HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

"ALEXANDER FLEMING"
Biomedical Sciences Research Center

DIPLOMA THESIS:

DEVELOPMENT OF INTEGRATIVE STATISTICAL ALGORITHMS FOR THE ANALYSIS OF GENE EXPRESSION DATA.

DIONYSIOS FANIDIS

ID number 20170227

SUPERVISOR:

Pantelis Hatzis, PhD, Researcher B'
B.S.R.C Alexander Fleming

ATHENS 2019

Supervisor	Pantelis Hatzis, PhD, Researcer B', BSRC Alexander Fleming
Co-Supervisor	Panagiotis Moulos, PhD, BSRC Alexander Fleming
Thesis committee member A'	Despoina Sanoudou, PhD, Associate Professor, 4rth Department of Internal Medicine, Medical School, National and Kapodistrian University of Athens
Thesis committee member B'	Aristotelis Chatziioannou, Researcher B', National Hellenic Research Foundation

To Elena

Table of Contents

Prologue 7
Abstract 8
1. INTRODUCTION 9
1.1 RNA-sequencing 9
1.2 Applications of RNA-seq technology 9
1.3 Handling of systematic variability 9
1.4 More RNA-seq DEA biases 12
1.5 Modeling RNA-seq counts 12
1.6 metaseqR and PANDORA 15
1.7 Aim and contributions 15
2. METHODS 17
2.1 metaseqR2 integrated p-value combination algorithms 17
2.2 RNA-seq datasets 18
2.3 New statistical analysis tools and model organism selection criteria 19
2.4 Simulation data 19
2.5 Performance metrics 19
2.6 Hierarchical clustering 20
2.7 PA input preparation 21
3. RESULTS 22
3.1 PANDORA versus other metaseqR2-implemented statistical analysis tools performance (I) 22
3.2 PANDORA versus other metaseqR2-implemented statistical analysis tools performance (II) 25
3.3 Effects of different normalization methods upon downstream DEA. 27
3.4 Gene length propagation into PA. 29
3.5 DEA of lncRNAs. 30
3.6 metaseqR2 tools concordance analysis using three different biotype designs. .33
4. DISCUSSION 38
5. REFERENCES 41
6. APPENDIX I 44
7. Appendix II 58
8. Appendix III 70

Table of Main Figures

Simulation Data Figures

Figure 1: False Discovery Curves (FDC) using raw p-values after EDASeq normalization. 22
Figure 2: Area under the ROC curve (AUC) using raw p-values after Edaseq normalization. 23
Figure 3: False Discovery Rate (FDR) using adjusted p-values after EDASeq normalization. 24
Figure 4: F_1-score (precision-sensitivity tradeoff) using raw p-values after EDASeq normalization. 25

Real Data Figures

Figure 5: ROC and F1-score analysis of real datasets using raw p-values after EDASeq normalization. 26

Different Normalization Methods Figures

Figure 6:ROC analysis for DESeq and TMM normalized simulation data.....28
Figure 7: F1-score analysis for DESeq and TMM normalized simulation data 28

Gene Length Propagation into Pathway Analysis Figures

Figure 8: Investigation of gene length bias presence within the human-based simulation dataset....29
Figure 9: Kolmogorov-Smirnov analysis for all unique pathway genes and unique leading edge genes. 30

DEA of lncRNAs - Simulation Figures

Figure 10: False Discovery Curves (FDC) using raw p-values after EDASeq normalization. 31
Figure 11: False Discovery Rate (FDR) using adjusted p-values after EDASeq normalization. 32
Figure 12: F1-score (precision-sensitivity tradeoff) using raw p-values after EDASeq normalization. 33

metaseqR2 Tools Concordance Figures

Figure 13: Number of DEGs per tool and biotype, using unadjusted p-values. 34
Figure 14: Biotype representativeness ratio, using unadjusted p-values. 34
Figure 15: Mean overlap proportion of DEGs for each tool and biotype scheme, using unadjusted p-values.....35
Figure 16: Mean DEG ranking correlation between DEA tools for all biotype scheme, using unadjusted p-values. 36
Figure 17: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. 36
Figure 18: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. 37
Figure 19: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. 37

Table of Appendix I Figures

Supplementary Figure 1: metaseqR2 workflow. 44

Simulation Data Figures

Supplementary Figure 2: False Discovery Curves (FDC) usign adjusted p-values after EDASeq normalization. 45
Supplementary Figure 3: Area under the ROC curve (AUC) using adjusted p-values and EDASeq normalization. 46
Supplementary Figure 4: False Discovery Tradeoff (FDT) using raw p-values after EDASeq normalization. 47
Supplementary Figure 5: F_1-score (precision-sensitivity tradeoff) using adjusted p-values after EDASeq normalization 48
Supplementary Figure 6: False Negative Curves (FNC) using raw p-values after EDASeq normalization. 49
Supplementary Figure 7: False Negative Curves (FNC) using adjusted p-values after EDASeq normalization. 50
Supplementary Figure 8: FN, FP and TP hits for all datasets, simulated replicate designs and statistical analysis methods using unadjusted p-values 51

Real Data Figures

Supplementary Figure 9: ROC and F1-score analysis of real datasets using adjusted p-values after EDASeq. 52

DEA of lncRNAs - Simulation Figures

Supplementary Figure 10: False Discovery Curves (FDC) using adjusted p-values after EDASeq normalization. 53
Supplementary Figure 11: False Discovery Curves (FDC) using adjusted p-values of all simulated genes after EDASeq normalization. 53
Supplementary Figure 12: F1-score (precision-sensitivity tradeoff) using adjusted p-values after EDASeq normalization 54

metaseqR2 Tools Concordance Figures

Supplementary Figure 13: Number of DEGs per tool and biotype, using adjusted p-values after EDASeq normalization. 54
Supplementary Figure 14: Biotype representativeness ratio, using adjusted p-values after EDASeq normalization. 55
Supplementary Figure 15: Mean overlap proportion of DEGs for each tool and biotype scheme, using adjusted p-values and EDASeq normalization. 55
Supplementary Figure 16: Mean DEG ranking correlation between DEA tools for all biotype scheme, using adjusted p-values after EDASeq normalization. 56
Supplementary Figure 17: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. 56
Supplementary Figure 18: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. 57
Supplementary Figure 19: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. 57

Prologue

This MSc thesis entitled "Development of integrative statistical algorithms for the analysis of gene expression data" was conducted at Dr. Panagiotis Moulos laboratory, *Biomedical Sciences Research Center Alexander Fleming*, as part of the *International MSc in Molecular Biomedicine*, *Department of Medicine, UOA* and *BSRC Alexander Fleming*.

First of all, I would like to give Dr. P. Moulos my heartfelt thanks for trusting me with this Bioinformatics topic, for his valuable help and advice whenever I needed him throughout this thesis elaboration, as well as for our excellent collaboration. I really feel like I have stepped on the shoulders of a giant!

Moreover, I owe thanks to Dr. Pantelis Hatzis for his supervision and his willing character to assist me with everything that I needed. In particular I must gratefully acknowledge him for recommending me to Dr. Moulos for the first time.

I would also like to thank Prof. Despoina Sanoudou and Dr. Aristotelis Chatziioannou first of all for accepting to participate in the composition of my MSc thesis committee and secondly for their interest in the progress of my dissertation.

In addition, I thank Dr. Moulos laboratory members: Dr. Alexandros Dimopoulos (Elixir-GR post-doctoral fellow), Marielena Georgaki (PhD candidate) and Popi Markopoulou (former lab member), along with Dr. Dimitris Konstantopoulos (Post-Doctoral scientist, Dr. Fousteri lab), Alexandros Galaras (PhD candidate; Dr. Hatzis lab) and Christos Tzaferis (PhD candidate; Prof. Kollias lab) for their support and advice.

Finally, I am much obliged to my family and friends for helping me behind-the-scenes to complete another step in my scientific career.

Abstract

In the past few years, RNA-seq has become the technology of choice for monitoring gene expression at massive scales. Although its benefits outmatch its potential pitfalls, RNA-seq exhibits certain technical and systematic biases like every high-throughput technique. Such biases become more evident in real-life experimental settings such as searching for a signature that differentiates healthy and disease tissues or finding a set of genes whose expression significantly varies across a time-course or a drug dosage study. In an attempt to confront RNA-seq data inherent biases, many different statistical analysis approaches have been proposed, each one with its own advantages and drawbacks. Taking into consideration the limited research dedicated in developing meta-analysis pipelines capable of ameliorating the results yielded by individual methods, we hereby present metaseqR2, an upgraded version of the previously released metaseqR Bioconductor package. Including some of the best performing and most popular differential expression analysis statistical tools, as well as a new supported organism, metaseqR2 is an all-in-one, powerful tool for RNA-seq data analysis. Moreover, we demonstrate that PANDORA, the main p-value combination method behind the metaseqR2 package, not only continues to greatly perform under metaseqR2 statistical environment, but it is also characterized by a very robust behavior under different analysis pipelines. Finally, in the presence of RNA-seq biases such as the gene length bias and the recently discovered bias in the detection of differentially expressed lncRNAs, PANDORA is probably the most reliable solution to work with.

1. INTRODUCTION

1.1 RNA-sequencing

Historical facts, wet lab protocol and following data analysis outline.

RNA-sequencing (RNA-seq) is a high-throughput method for sequencing RNA populations. It was first applied on 2008 by three separate teams (Lister et al. 2008; Nagalakshmi et al. 2008; Mortazavi et al. 2008) and it gradually replaced microarrays in common laboratory practice due to demonstrating among many other advantages higher resolution, lower noise and greater dynamic range (Z. Wang, Gerstein, and Snyder 2009; Wu, Wang, and Wu 2013).

An RNA-seq protocol has two major steps: that of the wet and that of the dry lab. Beginning from the wet lab, an RNA species population is isolated and reverse transcribed into cDNA. If the RNA molecules are not prior to reverse transcription fragmented, then the cDNA molecules follow a fragmentation procedure and vice versa. Subsequently, sequencing adapters are added to the cDNA fragments and library amplification is performed or not according to sequencing technology used and/or experimenter's needs (Z. Wang, Gerstein, and Snyder 2009; Quail et al. 2012). Finally, short read sequencing takes place to obtain either single- or paired-end reads (Z. Wang, Gerstein, and Snyder 2009).

These reads are the input of the subsequent computational analysis pipeline. After a first step of quality control, reads are mapped to an available reference genome, transcriptome or exome (Han et al. 2015). In the absence of a reference genome reads are *de novo* assembled so as to infer about the identity of the expressed transcripts (Conesa et al. 2016). Finally, reads overlapping any genomic feature of interest are bioinformatically quantified and a read counts table is delivered. The number of counted reads for each genomic feature can then be used to estimate the feature's true expression levels (Han et al. 2015). While, the examined genomic feature can be a gene, transcript or exon, the same principles discussed apply for all of them. Thus, for the rest of this dissertation thesis we will generally refer to them as the "gene" for simplicity.

1.2 Applications of RNA-seq technology

Brief reference to existing applications.

Since its appearance, RNA-seq technology has been used in a great number of applications. For example, the detection of alternative splicing, limited by former applied technologies, was achieved for the first time thanks to RNA-seq (Mortazavi et al. 2008). Additionally, it has been used among other techniques to identify tissue-specific alternative transcripts (E. T. Wang et al. 2008), to discover *de novo* transcripts (Roberts et al. 2011), to detect and quantify allele-specific expression (Tuch et al. 2010). In addition, it has been applied to find *de novo* gene fusion events in cancer cells (Edgren et al. 2011) and to report statistically significant changes in gene expression between two or more conditions (control vs treatment, health vs disease) and/or time-points. This latter application is called differential expression analysis (DEA) and will be the current topic of interest in the present dissertation.

1.3 Handling of systematic variability

Presentation of RNA-seq systematic biases and normalization methods

Despite the initial optimistic expectation that RNA-seq data would require minimal normalization (Z. Wang, Gerstein, and Snyder 2009), there are a lot of confounding factors/biases introducing systematic variability to the data that need to be taken into consideration before differential expression (DE) statistical analysis is applied.

To begin with, intra-sample sources of systematic variability include the gene length and the GC-content bias. Gene length bias became obvious from the early steps of RNA-seq technology (Mortazavi et al. 2008). It describes the fact that longer genes have higher probabilities of being assigned more sequencing reads than the shorter ones (Oshlack and Wakefield 2009). The more reads a gene has accumulated, the higher its statistical power becomes and thus, the more probable it is to be called differentially expressed even if it is not. Interestingly, it has been proposed that gene length bias is more intense when the actual gene expression is low (Oshlack and Wakefield 2009) a finding partially replicated recently by (Yoon and Nam 2017). As far as GC-content is concerned, it fluctuates not only between genes of a sample, as it is usually assumed, but between samples as well (Pickrell et al. 2010) affecting DE analysis if left untreated (Hansen, Irizarry, and Wu 2012). Finally, a group of normalization methods also alleviate less systematic biases concerning transcript library composition. Genes of disproportionally large expression tend to monopolize read counts in expense of less active genes, exactly as gene length bias does, leading to more unreliable downstream DE results (read count bias) (Anders and Huber 2010; Robinson and Oshlack 2010; Bullard et al. 2010).

Between-sample biases should also be taken into consideration. For instance, samples sequenced deeper than others (higher library size) are characterized by a larger count, making library size normalization necessary for proper comparison of counts between the examined groups (Dillies et al. 2012). Additionally, condition/tissue specific transcripts have the tendency to obscure real DE patterns between samples and thus several analysis platforms focus mainly on housekeeping genes (Anders and Huber 2010).

Many normalization methods have been developed for the alleviation of such sources of variability and each one of them takes in account different working assumptions. DEA tools may allow the use of more than one normalization approaches either as they were originally proposed in literature or after tool specific modifications, mainly for consistency reasons. In the current section we will focus on the normalization methods implemented by RNA-seq analysis tools used for the purposes of this dissertation.

To begin with, the DESeq package has its own method of normalization driven by the assumption that in biological samples, the vast majority of genes are not differentially expressed (Anders and Huber 2010). It accounts for sequencing depth as well as for varying library composition (tissue/condition specific genes) using sample-specific size factors computed with shared inter-sample information: the median of the ratio of each sample's read counts to the geometric mean of read counts across samples. By using the median, DESeq normalization is less sensitive to library composition effects able to skew expression patterns and with the use of the log it removes condition-specific genes to focus mainly on housekeeping ones. Finally, the geometric mean applied makes the method unbiased to outliers.

DESeq2 software (Love, Huber, and Anders 2014) implements the original DESeq normalization method as well as two modifications of it, suitable for dealing with zero expression values. In addition, it supports the integration of user-provided, gene-wise, size factors matrix, so as to compensate for more specific intra-sample biases like GC-content bias.

Just like DESeq and DESeq2, edgeR's TMM method proposed by Robinson and Oshlack (2010) normalizes sample read counts for sequencing depth and library composition, while it lefts untreated other types of biases, like GC-content and gene length bias, because they are assumed to be of relatively equal effect across samples (Robinson and Oshlack 2010). TMM values (Trimmed Mean of M values) are used to compute sample specific scaling factors, which are then multiplied with their respective library size to give an effective library size that will be used in downstream analysis. As with DESeq normalization, it

assumes that most genes in a sample are not differentially expressed and moreover, that the differentially expressed ones are uniformly separated between the upregulated and downregulated groups. Notably, edgeR accepts user-provided gene-wise normalization factors to adjust for intra-sample sources of variance, while DESeq and the upper quartile (UQ) normalization method proposed by Bullard et al. (2010) are also included as normalization alternatives (Anders and Huber 2010; Bullard et al. 2010). Importantly, UQ method normalizes read count data for sequencing depth and expression outliers, by matching the upper quartile of lane-specific, read count distributions (Bullard et al. 2010).

Unlike TMM method (Robinson and Oshlack 2010), NBPSeq assumes that relative frequency of gene counts in respect of total library size (sum of library counts) alone is an effective approximation of gene expression (Di et al. 2011). Due to the fact that relative frequencies of each sample must sum up to one, they are computed based on "effective" library sizes: the real library size multiplied by a sample-specific normalization factor, either computed as in DESeq package or provided by the user. Before effective library sizes are computed, counts are downsampled in random to achieve equal active library sizes, a necessary assumption for subsequent NBPSeq statistical testing (Di et al. 2011).

NOISeq package (Tarazona et al. 2015) provides three different normalization procedures: RPKM (Reads Per Kilobase of exon model per Million mapped reads), UQ and TMM of edgeR package. RPKM was introduced by one of the first papers that ever used RNA-seq technology for alleviating gene length bias and differential sequencing depth between experimental conditions (Mortazavi et al. 2008). However, it is noteworthy that scaling raw read counts by gene length alone was found insufficient as a method to reduce gene length bias (Oshlack and Wakefield 2009; Bullard et al. 2010).

Initially launched in 2004 as a Bioconductor package for the analysis of microarray and PCR data, limma was expanded in 2015 to support statistical analysis of RNA-seq data via *voom()* function (Gentleman et al. 2004; Ritchie et al. 2015). However, no RNA-seq specific normalization method was developed. Thus, the user is advised to use either a quartile normalization method of microarray logic or normalization factors given by TMM method of the edgeR tool before statistical analysis takes place (Ritchie et al. 2015). For the record, quartile normalization has been shown to introduce undesirable intra-sample variability (Dillies et al. 2012) and thus must better be avoided.

Three of the remaining packages used, provide RNA-seq normalization facilities already described. Thus, baySeq (Hardcastle and Kelly 2009) uses sample-specific scaling factors given by the methods of (Bullard et al. 2010) and (Robinson and Oshlack 2010); ABSSeq (Yang, Rosenstiel, and Schulenburg 2016) includes the total read counts, DESeq, TMM, upper quartile and qtotal normalization procedures and finally, DSS (Wu, Wang, and Wu 2013) includes the median of log ratio of counts (a TMM method modification), UQ, total counts and the median counts method.

To conclude with, while the above tools use correction methods for the biases of library composition, library size and gene length, the GC-content bias present in both intra- and inter-sample level is treated only by the last tool used; EDASeq (Risso et al. 2011). More specifically, EDASeq performs two rounds of normalization: a within-lane-specific to treat intra-sample biases and then a between-lane specific to remove inter-lane biases. More specifically, GC-content normalization can be performed using any of three proposed normalization methods. *Loess*, the default method, regresses log-scale gene counts to GC-content using the loess robust local regression method, *global scaling normalization* bins genes according to their sequence composition and then forces median or upper-quartile to be equal across bins and *full-quartile* normalization after binning genes according to sequence composition, it then matches bin distributions using a microarrays similar full quartile normalization. For between-lane normalization

EDASeq uses either full quartile normalization (the default), median or the upper quartile approach of (Bullard et al. 2010).

1.4 More RNA-seq DEA biases

Gene length bias & pathway analysis; lncRNA bias.

Unfortunately, gene length bias described at the beginning of section 1.3 does not affect only DEA, but has been reported in many independent publications such as (Oshlack and Wakefield 2009; Young et al. 2010) to propagate into subsequent Pathway Analysis (referred hereafter as PA), too.

In short, PA (aka functional enrichment analysis of pathways) is a computational procedure used to make sense out of high-throughput sequencing technology results. For example, a list of thousands of differentially expressed genes (DEG) cannot be interpreted if not previously organized in a human friendly format using PA. Note that the term "pathway" as used in PA does not have the strict, classical meaning of a molecular pathway, but rather the more loose definition of a functional biological entity consisted of several pathway components such as genes (García-Campos, Espinal-Enríquez, and Hernández-Lemus 2015). Components' interaction define pathways' functionality and all pathways can be organized in a super-complex network. Although different kind of pathways and PA methods exist, they all have the same main objective: given the high-throughput sequencing data and a list of pathways, to explore whether there are statistically significant functionality patterns in the high-throughput data examined (García-Campos, Espinal-Enríquez, and Hernández-Lemus 2015).

As far as the gene-length bias–PA connection is concerned, pathways that contain many genes longer than average tend to be called enriched more often that the rest, when gene length bias is present in the RNA-seq dataset (Oshlack and Wakefield 2009). Interestingly, the bias is present even when using different PA tools and/or pathway databases (Gao et al. 2011; Oshlack and Wakefield 2009) introducing the need to take it seriously into consideration.

Finally, another RNA-seq bias affecting DEA is the lncRNA bias. In particular, it was very recently shown that lncRNAs and low expression genes in general are under-represented by RNA-seq DEA algorithms due to low read counts, high noise levels and condition specific expression (Assefa et al. 2018). This is why the majority of these tools propose the filtering of such genomic features before downstream analysis is performed, a step that inhibits the analysis of almost 70% of a cells transcriptome.

1.5 Modeling RNA-seq counts

Statistical approaches to describe RNA-seq counts data.

Due to the fact that any RNA-seq experiment contains a finite number of replicates, the read counts of each condition must be modeled, parametrically or not, in order to calculate the statistics needed for downstream DE hypothesis testing.

One of the first distributions used for this purpose was the Poisson distribution (Bloom et al. 2009), as its core assumption that counts' mean (μ) equals their variance (σ^2) ($\mu = \sigma^2$) was witnessed to hold true for RNA-seq technical replicates (Marioni et al. 2008). However, it was already proven for SAGE (Serial Analysis of Gene Expression) and later on for RNA-seq data as well, that a Poisson distribution accounts only for sequencing noise and not for biological or technical sources of variability (Baggerly et al. 2004; Lu, Tomfohr, and Kepler 2005). Consequently, data modeled by such a distribution are overdispersed ($\mu < \sigma^2$) and analysis' results are not reliable.

A very popular alternative to the Poisson modeling is the Negative Binomial (NB) distribution. Being described by two parameters, NB(μ,σ^2), it enables for read count modeling without constraining their variance range (Anders and Huber 2010) and has been shown to apply reliably even for data that are not actually NB distributed (Lu, Tomfohr, and Kepler 2005). On the other hand, it introduces an uncertainty of both μ and σ^2 estimation for small biological replicate numbers (Di et al. 2011). To account for it or for any other RNA-seq data statistical analysis bias, like the big log fold change (LFC) variability of low count genes (Yang, Rosenstiel, and Schulenburg 2016), NB distribution's parameters are estimated by each tool under different assumptions and working hypotheses.

In addition to the NB, normal and non-parametric distributions are also used. First of all, continuous data statistics are exploited in order to alleviate several drawbacks of discrete statistics describing RNA-seq data, such as the less tractable count distributions statistics theory, the difficulty of NB distributions to adapt to data with different degrees of heterogeneity and the limited number of available statistical analysis tools (Law et al. 2014). Moreover, non-parametric approaches are mainly applied to cover cases where assumptions made about NB distribution parameters does not hold true, as well as in order to better handle transcriptomic features with very low count reads (Tarazona et al. 2011).

This section is dedicated to a brief presentation of the RNA-seq data DEA approaches used for the purposes of the current dissertation. For simplicity reasons, when referring to read counts, normalized read counts will be implied, except otherwise stated. A more detailed description of statistics employed by each tool can be found to the respective publications.

To begin with, the DESeq method assume that genes with the same read counts have the same variance, by letting raw gene-wise variance to be a smooth function of the condition's mean gene counts (Anders and Huber 2010). While this global dispersion trend allows sharing information across genes to increase statistical analysis power given small sample sizes (Anders and Huber 2010), it does not take into account gene-wise expression variability (McCarthy, Chen, and Smyth 2012). At last, both μ and φ (dispersion) are dependent to a sample-wise size factor s calculated as explained in the previous section, to normalize for systematic biases (Anders and Huber 2010).

