5-9-2016

Methods for Differential Analysis of Gene Expression and Metabolic Pathway Activity

Yvette Charly B. Temate Tiagueu

METHODS FOR DIFFERENTIAL ANALYSIS OF GENE EXPRESSION AND METABOLIC PATHWAY ACTIVITY

by

YVETTE CHARLY BLANCHE TEMATE-TIAGUEU

Under the Direction of Alexander Zelikovsky, PhD

ABSTRACT

RNA-Seq is an increasingly popular approach to transcriptome profiling that uses the capabilities of next generation sequencing technologies and provides better measurement of levels of transcripts and their isoforms. In this thesis, we apply RNA-Seq protocol and transcriptome quantification to estimate gene expression and pathway activity levels. We present a novel method, called IsoDE, for differential gene expression analysis based on bootstrapping. In the first version of IsoDE, we compared the tool against four existing

methods: Fisher's exact test, GFOLD, edgeR and Cuffdiff on RNA-Seq datasets generated using three different sequencing technologies, both with and without replicates. We also introduce the second version of IsoDE which runs 10 times faster than the first implementation due to some in-memory processing applied to the underlying gene expression frequencies estimation tool and we also perform more optimization on the analysis.

The second part of this thesis presents a set of tools to differentially analyze metabolic pathways from RNA-Seq data. Metabolic pathways are series of chemical reactions occurring within a cell. We focus on two main problems in metabolic pathways differential analysis, namely, differential analysis of their inferred activity level and of their estimated abundance. We validate our approaches through differential expression analysis at the transcripts and genes levels and also through real-time quantitative PCR experiments. In part Four, we present the different packages created or updated in the course of this study. We conclude with our future work plans for further improving IsoDE 2.0.

INDEX WORDS: Bootstrapping algorithm, Next generation sequencing, Gene expression, RNA-Seq data, Expectation maximization, Graph analysis, Metabolic pathway activity level, Metabolic pathways, Metabolic pathway abundance, KEGG, Differential gene expression analysis

METHODS FOR DIFFERENTIAL ANALYSIS OF GENE EXPRESSION AND

METABOLIC PATHWAY ACTIVITY

by

YVETTE CHARLY BLANCHE TEMATE-TIAGUEU

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2016

METHODS FOR DIFFERENTIAL ANALYSIS OF GENE EXPRESSION AND

METABOLIC PATHWAY ACTIVITY

by

YVETTE CHARLY BLANCHE TEMATE-TIAGUEU

Committee Chair:	Alexander Zelikovsky
Committee:	Ion Mandoiu
	Robert Harrison
	Yanqing Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2016

DEDICATION

To my late father, who always encouraged me to study hard and instilled in me the desire to succeed. To my late sister Christiane, who passed away in June 2014, may your soul rest in peace. To my family, for all the time I was away fom them because of this study, thank you.

ACKNOWLEDGEMENTS

I would like to take this opportunity to first and foremost thank God for being my strength and guide in the writing of this thesis. Without Him, I would not have had the wisdom or the physical ability to do so.

I express my gratitude to my advisor Dr. Alex Zelikovsky for his encouragement, patience and overall guidance. You made me a better scientist and researcher. Your open mind and friendliness went a long way in making me feel less stress in my studies.

I would also like to thank Dr. Ion Mandoiu who has been there almost from the beginning, always very responsive; your methodical approach to research coupled with your vast knowledge of technologies made me learn faster. My special thanks go to my graduate committee members Dr. Robert Harrison and Dr. Yanqing Zhang, for helping to guide this research and provide exceptional feedback and encouragement, by your commitment to research, you become a model of inspiration to me.

I am grateful for all the help that Dr. Raj Sunderraman has provided to me over the years. You have been a true mentor to me through your advice and supervision on and off academic matters.

I would also like to thank Dr. Sahar Al Seesi for her patience, her diligence and hard work that motivates me to improve myself and for the excellent discussions and explanations regarding gene expression analysis.

Special thanks to all of my friends in the department: Mrs Tammie Dudley, Mrs Venette Rice, Mrs Celena Pittman, Mrs Adrienne Martin, Dr. Pavel Skums, Nick Mancuso, Adrian Caciula, Bassam Tork, Igor Mandric, Katia Nenastyeva, Olga Glebova, Sasha Artyomenko, Chinua Umoja, Evrim Guler, Melinda McDaniel and Dhara Shah and to the entire department of Computer Science at Georgia State University. All of you guys contributed immensely in my academic and social growth.

I cannot thank my family enough for their unrelenting support throughout my life.

A special recognition goes to my late father, my mother and all my brothers and sisters for their continuous encouragements and prayers and for believing in me. My gratitude also goes to those I now call family: Evelyne and Roland Temah, Gertrude Kowa, Justine Siewe, Odelia Mbah, Emma Sapim, Paul and Phillis Montello, Myriam Balde and all the members of my prayer group and CRHP sisters for all the support and encouragement. All of you motivated me to never give up.

Finally, I can't thank my husband and kids enough for the sacrifice they made for me to go through this program. Your constant encouragement and unconditional understanding help me more than you can think.

TABLE OF CONTENTS

LIST OF TABLES

xiii

LIST OF FIGURES

LIST OF ABBREVIATIONS

- NGS - Next Generation Sequencing

- PPV - Predictive Positive Value

- EM - Expectation Maximization

- DE - Differential Expression

- FPKM -Fragment Per Kilobase of gene length per Million reads

- KEGG - Kyoto Encyclopedia of Genes and Genomes

- qPCR - Quantitative Polymerase Chain Reaction

PART 1

INTRODUCTION

1.1 RNA-Seq protocol for transcriptome quantifications

RNA-Seq is an increasingly popular approach to transcriptome profiling that uses the capabilities of next generation sequencing (NGS) technologies and provides better measurement of levels of transcripts and their isoforms. One issue plaguing RNA-Seq experiments is reproducibility. This is a central problem in bioinformatics in general. It is not easy to benchmark the entire RNA-Seq process [1], and the fact that there are fundamentally different ways of analyzing the data (assembly, feature counting, etc) make it more difficult. Nevertheless RNA-Seq offers huge advantages over microarrays since there is no limit on the numbers of genes surveyed, no need to select what genes to target, and no requirements for probes or primers and it is the tool of choice for metagenomics studies. Also, RNA-seq has the ability to quantify a large dynamic range of expression levels, this lead to transcriptomics and metatranscriptomics.

Rapid advances in NGS have enabled shotgun sequencing of total DNA and RNA extracted from complex microbial communities, ushering the new fields of metagenomics and metatranscriptomics. Depending on surrounding conditions e.g. food availability, stress or physical parameters, the gene expression of organisms can vary widely. The aim of transcriptomics is to capture the gene activity. Transcriptomics helps perform gene expression profiling to unravel gene functions. It can tell us, which metabolic pathways are in use under the respective conditions and how the organisms interact with the environment. Hence, it can be applied for environmental monitoring and for the identification of key genes. Transcriptomics also play a role in clinical diagnosis and in screening for drug targets or for genes, enzymes and metabolites relevant for biotechnology [2–4].

While transcriptomics deals with the gene expression of single species, metatran-

scriptomics covers the gene activity profile of the whole microbial community. Metatran-scriptomics studies changes in the the function and structure of complex microbial com-munities as it adapts to environments such as soil and seawater. Unfortunately, as in all "meta" approaches, only a small percentage of the vast number of ecologically important genes has been correctly annotated[5].

In this thesis, we apply RNA-Seq protocol and transcriptome quantification to esti-mate gene expression and pathway activity levels.

1.1.1 Differential gene expression analysis

Gene expression is the process by which the genetic code (the nucleotide sequence) of a gene becomes a useful product. The motivation behing analyzing gene expression is to identify genes whose patterns of expression differ according to phenotype, disease, exper-imental condition (e.g.disease and control) or even from different organisms. Important factors to consider while analyzing differentially expressed genes are: normalization, ac-curacy of differential expression detection and differential expression analysis when one condition has no detectable expression. A very popular domain of application of gene expression analysis is time-series gene expression data [6], this stems from the fact that biological processes are often dynamic.

In general, one of the main goals of differential expression (DE) analysis is to iden-tify the differentially expressed genes between two or more conditions. Such genes are selected based on a combination of expression level threshold and expression score cut-off, which is usually based on p-values generated by statistical modeling. The expression level of each RNA unit is measured by the number of sequenced fragments that map to the transcript, which is expected to correlate directly with its abundance level [7].

The outcome of DE analysis is influenced by the way primary analysis (mapping, mapping parameters, counting) is conducted [7]. In addition, the overall library prepa-ration protocol and quality is also an important factor of bias [8–10]. As described in the next chapters, DE analysis methods differ in how to deal with these pre-analysis phases.

Furthermore, RNA Seq experiments tend to be underpowered (too few replicates) and we need methods to perform DE under these circumstances.

In this thesis, we will address the following problems:

Differential gene expression problem

>   **Given:** RNA-Seq reads from two or more samples from two conditions and gene annotation.
>
>   **Find:** Differentially expressed genes across both conditions.

1.1.2   Analysis of metabolic pathway activity

Metabolic pathways are series of chemical reactions occurring within a cell. They referred to any of the sequences of biochemical reactions, catalysed by enzymes, that occur in all living cells. In each pathway, a principal chemical is modified by a series of chemical reactions. Figure 1.1 presents a part of the Glycolysis/Gluconeogenesis pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG). In analyzing the pathways, we will address the following problems:

Quantification of pathway activity level problem

>   **Given:** RNA-Seq reads from two or more samples from two conditions and a pathway database.
>
>   **Find:** The activity levels of all pathways.

Differential pathway activity analysis problem

>   **Given:** RNA-Seq reads from two or more samples from two conditions and a pathway database.
>
>   **Find:** Pathways with differential activity levels.
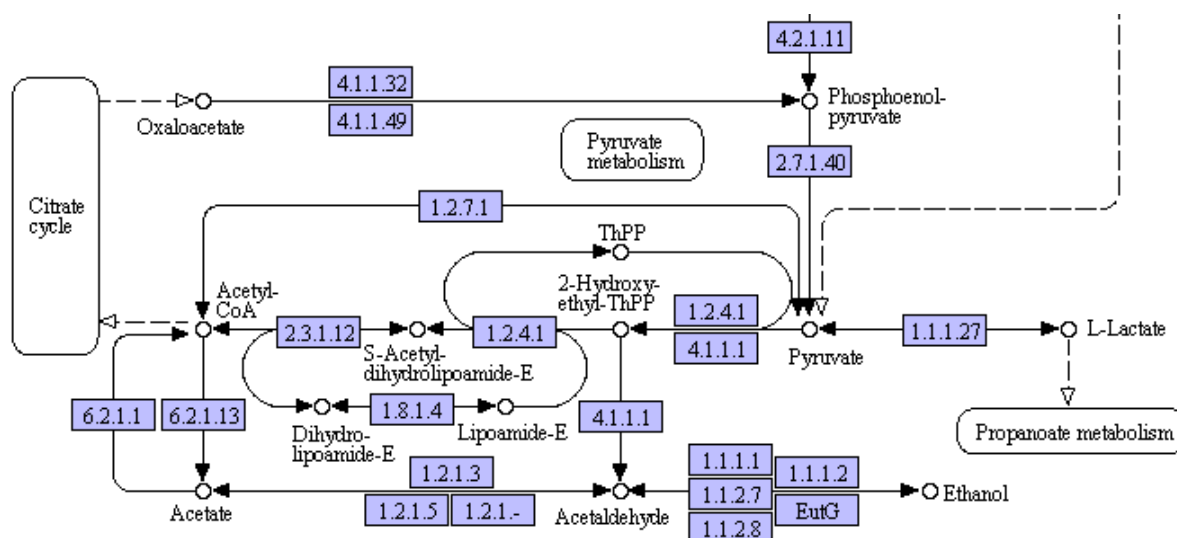
Figure 1.1 Part of the Glycolysis/Gluconeogenesis pathway from KEGG

1.2 Contributions

- IsoDE a novel method for differential gene expression analysis for RNA-Seq data. Our method uses the traditional bootstrapping approach to resample RNA-Seq reads, in conjunction with the accurate Expectation-Maximization IsoEM algorithm to estimate gene expression levels from the samples.

- Experimental study on RNA-Seq datasets generated using three different technologies (Illumina, ION Torrent, and 454) from two well-characterized MAQC samples. We show that IsoDE has consistently high accuracy, comparable or better than that of Fisher's exact test, GFOLD, Cuffdiff, and edgeR (we did not compare directly with NPEBSeq since installation was not successful).

  - Unlike other methods, IsoDE maintains high accuracy (sensitivity and PPV around 80%) on low coverage RNA-Seq datasets and at lower fold change thresholds.

- Application of IsoDE to multi-replicates studies. We explored the effect of the number of replicates on prediction accuracy using a RNA-Seq dataset with 7 replicates

for each of two conditions (control and E2-treated MCF-7 cells).

- – We show that all methods generally benefit from the use of additional replicates, GFOLD and edgeR show a marked discontinuity when transitioning from 1 to 2 replicates.

- – In contrast, IsoDE accuracy varies smoothly with changes in the number of replicates.

- A novel graph-based approach to analyze pathways significance. We represent metabolic pathways as graphs that use nodes to represent biochemical compounds, with enzymes associated with edges describing biochemical reactions.

- An implementation of an EM algorithm, in which pathways are viewed as sets of orthologs.

- The validation of the two approaches through differential expression analysis at the transcripts and genes levels and also through real-time quantitative PCR experiments.

1.3 Future work

- A new release of IsoDE with an improved pre-processing step extracting more information from bootstrap samples to tune the analysis phase further.

1.4 Roadmap

The dissertation proposal is organized as follows. Part 2 describes IsoDE tool and the motivation behind it. We first presents the state of the art in DE analysis, then we introduce our bootstrapping-based method IsoDe, we finish by describing the experimental setup, the results and discussions. Different versions of IsoDE are discussed and compare at the end of the chapter as well as our plans for improving upgrade on the

tool. Part 3 describes metabolic pathways analysis, namely inferring pathway activity level and abundance from RNA-Seq data. This section especially provides details on our expected maximization based model to estimate pathway activity and details on our topology-based approach to estimate pathway significance. In Part 4, we give a brief description of different software and tools implemented, updated or published during this work. Our plans for future work are described in Part 5.

1.5 Publications

Journal Papers

1. Meril Mathew, Kayla I Bean, **Yvette Temate-Tiagueu**, Adrian Caciula, Ion I Mandoiu, Alex Zelikovsky and Nicole B Lopanik. "Influence of symbiont-produced bioactive natural product on holobiont fitness in the marine bryozoan, Bugula neritina". Marine Biology 163.2 (2016): 1-17

2. **Yvette Temate-Tiagueu**, Meril Mathew, Igor Mandric, Sahar Al Seesi, Alex Rodriguez, Kayla Bean, Qiong Cheng, Olga Glebova, Ion Mandoiu, Nicole B. Lopanik and Alex Zelikovsky. "Inferring metabolic pathway activity levels from RNA-Seq data" [Accepted, in BMC Genomics, Dec 2015]

3. S. Al Seesi, **Yvette Temate-Tiagueu**, A. Zelikovsky, and I.I. Mandoiu, "Bootstrap-based differential gene expression analysis for RNA-Seq data without replicates". BMC Genomics 15(Suppl 8):S2 , 2014.

Book Chapter

1. Olga Glebova, **Yvette Temate-Tiagueu**, Adrian Caciula, Sahar Al Seesi, Alexander Artyomenko, Serghei Mangul, James Lindsay, Ion I. Mandoiu and Alex Zelikovsky. Book chapter. Computational Methods for Next Generation Sequencing Data Analysis. Wiley. December 2015.

