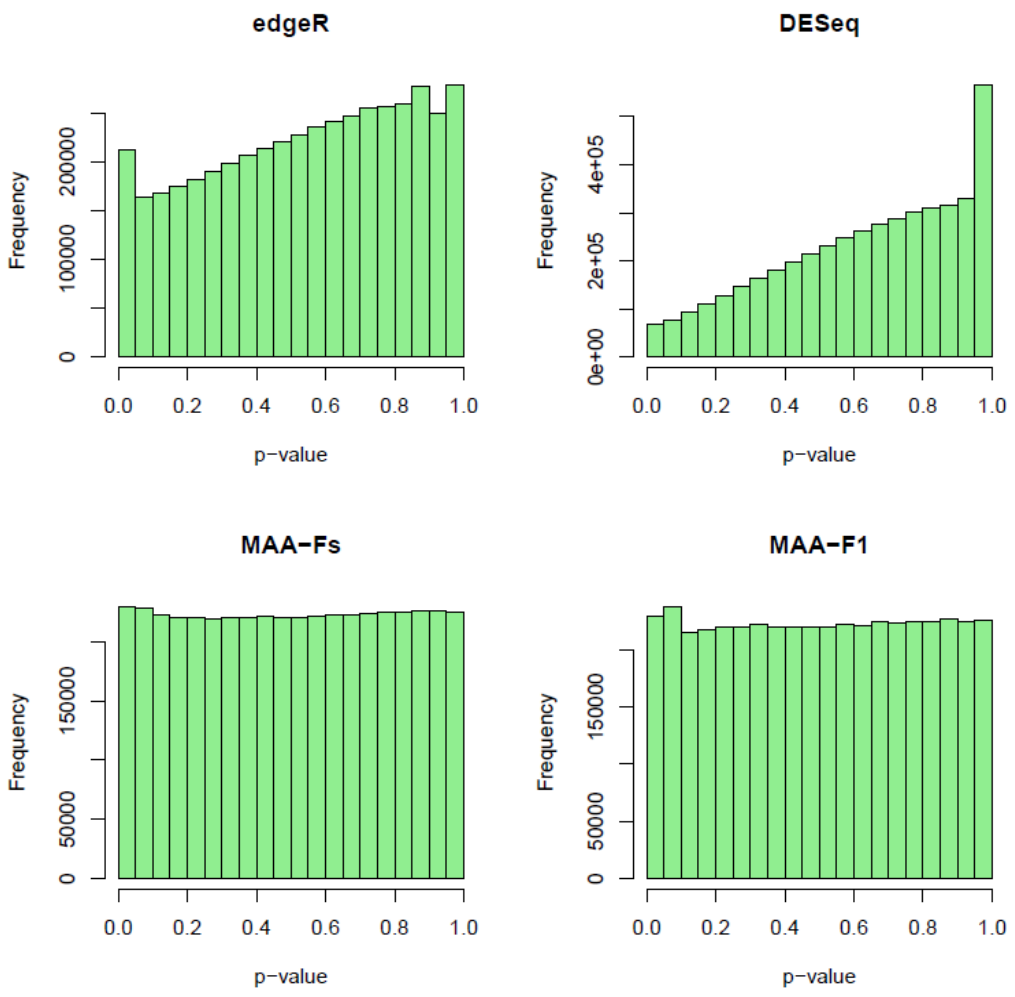


Figure 11 P-value distribution of non differentially expressed transcripts for a simulated scenario with three biological replicates

P-value distribution of non differentially expressed transcripts for a simulated scenario with 3 biological replicates using: 1) edgeR, 2) DESeq, 3) MAA-Fs: MAANOVA Fs moderated test (permutation), and 4) MAA-F1: MAANOVA F1 transcript by transcript test (permutation)

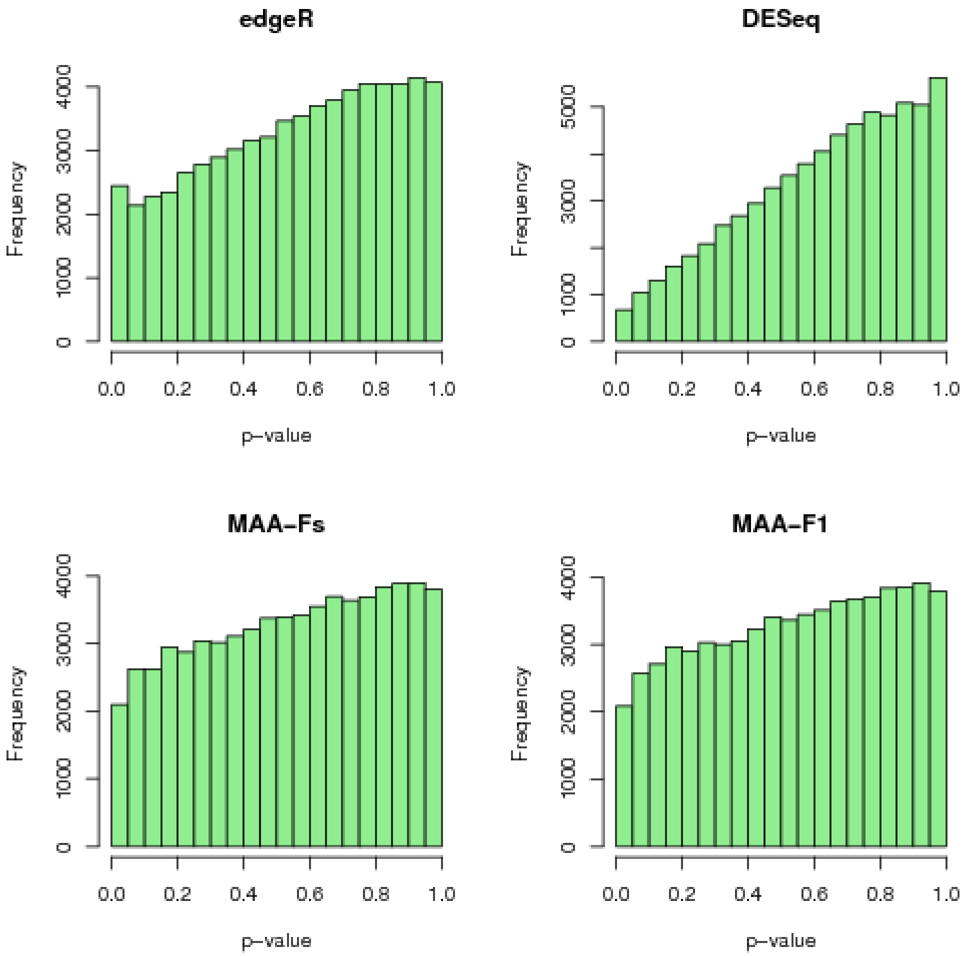


Figure 12 P-value distribution of non differentially expressed transcripts for plasmodes generated from Cheung dataset

P-value distribution of non differentially expressed transcripts for plasmodes generated from Cheung dataset using: 1) edgeR, 2) DESeq, 3) MAA-Fs: MAANOVA Fs moderated test (permutation), and 4) MAA-F1: MAANOVA F1 transcript by transcript test (permutation)

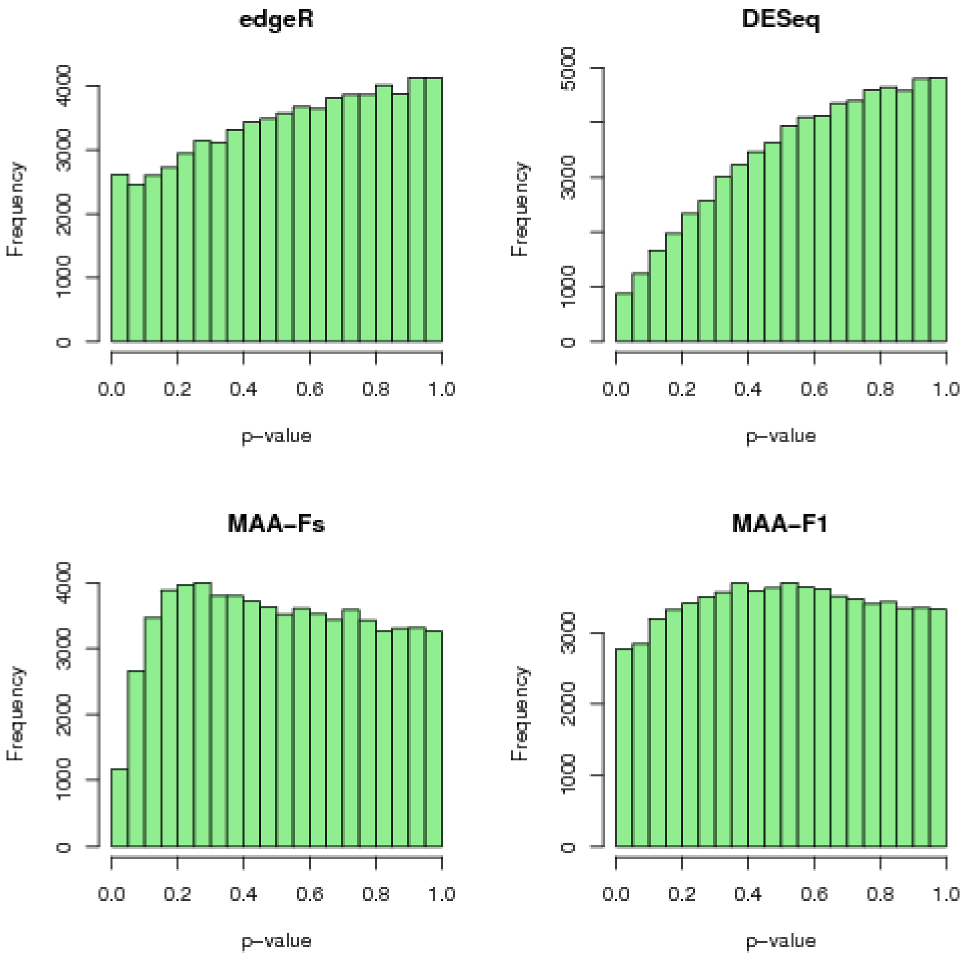


Figure 13 P-value distribution of non differentially expressed transcripts for plasmodes generated from Bottomly dataset

P-value distribution of non differentially expressed transcripts for plasmodes generated from Bottomly dataset using: 1) edgeR, 2) DESeq, 3) MAA-Fs: MAANOVA Fs moderated test (permutation), and 4) MAA-F1: MAANOVA F1 transcript by transcript test (tabulated)

Chapter 3

Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using plasmode datasets

Reeb, P. D., Bramardi, S. J., and Steibel, J. P. (2015) Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using plasmode datasets. *PLoS One 10(7):* e0132310. doi:10.1371/journal.pone.0132310

3.1 Introduction

Hierarchical cluster analysis has been a popular method for finding patterns in data and for representing results of gene expression analysis (Liu and Si, 2014). Clustering algorithms have been widely studied for analyzing microarray data (Jiang et al., 2004; Dalton et al., 2009), however, such technology is being rapidly replaced by RNA sequencing technology (RNA-seq) (Wang et al., 2009). In contrast to microarray experiments, RNA-seq generates count data of discrete nature that may call for different analysis methods. One of the most obvious differences between clustering gene expression data from RNA-seq or microarray is the choice of a dissimilarity measure, or the need to transform and normalize RNA-seq data in order to use dissimilarity measures commonly used for microarray data (Liu and Si, 2014).

Before implementing any statistical analysis of RNA-seq data, normalization and transformation have to be performed. (Bullard et al., 2010; Law et al., 2014; Liu and Si, 2014). Normalization aims at reducing non-systematic variation within and between samples, such as sequencing depth and library preparation. Data transformation could be very important because it aims at reducing the effects of skewness, scale and presence of outliers that can be found in read count data that usually follow a Poisson (Marioni et al., 2008) or negative binomial distribution (Robinson et al., 2010; Anders and Huber, 2010). Through appropriate transformation, dissimilarity measures that are sensitive to asymmetric distributions and scale magnitude, such as Euclidean and 1 – Pearson correlation (Liu and Si, 2014; Jiang et al., 2004; Johnson and Wichern, 2002) could be used for clustering RNA-seq data.

Although a Gaussian distribution assumption is not required to compute Euclidean and correlation based distances, transformations that convert count data into a continuous and almost Gaussianly distributed variable (Law et al., 2014) could be used for hierarchical clustering. For instance, besides the classical logarithmic transformation, several functions have been proposed to model the mean-variance relationship of RNA-seq data (Anders and Huber, 2010; Love et al., 2014a; Law et al., 2014), while accounting for over-dispersion. But the properties of those transformations need to be tested.

Finally, instead of using transformations to approximate the data to a pre-specified distribution where available dissimilarity measures perform well, model based methods can be directly used to compute dissimilarity measures (Witten, 2011).

Evaluating the adequacy of alternative dissimilarity measures for hierarchical clustering requires the fundamental step of choosing reference datasets (Handl et al., 2005). An ideal reference dataset should mimic the technical and biological variability found in experimental data, and it should also have some *a priori* known structure in order to assess the goodness of results from alternative analyses. Parametric simulations, exemplar datasets, and permutation sampling have been used to generate such datasets in clustering analysis of biological data (Sloutsky et al., 2013). Similarly, plasmode datasets (Mehta et al., 2004) have been proposed for evaluating differential expression analysis in RNA-seq experiments (Reeb and Steibel, 2013). A plasmode is a dataset obtained from experimental data from which some truth is known, thus, it is an ideal way to generate data with an a priori defined structure that realistically mimics RNA-seq data. Plasmodes were originally proposed for assessing multivariate analysis methods (Cattell and Jaspers, 1967) and have been used in behavioral science (Waller et al., 2010) and also in genomics (Gadbury et al., 2008; Steibel et al., 2009).

In this paper, we propose the use of plasmode datasets to assess the properties of dissimilarity measures for agglomerative hierarchical clustering or RNA-seq data. We present two possible ways of creating plasmode datasets that depend on the available data structure, and we use the resulting reference datasets to compare several commonly used dissimilarity measures.

3.2 Material and Methods

3.2.1 Datasets

Two experimental datasets were used in this study to create reference datasets. The first dataset, *"*Bottomly*",* corresponds to an experiment described elsewhere (Bottomly et al., 2011). Briefly, 21 samples of *striatum* tissue from two inbred mouse strains (C57BL/6J (B6), n=10; and DBA/2J (D2), n=11) were sequenced in three Illumina GAIIx flowcells. Data were downloaded from ReCount website (Frazee et al., 2011). After filtering out genes with zero counts in all samples, the count matrix contained 13932 rows (transcripts) and 21 columns (samples). The second dataset, "MSUPRP", corresponds to 24 samples of *longissimus* muscle selected from the MSU Pig Resource Population (Steibel et al., 2011) and sequenced by our collaborators (Steibel et al., 2014). Total RNA from 24 F2 female pigs of Duroc by Pietrain ancestry was barcoded and sequenced on Illumnina HiSeq 2000. Read mapping, gene modelling and read counting were performed using Tophat (Trapnell et al., 2009), Cufflinks (Trapnell et al., 2012) and HTSeq (Anders et al., 2015), respectively. After processing the sequence reads, we obtained a count matrix with 26740 rows (transcripts) and 24 columns (samples). (For details, see supplementary material). The count matrix of the five samples (animals) used in this paper is available as supporting information at PLoS ONE website.

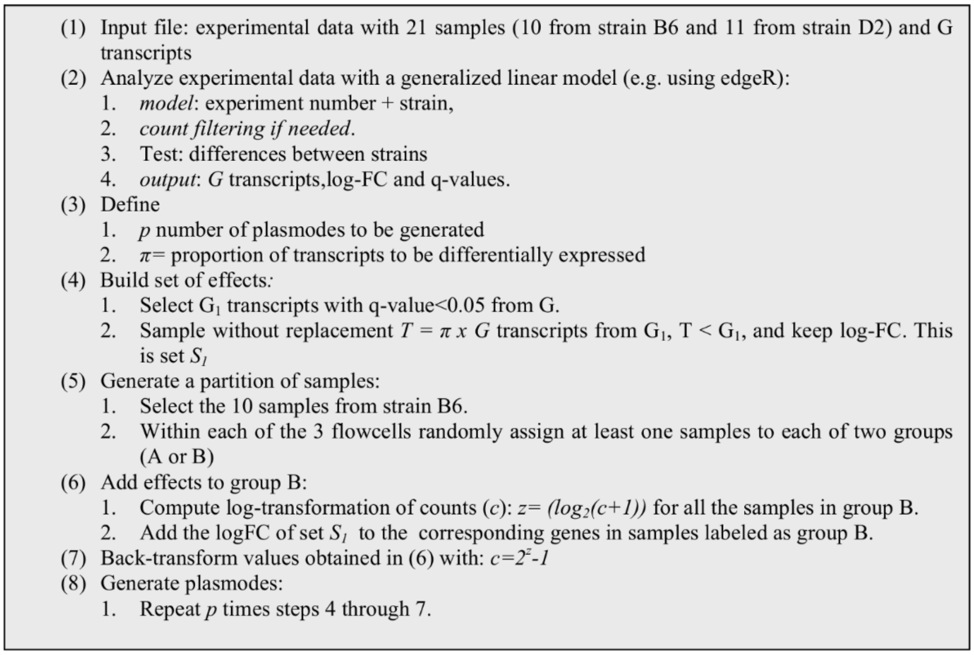
3.2.2 Plasmodes

Plasmodes are synthetic datasets generated from experimental data for which some true characteristic is known (Mehta et al., 2004). For instance, we may know *a priori* which genes are not differentially expressed or we may know group membership of each sample. Then, we build a plasmode by re-shuffling the existing data without assuming any probability distributions or correlation structures. Thus, we can use the known characteristic of the synthetic dataset to assess properties of analysis methods. For instance we can apply resampling-based methods to create plasmodes consistent with the null hypothesis (no differential expression) and use them to evaluate the type I error rate hypothesis of testing procedures (Mehta et al., 2006), or we can use the known group memberships to assess the accuracy of clustering methods, as we do in this paper. Thus, plasmodes need to be constructed according to the validation objectives (i.e. considering the statistical method that is being evaluated) and considering the available experimental data.

In this paper, we present two examples on how to create plasmodes to assess the effect of choice of dissimilarity measures on the results of hierarchical clustering of RNA-seq data. In the first experimental dataset, the natural structure of the data is known *a priori* and it was generated through the experimental design (sequencing flowcells and mice strains), while in the second experimental dataset there is not an *a priori* known structure, so we create a set of artificial samples where the structure is generated by construction.