Being DESeq's successor, DESeq2 extends its predecessor's main modeling ideas (Love, Huber, and Anders 2014). At first, a Generalized Linear Model (GLM) following NB distribution is fitted to each gene offering the ability to analyze complex experimental designs. Furthermore, gene-wise φ is shrunk towards the values reported by μ-φ linear regression fitted as proposed by DESeq (Anders and Huber 2010). To accommodate sample size and distance from the reported trend, an empirical Bayesian model is used to regulate φ shrinkage degree. An empirical Bayes method is also used to shrink log-Fold-Change (logFC) values and reduce FPs for genes of inadequate information.

edgeR, initially a Bioconductor package used for DEA of SAGE data, is one of the first tools published to use the NB distribution for statistical description of RNA-seq read counts (Robinson, McCarthy, and Smyth 2009). edgeR computes NB's dispersion using a maximum likelihood method and then, in order to tackle with small replicate sizes it uses the empirical Bayes method of (Robinson and Smyth 2007) to shrink gene-wise dispersion values towards a common φ value (Robinson, McCarthy, and Smyth 2009). Alternatively, it enables the shrinkage of φ towards a common "trend" allowing gene-specific dispersion variation. In 2012 edgeR was expanded with the addition of generalized linear models (GLM) so that it can also address complex experimental designs (McCarthy, Chen, and Smyth 2012).

Hardcastle et al. proposed in 2009 the baySeq method (Hardcastle and Kelly 2009). As its name implies, baySeq estimates NB models' prior and posterior probabilities using an empirical Bayes method, while NB parameters are estimated by a quasi-likelihood approach that takes into account gene-specific variability. At the time of its publication, baySeq was the only high-performing tool that could handle DEA between more than two experimental conditions. As an exchange for its good performance, baySeq algorithm is more computationally intensive when compared with other tools (Hardcastle and Kelly 2009; Zhang et al. 2017).

A more general, over-parameterized version of the commonly used NB distribution is implemented in the NBPseq package (Di et al. 2011). NBP introduces an additional parameter a to the classical NB model, allowing for an increased flexibility in the description of dispersion's dependency from the mean. Statistical testing for DE is performed by a modification of Robinson's and Smyth's exact test (Robinson and Smyth 2007, 2008).

Similarly to DESeq2, the DSS package team developed their own Bayesian model to shrink NB dispersion parameter and thus estimate more precisely large variations in data heteroskedastisity (aka gene-specific dispersion) (Wu, Wang, and Wu 2013). Prior φ probabilities are derived from a log-normal distribution while posteriors are obtained from maximizing an approximate of the original conditional posterior distribution. At last, as a hypothesis testing a Wald test is used.

The last of the NB based tools used in this project applies a rather different approach in terms of measuring and testing for DE. While other tools like DESeq, DESeq2 and edgeR use the difference of the mean counts between two conditions when testing for DE, ABSSeq uses the absolute difference in read counts instead (Yang, Rosenstiel, and Schulenburg 2016). Thus, it is this absolute difference that is modeled as NB. Furthermore, To take account of small replicate number biases, ABSSeq adds pseudocounts to real data based on gene-specific dispersion, a process in which mean and variance relationship is established as proposed in DESeq package (Anders and Huber 2010). Notably, the most important feature of ABSSeq is its ability to reduce FPs for low count genes by shrinking logFC towards the mean, a method conditioned by both expression levels and gene-specific dispersion. The difference between DESeq2's and ABSSeq's logFC shrinkage is that the latter uses p-values to do so, a strategy that does not affect the number of significant DE genes reported (Yang, Rosenstiel, and Schulenburg 2016).

NOISeq and limma-voom are an exception to the rule making use of non-parametric and normal distribution statistics, respectively to model RNA-seq read counts. More specifically, after computing the log ratio (M) and the read counts absolute difference (D) for each gene between the two conditions of interest, NOISeq compares these two statistics with M' and D' noise distributions to infer for DE (Tarazona et al. 2011). Noise threshold for both metrics is empirically computed by intra-conditionally contrasting read counts of either real (NOIseq-real) or simulated data (NOIseq-sim). Finally, because statistics of continuous distributions is better established than that of discrete data, voom algorithm processes read counts so as to be compatible with the limma microarray analysis method (Law et al. 2014). At first, read counts are used to estimate a condition-wide smooth μ-σ^2 trend for each gene and then the fitted curve is used to compute sample-specific gene variance values, which are then embodied into an inverse weight for each count value. At last, weights are passed along with read counts into limma's linear modeling procedure.

1.6 metaseqR and PANDORA

Basic concept and brief description.

While there is such a wealth of RNA-seq DEA tools, to our knowledge there is not much active research on meta-analysis algorithms that could combine individual tools' assets. However, given methods' differential performance under various experimental circumstances (Franck et al. 2013; Soneson and Delorenzi 2013), the development of method-combination tools becomes almost imperative.

In an attempt to explore such an approach, Moulos and Hatzis created PANDORA (PerformANce Driven scOring of RNA-seq stAtistics), a weighted p-value combination algorithm implemented in the metaseqR Bioconductor package (Moulos and Hatzis 2015). metaseqR is an easy to use, powerful RNA-seq DEA tool, that combines multiple normalization methods with six DE statistical analysis tools, six meta-analysis algorithms and a comprehensive, self-explanatory report. Normalization and subsequent statistical analysis methods can be chosen at will by the user. In addition, further facilities are offered like gene-level, exon-level or 3' UTR-level read quantification (used for analysis of data generated by Lexogen QuantSeq 3' mRNA-Seq), creation of simulated data based on given real ones and easy access to any of the five supported (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Drosophila melanogaster and Arabidopsis thaliana*) model organisms.

Interestingly, during statistical analysis performance examination of metaseqR methods using both simulated and real datasets, PANDORA was shown to behave if not better at least equally well with the top performing DEA algorithms for all metrics. In particular, PANDORA had the best precision-sensitivity trade-off among all algorithms as measured by the F_1-score (Moulos and Hatzis 2015). These data is proof of concept that indeed meta-analysis tools for DEA of RNA-seq data must be more actively developed.

1.7 Aim and contributions

Dissertation's objectives.

Embracing the idea of meta-analysis power to boost RNA-seq DEA and wanting to further compare the behavior of PANDORA with that of other tools under DEA biases potentially "lurking" within a real dataset, we set out to:

- Upgrade metaseqR into metaseqR2 by incorporating more statistical analysis tools and supported model organisms.
- Examine metaseqR2 implemented statistical analysis methods' performance under:
    - real data-based simulated datasets of various configurations
    - real datasets coupled with experimental validation such as qPCR and spike-ins
    - different normalization procedures applied
    - gene length and lncRNA biases presence.

As a result, we deliver metaseqR2, an up-to-date RNA-seq DEA package. Among others, it includes three new DEA tools (DESeq2, ABSSeq and DSS) and PANDORA weights for six different model species. Furthermore, we provide the results of an extended metaseqR2 tools evaluation based on various computational experiments. In brief, based on both simulation and real dataset analysis results, PANDORA presents in most cases the best precision-sensitivity tradeoff and is also among others characterized by the capability of simultaneously controlling both FPs and FNs at an adequate level. Additionally, PANDORA behaves robustly when coupled with different normalization methods.

Furthermore, as far as gene length bias is concerned, PANDORA is the tool that better controls its propagation into subsequent pre-ranked PA. At the same time, while all metaseqR2 tools are biased during DEA of lncRNAs, PANDORA is if not the best, one of the best performing statistical methods in the vast majority of metrics calculated. At last but not least, data supporting an already existent hypothesis about the exact nature of gene length bias are provided.

Conclusively, we propose the p-value combination method PANDORA as one of the best methods for RNA-seq DEA, combining a very good performance with robust and reliable behavior under different experimental conditions and RNA-seq technology inherent biases.

2. METHODS

2.1 metaseqR2 integrated p-value combination algorithms

Brief description of the p-value combination algorithms used.

metaseqR2 incorporates six different p-value combination algorithms. Given p-values from more than one DEA tools for any gene *i,* a combined p-value (p_i^*) is returned. In this section we will shortly describe each p-value combination approach. Let p_{ij} be the p-value for any gene *i* after application of the statistical test *j*. For clarity reasons the term "p-value" will be defined more loosely to include the baySeq (*1 – posterior probability of differential expression*) and NOISeq (*1 – q statistic*) statistics. An in more depth portrayal of the methods can be found in metaseqR publication (Moulos and Hatzis 2015).

Simes algorithm

If p_{i1}, p_{i2}, …, p_{im} are the p-values reported by m statistical analysis tools for a given gene *i* and $p_{i(1)}$, $p_{i(2)}$, …, $p_{i(m)}$ the same p-values ranked in an increasing order, then according to a modified Simes method the probability:

$$p_i^* = \min_k \left\{ p_{i(k)} / k \right\},\ k \in (1,\ldots,m)$$

can be used either as an exact or an approximate combined p-value for all m tools' (Simes 1986).

Union algorithm

Combined p-value is given by:

$$p_i^* = \min_j \left\{ p_{ij} \right\},\ j \in (1,\ldots,m)$$

For any examined gene *i* and a given p-value threshold (α), all significant p-values reported by *m* statistical tools are taken into consideration. From them, the minimum p-value is returned as the combined p-value in order to increase TPs in exchange for an also increased number of type I errors (FPs).

Intersection algorithm

Combined p-value is given by:

$$p_i^* = \max_j \left\{ p_{ij} \right\},\ j \in (1,\ldots m)$$

For any examined gene *i* and a given p-value threshold (α) the intersection of the statistically significant DE genes reported by m DEA tools is taken into consideration. From them, the maximum p-value is returned as the combined p-value. As a result DEG list contains less type I errors at the cost of less TPs.

PANDORA algorithm

PANDORA's combined p-value is calculated by:

$$p_i^* = \prod_{j=1}^{m} p_{ij}^{w_j},\ \text{with} \sum_{j=1}^{m} w_j = 1$$

where w_j represents the weight attributed to *j* statistical algorithm performed. Weights can be either automatically estimated using the area under the false discovery curve (AUFC) for the results of test *j*

applied on a real dataset (see next formula) or can be user specified. In any case, weights must return a unit's sum.

$$w_j = \frac{\sum_{j=1}^{m} AUFC_j \Big/ AUFC_j}{\sum_{j=1}^{m}\left(\sum_{j=1}^{m} AUFC_j \Big/ AUFC_j\right)}$$

Fisher's algorithm

Fisher's method uses the *f* statistic (described from the following formula) to perform p-value combination:

$$f_i = -2\sum_{j=1}^{m} \ln p_{ij}$$

f has been proven (Rödel 2007) to follow a chi-square distribution with *2m* degrees of freedom and this distribution is used to infer combination p-values. Notably, the initial method was developed to combine p-values reported by a single tool after the analysis of multiple different datasets. However, the same statistical concept applies as well in case of combining p-values returned by multiple DEA packages after analysis of the same dataset.

Whitlock's algorithm

Combined p-values are derived from the weighted Z statistic's (Whitlock 2005) (following formula) normal distribution.

$$Z_j^w = \sum_{j=1}^{m} w_j Z_j \Big/ \sqrt{\sum_{j=1}^{m} w_j^2}$$

As with Fisher's, Whitlock's approach is also not designed for p-value combination across tools analyzing the same dataset, but will be assumed as thus.

2.2 RNA-seq datasets

Brief description of the datasets used.

For the current MSc thesis nine different datasets were used. Seven out of them, the human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*), fruit-fly (*Drosophila melanogaster*), arabidopsis (*Arabidopsis thaliana),* TaqMan and SEQC datasets are the ones also used by (Moulos and Hatzis 2015).

Rat (*Rattus* norvegicus) dataset was generated by (Heyne et al. 2014) and is publicly available through ArrayExpress, accession number ERP006055. The brains of 75 most tame and 75 most aggressive mice against humans were selected for RNA-seq. The original dataset was subsetted and 15 mice from each condition were kept. Each animal's RNA-seq data were downloaded as several fastq files that were then merged prior to two-round alignment against the rn6 (Rnor_6.0) rat genome build, using Hisat2 (Kim, Langmead, and Salzberg 2015) and Bowtie2 (Langmead and Salzberg 2012) aligners via custom scripts. RNA-seq data were quantified using metaseqR counting feature.

NGP-nutlin dataset was created by (Assefa et al. 2018) and can be found in GEO, GSE104756. It consists of total RNA-seq data from five control and equal number of nutlin-3 treated human NGP neuroblastoma

cell cultures. SRA files were downloaded and subsequently transformed into fastq files using the SRA toolkit (http://ncbi.github.io/sra-tools/). Two-round alignment against the hg38 (GRCh38) human genome build and subsequent quantification were carried out as for the rat dataset.

2.3 New statistical analysis tools and model organism selection criteria

Criteria applied for metaseqR2's new tools and model organism literature selection.

Literature was reviewed for RNA-seq DEA statistical packages not included in metaseqR. 19 tools were found and were further filtered according to the following criteria:

- proper maintenance (deposited into Bioconductor, version updates)
- metaseqR compatible (accept read counts table as input and be compatible with different normalization algorithm's output)
- following the basic metaseqR concept (DEA between given biological conditions)
- tested under various experimental conditions
- at least as good performance as the metaseqR implemented individual DEA tools
- popularity (number of Bioconductor downloads)

Only three of them, DESeq2 (Love, Huber, and Anders 2014), ABSSeq (Yang, Rosenstiel, and Schulenburg 2016) and DSS (Wu, Wang, and Wu 2013) full-filled almost all standards and was thus implemented in metaseqR2.

As far as the new model organisms are concerned, *Rattus norvegicus* was selected due to its extensive laboratory use. For the selection of a proper dataset, replicate size was the main selection criterion, while pairwise comparison of RNA-seq data was also required, as metaseqR2 is currently limited in examining pairwise comparisons.

2.4 Simulation data

Description of all simulation data configurations.

All simulations used have similar configurations as in (Moulos and Hatzis 2015). More specifically, two simulation designs of 10k genes and two conditions each were created using metaseqR2's *make.sim.data.sd* function, which uses real data to estimate NB parameters. In the first design, named *3x replicate design* (aka *3 replicates – balanced DEG*), three biological replicates were assigned per condition and 1k genes were set to be differentially expressed (half upregulated and half downregulated). In the second one, named *7x replicate design* (aka *7 replicates – unbalanced DEG*), seven biological replicates were developed per condition and 1.2k of the total genes were defined as differentially expressed (700 up-regulated and 500 down-regulated).

For the gene-length bias propagation into PA simulation (Section 3.3), gene length bias was also introduced into *3x* and *7x replicate designs* configurations.

Ten iterations were run for each of the above mentioned simulation designs and datasets.

2.5 Performance metrics

Metrics used for simulation/real dataset evaluation and lncRNA concordance analysis.

For the analysis of all simulated/real data, the metrics used by (Moulos and Hatzis 2015) were applied over 10 simulations average output:

- False discovery (FDC) and false negative curves (FNC) were used to record FPs and FNs emergence, while traversing gene lists ranked in respect to their statistical significance from top to bottom and bottom to top, respectively.
- ROC (Receiver Operating Characteristics) curves (and their respective area under the curve; AUC) were created to visualize each methods ability to rank DEG ahead of non-DEGs.
- F_1-score (and the area under the F_1-score), as well as the *ad hoc* False Discovery Trade-off (FDT) metric were applied to measure each method's precision-sensitivity tradeoff.

Finally, 'same-versus-same' mock comparisons were used as a negative control for all individual and p-value combination tools exactly as in (Moulos and Hatzis 2015).

For the purposes of metaseqR2's tools concordance analysis based on the NGP-nutlin dataset (see Section 2.4), the following metrics were used:

- Number of DEG reported was used to examine each algorithm's strictness or looseness with regard to the others applied. Metric used as in (Assefa et al. 2018).
- Mean DEG overlap was used to examine the agreement between tools' results. For its calculation, the tool-wise proportion of common DEG was computed and then averaged across all comparisons for any given tool. Metric used as in (Assefa et al. 2018).
- Mean DEG ranking correlation was used to infer about inter-tool DEG list consistency. For its calculation, DEGs were ranked according to their p-value and ranked lists obtained were correlated in a tool-wise manner using Spearman's rank correlation statistic. The last was finally averaged across all comparisons for any given tool. Metric inspired by (Assefa et al. 2018).
- Biotype representativeness is an *ad hoc* metric developed to investigate potential biotype over- or under-representation in the final DEG lists. If *b* is a given biotype and G a subset of the examined genes then biotype representativeness is computed as:

$$log\left[\frac{\sum DEG_b / \sum DEG_{all}}{\sum G_b / G_{all}}\right]$$

  A positive representativeness ratio denotes biotype's over-representation, while a negative value denotes under-representation. The greater the deviation from zero, the bigger the existing bias is. A zero value can be returned in two occasions: when a biotype is perfectly represented or when all biotypes (namely all examined genes) are used (control).

All metrics of both simulation and concordance analysis evaluation were also calculated after p-value adjustment under a Benjamini–Holchberg (BH) threshold of 0.05 (Benjamini and Hochberg 1995). Adjustment was not performed for baySeq and NOISeq results, because they do not report a classical p-value score (Hardcastle and Kelly 2009; Sonia Tarazona and Fernando Garc a-Alcalde 2011).

2.6 Hierarchical clustering

Concordance analysis summarization.

Hierarchical clustering was used to summarize and visualize in a human-readable fashion metaseqR2's tools concordance analysis results. For each of the respective four metrics applied (see Section 2.5) mean values previously calculated were now used to compute tool- and metric- specific z-scores. Subsequent hierarchical clustering using Ward's criterion was based on the aforementioned z-scores. Inter-cluster differences were measured by the squared Euclidean distances of their metric-specific z-score means.

2.7 PA input preparation

Tool used, pathway list preparation, pre-ranked GSEA pipeline.

PA analysis, and more specifically pre-ranked GSEA (Gene Set Enrichment Analysis), was performed using the fgsea (fast gsea) Bioconductor tool, release 3.9 (Sergushichev 2016). fgsea was selected due to consistency with the other tools used (R-implemented), speed of analysis and because it provides the core functions for other GSEA R packages like clusterProfiler (Yu et al. 2012).

For the pre-ranked GSEA analysis, a decreasing list of p-values transformed by their negative common algorithm was fed to fgsea along with a custom developed (for consistency with the dataset's genome build) GO pathway list. The aforementioned pipeline was applied for all reported DEG lists.

3. RESULTS

3.1 PANDORA versus other metaseqR2-implemented statistical analysis tools performance (I)

Simulated data

We evaluated the capability of metaseqR2-implemented statistical analysis methods to correctly report DEGs using simulated data according to the *3x* and *7x replicate designs* (see Section 2.4) for the human,

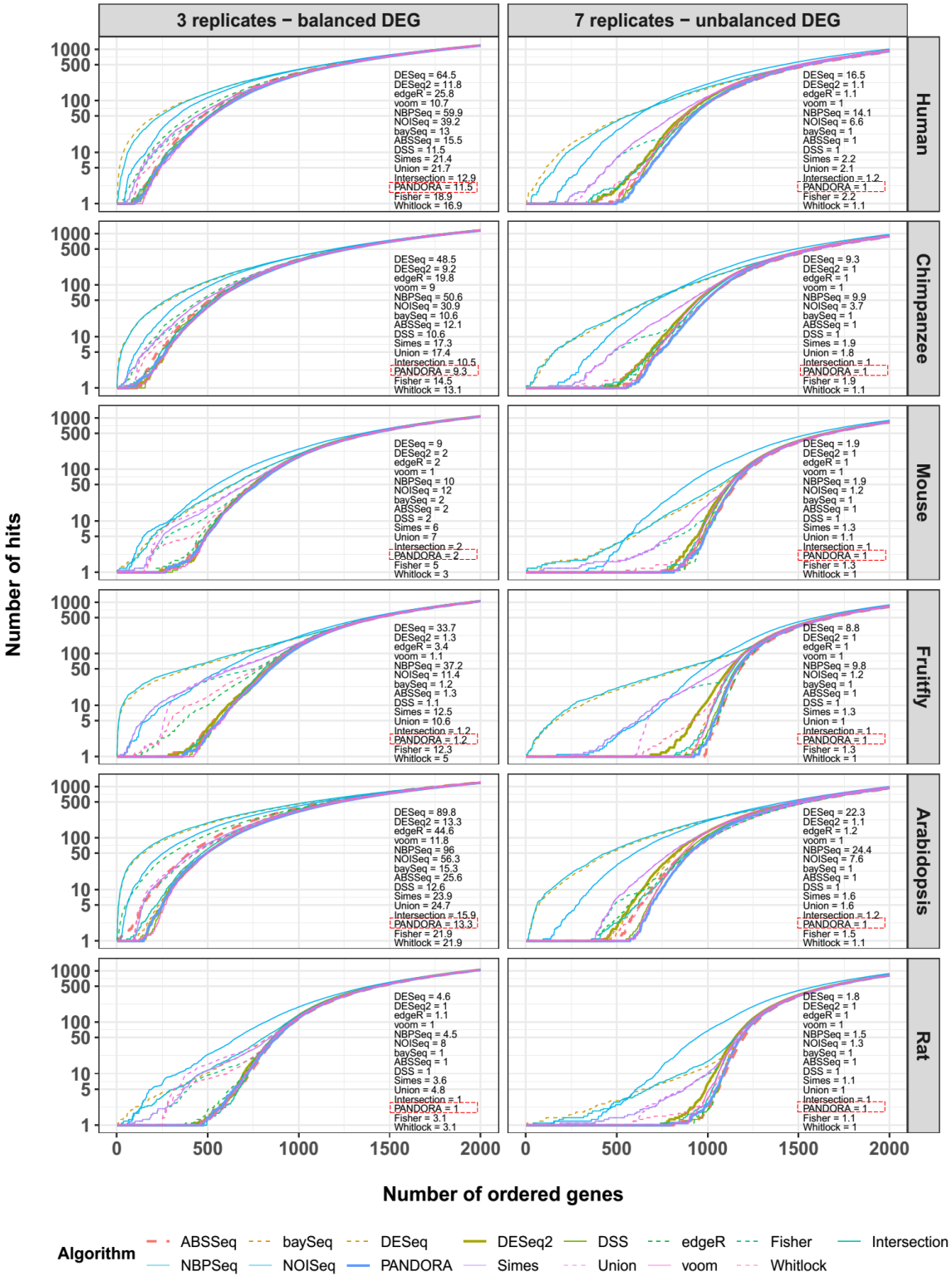


Figure 1: False Discovery Curves (FDC) using raw p-values after EDASeq normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. PANDORA is consistently among the best performing algorithms. (significant p-value threshold: 0.05, EDASeq normalization)

chimpanzee, mouse, arabidopsis, fruit-fly and rat datasets. EDASeq (default) was used as a common normalization baseline for all tools, but another round of DEA after tool specific normalizations was also realized (see Appendix II). Performance evaluation metrics (see Section 2.5) were calculated over 10 iterations for each simulation design–dataset combination.

The ability of tools to rank true DEGs ahead of FPs is visualized using False Discovery Curves (FDCs) and quantified by computing the area under them (AUFC) (*Figure 1*). As far as the *3x replicate design* is concerned, from the nine metaseqR2-implemented, individual statistical analysis tools, limma-voom reports in all datasets the less FPs, followed closely by DESeq2, baySeq & DSS and in some occasions by ABSSeq and edgeR methods, too. PANDORA, although not of the same capacity with limma-voom in dissecting DEGs from non-DEGs, is however always among the top three performing algorithms, surpassing the other p-value combination methods. When biological replicates number is increased (*7x replicate design*), then all tools perform notably better, with DESeq, NBPSeq and NOISeq demonstrating the bigger AUFC. Trends remain mostly the same after p-value adjustment under a BH threshold of 0.05 (Appendix I; *Supplementary Figure 2*: **False Discovery Curves (FDC) usign adjusted p-values after EDASeq normalization.** FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. Trends remain the same after p-value adjustment. (BH p-value threshold: 0.05, EDASeq normalization)). From all p-value corrections NOISeq was excluded for reasons explained in (Soneson and Delorenzi 2013).