Conference Papers

1. **Yvette Temate-Tiagueu**, Meril Mathew, Adrian Caciula, Qiong Cheng, Olga Glebova, Nicole Beth Lopanik, Ion Mandoiu and Alex Zelikovsky " Metabolic pathway activity estimation from RNA-Seq data". 11th International Symposium On Bioinformatics Research And Applications (ISBRA 2015)

2. Sahar Al Seesi, **Yvette Temate-Tiagueu**, Alex Zelikovsky, Ion Mandoiu "Bootstrapping-based differential gene expression analysis for RNA-Seq data with and without replicates" 4th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2014)

Posters presentation

1. Igor Mandric, **Yvette Temate-Tiagueu**, Adriano Senatore, Paul Katz and Alex Zelikovsky. "DORFA: Database-guided ORFeome Assembly from RNA-Seq Data ". Proceedings ICCABS, Miami, FL, 2015

2. **Yvette Temate-Tiagueu**, Meril Mathew, Adrian Caciula, Sahar Al Seesi, Olga Glebova, Nicole Beth Lopanik, Ion Mandoiu and Alex Zelikovsky "Bioinformatics Analysis of RNA-Seq Data for Bugula neritina" 4th IEEE International Conference on Computational Advances in Bio and Medical Sciences (CANGS 2014)

3. Leos G. Kral, Adrian Caciula, **Yvette Temate-Tiagueu**, Alexander Zelikovsky "Assembly and Annotation of the Etheostoma tallapoosae Genome" Plant and Animal Genomes conference (PAG 2014)

PART 2

DIFFERENTIAL GENE EXPRESSION ANALYSIS

2.1 Introduction

RNA-Seq has become the new standard for the analysis of differential gene expression [11] [12] [13] due to its wider dynamic range and smaller technical variance [14] compared to traditional microarray technologies. However, simply using the raw fold change of the expression levels of a gene across two samples as a measure of differential expression can still be unreliable, because it does not account for read mapping uncertainty or capture, fragmentation, and amplification variability in library preparation and sequencing. Therefore, the need for using statistical methods arises. Traditionally, statistical methods rely on the use of replicates to estimate biological and technical variability in the data. Popular methods for analyzing RNA-Seq data with replicates include edgeR[15], DESeq [16], Cuffdiff [1], and the recent NPEBSeq [17].

Unfortunately, due to the still high cost of sequencing, many RNA-Seq studies have no or very few replicates [18]. Methods for performing differential gene expression analysis of RNA-Seq datasets without replicates include variants of Fisher's exact test [14]. Recently, Feng et al. introduced GFOLD [19], a non-parametric empirical Bayesian-based approach, and showed that it outperforms methods designed to work with replicates when used for single replicate datasets.

A simple approach is to select genes using a fold-change criterion. This may be the only possibility in cases where no, or very few replicates, are available. An analysis solely based on fold change however does not allow the assessment of significance of expression differences in the presence of biological and experimental variation, which may differ from gene to gene. This is the main reason for using statistical tests to assess differential expression.[20]

2.1.1   State of the art

2.1.1.1   GFOLD

GFOLD [19] is a generalized fold change algorithm which produces biologically meaningful rankings of differentially expressed genes from RNA-Seq data. GFOLD assigns reliable statistics for expression changes based on the posterior distribution of log fold change. The authors show that GFOLD outperforms other commonly used methods when used for single replicate datasets.

2.1.1.2   Cuffdiff

Cuffdiff [1] uses a beta negative binomial distribution model to test the significance of change between samples. The model accounts for both uncertainty resulting from read mapping ambiguity and cross-replicate variability. Cuffdiff reports fold change in gene expression level along with statistical significance.

Cufflinks includes a separate program, Cuffdiff, which calculates expression in two or more samples and tests the statistical significance of each observed change in expression between them. The statistical model used to evaluate changes assumes that the number of reads produced by each transcript is proportional to its abundance but fluctuates because of technical variability during library preparation and sequencing and because of biological variability between replicates of the same experiment. Cufflinks is a transcript-level fragment count estimates. Cuffdiff uses an algorithm to model the expression of a gene G. Following this algorithm, it is able to get a distribution for the expression of a G. In the presence of replicates, to estimate the distribution of the log-fold-change in expression for G under the null hypothesis, Cuffdiff compute the average of these distributions and takes their log ratio. The process is repeated thousand times across the two conditions. To calculate a p-value of observing the real log-fold-change, they sort all the samples and count how many of them are more extreme than the log fold change they

actually saw in the real data. This number divided by the total number of draws is the estimate for the p-value. [1].

2.1.1.3 *edgeR and DESeq*

edgeR [15] is a statistical method for differential gene expression analysis which is based on the negative binomial distribution. Although edgeR is primarily designed to work with replicates it can also be run on datasets with no replicates. We used edgeR on counts of uniquely mapped reads, as suggested in [21].

EdgeR as well as DESeq are downstream count-based analysis tools like both avaioable as R/Bioconductor packages. The edgeR can be used to analyze replicates data set (highly recommended) and non-replicate.

A particular feature of edgeR functionality, are empirical Bayes methods that permit the estimation of gene-specific biological variation, even for experiments with minimal levels of biological replication. edgeR can be applied to differential expression at the gene, exon, transcript or tag level. In fact, read counts can be summarized by any genomic feature. edgeR analyses at the exon level are easily extended to detect differential splicing or isoform-specific differential expression.

DESeq and edgeR are two methods and R packages for analyzing quantitative read-outs (in the form of counts) from high-throughput experiments such as RNA-seq. After alignment, reads are assigned to a feature, where each feature represents a target transcript, in the case of RNA-Seq. An important summary statistic is the count of the number of reads in a feature (for RNA-Seq, this read count is a good approximation of transcript abundance).

Methods used to analyze array-based data assume a normally distributed, continuous response variable. However, response variables for digital methods like RNA-seq and ChIP-seq are discrete counts. Thus, both DESeq and edgeR methods are based on the negative binomial distribution.

EdgeR and DESeq use a model where they separate out the shot noise (aka counting

noise, sampling noise, Poisson noise) that arises from the count nature of the data from the variance introduced by other types of noise (technical variance and biological variance). Then the extra variance is modeled either as uniform (as in the case of EdgeR - e.g. all biological/technical variances for all transcripts are set to the same over dispersion from Poisson) or as quasi-correlated with read depth for DESeq. DESeq assumes that there is a correlation between read depth and extra-Poisson noise while EdgeR assumes no correlation.

2.1.2 Roadmap

In the next sections, we will compare our novel approach with the existing ones described here using the following measures described by Rapaport et al.[7]: i) normalization of count data; ii) sensitivity and specificity of DE detection; iii) performance on the subset of genes that are expressed in one condition but have no detectable expression in the other condition and, finally, iv) the effects of reduced sequencing depth and number of replicates on the detection of differential expression.

2.2 The IsoDE method

2.2.1 Bootstrap sample generation

As most differential expression analysis packages, IsoDE starts with a set A of RNA-Seq read alignments for each condition. Bootstrapping can be used in conjunction with any method for estimating individual gene expression levels from aligned RNA-Seq reads, estimation typically expressed in *fragment per kilobase of gene length per million reads* (FPKM). In IsoDE, we use the IsoEM algorithm [22], an expectation-maximization (EM) algorithm that takes into account gene isoforms in the inference process to ensure accurate length normalization. Unlike some of the existing estimation methods, IsoEM uses non-uniquely mapped reads, relying on the distribution of insert sizes and base quality scores (as well as strand and read pairing information if available) to probabilistically

infer their origin. Previous experiments have shown that IsoEM yields highly accurate FPKM estimates with lower runtime compared to other commonly used inference algorithms [23].

The first step of IsoDE is to generate M bootstrap samples by randomly resampling with replacement from the reads represented in A. When a read is selected during resampling, all its alignments from A are included in the bootstrap sample. The number of resampled reads in each bootstrap sample equals the total number of reads in the original sample. However, the total number of alignments may differ between bootstrap samples, depending on the number of alignments of selected reads and the number of times each read is selected. The IsoEM algorithm is then run on each bootstrap sample, resulting in M FPKM estimates for each gene. The bootstrap sample generation algorithm is summarized below:

1. Sort the alignment file A by read ID

2. Compute the number N of reads and generate a list $\mathcal{L}$ containing read IDs in the alignment file A

3. For $i = 1, \ldots, M$ do:

    (a) Randomly select with replacement N read IDs from $\mathcal{L}$, sort selected read IDs, and extract in A_i all their alignments with one linear pass over A (if a read is selected m times, its alignments are repeated m times in A_i)

    (b) Run IsoEM on A_i to get the ith FPKM estimate for each gene

2.2.2 Bootstrap-based testing of differential expression

To test for differential expression, IsoDE takes as input two folders which contain FPKM estimates from bootstrap samples generated for the two conditions to be compared. In case of replicates, a list of bootstrap folders can be provided for each condition (one folder per replicate, normally with an equal number of bootstrap samples) – IsoDE

will automatically merge the folders for the replicates to get a combined folder per condition, then perform the analysis as in the case without replicates.

In the following we assume that a total of M bootstrap samples is generated for each of the compared conditions. We experimented with two approaches for pairing the FPKMs estimated from the two sets of bootstrap samples. In the "matching" approach, a random one-to-one mapping is created between the M estimates of first condition and the M estimates of the second condition. This results in M pairs of FPKM estimates. In the "all" approach, M^2 pairs of FPKM estimates are generated by pairing each FPKM estimate for first condition with each FPKM estimate for second condition. When pairing FPKM estimate a_i for the first condition with FPKM estimate b_j for the second condition, we use a_i/b_j as an estimate for the fold change in the gene expression level between the two conditions. The "matching" approach thus results in $N = M$ fold change estimates, while the "all" approach results in $N = M^2$ fold change estimates.

The IsoDE test for differential expression requires two user specified parameters, namely the minimum fold change f and the minimum bootstrap support b. For a given threshold f (typically selected based on biological considerations), we calculate the percentage of fold change estimates that are equal to or higher than f when testing for overexpression, respectively equal to or lower than $1/f$ when testing for underexpression. If this percentage is higher than the minimum bootstrap support b specified by the user then the gene is classified as differentially expressed (DE), otherwise the gene is classified as non-differentially expressed (non-DE). The actual bootstrap support for fold change threshold f, as well as the minimum fold change with bootstrap support of at least b are also included in the IsoDE output to allow the user to easily increase the stringency of the DE test.

As discussed in the results section, varying the bootstrap support threshold b allows users to achieve a smooth tradeoff between sensitivity and specificity for a fixed fold change f (see, e.g., Figure 1). Since different tradeoffs may be desirable in different biological contexts, no threshold b is universally applicable. In our experiments we computed

b using a simple binomial model for the null distribution of fold change estimates and a fixed significance level $\alpha = 0.05$. Specifically, we assume that under the null hypothesis the fold changes obtained from bootstrap estimates are equally likely to be greater or smaller than f. We then compute b as x_{min}/N, where $x_{min} = \min\{x : P(X \geq x) \leq \alpha\}$ and X is a binomial random variable denoting the number of successes in N independent Bernoulli trials with success probability of 0.5. For convenience, a calculator for computing the bootstrap support needed to achieve a desired significance level given the (possibly different) numbers of bootstrap samples for each condition has been made available online (see Availability).

The number M of bootstrap samples is another parameter that the users of IsoDE must specify. As discussed in the results section, computing the bootstrap support for all genes takes negligible time, and the overall running time of IsoDE is dominated by the time to complete the 2M IsoEM runs on bootstrap samples. Hence, the overall runtimes grows linearly with M. Experimental results suggest that the "all" pairing approach produces highly accurate results with relatively small values of M (e.g., M = 20), and thus results in practical runtimes, independent of the number of replicates. We also note that for studies involving pairwise DE analysis of more than two conditions, IsoDE only requires M independently generated bootstrap samples per condition. Since the time for computing pairwise bootstrap support values is negligible, the overall running time will grow linearly with the number of conditions.

2.3 Settings of compared methods

The four methods that were compared to IsoDE are briefly described below.

Fisher's exact test: Fisher's exact test is a statistical significance test for categorical data which measures the association between two variables. The data is organized in a 2x2 contingency table according to the two variables of interest. We use Fisher's exact test to measure the statistical significance of change in gene expressions between two conditions A and B by setting the two values in the first row of the table to the estimated number

of reads mapped per kilobase of gene length (calculated from IsoEM estimated FPKM values) in conditions A and B, respectively. The values in the second row of the contingency table depend on the normalization method used. We compared three normalization methods. The first one is total read normalization, where the total number of mapped reads in conditions A and B are used in the second row. The second is normalization by a housekeeping gene. In this case, the estimated number of reads mapped per kilobase of housekeeping gene length in each condition is used. We also test normalization by ERCCs RNA spike-in controls [24]. FPKMs of ERCCs are aggregated together (similar to aggregating the FPKMs of different transcripts of a gene), and the estimated number of reads mapped per kilobase of ERCC are calculated from the resulting FPKM value and used for normalization. In our experiments, we used POLR2A as a housekeeping gene.

The calculated p-value, which measures the significance of deviation from the null hypothesis that the gene is not differentially expressed, is computed exactly by using the hypergeometric probability of observed or more extreme differences while keeping the marginal sums in the contingency table unchanged. We adjust the resulting p-values for the set of genes being tested using the Benjamini and Hochberg method [25] with 5% false discovery rate (FDR).

GFOLD: We used GFOLD v1.0.7 with default parameters and fold change significance cutoff of 0.05.

Cuffdiff: In our comparison, we used Cuffdiff v2.0.1 with default parameters.

edgeR: We followed the steps provided in the edgeR manual for RNA-Seq data. calcNormFactors(), estimateCommonDisp(), estimateTagwiseDisp(), and exactTest() were used with default parameter, when processing the MCF-7 replicates. When processing MAQC data and a single replicate of MCF-7 data, estimateTagwiseDisp() was not used, and the dispersion was set to 0 when calling exactTest(). The results where adjusted for multiple testing using the Benjamini and Hochberg method with 5%.

2.3.1   Mapping RNA-Seq reads

MAQC Illumina reads were mapped onto hg19 Ensembl 63 transcript library; all other datasets were mapped onto hg19 Ensembl 64 transcript library. Illumina datasets (MAQC and MCF-7) were mapped using Bowtie v0.12.8 [26]. ION Torrent reads were mapped using TMAP v2.3.2, and 454 reads were mapped using MOSAIK v 2.1.33 [27]. For edgeR, non-unique alignments were filtered out, and read counts per gene were generated using coverageBed (v2.12.0). Read mapping statistics are detailed in Table S1 in Additional File 1. Number of mapped reads per kilobase of gene length used in Fisher's exact test calculation are based on IsoEM FPKMs.

2.4   Ground truth definition

On MAQC dataset the ground truth was defined based on the available qPCR data from [28]. Each TaqMan assay was run in four replicates for each measured gene. POLR2A (ENSEMBL gene ID ENSG00000181222) was chosen as the reference gene and each replicate CT was subtracted from the average POLR2A CT to give the log2 difference (delta CT). For delta CT calculations, a CT value of 35 was used for any replicate that had CT >35. The normalized expression value of a gene g would be: $2^{(\text{CT of POLR2A})-(\text{CT of }g)}$. We filtered out genes that: (1) were not detected in one or more replicates in each samples or (2) had a standard deviation higher than 25% for the four TaqMan values in each of the two samples. Of the resulting subset, we used in the comparison genes whose Taq-Man probe IDs unambiguously mapped to Ensemble gene IDs (686 genes). A gene was considered differentially expressed if the fold change between the average normalized TaqMan expression levels bin the two conditions was greater than a set threshold with the p-value for an unpaired two-tailed T-test (adjusted for 5% FDR) of less than 0.05. We ran the experiment for fold change thresholds of 1, 1.5, and 2.

For experiments with replicates we used the RNA-Seq data generated from E2-treated and control MCF-7 cells in [21]. In this experiment, we compared IsoDE with

Predicted	Ground truth		
	Over-Expressed (TOE)	Non-Differential (TND)	Under-Expressed (TUE)
Over-Expressed (POE)	TPOE		
Non-Differential (PND)		TPND	
Under-Expressed (PUE)			TPUE

Table 2.1 Confusion matrix for differential gene expression

GFOLD and edgeR. The predictions made by each method when using all 7 replicates for each condition were used as its own ground truth to evaluate predictions made using fewer replicates. The ground truth for IsoDE was generated using a total of 70 bootstrap samples per condition.

2.5   Evaluation metrics

For each evaluated method, genes were classified according to the differential expression confusion matrix detailed in Table 2.1. Methods were assessed using sensitivity, positive predictive value (PPV), F-score, and accuracy, defined as follows:

$$
\begin{aligned}
Sensitivity &= \frac{(\text{TPOE} + \text{TPUE})}{(\text{TOE} + \text{TUE})} \\
PPV &= \frac{(\text{TPOE} + \text{TPUE})}{(\text{POE} + \text{PUE})} \\
Accuracy &= \frac{(\text{TPOE} + \text{TPND} + \text{TPUE})}{(\text{TOE} + \text{TND} + \text{TUE})} \\
F-score &= 2 \times \frac{Sensitivity \times \text{PPV}}{Sensitivity + \text{PPV}}
\end{aligned}
$$

2.6 Experimental setup

2.6.1 Datasets

We conducted experiments on publicly available RNA-Seq datasets generated from two commercially available reference RNA samples and a breast cancer cell line.