3.2.2.1 Plasmodes from Bottomly dataset

We built plasmodes for this dataset by using samples from B6 strain, partitioning them in two groups and adding known effects for selected genes taken from the difference in gene expression with strain D2. Figure 14 presents the algorithmic steps used to generate the plasmodes. Two main effects, strain and flowcell, were used to classify the 21 samples (Step 2.1 in Figure 14) given the importance of both sources of variation has been described before (Reeb and Steibel, 2013; Law et al., 2014). Then, a differential expression analysis including all the samples (both strains) was conducted with edgeR (Robinson et al., 2010) and transcripts with q-value <0.05 were identified as differentially expressed (set G_1 in step 4 of Figure 14). Subsequently, samples from strain B6 were randomly assigned to two groups (A or B) within each flowcell, and a subset (S_1) of effects randomly selected from G_1 was added to the corresponding genes in samples labelled as B (Steps 5-6) . Therefore, samples from group A and B differ due to the strain effect added by the subset (S_1) of differentially expressed genes, while samples within each group differ due to the flowcell effect. We generated 50 plasmodes with 10% of differentially expressed transcripts by defining p=50 and π=0.10 in step 3 and by randomly assigning 2 samples to group B and one or two samples, if available, to group A within each flowcell (Step 5.2 in Figure 14). As a result, in each plasmode generation we obtained a total of 10 samples under two artificial treatments (A or B) and three flowcell effects (1, 2 or 3), resulting in a set of samples indexed by such factors as:{(A_1, A_1, B_1, B_1),(A_2,B_2,$B_{2)}$,(A_3,B_3,B_3)}. If we use only differentially expressed genes, we expect the samples with same letter to cluster together because of the treatment effect, but as we add a large number of non differentially expressed genes, we can expect that samples with the same subindex (flowcell) will tend to cluster together because it has been shown before that there is a strong flowcell effect in this experiment (Law et al., 2014; Reeb and Steibel, 2013). To evaluate the performance of dissimilarity measures under various differentially expressed / non differentially expressed ratios (DE/nonDE), we analyzed three scenarios for each plasmode: 1) only DE transcripts ($DE_{[100\%]}$), 2) DE transcripts + all nonDE transcripts ($DE_{[10\%]}$+$nonDE_{[90\%]}$), and 3) DE transcripts + a random sample of 50% from nonDE transcripts ($DE_{[20\%]}$+$nonDE_{[80\%]}$).



```
(1) Input file: experimental data with 21 samples (10 from strain B6 and 11 from strain D2) and G transcripts
(2) Analyze experimental data with a generalized linear model (e.g. using edgeR):
    1. model: experiment number + strain,
    2. count filtering if needed.
    3. Test: differences between strains
    4. output: G transcripts,log-FC and q-values.
(3) Define
    1. p number of plasmodes to be generated
    2. $\pi$= proportion of transcripts to be differentially expressed
(4) Build set of effects:
    1. Select $G_1$ transcripts with q-value<0.05 from G.
    2. Sample without replacement $T = \pi \, x \, G$ transcripts from $G_1$, $T < G_1$, and keep log-FC. This is set $S_1$
(5) Generate a partition of samples:
    1. Select the 10 samples from strain B6.
    2. Within each of the 3 flowcells randomly assign at least one samples to each of two groups (A or B)
(6) Add effects to group B:
    1. Compute log-transformation of counts (c): $z= (log_2(c+1))$ for all the samples in group B.
    2. Add the logFC of set $S_1$ to the  corresponding genes in samples labeled as group B.
(7) Back-transform values obtained in (6) with: $c=2^z-1$
(8) Generate plasmodes:
    1. Repeat p times steps 4 through 7.
```

Figure 14 Algorithm used to generate plasmodes from Bottomly dataset

3.2.2.2 Plasmodes from MSUPRP dataset

Since this dataset did not have a natural sample structure derived from experimental conditions, a structure had to be induced in order to know a priori the expected clustering configuration. From a descriptive multidimensional scaling analysis of the 24 pig samples (animals), we selected 5 dissimilar samples (A, B, C, D, E) according to their configuration in the main plane (Figure 15). Synthetic samples were generated by combining a known proportion of randomly sampled read counts of individual genes from each of two of the five selected samples. For instance, a new synthetic sample named AAC was generated combining 2/3 and 1/3 of read counts of individual genes from A and C respectively. A full plasmode consisted of 12 samples that included the five selected samples {AAA, BBB, CCC, DDD, EEE}, five synthetic samples {AAC, BBC, CCB, DDE, EED} obtained by combining 2/3:1/3 proportions from two of the selected samples, and two synthetic samples {CxB, ExD} obtained by combining 1/2:1/2 proportions of two of the selected samples (see Figure 21.in supplemental material with a representation of the relationships among the 12 samples of each plasmode). Following this procedure, a total of 50 replicated plasmodes were generated. As a result we created a synthetic dataset where the samples were expected to resemble each other to a known degree given the proportions of shared reads.

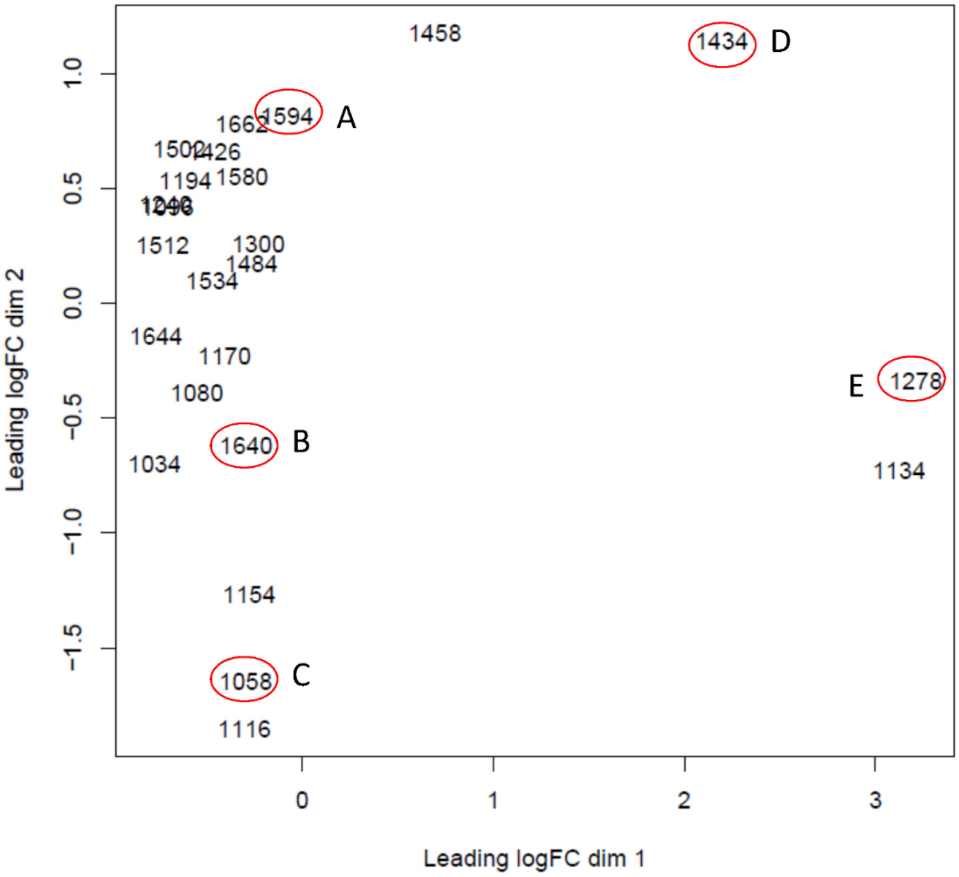


Figure 15 Multidimensional scaling analysis of MSUPRP dataset

Twenty four samples were represented in the main plane (dimension 1 and dimension 2 explained 22.4% and 13.8% respectively) and five distant samples (A, B, C, D, E, marked with ovals) were selected as input samples to generate plasmode datasets.

3.2.3 Clustering

Defining a dissimilarity measure and a linkage method are the two key decisions for performing hierarchical cluster analysis. We focused on assessing the adequacy of dissimilarity measures that have been commonly used for clustering gene expression data. We also include a recently proposed dissimilarity measure for RNA-seq count data (Witten, 2011). As linkage method, we decided to use complete linkage because it is invariant under monotone transformations (Izenman, 2008), and hence dissimilarity measures that have the same relative ranking result in the same cluster structure (Liu and Si, 2014). This robustness reduces the effect of linkage method when comparing dendrograms and allowed us to concentrate in the evaluation of dissimilarity measures. Hierarchical cluster analysis was applied to each plasmode using the agglomerative procedure implemented in function hclust from R (R Development Core Team, 2014) to concatenate samples and to generate dendrograms.

Eight dissimilarity measures were compared, including 4 variants based on Euclidean distance, 3 correlation based approaches, and one Poisson based measure. Euclidean distances were computed between samples following one of 4 approaches: i) using raw count data (*raw*), ii) after normalizing samples using the median ratio size factor proposed by Anders and Huber (Anders and Huber, 2010) (*rnr*), iii) after applying a variance stabilizing transformation computed with DESeq2 (Love et al., 2014b) (*vsd*), and iv) after applying a regularized logarithm transformation implemented in DESeq2 (Love et al., 2014b) (*rld*). Correlation based dissimilarities comprised: i) 1- Pearson correlation between samples using raw counts (*pea*), ii) 1- Pearson correlation between samples using counts transformed by logarithm of raw counts +1 (*plg*), and iii) 1- Spearman correlation between samples using raw counts (*spe*). The Poisson dissimilarity (*poi*), which is based on a log likelihood ratio statistic for a Poisson model (Witten, 2011), was computed on data that were transformed by a power function to account for overdispersion, and normalized by total sum of counts for each sample.

3.2.4 Cluster validation using results from plasmodes

Cluster validation can be assessed using several indices (Halkidi et al., 2001; Xiong and Li, 2013) and the choice of a particular measure is application dependent (Jiang et al., 2004). Cophenetic distances provide a way to quantify similarities among dendrograms in hierarchical clustering. The cophenetic distance is the distance from the bottom of the tree at which two elements (samples in this paper) are grouped in the same cluster for the first time in the hierarchy. To represent a dendrogram in terms of a set of cophenetic distances, the distances between all pairs of elements is computed and arranged into a matrix called cophenetic matrix that represents the whole hierarchy, as illustrated in Figure 22 and Table 5 in supplementary material. Cophenetic matrices can be used to compare dendrograms (Sokal and Rohlf, 1962). For instance, to compare how similar are two dendrograms, the Pearson correlation between the lower triangular portions of two cophenetic matrices can be used.

We computed the correlation between cophenetic matrices (Handl et al., 2005) to compare dendrograms obtained with different dissimilarity measures (between dissimilarity measure comparison) as well as to compare all dendrograms obtained with a particular dissimilarity measure (within dissimilarity measure comparison). Mean and standard deviation of correlations between dissimilarity measures were used as a measure of agreement while mean and standard deviation of correlations within a dissimilarity measure were used as a measure of consistency.

We also visually compared the obtained dendrograms to a reference dendrogram built according to the sample structure known *a priori* from the plasmode generation process in the MSUPRP dataset. For the MSUPRP dataset, we defined the expected similarity between two samples (s_{ij}) as the maximum proportion of shared reads, and we defined 1- s_{ij} as a reference dissimilarity (see Tables Table 3 and Table 4 in supplementary material). With the correlation between each of the dissimilarity matrices and the reference dissimilarity, we assessed how well each dissimilarity measure recovered the expected sample structure. An equivalent reference dissimilarity matrix and reference dendrogram cannot be easily built for the Bottomly dataset because we did not exploit relationships between samples to build the plasmode, except for their group membership. In this case, we compared typical dendrograms obtained from plasmodes to the known strain and experiment membership in the original data.

3.3 Results

3.3.1 Bottomly

Figure 16 shows the typical dendrograms obtained for plasmode datasets using two dissimilarity measures, *poi* and *rnr*, which are representative examples of two sets of results under the three different scenarios ($DE_{[100\%]}$, $DE_{[10\%]}$+non$DE_{[90\%]}$, and $DE_{[20\%]}$+non$DE_{[80\%]}$,). On the one hand, scenario 1 ($DE_{[100\%]}$) uses only differentially expressed transcripts, therefore the expected hierarchy should arrange samples in two separate groups according to main treatment labels. Such is the structure obtained utilizing the Poisson (*poi*) dissimilarity measure (Figure 16a). Using the Poisson dissimilarity measure, samples were clustered in two groups corresponding to treatments A or B, and within each of the groups, samples were arranged according to block numbers (4, 6, or 7). Differently, the dendrogram based on Euclidean distance calculated from raw normalized data (*rnr*) (Figure 16b) mixed treatment labels and did not recover any expected structure. On the other hand, scenario 2 ($DE_{[10\%]}$+non$DE_{[90\%]}$), uses information from differentially (10%) and non differentially expressed (90%) transcripts. As a result, we expected that the dissimilarity measures would tend to represent other aspects of samples in addition to the treatment effect. In concordance with such expected structure, dendrogram obtained using the Poisson dissimilarity (Figure 16c) firstly separated samples according to block labels, block 4 being the most different group. Subgroups for treatments A and B were arranged within each block. Conversely, dendrogram based on Euclidean distance calculated from raw normalized data (*rnr*) (Figure 163d) did not present any expected structure. Finally, scenario 3 (DE$_{[20\%]}$+nonDE$_{[80\%]}$) represents an intermediate case that is useful to further explore the performance of dissimilarity measures because it is enriched in DE genes with respect to scenario 1, but it still conserves 80% of background (nonDE) genes. The dendrogram based on the Poisson dissimilarity (Figure 16e) presented an intermediate structure where we observed that samples from block 4 were clustered together while the remaining samples were clustered in a separate group mainly classified by treatment effect. Yet again, dendrogram based on *rnr* (Figure 16f) did not characterize any expected configuration. To sum up, for this dataset, dendrograms generated from a Poisson dissimilarity resemble the expected hierarchical structures in all three scenarios, however, dendrograms based on Euclidean distance computed on raw normalized data did not. Comparison of hierarchies between clusters constructed using *poi* and *rnr* dissimilarities across 50 plasmodes presented correlation of cophenetic matrices with low means and high standard deviations (0.52±0.24, , 0.65±0.17, 0.59±0.23 , for scenarios 1, 2 and 3, respectively) (Figure 17). These results emphasize a poor correspondence between hierarchies constructed upon *poi* and *rnr* dissimilarities.

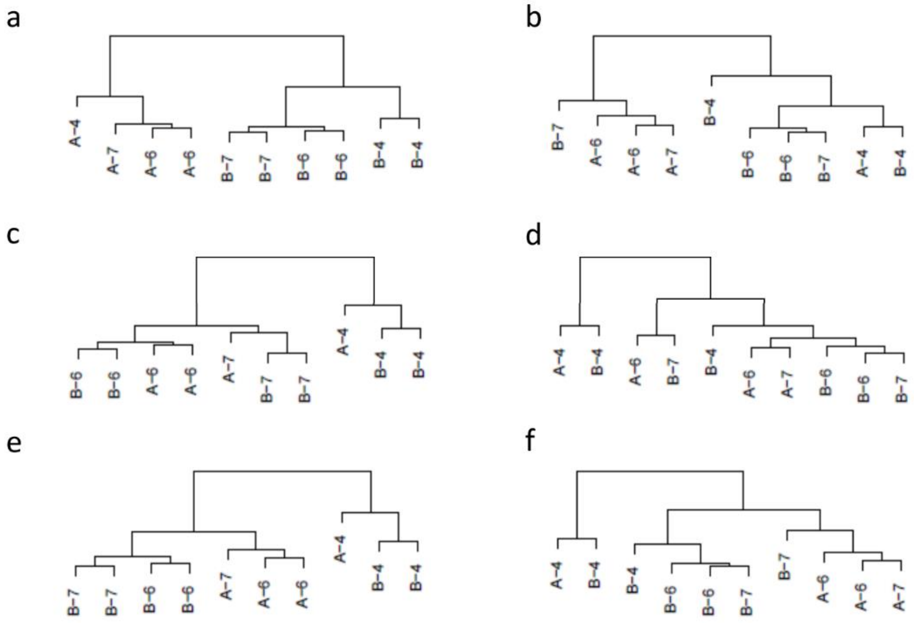
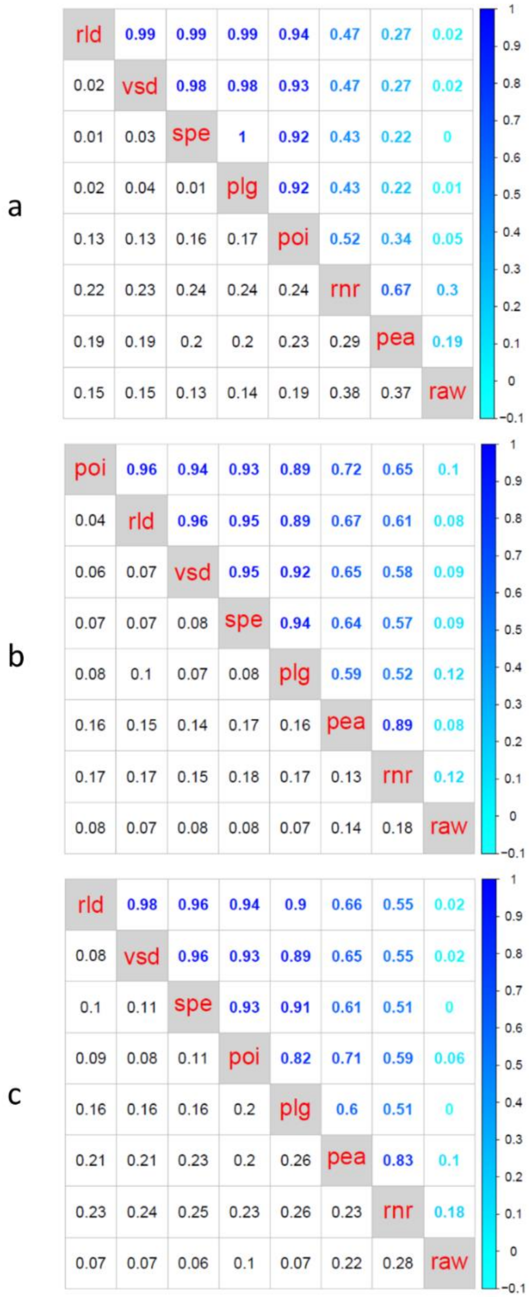


Figure 16 Typical dendrograms obtained for plasmode datasets from Bottomly experimental data with two dissimilarity measures under three scenarios