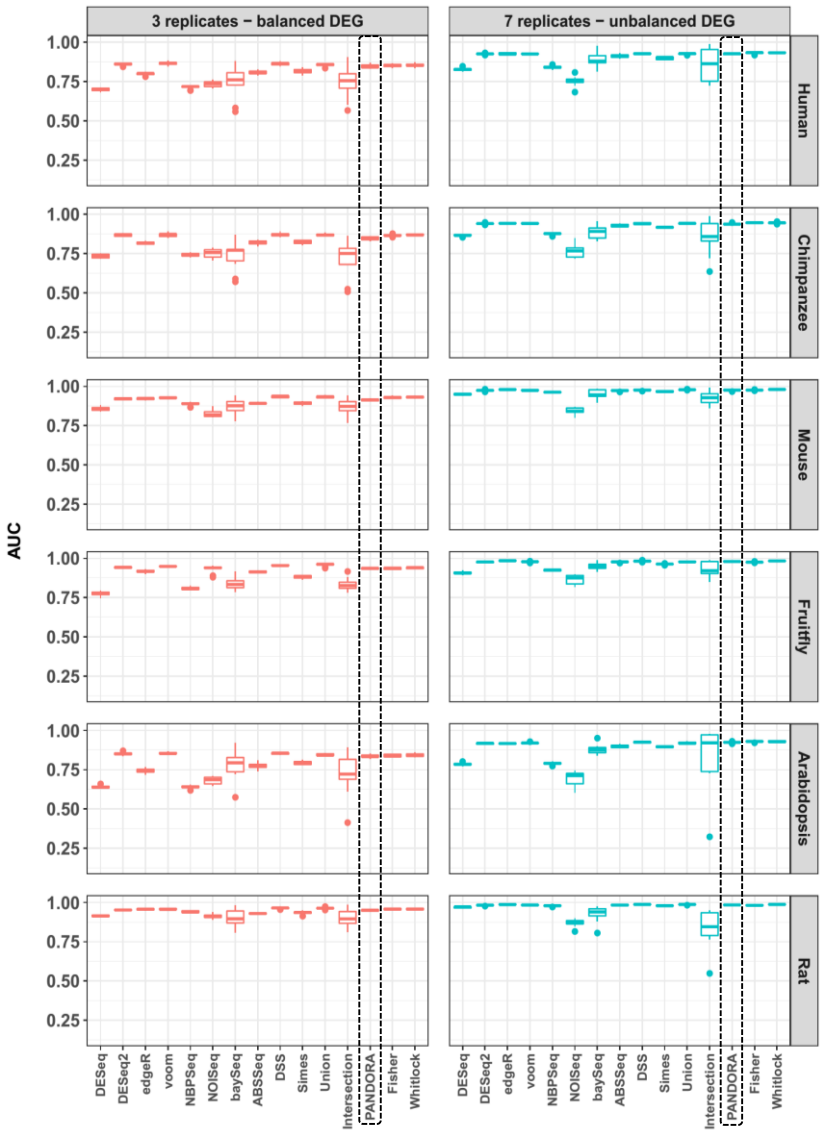


Figure 2: Area under the ROC curve (AUC) using raw p-values after Edaseq normalization. AUC are summarized across ten iterations for each tool and simulation design using unadjusted p-values. PANDORA and most other tools demonstrate an almost excellent performance, with DESeq, NBPSeq, NOISeq and baySeq being the exceptions. (significant p-value threshold: 0.05, EDASeq normalization)

As far as receiver operating characteristic (ROC) analysis summarized across 10 simulations is concerned (*Figure 2*), the majority of individual tools showed an adequate relationship between reported sensitivity and specificity for the *3x replicate design*. limma-voom, DESeq2 and DSS returned the bigger AUC, while on the contrary DESeq returned one of the smallest AUC, an observation that validates the findings in (Moulos and Hatzis 2015). Amid p-value combination algorithms, the poorer Intersection, whereas the

remaining four performed equivalently well with the top performing individual tools. At last, accumulating more replicates (*7x replicate design*) was beneficial for all tools except for NOISeq that remained practically unaffected. DESeq, NBPSeq, NOISeq and baySeq were again characterized by the smaller ROC analysis AUC. Trends remained the same after p-value adjustment under a BH threshold of 0.05. (Appendix I; *Supplementary Figure 3*).

As can be seen from the data in *Figure 3*, when examining False Discovery Rate (FDR) under a BH threshold of 0.05, baySeq and Intersection report in all cases the smallest FDR score, an expected phenomenon due to their inherent stringency.

Notably, PANDORA is constantly among the two or three second best performing algorithms along with limma-voom and ABSSeq. Fisher's and Whitlock's methods are constantly the worse two, not having the power to increase TPs without a simultaneous increase in FPs.

The next metric assayed, was the F_1-score, a precision–sensitivity tradeoff measurement. When the complete DEG list is taken into consideration (*Figure 4*), it is obvious that for the *3x replicate design* PANDORA is the best performing method for the human, chimpanzee and Arabidopsis datasets, while for the remaining three datasets it holds the second place, falling slightly behind ABSSeq. From the other p-value combination algorithms Simes is the one that always stands out and from the remaining eight individual statistical analysis tools baySeq has repeatedly the lowest F_1-score.

Consequently to increasing the biological replicates number (*7x replicate design*), F_1-score of most tools and datasets also tends to increase. PANDORA if not leading is still on top along with DESeq, edgeR and ABSSeq, while Simes maintains its relative position. Finally, baySeq shows unreliable performance as in the mouse, fruit-fly and rat datasets exhibits a satisfactory precision–sensitivity tradeoff that cannot be elsewhere replicated.Complementary to F_1-score findings, False Discovery Tradeoff (FDT) analysis further proves PANDORA's ability to find the golden ratio between precision and sensitivity (Appendix I; *Supplementary Figure 4*).

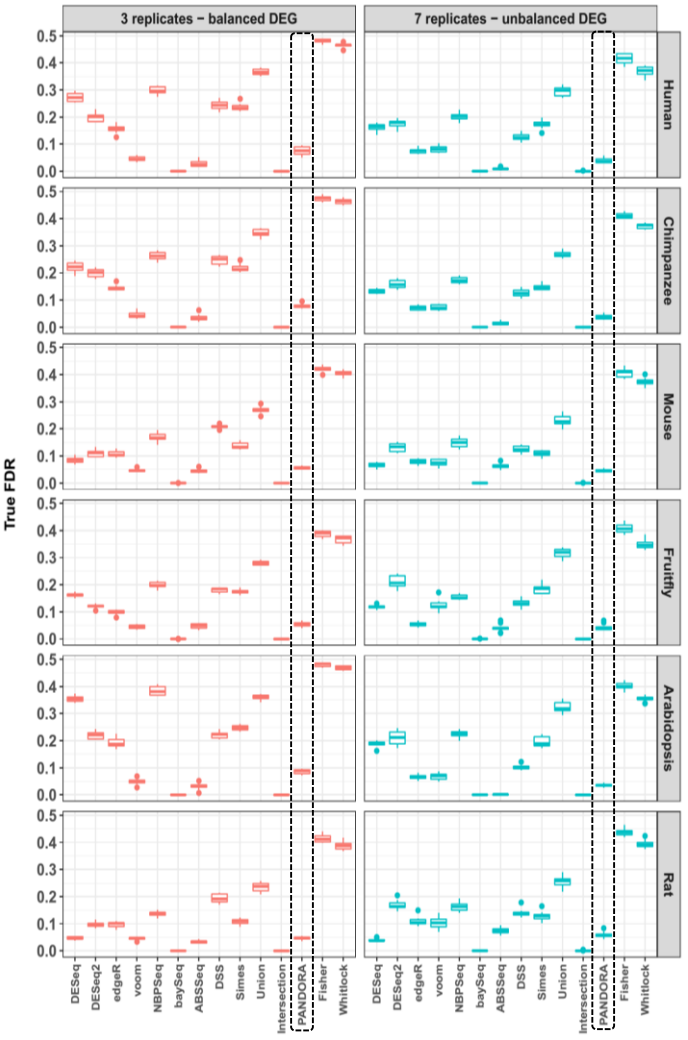


Figure 3: False Discovery Rate (FDR) using adjusted p-values after EDASeq normalization. FDR is summarized across ten iterations for each tool and simulation design at a BH adjusted p-value threshold of 0.05. PANDORA and ABSSeq share in most cases the second best FDR score behind the stringent baySeq and Intersection methods. NOISeq was excluded for reasons explained further above (EDASeq normalization)

The most interesting aspect of the F_1-score analysis, though, emerges from comparison of its values before and after p-value adjustment under a BH threshold of 0.05 (*Figure 4* versus *Supplementary Figure 5*; Appendix I). While most tools' performance is affected in a greater or lesser extend from p-value correction, PANDORA is perhaps the only DEA

procedure that behaves robustly both before and after p-value adjustment, rendering itself a dependable statistical analysis method to work with.

Finally, as would have been expected, methods highly capable in properly ranking FPs are not that capable in ranking FNs' below true hits (*Figure 1* versus *Supplementary Figure 6*; Appendix I). For example while limma-voom is in all datasets of the FDC's *3x replicate design* the tool that reports the less FPs, in the FNC case it delivers the most FNs for the human and Arabidopsis dataset and the second most (below baySeq) for all the other datasets. Strikingly, while similar trends can also be witnessed for most tools, they are not manifested to such an extent for PANDORA, suggesting that it can adequately control both FPs and FNs. Last but not least, it must be noted that after p-value adjustment with a BH threshold of 0.05, PANDORA performance significantly ameliorates surpassing that of other tools (Appendix I, *Supplementary Figure 7*).

To conclude with, by taking into consideration all the above simulation evaluation metrics we can infer that PANDORA is not only one of the best performing DEA methods under different organisms and dataset properties, but perhaps the most robustly behaving one, too.

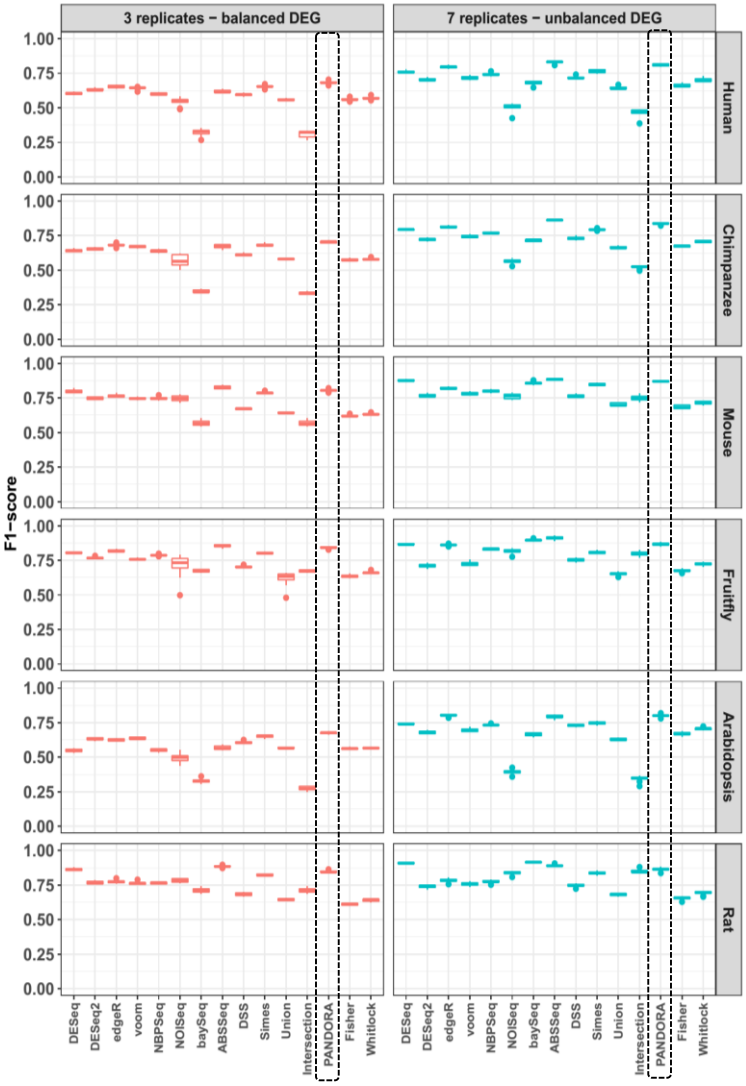


Figure 4: F_1-score (precision-sensitivity tradeoff) using raw p-values after EDASeq normalization. F_1-score summarized across ten iterations for each tool and simulation design, using unadjusted p-values. ABSSeq and PANDORA present consistently the best precision-sensitivity tradeoff, with only a few exceptions. (significant p-value threshold: 0.05, EDASeq normalization)

3.2 PANDORA versus other metaseqR2-implemented statistical analysis tools performance (II)

Real data

Evaluating tools' performance under the controlled "environment" of a simulation does not account for all the confounding factors that co-exist within a real dataset and can possibly affect DEA results. For this reason two real datasets, the SEQC and the TaqMan datasets, were also analyzed (see section 2.4 for more details) and the ROC and F_1-score exploratory diagrams were plotted. It must be noted that weights used for the analysis were the estimated ones from the human simulated dataset, in order to avoid unwanted biases introduced by technical replicates. As with the simulated data analysis, EDASeq was chosen as a global normalization procedure, but again, individual tools' normalization methods were also separately applied (see Appendix II).

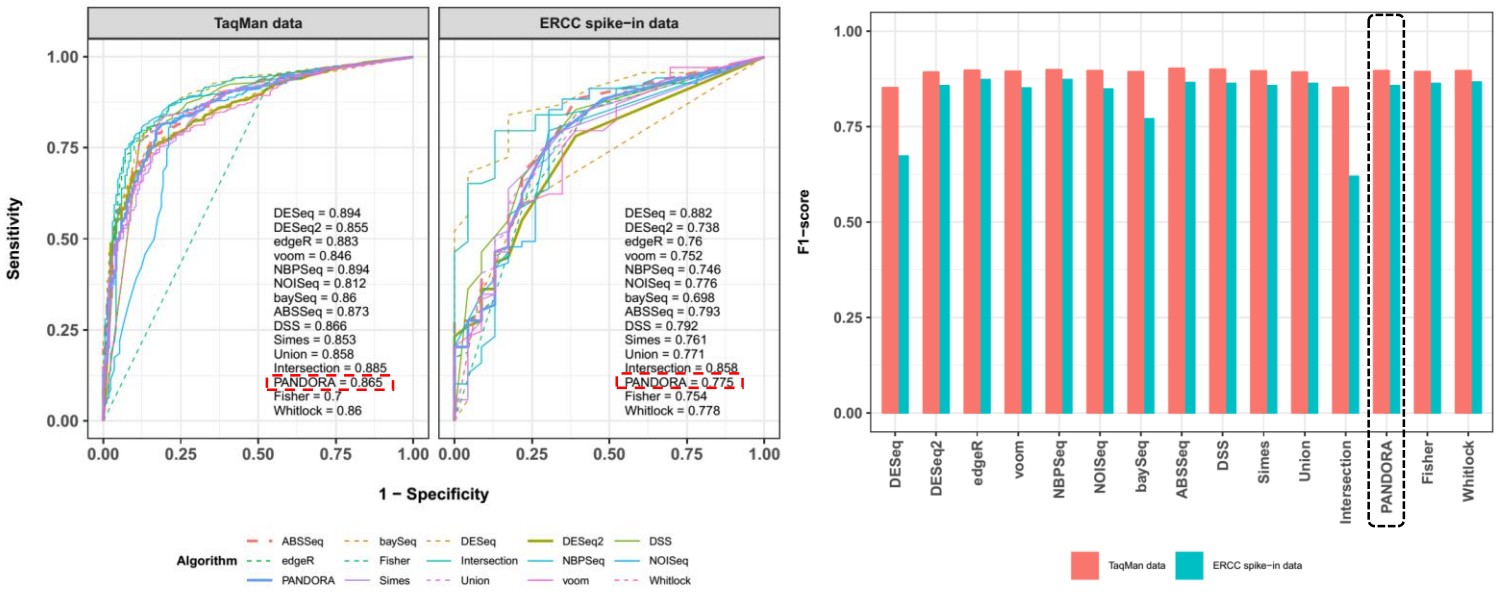


Figure 5: ROC and F1-score analysis of real datasets using raw p-values after EDASeq normalization (on the left and right respectively). AUC can be found at the bottom right corner of the ROC figures. PANDORA shows an adequate sensitivity-specificity relationship when examining the two real datasets. Its precision-specificity threshold is equally good with that of the top performing tools for both datasets. (significant p-value threshold: 0.05, EDASeq normalization)

ROC analysis and F_1-score results for both datasets are depicted on *Figure 5*. Most AUCs, whilst prima facie the same, do present some differences. For instance, DESeq has in both TaqMan and ERCC data the higher AUC and NOISeq along with baySeq the smaller ones, respectively. In addition, NBPSeq achieves an equal to DESeq's AUC when TaqMan data are analyzed. PANDORA, although not the best method, performs adequately well in both cases occupying a median AUC score in comparison with the rest of the tools. From the other p-value combination methods, Fisher is constantly the worst performing one. Last but not least, after p-value adjustment using a BH threshold of 0.05 (Appendix I; *Supplementary Figure 8*), trends are kept the same.

Table 1. False Discovery Rates Approximation using three "same versus same" comparisons.
(BH adjustment threshold 0.05; EDASeq normalization)

	DESeq	DESeq2	edgeR	voom	NBPSeq	baySeq	ABSSeq
SEQC_A	0,0239	< 0,0001	0,0007	< 0,0001	0,0177	< 0,0001	< 0,0001
SEQC_B	0,0002	0,0002	0,0012	< 0,0001	0,0019	0,0001	< 0,0001
Chimpanzee	0,0028	0,0004	0,0004	< 0,0001	0,003	< 0,0001	< 0,0001
	DSS	**Simes**	**Union**	**Intersection**	**PANDORA**	**Fisher**	**Whitlock**
SEQC_A	0,0029	0,0214	0,0355	< 0,0001	0,0004	0,0588	0,0401
SEQC_B	0,0003	0,0014	0,0048	< 0,0001	< 0,0001	0,0379	0,0104
Chimpanzee	0,0061	0,0033	0,0147	< 0,0001	< 0,0001	0,0744	0,0642

With regards to the TaqMan data F_1-score, it is almost the same for the vast majority of DEA methods, except for DESeq and Intersection that appear of poorer capability to achieve a good balance between precision and sensitivity. On the other hand, F_1-score for the ERCC dataset is more variable with several tools presenting slight differences among each other. Markedly, DESeq's and Intersection's substandard performance is even worse than in the TaqMan dataset. Not surprisingly, baySeq's varying performance

validates the simulated data reported previously (*Figure 4*). Finally, as was also seen for the ROC analysis, F_1-score trends did not change under a p-value BH adjustment threshold of 0.05 (Appendix I; *Supplementary Figure 9*).

Additionally to the ROC analysis and the F_1-score, the final metric computed is the approximate true False Discovery Rates (aFDR) under a BH threshold of 0.05 for all methods applied except NOISeq for reasons explained in (Soneson and Delorenzi 2013) (Table 1). aFDR rates were computed across tools using three different "same versus same" mock analyses: one by using data from the real chimpanzee dataset and two by splitting each SEQC dataset group (SEQC_A and SEQC_B) into two subgroups. Thus, from the results summarized in Table 1 it is obvious that voom and ABSSeq, closely followed by baySeq and DESeq2 are the best performing individual tools, while Intersection and PANDORA the first and second best from the p-value combination methods, respectively.

Overall, these results suggest that most tools perform adequately well in both real dataset analyses and that PANDORA behaves as good as the top performing algorithms, validating the simulation results of the previous section.

3.3 Effects of different normalization methods upon downstream DEA.

Simulation data.

As already mentioned, normalization is one of the most crucial steps in DEA of RNA-seq data and a great variety of methods have been developed to cope with this need. As most DEA tools suggest specific normalization method(s) to be coupled with their implemented statistical analysis, it is logical to speculate that there may be a dependence between them in order to attain optimal performance.

To address this hypothesis, we normalized in parallel simulated data of the human, chimpanzee, mouse, fruitfly, arabidopsis and rat datasets with DESeq or TMM (edgeR) normalization algorithms, prior to differential expression statistical analysis, performance evaluation and subsequent comparison with EDASeq normalization results. It must be noted that baySeq statistical analysis was not coupled with DESeq normalization due to extensive computational time required. For simplicity reasons only the F_1-score (*Figure 7*) and the area under the ROC curve (*Figure 6*) as computed after DESeq and TMM normalization will be here described. All the other metrics' figures created using the alternative two normalizations can be found in Appendix III.

The comparison of *Figure 7* with *Figure 4* is revealing in several ways. First of all, it is obvious that TMM is not as powerful as EDASeq or DESeq normalization methods in controlling outliers. Secondly, while the general trends seem to remain generally stable between the three normalization approaches, there are tools like ABSSeq that are tremendously affected by the normalization procedure chosen. On the contrary, PANDORA is one of the few methods that performs robustly over all three normalization schemes. In addition, PANDORA is not that heavily affected by outliers under TMM normalization.

A similar, but less intense dependence between normalization and statistical analysis methods can also be observed when ROC analysis results (*Figure 6*) are compared with their EDASeq counterparts (*Figure 2*).

In summary, given metaseqR's and metaseqR2's concept of providing various normalization and statistical analysis tools that can be differentially combined by the user during an analysis, PANDORA might consist the safer DEA choice for most everyday RNA-seq data analyses.