To compare the accuracy of different methods, we used RNA-Seq data RNA samples that were well-characterized by quantitative real time PCR (qRT-PCR) as part of the MicroArray Quality Control Consortium (MAQC) [28]; namely an Ambion Human Brain Reference RNA, Catalog # 6050), henceforth referred to as HBRR and a Stratagene Universal Human Reference RNA (Catalog # 740000) henceforth referred to as UHRR. To assess accuracy, DE calls obtained from RNA-Seq data were compared against those obtained as described in the Methods section from TaqMan qRT-PCR measurements collected as part of the MAQC project (GEO accession GPL4097).

We used RNA-Seq data generated for HBRR and UHRR using three different technologies: Illumina, ION-Torrent, and 454. Details about the datasets and their SRA accession numbers (or run IDs for ION Torrent datasets) are available in Table S1 in Additional File 1.

The MCF-7 RNA-Seq data was generated (from the MCF-7 ATCC human breast cancer cell line) by Liu et al. [21] using Illumina single-end sequencing with read length of 50bp. A total of 14 biological replicates were sequenced from two conditions: 7 replicates for the control group and 7 replicates for E2-treated MCF-7 cells. Sequencing each replicate resulted produced between 25 and 65 millions of mapped reads. Details about this dataset and accession numbers are also available in Table S1 in Additional File 1.

2.6.2 Bootstrapping support and pairing strategy effects on IsoDE accuracy and runtime

We evaluated both the "matching" and "all" pairing strategies of IsoDE (referred to as IsoDE-Match and IsoDE-All) for fold change threshold f of 1, 1.5, respectively 2, and
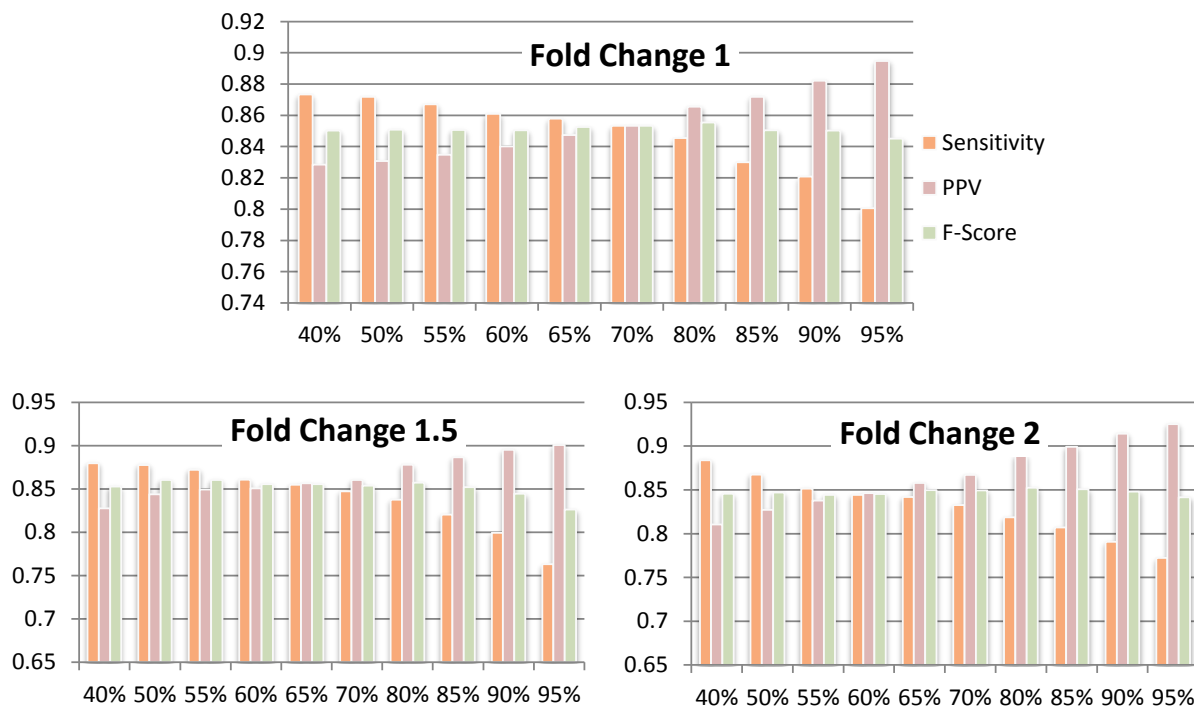
Figure 2.1 Sensitivity, PPV, and F-Score of IsoDE-Match (M=200 bootstrap samples per condition) on the Illumina MAQC data, with varying bootstrap support threshold.

bootstrap support threshold b between 40% and 95%. The results of IsoDE-Match with $M = 200$ bootstrap replicates per condition are shown in Figure 2.1. The results show that, for each tested value of f, varying b results in a smooth tradeoff between sensitivity and PPV, while the F-score changes very little. For the remaining experiments we used a bootstrap support level b computed using a significance level of 0.05 under the binomial null model detailed in the Methods section. Note the value of b selected in this way depends on the number of number N of fold change estimates, which in turn depends on both M and the pairing strategy (N is equal to M for IsoDE-Match, respectively to M^2 for IsoDE-All).

To determine the best pairing strategy, we ran IsoDE-Match and IsoDE-All with number of bootstrap samples M varying between 10 and 200 (results not shown). For the considered measures, IsoDE-All achieved an accuracy very close to that of IsoDE-Match

when run with a comparable value of N. For example, as shown in Tables 2-4, IsoDE-All run on M = 20 bootstrap samples (N = 400) had similar accuracy with the largest number of bootstrap samples we could use with IsoDE-Match (M = N = 200).

Since for a fixed N IsoDE-Match requires 2N bootstrap samples while IsoDE-All requires only $2\sqrt{N}$ of them, using IsoDE-All is significantly faster in practice. Indeed, most of the IsoDE time is spent generating bootstrap samples and estimating expression levels for each of them using the IsoEM algorithm, with bootstrap support computation typically taking a fraction of a minute. Figure 2.2 shows the time required to generate M = 20, respectively M = 200, bootstrap samples for both conditions of several MAQC datasets. All timing experiments were conducted on a Dell PowerEdge R815 server with quad 2.5GHz 16-core AMD Opteron 6380 processors and 256Gb RAM running under Ubuntu 12.04 LTS. IsoEM is run on bootstrap samples sequentially, but for each run its multi-threaded code takes advantage of all available cores (up to 64 in our experimental setup). As expected, the running time scales linearly with the number of bootstrap samples per condition, and thus generating M = 20 bootstrap samples per condition is nearly 10 times faster than generating M = 200 of them. Overall, IsoDE-Match with M = 20 has reasonable running time, varying between 1 hour for the smallest 454 dataset to 3.5 hours for the Illumina dataset.

2.7   Results and discussion

2.7.1   Results for DE prediction without replicates

We compared IsoDE against GFOLD, Cuffdiff, edgeR, and different normalization methods for Fisher's exact test; namely total normalization, housekeeping gene (POLR2A) normalization, and normalization using External RNA Controls Consortium (ERCC) RNA spike-in controls [24]. Cuffdiff results were considerably worse on the Illumina MAQC dataset, compared to other methods. Consequently, Cuffdiff was not included in other comparisons. edgeR was also not included in further comparisons due
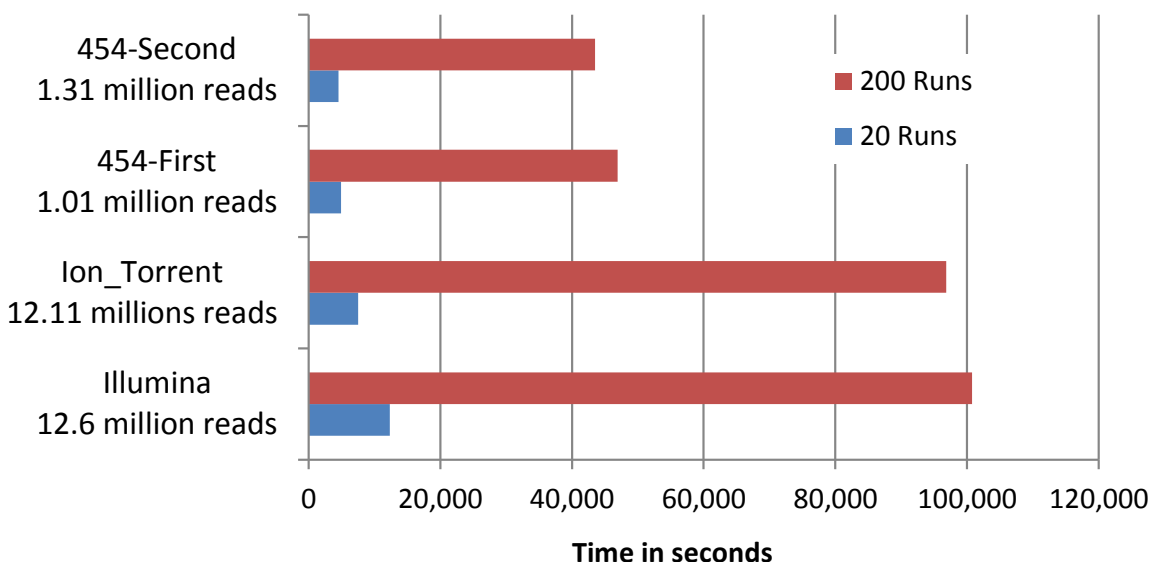
Figure 2.2 Running times (in seconds) of IsoDE-Match with $M = 200$ and IsoDE-All with $M = 20$ on several MAQC datasets. The indicated number of reads represents the total number of mapped reads over both conditions of each dataset, for more information on the datasets see Table S1.

to lack of clear definition of uniquely mapped reads for ION-Torrent and 454 datasets which were mapped using local read alignment tools. ERCC spike-ins were available only for ION Torrent samples; therefore, ERCC normalization for Fisher's exact test was conducted only for ION Torrent datasets.

Table 2.2 shows the results obtained for the MAQC Illumina dataset using minimum fold change threshold f of 1, 1.5, and 2, respectively. Table 2.3 shows the results obtained from combining the ION Torrent runs listed in Table S1 (Additional File 1) for each of the MAQC datasets for the same values of f. Table 2.4 shows the results for the First 454 MAQC dataset, while results for the Second 454 dataset are presented in Table S2 in Additional File 1. For each fold change threshold, the best performing method for each statistic is highlighted in bold.

IsoDE has very robust performance, comparable or better than that of the other methods for differential gene expression. Indeed, IsoDE outperforms them in a large number

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
	FishersTotal	70.41%	70.79%	91.24%	79.72%
	FishersHousekeeping	65.60%	65.22%	95.05%	77.36%
1	GFOLD	78.13%	80.06%	92.67%	**85.90%**
	Cuffdiff	11.37%	6.96%	**100.00%**	13.01%
	edgeR	73.03%	73.26%	95.56%	82.94%
	IsoDE-Match	**82.63%**	**87.46%**	83.70%	85.54%
	IsoDE-All	82.22%	87.17%	82.82%	84.94%
	FishersTotal	74.05%	78.20%	84.85%	81.39%
	FishersHousekeeping	76.68%	73.61%	93.67%	82.44%
1.5	GFOLD	79.15%	79.35%	90.41%	84.52%
	Cuffdiff	28.43%	8.60%	**100.00%**	15.85%
	edgeR	**80.01%**	79.92%	92.07%	**85.57%**
	IsoDE-Match	78.98%	86.23%	84.62%	85.42%
	IsoDE-All	79.01%	**86.42%**	84.49%	85.44%
	FishersTotal	78.43%	81.86%	82.44%	82.15%
	FishersHousekeeping	81.20%	80.00%	88.21%	83.90%
2	GFOLD	82.94%	78.84%	92.37%	85.07%
	Cuffdiff	40.96%	10.47%	**100.00%**	18.95%
	edgeR	**83.67%**	81.63%	91.17%	**86.13%**
	IsoDE-Match	82.04%	85.58%	85.19%	85.38%
	IsoDE-All	81.20%	**86.74%**	83.07%	84.87%

Table 2.2 Accuracy, sensitivity, PPV and F-Score in % for MAQC Illumina dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is M = 200 for IsoDE-Match and M = 20 for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
	FisherTotal	71.68%	72.76%	90.56%	80.69%
	FisherHousekeeping	67.15%	66.87%	**94.74%**	78.40%
1	FisherERCC	71.39%	72.45%	88.97%	79.86%
	GFOLD	75.77%	77.55%	90.43%	83.50%
	IsoDE-Match	**81.75%**	**86.38%**	82.18%	**84.05%**
	IsoDE-All	81.46%	86.07%	82.13%	**84.05%**
	FisherTotal	74.16%	78.39%	85.06%	81.59%
	FisherHousekeeping	76.06%	73.23%	**92.96%**	81.93%
1.5	FisherERCC	74.31%	78.59%	85.45%	81.87%
	GFOLD	75.47%	77.63%	87.88%	82.44%
	IsoDE-Match	77.66%	83.94%	84.75%	**84.34%**
	IsoDE-All	**77.81%**	**84.13%**	84.45%	84.29%
	FisherTotal	79.71%	83.02%	84.00%	83.51%
	FisherHousekeeping	**81.75%**	80.70%	88.75%	84.53%
2	FisherERCC	79.42%	82.56%	84.12%	83.33%
	GFOLD	80.58%	76.74%	**90.66%**	83.12%
	IsoDE-Match	**81.75%**	85.81%	84.63%	**85.22%**
	IsoDE-All	81.61%	**86.28%**	84.13%	85.19%

Table 2.3 Accuracy, sensitivity, PPV and F-Score in % for Ion Torrent dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is M = 200 for IsoDE-Match and M = 20 for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
	FisherTotal	34.01%	30.50%	95.63%	46.24%
	FisherHousekeeping	24.52%	20.12%	**94.74%**	33.38%
1	GFOLD	55.62%	54.18%	92.11%	68.23%
	IsoDE-Match	75.33%	79.57%	77.41%	78.47%
	IsoDE-All	**78.85%**	**84.67%**	81.04%	**82.82%**
	FisherTotal	48.18%	35.37%	89.81%	50.75%
	FisherHousekeeping	42.48%	24.86%	**97.74%**	39.63%
1.5	GFOLD	62.19%	58.13%	85.39%	69.17%
	IsoDE-Match	64.09%	74.19%	72.52%	73.35%
	IsoDE-All	**72.85%**	**79.54%**	80.62%	**80.08%**
	FisherTotal	57.96%	39.53%	85.43%	54.05%
	FisherHousekeeping	55.33%	29.30%	**97.67%**	45.08%
2	GFOLD	69.05%	61.16%	83.49%	70.60%
	IsoDE-Match	67.15%	76.51%	70.30%	73.27%
	IsoDE-All	**75.18%**	**80.93%**	78.03%	**79.45%**

Table 2.4 Accuracy, sensitivity, PPV and F-Score in % for the First 454 dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and $M = 20$ for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

of cases, across datasets and fold change thresholds. Very importantly, unlike GFOLD and Fisher's exact test, IsoDE maintains high accuracy (sensitivity and PPV around 80%) on datasets with small numbers of mapped reads such as the two 454 datasets. This observation is confirmed on results obtained for pairs of individual ION-Torrent runs, presented in Tables S3 and S4 in Additional File 1. This makes IsoDE particularly attractive for such low coverage RNA-Seq datasets.