Dendrograms obtained using complete linkage hierarchical clustering based on Poisson dissimilarity (*poi*) are presented in the left column (a, c and e), and dendrograms based on Euclidean distance calculated from raw normalized data (*rnr*) are presented in right column (b, d, f). The rows correspond to three scenarios with different percentage of differentially expressed (DE) transcripts: 1) $DE_{[100\%]}$ (a and b), 2) $DE_{[10\%]}$+$nonDE_{[90\%]}$ (c and d), and 3) $DE_{[20\%]}$+$nonDE_{[80\%]}$ (e and f). Sample labels correspond to main treatment (A or B) and flowcell number (4, 6 or 7). Dendrograms based on *poi* separates samples according to the expected sources of variation; in (a), only DE transcripts, samples are arranged in two separate groups following treatment labels; in (c), with a predominant number of non DE transcripts, the structure of groups is dominated by flowcell characteristics in addition to main treatment: and in (e) an in-between scenario, the dendrogram presents an intermediate group structure. Dendrograms based on *rnr* do not resemble any expected configuration.



rld	0.99	0.99	0.99	0.94	0.47	0.27	0.02
0.02	vsd	0.98	0.98	0.93	0.47	0.27	0.02
0.01	0.03	spe	1	0.92	0.43	0.22	0
0.02	0.04	0.01	plg	0.92	0.43	0.22	0.01
0.13	0.13	0.16	0.17	poi	0.52	0.34	0.05
0.22	0.23	0.24	0.24	0.24	rnr	0.67	0.3
0.19	0.19	0.2	0.2	0.23	0.29	pea	0.19
0.15	0.15	0.13	0.14	0.19	0.38	0.37	raw

poi	0.96	0.94	0.93	0.89	0.72	0.65	0.1
0.04	rld	0.96	0.95	0.89	0.67	0.61	0.08
0.06	0.07	vsd	0.95	0.92	0.65	0.58	0.09
0.07	0.07	0.08	spe	0.94	0.64	0.57	0.09
0.08	0.1	0.07	0.08	plg	0.59	0.52	0.12
0.16	0.15	0.14	0.17	0.16	pea	0.89	0.08
0.17	0.17	0.15	0.18	0.17	0.13	rnr	0.12
0.08	0.07	0.08	0.08	0.07	0.14	0.18	raw

rld	0.98	0.96	0.94	0.9	0.66	0.55	0.02
0.08	vsd	0.96	0.93	0.89	0.65	0.55	0.02
0.1	0.11	spe	0.93	0.91	0.61	0.51	0
0.09	0.08	0.11	poi	0.82	0.71	0.59	0.06
0.16	0.16	0.16	0.2	plg	0.6	0.51	0
0.21	0.21	0.23	0.2	0.26	pea	0.83	0.1
0.23	0.24	0.25	0.23	0.26	0.23	rnr	0.18
0.07	0.07	0.06	0.1	0.07	0.22	0.28	raw

Figure 17 Agreement between dissimilarity measures using Bottomly plasmode datasets

Each matrix contains means (upper triangle) and standard deviations (lower triangle) of correlation between cophenetic matrices of dendrograms (N= 50 plasmode datasets) for eight dissimilarity measures: Euclidean distances using raw count data (*raw*), Euclidean distances using normalized samples (*rnr*), Euclidean distances using variance stabilizing transformation (*vsd*), Euclidean distances using regularized logarithm (*rld*).) 1- Pearson correlation using raw counts (*pea*), 1- Pearson correlation using counts transformed by logarithm of raw counts +1 (plg), and 1- Spearman correlation using raw counts (*spe*), and Poisson dissimilarity (*poi*). Panel labels

Figure 17 (cont’d)

(a), (b) and (c) correspond to one of three scenarios of proportion of differential expressed genes: $DE_{[100\%]}$, $DE_{[10\%]}$+non$DE_{[90\%]}$, and $DE_{[20\%]}$+non$DE_{[80\%]}$, respectively. In all scenarios, we identified three sets of dissimilarity measures: 1) *raw*, 2) *rnr* and *pea*, and 3) *poi*, *rld*, *vsd*, *plg* and *spe*. Results from *raw*, set 1, were poorly related to results from any other dissimilarity measure. Dendrograms from dissimilarity measures in set 2 presented correlation of cophenetic matrices with medium to high means and high variability with each other, and low correlation with dendrograms from other dissimilarity measures. Dendrograms from dissimilarity measures in set 3 exhibited high correlations of cophenetic matrices and low to medium variability when compared to each other.

Correlations between hierarchies obtained with the eight dissimilarity measure approaches for each of three scenarios ($DE_{[100\%]}$, $DE_{[10\%]}$+non$DE_{[90\%]}$ and $DE_{[20\%]}$+non$DE_{[80\%]}$) are presented in Figure 17. Each matrix contains means and standard deviations of correlations between cophenetic matrices, in the upper and lower triangle respectively. Regardless of the scenario, dissimilarity measures can be apportioned to three groups with common patterns. First, Euclidean distance computed on raw data (*raw*) is poorly related to any other dissimilarity measure. Second, Euclidean distance computed on normalized data (*rnr*) and 1- Pearson correlation dissimilarity (*pea*) presented medium to high correlations of cophenetic matrices (mean from 0.67 to 0.89) with high variability (standard deviation from 0.13 to 0.29) with each other and low correlation values with other dissimilarity measures. Third, there is a subset comprising 1-Pearson correlation dissimilarity computed on log-transformed counts (*plg*), 1-Spearman correlation dissimilarity (*spe*), Euclidean distance computed on transformed counts after applying either a variance stabilizing function (*vsd*) or a regularized logarithm (*rld*), and the Poisson dissimilarity (*poi*). This last group of dissimilarity measures presents high correlations of cophenetic matrices (mean from 0.82 to 0.99) with low to medium variability (standard deviation from 0.01 to 0.2). Only hierarchies obtained with dissimilarity measures from the third group consistently presented the expected natural structure created by design in the plasmode generation process, being *rld*, *vsd*, *spe* and *plg* the more consistent across all scenarios (Table 2, columns 2, 3 and 4).

Table 2 Consistency for each dissimilarity measure

Dissimilarity	Bottomly			MSUPRP
	DE$_{[100\%]}$	**DE$_{[10\%]}$+nonDE$_{[90\%]}$**	**DE$_{[20\%]}$+nonDE$_{[80\%]}$**	
raw	0.58 (0.22)	0.91 (0.13)	0.75 (0.20)	0.56 (0.23)
rnr	0.35 (0.27)	0.43 (0.28)	0.33 (0.28)	0.40 (0.22)
rld	0.98 (0.01)	0.90 (0.11)	0.88 (0.13)	0.99 (0.01)
vsd	0.96 (0.04)	0.91 (0.10)	0.86 (0.15)	0.99 (0.01)
pea	0.37 (0.31)	0.53 (0.30)	0.45 (0.30)	0.43 (0.28)
plg	0.98 (0.01)	0.89 (0.13)	0.75 (0.19)	0.98 (0.01)
spe	0.99 (0.01)	0.86 (0.14)	0.88 (0.14)	0.99 (0.01)
poi	0.86 (0.21)	0.92 (0.09)	0.88 (0.14)	0.99 (0.01)

Mean and standard deviation of correlation between cophenetic matrices of dendrograms (N=50 plasmode datasets) for each of eight dissimilarity measures: Euclidean distances using raw count data (*raw*), Euclidean distances using normalized samples (*rnr*), Euclidean distances using regularized logarithm (*rld*) Euclidean distances using variance stabilizing transformation (*vsd*), 1-Pearson correlation using raw counts (*pea*), 1- Pearson correlation using counts transformed by logarithm (plg), and 1- Spearman correlation using raw counts (*spe*), and Poisson dissimilarity (*poi*). Columns correspond to the three scenarios generated for Bottomly (with different proportion of DE genes) and the MSUPRP dataset. We considered a clustering from a dissimilarity measure to be consistent if hierarchies obtained for different plasmode datasets within each dissimilarity measure were highly correlated and presented a low standard deviation. Clustering based on *raw*, *rnr* and *pea* were generally inconsistent presenting a number of very different hierarchical structures.

We considered a dissimilarity measure to be consistent if hierarchies obtained for different plasmode datasets within each dissimilarity measure were highly correlated and presented a low standard deviation. Consequently, we computed correlations of cophenetic matrices for dendrogram within each dissimilarity measure and calculated the mean and standard deviation for each ensemble of 50 plasmodes (Table 2, columns 2, 3 and 4). Dissimilarity measures *raw*, *rnr* and *pea* were generally inconsistent, resulting in a number of different hierarchical structures. For instance, *rnr* presented mean correlation values of 0.35±0.27, 0.43±0.28, and 0.33±0.28 for the three respective scenarios. Conversely, all the other dissimilarity measures were much more consistent. For example, *rld* presented mean correlation values of 0.98±0.01, 0.90±0.11, and 0.88±0.13 for the three respective scenarios. Such high values mean that hierarchies obtained with *rld* for the 50 plasmodes were all very similar to each other.

3.3.2 MSUPRP

Plasmodes from MSUPRP were constructed by combining known proportions of sequence reads from pairs of samples, including nonDE as well as potentially DE transcripts across individual. We expect that dendrograms cluster the samples according to the known proportions of shared reads as presented in the reference dendrogram in Figure 18c (see Table 4 in supplementary material with the corresponding reference dissimilarity matrix). Figure 18a-b present the typical dendrograms obtained for plasmode datasets using *rnr* and *poi*, which are representative examples of the 8 dissimilarity metrics. Dendrogram based on the Poisson dissimilarity (Figure 18a) clustered the original samples A, B, and C and their synthetic combinations in one group, and original samples E and D and their synthetic combinations in a distinct group. The hierarchical structure of each of these two groups represented the degree of shared reads between samples by joining first samples that shared ⅓ of reads and then samples that shared ½ of reads. Additionally, the separation between samples {A, B, C} and {D,E} agreed with positions along the most important dimension (dim1) in Figure 15. In contrast, dendrogram based on *rnr* (Figure 18b) did not cluster samples according to the anticipated configuration. Comparison of hierarchies between clusters constructed from *rnr* and *poi* dissimilarities for all plasmodes presented low mean and high standard deviations (0.44±0.22) of correlation between cophenetic matrices (Figure 19).

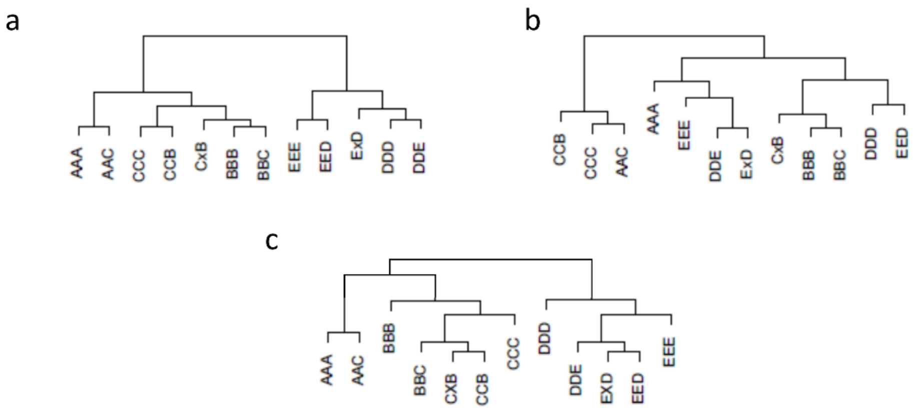
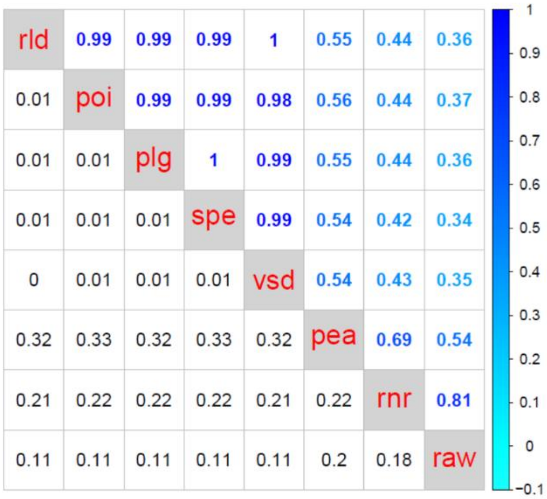


Figure 18 Typical dendrograms obtained for plasmode datasets from MSUPRP experimental data with two dissimilarity measures

Dendrograms using complete linkage based on (a) Poisson dissimilarity (*poi*), (b) Euclidean distance calculated from raw normalized data (*rnr*), and (c) reference dissimilarity based on maximum proportion of shared reads. Original samples are labeled with 3 same letters (AAA, BBB, CCC, DDD or DDD), synthetic samples are labelled with 2 or 3 letters symbolizing the proportion of transcripts, ½ or ⅓ respectively, taken from the original samples. Dendrogram (a) clustered original samples A, B and C and their synthetic combinations in one group, and original samples E and D and their synthetic combinations in another group. The hierarchical structure of each of these two groups represented the degree of shared reads between samples by joining first samples that shared ⅓ of reads and then samples that shared ½ of reads. Dendrogram (b) did not cluster samples according to the expected configuration.



rld	0.99	0.99	0.99	1	0.55	0.44	0.36
0.01	poi	0.99	0.99	0.98	0.56	0.44	0.37
0.01	0.01	plg	1	0.99	0.55	0.44	0.36
0.01	0.01	0.01	spe	0.99	0.54	0.42	0.34
0	0.01	0.01	0.01	vsd	0.54	0.43	0.35
0.32	0.33	0.32	0.33	0.32	pea	0.69	0.54
0.21	0.22	0.22	0.22	0.21	0.22	rnr	0.81
0.11	0.11	0.11	0.11	0.11	0.2	0.18	raw

Figure 19 Agreement between dissimilarity measures using MSUPRP plasmode datasets

The matrix contains means (upper triangle) and standard deviations (lower triangle) of correlation between cophenetic matrices of dendrograms (N= 50 plasmode datasets) for eight dissimilarity measures: Euclidean distances using raw count data (*raw*), Euclidean distances using normalized samples (*rnr*), Euclidean distances using variance stabilizing transformation (*vsd*), Euclidean distances using regularized logarithm (*rld)* 1- Pearson correlation using raw counts (*pea*), 1- Pearson correlation using counts transformed by logarithm of raw counts +1 (*plg*), and 1- Spearman correlation using raw counts (*spe*). We identified the same three sets of dissimilarity measures described before: 1) *raw*, 2) *rnr* and *pea*, and 3) *poi*, *rld*, *vsd*, *plg* and *spe*.

Correlations between hierarchies for the eight dissimilarity measures are summarized in Figure 19. It contains mean and standard deviations, in the upper and lower triangle respectively, of correlations between cophenetic matrices computed on 50 plasmode datasets. As observed in the three scenarios for the Bottomly experiment, dissimilarity measures can be apportioned to three groups: 1) *raw*, 2) *rnr* and *pea*, and 3) *poi*, *rld*, *vsd*, *plg* and *spe*. Dendrograms from *raw* did not agree with dendrograms from other groups. Hierarchies from dissimilarity measures in group 2 presented a medium correlation of cophenetic matrices with high variability (0.69±0.22). Dendrograms from the dissimilarity measures in group 3 presented high correlation values with each other (>0.98) and low variation (<0.01).

The correlation of hierarchies within each of the dissimilarity measures (Table 2, column 5) was low for *raw*, *rnr* and *pea*, whereas clusters were much more consistent (*r*>0.98) for *poi*, *rld*, *vsd*, *plg* and *spe* dissimilarities. Additionally, dissimilarity measures *raw*, *rnr*, and *pea* were poorly correlated with the reference dissimilarity (0.57±0.07, 0.53±0.07, and 0.51±0.04, respectively) while *poi*, *rld*, *vsd*, *plg*, and *spe* were highly correlated with the reference dissimilarity (r>0.8±0.001, see Table 6 in supplementary material). Consequently, dissimilarities raw, *rnr*, and *pea* did not resemble the expected sample structure and resulted in dendrograms that were very inconsistent over repeated sampling of the same dataset. In contrast, dissimilarities *poi*, *rld*, *vsd*, *plg*, and *spe* maintained the sample structure and produced highly reproducible results in hierarchical dendrograms.

3.4 Discussion

Hierarchical cluster analysis is one of the most used techniques for exploring expression patterns in sequencing data (Liu and Si, 2014). In this paper, we showed how to assess the adequacy of dissimilarity measures for clustering samples from RNA-seq experiments by generating plasmode datasets from experimental data.

Plasmode datasets are useful alternatives to parametric simulations for assessing statistical methodologies as data are generated on more realistic conditions and do not depend on a specific parametric model (Gadbury et al., 2008). The algorithm used to build a plasmode dataset depends on the characteristics of the experimental data and the objective of the study. We presented two examples on how to build plasmode datasets from two experiments with different conditions.

The Bottomly dataset had an experimental design with two main sources of variation and used highly inbred individuals. Such context allowed us to generate plasmode datasets with known proportions of differentially expressed transcripts (Figure 14) and focused on assessing the adequacy of dissimilarity measures in recovering the main sources of variation in the hierarchical structure (Figure 16). Analogous plasmode generation algorithms have been used with different objectives, for example to validate differential expression methods for RNA-seq (Reeb and Steibel, 2013), microarray analysis (Gadbury et al., 2008), and qPCR (Steibel et al., 2009), but this is the first time that they are used to assess the properties of sample-based clustering. These procedures are by no means exhaustive of the possible ways of creating plasmodes for clustering. For instance, the algorithm presented in Figure 14 preserves the correlation among genes (Reeb and Steibel, 2013) when generating plasmodes, but other sampling strategies could purposefully select groups of genes with specific correlation patterns. For instance, instead of sampling from DE and nonDE groups, transcripts could be sampled from blocks of co-expressed genes, resulting in more realistic datasets especially if the study is focused on gene-based clustering and co-expression analysis (Si et al., 2014; Rau et al., 2015).

Different from the Bottomly dataset, the MSUPRP did not present any experimental treatment, and individual characteristics were more important. Under these circumstances, we built plasmodes creating synthetic individuals, by combining known proportions of read counts from original individuals and we evaluated the adequacy of dissimilarity measures in resembling the different mixture proportions in the hierarchical sample structure (Figure 18). A similar plasmode generation algorithm was proposed (Vaughan et al., 2009) to evaluate admixture estimation methodologies where the objective is to estimate the proportion of an individual's genome that originates from different founding populations, but using SNP genotypes instead of sequence read counts.