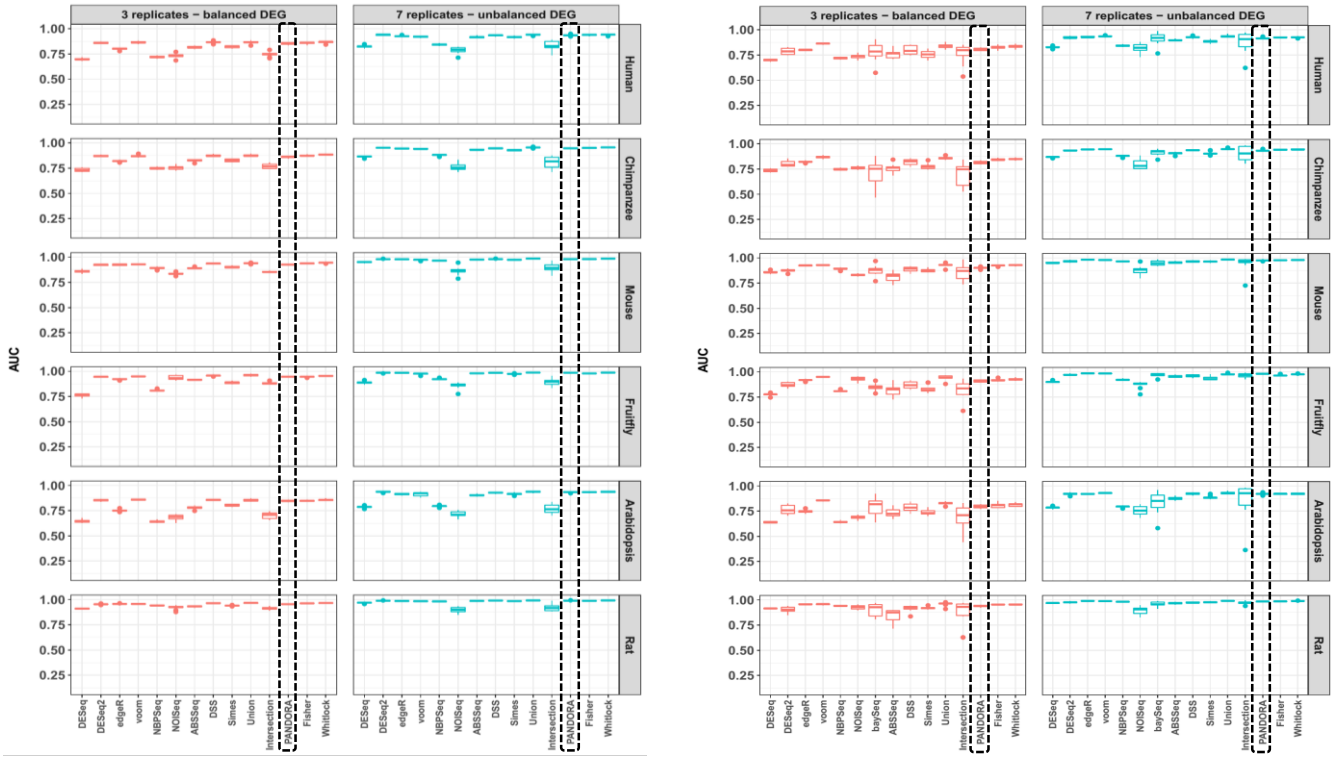


Figure 7: ROC analysis for DESeq and TMM normalized simulation data (on the left and right respectively) across ten iterations for each tool and simulation design. (significant p-value threshold: 0.05)

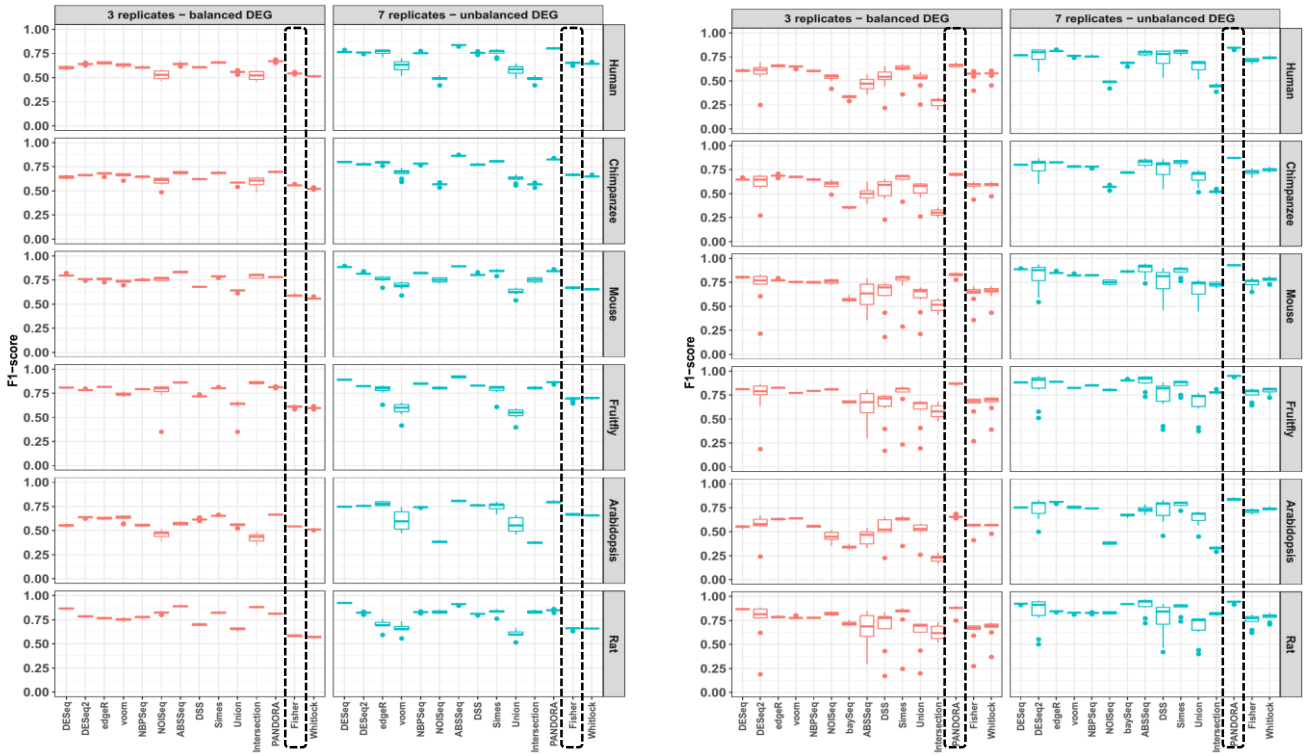


Figure 6: F_1-score analysis for DESeq and TMM normalized simulation data (on the left and right respectively) across ten iterations for each tool and simulation design. In contrast to many tools like ABSSeq, PANDORA behaves robustly under all three normalization procedures (significant p-value threshold: 0.05)

3.4 Gene length propagation into PA.

Simulation data.

RNA-seq technology is prone to multiple bias sources. Gene length bias might be the first one recognized by the scientific community and it has been shown to affect not only DEA itself, but also to propagate into downstream PA (for more details see Section 1.4). While several tools, like these in (Young et al. 2010; Gao et al. 2011; Mi et al. 2012), have been published to correct for this bias during PA, to our knowledge there is no previous research on whether p-value combination algorithms like PANDORA, have the ability to control gene length bias's effects on PA.

With the aim of witnessing if the "corrected" DEG lists reported by metaseqR2-implemented p-value combination methods are a more reliable "starting material" for PA, we simulated gene length bias using the human dataset (see Section 2.4 for more details) and then, we performed DEA using newly estimated tool-specific PANDORA weights (practically the same with the default ones). EDASeq normalization was again globally applied.

Prior to continuing into PA, gene length bias's existence at the DEA level should be validated. Thus, every DEG list reported was first separated into 50 equal bins according to the binary logarithm of its genes' length. Then the per bin mean gene length was plotted against each bin's DEG percentage, in an attempt to specify those simulation iterations where all tools were simultaneously affected by the bias (positive loess curve's slope accompanied by a p-value < 0.05 for each tool). Unfortunately, only a few simulation iterations were significantly affected, leading us to continue our benchmarking with just one of them

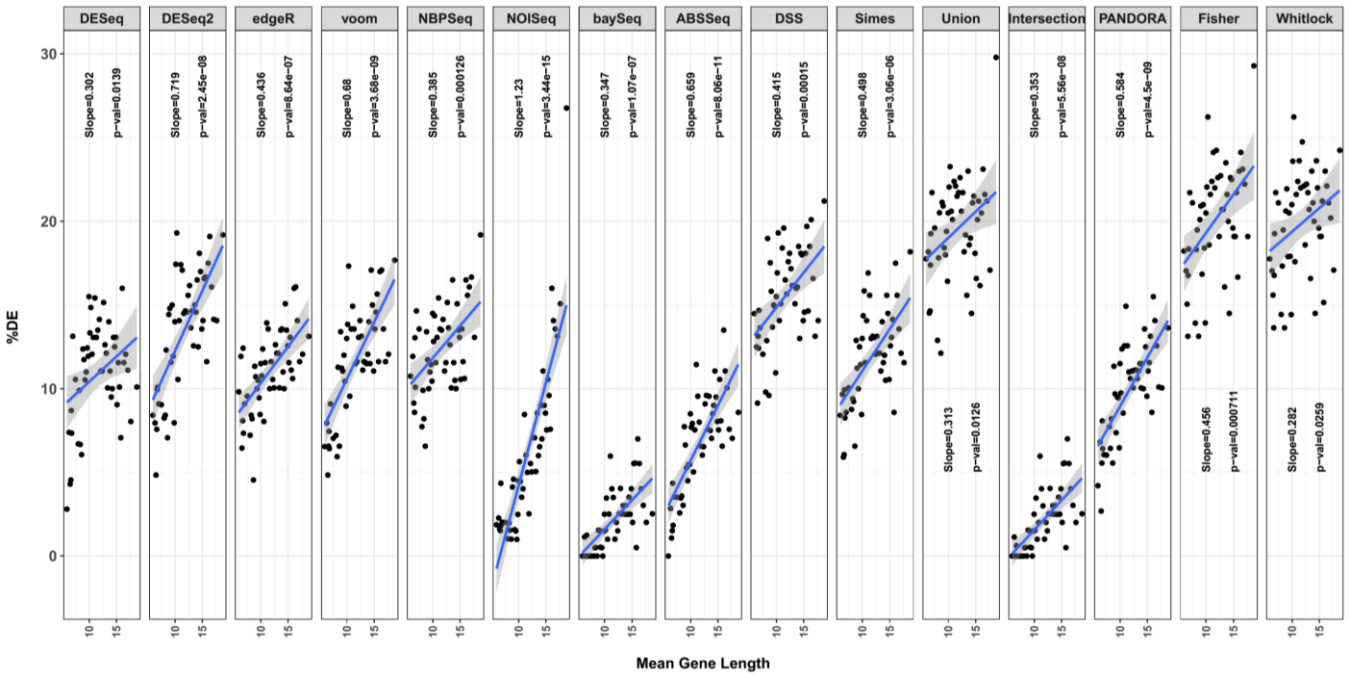


Figure 8: Investigation of gene length bias presence within the human-based simulation dataset. Genes were binned into 50 groups by their length binary logarithm. On the x-axis each bin's mean gene length is depicted and on the y the per bin DEG percentage respectively. It is apparent that all tools are polarized towards reporting more often longer genes as DE. (significant p-value threshold: 0.05)

belonging to the *3x replicate design*. *Figure 8*, is created based on data derived from this iteration

Once pre-ranked GSEA was performed as described in Section 2.7, non-significantly over-represented pathways (p-value > 0.05) were filtered out and the remaining ones undergone another filtering round to

keep those characterized by a positive enrichment score (ES). Downstream analysis was based on the assumption that if there is no length bias present, then the gene length distribution of the genes belonging to significant pathways should be similar to the length distribution of all genes tested. Consequently, unique pathway genes were kept and kernel densities of their log transformed length distributions were calculated. The same process was also followed for all simulated genes of the respective iteration and the distributions differences were quantified using the Kolmogorov-Smirnov test statistic *(Figure 9*). This pipeline was in parallel applied for both all pathway genes and for only the pathways' leading edge genes so as to reveal any masking effects (leading edge genes are a pathway's genes that contributed the most at the establishment of this pathway's enrichment within a given DEG list; therefore, they can be assumed as of major biological importance).

As can be seen from *Figure 9*, when all unique pathway genes are taken into consideration, PANDORA and edgeR followed by DESeq are the most capable tools in controlling gene length bias propagation into PA, whereas voom, NOISeq and Simes are the worst performing ones among the individual and p-value combination methods, respectively. Furthermore, when only leading edge genes are analyzed, more sharp differences can be indeed observed. PANDORA again gives the most suitable DEG list to perform pre-ranked GSEA with, whereas this time DSS, Intersection and baySeq are the worst ones.

Taken together, these results suggest that, although all metaseqR2-incorporated tools can be affected by gene length bias that may be present in a dataset, some of them exhibit better bias handling leading to more reliable PA results, when pre-ranked GSEA is performed.

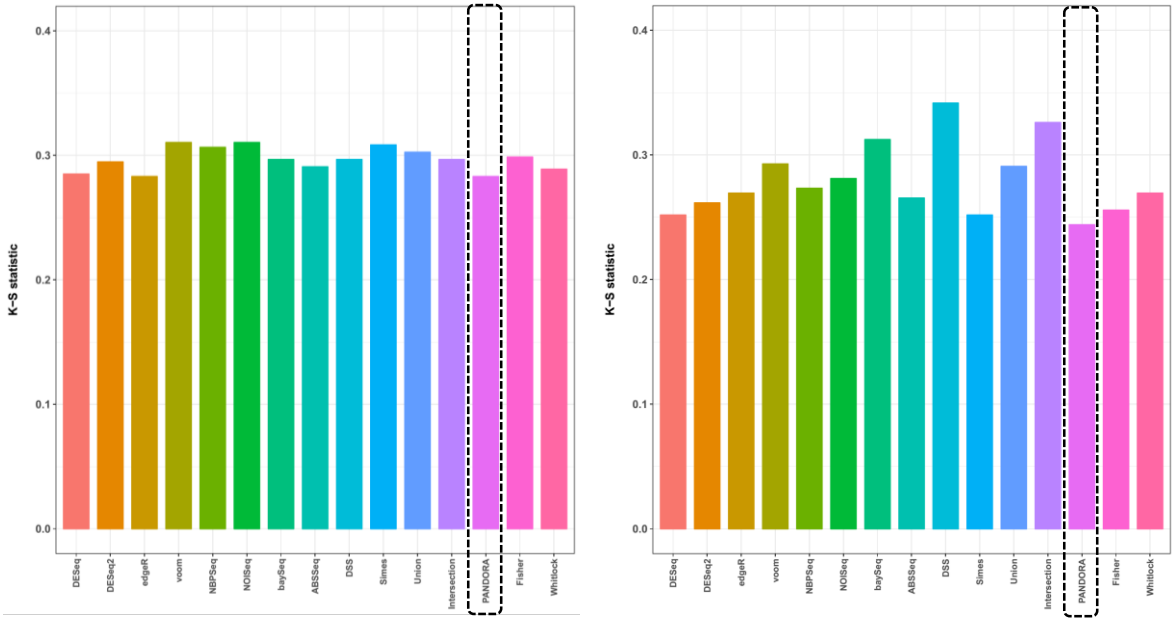


Figure 9: Kolmogorov-Smirnov analysis for all unique pathway genes and unique leading edge genes (left and right respectively). Kolmogorov-Smirnov test reveals that PANDORA returns the most reliable DEG list to use in pre-ranked GSEA, suggesting an advantage of our method against all others when gene length bias is present. (significant p-value threshold: 0.05)

3.5 DEA of lncRNAs.

Simulation data.

Apart from gene-length bias, another RNA-seq data analysis challenge has recently attracted the attention of experts: that of lncRNA DEA (for more details see Section 1.4). In a recent publication, very popular DEA algorithms (including metaseqR2-implemented ones) showed inferior ability to analyze lncRNAs'

differential expression, while at the same time that of mRNAs was better approached. Towards, investigating if lncRNAs are better represented on final DEG lists using PANDORA or any other metaseqR2-incorporated tool, we performed DEA of NGP-nutlin-based simulation data and evaluated performance of all tools for mRNAs and lncRNAs separately. No gene or exon filters were applied during the analysis, as many lncRNAs would had been filtered out. Once again, EDASeq normalization was applied.

To begin with, FDCs of all metaseqR2 tools under both simulation designs and RNA biotypes, as illustrated in *Figure 10*, validate (Assefa et al. 2018) observation that lncRNAs are indeed underrepresented during DEA (they are characterized by bigger AUFC with respect to mRNAs). However, it must be also noted that relative tools' performance is almost retained as previously reported (*Figure 1*) and that increase of biological replicates number do somehow alleviate the bias. In addition, PANDORA is one of the best performing tools for both biotypes and especially under the *3x replicate design.* P-value adjustment using a BH threshold of 0.05 results in no major changes (Appendix I; *Supplementary Figure 10*).

Commenting on the fact that AUFC scores in *Figure 10* are much bigger than those reported in *Figure 1* or any other relative figure, we must highlight the fact that AUFC score is highly depended on the subset of the original dataset analyzed. For example, if the whole dataset is taken into consideration (Appendix I; *Supplementary Figure 11*) then AUFC scores are of the same order of magnitude with those reported in *Figure 1*.

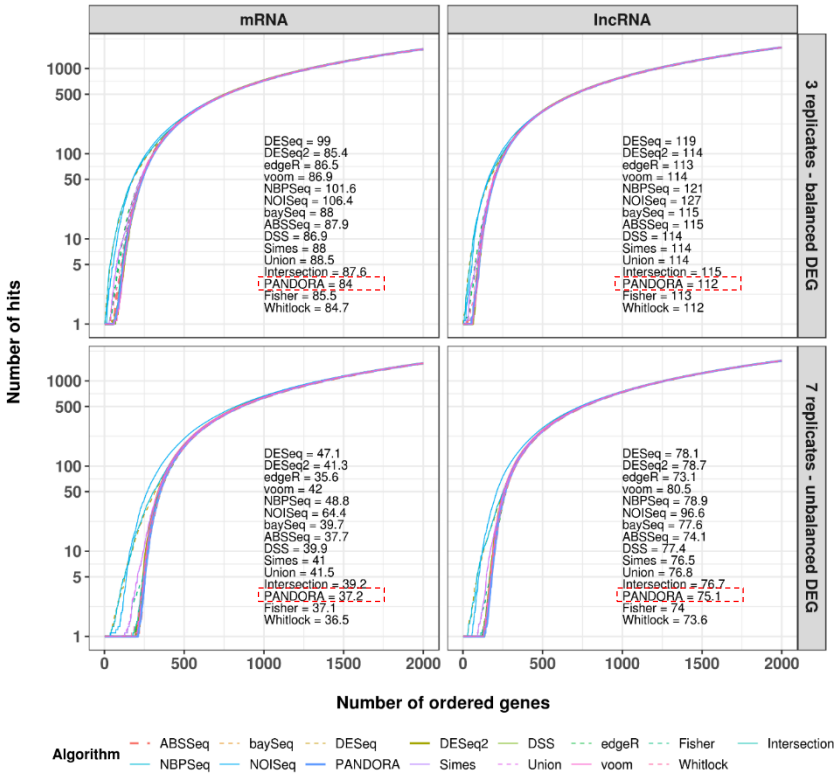


Figure 10: False Discovery Curves (FDC) using raw p-values after EDASeq normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. Performance was separately assayed for mRNAs and lncRNAs. Indeed all tools performance is compromised when lncRNAs are analyzed.(significant p-value threshold: 0.05, EDASeq normalization)

Therefore, *Figure 10* AUFC scores are intended only for relative comparisons between the two biotypes and/or simulation designs, and must not be considered as proper metrics for inferring about general tools' performance.

As far as FDR analysis is concerned (*Figure 11*), mRNA versus lncRNA differences are not that obvious and more strikingly, some tools report a smaller, although more disperse, FDR score for lncRNAs than for mRNAs (e.g. NBPSeq in *3x replicate design*). Another very interesting observation is that in contrast with most cases of *Figure 3*, PANDORA performs better than ABSSeq when seven biological replicates are used. Finally, very low TP to (FP+FN) ratio rates of Intersection and baySeq is once again expected for reasons explained in Section 3.1

Furthermore, adding to data reported in *Figure 4*, PANDORA not only continues to demonstrate an excellent F_1-score along with ABSSeq for both biotypes, but it even surpasses that of its "competitor" in the *3x replicate design* (*Figure 12*). In all cases, second best from the individual tools comes edgeR and Simes from the p-value combination ones. At last, given adjusted p-values (Appendix I; *Supplementary Figure 12*), many tools' performance is compromised. However, PANDORA demonstrates once more a robust behavior that is especially apparent in the *7x replicate design* of both biotypes.

To conclude with, the above data suggest that lncRNA bias is indeed present when analysis is performed with the existent DEA algorithms. However, it is obvious that not all of them are affected with the same severity and on the contrary there are even several methods, like edgeR, ABSSeq and PANDORA that are capable of alleviating lncRNA bias at a sufficient degree.

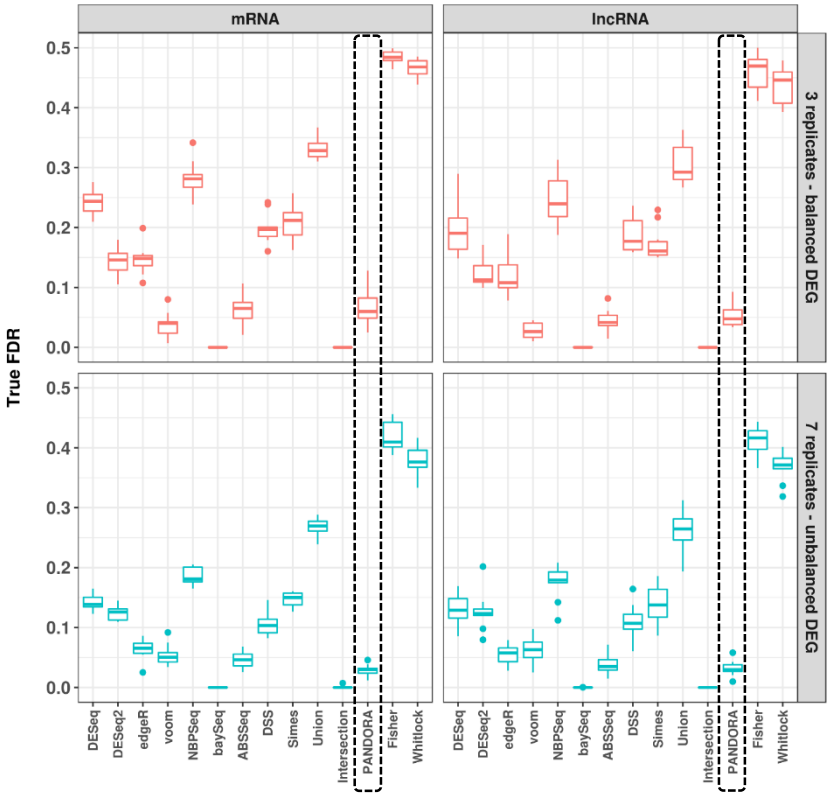


Figure 11: False Discovery Rate (FDR) using adjusted p-values after EDASeq normalization. FDRs are summarized across ten iterations for each tool and simulation design at a BH adjusted p-value threshold of 0.05. In contrast with *Figure 3* data PANDORA outperforms ABSSeq both for mRNA but most importantly for the lncRNA biotype. (EDASeq normalization)

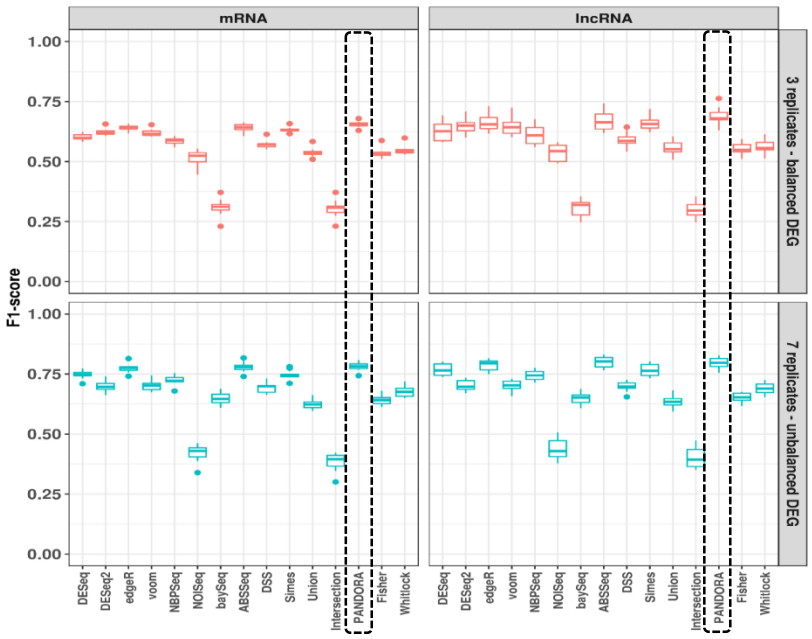


Figure 12: F1-score (precision-sensitivity tradeoff) using raw p-values after EDASeq normalization. F_1-score is summarized across ten iterations for each tool and simulation design, using unadjusted p-values. ABSSeq and PANDORA present in all cases the best precision-sensitivity tradeoff, with the later one even surpassing the former in the *3x replicate design.* (significant p-value threshold: 0.05, EDASeq normalization)

3.6 metaseqR2 tools concordance analysis using three different biotype designs.