2.7.2   DE prediction with replicates

We also studied the effect of the number of biological replicates on prediction accuracy using the MCF-7 dataset. We performed DE predictions using an increasing number of replicates. IsoDE was run with a total of 20 bootstrap samples per condition, distributed equally (or as close to equally as possible) among the replicates, as detailed in Table 2.5. GFOLD and edgeR were evaluated for 1 through 6 replicates using as ground
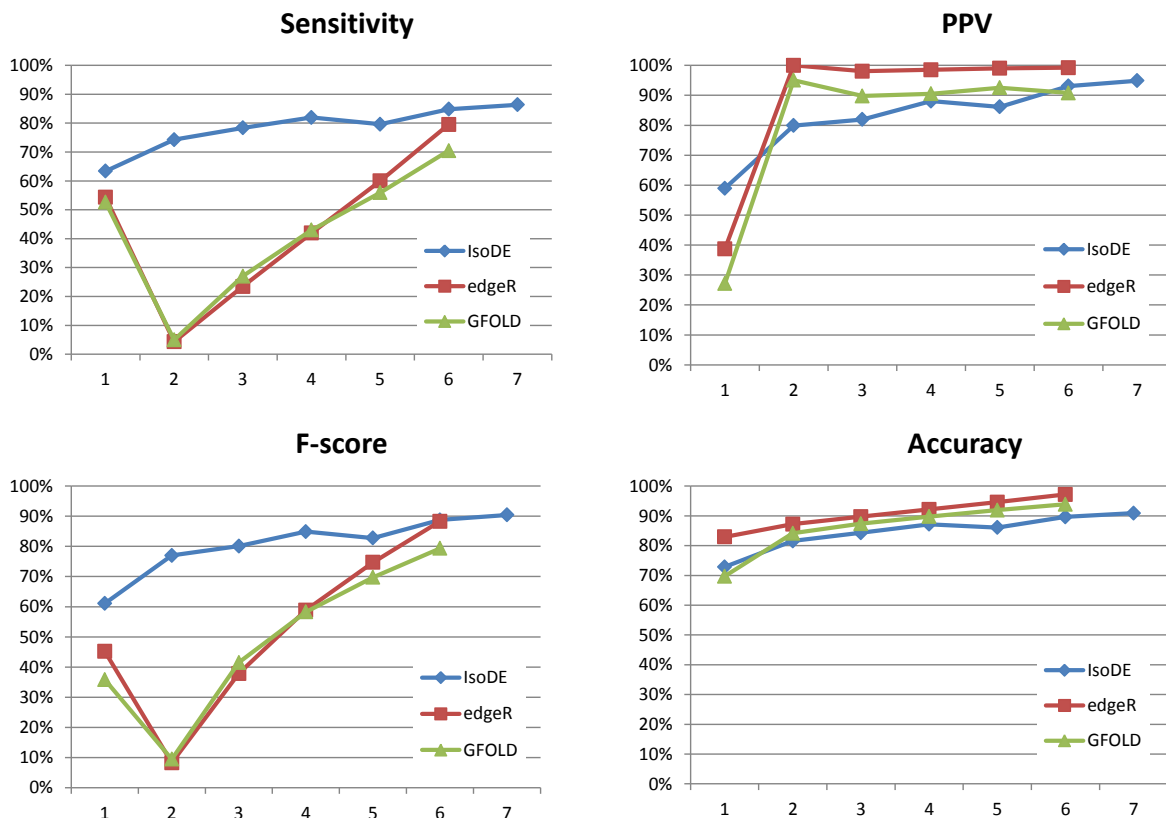
Figure 2.3 Sensitivity, PPV, F-Score, and accuracy of IsoDE-All (with 20 bootstrap runs per condition), edgeR, and GFOLD on the Illumina MCF-7 data with minimum fold change of 1 and varying number of replicates.

truth the results obtained by running each method on all 7 replicates (see the Methods section). For IsoDE, we also include the results using $M = 20$ bootstrap samples from all 7 replicates as its ground truth is generated using a much larger number of bootstrap samples ($M = 70$). Figure 2.3 shows the results of the three compared methods for a fold change threshold of 1, results for fold change thresholds 1.5 and 2 are shown in Figures S1 and S2 in Additional File 1.

Since for this experiment the ground truth was defined independently for each method, it is not meaningful to directly compare accuracy metrics of different methods. Instead, we focus on the rate of change in the accuracy of each method as additional replicates are added. Generally, all methods perform better when increasing the number

# Replicates	Rep1	Rep2	Rep3	Rep4	Rep5	Rep6	Rep7	Bootstraps Per Condition
1	20							20
2	10	10						20
3	7	7	6					20
4	5	5	5	5				20
5	4	4	4	4	4			20
6	4	4	3	3	3	3		20
7	3	3	3	3	3	3	2	20

Table 2.5 IsoDE setup for experiments with replicates. IsoDE experiments on the MCF-7 dataset was performed as follows. First we generated, for each of the 7 replicates of each condition 20, 10, 6, 5, 4, 3, respectively 2 bootstrap samples. We then used subsets of these bootstrap samples as input for IsoDE to perform DE analysis with varying number of replicates and a fixed total number $M = 20$ of bootstrap samples per condition. In experiment 1 we used 20 bootstrap samples from first replicate of each condition, in experiment 2 we used 10 bootstrap samples for each of the first 2 replicates of each condition, and so on.

of replicates. However, the accuracy of IsoDE varies smoothly, and is much less sensitive to small changes in the number of replicates. Surprisingly, this is not the case for GFOLD and edgeR sensitivity, which drops considerably when transitioning from 1 to 2 replicates, most likely due to the different statistical models employed with and without replicates. Although we varied the number of replicates without controlling the total number of reads as Liu et al. [21], our results strongly suggest that cost effectiveness metrics such as those proposed in [21] are likely to depend on to the specific method used for performing DE analysis. Therefore, the analysis method should be taken into account when using such a metric to guide the design of RNA-Seq experiments.

2.7.3   Effect of gene abundance

We also studied the effect of gene abundance on the IsoDE, GFOLD, and edgeR prediction accuracy. We selected the subset of genes that are expressed in at least one of the two RNA samples. We sorted these genes by the average of the gene's expression. We used the FPKM values predicted by IsoEM, the FPKM values predicted by GFOLD, and the number of uniquely mapped reads, for IsoDE, GFOLD, and edgeR, respectively. The

genes were then divided into quintiles, for each method independently, where quintile 1 had the genes with the lowest expression levels, and quintile 5 had the genes with the highest expression levels. Sensitivity, PPV, and F-score where calculated for each quintile separately.

Figure 2.4 shows that, for results with both 1 and 6 replicates, sensitivity, PPV, and F-score of IsoDE are only slightly lower on genes with low expression levels compared to highly expressed genes (similar results are achieved for intermediate numbers of replicates and higher fold change thresholds). In contrast, GFOLD shows a marked difference in all accuracy measures for genes in the lower quintiles compared to those in the higher quintiles. The sensitivity of edgeR is also lower for genes expressed at low levels, however it's PPV is relatively constant across expression levels.

2.8 Taking in account inter-replicates variations

2.8.1 Estimation of inter-replicates variations

2.8.1.1 Experimental setup

We conducted an analysis to estimate false positive rate. Our expectation is to have no DE genes or a very few number of them when we perform DE within the same condition. For this purpose, we used once again the MCF-7 dataset. We have 7 MCF treated and 7 MCF untreated replicates. In this experiment we would compare replicates of treated with replicates of treated, and replicates of untreated with replicates of untreated. We can only have 3 replicates per sub-conditions.

We designed three scenarios all of them comparing control vs control and treated vs treated from MCF7 replicates. To follow a similar setting as in the experiment in Table S5, we ran IsoDE using an increasing number of replicates, all from the same conditions. A total of 20 bootstrap samples were used for each sub-conditions while varying the number of replicates from 1 to 7. Let us call the replicates E1,...,E7 and C1,...,C7. The details of the experiments are presented in table 2.6.
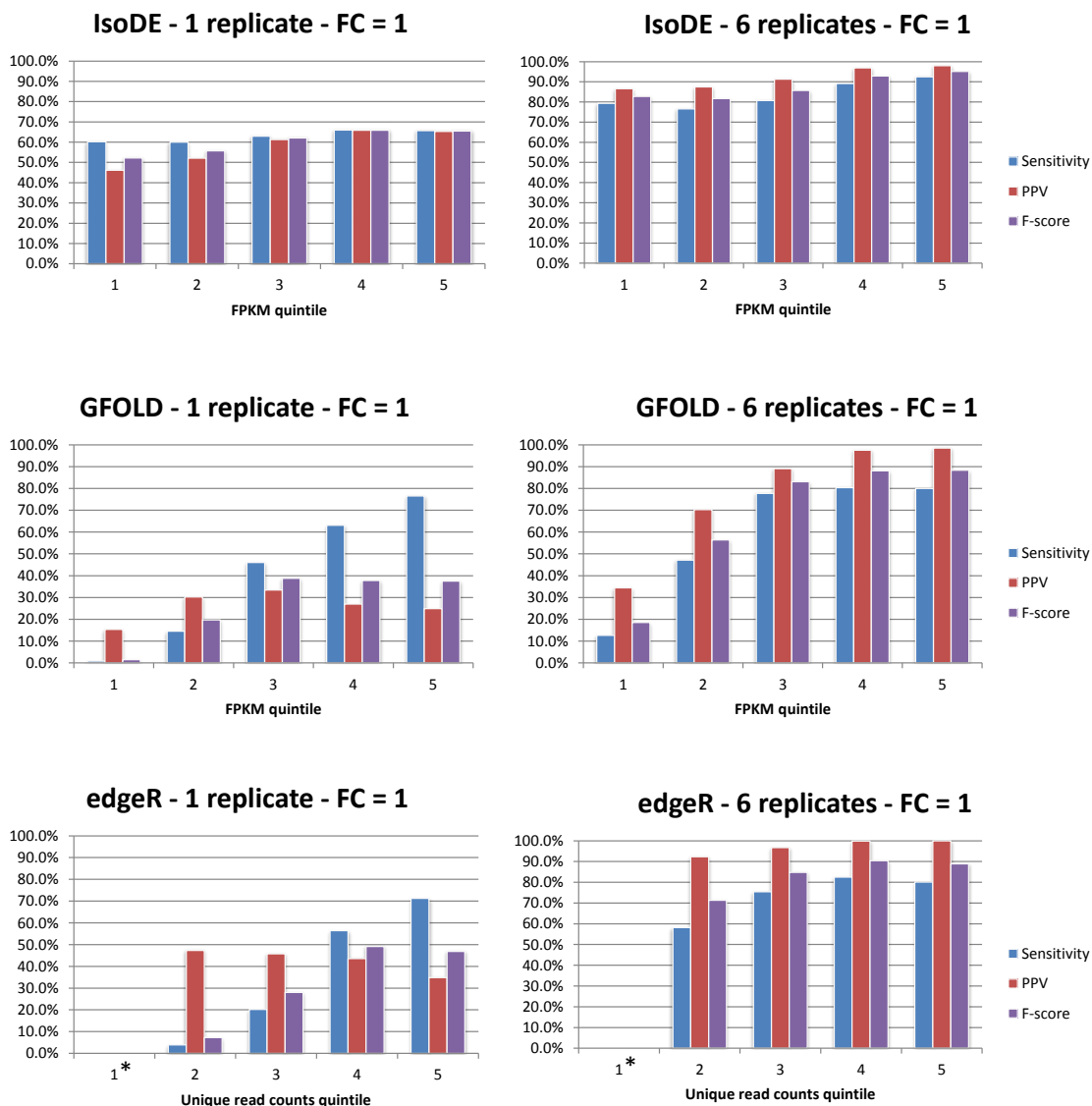
Figure 2.4 Sensitivity, PPV, and F-Score of IsoDE-All (with 20 bootstrap runs per condition), edgeR, and GFOLD on the Illumina MCF-7 data, computed for quintiles of expressed genes after sorting in non-decreasing order of average FPKM for IsoDE and GFOLD and average count of uniquely aligned reads for edgeR. First quintile of edgeR had 0 differentially expressed genes according to the ground truth (obtained by using all 7 replicates).

# Replicates	E1	E2	E3	E4	E5	E6	E7	Bootstraps Per Sub-condition
1	20			20				20
2	10	10		10	10			20
3	7	7	6	7	7	6		20

Table 2.6 IsoDE setup for experiments with replicates within each condition. IsoDE experiments on the MCF-7 dataset was performed as follows. First we generated, for each of the 6 replicates used in each condition 20, 10, 7, respectively 6 bootstrap samples. We then used subsets of these bootstrap samples as input for IsoDE to perform DE analysis with varying number of replicates and a fixed total number $M = 20$ of bootstrap samples per condition. Each condition consists of 3 replicates. In experiment 1 we used 20 bootstrap samples from first replicate and fourth replicate, in experiment 2 we used 10 bootstrap samples for each of the first 2 replicates and 10 bootstrap samples for replicates 4th and 5th, and so on.

2.8.1.2   Results

Figure 2.5 presents the results of IsoDE for the experiments. The chart shows the total number of genes, the number of genes which have FPKM equals to 0 - without sampling, that is using all the reads - and a bar showing genes which are DE with FPKM = 0 in at least one condition.
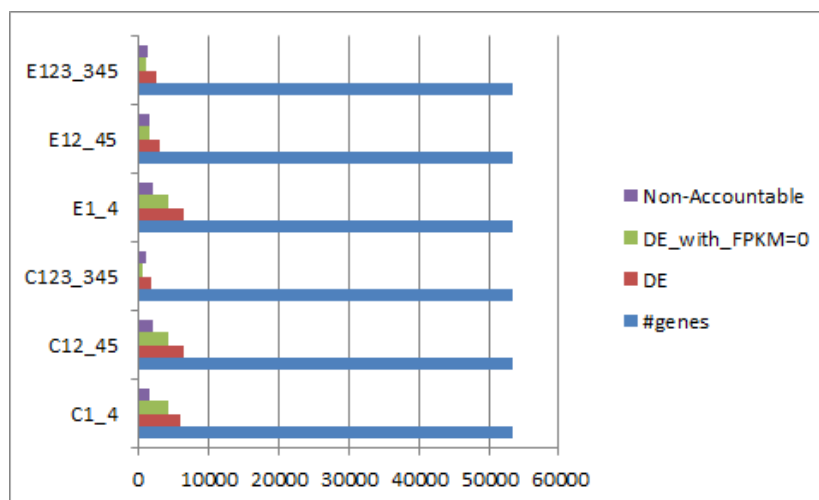
Figure 2.5 False positive rate in replicates experiment.

2.8.1.3 Discussion

The first observation to catch our attention is that for a handful of genes, IsoDE is (1) reporting non negligible fold changes (e.g. fold change 2), and (2) not calling them statistically significant based on their very low FPKM bootstrap distribution value. How can we explain these results?

We can argue that in case of (2) these genes have extremely low counts (usually very few counts in perhaps only one or two samples: experiment involving 2 or 3 replicates). The logic of IsoDE does not apply any independent filtering [29], such that the likelihood of a gene being significantly differentially expressed is related to how strongly it is expressed. This advocates for discarding extremely lowly expressed genes, because differential expression is likely not statistically detectable.

To reduce these cases of false positive, we can apply a higher bootstrap threshold such that only the genes which have a high fold change will be declared DE. When we have high count, the results are not surprising but with low count, there is too much unpredictability. We can also filter some genes based on their reads frequencies. For example, we can remove genes without at least 2 counts per million in at least two samples. After filtering, we expect the number of genes showing apparently large fold changes to reduce.

Even with filtering and thesholding, we still expect to see some DE genes. These can account for variability or noise in the preparation of replicates which is a very common issue in RNA-Seq experiments.

2.8.2 Factoring in inter-replicate variations

Future update to the tool will be, in case of replicates, to first perform some preprocessing. This will consist of setting up a data-generated bootstrap support to get rid of genes which account for noise and remove them from the analysis. This adjustement of our tool to deal with replicates is expected to reduce the false positive rate.

2.9   Update IsoDE: P-value computation

2.9.1   IsoDE - New IsoEM

The new version of the tool presents a superfast IsoEM generating bootstrap samples in a matter of few minutes. In the old version, bootstrapping was external to IsoEM, thus to compute x bootstrap samples, we were calling IsoEM x times. The new version eliminates this multiple calling of IsoEM algorithm and now everything is done internally. A typical IsoEM command now include as option the number of bootstrap sample along with confidence interval. Providing those two options enable IsoDE to run at least 10 times faster.

For example, to generate 20 bootstrap from a sample with about 65 millions reads can now require only 12 minutes while we needed 2 hours with the former version. For the same sample, 200 bootstrap can now be available after just 2 hours and not in about 2 days like in the past. In addition to the FPKM (Fragments Per Kilobase Million) format the former version used to report gene and isoforms frequencies, different output formats for abundance are now available:

- The eCPM (expected Count Per Million) is one of the new measure introduced. eCPM takes into account multireads by using the fractional allocation of reads to transcripts. To compute them we take the final expected read counts n(j) computed by IsoEM algorithm and normalize them to a sum of 1 million, i.e.,

  $eCPM(i) = 10^6 * n(i)/sum_j n(j)$

  This gives eCPMs for isoforms. The eCPMs for genes are computed by just summing over the isoforms of each gene.