Although the utility of plasmode datasets has been recently highlighted in RNA-seq studies (Reeb and Steibel, 2013; Zhou et al., 2014), only parametric simulations or exemplar experimental datasets have been used to compare dissimilarity measures and clustering methods (Witten, 2011; Ma and Wang, 2012). While extremely useful, parametric simulations are often criticized as being too simplistic to appropriately capture the complexity in gene expression data (Gadbury et al., 2009), thus limiting the scope and validity of the resulting conclusions. On the other hand, using a single exemplar dataset with unknown properties is not an appropriate approach for comparing statistical methods (Mehta et al., 2004). As a partial solution to these limitations, in this paper we show that plasmodes can supplement the evaluation of clustering algorithms by including agreement and consistency measures based on datasets that mimic read count distributions more realistically. One likely criticism of plasmode is that the results may heavily depend on the original dataset (Reeb and Steibel, 2013). However, this does not invalidate their use. Moreover, as shown in this paper, using two very different datasets, some properties of alternative metrics remain consistent, which encourages the use of the plasmode generation methods presented here using alternative datasets.

We built plasmodes to evaluate alternative dissimilarity measures. The selected dissimilarity measures allowed i) the comparison of traditional dissimilarity measures and dissimilarity measures based on discrete count distributions specifically proposed for RNA-seq, and ii) studying the effect of normalization and transformation prior to computing dissimilarities.

The Euclidean distance or the Pearson correlation based dissimilarity computed after transforming data is a routine method adopted from microarray gene expression analysis (Jiang et al., 2004; Liu and Si, 2014). In fact, the Pearson correlation based dissimilarity is equivalent to squared Euclidean distance of standardized data (Hastie et al., 2009). On the other hand, the most common transformations used for RNA-seq are the logarithm of counts, or logarithm of counts plus a constant (Severin et al., 2010), but other variance stabilizing and regularized logarithm transformation functions have been proposed to model the mean-variance relationship of RNA-seq counts (Anders and Huber, 2010; Law et al., 2014; Love et al., 2014b). The Spearman correlation based dissimilarity uses the rank of the read count instead of the counts themselves to compute correlation; consequently it could be applied without transforming the data. Although the use of Spearman correlation based dissimilarity has been discouraged for gene-based clustering of RNA-seq data (Liu and Si, 2014), because it uses a small number of grouped samples to compute ranks, we have used it for sample-based clustering where the number of genes is potentially large enough to obtain more precise ranks. Finally, the Poisson dissimilarity (Witten, 2011) was specifically proposed for clustering of sequencing data based on a Poisson log-linear model of normalized counts, and thus, it is a natural candidate to be included in this comparison.

The eight evaluated dissimilarity measures presented a common pattern of agreement and consistency in recovering the expected sample structure for both plasmode datasets. Dissimilarity measures with high level of agreement between them—correlations between cophenetic matrices with high mean and low standard deviation (Figures Figure 17 and Figure 19)—produce dendrograms with very similar hierarchical structures. However, if dissimilarity measures have correlations of cophenetic matrices with either low mean or high standard deviation (Figures Figure 17 and Figure 19), they generate dendrograms with different hierarchical structures. In addition, correlation between dendrograms obtained with a particular dissimilarity measure summarizes the consistency of such dissimilarity measure. If a dissimilarity measure has a within cophenetic correlation with high mean and low standard deviation, it consistently generates similar dendrograms.

To assess the adequacy of a dissimilarity measure, both agreement and consistency are important. We showed this with *poi*, *rld*, *vsd*, *plg* and *spe* dissimilarities, which presented similar level of agreement in both datasets (Figures Figure 17 and Figure 19). However, dendrograms based on these dissimilarity measures were consistent in the MSUPRP dataset, but showed different consistency under the three scenarios in the Bottomly datasets (Table 2). A counter example is dissimilarity measures *rnr* and *pea* that agreed with each other but were very inconsistent. This means that *rnr* and *pea* tended to reproduce similar clusters on each plasmode, and wide range of dendrograms structures. This has not been reported before, because consistency of clustering under repeated sampling has not been studied in previous works. But a reason for the agreement is that both *rnr* and *pea* are focusing on the same features, because they are essentially normalized Euclidean distances. The reason for the low consistency could be that these measures are expected to behave well with heterogeneous approximately Gaussian data, but not too well with extremely non-Gaussian data. On the other side, the measure *raw* is expected to be better suited for homogenous Gaussian data (all variances are of similar magnitude).

As mentioned before, we used plasmode to study the effect of normalizing data. Normalization is an essential data processing step in RNA-seq analysis that aims at removing systematic biases in order to make consistent comparisons within and between samples (Dillies et al., 2012). Although several methods (Bullard et al., 2010; Anders and Huber, 2010; Robinson and Oshlack, 2010) have been proposed to normalize data, especially for differential expression analysis, the impact of a particular normalization method seemed to be less important in classification and clustering analyses (Witten, 2011). We confirmed this, showing that normalizing counts to equal library sizes was not enough to capture the natural structure of samples when it was the only transformation applied. For instance, dissimilarity measures *raw* and *rnr* had low agreement with dissimilarity measures that resemble better the true structure of data, e.g. *rld*, *vsd*, *spe* or *plg* or *poi* (Figures Figure 17 and Figure 19).

Accounting for the discrete nature of read counts in RNA-seq data is the most important issue to consider when computing dissimilarity measures. For instance, Euclidean distance and

Pearson correlation based measures are known to be influenced by scale, skewness and outliers, thus, they may not work well for count data (Liu and Si, 2014). In support of this, we found that dissimilarity measures *raw*, *rnr*, *pea*, based directly on counts, regardless of normalization or standardization, did not resemble the expected dendrogram and were generally inconsistent. Dendrograms obtained with Pearson based correlation resembled the expected structures only when data were previously log-transformed. However, we found that the Spearman correlation based dissimilarity measure (*spe*) was suitable to represent the natural structure of samples even without normalizing data, possibly because it preserves the relative rank relationships, and it is less influenced by skewness and outliers (Kendall and Gobbons, 1990) when it is based on a large number of genes. The variance stabilizing (*vsd*) and regularized logarithm (*rld*) approaches consistently retrieved the expected dendrogram structure. Both transformations model the mean-variance relationship across all genes to stabilize the variance of counts across samples (Anders and Huber, 2010; Love et al., 2014b). The regularized logarithm transformation also accounts for variation in sequencing depth across samples (Love et al., 2014b). Both functions have been suggested as appropriate transformations for clustering and classification of RNA-seq data with less ambiguous results in hierarchical clustering than using simply log-transformed counts (Love et al., 2014b). Finally, directly using the Poisson dissimilarity (*poi*) generated dendrograms with the expected structure. This is not surprising considering that read counts are usually assumed to fit over dispersed Poisson distributions (Pachter, 2011). Similarly, Witten (Witten, 2011) obtained dendrograms with lower clustering error rates when using the Poisson dissimilarity rather than *vsd* or Euclidean distances on normalized data, but using overdispersed Poisson simulations. Our results are encouraging because we did not use a parametric model to produce similar outcomes.

Sample-based hierarchical cluster analysis can be used as a tool to present results after differential expression analysis or it can be used as an explorative technique for finding patterns in data. In the first approach only informative genes, i.e. differentially expressed genes (called signal in data mining literature) are used while in the second approach informative as well as non informative genes (also known as noise) are utilized (Jiang et al., 2004). As the signal-to-noise ratio (proportion of DE to nonDE genes) is usually less than 1:10 (Jiang et al., 2004), particular methods are applied to diminish the influence of non informative genes that can degrade the reliability of clustering results (Jiang et al., 2004). In RNA-seq analysis, cluster analysis is commonly applied only to differentially expressed genes or a subset of them (Liu and Si, 2014). We have assessed dissimilarity measures under scenarios that include not only a set of differentially expressed transcripts but we also combined differentially and non differentially expressed transcripts (signal-to-noise ratio 1:9 and 1:4), as well as a mixture of individuals. We found that *rld*, *vsd*, *plg*, *spe* and *poi* were highly consistent under all scenarios with a tendency to diminish consistency as the number of non informative genes increases. Although we focused our comparison on the effect of dissimilarity measures on hierarchical clustering results, the same plasmodes could be used to investigate the effect of other decisions made when performing sample-based clustering as the selection of the hierarchical clustering algorithm per-se or even the effect of pre-filtering transcript according to their level of expression (Rau et al., 2013; Bourgon et al., 2010; van Iterson et al., 2010). We did not explore those aspects of sample clustering, but their investigation will be facilitated by the plasmode building strategies described in this paper.

3.5 Conclusion

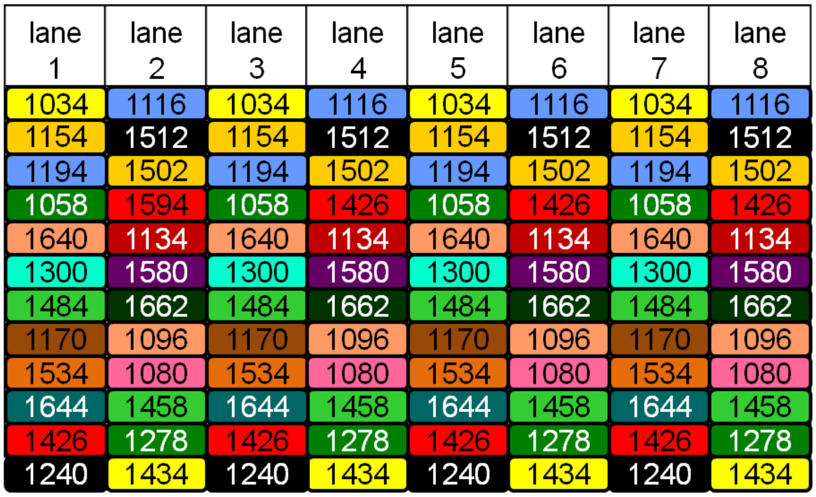
Generating plasmode datasets from experimental data is a reliable tool for evaluating dissimilarity measures in agglomerative hierarchical cluster analysis of RNA-seq data. Depending on the characteristics of the available datasets, several scenarios can be established to compare dissimilarity measures upon a broad spectrum of more realistic conditions than using other simulation approaches. Similar methodologies can be applied to study gene-based clustering as well as other clustering analysis methods.

Explorative sample-based hierarchical clustering of RNA-seq data needs as an input a dissimilarity matrix that accounts for the mean-variance relationship of the discrete nature of read counts. Euclidean distance calculated either on data that have been previously logarithm-transformed or regularized with more complex *ad hoc* functions, as well as model-based dissimilarity for RNA-seq data, were consistent in reproducing the expected sample structure in hierarchical dendrograms.

APPENDIX

Sample preparation, sequencing and mapping

The MSUPRP dataset corresponds to 24 samples of *longissimus dorsi* muscle extracted from 24 F2 females selected from the MSU Pig Resource Population (Steibel et al., 2011). Total RNA was obtained using TRIzol reagent (Invitrogen/Life Technologies, Carlsbad, CA, USA), and RNA quantity and quality were determined using an Agilent 2100 Bioanalyzer (RIN ≥ 7). RNA was reverse transcribed into cDNA, fragmented and labeled to generate 24 barcoded libraries that were sequenced on an Illumina HiSeq 2000 (100 bp, paired-end reads) at the Michigan State University Genomics Core Facility. Four technical replicates were collected from each library and arranged in four different lanes of a flowcell allowing up to 12 barcodes per lane as illustrated in Figure 20.



lane 1	lane 2	lane 3	lane 4	lane 5	lane 6	lane 7	lane 8
1034	1116	1034	1116	1034	1116	1034	1116
1154	1512	1154	1512	1154	1512	1154	1512
1194	1502	1194	1502	1194	1502	1194	1502
1058	1594	1058	1426	1058	1426	1058	1426
1640	1134	1640	1134	1640	1134	1640	1134
1300	1580	1300	1580	1300	1580	1300	1580
1484	1662	1484	1662	1484	1662	1484	1662
1170	1096	1170	1096	1170	1096	1170	1096
1534	1080	1534	1080	1534	1080	1534	1080
1644	1458	1644	1458	1644	1458	1644	1458
1426	1278	1426	1278	1426	1278	1426	1278
1240	1434	1240	1434	1240	1434	1240	1434

Figure 20 Sequencing layout of 24 barcoded samples

Each library has a tag with the sample number and a colour symbolizing its barcode. A same set of 12 tags is repeated 4 times (technical replicates).

The raw read data consisted of 96 pairs of fastq files (4 per sample) containing approximately 15 million short-reads (100bp) each. Those fastq files were pre-processed using FASTXtoolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess read quality. Then, Tophat (Trapnell et al., 2009) was used for mapping the reads to the reference genome (Sus scrofa 10.2.69 retrieved from the Ensembl database) using an index generated by Bowtie2 (Langmead and Salzberg, 2012). The aligned records were stored in BAM/SAM format (Li et al., 2009a). Alignment statistics and base coverage was calculated for each file using SAMtools (Li et al., 2009a). After that, Cufflinks software (Trapnell et al., 2012) was used to obtain gene models and to merge gene models from all samples and reference annotation. Finally, transcript specific read counts were estimated using HTSeq (Anders et al., 2014).

Consistently, about 85% of reads were successfully mapped to reference genome. We detected a total of 26740 transcripts with at least one read aligned. Average coverage per base across 96 pairs of fastq files was 45.79X.

To obtain the final count matrix, we filtered out transcript with zero expression in all samples and merged the 4 technical replicates from each sample. As a result we obtain a count matrix with 26740 transcripts and 24 samples.

Plasmode generation for MSUPRP samples

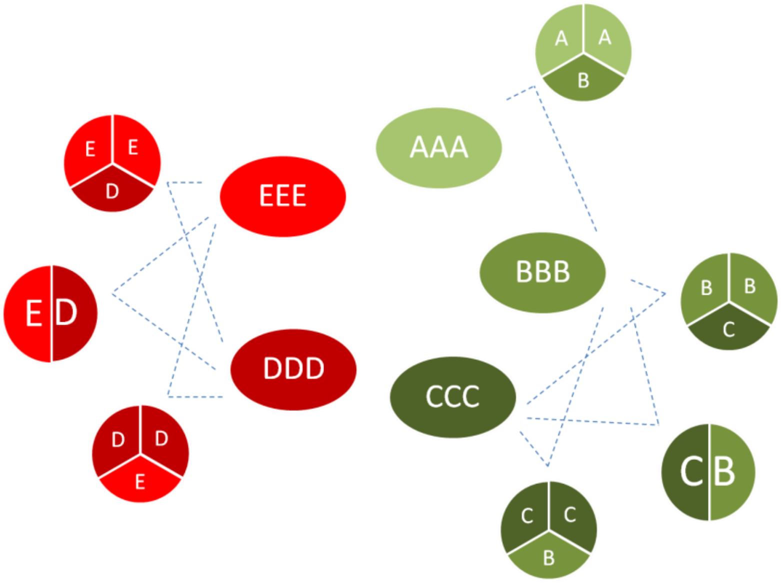


Figure 21 Plasmode generation for MSUPRP dataset

A singular plasmode dataset comprised 12 samples: 5 original samples, ovals labelled as {AAA,BBB,CCC,DDD,EEE}, and 7 synthetic samples, circles labelled as {AAC,BBC,CXB,CCB,DDE,EXD,EED}, obtained by combining known proportions of transcripts (½ , ⅓ , or ⅔) from the original samples.

Reference similarity and dissimilarity matrices for MSUPRP plasmodes

Table 3 Reference similarity matrix (S) for MSUPRP plasmodes

	AAA	AAC	BBB	BBC	CXB	CCB	CCC	DDD	DDE	EXD	EED	EEE
AAA	1.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AAC	0.66	1.00	0.00	0.33	0.33	0.33	0.33	0.00	0.00	0.00	0.00	0.00
BBB	0.00	0.00	1.00	0.66	0.50	0.33	0.00	0.00	0.00	0.00	0.00	0.00
BBC	0.00	0.33	0.66	1.00	0.83	0.66	0.33	0.00	0.00	0.00	0.00	0.00
CXB	0.00	0.33	0.50	0.83	1.00	0.83	0.50	0.00	0.00	0.00	0.00	0.00
CCB	0.00	0.33	0.33	0.66	0.83	1.00	0.66	0.00	0.00	0.00	0.00	0.00
CCC	0.00	0.33	0.00	0.33	0.50	0.66	1.00	0.00	0.00	0.00	0.00	0.00
DDD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.50	0.30	0.00
DDE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	1.00	0.83	0.66	0.33
EXD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.83	1.00	0.83	0.50
EED	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.66	0.83	1.00	0.66
EEE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.50	0.66	1.00

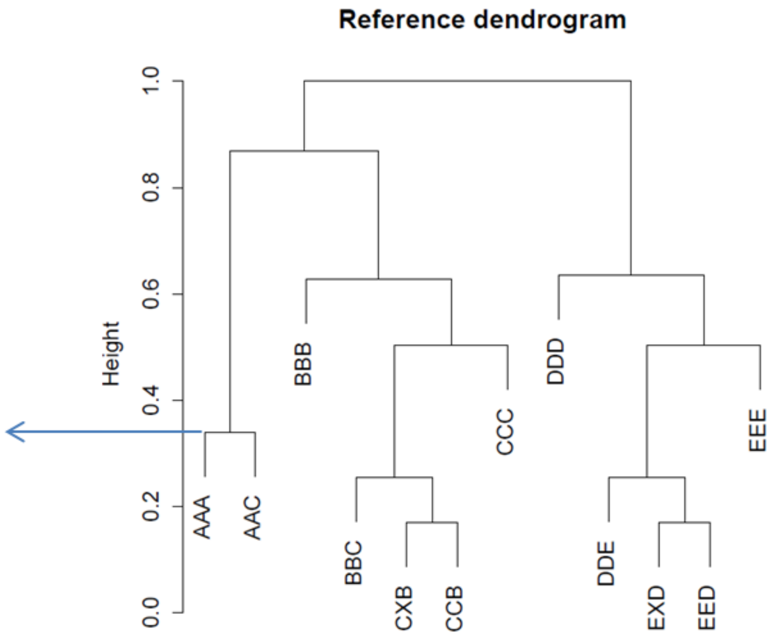
The similarity between two samples (s_{ij}) was calculated as the maximum proportion of original shared reads.

Table 4 Reference dissimilarity matrix (D) for MSUPRP plasmodes

	AAA	AAC	BBB	BBC	CXB	CCB	CCC	DDD	DDE	EXD	EED
AAC	0.34										
BBB	1.00	1.00									
BBC	1.00	0.67	0.34								
CXB	1.00	0.67	0.50	0.17							
CCB	1.00	0.67	0.67	0.34	0.17						
CCC	1.00	0.67	1.00	0.67	0.50	0.34					
DDD	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
DDE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.34			
EXD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.17		
EED	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.70	0.34	0.17	
EEE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.67	0.50	0.34

The dissimilarity between two samples (d_{ij}) was calculated as 1- s_{ij} .