Real data; individual metrics; hierarchical clustering.

After completing tools' evaluation using NGP-nutlin simulated data we moved in the final project's part to check concordance between metaseqR2 statistical analysis methods deploying DEA results of the actual NGP-nutlin dataset. Concordance analysis was performed in similar concept with that of (Assefa et al. 2018), applying the metrics described in Section 2.5.

When observing *Figure 13*, which depicts the number of DEG reported by each tool using unadjusted p-values, it is apparent that DESeq, Intersection, and baySeq are the more strict methods while DSS, Union, Fisher and Whitlock the most loose. PANDORA shows a median performance in all three biotype schemes possibly allowing for a more realistic representation of the real gene expression. The same trends remain after p-value adjustment, but the number of DEGs reported across all biotypes is globally reduced with the exception of Union, Fisher and Whitlock (Appendix I; *Supplementary Figure 13*).

Biotype representation ratio (*Figure 14*) comes to support previous section results. More specifically, all tools show a slight over-representation of mRNAs and a heavier under-representation of lncRNAs in the finally reported DEG list (ratio above and below zero respectively). Both biotypes were used as a successful formula control (see Section 2.5). Interestingly, whereas mRNA over-representation degree has minor inter-method differences, the same does not apply for lncRNA results More specifically, being the most permissive of all metaseqR2 tools, Union best represents differentially expressed lncRNAs in expense of encompassing many FPs among the reported discoveries. On the other hand, in an attempt to filter out as many FPs as possible, Intersection reports the smallest number of statistically significant DE lncRNAs, followed closely by baySeq. PANDORA although not the best it demonstrates an adequate lncRNA bias control.

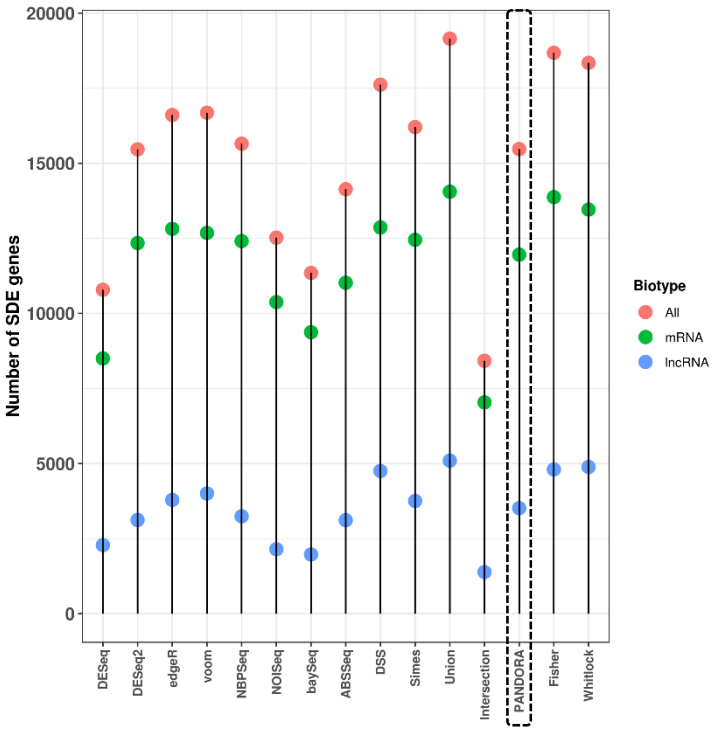


Figure 13: Number of DEGs per tool and biotype, using unadjusted p-values. PANDORA is neither too lose nor too strict as a DEA tool, for all three biotype designs. (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; significant p-value threshold: 0.05, EDASeq normalization)

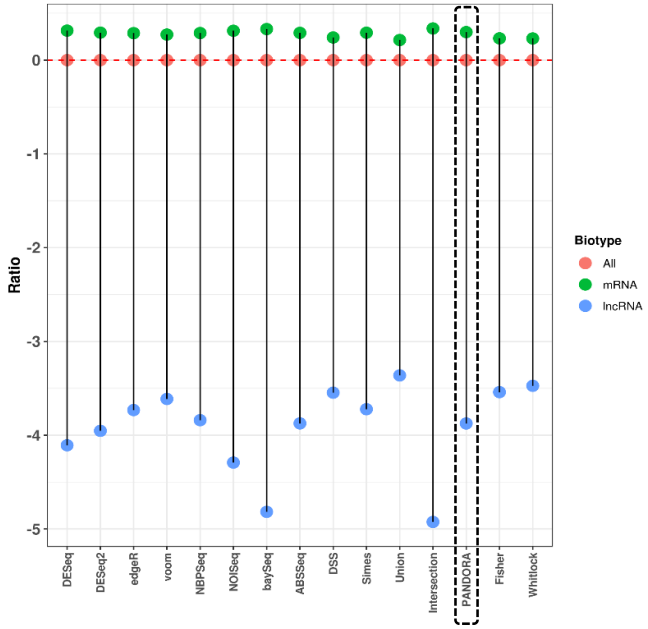


Figure 14: Biotype representativeness ratio, using unadjusted p-values. While mRNAs are slightly over-represented in all tools final DEG list, lncRNAs are heavily under-represented in all cases. (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; significant p-value threshold: 0.05, EDASeq normalization)

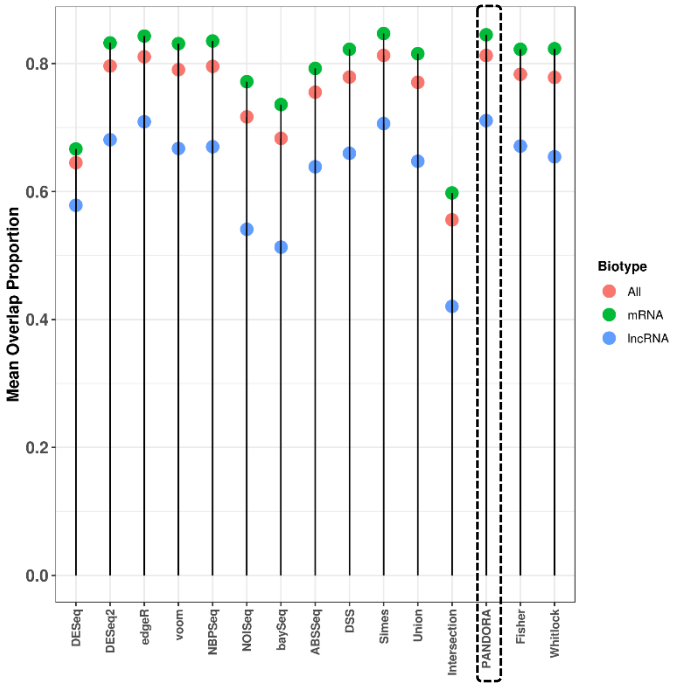


Figure 15: Mean overlap proportion of DEGs for each tool and biotype scheme, using unadjusted p-values. It is obvious that lncRNA results are more variable than these of mRNAs between examined methods. PANDORA along with edgeR and secondly Simes show the best accordance with the rest of the tools for all three biotype analyses. (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; significant p-value threshold: 0.05, EDASeq normalization)

Finally, after p-value adjustment (Appendix I; *Supplementary Figure 14*), all individual tools performance declined with regard to lncRNAs' analysis resulting in a subsequent drift of p-value combination methods as well.

Further differences in tools' behavior are shown when examining DEG list mean overlap proportion (*Figure 15*), which suggests a more significant accordance between mRNAs reported as DE and a much less one for lncRNAs. Reporting both biotypes concludes in a median overlap as expected. Notably, PANDORA and edgeR followed by Simes have the more extensive mean overlap with all other tools while DESeq and Intersection the more limited ones. After p-value correction, all tools DEG lists "experience" a broader divergence from one another, but general trends remain the same (Appendix I; *Supplementary Figure 15*).

The last metric used (that of mean ranking correlation) reveals that DEG ranking is highly variable among tools with PANDORA showing the best and almost identical for all biotype schemes ranking correlation degree, closely followed by that of edgeR (*Figure 16*). The more divergent individual tools' DEG rankings comes from DESeq, whereas for the combination methods from Fisher. Strikingly, for some tools like DESeq and DESeq2, ranking correlation is larger for lncRNAs than that for mRNAs, a phenomenon that becomes more apparent and most importantly global using corrected p-values (Appendix I; *Supplementary Figure 16*).

Lastly, in order to summarize concordance analysis results we performed hierarchical clustering of individual metrics' output in a similar way with that in (Assefa et al. 2018) (see Section 2.6). Dendrograms using raw p-values are presented in *Figure 17* to 19 and the respective ones after p-value adjustment are shown in Appendix I *Supplementary Figure 17* to 19.

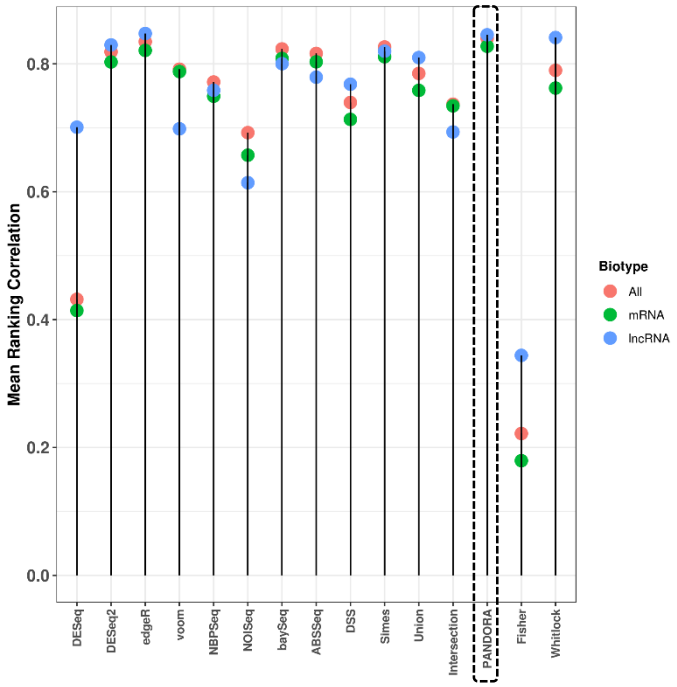


Figure 16: Mean DEG ranking correlation between DEA tools for all biotype scheme, using unadjusted p-values. Interestingly, lncRNAs ranking is more comprehensive than that of mRNAs for some tools including PANDORA, which also displays an excellent and almost identical degree or ranking correlation with other tools' results. (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; significant p-value threshold: 0.05, EDASeq normalization)

What is first of all shown by hierarchical clustering is that PANDORA always groups along with the same four or five individual methods: DESeq2, edgeR, voom, NBPSeq and ABSSeq. Simes is also part of this particular group. Secondly, DSS, Whitlock and Union form another cluster that constantly appears separate from all others. Finally, p-value correction gives no different results as compared to unadjusted ones, with the only exception of voom grouping together with DSS, Whitlock and Union in the lncRNA biotype (Figure 19 versus Appendix I; *Supplementary Figure 17*).

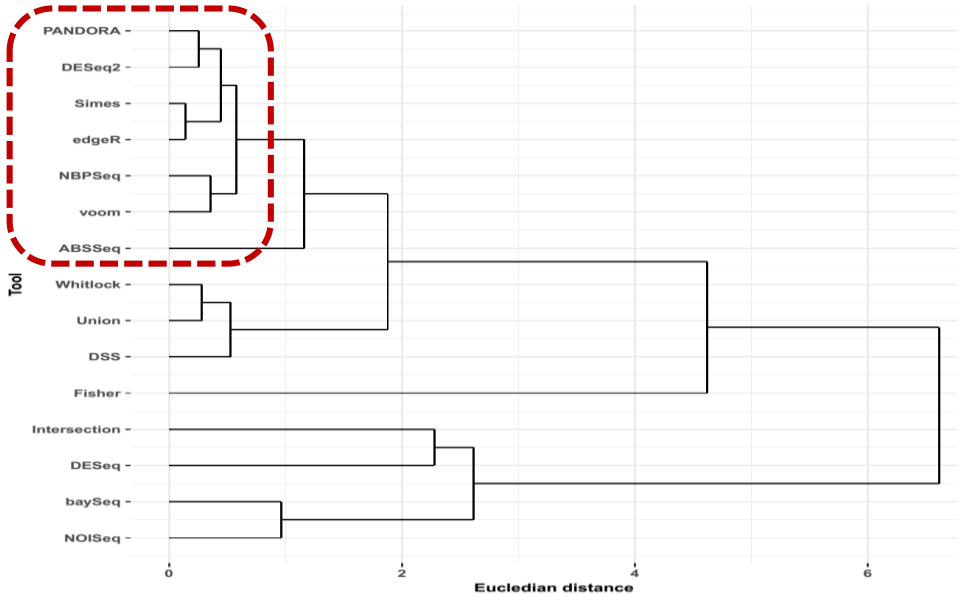


Figure 17: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. Both mRNA and lncRNA biotype scheme is here examined. (significant p-value threshold: 0.05, EDASeq normalization)

Conclusively, lncRNAs bias is evident for all metaseqR2 analysis tools. However, PANDORA's results are the most consistent and if not the best it is always among the best ones for all biotype designs (mRNAs, mRNAs-lncRNAs and lncRNAs).

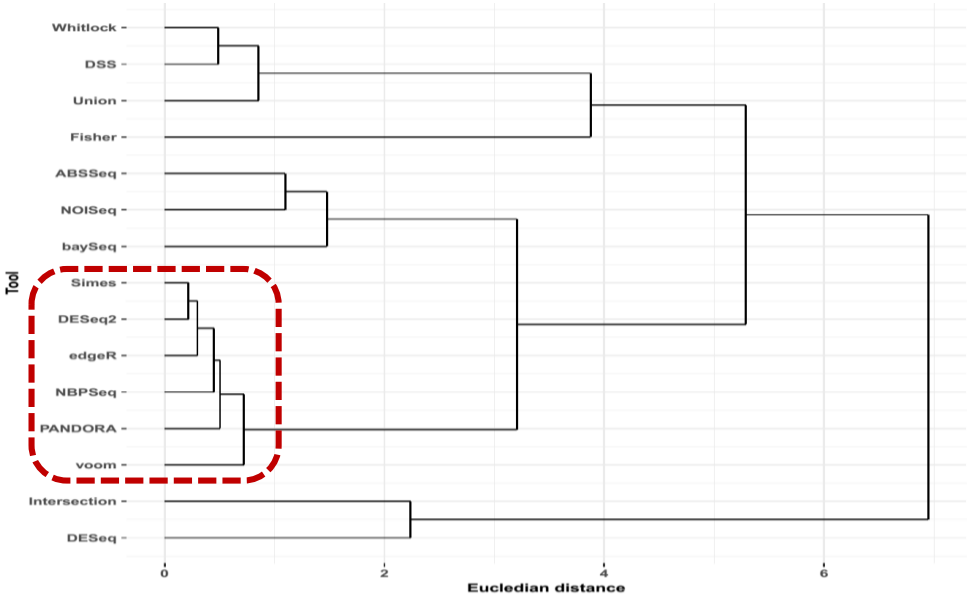


Figure 18: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. mRNA biotype scheme is here examined. (significant p-value threshold: 0.05, EDASeq normalization)

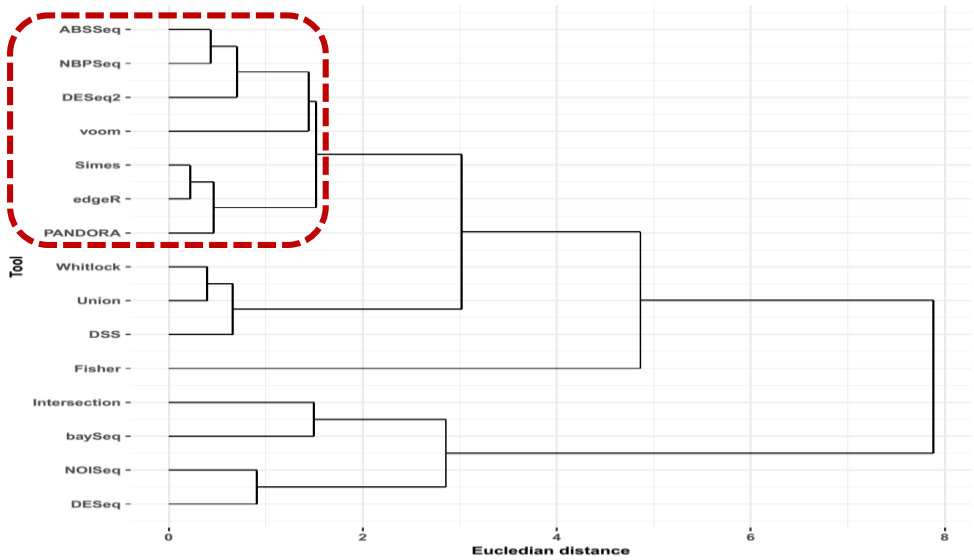


Figure 19: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. lncRNA biotype scheme is here examined. (significant p-value threshold: 0.05, EDASeq normalization)

4. DISCUSSION

RNA-seq has largely replaced previously used microarray technology in everyday laboratory practice due to both higher fidelity of results reported and to a wider range of possible applications. Unfortunately, RNA-seq data analysis has proven to be no less complicated than that of predecessor technologies as both extensive raw data normalization and sophisticated statistical analysis methods are required to tackle with "lurking" biases. More specifically, for the DE statistical analysis of RNA-seq data between two or more conditions and/or time series, many different approaches have been proposed for over one decade, but not enough effort has been invested for the development of combined approaches in the field.

Thus, in this dissertation thesis we focused on the upgrade of metaseqR, a powerful package for DEA of RNA-seq data, that apart from the individual methods included, it also allows access to 6 different p-value combination processes and several normalization methods. Furthermore, because PANDORA, one of the metaseqR-implemented p-value combination algorithms that was developed by our laboratory, had already shown enhanced performance in contrast to many other individual tools, further research on its behavior under different biases has been conducted.

To begin with, metaseqR2, the new package developed, includes three new statistical analysis tools: DESeq2, ABSSeq and DSS, which were chosen instead of many others due to proper maintenance, superior performance and compatibility with respect to the rest of individual tools shipped with metaseqR. Additionally, the list of supported organisms by computing PANDORA weights for *Rattus norvegicus.*

Next, by performing simulation studies based on six real datasets with two experimental configurations each, we observed that our PANDORA method is one of the best tools in ranking properly FP and FN relatively to TP and TN hits respectively. Most interestingly, it is also one of the few approaches that can achieve a good tradeoff between true and false hits at an adequate level. At the same time, ROC analysis placed PANDORA among the top performing tools, too. In addition, FDR score calculation using Benjamini- Hochberg corrected p-values showed that PANDORA along with ABSSeq and voom report an importantly small number of FPs with respect to total discoveries. On the other hand, the very good FDR control levels achieved by baySeq and Intersection can be first of all attributed to their inherent stringency and secondly it is not adequate in distinguishing these tools from the rest, if examined along with other metrics describing their general behavior and accuracy levels. Lastly, complementary to one another, results of the F_1 and the FDT measurements suggested a most promising precision-sensitivity tradeoff reported by PANDORA, ABSSeq and DESeq as well.

Eventually, real data analysis was performed for two datasets that bear qPCR or spike-in data as functional validation of the computationally reported differential gene expression. Via ROC analysis of both datasets, PANDORA demonstrated an adequate, sensitivity-specificity relationship, whereas its F_1-score was among the best delivered. Interestingly, baySeq, Intersection and even DESeq were characterized by an inferior precision-specificity tradeoff value especially for the ERCC spike-in dataset, a phenomenon consistent with these reported by (Moulos and Hatzis 2015). Finally, using three "same versus same" mock comparisons PANDORA along with voom, ABSSeq, baySeq and Intersection proved to have an amazing true FDR, which is even more important for the first three methods given their higher than baySeq and Intersection performance in other metrics.

Additionally, in order to assess consistency of results delivered by each metaseqR2 method after the application of different normalization procedures, we performed in parallel DEA of the previously reported simulation data after three different normalization methods: EDASeq (default), DESeq and TMM (of edgeR package). These three procedures were particularly chosen due to their popularity among bioinformaticians/ computational biologists (personal literature observation), as well as due to their

superior performance relatively to the other metaseqR2-offered normalization methods, like RPKM, quantile and total count (Dillies et al. 2012). Unexpectedly, while precision-specificity tradeoff of many tools such as ABSSeq, showed a high dependence on the normalization algorithm applied, PANDORA performed very robustly under all three normalization methods and experimental designs. Thereby, it proved itself to be a reliable solution for DEA in general and especially under metaseqR2's concept of enabling normalization and statistical analysis algorithms to be combined by the user at will. Finally, this experiment indicates once more the importance of proper normalization before RNA-seq DEA, in agreement with many previous researches such as (Dillies et al. 2012).

Afterwards, driven by literature findings indicating gene length bias propagation into PA, we set to investigate the possibility of this bias being controlled by any of the metaseqR2-implemented tools. Commenting on our choice to investigate gene length bias as it is presented in the PA results, we would like to stress out the fact that RNA-seq data are characterized by relatively large levels of underlying noise, as most high-throughput techniques. This noise is not completely eliminated even at the level of having come up with a short-list of differentially expressed genes through the usage of appropriate algorithms. Hence, making use of a DEG list directly may not have been the most appropriate approach for studying the bias. On the other hand, being able to check gene length bias handling in the context of a biologically important group of genes, like the ones reported by GSEA, is less noise dependent and might constitute a better option for such an investigation.

Thus, after simulating gene length bias using the human dataset, we first of all searched for simulation iterations where all thirteen tools were significantly affected by the bias. Unfortunately, too few iterations fulfilled this criterion. This phenomenon although prima facie awkward could be possibly in agreement with the findings of (Yoon and Nam 2017). Briefly, Yoon and his colleagues propose that gene length bias is only a small source of the total observed RNA-seq data variability and that it is specifically introduced when NB modelling is used for the analysis. Consequently, if a dataset's replicates are independent biological entities, huge inter-sample variability will mask gene length bias and the latter will not be detectable. However, in the opposite case where replicates are genetically identical (or similarly technical), inter-replicate variability will be low and gene length bias will then become noticeable. The above observations could, thus, probably explain why our simulations, which were created by estimating NB distribution parameters from the real human dataset (which contains different individuals as biological replicates), demonstrate the prior reported behavior. Finally, the fact that almost all previous publications investigating the same topic use technical and/or genetically identical replicates to show existence of gene length bias, further supports our hypothesis (Oshlack and Wakefield 2009; Gao et al. 2011; Mi et al. 2012) and poses a question to whether gene length bias correction is indeed necessary when independent biological entities are examined.

In order to continue research on the topic we focused on one of the *3x replicate design* iterations for which all tools' results were significantly biased. GSEA was chosen as a PA method for consistency reasons with most previous publications (Gao et al. 2011; Mi et al. 2012), while the pre-ranked option was selected because we wanted genes to be ranked according to their p-value, the statistic where PANDORA and the other combination methods do act. At last, according to Kolmogorov-Smirnov test, we suggest that PANDORA as well as DESeq and Simes methods somehow alleviate bias's propagation into pre-ranked GSEA. Yet, these data must be interpreted with caution due to lacking replication.