- TPM (Transcripts Per Kilobase Million) the last format added in IsoEM is normalized over all reads count in a sample. Hence from the FPKMs, which are already available, to TPM is straightforward. For a gene i having j isoform(s)::

  $TPM_i = 10^6 FPKM_i/(sum_j FPKM_j)$

The new version of IsoEM includes not only a zipped file containing all bootstrap sample of isoform and gene frequencies estimation but also confidence interval for each gene expression level.Bootstrapped confidence intervals are an important feature since, in the most simplest case, it directly allow to flag some geneâĂŹs expression. The standard error/variance from bootstrap abundance estimates will tell us how reliable our expression estimates from aligned reads. For example, high variance of a geneâĂŹs expression informs us that our abundance estimate is not reliable. This could typically occur for genes with lower expression abundance and genes with large contributions from multi-mapping reads. For our DE analysis it is a very useful information.

All these new features along with some optimization on the code enable IsoDE 2.0 to drastically reduce running time, and also improves performance over the existing technologies, sequencing depths, and minimum fold change thresholds.

2.9.2 IsoDE - Kallisto

Kallisto [30] is a new RNA-Seq quantification approach. The results in the kallisto paper indicate that the software is not just fast, but also very accurate. Kallisto underlying RNA-Seq analysis are the alignments, and although kallisto is pseudoaligning instead, it is almost always only the compatibility information that is used in actual applications. This is the same approach perform in the Sailfish paper [31].As we the authors show in their paper, from the point of view of compatibility, the pseudoalignments and alignments are almost the same.

The way Kallisto does this is to directly assess, for each read, which transcripts it is compatible with, by checking the compatibility of the k-mers in each read (for some suitable k=31).

Essentially, pseudoalignments define a relationship between a read and a set of compatible transcripts (this relationship is computed based on *mapping* the k-mers to paths in a transcript De Bruijn graph). The pseudoalignment here gives us more information than the set of individual k-mers, since the k-mers in a read remain coupled when the read is

assigned to a transcript equivalence class. As the pseudoalignments are generated, equivalence classes are computed and maintained similar to the concept of equivalence classes among reads as introduced in the IsoEM paper or to equivalence classes among k-mers as used in Sailfish [31] or in Salmon [32]. Another important speed-inducing feature is that kallisto only looks for exact matches of k-mers to the transcriptome.

The speed of kallisto speed also enables bootstrap implementation to estimate the confidence interval for each transcript expression abundance

Because Kallisto compares with IsoEM and it does already generate bootstrap samples, it naturally becomes a choice for a new version of IsoDE analysis. Thus, this thesis also presents a configuration of IsoDE which integrates Kallisto.

2.9.3   Comparison IsoDE: IsoEM-Old, IsoEM-New, Kallisto

2.10   Conclusions

A practical bootstrapping based method, IsoDE, was developed for analysis of differentially expressed genes in RNA-Seq datasets. Unlike other existing methods, IsoDE is non-parametric, i.e., does not assume an underlying statistical distribution of the data. Experimental results on publicly available datasets both with and without replicates show that IsoDE has robust performance over a wide range of technologies, sequencing depths, and minimum fold changes. IsoDE performs particularly well on low coverage RNA-Seq datasets, at low fold change thresholds, and when no or very few replicates are available.

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
1	FisherTotal	34.01%	30.49%	95.63%	46.24%
	FisherHousekeeping	24.53%	20.12%	**97.74%**	33.38%
	GFOLD	55.62%	54.18%	92.11%	68.23%
	IsoDE-IsoEM-Match	**78.54%**	82.9%	80.36%	81.65%
	IsoDE-IsoEM-All	79.3%	**84.1%**	79.9%	**81.9%**
	IsoDE-Kallisto-Match	77.81%	82.04%	80.06%	81.04%
	IsoDE-Kallisto-All	78.39%	82.97%	79.41%	81.15%
1.5	FisherTotal	48.18%	35.37%	89.81%	50.75%
	FisherHousekeeping	42.48%	24.86%	**97.74%**	39.63%
	GFOLD	62.19%	58.13%	85.39%	69.17%
	IsoDE-IsoEM-Match	**73.14%**	**82.79%**	76.10%	79.30%
	IsoDE-IsoEM-All	71.5%	80.3%	78.7%	79.5%
	IsoDE-Kallisto-Match	70.80%	79.54%	78.79%	79.16%
	IsoDE-Kallisto-All	71.53%	81.84%	77.68%	**79.70%**
2	FisherTotal	57.96%	39.53%	85.43%	54.05%
	FisherHousekeeping	55.33%	29.30%	**97.67%**	45.08%
	GFOLD	69.05%	61.16%	83.49%	70.60%
	IsoDE-IsoEM-Match	**74.60%**	80.47%	78.10%	**79.27%**
	IsoDE-IsoEM-All	72.0%	80.9%	74.5%	77.6%
	IsoDE-Kallisto-Match	74.16%	80.23%	78.05%	79.13%
	IsoDE-Kallisto-All	72.41%	79.30%	76.12%	77.68%

Table 2.7 Accuracy, sensitivity, PPV and F-Score in % for the Second 454 dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and $M = 20$ for IsoDE-All, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

Fold Change	Method	Accuracy %	Sensitivity %	PPV %	F-Score %
	FishersTotal	70.41%	70.79%	91.24%	79.72%
	FishersHousekeeping	65.60%	65.22%	95.05%	77.36%
1	GFOLD	78.13%	80.06%	92.67%	**85.90%**
	Cuffdiff	11.37%	6.96%	**100.00%**	13.01%
	edgeR	73.03%	73.26%	95.56%	82.94%
	IsoDE-Kallisto-All	65.8%	65.9%	94.5%	77.7%
	IsoDE-IsoEM-Match	82.3%	86.7%	83.8%	85.2%
	IsoDE-IsoEM-All	**82.9%**	**87.9%**	83.0%	85.4%
	FishersTotal	74.05%	78.20%	84.85%	81.39%
	FishersHousekeeping	76.68%	73.61%	93.67%	82.44%
1.5	GFOLD	79.15%	79.35%	90.41%	84.52%
	Cuffdiff	28.43%	8.60%	**100.00%**	15.85%
	edgeR	80.01%	79.92%	92.07%	85.57%
	IsoDE-Kallisto-All	76.2%	77.1%	89.4%	82.8%
	IsoDE-IsoEM-Match	**80.1%**	87.2%	85.2%	**86.2%**
	IsoDE-IsoEM-All	79.9%	**88.0%**	83.8%	85.8%
	FishersTotal	78.43%	81.86%	82.44%	82.15%
	FishersHousekeeping	81.20%	80.00%	88.21%	83.90%
2	GFOLD	82.94%	78.84%	92.37%	85.07%
	Cuffdiff	40.96%	10.47%	**100.00%**	18.95%
	edgeR	**83.67%**	81.63%	91.17%	**86.13%**
	IsoDE-Kallisto-All	79.3%	85.6%	81.6%	83.5%
	IsoDE-IsoEM-Match	82.3%	86.3%	84.9%	85.6%
	IsoDE-IsoEM-All	80.9%	**87.0%**	82.2%	84.5%

Table 2.8 Accuracy, sensitivity, PPV and F-Score in % for MAQC Illumina dataset and fold change threshold f of 1, 1.5, and 2. The number of bootstrap samples is $M = 200$ for IsoDE-Match and IsoDE-All and $M = 20$ for IsoDe-Kallisto, and bootstrap support was determined using the binomial model with significance level $\alpha = 0.05$.

PART 3

INFERRING METABOLIC PATHWAY ACTIVITY LEVELS FROM RNA-SEQ DATA

3.1 Introduction

For the past several years, RNA-Seq has revolutionized biological research through the many advantages it provides. Because of RNA-Seq, it is easier to characterize transcripts and their isoforms, to detect genes without need of prior information in the form of probes or primers, and estimate expression level of transcript with good precision.

In contrast to microarray data, RNA-Seq data allows frequency of expression of all transcripts without a priori knowledge of the gene sequence. RNA-Seq data can account for the entire RNA volume producing enzyme for a given pathway. When applied to metatranscriptome data, the first challenge of pathway analysis is to decide which metabolic pathways are active in the sampled community (i.e., pathway activity detection). Recent software tool (*MEGAN4* [33] and *MetaPathways* [34] using SEED and KEGG (Kyoto Encyclopedia of Genes and Genomes) [35] annotations) enable the organization of transcripts or reads into ortholog groups and pathways by collecting all transcripts or reads represented by at least one ortholog group and providing that collection to the user. The parsimonious approach *MinPath* [36] identifies the smallest family of pathways covering all expressed ortholog groups. A more elaborate Markov Chain Monte Carlo (MCMC) approach takes into account the co-occurrences of genes in more than one pathway for analyzing metagenomic data [37]. Following pathway detection, the second major challenge of pathway analysis is to infer pathway activity levels to enable detection of differential expression. Few existing tools incorporate this step, a major focus of this paper.

Methods that treat pathways as simple gene sets [38, 39] are popular even though they do not use all information available. In recent years, a number of pathway analysis

methods have been developed that combine knowledge of pathway topology (e.g., gene position on the pathway, gene-gene interactions, etc.) with gene expression data based on comparative analyses (reviewed in [40]). Such methods have been applied primarily to experimental studies of single organisms. Despite the inherent pathway architecture of microbial biochemical function, relatively few analyses of complex metatranscriptomic datasets incorporate pathway-level inference of metabolic activity.

The new analysis techniques, presented here, is suitable for RNA-Seq data analysis to investigate the underlying metabolic differences between species living in separate environments.

Our contribution consists of the following:

1. A novel graph-based approach to analyze pathways significance. We represent metabolic pathways as graphs that use nodes to represent biochemical compounds, with enzymes associated with edges describing biochemical reactions.

2. An implementation of an EM algorithm, in which pathways are viewed as sets of orthologs.

3. The validation of the two approaches through differential expression analysis at the transcripts and genes levels and also through real-time quantitative PCR experiments.

Pathways can also be regarded as a set of ortholog group on which we can apply a set cover. We will use a binary ortholog group expression model to determine if there is or not RNA-Seq evidence for the expression of a given ortholog group in a given sample.

The validation step of these methods consists of extracting the genes involved in our estimated differential pathways activity levels, and analyzing their expression levels or transcript frequency estimation. We expect to see the differential pathway activity confirmed at the protein and contigs level. We carry this final analysis through the novel bootstrapping tool IsoDE [41].

Our experimental study was performed with RNA-Seq data from the marine bry-ozoan, *Bugula neritina*. Using the two novel computational approaches we implemented, we were able to find differentially expressed pathways from the data. This result is been validated by quantitative PCR (qPCR) conducted using a housekeeping gene also iden-tified in the data. The housekeeping gene, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), was chosen to normalize the qPCR data, as is standard practice. Based on our results, we applied qPCR experiments to quantify transcripts in the following identified pathways: Fatty acid elongation (ko00062), Peroxisome (ko04146) and Ribosome biogen-esis in eukaryotes (ko03008) [35].

The rest of the paper is organized as follows. Our formal models to analyze path-ways and to infer pathway activity are presented in the next section entitled Methods. Following Methods is the Differential-analysis section in which we present how we com-pute differential activity between pathways. We finish by presenting and discussing our results on *Bugula neritina* data in the Results section. The paper is concluded with possible future work.

3.2   Methods

In this paper we introduce a graph-based and an expectation maximization approach to identify specific differences between biological systems on the level of ortholog groups and pathways.

Figure 1 presents the entire flow of XPathway tools. In the graph-based approach, we compute a p-value using parameters extracted from the network to answer two dif-ferent statistical questions: (1) When and based on what parameter can we say that a set of proteins significantly map to a pathway? (2) What is the probability of finding such a mapping by chance given the data (transcripts/reads/proteins) and a pathway topol-ogy? Finally, significant metabolic pathways are selected by comparing the p-value of the original pathway with the ones from different bootstrapped samples. The expectation maximization method on the other hand uses the interaction among identified ortholog
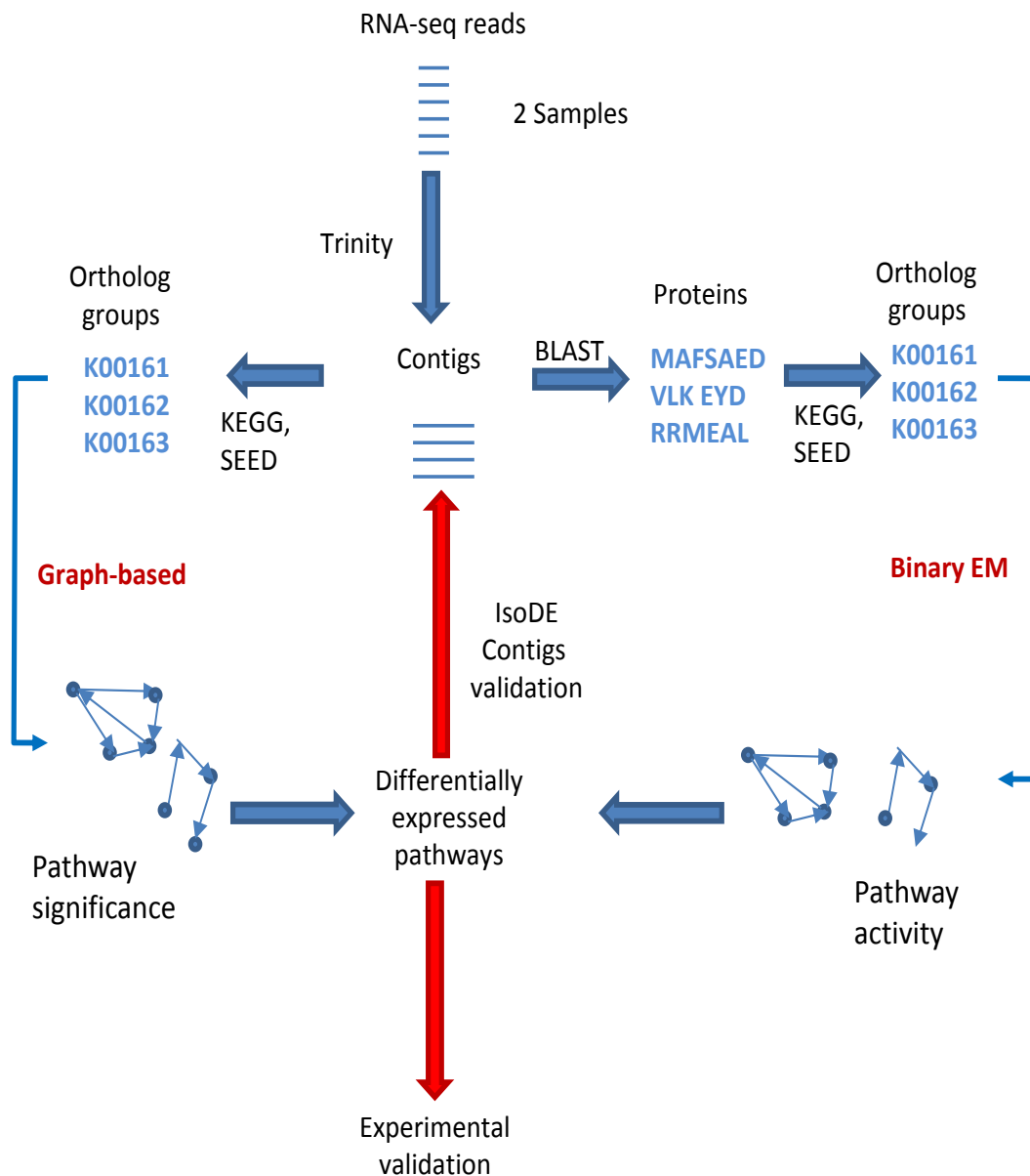
Figure 3.1 The XPathway tools analysis flow. The branches represent the two approaches used to compute pathway significance in the case of graph-based on the left and pathway activity level in the case of the expectation maximization approach on the right. Both methods are validated by computing contigs/transcripts differential expressions and qPCR as the last step of the flow.

groups to decide on a pathway activity. The last part of the flow consists of validating both branches. First, we compute differential expression analysis on all contigs extracted from pathways output by both branches. Secondly, a qPCR experiment is carried out on

the contigs which have a fold change of 1.5 or more.