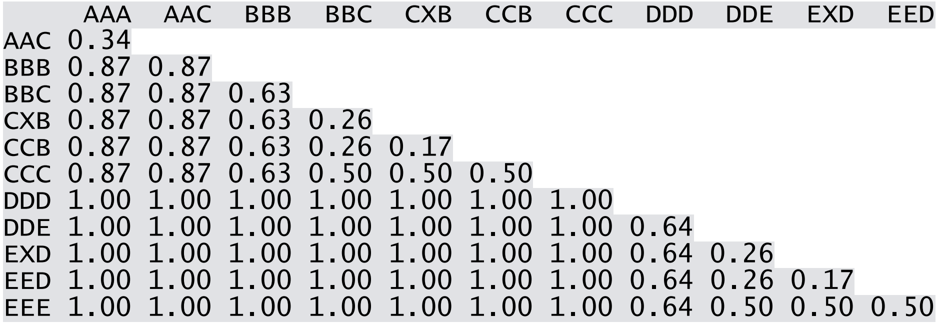
Reference dendrogram and cophenetic matrix for MSUPRP plasmodes



Cophenetic distance
between
AAA and AAC = 0.34

Figure 22 Reference dendrogram for MSUPRP plasmodes

Table 5 Cophenetic matrix of the reference dendrogram for MSUPRP plasmodes.



	AAA	AAC	BBB	BBC	CXB	CCB	CCC	DDD	DDE	EXD	EED
AAC	0.34										
BBB	0.87	0.87									
BBC	0.87	0.87	0.63								
CXB	0.87	0.87	0.63	0.26							
CCB	0.87	0.87	0.63	0.26	0.17						
CCC	0.87	0.87	0.63	0.50	0.50	0.50					
DDD	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
DDE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.64			
EXD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.64	0.26		
EED	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.64	0.26	0.17	
EEE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.64	0.50	0.50	0.50

Correlation between dissimilarity measures and reference dissimilarity for the MSUPRP plasmodes

Table 6 Correlation between dissimilarity measures and reference dissimilarity

Dissimilarity	Correlation with reference dissimilarity
raw	0.57 (0.075)
rnr	0.53 (0.069)
rld	0.83 (0.001)
vsd	0.82 (0.001)
pea	0.51 (0.045)
plg	0.82 (0.001)
spe	0.81 (0.001)
poi	0.81 (0.01)

Mean and standard deviation of correlation between each of the eight dissimilarity matrices and the reference dissimilarity for MSUPRP plasmodes (N=50 plasmode datasets). Euclidean distances using raw count data (*raw*), Euclidean distances using normalized samples (*rnr*), Euclidean distances using regularized logarithm (*rld*), Euclidean distances using variance stabilizing transformation (*vsd*), 1- Pearson correlation using raw counts (*pea*), 1- Pearson correlation using counts transformed by logarithm (plg), and 1- Spearman correlation using raw counts (*spe*), and Poisson dissimilarity (*poi*). Distances measures *raw*, *rnr*, and *pea* poorly preserved the expected sample structure (r<0.57) while *poi*, *rld*, *vsd*, *plg*, and *spe* highly preserved the expected sample structure (r>0.8).

MSUPRP dataset. This file contains raw count data (6 columns x 25798 rows) for the five animals used to generate the MSUPRP plasmode datasets. Available online at PLoS ONE website.

Chapter 4

Assessing genotype call accuracy from RNA sequencing data

4.1 Introduction

RNA sequencing technology (RNA-seq) (Wang et al., 2009) has become the platform of choice for gene expression profiling (Oshlack et al., 2010). In addition to quantifying total gene expression measures, RNA-seq experiments can be used for calling SNP genotypes to investigate allele-specific expression (ASE) (Pirinen et al., 2015; Steibel et al., 2015) , and for studying RNA-editing (Ramaswami et al., 2013). Another proposed use of genotypes called from RNA-seq data is for performing population genetics studies in nonmodel organisms (De Wit et al., 2012; Schunter et al., 2013; Lemay et al., 2013; Yu et al., 2014) and for performing genetic associations with phenotypes (Cánovas et al., 2013). These experiments, which study and quantify variation in transcripts nucleotide sequence rely on accurately calling genomic variants and genotyping individuals. Moreover, calling variants from RNA-seq data poses challenges that are not present in variant calling from DNA sequence (Piskol et al., 2013; Quinn et al., 2013). For instance, alternative splicing (Piskol et al., 2013; Quinn et al., 2013) and ASE (Steibel et al., 2015) may substantially complicate genotype calling from RNA-seq data. Another difference between RNA-seq and DNA-seq data is that the sequence coverage of certain genes across samples may differ substantially, due to gene-specific and subject to subject variation in gene expression.

Despite these challenges, identification of genetic variants and calling genotypes from RNA-seq data is commonly performed using tools developed for whole genome sequencing (WGS) of DNA samples (Altmann et al., 2012; Nielsen et al., 2011). Consequently, the performance of such algorithms have to be studied separately for RNA-seq experiments.

Assessing the performance of variant calling programs from RNA-seq data is usually done by comparing the called genotypes to a gold standard such as genotypes from a SNP chip (Djari et al., 2013; Cánovas et al., 2013) or from WGS (Piskol et al., 2013; Quinn et al., 2013) and estimating genotyping accuracy. But in the context of calling SNP genotypes from RNA-seq data, estimating genotyping accuracies separately for heterozygous and homozygous is important due to the different use of those genotypes. For instance: false heterozygous genotypes (a true homozygous site that is called heterozygous in RNA-seq data) will likely lead to biases in estimating ASE (Steibel et al., 2015; Pirinen et al., 2015) and they will lead to false positives in RNA editing studies. On the other side, false homozygous genotypes (true heterozygous sites called homozygous) will likely result in lack of power to detect ASE and RNA editing. In general, notable exceptions (Quinn et al., 2013), this issue has been ignored, and usually omnibus genotype calling accuracies or error rates are defined and reported.

Calling reliable SNP genotypes may have other applications such as conducting population genetics studies for estimating effective population size, studying population substructure, and performing genetic associations with ecologically or economically important phenotypes. In those cases, estimating SNP- specific genotyping call accuracies and error rates may not be relevant. For those cases it is more important to correctly estimate the relatedness between individuals. For instance, reconstructing genomic relationship matrices (VanRaden, 2008) with SNPs called from RNA-seq data and comparing them to a reference relationship matrix would definitely be more informative for animal breeding applications (Habier et al., 2007), as well as for population history and demography studies of nonmodel organisms (Ekblom and Galindo, 2011).

In this work, we compared sensitivity and false discovery rates for SNP calling using two well-established variant calling programs BCFtools (Li, 2011), and VarScan2 (Koboldt et al., 2012). We also used Beagle (Browning and Browning, 2007) to impute genotype calls from BCFtools, as previously recommended (Nielsen et al., 2011; Li et al., 2013). We apply these algorithms to RNA-seq data from a pig resource population that has been extensively genotyped and phenotyped. We report homozygous and heterozygous genotype call rates and we estimate error rates by comparing called genotypes against genotypes obtained from SNP chip array and DNA sequencing.

4.2 Material and Methods

4.2.1 Resource population

The Michigan State University pig resource population (MSUPRP) is an F_2 cross originated from 4 F_0 Duroc sires and 16 F_0 Pietrain dams. The full pedigree includes 20 F_0, 56 F_1 and 954 F_2 animals. This population has been described in detail elsewhere (Edwards et al., 2008). For this study we selected 24 F2 females that presented extreme phenotypes in loin eye area (Steibel et al., 2011).

Total RNA was extracted from *longissimus dorsi* muscle of the 24 F2 females using TRIzol reagent (Invitrogen/Life Technologies, Carlsbad, CA, USA), and RNA quantity and quality were determined using an Agilent 2100 Bioanalyzer (RIN ≥ 7). RNA was reverse transcribed into cDNA, fragmented and labeled to generate 24 barcoded libraries that were sequenced on an Illumina HiSeq 2000 (100 bp, paired-end reads) at the Michigan State University Genomics Core Facility. Four technical replicates were collected from each library and arranged in four different lanes of a flowcell allowing up to 12 barcodes per lane. The raw read data consisted of 96 pairs of fastq files (4 per sample) containing approximately 15million short-reads (100bp) each. Those fastq files were pre-processed using FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess read quality. Then, Tophat (Trapnell et al., 2009) was used for mapping the reads to the reference genome (Sus scrofa 10.2.69 retrieved from the Ensembl database) using an index generated by Bowtie2 (Langmead and Salzberg, 2012). The aligned records were stored in BAM/SAM format

SNP chip genotypes were obtained with the SNP60K Illumina chip (Gualdrón Duarte et al., 2013) and used as gold standard to compare genotypes called from mRNA sequencing data.

In addition to the SNP chip genotypes and muscle tissue RNA-seq data, genomic DNA sequence, liver and fat tissue RNA-seq data were available for one of the 24 animals. Genomic DNA was purified from white blood cells using the Invitrogen Purelink Genomic DNA Mini Kit. Library preparation was performed with the Illumina TruSeq Nano DNA Library Preparation Kit HT and sequenced on 2 lanes of an Illumina HiSeq 2500 Rapid Run flow cell and sequenced in a 2x150bp (PE150) configuration Rapid SBS reagents. Reads were trimmed with Condetri (Smeds and Künstner, 2011) and aligned using Bowtie 2 (Langmead and Salzberg, 2012). For this animal, RNA-seq was performed slightly differently from the other 23 pigs. Total RNA was extracted from subcutaneous fat, liver and *longissimus dorsi* muscle tissue. The samples were prepared for sequencing using the Illumina TruSeq Stranded mRNA Library Preparation Kit. After validation and quantitation the libraries were pooled and loaded on both lanes of an Illumina HiSeq 2500 Rapid Run flow cell (v1). Base calling was performed by Illumina Real Time Analysis (RTA) v1.18.61 and output of RTA was demultiplexed and converted to FastQ files with Illumina Bcl2fastq v1.8.4.

Reads were trimmed with Condetri (Smeds and Künstner, 2011) and aligned to the reference genome (Sus scrofa 10.2.69) with Tophat (Trapnell et al., 2009). The BAM files were merged and unique alignments were extracted and separated by strand.

4.2.2 Variant calling

First, base alignment files were obtained using the mpileup command of SAMtools version 1.0 (Li et al., 2009a) using options -Q20 -C50 -B of mpileup to eliminate reads with quality scores below 20, to downgrade mapping quality for reads with excessive mismatches, and to disable the calculation of base realignment quality in order to diminish false SNPs generated by misalignments. For the single animal with mRNA and DNA sequence data, reads with quality scores below 25 were filtered and the coefficient for downgrading mapping quality with excessive mismatches was applied, but base realignment quality was allowed (option -Q25 -C50 -E). These options have been used before for the study of study RNA editing (Chen et al., 2014).

The multiallelic calling version of BCFtools was used (bcftools call -vm) to call genotypes, while to call SNP with VarScan2, the mpileup2snp command was used with default settings (a minimum of 8 reads to make a call, with at least 2 supporting reads to call a variant, a minimum average base quality of 15, and a p-value threshold of 0.01). Finally, Beagle (Browning and Browning, 2007) was used to refine genotype calls from BCFtools using genotype imputation. Beagle used as input vcf files obtained by BCFtools. We used default parameters for Beagle version 4.0 (java -Xmx4000m -jar beagle.r1398.jar gl=bcfvariants).

4.2.3 Assessing genotype call accuracy

SNP genotypes obtained with each variant calling program were compared to genotypes obtained from a DNA SNP60K Illumina chip to estimate genotype call rates. For each SNP genotyped with both technologies, we built a contingency table (Table 7) and we computed several genotyping accuracy and error measures. For instance: true homozygous called homozygous as well as true heterozygous called heterozygous contribute to the accuracy of the variant calling method and they are the basis to estimate sensitivity. On the other hand, true homozygous called heterozygous and true heterozygous called homozygous contribute to genotyping errors.

Table 7 Contingency table with classification rule used to compute sensitivity and error rate of genotype calling

RNA-seq Variant calling Genotype	SNP chip Genotype		Total
	Heterozygous	Homozygous	
Heterozygous	True He[1]	False He[2]	Called He[3]
Homozygous	False Ho[4]	True Ho[5]	Called Ho[6]
Non Called	NC He[7]	NC Ho[8]	NC[9]
Total	He[10]	Ho[11]	

[1] Number of heterozygous genotyped in the SNP chip called heterozygous from RNA-seq;
[2] Number of homozygous genotyped in the SNP chip called heterozygous from RNA-seq;
[3] Total Number of called heterozygous from RNA-seq;
[4] Number of heterozygous genotyped in the SNP chip called homozygous from RNA-seq;
[5] Number of homozygous genotyped in the SNP chip called homozygous from RNA-seq;
[6] Total Number of called homozygous from RNA-seq;
[7] Number of heterozygous genotyped in the SNP chip non-called from RNA-seq;
[8] Number of homozygous genotyped in the SNP chip non-called from RNA-seq;
[9] Total number of non-called genotypes from RNA-seq;
[10] Total number of heterozygous genotyped in the SNP chip
[11] Total number of homozygous genotyped in the SNP chip

To assess the performance of the variant calling methods, we computed true genotype rates (sensitivity) and false discovery rates (FDR) for each genotype as:

$$Sensitivity_{He} = \frac{True\ He}{He}$$

$$Sensitivity_{Ho} = \frac{True\ Ho}{Ho}$$

$$FDR_{He} = \frac{False\ He}{Called\ He}$$

$$FDR_{Ho} = \frac{False\ Ho}{Called\ Ho}$$

4.2.4 Equivalence between relationship matrices and correlation with SNP chip data

The marker-based relationship matrix among individuals (VanRaden, 2008) was calculated using genotypes from the 60K SNPchip ($\boldsymbol{G}_{SNPchip}$) and compared to the marker-based relationship using genotypes from RNA-seq ($\boldsymbol{G}_{RNA-seq}$). In order to compute $\boldsymbol{G}$ we selected SNPs called from RNA-seq or from the chip and built a genotype matrix $\boldsymbol{Z}$ containing standardized allelic dosages $Z_{ij} = \frac{g_{ij}-E(g_{ij})}{\sqrt{var(g_{ij})}}$, where $g_{ij}\{0,1,2\}$ are the allelic counts of the reference allele and its expectation and variance are computed assuming Hardy Weinberg equilibrium across SNP (VanRaden, 2008). Finally, following Van Raden (2008), $\boldsymbol{G} = \boldsymbol{ZZ'}\frac{g_{ij}-E(g_{ij})}{\sqrt{var(g_{ij})}}$

To generate $\boldsymbol{Z}_{RNA-seq}$, we used SNP called by BCFtools with at least 240 reads per base across the 24 animals. This level of total depth ensured that both heterozygous and homozygous genotypes were accurately called (see results section). Additionally, only monomorphic (minor allele frequency > 0.0) sites in autosomic chromosomes with genotyping data in all 24 samples were considered. After applying all the filters a total of 125,277 SNPs were used to compute the $\boldsymbol{G}_{RNA-seq}$. Finally, the maximum correlation between SNPs in the SNPchip and SNPs within 1 Mb distance in RNA-seq data were calculated to assess the redundancy of the genotype set obtained with RNA-seq with respect to the SNP set in the 60K chip.

4.3 Results

4.3.1 Analysis across variant calling programs for 24 animals

Sensitivity curves presented similar patterns for heterozygous and homozygous genotypes, but with specific quantitative differences. Figure 23a-b shows the sensitivity for calling heterozygous and homozygous genotypes, respectively, for three variant calling programs as a function of inverse cumulative read depth. Sensitivity for calling heterozygous genotypes was higher with BCFtools compared to VarScan2 for sequencing depth below 1000 reads per base (across all 24 samples). Using Beagle to impute missing genotypes called by BCFtools provided a slight increment in sensitivity of heterozygous genotype calling at the expense of FDR as we describe later. Regardless of the variant calling program, heterozygous genotypes were called with sensitivity > 94% when base read depth was equal or higher than 1000. When calling SNP genotypes at homozygous sites, the sensitivity was usually higher than corresponding to heterozygous sites, especially for sites covered by less than 1000 reads. Moreover, the reported advantage in sensitivity of BCFtools compared to VarScan2 for calling genotypes at heterozygous sites was still present but it was lower for homozygous sites. Consistently, imputing non-called genotypes using Beagle improved the sensitivity of homozygous calling with respect to BCFtools even more markedly than for heterozygous sites. Similarly to heterozygous, calling homozygous in sites covered by over 1000 reads resulted in sensitivity > 98%, regardless of the variant calling program.