The next and final step of our investigation focused on the suggestion that lncRNAs are underrepresented during DEA, in contrast for example with mRNAs, due to the formers' small expression levels (Assefa et al. 2018). Initially, we analyzed simulated data of the NGP-nutlin dataset, so as to find that while all tools are indeed underrepresenting lncRNAs in consistence with (Assefa et al. 2018) findings, some of them like PANDORA and/or ABSSeq show superior performance in the great majority of estimated metrics for

both mRNAs and lncRNAs. Notably, PANDORA showed once again an excellent precision-sensitivity tradeoff (F_1-score) and an also very competitive AUFC validating our first simulations' data.

Then, we performed a metaseqR2-supported tools concordance analysis based on the true NGP-nutlin dataset. Again PANDORA, although affected by lncRNA bias along with all other methods, proved itself a very reliable and trustworthy tool to work with (moderate DEG number calling, adequate lncRNA representation, great mean ranking correlation and one of the best mean DEG overlaps with the rest of the methods). Another observation made during concordance analysis was also surprising: lncRNAs mean ranking correlation was for some tools better than that of mRNAs, a phenomenon that became obvious for all tools after p-value correction. A possible explanation for this specific result would be the fact that the analyzed dataset contained less lncRNAs than mRNAs (almost half in numbers) thus augmenting the probabilities of the tools to achieve more similar lncRNA rankings with one another.

Hierarchical clustering was applied to summarize concordance analysis results. In particular, DESeq2 systematic clustering with PANDORA, edgeR, voom, ABSSeq and/or NBPSeq is partially consistent with the observations made in (Assefa et al. 2018), where DESeq2, edgeR and voom were again found in the same dendrogram's group. Particularly, PANDORA's clustering behavior can be attributed to the fact that top performing tools such as the aforementioned four are generally assigned big weights that "drift" PANDORA with them during the clustering process.

Summarizing all the above, metaseqR2 is an up-to-date, powerful tool for DEA of RNA-seq data. From the fifteen different statistical analysis methods included, PANDORA performs collectively better than most others and demonstrates a more robust performance, too. In addition, it must be noted that by using weights from all metaseqR2-implemented individual statistical analysis tools, PANDORA performs if not better (AUFC, aFDR, F_1-score with adjusted p-values) at least as good as with the weights of its previous environment of implementation (Moulos and Hatzis 2015). Furthermore, when RNA-seq biases like gene length and lncRNA bias are present, as well as when different normalization methods are applied, PANDORA is again one of the most reliable options to choose. Hence, it could conceivably be hypothesized that PANDORA is also beneficial for controlling read count and low count biases as well, since they are practically no different than the here examined gene length and lncRNA biases, respectively (Young et al. 2010; Assefa et al. 2018). At the end, as meta-analysis approaches to DEA of RNA-seq data seem most promising, further studies on the field are recommended so as to obtain an "enhanced resolution picture" of the true differences in gene expression patterns.

Taken together, all computational experiments conducted in the current MSc thesis points towards PANDORA as the best choice for DEA of RNA-seq data, not only because of its overall very good performance, but also due to its robustness and reliability. PANDORA is implemented in metaseqR2 package, which also offers access to many distinct analysis tools and to a user-friendly, self-explanatory report.

5. REFERENCES

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Assefa, Alemu Takele, Katrijn De Paepe, Pieter Mestdagh, Olivier Thas, Celine Everaert, and Jo Vandesompele. 2018. "Differential Gene Expression Analysis Tools Exhibit Substandard Performance for Long Non-Coding RNA-Sequencing Data." *Genome Biology* 19 (1): 1–16. https://doi.org/10.1186/s13059-018-1466-5.

Baggerly, Keith A., Li Deng, Jeffrey S. Morris, and C. Marcelo Aldaz. 2004. "Overdispersed Logistic Regression for SAGE: Modelling Multiple Groups and Covariates." *BMC Bioinformatics* 5: 1–16. https://doi.org/10.1186/1471-2105-5-144.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Bloom, Joshua S., Zia Khan, Leonid Kruglyak, Mona Singh, and Amy A. Caudy. 2009. "Measuring Differential Gene Expression by Short Read Sequencing: Quantitative Comparison to 2-Channel Gene Expression Microarrays." *BMC Genomics* 10. https://doi.org/10.1186/1471-2164-10-221.

Bullard, James H., Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. 2010. "Evaluation of Statistical Methods for Normalization and Differential Expression in MRNA-Seq Experiments." *BMC Bioinformatics* 11. https://doi.org/10.1186/1471-2105-11-94.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (1): 1–19. https://doi.org/10.1186/s13059-016-0881-8.

Di, Yanming, Daniel W Schafer, Jason S Cumbie, and Jeff H Chang. 2011. "The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq." *Statistical Applications in Genetics and Molecular Biology* 10 (1). https://doi.org/10.2202/1544-6115.1637.

Dillies, M.-A., G. Marot, L. Jouneau, C. Le Gall, J. Estelle, C. Keime, N. Servant, et al. 2012. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in Bioinformatics* 14 (6): 671–83. https://doi.org/10.1093/bib/bbs046.

Edgren, Henrik, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H Rye, Sandra Nyberg, Maija Wolf, and Olli Kallioniemi. 2011. "Identification of Fusion Genes in Breast Cancer By." *Genome Biology* 12: 1–13. https://doi.org/10.1186/gb-2011-12-1-r6.

Everaert, Celine, Manuel Luypaert, Jesper L.V. Maag, Quek Xiu Cheng, Marcel E. DInger, Jan Hellemans, and Pieter Mestdagh. 2017. "Benchmarking of RNA-Sequencing Analysis Workflows Using Whole-Transcriptome RT-QPCR Expression Data." *Scientific Reports* 7 (1): 1–11. https://doi.org/10.1038/s41598-017-01617-3.

Franck, Rapaport, Khanin Raya, Liang Yupu, Pirun Mono, Krek Azra, Zumbo Paul, E Mason Christopher, D Socci Nicholas, and Betel Doron. 2013. "Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data." *Genome Biology* 14 (9): R95. https://doi.org/10.1186/gb-2013-14-9-r95.

Gao, Liyan, Zhide Fang, Kui Zhang, Degui Zhi, and Xiangqin Cui. 2011. "Length Bias Correction for RNA-Seq Data in Gene Set Analyses." *Bioinformatics* 27 (5): 662–69. https://doi.org/10.1093/bioinformatics/btr005.

García-Campos, Miguel A., Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. 2015. "Pathway Analysis: State of the Art." *Frontiers in Physiology* 6 (DEC): 1–16. https://doi.org/10.3389/fphys.2015.00383.

Gentleman, Robert C, Vincent J Carey, Vincent J Carey, Douglas M Bates, Douglas M Bates, Ben Bolstad, Ben Bolstad, et al. 2004. "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology* 5 (10): R80. https://doi.org/10.1186/gb-2004-5-10-r80.

Han, Yixing, Shouguo Gao, Kathrin Muegge, Wei Zhang, and Bing Zhou. 2015. "Advanced Applications of RNA Sequencing and Challenges." *Bioinformatics and Biology Insights* 9: 29–46. https://doi.org/10.4137/BBI.S28991.

Hansen, Kasper D., Rafael A. Irizarry, and Zhijin Wu. 2012. "Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization." *Biostatistics* 13 (2): 204–16. https://doi.org/10.1093/biostatistics/kxr054.

Hardcastle, TJ, and KA Kelly. 2009. "Empirical Bayesian Methods for Differential Expression in Count Data." *BMC Bioinformatics* 11: 422. https://doi.org/1471-2105-11-422 [pii] 10.1186/1471-2105-11-422.

Heyne, Henrike O., Susann Lautenschläger, Ronald Nelson, François Besnier, maxime Rotival, Alexander Cagan, Rimma Kozhemyakina, et al. 2014. "Genetic Influences on Brain Gene Expression in Rats Selected for Tameness and Aggression." *Genetics* 198 (3): 1277–90. https://doi.org/10.1534/genetics.114.168948.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nature Methods* 12 (4): 357–60. https://doi.org/10.1038/nmeth.3317.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): R29. https://doi.org/10.1186/gb-2014-15-2-r29.

Lister, Ryan, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. 2008. "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis." *Cell* 133 (3): 523–36. https://doi.org/10.1016/j.cell.2008.03.029.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 1–21. https://doi.org/10.1186/s13059-014-0550-8.

Lu, Jun, John K. Tomfohr, and Thomas B. Kepler. 2005. "Identifying Differential Expression in Multiple SAGE Libraries: An Overdispersed Log-Linear Model Approach." *BMC Bioinformatics* 6: 1–14. https://doi.org/10.1186/1471-2105-6-165.

Marioni, John C., Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. 2008. "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Research* 18 (9): 1509–17. https://doi.org/10.1101/gr.079558.108.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97. https://doi.org/10.1093/nar/gks042.

Mi, Gu, Yanming Di, Sarah Emerson, Jason S. Cumbie, and Jeff H. Chang. 2012. "Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression." *PLoS ONE* 7 (10). https://doi.org/10.1371/journal.pone.0046128.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Moulos, Panagiotis, and Pantelis Hatzis. 2015. "Systematic Integration of RNA-Seq Statistical Algorithms for Accurate Detection of Differential Gene Expression Patterns." *Nucleic Acids Research* 43 (4): 1–12. https://doi.org/10.1093/nar/gku1273.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." *Science* 320 (5881): 1344–49. https://doi.org/10.1126/science.1158441.

Oshlack, Alicia, and Matthew J. Wakefield. 2009. "Transcript Length Bias in RNA-Seq Data Confounds Systems Biology." *Biology Direct* 4: 1–10. https://doi.org/10.1186/1745-6150-4-14.

Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. "Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing." *Nature* 464 (7289): 768–72. https://doi.org/10.1038/nature08872.

Quail, Michael A., Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." *BMC Genomics* 13 (1). https://doi.org/10.1186/1471-2164-13-341.

Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content Normalization for RNA-Seq Data." *BMC Bioinformatics* 12 (1): 480. https://doi.org/10.1186/1471-2105-12-480.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Roberts, Adam, Harold Pimentel, Cole Trapnell, and Lior Pachter. 2011. "Identification of Novel Transcripts in Annotated Genomes Using RNA-Seq." *Bioinformatics* 27 (17): 2325–29. https://doi.org/10.1093/bioinformatics/btr355.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3). https://doi.org/10.1186/gb-2010-11-3-r25.

Robinson, Mark D., and Gordon K. Smyth. 2007. "Moderated Statistical Tests for Assessing Differences in Tag Abundance." *Bioinformatics* 23 (21): 2881–87. https://doi.org/10.1093/bioinformatics/btm453.

Robinson, Mark D., and Smyth GK. 2008. "Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data." *Biostatistics* 9 (2): 321–32. https://doi.org/10.1093/biostatistics/kxm030.

Rödel, E. 2007. "Fisher, R. A.: Statistical Methods for Research Workers, 14. Aufl., Oliver & Boyd, Edinburgh, London 1970. XIII, 362 S., 12 Abb., 74 Tab., 40 S." *Biometrische Zeitschrift* 13 (6): 429–30. https://doi.org/10.1002/bimj.19710130623.

Sergushichev, Alexey. 2016. "An Algorithm for Fast Preranked Gene Set Enrichment Analysis Using Cumulative Statistic Calculation." *BioRxiv*, 60012. https://doi.org/10.1101/060012.

Simes, R J. 1986. "An Improved Bonferroni Procedure for Multiple Tests of Significance." *Biometrika* 73 (3): 751–54. https://doi.org/10.1093/biomet/73.3.751.

Soneson, Charlotte, and Mauro Delorenzi. 2013. "A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data." *BMC Bioinformatics* 14 (91): 1471–2015. https://doi.org/10.1038/nmeth.1226.

Sonia Tarazona, and Fernando Garc a-Alcalde. 2011. "Differential Expression in RNA-Seq: A Matter of Depth." *Genome Research*, 2213–23. https://doi.org/10.1101/gr.124321.111.Freely.

Tarazona, Sonia, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. 2015. "Data Quality Aware Analysis of Differential Expression in RNA-Seq with NOISeq R/Bioc Package." *Nucleic Acids Research* 43 (21). https://doi.org/10.1093/nar/gkv711.

Tarazona, Sonia, Fernando Garcia-Alcalde, Joaquin Dopazo, Alberto Ferrer, and Ana Conesa. 2011. "Differential Expression in RNA-Seq: A Matter of Depth." *Genome Research*, 2213–23. https://doi.org/10.1101/gr.124321.111.Freely.

Tuch, Brian B., Rebecca R. Laborde, Xing Xu, Jian Gu, Christina B. Chung, Cinna K. Monighetti, Sarah J. Stanley, et al. 2010. "Tumor Transcriptome Sequencing Reveals Allelic Expression Imbalances Associated with Copy Number Alterations." *PLoS ONE* 5 (2). https://doi.org/10.1371/journal.pone.0009317.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. "Alternative Isoform Regulation in Human Tissue Transcriptomes." *Nature* 456 (7221): 470–76. https://doi.org/10.1038/nature07509.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Whitlock, M C. 2005. "Combining Probability from Independent Tests: The Weighted Z-Method Is Superior to Fisher's Approach." *Journal of Evolutionary Biology* 18 (5): 1368–73. https://doi.org/10.1111/j.1420-9101.2005.00917.x.

Wu, Hao, Chi Wang, and Zhijin Wu. 2013. "A New Shrinkage Estimator for Dispersion Improves Differential Expression Detection in RNA-Seq Data." *Biostatistics* 14 (2): 232–43. https://doi.org/10.1093/biostatistics/kxs033.

Yang, Wentao, Philip C. Rosenstiel, and Hinrich Schulenburg. 2016. "ABSSeq: A New RNA-Seq Analysis Method Based on Modelling Absolute Expression Differences." *BMC Genomics* 17 (1): 1–14. https://doi.org/10.1186/s12864-016-2848-2.

Yoon, Sora, and Dougu Nam. 2017. "Gene Dispersion Is the Key Determinant of the Read Count Bias in Differential Expression Analysis of RNA-Seq Data." *BMC Genomics* 18 (1): 1–11. https://doi.org/10.1186/s12864-017-3809-0.

Young, Matthew D, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. 2010. "Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias." *Genome Biology*. https://doi.org/10.1186/gb-2010-11-2-r14.

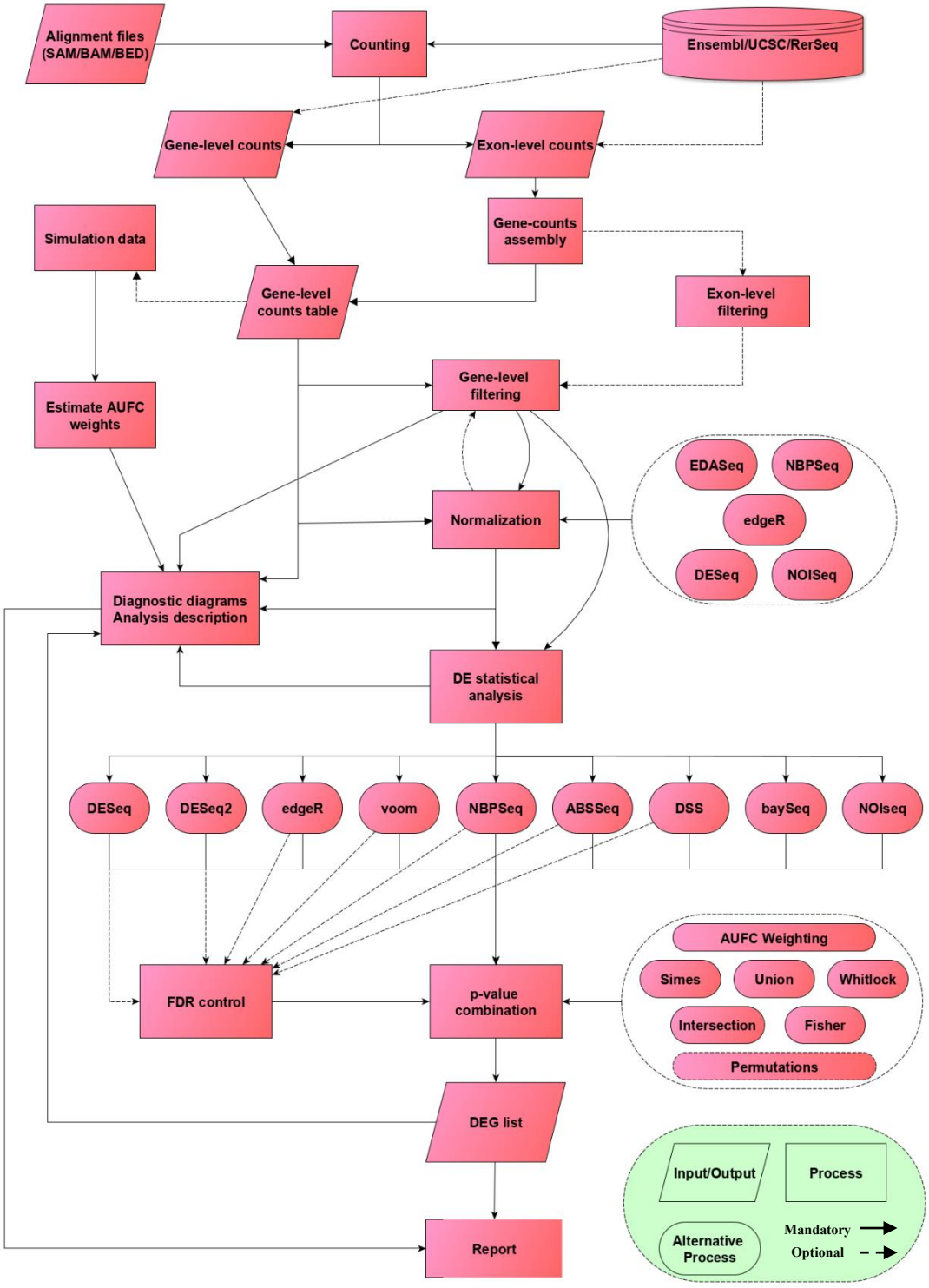
Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters." *OMICS: A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Zhang, Zhe, Yuanchao Zhang, Perry Evans, Asif Chinwalla, and Deanne Taylor. 2017. "RNA-Seq 2G: Online Analysis Of Differential Gene Expression With Comprehensive Options Of Statistical Methods." *BioRxiv*, no. 6: 122747. https://doi.org/10.1101/122747.

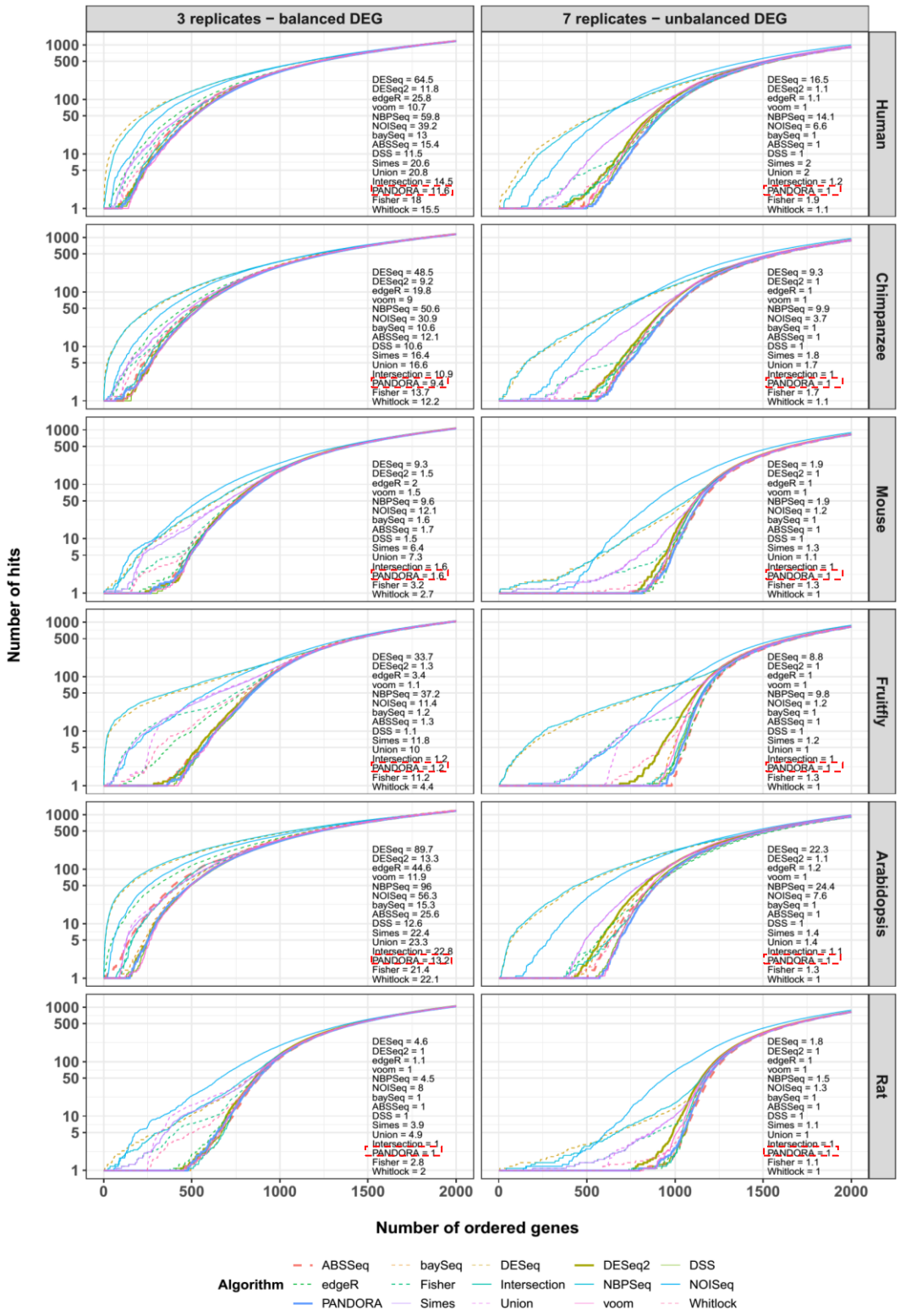
6. APPENDIX I

Simulated and real data evaluation supplementary figures using EDASeq normalization.