3.2.1 Expectation maximization model of pathway activity

In this section we present an EM-based algorithms for inferring pathway activity levels based on metatranscriptome sequence data. Let w be a pathway that is considered to be a set of enzymes represented by their ortholog groups $w = \{p_1, \ldots, p_k\}$. Since an ortholog group can have multiple functions and participate in multiple pathways, the pathways can be viewed as a family of subsets W of the set of all ortholog groups P. Below we start by introducing a uniform binary pathway activity model based on a discrete ortholog group expression model.

The uniform binary pathway activity model is based on the assumptions of *uniformity*, namely that each molecule from an ortholog group participates in each active pathway with the same probability (i.e., in equal proportions) and of *binary activity*, which postulates that a pathway is active if the level of ortholog group activity exceeds a certain (possibly pathway dependent) threshold. Formally, let $\delta(w)$ be a binary variable indicating the *activity status* of w, i.e., $\delta(w) = 1$ if w is active and $\delta(w) = 0$, otherwise. Also, let the *activity level* of pathway w be the summation over constituent ortholog groups g of their participation g_w in w. Since we assume that each ortholog group g is equally likely to participate in each pathway containing it, it follows that $g_w = (1 + \sum_{w' \ni p, w' \neq w} \delta(w'))^{-1}$ and the activity level f_w of pathway w is given by

$$f_w = \sum_{g \in w} g_w = \sum_{g \in w} \frac{1}{1 + \sum_{w' \ni g, w' \neq w} \delta(w')} \tag{3.1}$$

The binary activity status of w is computed from its activity level f_w and the threshold T_w as follows

$$\delta(w) = \begin{cases} 0 & \text{if } f_w < T_w \\ 1 & \text{if } f_w \geq T_w \end{cases} \tag{3.2}$$

The uniform binary model described by equations (3.1)-(3.2) can be solved using a sim-

ple iterative algorithm. The algorithm starts with assigning activity status $\delta(w) = 1$ to each pathway $w \in W$, i.e., $\Delta^0(W) = \{\delta^0(w)|w \in W\} \leftarrow 1$ and then repeatedly updates the activity level according to (3.1) and the activity status according to (3.2). The procedure terminates when the status sequence $\Delta^0(W) = 1, \Delta^1(W), \Delta^2(W), \ldots$ starts to oscillate $\Delta^{n+k}(W) = \Delta^n(W)$ or converges. In all our preliminary experiments, an oscillation with period $k = 2$ is achieved in at most 10 iterations. Also the threshold T_w does not significantly change the order of pathways sorted with respect to their activity levels estimated as the mean f_w after convergence. The model is represented in Figure 2.
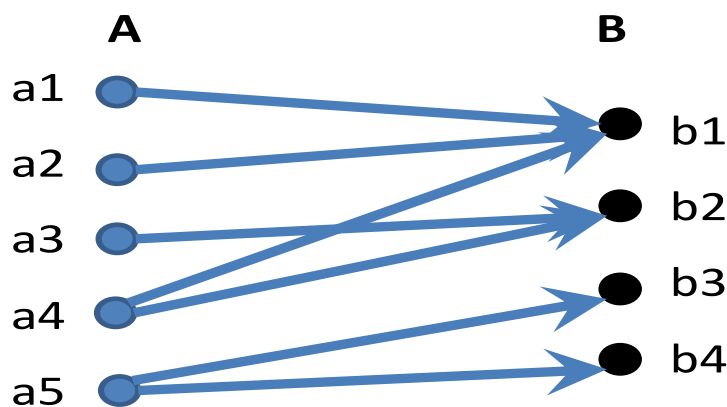
Figure 3.2 Expectation maximization approach to compute pathway activity. This bipartite graph consists of a set A representing reads/contigs/ORF/proteins and the set B is for ORFs/proteins/ortholog groups/EC (Enzyme Commission) numbers. The arcs represent mapping between elements of both sets. For our binary EM, the set A consists of contigs mapped to ortholog groups and the weight of each arc is 1.

Although the uniform binary model allows the computation of pathway activity by assigning ortholog groups to pathways, it does have some limitations hindering it for capturing specific attributes of the metabolic network. For example, the binary uniform model assigns only value 1 or 0, if the ortholog group belongs to a pathway or not, respectively. This yes or no assumption is not always true since there may be a fractional part of an ortholog group belonging to different pathways. Moreover, the uniformity model is not easily applicable to natural processes because all assignment are never equally likely. Finally the model is not completely stable but rather periodic with some subsets of or-

tholog groups fluctuating between pathways.

3.2.2 Graph-based estimation of pathway significance

Ideally, a comprehensive pathway analysis method would take into consideration the position and role of each gene in a pathway, the efficiency with which a certain reaction is carried out, and some limiting factors (e.g. dealing with metagenomics data or not). With genome data, it is possible to consider pathways size, gene length and overlap in gene content among pathways [37] to compute the relative abundance of pathways and pathway ranking, but this approach might not work with RNA-Seq data especially in the absence of a genome reference.

Henceforth, in our second approach, each pathway is viewed as a network of enzymes also called EC numbers (Enzyme Commission numbers) in order to compute their statistical significance. Significance of a pathway activity on the sample is measured by the randomness of the positions of matched enzymes in the corresponding KEGG pathway graph. The randomness is measured using a permutation model for finding significant pathway alignments and motifs [42].
This model assumes that the subset of expressed enzymes in an active annotated pathway should be connected. The enzyme permutation model finds the average vertex degree in the subgraph induced by expressed enzymes. Then the same parameter is computed for sufficiently many random permutations of enzyme labels. The statistically significant match should have density higher than 95% of permutations. Specific characteristics of the graph taken into account in our analysis are:

1. Number of nodes. A node represents a protein that got mapped during BLAST. KEGG usually assigns a green color to those proteins in the graph.

2. Density = (Number of edges) / (Number of vertex - 1)

3. Number of 0 in and out-degree vertex. Let call this number x. x is defined by:

x = ((number of vertex with out-degree = 0) + (number of vertex with in-degree=0))

/ 2 * (number of nodes)

4. We also consider other criteria such as (1) Number of green connected components, (2) Largest Number of nodes in connected component and (3) Largest Number of edges in connected component.

Using these metrics, we compute the density of the induced graph composed of only mapped proteins. We obtain the names of those proteins through EC numbers on the graph. Below, we present two graph-based models, the vertex label swapping and the edge swapping for random graph generation, to analyze pathways. This model is explained by the left side of Figure 1 and Figure 3 presents an example when we permute labels of two vertices.

Model 1: Vertex label swapping for random graph generation
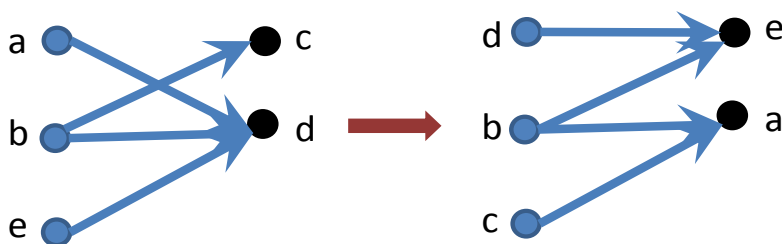
Figure 3.3 Vertex labels swapping model for random graph generation. We only swap vertices which have different labels. A label is an attribute of a vertex representing a mapped or not protein.

In this model, we keep the same topology but we allow swapping of labels between two vertices. One known issue of this approach is, vertices with high degree always get connected. This might lead to too many significant matches, thus increasing the false positive rate. The vertex label swapping algorithm can be represented as follows:

Model 2: Edge swapping for random graph generation

Because of the bias in the vertex label swapping model, we also implemented the edge swapping. Here, the idea is to keep the in-degree and out-degree of each node the same,

---

Algorithm 1 Vertex labels swapping algorithm for random graph generation

---

1: $G = (V, E) \leftarrow$ Original Graph
2: $i \leftarrow 1$
3: $j \leftarrow 1$
4: **for** $i \leq m$ **do**
5:    **for** $j \leq n$ **do**
6:       `randomly pick two vertices a and b from V`
7:       **if** no $(label(a) == label(b))$ **then**
8:          `swap label of a and b`
9:       **end if**
10:   **end for**
11: **end for**
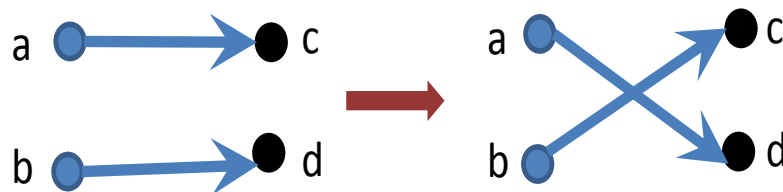
Figure 3.4 Edge swapping model for random graph generation. Before swapping the edges, we check that the in and out-degree of the vertices involved remain the same.

swapping nodes only if these values do not change. We keep vertex labels the same.

Figure 4 presents an example when we permute two edges.

The edge swapping algorithm can be represented as follows:

Algorithm 2 Edge swapping algorithm for random graph generation

---

1: $G = (V, E) \leftarrow$ Original Graph
2: $i \leftarrow 1$
3: $j \leftarrow 1$
4: **for** $i \leq m$ **do**
5:    **for** $j \leq n$ **do**
6:       `randomly pick two edges a` $(a0 \rightarrow a1)$ `and b`$(b0 \rightarrow b1)$ `from E`
7:       **if** no $(a0 == b0)$ or $(a0 == b1)$ or $(a1 == b0)$ or $(a1 == b1)$ and no $(a1 \rightarrow a0$ or $b1 \rightarrow b0)$ **then**
8:          `remove edges a and b`
9:          `create new edges:` $a0 \rightarrow b1$ `and` $b0 \rightarrow a1$
10:       **end if**
11:   **end for**
12: **end for**

3.2.3   Differential analysis of pathway activity and significance

3.2.3.1   *Differential analysis of pathway activity*

The goal of this analysis is to determine which pathway needs to be considered more closely to understand the difference in the metabolism of two organisms. For this purpose, we use the pathway expression computed from the binary model presented earlier. First we compute expression of each pathway present in the set of pathways we get from KEGG for a given sample. Then we compute the difference between the expression of each pathway. Under this model, the pathways selected as having differential activity are the ones where the ratio of their expression is greater than a certain threshold. Table 3 presents our results on differential analysis of pathway activity.

3.2.3.2   *Differential analysis of pathway significance*

Differential analysis of pathway significance is based on the $p\text{-}value$ described in the graph-based sub-section of Methods. We randomly permute each pathway graph generating m different graphs. Note that even the smallest pathway graphs contains at least 15 nodes and about 40 edges which is sufficient to generate default $m = 200$ distinct random graphs. A pathway is significant if the $p\text{-}value$ of the mapping is less than 5%. The $p\text{-}value$ is the position of the original graph when placed in the sorted list of all randomly generated graphs sorted first by "density" (largest to smallest) and then by the number of nodes having 0 in-degree or 0 out-degree (smallest to largest). A pathway is *significant* if its $p\text{-}value$ is less than 5%, *very significant* if its $p\text{-}value$ is less than 1% and *the most significant* if its $p\text{-}value$ is less or equal to 0.5%.

Let $p1$ be the $p\text{-}value$ for pathway X in sample 1 and let $p2$ be the $p\text{-}value$ for pathway X in sample 2. We say that pathway X is differentially significant between the two samples if the probability computed by the equation of $\text{probDiff}(X)$ below is greater than 50%.

$$
\text{probDiff}(X) = \begin{cases} (1-p1)*p2 & \text{if } p1 \geq p2 \\[2mm] (1-p2)*p1 & \text{if } p2 < p1 \end{cases}
$$

For example, let us consider $m = 200$ randomly generated graphs and the vertex label swapping model. In Figure 5 representing part of the Fatty acid elongation pathway (ko00062), the mapped enzymes (filled rectangles) in sample 1 form a sub-graph with density = 1.875 and the number of 0 in/out degree = 0.11 for that sub-graph. After sorting the graph, the position of our original graph is the first, hence p-value p1 = 0.5% (most significant pathway given the 200 graphs). In Sample 2, the mapped enzymes (filled rectangles) form a sub-graph with density = 1.375, number of 0 in/out degree = 0.22 for that sub-graph and its position after sorting is 148. This results in a p-value p2 = 74.5% (not a significant mapping).

Based on the value of p1 and p2, $\text{probDiff}(\text{ko00062}) = .74$ which is greater than 50%. We conclude that ko00062 is differentially significant in the two samples.
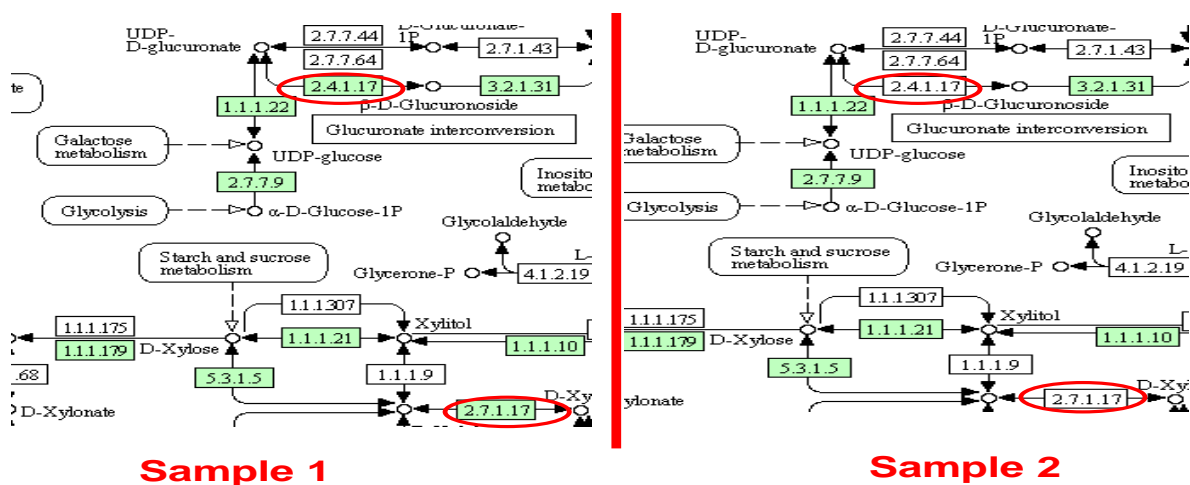
Figure 3.5 Pathway differential analysis. In sample 1, the mapped enzymes (filled rectangles) form a sub-graph with density = 1.475, the number of 0 in/out degree = 0.11 and p-value=0.5. In Sample 2, the mapped enzymes (filled rectangles) form a sub-graph with density = 1.375, the number of 0 in/out degree = 0.22 and p-value=.74. Based on these p-value, we say that this pathways is differentially significant.

3.3 Results and Discussion

3.3.1 *Bugula neritina* data preparation

Bugula neritina is a colonial marine invertebrate found in temperate waters around the world [43]. *B. neritina* associates with an uncultured microbial symbiont, âĂIJ*Candidatus* Endobugula sertulaâĂİ [44] that has been shown to produce bryostatins, bioactive compounds that protect the host larvae from predation [45, 46]. Because of the pharmaceutical potential of the bryostatins, and the inablilty to grow the symbiont in the laboratory, we chose to examine host gene expression in the presence and absence of the symbiont to understand more about the molecular underpinnings of this relationship. In addition, as symbiont endow the larvae with high concentrations of bryostatins compared to the adult [47, 48], we also wanted to examine host gene expression in portions of the colony that possess reproductive structures termed ovicells, and those without ovicells.

Adult colonies of *B. neritina* growing on floating docks were collected from four locations on the Eastern coast of the USA: Radio Island Marina and Yacht Basin Marina, Morehead City (North Carolina) in March 2012, and Oyster public docks, Oyster (Virginia) in June 2012. The colonies were rinsed in filtered sea water and preserved in TRIzol reagent (Invitrogen, Carlsbad, CA) at -80 degree celsius prior to RNA extraction. Total RNA was then extracted from the preserved samples (RNeasy Mini kit, Qiagen,Inc., Valencia, CA, USA). The RNA was purified and treated with RNase-free DNaseI to remove any contaminating DNA. The purified total RNA was processed according to standard operating procedure for preparation of mRNA library for sequencing (TruSeq RNA Sample Preparation Kit, Illumina, San Diego, CA, USA). The adapter-ligated cDNA library was hybridized to the surface of Illumina flow cell and sequenced on an Illumina HiSeq 2500 sequencing platform. The paired-end reads were assembled de novo using Trinity software package [49] and the assembled contigs were annotated by performing blastx searches (Translated Query-Protein Subject BLAST 2.2.26+) against the Swiss-Prot database. A complete description on how the samples were obtained, cleansed and sequenced is available in this

paper [50].