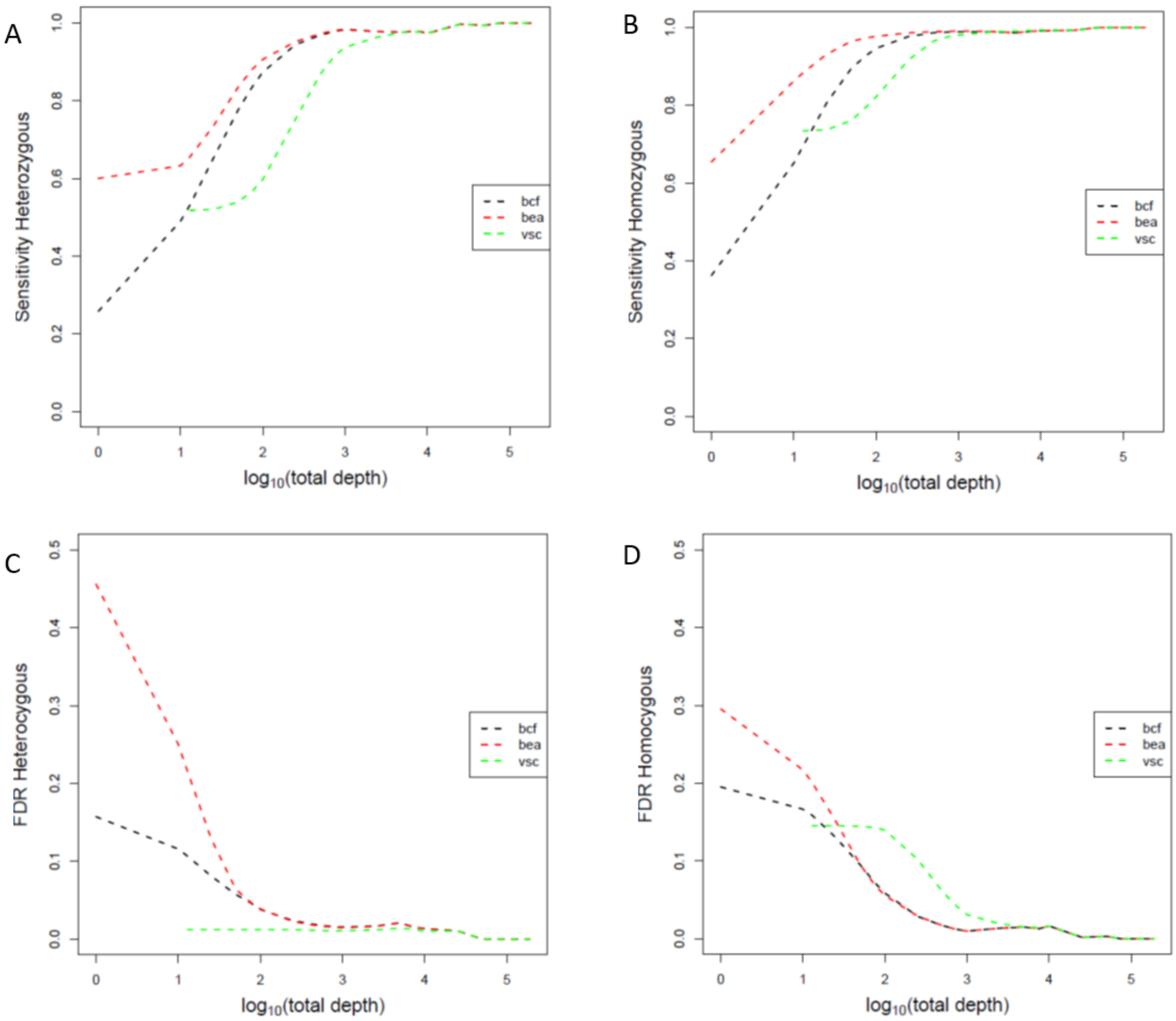


Figure 23 Assessing SNP calling accuracy of RNA-seq samples of *longisimus dorsi*

Sensitivity analysis for heterozygous **(A)** and homozygous **(B)**. False discovery rate for heterozygous **(C),** and homozygous **(D)**. Raw read depth in the x axis is inversely cumulative in the $\log_{10}$ scale. Varcalling software correspond to: bcf=BCFtools, bea=Beagle, and vsc=VarScan2

False discovery rates curves presented more distinct patterns between heterozygous and homozygous genotypes. Figure 23c-d shows the FDR for heterozygous and homozygous calls, respectively, for each of the three variant calling programs as a function of inverse cumulative read depth. Heterozygous sites called with VarScan2 were consistently correct regardless of the coverage depth (FDR very close to 0). However this comes as a price of very low sensitivity. For example, in sites covered by 10 to 100 reads across all 24 samples, the sensitivity of VarScan2 was 56% on average (Figure 23a). Contrastingly, at that same depth range, FDR with BCFtools was between 12% and 4%, but sensitivity was between 50% and 87%. At high sequencing depth, differences in FDR vanish and all programs presented FDR < 2% when base read coverage was equal or higher than 1000. As mentioned before, imputation with low coverage increased sensitivity at the cost of FDR especially for sites with less than 100 reads per base. For instance, the use of Beagle increased the FDR from 4% to 21% when coverage dropped below 10 reads per base. The proportion of incorrectly called homozygous was generally higher than the proportion of incorrectly called heterozygous (Figure 23d). Only sites with more than 1000 reads per base presented a FDR of homozygous below 3%. BCFtools had lower error rates than VarScan2 for base read depth between 10 and 1000. The use of Beagle increased the FDR for homozygous particularly for sites with very low coverage (less than 10 reads per base), and the increase in FDR was less than when imputing heterozygous sites.

4.3.2 Comparison of genotype calls from DNA-seq and RNA-seq from multiple tissues

Comparing SNP calling from RNA-seq to DNA-seq data, genotypes called from DNA-seq data showed uniformly high sensitivity (sensitivity > 0.99) and low FDR (2%). Conversely, error rates from genotypes called from RNA-seq were influenced by sequencing depth. Sensitivity for heterozygous sites called from RNA-seq data varied with base read depth while sensitivity for homozygous sites was always high (>96%) regardless of base read depth (Figure 24a-b) and comparable to the sensitivity of homozygous genotypes called from DNA-seq. Sensitivity of heterozygous genotype calls using RNA-seq for any of the tissues was below 90 % for sites with less than 10 reads per base and almost constant (94%-98%) for sites with more than 10 reads per base. A coverage of 10 reads per base in one animal is on average equivalent to 240 reads across 24 animals, at which point BCFtools provided a sensitivity of 94% across all samples (Figure 23a).

False discovery rates for heterozygous sites were below 2% regardless of base read depth either for DNA or RNA samples. Different patterns presented when analyzing homozygous sites. For homozygous sites, variant calling using RNA-seq incorrectly genotyped a large proportion of sites (between 59% and 10%) that were covered by 10 or fewer reads per base. Homozygous sites with more than 10 reads per base presented lower false discovery rates for RNA-seq samples from *longissimus dorsi* than from liver or fat tissue. FDR for homozygous was always below 1% when using the DNA sample.

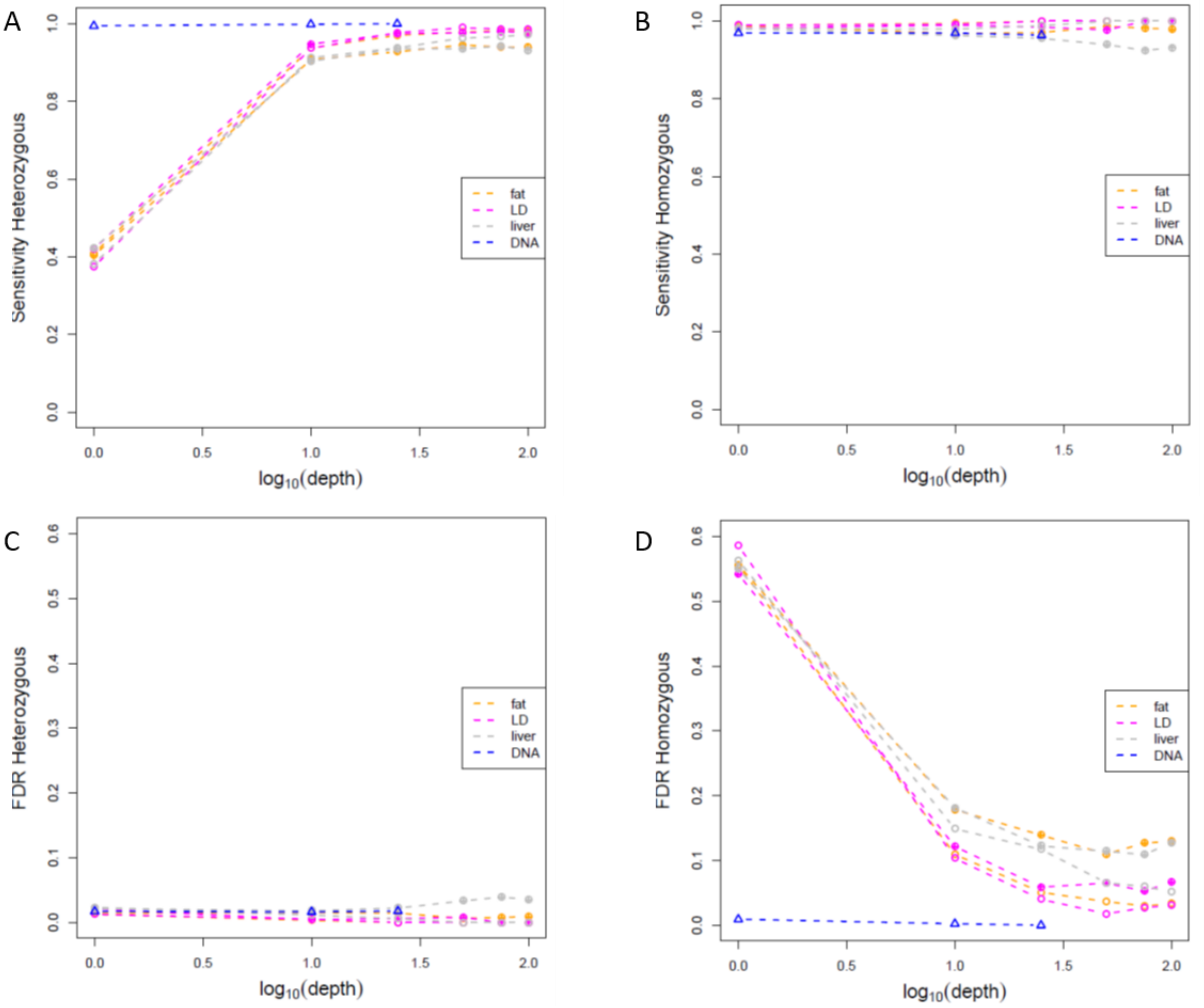


Figure 24 Assessing SNP calling accuracy of DNA and RNA-seq samples of three tissues from the same animal

Sensitivity analysis for heterozygous **(A)** and homozygous **(B)**. False discovery rate for heterozygous **(C),** and homozygous **(D)**. Raw read depth in the x axis is inversely cumulative in the $\log_{10}$ scale. Varcalling software correspond to: bcf=BCFtools, bea=Beagle, and vsc=VarScan2

4.3.3 Equivalence between relationship matrices

Correlation between relationship matrices $\boldsymbol{G}_{SNPchip}$ and $\boldsymbol{G}_{RNA-seq}$ (based on SNPchip markers and based on RNA-seq markers, respectively) was 0.99 (Figure 25). This value indicated that both matrices were virtually equivalent and either of them could be used to account for the relationship among the animals. In depth analysis of the correlation results showed that correlation between the diagonals of both relationship matrices was 0.66 and correlation between off-diagonals was 0.95 (Figure 25). These results are of interest to distinguish the use of the relationship matrices to compute genomic inbreeding coefficient from the diagonals or to compute genomic relationships between individuals from the off-diagonals (VanRaden, 2008).

An average of 110 SNPs called from RNA-seq were within the 1 Mb window defined for each SNP in the SNP60K chip and only nine percent of SNPs in the SNP 60K chip did not have any SNP called from RNA-seq. The median correlation between SNPs in the SNP60K chip and SNPs called from RNA-seq was 0.71 and increased to 0.78 for SNPs in the SNP60K chip with at least one SNP called from RNA-seq within the 1 Mb window. This result implied that on average the information contained in all the SNPs in the SNP60K chip could be replaced by information on SNPs called from RNA-seq.

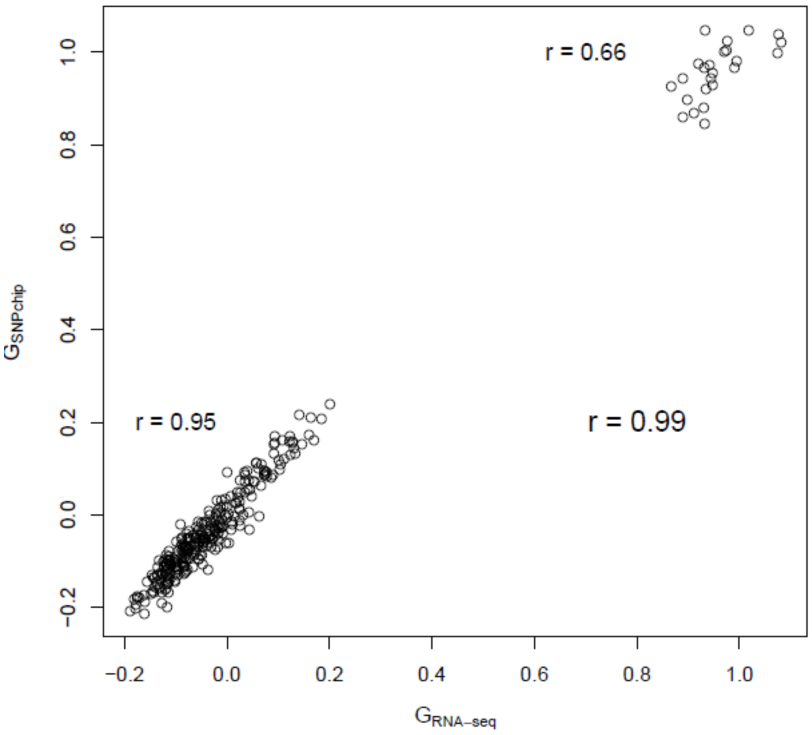


Figure 25 Scatterplot of correlation between relationship matrices for the 24 F2 females

The scatterplot shows a high equivalence (r=0.99) between the relationship matrix based on SNPchip genotypes ($\mathbf{G}_{\text{SNPchip}}$) and the relationship matrix based on SNPs genotypes called from RNA-seq sites ($\mathbf{G}_{\text{RNA-seq}}$) with depth >=240 reads per base. Correlation between relationship matrices for the same animal (diagonal values of the relationship matrices) correspond to values in the right top quadrant (r=0.99) while correlation between relationship matrices for the different animals (off-diagonal values of the relationship matrices) are represented in the left bottom quadrant (r=0.95)

4.4 Discussion

This work assessed genotype calling from RNA-seq and DNA-seq data. We compared genotypes called by two variant calling programs to genotypes obtained from a DNA SNP60K chip by computing sensitivity and false discovery rates separately. Accuracy of SNP calling varied between programs for different genotypes. Based on these results we make suggestions for implementing genotype calling from RNA-seq data, depending on its intended use, for instance when using heterozygous genotypes for allele-specific expression studies or when using homozygous for exploring RNA editing.

Identifying heterozygous SNP is the first step to conduct allele-specific expression studies (Pickrell et al., 2010). Heterozygous sites provide information to classify reads according to their haplotype transcript of origin and to quantify the allelic expression ratio (Lee et al., 2013). Two major consequences from genotyping error in ASE are:1) the omission of calling a heterozygous site, or 2) the incorrect assignment of a heterozygous genotype to a true homozygous genotype. In the case of missing a heterozygous call the direct consequence will be a loss in power, while when mislabeling the site the effect could be biases in estimation of the ASE ratio (Steibel et al., 2015). Quinn et al. (2013) found that sensitivity for heterozygous varied between 40% to 80% at coverage depth below 10X per sample, and it reached 90% for coverage depth above 10X per sample. Similarly, we found that for the single sample analysis more than 90% of heterozygous were called in all tissues when depth was above 10 reads per base (Figure 24A). For the 24 multiple samples, the same proportion (>90%) was obtained using BCFtools for an equivalent total depth of 240 reads or more per base across all samples (24 samples x 10X per sample, Figure 23a). However, VarScan2 called only 70% of heterozygous sites at the same depth of 240 reads per base. Interestingly, false discovery rates for heterozygous genotypes at 10 reads per base per sample, or equivalently a multi-sample total depth of 240 reads per base, were below 2%. Consequently, calling heterozygous from RNA-seq data with more than 10 reads on average per base and per sample can provide a good number of sites to conduct ASE studies with sufficient power and unbiased estimations of the allelic expression rate. Additionally, BCFtools and Beagle provided a larger set of true heterozygous genotypes (e.g. an average of 40% more SNP compared to VarScan2), due to increased sensitivity and to more stringent filtering criteria implemented in VarScan2. Yet, if a variant is called heterozygous by VarScan2, it will be likely be a true heterozygous as the FDR is very low regardless of the depth (Figures Figure 23c and

Figure 24c). However, more caution should be considered when using BCFtools and Beagle in sites with low depth (Figures Figure 23c and Figure 24c).

Homozygous sites provide an initial list of candidate sites for studying RNA editing (Li et al., 2009b; Chen et al., 2014). The typical approach to study RNA editing compares the mismatches between homozygous DNA and RNA sequences (Peng et al., 2012; Ramaswami et al., 2012) from a single individual. Besides identifying the type of editing process, e.g. A-to-I editing, it is possible to quantify editing levels in an analogous way as in ASE studies (Bahn et al., 2012; Lee et al., 2013). Therefore, similar implications as discussed above when analyzing sensitivity and FDR for calling heterozygous also applies when analyzing homozygous. Calling genotypes from DNA-seq is reliable and has been studied in detail (Nielsen et al., 2011; Li, 2014; Li et al., 2013; Li, 2011; Kumar et al., 2014). We confirmed previously published results for instance, when using DNA-seq, we found that sensitivity for homozygous was > 0.99 (Figure 24b) with a corresponding FDR < 2% (Figure 24D) irrespective of depth. Similar values and patterns were found for heterozygous genotypes (Figure 24a-b). However, calling genotypes from RNA-seq protocols is subject to diverse sources of false genotyping error, for instance: increased mapping errors due to the complexity of the transcriptome compared to the genome (Bass et al., 2012). For that reason, stringent mapping and filtering options are applied particularly when analyzing nucleotide variation profiles from RNA-seq (Quinn et al., 2013; Piskol et al., 2013; Lee et al., 2013). In this work, we found that calling homozygous from RNA-seq presented high sensitivity (>95%) in all tissues and irrespective of depth (Figure 24b), but it could result in a high number of false positives (Figure 24D). Similar results for calling homozygous were found in Quinn et al. (Quinn et al., 2013). Consequently, depth of RNA-seq reads should be considered when analyzing homozygous sites using RNA-seq. This is usually not considered in RNA-editing studies where protocols for including editing sites emphasize requirements in DNA depth but not in RNA-seq depth (Chen et al., 2014; Lee et al., 2013). Adding a depth filter to RNA-seq reads could help to lower the number of false positives in addition to other filtering strategies that are currently used (Chen et al., 2014; Lee et al., 2013).

Imputation of genotypes using Beagle after calling variants with BCFtools neither improve sensitivity nor worsen FDR in either heterozygous or homozygous sites covered by more than 100 reads across 24 samples (Figure 23). At lower coverage, the effect was detrimental as imputation provided little increase in sensitivity at the cost of expanding FDR in both genotypes (Figure 23c-d). These F2 pigs were sampled from an admixed population originated from two different founding populations, thus it is expected long range persistence of linkage disequilibrium (Vaughan et al., 2009) and heterogeneity of haplotype blocks (Gualdrón Duarte et al., 2013) in comparison to a panmictic population. Consequently, imputation based only on linkage disequilibrium may not result in high imputation accuracy (Gualdrón Duarte et al., 2013). However, if RNA-seq data alone or a mixture of DNA-seq and RNA-seq data were available across multiple generations, pedigree based imputation could result in a more accurate imputation (Gualdrón Duarte et al., 2013). We did not explore this aspect in this paper due to small sample size and lack of sequence genotypes in ancestors of the F2 pigs.