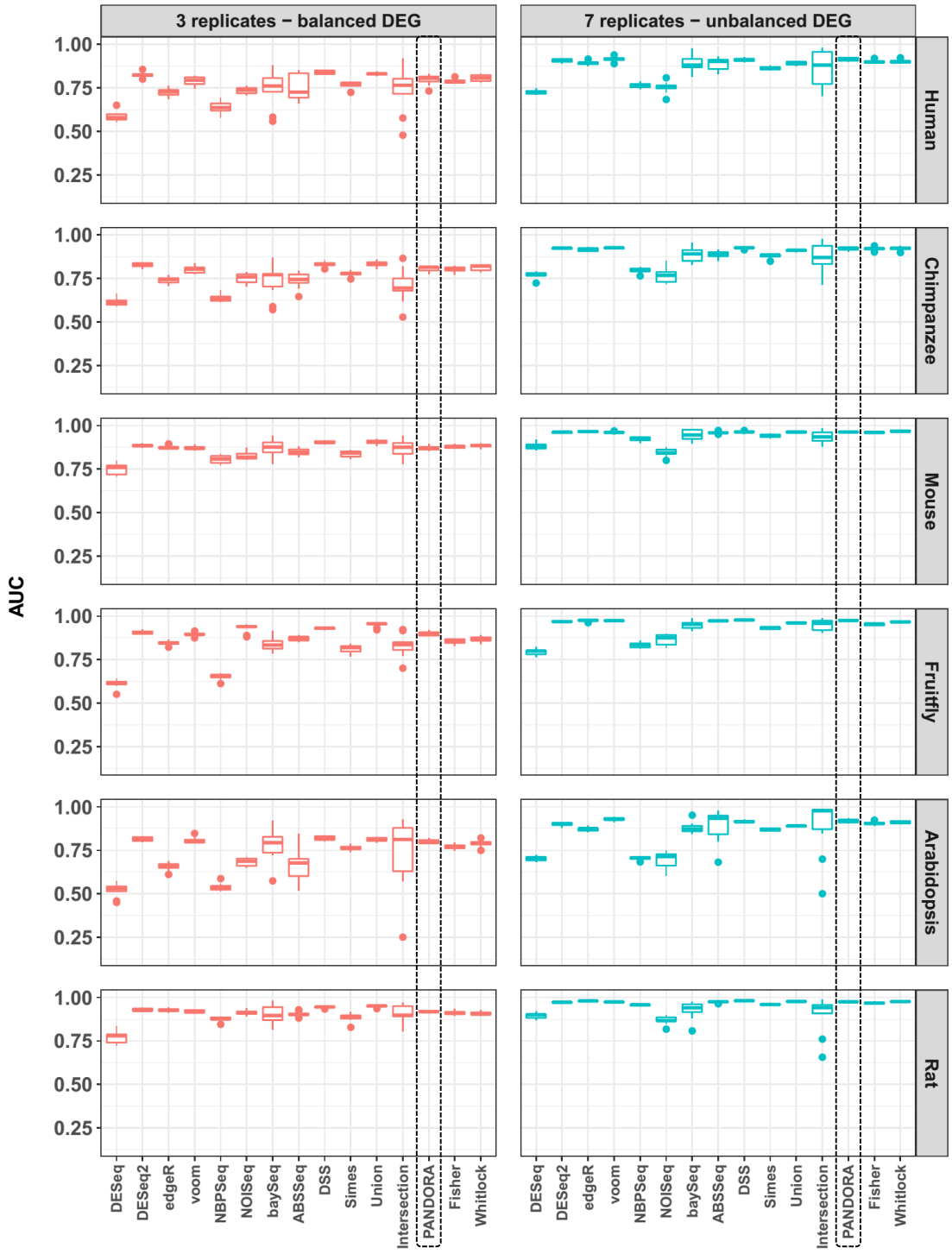
Note: In all BH p-value adjustments NOISeq is presented with its unadjusted p-value, mainly for reasons of completeness.



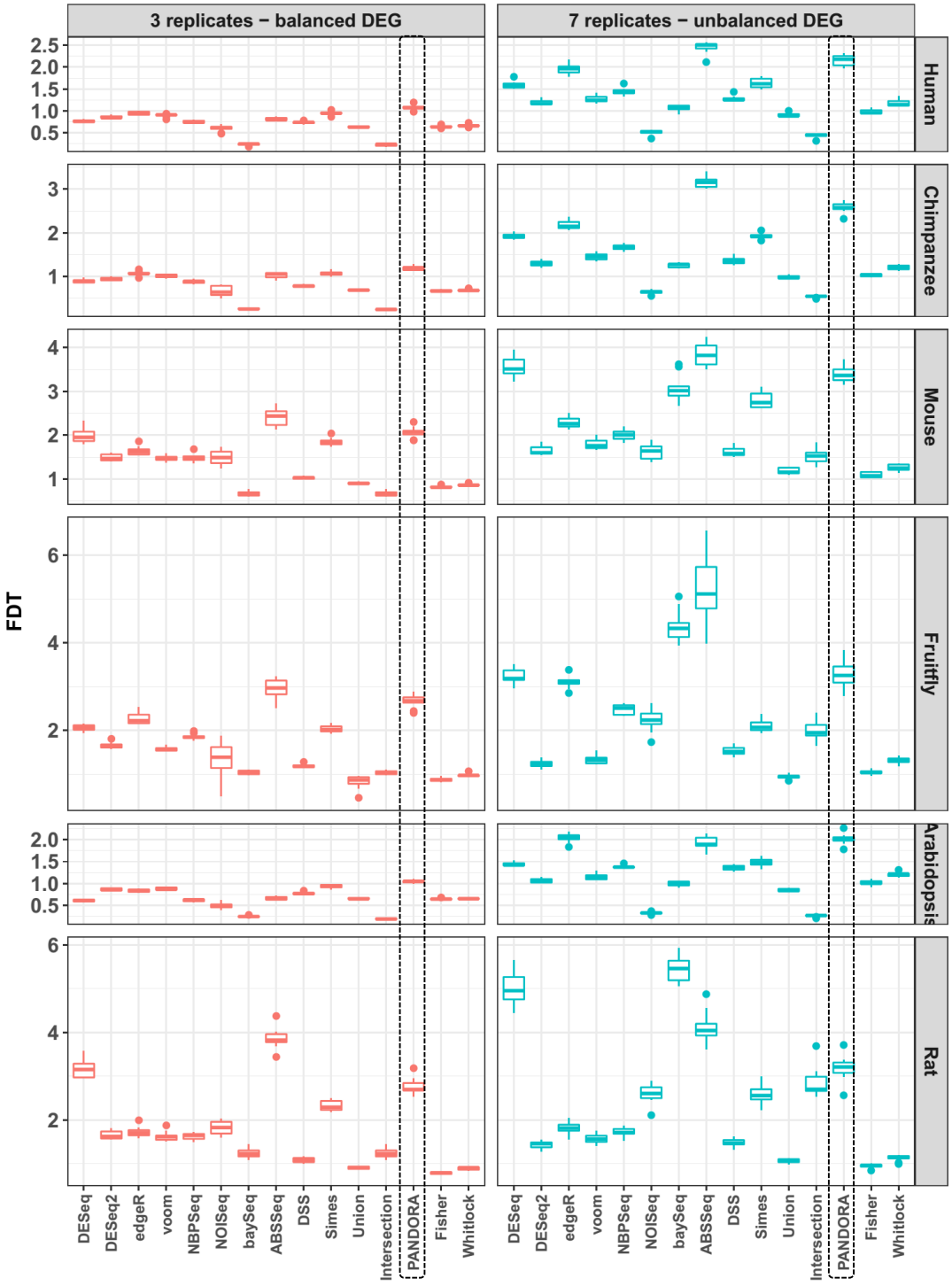
Supplementary Figure 1: metaseqR2 workflow.



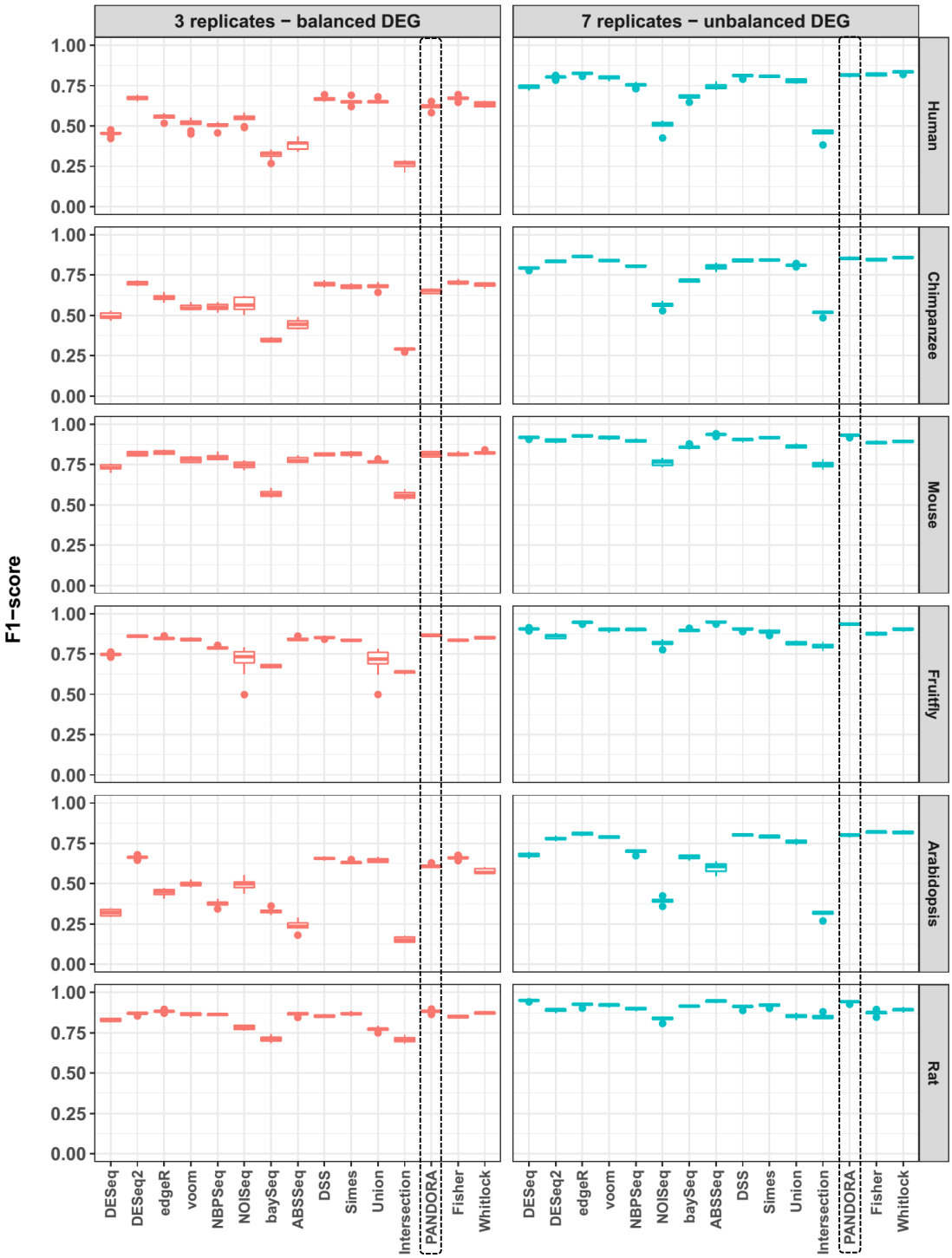
Supplementary Figure 2: False Discovery Curves (FDC) usign adjusted p-values after EDASeq normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. Trends remain the same after p-value adjustment. (BH p-value threshold: 0.05, EDASeq normalization)



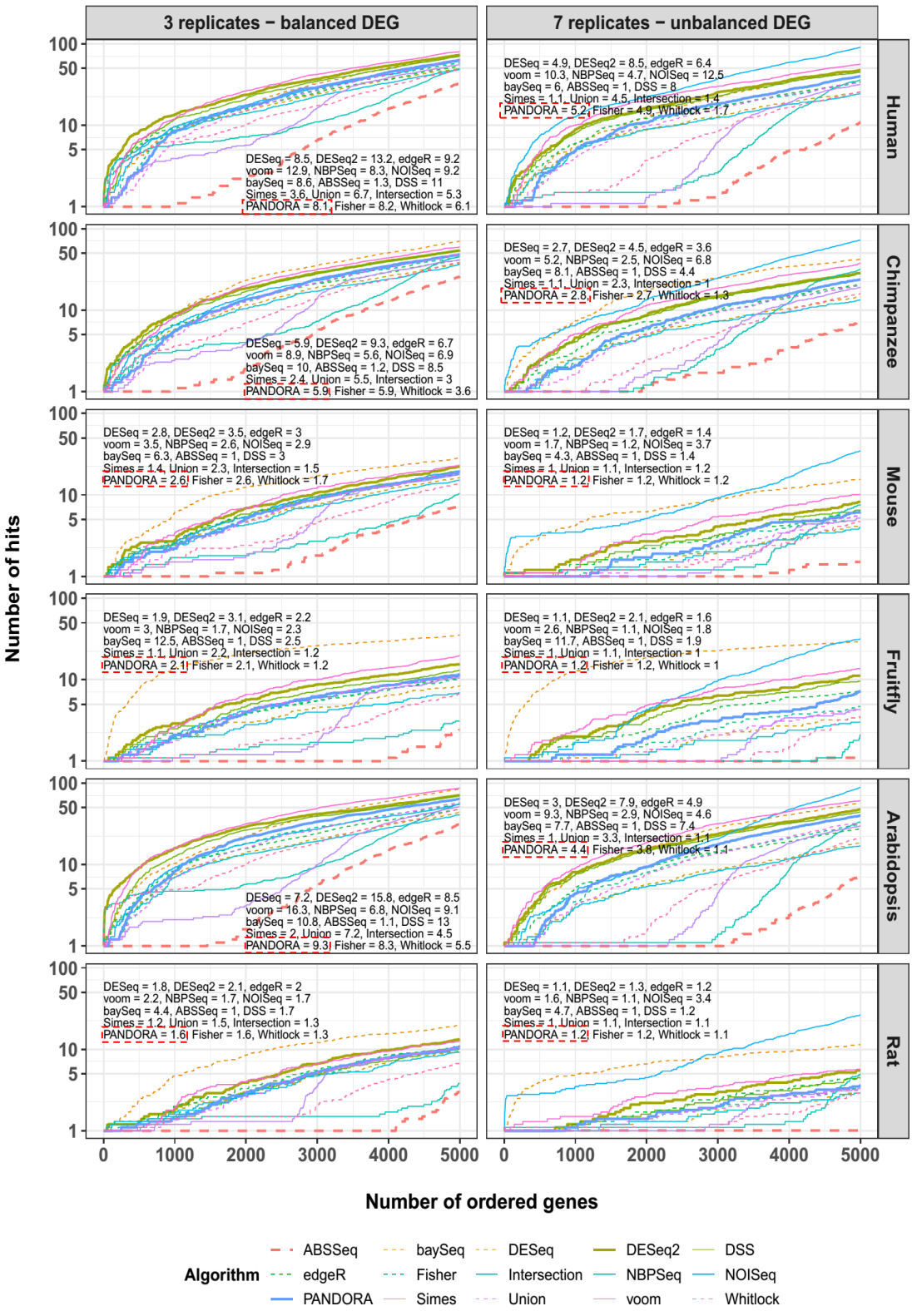
Supplementary Figure 3: Area under the ROC curve (AUC) using adjusted p-values and EDASeq normalization. AUC are summarized across ten iterations for each tool and simulation design, using adjusted p-values. Trends remain the same after p-value adjustment. (BH p-value threshold: 0.05, EDASeq normalization)



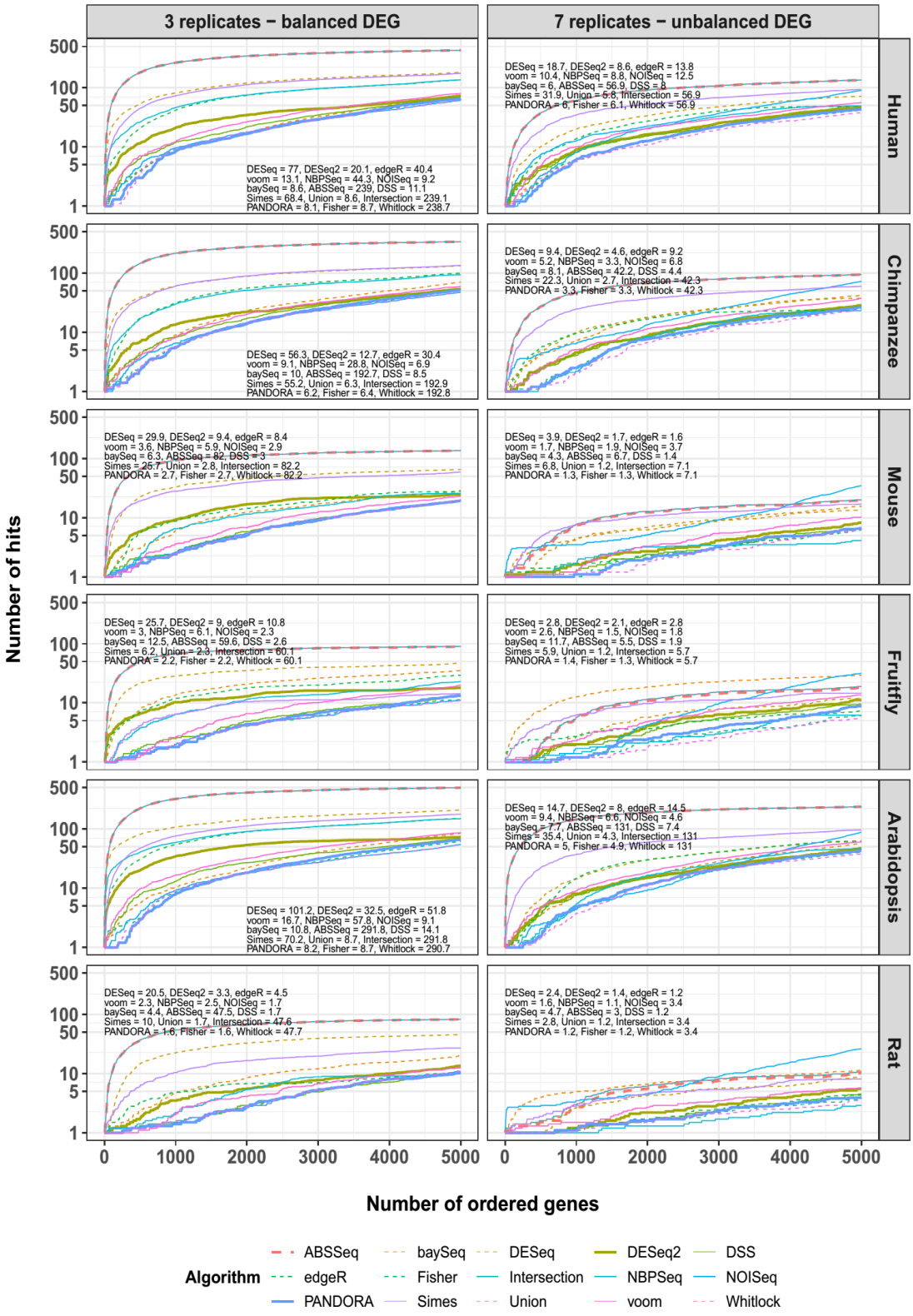
Supplementary Figure 4: False Discovery Tradeoff (FDT) using raw p-values after EDASeq normalization. FDT is summarized across ten iterations for each tool and simulation design using unadjusted p-values. FDT analysis replicates PANDORA's ability to favorably control TP versus FP and FN numbers. (statistical significant threshold: 0.05, EDASeq normalization)



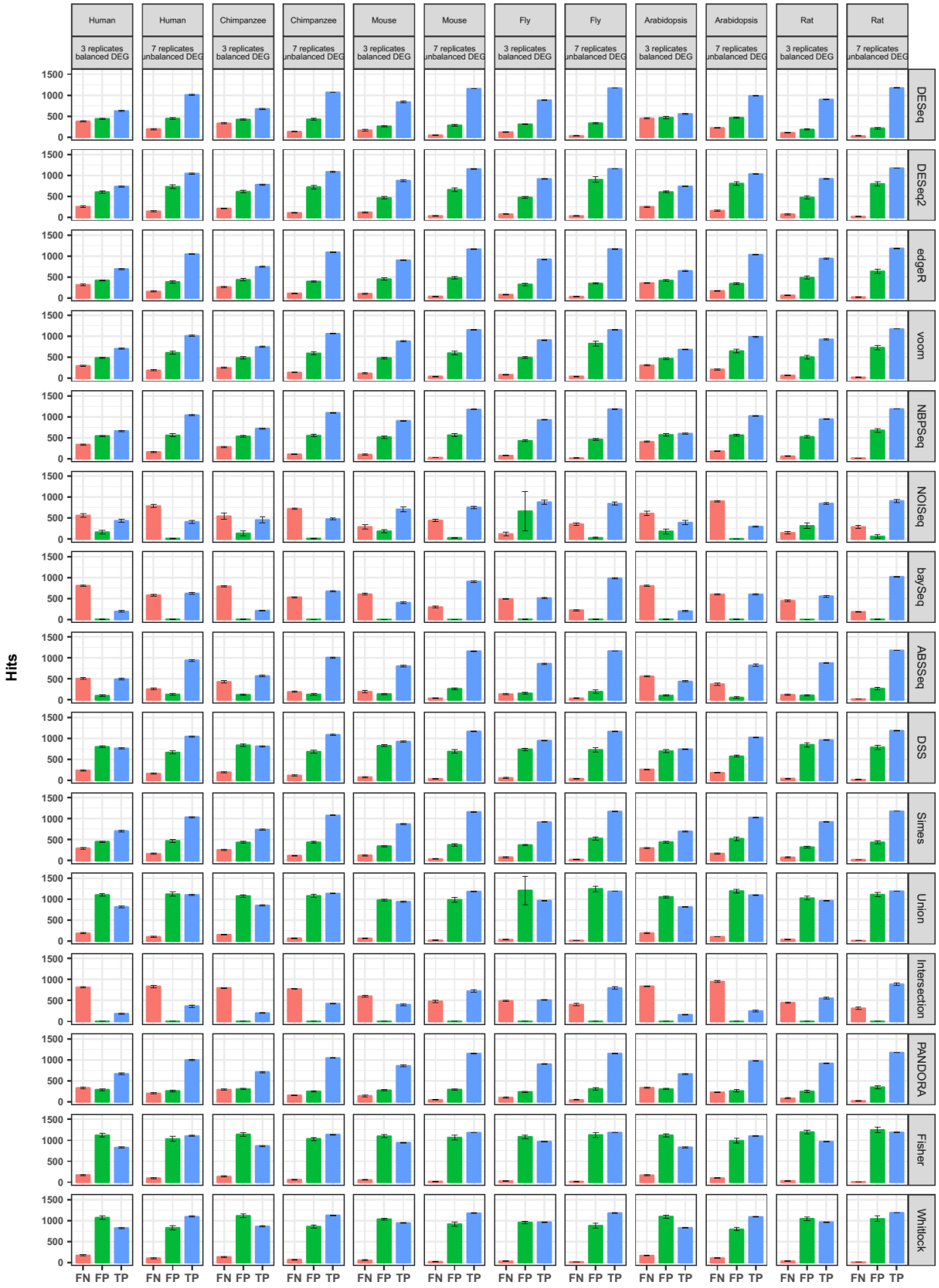
Supplementary Figure 5: F_1-score (precision-sensitivity tradeoff) using adjusted p-values after EDASeq normalization. F_1-score is summarized across ten simulations for each tool and simulation design, using adjusted p-values. PANDORA demonstrates a robust F1-score both before (Figure 4) and after p-value adjustment. (BH p-value threshold: 0.05, EDASeq normalization)