The samples used for analysis include: Lane 1: symbiotic; Lane 2: non symbiotic (aposymbiotic) [51]; Lane 3: symbiotic, with ovicells; and Lane 4: symbiotic, without ovicells. The symbiotic relationship was assessed in the collected colonies by PCR analysis for the presence of the bryostatin biosynthetic gene cluster gene, bryS [51, 52]. We report only the values averaged over all 4 lanes. The reads were assembled into contigs by Trinity. We also conducted a Trinity assembly on the union of all reads - about 221,818,850 of them - 2x50 bp using 22 Gb of space. We obtained 166,951 contigs, after filtering with RSEM-isopct-cutoff=1.00, also 76,769 ORFs, 37,026 BLAST hits of translated ORFs against the SwissProt database and around 12,748 proteins hits. This translates to 59.37% ORFs hits and 63.35% contigs hits. Using IsoDE, we were able to identify 1485 differential expressed genes between the two different conditions, the symbiotic and aposymbiotic *B. neritina*.

3.3.2   Pathway extraction and graph generation

By *de novo* co-assembly of RNA-Seq data and BLAST-ing resulting contigs against protein databases, with a certain confidence, we can infer the ortholog groups expressed in the sample. This is an important attribute of KEGG. We use KEGG to generate pathways from Trinity contigs and proteins. From the pathway databases we can easily extract the enzyme information associated with each pathway. We actually extracted all pathways along with all mapped proteins.

KEGG represents proteins as KO numbers and we also follow this representation. It uses KGML, an exchange format of KEGG pathway maps, to interact with external applications. The next step was to download all KGML files associated with the pathways using the API provided by KEGG. To convert KGML files to graph of node and vertices, we implemented and ran a novel tool called KGMLPathway2Graph. Mapping the output of KGMLPathway2Graph with KO numbers from KEGG analysis of our data, allowed us to compute pathway significance through $p\text{-}value$.

3.3.3 Results

Pathway expression differences in symbiotic and aposymbiotic *Bugula neritina* (Lanes 1 and 2) are shown in Table 1 and in Table 2 respectively following vertex labels permutation and the edges permutation models. We ran our graph-based models algorithm 1 with $m = 200$ graphs each generated after $n = 1000000$ permutations of labels or edges. This high number of permutations is necessary to introduce sufficient differences in the generated graphs. Table 3 presents the results on differential analysis of pathway activities between the same two *Bugula neritina* conditions.

Pathway	p-value from symbiotic Bugula	p-value from apo-symbiotic Bugula	ProbDiff
ko04146	.99	.05	.94
ko03008	.99	.05	.94
ko03013	.99	.05	.94
ko00983	.99	.05	.94
ko04530	.99	.05	.94
ko00062	.01	.75	.74
ko00400	.01	.99	.98
ko00071	.99	.01	.98
ko00100	.99	.01	.98
ko00910	.04	.99	.95
ko04122	.99	.03	.97
ko04713	.99	.01	.99

Table 3.1 Vertex label permutation: the p-values of pathways are computed from the symbiotic (Lane 1) and aposymbiotic (Lane 2) *B. neritina* data. This table presents the most significant divergence in pathway results, using the criteria described in the Methods section, they are declared differentially significant.

In the additional file, we present a summary of transcripts differential expression (DE) analysis results using IsoDE [41] and pathway activity inference results.

Pathway	p-value from sym-biotic Bugula	p-value from apo-symbiotic Bugula	ProbDiff
ko04146	.99	.05	.94
ko03008	.99	.05	.94
ko03013	.99	.05	.94
ko00983	.99	.05	.94
ko04530	.99	.05	.94
ko00400	.01	.99	.98
ko04122	.99	.03	.97
ko04713	.99	.01	.99
ko00130	.01	.75	.74
ko00120	.01	.99	.98
ko00072	.99	.01	.98
ko00120	.99	.01	.98
ko00230	.04	.99	.95
ko00627	.99	.03	.97
ko00770	.99	.01	.99
ko00980	.99	.03	.97
ko04630	.99	.01	.99

Table 3.2 Edge permutation: the p-values of pathways are computed from the symbiotic (Lane 1) and aposymbiotic (Lane 2) Bugula data. This table presents the most significant divergence in pathway results, using the criteria described in the Methods section, they are declared differentially significant.

3.3.4 Validation

From our statistical analysis, we identified some pathways that were differentially expressed (DE) by all methods. The next step was to experimentally validate these results. The first validation step is done through IsoDE, a software to analyze differentially expressed genes. Through KEGG, we are able to get all the proteins (contigs) participating in a pathway. IsoDE then indicates which of those contigs are also DE. From those DE contigs, we extracted the genes that will be tested via quantitative PCR (qPCR) experiment, the next validation step.

The goal of qPCR is to quantify the level of expression in the symbiotic and aposymbiotic *B. neritina*. It is used to validate the gene expression given by IsoDE. Primers were

Pathway	Expression from-symbiotic Bugula	Expression from-aposymbiotic Bugula	Ratio between-pathway expressions
ko00300	2.75	0.38	7.27
ko00290	4.26	1.71	2.49
ko04712	1.77	0.77	2.30
ko00903	1.88	0.84	2.25
ko01220	2.24	1.10	2.04
ko00981	2.00	1.00	2.00
ko04744	3.55	1.82	1.95
ko00626	1.48	0.76	1.93
ko00624	1.17	0.67	1.76
ko00072	2.17	1.25	1.74
ko00730	2.50	1.50	1.67
ko04730	3.80	2.40	1.58
ko00363	2.92	1.88	1.56
ko05150	2.13	1.38	1.55
ko04112	3.00	2.00	1.50
ko05219	1.67	2.50	0.66
ko00625	1.00	1.98	0.50
ko00984	1.00	2.00	0.50
ko00592	1.25	3.02	0.42
ko00965	0.66	1.66	0.40
ko00940	1.42	3.71	0.38
ko00460	0.60	2.02	0.30
ko00944	0.17	1.17	0.14

Table 3.3 Pathway activities levels with ratio. Expression represents the expression level of the pathway activity in symbiotic (Lane 1) and aposymbiotic (Lane 2) *B. neritina* data. This table presents pathways with a ratio of 1.5 or higher in their activity level or pathways with a ratio of 0.66 or lower from the opposite direction. Using the criteria described in section 2, they are found to significantly differ in activities level.

designed using Primer 3 Plus [53]. Total RNA from recently collected symbiotic and aposymbiotic *B. neritina* colonies was converted to cDNA using Superscript III (Invitrogen, Carlsbad, CA, USA) using random hexamers according to the manufacterâĂŹs instructions. cDNA was subjected to qPCR analysis after and expression in the samples was compared using the 2△Ct method [54]. The reference gene used with the glyceraldehyde-3-phosphate dehydrogenase gene, a housekeeping gene identified from the *B. neritina*

transcriptome (contig m.4423) [55, 56]. The expression of three genes identified as being differentially expressed in symbiotic and aposymbiotic animals from the fatty acid elongation pathway (ko00062) were compared.

Using IsoDE, nine gene pathways were chosen from KEGG and 2485 top differentially expressed contigs were taken from the list of all contigs. Within the selected pathways, there was a total of 637 contigs extracted. Each gene in these contigs was checked for fold change 1.2 or higher. Next, the number of genes that had significant fold change was compared to the total number of genes in the pathway. If the number of genes with fold change of 1.2 or higher account for at least 15% of the total number of genes in that pathway, then this pathway was chosen to be further investigated. The pathways chosen are fatty acid elongation (ko00062), peroxisome (ko04146), ribosome biogenesis in eukaryotes (ko03008), RNA transport (ko03013) and drug metabolism - other enzymes (ko00983) [35].

The fatty acid elongation pathway contains fourteen (14) KEGG mapped contigs and three (3) of those were found significantly differentially expressed. They are very-long-chain 3-oxoacyl-CoA reductase (DHB12), 3-ketoacyl-CoA thiolase (fadA) and 3-ketoacyl-CoA thiolase B, peroxisomal (THIKB). Once all the contigs in the pathway were checked, additional information was compiled: KEGG pathway and protein numbers, contig number, UniProt accession number and predicted fold change between symbiotic and aposymbiotic *B. neritina* [57]. qPCR primers were designed using Primer 3 Plus and ordered from Integrated DNA Technology [53]. The primers were tested using cDNA at concentrations from 1 ng/μL to 0.1 pg/μL. In order to use $\triangle\triangle$Ct method, every primer had to have the same efficiency and efficiency around 100% [54].

RNA was extracted from symbiotic and aposymbiotic *B. neritina* colonies. Following the Direct-zol RNA MiniPrep protocol (Zymo Research Corp., Irvine, California, USA), 50 mg of *B. neritina* tissue was homogenized and RNA was extracted. Then the RNA was further purified using the OneStep PCR Inhibitor Removal Kit (Zymo Research Corp.,

Irvine, California, USA). To eliminate any contaminating genomic DNA, a DNase I treatment was performed according to the manufacturerâĂŹs protocol (DNase I Recombinant, RNase-free, Roche, Mannheim, Germany). Finally, the RNA was further purified with the RNA Clean-up and Concentrator Kit (Zymo Research Corp., Irvine, California, USA). The concentration of RNA was quantified in triplicate using a Nanodrop spectrophotometer.

Both symbiotic and aposymbiotic cDNA were synthesized using the Superscript III protocol (Superscript III First Strand Synthesis System for RT-PCR, Invitrogen by Life Technologies, Carlsbad, California, USA) with random hexamers. qPCR primer efficiencies were determined using qPCR. All qPCR reactions were performed using 7500 Fast Real-Time PCR system (Applied Biosystems) with hot-start Taq polymerase, SYBR Green fluorescent dye and ROX passive reference dye (Maxima SYBR Green/ROX qPCR Master Mix (2X), Life Technologies, Carlsbad, California, USA). The efficiencies for each primer pair were calculated using the slope of amplification curve in the equation E = 10(-1/slope) [58].

The expression levels of three genes, *fadA*, DHB12, and THIKB were measured in symbiotic and aposymbiotic *B. neritina* with G3P acting as an endogenous control. The reactions were run in triplicate for symbiotic and aposymbiotic cDNA along with a negative template control. The Ct averages and standard deviations were calculated to find the Ct differences between the target gene and the control (△Ct) and △Ct standard deviation. △△Ct was calculated by subtracting the symbiotic or aposymbiotic △Ct by the symbiotic △Ct. This resulted in symbiotic △△Ct equal to 0 to compared the fold change between symbiotic and aposymbiotic expression levels.

As presented in Table 4, the fold change predicted by differential expression analysis, using IsoDE, for these three genes indicated that expression was higher in the aposymbiotic B. neritina. *fadA* had a predicted fold change of 2.91, while DHB12 had a value of 1.90 (non-significant difference), and THIKB equaled 2.84. qPCR analysis showed that when aposymbiotic gene expression was compared to symbiotic gene expression, *fadA*

had 6.88 higher expression in aposymbiotic *B. neritina*. DHB12 had 0.66 times lower expression and THIKB had 2.52 higher expression, indicating that computational method closely predicted expression in independent biological samples.

Genes	fadA	DHB12	THIK
Fold change of gene expression in aposymbiotic *B. neritina* compared to symbiotic by FPKM analysis	2.91	1.90	2.84
Gene expression in symbiotic *B. neritina* by qPCR analysis	2.46	4.34	4.34
Gene expression in aposymbiotic *B. neritina* by qPCR analysis	29.32	2.85	10.95
Fold change of gene expression in aposymbiotic *B. neritina* compared to symbiotic by qPCR analysis	6.88	0.66	2.52

Table 3.4 Experimental quantification of fatty acid elongation gene expression by qPCR in symbiotic and naturally aposymbiotic *B. neritina*.

3.3.5   Discussion

Although all the EM and the graph-based methods worked on the same data generated by KEGG, the input to each approach were very different. For example, the Trinity output of sample1 on KEGG generates about 306 pathways. All of these pathways were considered for EM methods while only a small subset of 80 was used as input to each of the graph-based model. Different factors contributed to this reduced number of pathways been analyzed in the edge/vertex swapping model: (1) We were not able to extract the KGML of all pathways from from KEGG; (2) We were not able to convert all KGML to actual graph and (3) Some graphs did not carry enough mapping to be significant (we excluded pathways with less than 3 ortholog groups mapped).

Consequently, the graph-based approaches yield considerable less differentially expressed pathways than EM methods although results from both models in the graph-based approaches were very consistent. Also, the graph-based analysis appears to be

more stringent selecting only the pathways which are the farthest apart according to our statistic criteria.

Looking at the overall small number of differential pathways, we can say that Bugula n. with or without the symbiotic relationship still exhibits very similar reaction. As shown in Table 5, the over all difference in the pathway activity and differentially expressed transcripts between the two samples is very small.

Because working with non-model organisms, such as *B. neritina*, is more challenging due to the lack of tools for genetic manipulation, for future works, we plan on the following. First, run XPathways tools in other organisms, including model organisms to further verify their efficacy and second, extend the model to handle not only metabolic pathways put also signaling pathways.

	FC = 2	FC = 1.5
Pathway	8%	12%
Contigs	13%	28%

Table 3.5 Percentage of differentially expressed contigs with fold change (FC) of 2 and 1.5 respectively.

3.4   Conclusions

XPathway tools with its two computational approaches is able to efficiently infer pathway activity as well as pathway significance, respectively an expectation maximization and a graph-based approach. Rather than trying directly to identify differentially expressed genes from RNA-Seq data for a non-model organism, XPathway tools allows to more accurately predict differential expression of genes using wealth of information collected in the KEGG database for related organisms. Our experimental comparisons on *Bugula neritina* RNA-Seq data with or without the symbiotic bacteria enable the detection and comparison of pathways with metabolic difference. qPCR experiment successfully validated our findings.

PART 4

SOFTWARE PACKAGES AND TOOLS

4.1   Software package

4.1.1   XPahway Tools

- **XPahway Tools** constitutes a set of tools that compares metabolic pathway activity analyzing mapping of contigs assembled from RNA-Seq reads to KEGG pathways. The XPathway analysis of pathway activity is based on expectation maximization and topological properties of pathway graphs. The different tools that constitute XPathway are:

  - KGMLPathway2Graph: Extraction tool for metabolic network. KGMLPathway2Graph aims at extracting metabolic pathways from KGML flatfile database.

  - Link Gopher 1.3.3: Mozilla Firefox add-ons Link gopher is used to copy all green nodes in each pathway from KEGG output. These nodes are part of the pathway urls.

  - java code: To extract all green nodes per pathways for further analysis.

  - Python code: To compute pathway activity level and significance.

  - shell script: To download all KGML file from Kegg using wget. This is a one time operation since ko xml file do not change.

  - Infer Pathway activity level pipeline and Pathway significance pipeline All steps for the analysis are provided and explained in the Readme file.

  $http://alan.cs.gsu.edu/NGS/?q = content/xpathway$

4.1.2   IsoDE

- **IsoDE** is a software package that can be used to perform differential gene expression

analysis for RNA-Seq data both with and without replicates. IsoDE is based on boot-strapping, which provides a principled way to test for differential expression based on fold changes obtained from FPKM estimates obtained by resampling the original read alignments. This strategy can be used in conjunction with any method for esti-mating individual gene expression levels from aligned RNA-Seq reads; in the IsoDE implementation we rely on the IsoEM algorithm, a scalable expectation-maximization algorithm that takes into account gene isoforms in the inference process to ensure ac-curate length normalization. Experiments on MAQC RNA-Seq datasets without repli-cates show that IsoDE has consistently high accuracy as defined by the qPCR ground truth, frequently outperforming existing methods such as Fishers exact test, edgeR, GFOLD, and Cuffdiff, particularly at for low coverage data and at lower fold change thresholds. In experiments on MCF-7 RNA-Seq datasets with up to 7 replicates IsoDE also achieved high accuracy that varies smoothly with the number of replicates and is relatively uniform across the entire range of gene expression levels.

The software is written in Java so it can be run on any platform with a java virtual machine. The source code is distributed with the installation package.