Accurate genotypes obtained by RNA-seq can be used to estimate genetic relationships between individuals in a similar way to using genotypes from a SNP60K chip (Figure 25). Even though the SNP sets from the chip and the RNA-seq analysis did not completely overlap, we found that they are highly correlated and they represent the relationships among individual in a very similar way (Figure 25). We know from previous analyses that the SNP60K chip represents the relationships among individuals in this population with high accuracy (Gualdrón Duarte et al., 2013). Additionally, RNA-seq derived SNPs may be used to estimate relationships in populations that are not well represented by SNPs in the SNP60K chip to avoid problems related to ascertainment bias. In nonmodel organisms of little agricultural interest SNP chips are generally not available. In this situations, RNA-seq data could be used to estimate relationships among individuals (Seeb et al., 2011; Helyar et al., 2011; Weinman et al., 2014) or to complement genotyping by sequencing studies (Narum et al., 2013), and to improve genome-wide association studies by supplementing randomly selected SNP sets with more likely functional SNP sets (Cánovas et al., 2013).

Chapter 5

General discussion

5.1 Introduction

RNA sequencing (RNA-seq) (Wang et al., 2009) has emerged as a revolutionary technology to study transcriptomes. Both quantitative and qualitative aspects are being intensively studied in a broad spectrum of biological applications, such as gene differential expression analysis and single nucleotide variation profiling in both model and nonmodel organisms (Wickramasinghe et al., 2014; Qian et al., 2014). Simultaneously, a number of methods are being proposed for analyzing RNA-seq data in each of those applications (Nookaew et al., 2012). Moreover, methods are being periodically updated and new algorithms are being continuously published to improve the analyses (Seyednasrollah et al., 2013; Kvam et al., 2012). As a consequence, researchers often find themselves having to decide between competing models and assessing the reliability of results obtained with a designated analysis program. Choosing the most appropriate software can help to get better results and to achieve the goals of the investigation (Mehta et al., 2006). The overarching goal of this dissertation was to propose and implement validation procedures based on experimental data to estimate the properties of widely used statistical analysis methods of RNA-seq experiments. Using synthetic and reference datasets, I compared statistical models to perform differential expression analysis, sampled-base hierarchical cluster analysis, and variant calling and genotyping.

5.2 Objectives revisited

1. To evaluate statistical models for differential expression analysis in RNA-seq experiments

Evaluation of statistical models for differential expression analysis has been based on parametric simulated datasets using count data distributions, such as Poisson and negative binomial. In this dissertation, I provided a more comprehensively approach using plasmodes from experimental datasets as well as parametric simulations. Two methods based on negative binomial (edgeR and DESeq) presented higher type I error rates than a transformed Gaussian model (MAANOVA) despite the use of simulations or plasmodes. The complement between the two types of reference datasets was particularly useful when comparing the methods. Since parametric simulation can benefit a model that is based on the same distribution, the use of plasmodes provided an independent reference to supplement the analysis.

Additionally, I presented the comparison of models by exploring the joint null distribution of p-values that is expected to resemble a uniform distribution. The MAANOVA program used to fit the transformed Gaussian model can generate p-values using different strategies, such as using moderated test or transcript-by-transcirp test. The best approach to compute the p-values differed between datasets. The comparison of the p-value distributions helped to decide on the best set of specifications to use in each dataset. Usually, this measure of comparison is not reported when evaluating the adequacy of a model. However, researchers typically correct p-values by FDR as proposed by Benjamini and Hochber (Benjamini and Hochberg, 1995) and such correction may not be appropriate if the uniform distribution is not supported, thus, leading to wrong decisions when comparing treatments.

Overall, Gaussian models had p-values closer to nominal significance levels and presented p-value distributions closer to the uniform distribution. Researchers using models with these characteristics will have less false positive when distinguishing differentially expressed transcripts. In addition, Gaussian transformed model can include random effects and hierarchical structures that arise in complex experiments, such as when collecting data in fisheries or wildlife studies. For instance, landscape genomics has been proposed as a framework for studying adaptive and neutral genetic variation at the population level in a spatially explicit context (Joost et al., 2007), however, flexible models that can account for environmental and spatial autocorrelations are needed (Schoville et al., 2012). Similar to macroscale biology applications, more complex analysis methods are also required at the molecular scale for analyzing gene families or regulatory networks that need to account for time course and gene-set correlations (Yang and Wei, 2015; Emmert-Streib, 2013).

Additionally, the proposed plasmode algorithm is an alternative to previous implemented plasmodes. The algorithm allows incorporation of biological variation by making several partitions from the original set of samples instead of only one as implemented in Gadbury et al. (2008). This way of creating plasmodes can be of interest in field studies where the individual variability can be studied *a priori* to improve experimental designs aspects such as sample size. In conclusion, I proposed and implemented a general validation method that was applied to specifically compare count data based models versus a Gaussian transformed data model. Furthermore, the procedure allowed to set the best analysis option for the Gaussian transformed model according to specific datasets, and presented other uses of the information generated by the validation process that can help researchers to plan data collection.

2. To assess the properties of dissimilarity measures for agglomerative hierarchical clustering of RNA-seq data

Hierarchical cluster analysis is one of the most used techniques for representing pattern recognition and representation of results from sequencing data (Liu and Si, 2014). Arguably, selection of a distance or similarity metrics is one of the most important decision in the implementation of hierarchical clustering (Izenman, 2008). As discussed for differential expression methods, the validity of alternatives metrics for hierarchical clustering has been previously assessed using parametric simulations. In this dissertation, I presented two ways for creating plasmode datasets for assessing the suitability of distance metrics for clustering RNA-seq expression data. For that, I used data from two experiments conducted under different conditions. One experiment was the result of performing RNA-seq from animals in a pre-designed experiment based on two inbred strains of mice, while the other experiment consisted in sequencing the transcriptome of randomly sampled (untreated) individuals from an outbred population. The common goal of performing clustering in both datasets was to uncover population structure due to genetic differences in gene expression (differences between breeds of mice or differences within a crossbred population). In both situations, the proposed plasmode building strategy allowed for individual-to-individual variability and it extended the validation of cluster analysis for datasets with different experimental structures, e.g the first dataset is a typical example of wet lab experiment for studying the influence of controlled factors, e.g. Smith et al. (2013), whereas the second is a common example of field studies for establishing relationships among individuals, e.g. Lamichhaney et al. (2012), or for studying evolution of gene differential expression, e.g. Gu et al. (2013).

The validation included the comparison of a wide range of dissimilarity measures, including the one based on the traditional Euclidean distance, correlation-based dissimilarities using raw, transformed and normalized count data, as well as measures specifically developed for discrete count data. Dissimilarity measures based on non-transformed count data (Euclidean and Pearson correlation), resulted in dendrograms that did not resemble the expected sample structure. Normalization did not help with the use of those measures either. The dissimilarity based on Spearman correlation was the only correlation-based dissimilarity that recovered the natural sample structure. Dissimilarities calculated using appropriate transformations for count data were consistent in reproducing the expected dendrograms.

Additionally, the validation procedure proposed two metrics, agreement and consistency, for measuring the adequacy of dissimilarity measures. Agreement measures the correlation between dissimilarities, then two dissimilarities with high agreement produce dendrograms with similar hierarchical structures. Consistency measures the correlation between dendrograms obtained with a particular dissimilarity, then a dissimilarity with high consistency always generates similar dendrograms. These type of objective measures had not been reported before when comparing dendrograms in RNA-seq studies.

Finally, plasmodes allowed the analysis to be extended to different hypothetical scenarios. In one scenario only differentially expressed genes were used in clustering analysis while in another scenario all genes were included (differentially and non differentially expressed). All the dissimilarities that accounted for appropriate transformations were highly consistent in all scenarios. This dual use of representing information is useful to explore relationships but has to be interpreted in the correct context. For instance, Lamichhaney et al. (2012) presented philogenetic trees using all markers or using only markers that showed significantly genetic differentiation when studying local adaptation to salinity levels in Atlantic herring populations. The study failed to uncover the population structure, except when the gene set was restricted to pre-selected genes known for harboring SNP segregating at extreme frequencies among diverse populations.

All in all, the proposed validation method compared the most common dissimilarity measures used in hierarchical clustering and dissimilarity measures potentially appropriate for count overdispersed data. The validation included a variety of scenarios suitable for RNA-seq applications in biology studies and specifically in fisheries and wildlife applications, and proposed agreement and consistency metrics to objectively compare dendrograms structures.

3. To compare sensitivity and false discovery rates for SNP genotyping in variant calling programs

Calling and genotyping variants from DNA-seq has been actively studied (Nielsen et al., 2011; Li et al., 2014, 2013) and the same programs are being used for calling variant from RNA-seq (Piskol et al., 2013; Lee et al., 2013). In this thesis I compared three well known programs, namely

BCFtools, Beagle and VarScan2, in terms of sensitivity and false discovery rate when genotyping heterozygous and homozygous sites using as reference genotypes obtained from a 60K SNPchip. Additionally, the comparison considered multi samples and multi tissue scenarios, as well as studied the influence of read depth. Calling heterozygous with more than 10 reads per base and per sample provided sensitivity of 70% for VarScan2 and more that 90% for BCFtools and Beagle with low FDR in all programs. Homozygous were called with higher sensitivity in all tissues and samples irrespective of depth but presented higher FDR.

Distinguishing the sensitivity and false genotyping rate of homozygous from heterozygous sites separately allows using the information for designing analysis for different applications such as allele-specific expression and RNA editing. This type of study is emerging in fish and wildlife populations. For instance, Wang et al. (2013) used RNA-seq to study differential expression analysis and allele-specific expression related to disease resistance against enteric septicemia of catfish. Furthermore, other uses of SNP genotyping from RNA-seq are expected to increase in the future. In nonmodel organisms, SNP genotyping will be relevant for ecological and evolutionary studies and population genetics (Seeb et al., 2011; Helyar et al., 2011). For instance, Weinman et al. (2014) used RNA-seq for discovering SNP and establishing parentage and kinship in cooperatively breeding super starlings that live in highly kin-structured groups. In this sense, I demonstrated in this dissertation that accurate genotypes obtained by RNA-seq can be used to estimate genetic relationships between individuals in a similar way to using genotypes from a SNPchip. SNPchip are usually not available or are expensive to design for nonmodel organisms of little agricultural interest, thus RNA-seq can be used to complement other genotyping approaches like genotyping-by-sequencing (Narum et al., 2013) or genotyping-in-thousands by sequencing (Campbell et al., 2015; Pavey, 2015).

In summary, the comparison of variant calling programs provided information on the accuracy of calling and genotyping homozygous and heterozygous sites including multiple samples and multiple tissues scenarios, and considering the effect of read depth. The evaluation explored the results in the context of applications for allele-specific and RNA-editing studies, as well as for studying relationships in a population.

5.3 Future research directions

The genomics era is providing unprecedented opportunities for researchers to study mechanisms underlying biological processes. The amount and type of data that are being generated by ‘omics’ technologies are virtually inundating labs and research institutions. Yet, having data is just the first step towards generating information that can help to elucidate the biological processes. This context requires researchers that are able to understand the biology and to develop and apply valid statistical methods for extracting information (Gomez-Cabrero et al., 2014).

Systems biology is a field of study in which disciplines converged to study biological systems in a holistic approach. Neuroscience, physiology, and ecology all converged on the idea that it is as important to study and model the whole system and its interactions as to analyze the individuals parts alone (Conesa and Mortazavi, 2014). For instance the molecular systems biology needs to integrate data from complex interactions among biomolecules such as DNA, different types of RNA, proteins, and metabolites but similar integrative challenges exist at other multiple scales, such as cellular, organismal and ecological organization (Conesa and Mortazavi, 2014). Moreover, any integrative approach in omics data has to deal with high dimensional data and in most cases *a priori* uncertainty in the definition of the hypotheses to be tested. Thus, a field of active research is the improvement of analysis methods that allow integration of multiple data types to produce statistically valid and biologically relevant interpretations.

The procedures studied and developed in this dissertation are extremely useful to continue research in integrative analysis method in life sciences. The flexibility of plasmode construction, as presented in chapters 2 and 3, can be extended to validate further analysis methods. In chapter 2, the algorithm to generate plasmodes repeatedly sampled a random set of genes. However, it may be customized to sample particular related sets of genes in order to validate explorative and inference methodologies of gene and molecular networks such as pathway analysis. In chapter 3, the assessment of dissimilarity measures for sample-based clustering analysis can be extended to gene-based or bi-clustering, and other integrative clustering analysis for multiple omics datasets (Milone et al., 2014; Consortia, 2014). Importantly, the validation could be easily tested in dimension reduction techniques that are important to reduce the noisy high-dimensional characteristic of RNA-seq and other omics. Results obtained with principal components analysis, multidimensional scaling or correspondence analysis can be compared and validated including new approaches of dimensionality reduction recently proposed for integrative analysis (Reverter et al., 2014; Reshetova et al., 2014; Simmons et al., 2015; Tomescu et al., 2014). Chapter 3, provided and in-depth exploration of calling and genotyping SNPs. SNP datasets are one of the fundamental data sources used when integrating datasets in genomics. For instance it is typical to integrate SNP with gene expression (Conesa and Mortazavi, 2014) as in eQTL studies.

REFERENCES

REFERENCES

Alamancos, G., Agirre, E., and Eyras, E. (2014). “Methods to Study Splicing from High-Throughput RNA Sequencing Data,” in *Spliceosomal Pre-mRNA Splicing* Methods in Molecular Biology., ed. K. J. Hertel (Totowa, NJ: Humana Press), 357–397. doi:10.1007/978-1-62703-980-2.

Altmann, A., Weber, P., Bader, D., Preuss, M., Binder, E. B., and Müller-Myhsok, B. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.* 131, 1541–54. doi:10.1007/s00439-012-1213-z.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106.

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. *bioRxiv*, 0–4. doi:10.1101/002824.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1101/002824.

Auer, P. L., and Doerge, R. W. (2011). A Two-Stage Poisson Model for Testing RNA-Seq Data. *Stat. Appl. Genet. Mol. Biol.* 10, 1–28. doi:10.2202/1544-6115.1627.

Auer, P. L., and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics* 185, 405–416. doi:10.1534/genetics.110.114983.

Babbitt, C. C., Tung, J., Wray, G. A., and Alberts, S. C. (2012). Changes in Gene Expression Associated with Reproductive Maturation in Wild Female Baboons. *Genome Biol. Evol.* 4, 102–109. doi:10.1093/gbe/evr134.

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., Van Tassell, C. P., and Steibel, J. P. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet.* 14, 8. doi:10.1186/1471-2156-14-8.

Bahn, J. H., Lee, J., Li, G., Greer, C., and Peng, G. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. 142–150. doi:10.1101/gr.124107.111.

Bass, B., Hundley, H., Li, J. B., Peng, Z., Pickrell, J., Xiao, X. G., and Yang, L. (2012). The difficult calls in RNA editing. *Nat. Biotechnol.* 30, 1207–1209. doi:nbt.2452 [pii]\r10.1038/nbt.2452.

Bates, D., Maechler, M., and Bolker, B. (2013). lme4: Linear and mixed-effects models using S4 classes.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B-Methodological* 57, 289–300.

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics* 5, 433–8. doi:10.1517/14622416.5.4.433.

Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M., and Gilad, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* 20, 180–189. doi:10.1101/gr.099226.109.

Bottomly, D., Walter, N. A. R., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011). Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PLoS One* 6, e17820. doi:10.1371/journal.pone.0017820.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci.* 107, 9546–9551. doi:10.1073/pnas.0914005107.

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi:10.1086/521987.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94.

Campbell, N. R., Harmon, S. a., and Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour.* 15, 855–867. doi:10.1111/1755-0998.12357.

Cánovas, a, Rincón, G., Islas-Trejo, a, Jimenez-Flores, R., Laubscher, a, and Medrano, J. F. (2013). RNA sequencing to study gene expression and single nucleotide polymorphism variation associated with citrate content in cow milk. *J. Dairy Sci.* 96, 2637–48. doi:10.3168/jds.2012-6213.

Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S., and Medrano, J. F. (2010). SNP discovery in the bovine milk transcriptome using RNA-Seq technology. *Mamm. Genome* 21, 592–8. doi:10.1007/s00335-010-9297-z.

Cattell, R. B., and Jaspers, J. (1967). General Plasmode No. 30-10-5-2 for factor analytic exercises and research. *Multivariate Behav. Res.*

Chen, G., Wang, C., and Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Sci. China. Life Sci.* 54, 1121–8. doi:10.1007/s11427-011-4255-x.

Chen, J. Y., Peng, Z., Zhang, R., Yang, X. Z., Tan, B. C. M., Fang, H., Liu, C. J., Shi, M., Ye, Z. Q., Zhang, Y. E., et al. (2014). RNA Editome in Rhesus Macaque Shaped by Purifying Selection. *PLoS Genet.* 10, 8–12. doi:10.1371/journal.pgen.1004274.

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., and Spielman, R. S. (2010). Polymorphic <italic>Cis</italic>- and <italic>Trans</italic>-Regulation of Human Gene Expression. *PLoS Biol* 8, e1000480. doi:10.1371/journal.pbio.1000480.

Conesa, A., and Mortazavi, A. (2014). The common ground of genomics and systems biology. *BMC Syst. Biol.* 8 Suppl 2, S1. doi:10.1186/1752-0509-8-S2-S1.

Consortia, Stat. (2014). STATegRa: Classes and methods for multi-omics data integration. R package v 1.2.1.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 6, 59–75. doi:10.1093/biostatistics/kxh018.

Dalton, L., Ballarin, V., and Brun, M. (2009). Clustering algorithms: on learning, validation, performance, and applications to genomics. *Curr. Genomics* 10, 430–45. doi:10.2174/138920209789177601.

Di, Y., Schafer, D. W., Cumbie, J. S., and Chang, J. H. (2011). The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* 10. doi:10.2202/1544-6115.1637.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* doi:10.1093/bib/bbs046.

Djari, A., Esquerré, D., Weiss, B., Martins, F., Meersseman, C., Boussaha, M., Klopp, C., and Rocha, D. (2013). Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* 14, 307. doi:10.1186/1471-2164-14-307.

Edwards, D. B., Ernst, C. W., Tempelman, R. J., Rosa, G. J. M., Raney, N. E., Hoge, M. D., and Bates, R. O. (2008). Quantitative trait loci mapping in an F2 Duroc x Pietrain resource population: I. Growth traits. *J. Anim. Sci.* 86, 241–253. doi:10.2527/jas.2006-625.

Ekblom, R., and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity (Edinb).* 107, 1–15. doi:10.1038/hdy.2010.152.

Emmert-Streib, F. (2013). Influence of the experimental design of gene expression studies on the inference of gene regulatory networks: environmental factors. *PeerJ* 1, e10. doi:10.7717/peerj.10.

Ernst, C. W., Steibel, J. P., Sollero, B. P., Strasburg, G. M., Guimarães, J. D., and Raney, N. E. (2011). Transcriptional profiling during pig fetal skeletal muscle development using direct high-throughput sequencing and crossplatform comparison with gene expression microarrays. in *Annual Meeting American Dairy Science Association and American Society of Animal Science.* (New Orleans, Louisiana: J. Anim. Sci).