IsoDE is available in three different packages:

- IsoDE coupled with IsoEM with external bootstrap generation

- IsoDE coupled with IsoEM with internal bootstrap generation

- IsoDE coupled with Kallisto using kallisto bootstrap sample and IsoDE analysis.

$http : //dna.engr.uconn.edu/?page_id = 517$.

4.1.3 IsoEM version 1.1.4

- **IsoEM** package can be used to infer isoform and gene expression levels from high-throughput transcriptome sequencing RNA-Seq data. IsoEM uses a novel expectation-maximization algorithm that exploits read disambiguation information provided by the distribution of insert sizes generated during sequencing library preparation, and

takes advantage of base quality scores, strand, and read pairing information (if available). Empirical experiments on synthetic datasets show that the algorithm significantly outperforms existing methods of isoform and gene expression level estimation from RNA-Seq data.

The software is written in Java so it can be run on any platform with a java virtual machine. The source code is distributed with the installation package.

$\text{http}://\text{dna.engr.uconn.edu}/?\text{page}_i\text{d} = 105$

4.1.4 DORFA

- **DORFA** translated as Database-guided ORFeome Assembly from RNA-Seq Data is a novel tool for protein database guided ORF assembly. DORFA takes as input the set of partial ORFs produced by an RNA-Seq assembler and builds from them complete ORFs. The biological value of the tool is very important, since it complements the output of a de-novo RNA-Seq assembler. Finding the exhaustive set of ORFs can be crucial for accurate protein activity level estimation or for pathway reconstruction (i.e. missing some proteins one can not tell exactly whether a certain pathway is present in an organism or not).

4.2 Tools and Applications

4.2.1 The Etheostoma tallapoosae Genome Annotation pipeline

The family Percidae contains over 200 species, most of which are within the subfamily Etheostomatinae. This subfamily (the darters) represents a species rich radiation of freshwater fishes in North America. Evolutionary relationships between the various darter species have been deduced from morphological, mitochondrial DNA sequence and limited nuclear DNA sequence comparisons. However, a thorough understanding of the evolution of the darter species will require comparisons at the whole genome level.

DNA was extracted from a single Tallapoosa darter utilizing the Qiagen DNeasy

Blood and Tissue kit. This DNA was then prepared for sequencing and sequenced with an Illumina MiSeq by the Georgia Genomics Facility at the University of Georgia. The DNA was sheared to an average size of about 700 bp and libraries were prepared with the TruSeq sample preparation kit. Two 250PE runs were performed on a MiSeq instrument. Two barcoded libraries were sequenced in the first run (this was necessary because this run also contained a small percentage of barcoded amplicons for a different study). The second run just sequenced a single genomic DNA library.

As a first step, the genome of the Tallapoosa darter (Etheostoma tallapoosae) has been sequenced utilizing two Illumina MiSeq 250-PE runs generating 52 million reads. This provided an average 12 fold coverage of the estimated 1 billion nucleotide genome. The sequences were assembled with Minia into contigs and these were assembled into scaffolds with SSPACE.

For an initial assembly of the sequences into longer contigs, the three libraries were combined into one fasta file. This file was used as the input for the Minia short read assembler software. The only parameters that can be varied are the k-mer length and minimum abundance. The sequences were assembled for most k-mer lengths between 31 and 75 at minimum abundance settings of 2 and 3. The longest contigs and the most nucleotides assembled into contigs were obtained at k-mer 61 with minimum abundance setting of 3 and at k-mer 73 with minimum overlap abundance of 2.

An initial assessment to determine if the sequence reads were correctly assembled was performed by finding scaffolds which shared sequence similarity with the LPLII, PLC and RPS6 Tallapoosa darter genomic sequences described above. A BLAST server was set up and databases were created from each of the scaffold assemblies. A BLAST server allows for the identification of Tallapoosa darter scaffolds homologous to sequences of interest.Scaffolds that contain sequences homologous to each of the three Tallapoosa darter reference sequences were found with blastn searches. These scaffolds were then paired with the reference sequences.

The scaffolds were also imported into an instance of WebApollo along with gene

evidence tracks generated by fgenesh. A set of scripts were developed to facilitate the formatting and import of these tracks and scaffold sequences into WebApollo that will make it simple for labs to set up WebApollo instances for their own genome data without extensive computer system experience. A web site has been developed that gives access to both the BLAST and WebApollo servers to the public to spur interest in darter genomics and to enable annotation of the Tallapoosa darter genome by a community of darter researchers.

$http://www.dartergenomics.org/tallapoosa-darter-genome$

$https://pag.confex.com/pag/xxii/webprogram/Paper12567.html$

4.2.2 Vispa Plugin

VISPA [59] algorithm has seen substantial changes in the past years. The motivation behind the adjustments was to:

1. Normalize the frequencies of reconstructed quasispecies

2. Produce better alignment and ultimately to speed the running time of the algorithm.

The adjusted VISPA software now incorporates an enhanced approach of estimating frequencies using expected maximization. SEGEMEHL, the previous aligner, was replaced by MOSAIK and two versions of VISPA are now available. The first one accepts fasta file as input and runs with MOSAIK and the second version takes bam/sam file as input and runs with any other aligner.

Alongside the improvements on the code, VISPA plugin was developed and installed on Ion Torrent on December 2012; Torrent Suite version 3.2. The plugin uses the internal aligner provided by Ion Torrent, TMAP, to map reads to a genome reference. However, in January 2013, Ion Torrent released version 3.4 of its server. This new Torrent Suite version triggered the need to upgrade the plugin since fasta file were no longer directly available from the experiment. This task was completed on February 10th, 2013 and the plugin

made available for download on Ion Torrent Community website.

$http://mendel.iontorrent.com/ion-docs/visa-Plugin-15007958.html$

PART 5

DISCUSSION AND FUTURE WORK

In ongoing work, we are exploring possibility of integrating multiple samples to perform gene differential analysis. This addition will boost comparative analysis from samples sequenced under similar conditions such as outbreak. Also, we plan to explore a post-processing analysis on each generated bootstrap sample before computing gene abundance to check their biological significance by controlling the false discovery rate. The direct outcome of this new feature will be, depending on the size of the sample, a maximal number of bootstrap sample will be recommended. Further more, we are exploring how to incorporate elements of the A-Seq-2 protocol in IsoEM. The A-Seq-2 analysis is an independently estimated abundances for transcripts expression level estimates that can be used as ground truth as describes in the following paper [60]. This protocol can be used to assess the accuracy of abundance estimation for RNA-Seq data for real data and can be used to further improve IsoEM estimates.

REFERENCES

[1] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with rna-seq," *Nature biotechnology*, vol. 31, no. 1, pp. 46–53, 2012.

[2] J. C. Alwine, D. J. Kemp, and G. R. Stark, "Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5350–5354, 1977.

[3] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[4] A. M. Wang, M. V. Doyle, and D. F. Mark, "Quantitation of mrna by the polymerase chain reaction," *Proceedings of the National Academy of Sciences*, vol. 86, no. 24, pp. 9717–9721, 1989.

[5] M. A. Moran, "Metatranscriptomics: eavesdropping on complex microbial communities," *Issues*, 2010.

[6] A. Fujita, P. Severino, K. Kojima, J. R. Sato, A. G. Patriota, and S. Miyano, "Functional clustering of time series gene expression data by granger causality," *BMC systems biology*, vol. 6, no. 1, p. 137, 2012.

[7] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for rna-seq data," *Genome biology*, vol. 14, no. 9, p. R95, 2013.

[8] J. Li, H. Jiang, and W. Wong, "Method modeling non-uniformity in short-read rates in rna-seq data," *Genome Biol*, vol. 11, no. 5, p. R25, 2010.

[9] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in illumina transcriptome sequencing caused by random hexamer priming," *Nucleic acids research,* vol. 38, no. 12, pp. e131–e131, 2010.

[10] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter, "Improving rna-seq expression estimates by correcting for fragment bias," *Genome biology,* vol. 12, no. 3, p. R22, 2011.

[11] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods,* vol. 5, no. 7, pp. 621–628, 2008.

[12] O. Morozova, M. Hirst, and M. A. Marra, "Applications of new sequencing technologies for transcriptome analysis," *Annual review of genomics and human genetics,* vol. 10, pp. 135–151, 2009.

[13] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics,* vol. 10, no. 1, pp. 57–63, 2009.

[14] J. Bullard, E. Purdom, K. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics,* vol. 11, no. 1, p. 94, 2010.

[15] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, no. 1, pp. 139–140, 2010.

[16] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol,* vol. 11, no. 10, p. R106, 2010.

[17] Y. Bi and R. V. Davuluri, "Npebseq: nonparametric empirical bayesian-based procedure for differential expression analysis of rna-seq data," *BMC bioinformatics,* vol. 14, no. 1, p. 262, 2013.

[18] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman *et al.*, "Ncbi geo: archive for functional genomics data setsâĂŤ10 years on," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D1005–D1010, 2011.

[19] J. Feng, C. A. Meyer, Q. Wang, J. S. Liu, X. S. Liu, and Y. Zhang, "Gfold: a generalized fold change for ranking differentially expressed genes from rna-seq data," *Bioinformatics*, vol. 28, no. 21, pp. 2782–2788, 2012.

[20] D. Scholtens and A. Von Heydebreck, "Analysis of differential gene expression studies," in *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 229–248.

[21] Y. Liu, J. Zhou, and K. P. White, "Rna-seq differential expression studies: more sequence or more replication?" *Bioinformatics*, vol. 30, no. 3, pp. 301–304, 2014.

[22] M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from rna-seq data," *Algorithms for Molecular Biology*, vol. 6:9, 2011. [Online]. Available: http://www.almob.org/content/6/1/9

[23] C. N. D. Bo Li, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," p. 323, 2011.

[24] L. H. Reid, "Proposed methods for testing and selecting the ercc external rna controls," *BMC genomics*, vol. 6, no. 1, pp. 1–18, 2005.

[25] Y. H. Yoav Benjamini, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[26] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, p. R25, 2009.

[27] W.-P. Lee, M. P. Stromberg, A. Ward, C. Stewart, E. P. Garrison, and G. T. Marth, "MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping," *PLoS ONE*, vol. 9, no. 3, pp. e90 581+, 2014.

[28] M. Consortium, "The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[29] T. Stephen. (2012, Jun.) This is a test entry of type @ONLINE. [Online]. Available: www.http://gettinggeneticsdone.blogspot.com/2012/09/deseq-vs-edger-comparison.html

[30] N. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal rna-seq quantification," *arXiv preprint arXiv:1505.02710*, 2015.

[31] R. Patro, S. M. Mount, and C. Kingsford, "Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms," *Nature biotechnology*, vol. 32, no. 5, pp. 462–464, 2014.

[32] R. Patro, G. Duggal, and C. Kingsford, "Salmon: Accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment," *bioRxiv*, p. 021592, 2015.

[33] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster, "Integrative analysis of environmental sequences using MEGAN4," *Genome research*, vol. 21, no. 9, pp. 1552–1560, 2011.

[34] K. M. Konwar, N. W. Hanson, A. P. Pagé, and S. J. Hallam, "Metapathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information," *BMC bioinformatics*, vol. 14, no. 1, p. 202, 2013.

[35] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[36] Y. Ye and T. G. Doak, "A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes," *PLoS computational biology*, vol. 5, no. 8, p. e1000465, 2009.

[37] I. Sharon, S. Bercovici, R. Y. Pinter, and T. Shlomi, "Pathway-based functional analysis of metagenomes," *Journal of Computational Biology*, vol. 18, no. 3, pp. 495–505, 2011.

[38] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[39] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The annals of applied statistics*, pp. 107–129, 2007.

[40] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichita, and S. Drăghici, "Methods and approaches in the topology-based analysis of biological pathways," *Frontiers in physiology*, vol. 4, 2013.

[41] S. Al Seesi, Y. T. Tiagueu, A. Zelikovsky, and I. I. Măndoiu, "Bootstrap-based differential gene expression analysis for rna-seq data with and without replicates," *BMC genomics*, vol. 15, no. Suppl 8, p. S2, 2014.

[42] Q. Cheng and A. Zelikovsky, "Combinatorial optimization algorithms for metabolic networks alignments and their applications," *IJKDB*, vol. 2, no. 1, pp. 1–23, 2011.

[43] A. E. Trindade-Silva, G. E. Lim-Fong, K. H. Sharp, and M. G. Haygood, "Bryostatins: biological context and biotechnological prospects," *Current opinion in biotechnology*, vol. 21, no. 6, pp. 834–842, 2010.

[44] M. G. Haygood and S. K. Davidson, "Small-subunit rrna genes and in situ hybridization with oligonucleotides specific for the bacterial symbionts in the larvae of the bryozoan bugula neritina and proposal of" candidatus endobugula sertula"." *Applied and environmental microbiology*, vol. 63, no. 11, pp. 4612–4616, 1997.

[45] S. Davidson, S. Allen, G. Lim, C. Anderson, and M. Haygood, "Evidence for the biosynthesis of bryostatins by the bacterial symbiont âĂIJcandidatus endobugula sertulaâĂİ of the bryozoanbugula neritina," *Applied and Environmental Microbiology*, vol. 67, no. 10, pp. 4531–4537, 2001.

[46] N. Lopanik, N. Lindquist, and N. Targett, "Potent cytotoxins produced by a microbial symbiont protect host larvae from predation," *Oecologia*, vol. 139, no. 1, pp. 131–139, 2004.

[47] N. Lindquist and M. E. Hay, "Palatability and chemical defense of marine invertebrate larvae," *Ecological Monographs*, pp. 431–450, 1996.

[48] N. B. Lopanik, N. M. Targett, and N. Lindquist, "Ontogeny of a symbiont-produced chemical defense in bugula neritina (bryozoa)," *MARINE ECOLOGY-PROGRESS SERIES-*, vol. 327, p. 183, 2006.

[49] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng *et al.*, "Full-length transcriptome assembly from rna-seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[50] M. Mathew, K. I. Bean, Y. Temate-Tiagueu, A. Caciula, I. I. Mandoiu, A. Zelikovsky, and N. B. Lopanik, "Influence of symbiont-produced bioactive natural products on holobiont fitness in the marine bryozoan, bugula neritina via protein kinase c (pkc)," *Marine Biology*, vol. 163, no. 2, pp. 1–17, 2016.

[51] J. Linneman, D. Paulus, G. Lim-Fong, and N. B. Lopanik, "Latitudinal variation of a defensive symbiosis in the bugula neritina (bryozoa) sibling species complex," 2014.

[52] S. Sudek, N. B. Lopanik, L. E. Waggoner, M. Hildebrand, C. Anderson, H. Liu, A. Patel, D. H. Sherman, and M. G. Haygood, "Identification of the putative bryostatin polyketide synthase gene cluster from âĂIJcandidatus endobugula sertulaâĂİ, the uncultivated microbial symbiont of the marine bryozoan bugula neritina," *Journal of natural products*, vol. 70, no. 1, pp. 67–74, 2007.

[53] A. Untergasser, H. Nijveen, X. Rao, T. Bisseling, R. Geurts, and J. A. Leunissen, "Primer3plus, an enhanced web interface to primer3," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W71–W74, 2007.

[54] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative pcr and the 2- δδct method," *methods*, vol. 25, no. 4, pp. 402–408, 2001.

[55] P. D. Lee, R. Sladek, C. M. Greenwood, and T. J. Hudson, "Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies," *Genome Research*, vol. 12, no. 2, pp. 292–297, 2002.

[56] O. Thellin, W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen, "Housekeeping genes as internal standards: use and limits," *Journal of biotechnology*, vol. 75, no. 2, pp. 291–295, 1999.

[57] M. Magrane, U. Consortium *et al.*, "Uniprot knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, p. bar009, 2011.

[58] M. W. Pfaffl, "A new mathematical model for relative quantification in real-time rt–pcr," *Nucleic acids research*, vol. 29, no. 9, pp. e45–e45, 2001.

[59] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Măndoiu, P. Balfe, and A. Zelikovsky, "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC bioinformatics*, vol. 12, no. 6, p. 1, 2011.

[60] A. Kanitz, F. Gypas, A. J. Gruber, A. R. Gruber, G. Martin, and M. Zavolan, "Comparative assessment of methods for the computational inference of transcript isoform abundance from rna-seq data," *Genome biology*, vol. 16, no. 1, pp. 1–26, 2015.