Fang, Z., and Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Brief. Bioinform.* 12, 280–287. doi:10.1093/bib/bbr004.

Frazee, A., Langmead, B., and Leek, J. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* 12, 449.

Gadbury, G., Garrett, K., and Allison, D. (2009). “Challenges and Approaches to Statistical Design and Inference in High-Dimensional Investigations,” in *Plant Systems Biology SE - 9* Methods in Molecular Biology^{TM}., ed. D. A. Belostotsky (Humana Press), 181–206 LA – English. doi:10.1007/978-1-60327-563-7_9.

Gadbury, G. L., Xiang, Q., Yang, L., Barnes, S., Page, G. P., and Allison, D. B. (2008). Evaluating statistical methods using plasmode data sets in the age of massive public databases: an illustration using false discovery rates. *PLoS Genet.* 4, e1000098. doi:10.1371/journal.pgen.1000098.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8, I1. doi:10.1186/1752-0509-8-S2-I1.

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, a S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., Pugh, T. J., et al. (2010). Alternative expression analysis by RNA sequencing. *Nat. Methods* 7, 843–7. doi:10.1038/nmeth.1503.

Gu, X. (2015). Statistical Detection of Differentially Expressed Genes based on RNA-seq: from Biological to Phylogenetic Replicates. *Brief. Bioinform.*, bbv035. doi:10.1093/bib/bbv035.

Gu, X., Zou, Y., Huang, W., Shen, L., Arendsee, Z., and Su, Z. (2013). Phylogenomic distance method for analyzing transcriptome evolution based on RNA-seq data. *Genome Biol. Evol.* 5, 1746–1753. doi:10.1093/gbe/evt121.

Gualdrón Duarte, J. L., Bates, R. O., Ernst, C. W., Raney, N. E., Cantet, R. J. C., and Steibel, J. P. (2013). Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet.* 14, 38. doi:10.1186/1471-2156-14-38.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi:10.1534/genetics.107.081190.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On Clustering Validation Techniques. *J. Intell. Inf. Syst.* 17, 107–145.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 3201–12. doi:10.1093/bioinformatics/bti517.

Hardcastle, T. J., and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11, 422. doi:10.1186/1471-2105-11-422.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning. data Mining, Inference, and Prediction*. Second Edi. New York, New York, USA: Springer.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, a., Maes, G. E., Diopere, E., Carvalho, G. R., et al. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Mol. Ecol. Resour.* 11, 123–136. doi:10.1111/j.1755-0998.2010.02943.x.

Hu, M., Zhu, Y., Taylor, J. M. G., Liu, J. S., and Qin, Z. S. (2011). Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics* 28, 63–68. doi:10.1093/bioinformatics/btr616.

Van Iterson, M., Boer, J. M., and Menezes, R. X. (2010). Filtering, FDR and power. *BMC Bioinformatics* 11, 450. doi:10.1186/1471-2105-11-450.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. New York, New York, USA: Springer.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* 16, 1370–1386. doi:10.1109/TKDE.2004.68.

Johnson, R. A., and Wichern, D. W. (2002). *Applied multivariate statistical analysis*. 5th ed. Upper Saddle River, N.J.: Prentice Hall.

Joost, S., Bonin, a., Bruford, M. W., Després, L., Conord, C., Erhardt, G., and Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16, 3955–3969. doi:10.1111/j.1365-294X.2007.03442.x.

Kendall, M. G., and Gobbons, J. D. (1990). *Rank Correlation Methods*. 5th ed. USA: Oxford University Press.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., Miller, C. a, Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. 568–576. doi:10.1101/gr.129684.111.

Kumar, P., Al-shafai, M., Ahmed, W., Muftah, A., Chalhoub, N., Elsaid, M. F., Aleem, A. A., and Suhre, K. (2014). Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and Mendelian inheritance. 7, 1–13. doi:10.1186/1756-0500-7-747.

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E. R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M. T., et al. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proc. Natl. Acad. Sci. U. S. A.* 109, 19345–50. doi:10.1073/pnas.1216128109.

Langmead, B., Hansen, K. D., and Leek, J. T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* 11, R83. doi:10.1186/gb-2010-11-8-r83.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–360. doi:10.1038/nmeth.1923.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15, R29. doi:10.1186/gb-2014-15-2-r29.

Lee, J.-H., Ang, J. K., and Xiao, X. (2013). Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA*, rna.037903.112–. doi:10.1261/rna.037903.112.

Leek, J. T., and Storey, J. D. (2011). The Joint Null Criterion for Multiple Hypothesis Tests. *Stat. Appl. Genet. Mol. Biol.* 10. doi:10.2202/1544-6115.1673.

Legendre, P., and Legendre, L. (2012). *Numerical ecology*. Third. Amsterdam: Elsevier.

Lemay, M. A., Donnelly, D. J., and Russello, M. A. (2013). Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon. *BMC Genomics*. doi:10.1186/1471-2164-14-308.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi:10.1093/bioinformatics/btr509.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30, 2843–2851. doi:10.1093/bioinformatics/btu356.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. 1–2.

Li, J. B., Levanon, E. Y., Yoon, J.-K., Aach, J., Xie, B., LePoust, E., Zhang, K., Gao, Y., and Church, G. M. (2009b). Genome-Wide Identification of Human RNA editing sites by parallel DNA capturing and sequencing. *Science (80-.).* 324, 1210:1213.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13, 523–38. doi:10.1093/biostatistics/kxr031.

Li, S., Łabaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.-Y., Wang, M., Wang, C., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* 32. doi:10.1038/nbt.3000.

Li, Y., Chen, W., Liu, E. Y., and Zhou, Y. H. (2013). Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data. *Stat. Biosci.* 5, 3–25. doi:10.1007/s12561-012-9067-4.

Liu, P., and Si, Y. (2014). “Cluster Analysis of RNA-Sequencing Data,” in *Statistical Analysis of Next Generation Sequencing Data SE - 10* Frontiers in Probability and the Statistical Sciences., eds. S. Datta and D. Nettleton (Springer International Publishing), 191–217. doi:10.1007/978-3-319-07212-8_10.

Love, M. I., Huber, W., and Anders, S. (2014a). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Love, M. I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*. doi:10.1101/002832.

Ma, C., and Wang, X. (2012). Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis. *Plant Physiol.* 160, 192–203. doi:10.1104/pp.112.201962.

Marguerat, S., Bähler, J., and Bahler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* 67, 569–579. doi:10.1007/s00018-009-0180-6.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, gks042–. doi:10.1093/nar/gks042.

Mehta, T. S., Zakharkin, S. O., Gadbury, G. L., and Allison, D. B. (2006). Epistemological issues in omics and high-dimensional biology: give the people what they want. *Physiol. Genomics* 28, 24–32. doi:10.1152/physiolgenomics.00095.2006.

Mehta, T., Tanik, M., and Allison, D. B. (2004). Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.* 36, 943–947. doi:10.1038/ng1422.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi:10.1038/nrg2626.

Milone, D. H., Stegmayer, G., Lopez, M., Kamenetzky, L., and Carrari, F. (2014). Improving clustering with metabolic pathway data. *BMC Bioinformatics* 15, 101. doi:10.1186/1471-2105-15-101.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi:http://www.nature.com/nmeth/journal/v5/n7/suppinfo/nmeth.1226_S1.html.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., and Hohenlohe, P. a. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.*, n/a–n/a. doi:10.1111/mec.12350.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–51. doi:10.1038/nrg2986.

Oshlack, A., Robinson, M., and Young, M. (2010). From RNA-seq reads to differential expression results. *Genome Biol.* 11, 220.

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi:10.1038/nrg2934.

Pachter, L. (2011). Models for transcript quantification from RNA-Seq. *ArXiv* 1104.3889, 1–28.

Pavey, S. a (2015). High-throughput SNPs for all: genotyping-in-thousands. *Mol. Ecol. Resour.* 15, 685–687. doi:10.1111/1755-0998.12405.

Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* 30, 253–260. doi:10.1038/nbt.2122.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772. doi:http://www.nature.com/nature/journal/v464/n7289/suppinfo/nature08872_S1.html.

Pirinen, M., Lappalainen, T., Zaitlen, N. A., Dermitzakis, E. T., Donnelly, P., McCarthy, M. I., and Rivas, M. A. (2015). Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics*, btv074–. doi:10.1093/bioinformatics/btv074.

Piskol, R., Ramaswami, G., and Li, J. B. (2013). Reliable Identification of Genomic Variants from RNA-Seq Data. *Am. J. Hum. Genet.* 93, 641–51. doi:10.1016/j.ajhg.2013.08.008.

Qian, X., Ba, Y., Zhuang, Q., and Zhong, G. (2014). RNA-Seq technology and its application in fish transcriptomics. *OMICS* 18, 98–110. doi:10.1089/omi.2013.0110.

Quinn, A., Juneja, P., and Jiggins, F. M. (2014). Estimates of allele-specific expression in Drosophila with a single genome sequence and RNA-seq data. *Bioinformatics* 30, 2603–10. doi:10.1093/bioinformatics/btu342.

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., Corvin, A. P., and Morris, D. W. (2013). Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS One* 8. doi:10.1371/journal.pone.0058815.

R Development Core Team (2014). R: A language and environment for statistical computing.

Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C., and Li, J. B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods* 9, 579–581. doi:10.1038/nmeth.1982.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O’Connell, M. a, and Li, J. B. (2013). Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods* 10, 128–32. doi:10.1038/nmeth.2330.

Rapaport, F., Khanin, R., Liang, Y., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential expression analysis methods for RNA-seq data. 1–21.

Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 29, 2146–2152. doi:10.1093/bioinformatics/btt350.

Rau, A., Maugis-Rabusseau, C., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics* 31, 1420–1427. doi:10.1093/bioinformatics/btu845.

Reeb, P. D., and Steibel, J. P. (2013). Evaluating statistical analysis models for RNA sequencing experiments. *Front. Genet.* 4, 1–9. doi:10.3389/fgene.2013.00178.

Reshetova, P., Smilde, A. K., van Kampen, A. H., and Westerhuis, J. a (2014). Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. *BMC Syst. Biol.* 8 Suppl 2, S2. doi:10.1186/1752-0509-8-S2-S2.

Reverter, F., Vegas, E., and Oller, J. M. (2014). Kernel-PCA data integration with enhanced interpretability. *BMC Syst. Biol.* 8 Suppl 2, S6. doi:10.1186/1752-0509-8-S2-S6.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27, 2325–9. doi:10.1093/bioinformatics/btr355.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25. doi:10.1186/gb-2010-11-3-r25.

Robinson, M. D., and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332. doi:10.1093/biostatistics/kxm030.

Robles, J. a, Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., and Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* 13, 484. doi:10.1186/1471-2164-13-484.

Rosa, G. J. M., Steibel, J. P., and Tempelman, R. J. (2005). Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comp. Funct. Genomics* 6, 123–31. doi:10.1002/cfg.464.

Salem, M., Vallejo, R. L., Leeds, T. D., Palti, Y., Liu, S., Sabbagh, A., Rexroad, C. E., and Yao, J. (2012). RNA-seq identifies SNP markers for growth traits in rainbow trout. *PLoS One* 7. doi:10.1371/journal.pone.0036264.

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive Genetic Variation on the Landscape: Methods and Cases. *Annu. Rev. Ecol. Evol. Syst.* 43, 23–43. doi:10.1146/annurev-ecolsys-110411-160248.

Schunter, C., Garza, J. C., Macpherson, E., and Pascual, M. (2013). SNP development from RNA-seq data in a nonmodel fish: how many individuals are needed for accurate allele frequency prediction? *Mol. Ecol. Resour.*, n/a–n/a. doi:10.1111/1755-0998.12155.

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour.* 11, 1–8. doi:10.1111/j.1755-0998.2010.02979.x.

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., Muehlbauer, G. J., Nelson, R. T., Grant, D., Specht, J. E., et al. (2010). RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* 10, 160. doi:10.1186/1471-2229-10-160.

Si, Y., Liu, P., Li, P., and Brutnell, T. P. (2014). Model-Based Clustering for RNA-Seq Data. *Bioinformatics* 30, 197–205. doi:10.1093/bioinformatics/btt632.

Simmons, S., Peng, J., Bienkowska, J., and Berger, B. (2015). Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data. *J. Comput. Biol.* 22, 150622063210002. doi:10.1089/cmb.2015.0085.

Sloutsky, R., Jimenez, N., Swamidass, S. J., and Naegle, K. M. (2013). Accounting for noise when clustering biological data. *Brief. Bioinform.* 14, 423–36. doi:10.1093/bib/bbs057.

Smeds, L., and Künstner, A. (2011). ConDeTri--a content dependent read trimmer for Illumina data. *PLoS One* 6, e26314. doi:10.1371/journal.pone.0026314.

Smith, S., Bernatchez, L., and Beheregaray, L. B. (2013). RNA-seq analysis reveals extensive transcriptional plasticity to temperature stress in a freshwater fish species. *BMC Genomics* 14, 375. doi:10.1186/1471-2164-14-375.

Smyth, G. K. (2005). “Limma : Linear Models for Microarray Data,” in *Bioinformatics* Statistics for Biology and Health., eds. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber (New York: Springer), 397–420. doi:10.1007/0-387-29362-0_23.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, Article3. doi:10.2202/1544-6115.1027.

Smyth, G. K., Ritchie, M., and Thorne, N. (2012). limma: Linear Models for Microarray Data User ' s Guide (Now Including RNA-Seq Data Analysis). Melbourne, Australia: Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research.

Sokal, R. R., and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon* 11, 33–40.

Sollero, B. P., Guimarães, S. E. F., Rilington, V. D., Tempelman, R. J., Raney, N. E., Steibel, J. P., Guimarães, J. D., Lopes, P. S., Lopes, M. S., and Ernst, C. W. (2011). Transcriptional profiling during foetal skeletal muscle development of Piau and Yorkshire–Landrace cross-bred pigs. *Anim. Genet.* 42, 600–612. doi:10.1111/j.1365-2052.2011.02186.x.

Srivastava, S., and Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* 38, e170. doi:10.1093/nar/gkq670.

Steibel, J. P., Bates, R. O., Rosa, G. J. M., Tempelman, R. J., Rilington, V. D., Ragavendran, A., Raney, N. E., Ramos, A. M., Cardoso, F. F., Edwards, D. B., et al. (2011). Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs. *PLoS One* 6, e16766. doi:10.1371/journal.pone.0016766.

Steibel, J. P., Poletto, R., Coussens, P. M., and Rosa, G. J. M. (2009). A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data. *Genomics* 94, 146–52. doi:10.1016/j.ygeno.2009.04.008.

Steibel, J. P., Reeb, P. D., Ernst, C. W., and Bates, R. O. (2014). Mapping cis and trans-acting eQTL in swine populations. in *10th WCGALP* (Vancouver, Canada).

Steibel, J. P., Wang, H., and Zhong, P.-S. (2015). A hidden Markov approach for ascertaining cSNP genotypes from RNA sequence data in the presence of allelic imbalance by exploiting linkage disequilibrium. *BMC Bioinformatics* 16, 1–12. doi:10.1186/s12859-015-0479-2.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 64, 479–498. doi:10.1111/1467-9868.00346.

Storey, J. D., and Tibshirani, R. (2003). Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol. Biol.* 224, 149–57. doi:10.1385/1-59259-364-X:149.

Tomescu, O. a, Mattanovich, D., and Thallinger, G. G. (2014). Integrative omics analysis. A study based on Plasmodium falciparum mRNA and protein data. *BMC Syst. Biol.* 8 Suppl 2, S4. doi:10.1186/1752-0509-8-S2-S4.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi:10.1038/nprot.2012.016.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–5. doi:10.1038/nbt.1621.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980.

Vaughan, L. K., Divers, J., Padilla, M., Redden, D. T., Tiwari, H. K., Pomp, D., and Allison, D. B. (2009). The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Comput. Stat. Data Anal.* 53, 1755–1766. doi:10.1016/j.csda.2008.02.032.

Vijay, N., Poelstra, J. W., Künstner, A., and Wolf, J. B. W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.* 22, 620–34. doi:10.1111/mec.12014.

Waller, N. G., Underhill, J. M., and Heather, A. (2010). Multivariate Behavioral A Method for Generating Simulated Plasmodes and Artificial Test Clusters with User-Defined Shape , Size , and. 37–41.

Wang, R., Sun, L., Bao, L., Zhang, J., Jiang, Y., Yao, J., Song, L., Feng, J., Liu, S., and Liu, Z. (2013). Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish. *BMC Genomics* 14, 929. doi:10.1186/1471-2164-14-929.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57–63. doi:10.1038/nrg2484.

Weinman, L. R., Solomon, J. W., and Rubenstein, D. R. (2014). A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. *Mol. Ecol. Resour.*, n/a–n/a. doi:10.1111/1755-0998.12330.

Wickramasinghe, S., Cánovas, A., Rincón, G., and Medrano, J. F. (2014). RNA-Sequencing: A tool to explore new frontiers in animal genetics. *Livest. Sci.* 166, 206–216. doi:10.1016/j.livsci.2014.06.015.

Van De Wiel, M. a, Leday, G. G. R., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 113–128. doi:10.1093/biostatistics/kxs031.

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., Therkildsen, N. O., Morikawa, M., and Palumbi, S. R. (2012). The simple fool’s guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Mol. Ecol. Resour.* 12, 1058–67. doi:10.1111/1755-0998.12003.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* 5, 2493–2518. doi:10.1214/11-AOAS493.

Xiong, H., and Li, Z. (2013). “Clustering Validation Measures,” in *Data Clustering* Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. (Chapman and Hall/CRC). doi:doi:10.1201/b15410-24.

Yang, C., and Wei, H. (2015). Designing Microarray and RNA-Seq Experiments for Greater Systems Biology Discovery in Modern Plant Genomics. *Mol. Plant* 8, 196–206. doi:10.1016/j.molp.2014.11.012.

Yang, H., and Churchill, G. (2007). Estimating p-values in small microarray experiments. *Bioinformatics* 23, 38–43. doi:10.1093/bioinformatics/btl548.

Yu, Y., Wei, J., Zhang, X., Liu, J., Liu, C., Li, F., and Xiang, J. (2014). SNP discovery in the transcriptome of white Pacific shrimp Litopenaeus vannamei by next generation sequencing. *PLoS One* 9, e87218. doi:10.1371/journal.pone.0087218.

Zhou, X., Lindsay, H., and Robinson, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.*, 1–10. doi:10.1093/nar/gku310.

Zhou, Y.-H., Xia, K., and Wright, F. a (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* 27, 2672–8. doi:10.1093/bioinformatics/btr449.