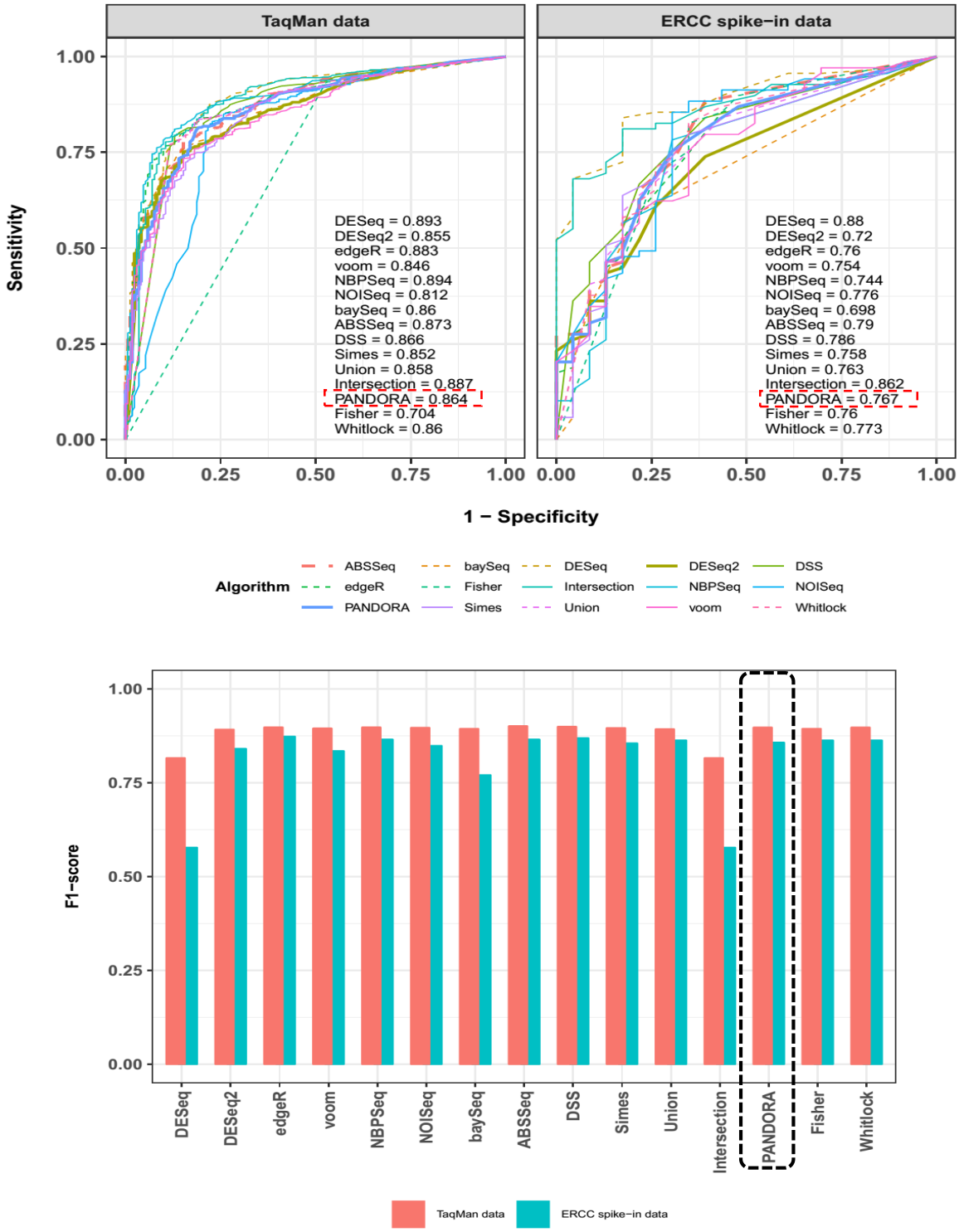
Supplementary Figure 6: False Negative Curves (FNC) using raw p-values after EDASeq normalization. FNCs are summarized across ten iterations for each tool and simulation design examining the last 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the top left or bottom right corner of each plot. Tools like voom that successfully control FPs (Figure 1) cannot control FNs at the same time. PANDORA is one of the few exceptions that can simultaneously control both FPs and FNs. (significant p-value threshold: 0.05, EDASeq normalization)



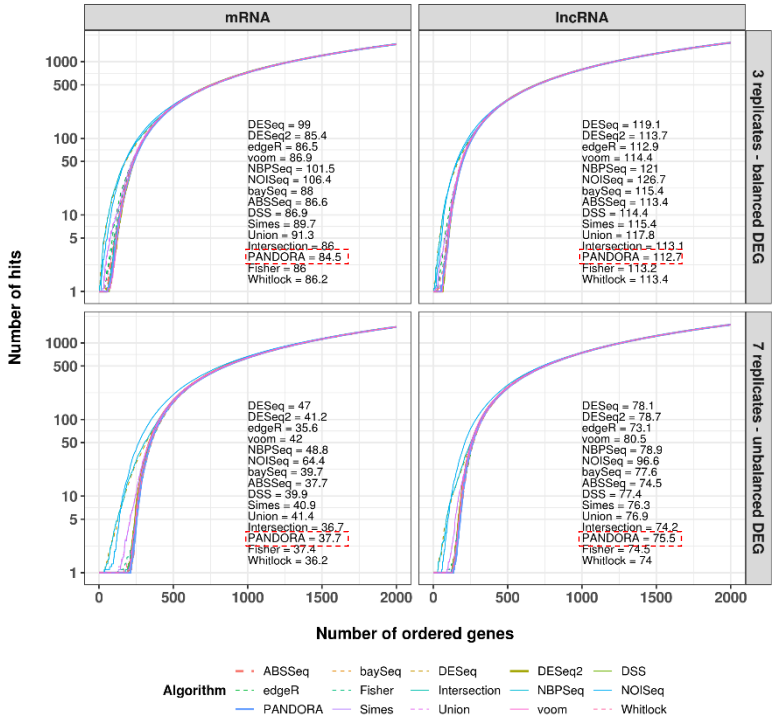
Supplementary Figure 7: False Negative Curves (FNC) using adjusted p-values after EDASeq normalization. FNCs are summarized across ten iterations for each tool and simulation design examining the last 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the top left or bottom right corner of each plot. PANDORA performance is much better when using adjusted p-values. (BH p-value threshold: 0.05, EDASeq normalization)



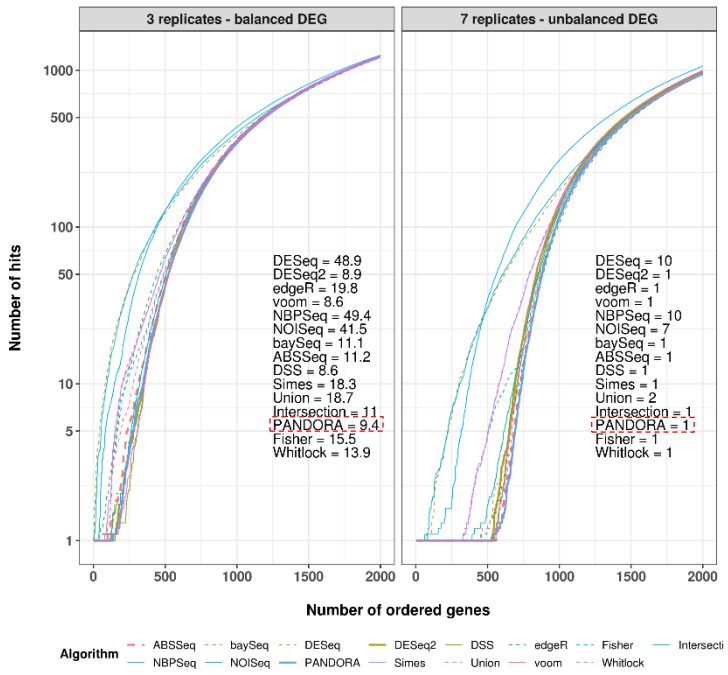
Supplementary Figure 8: FN, FP and TP hits for all datasets, simulated replicate designs and statistical analysis methods using unadjusted p-values. (significant p-value threshold: 0.05, EDASeq normalization)



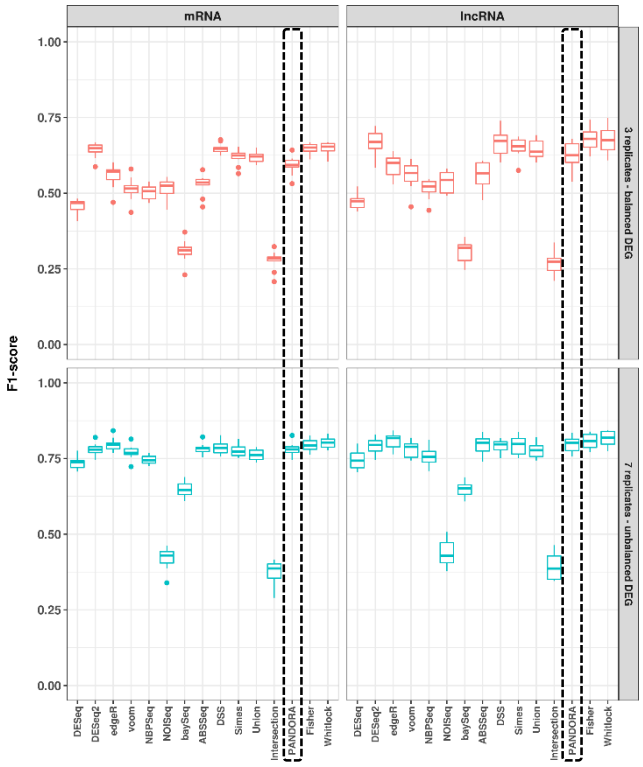
Supplementary Figure 9: ROC and F1-score analysis of real datasets using adjusted p-values after EDASeq (top and bottom respectively). AUC can be found at the bottom right corner of the ROC figures. Trends are the same as before normalization for both ROC and F1-score analysis. DESeq's, baySeq's and Intersection's tools F1-score for the ERCC dataset is even more compromised than with the unadjusted p-values. (BH p-value threshold: 0.05, EDASeq normalization)normalization



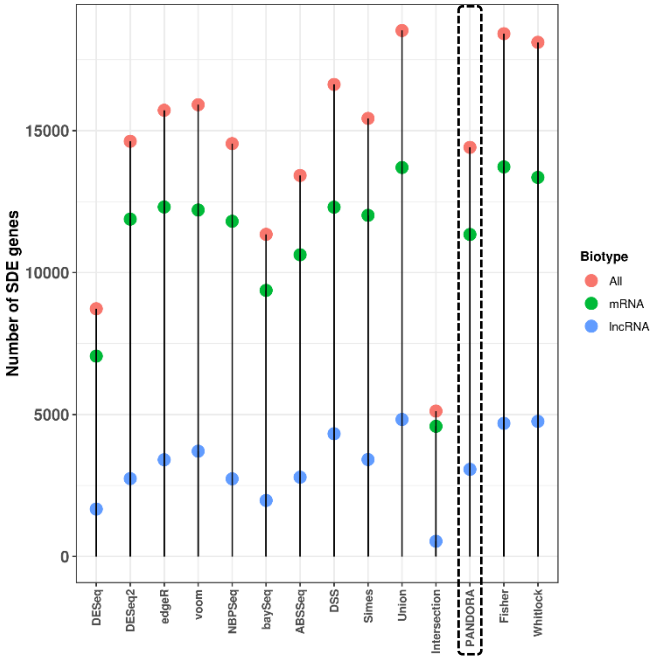
Supplementary Figure 10: False Discovery Curves (FDC) using adjusted p-values after EDASeq normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. p-value adjustment has no actual effect on AUFC of both biotypes and simulation configurations. (BH p-value threshold: 0.05, EDASeq normalization)



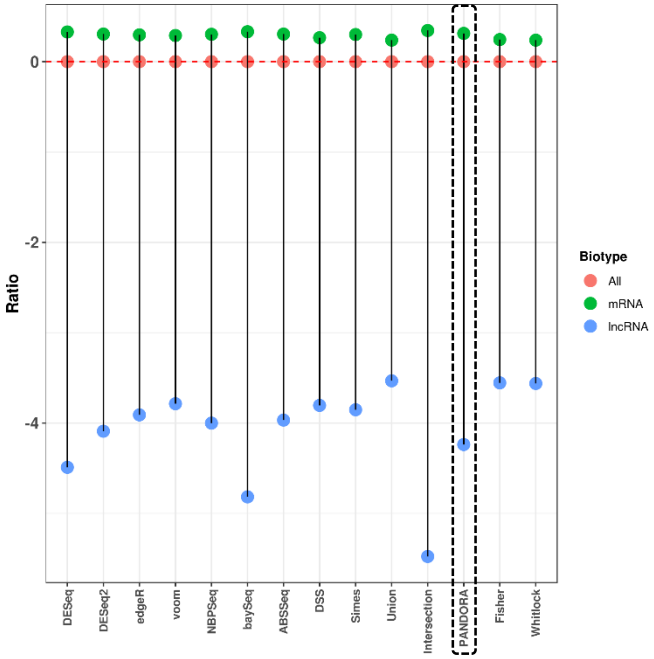
Supplementary Figure 11: False Discovery Curves (FDC) using adjusted p-values of all simulated genes after EDASeq normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. When data are calculated for all simulated genes, AUFC values are of the same magnitude as in Figure 1. (BH p-value threshold: 0.05, EDASeq normalization)



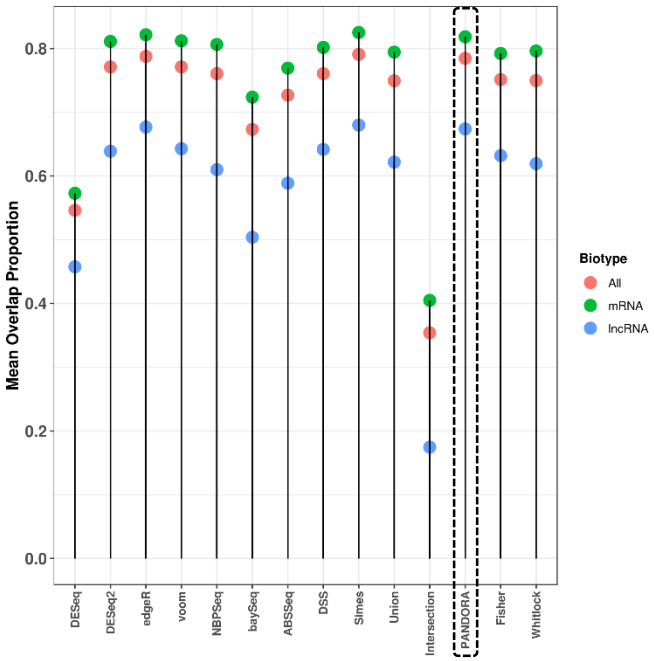
Supplementary Figure 12: F1-score (precision-sensitivity tradeoff) using adjusted p-values after EDASeq normalization. F_1-score is summarized across ten iterations for each tool and simulation design, using adjusted p-values. PANDORA behaves robustly in most cases relative to its score using unadjusted p-values (Figure 12). (BH p-value threshold: 0.05, EDASeq normalization)



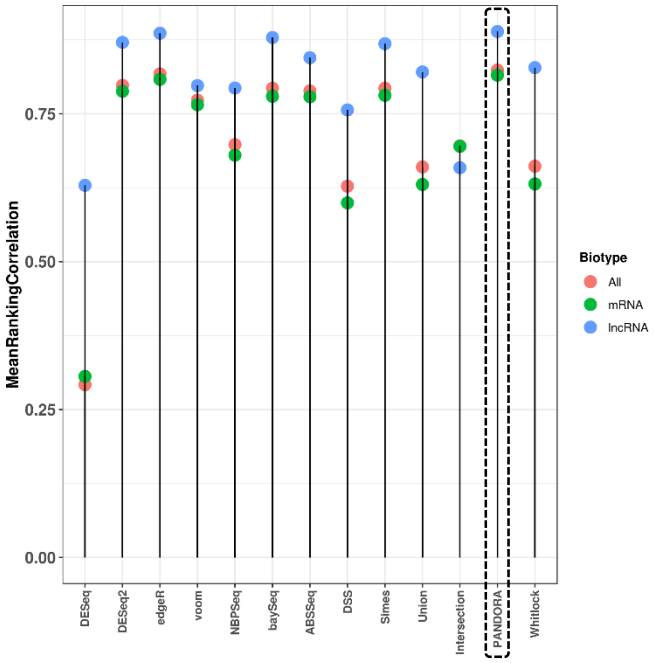
Supplementary Figure 13: Number of DEGs per tool and biotype, using adjusted p-values after EDASeq normalization. PANDORA is neither too lose nor too strict as a DEA tool, for all three biotype designs (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; BH p-value threshold: 0.05, EDASeq normalization)



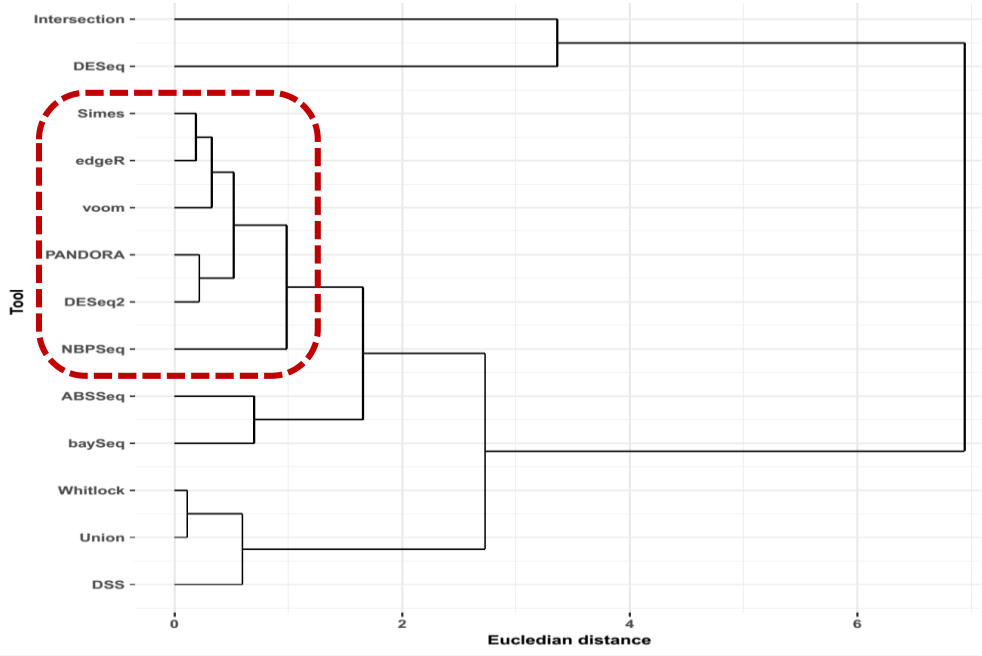
Supplementary Figure 14: Biotype representativeness ratio, using adjusted p-values after EDASeq normalization. While mRNAs are slightly over-represented in all tools final DEG list, lncRNAs are heavily under-represented in all cases. All tools performance regarding lncRNAs is compromised following p-value adjustment. (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; BH p-value threshold: 0.05, EDASeq normalization)



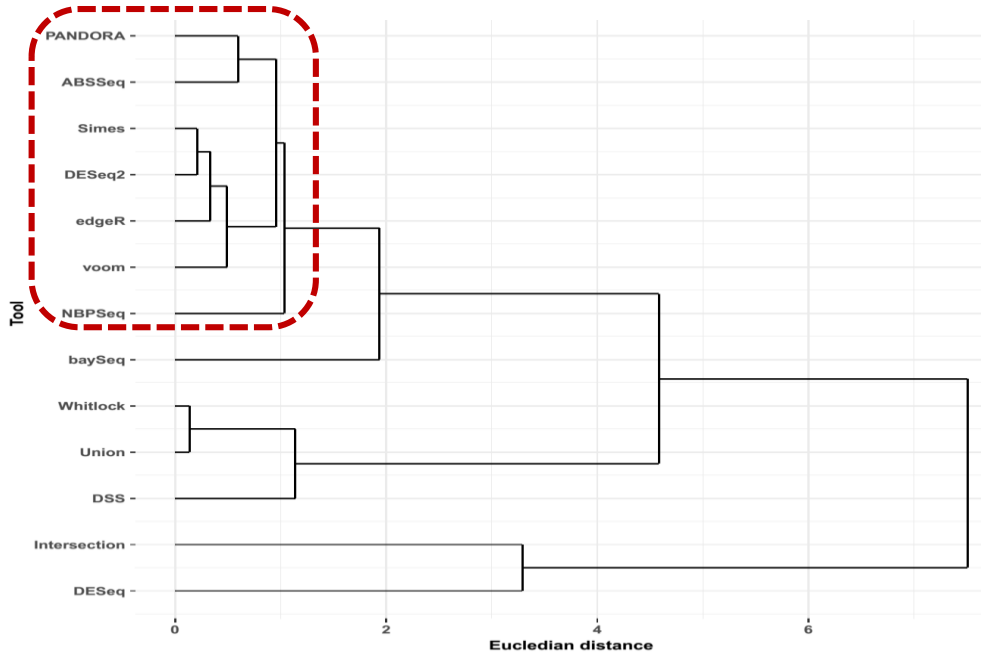
Supplementary Figure 15: Mean overlap proportion of DEGs for each tool and biotype scheme, using adjusted p-values and EDASeq normalization. Trends are the same as those before p-value correction. (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; BH p-value threshold: 0.05, EDASeq normalization)



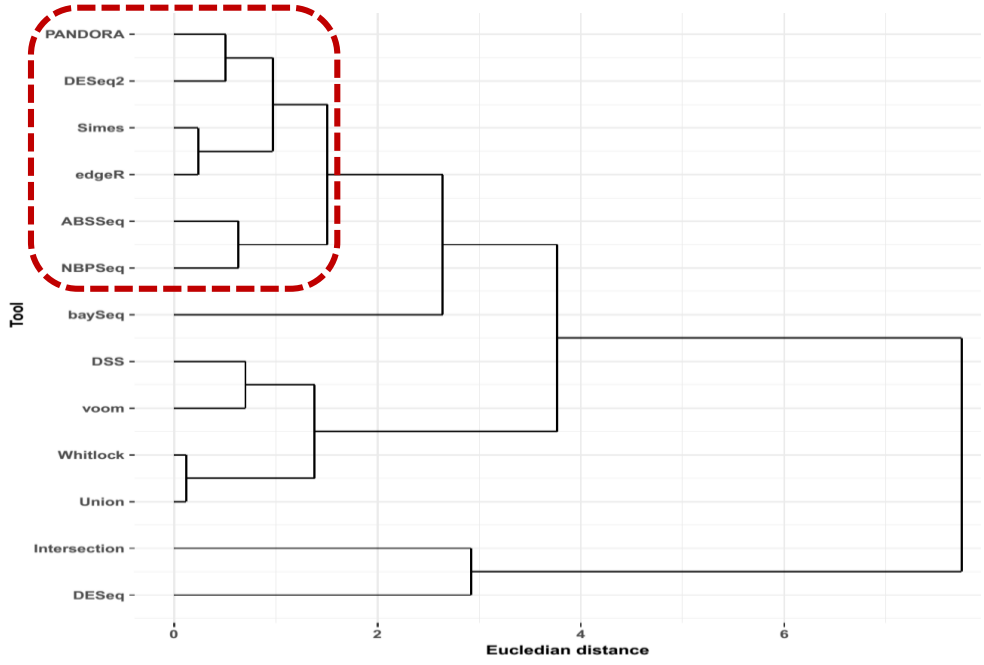
Supplementary Figure 16: Mean DEG ranking correlation between DEA tools for all biotype scheme, using adjusted p-values after EDASeq normalization. While general trends remain the same, lncRNAs ranking becomes better than that of mRNAs after p-value adjustment. NOISeq is not included for reasons explained in (Soneson and Delorenzi 2013) and Fisher due to the big number of ties reported (mRNAs are represented by green, lncRNAs by blue and both biotypes by orange dots; BH p-value threshold: 0.05, EDASeq normalization)



Supplementary Figure 17: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. Both mRNA and lncRNA biotype scheme is here examined. NOISeq and Fisher are not included for reasons explained in Sup. Figure 14 legend.
(BH p-value threshold: 0.05, EDASeq normalization)



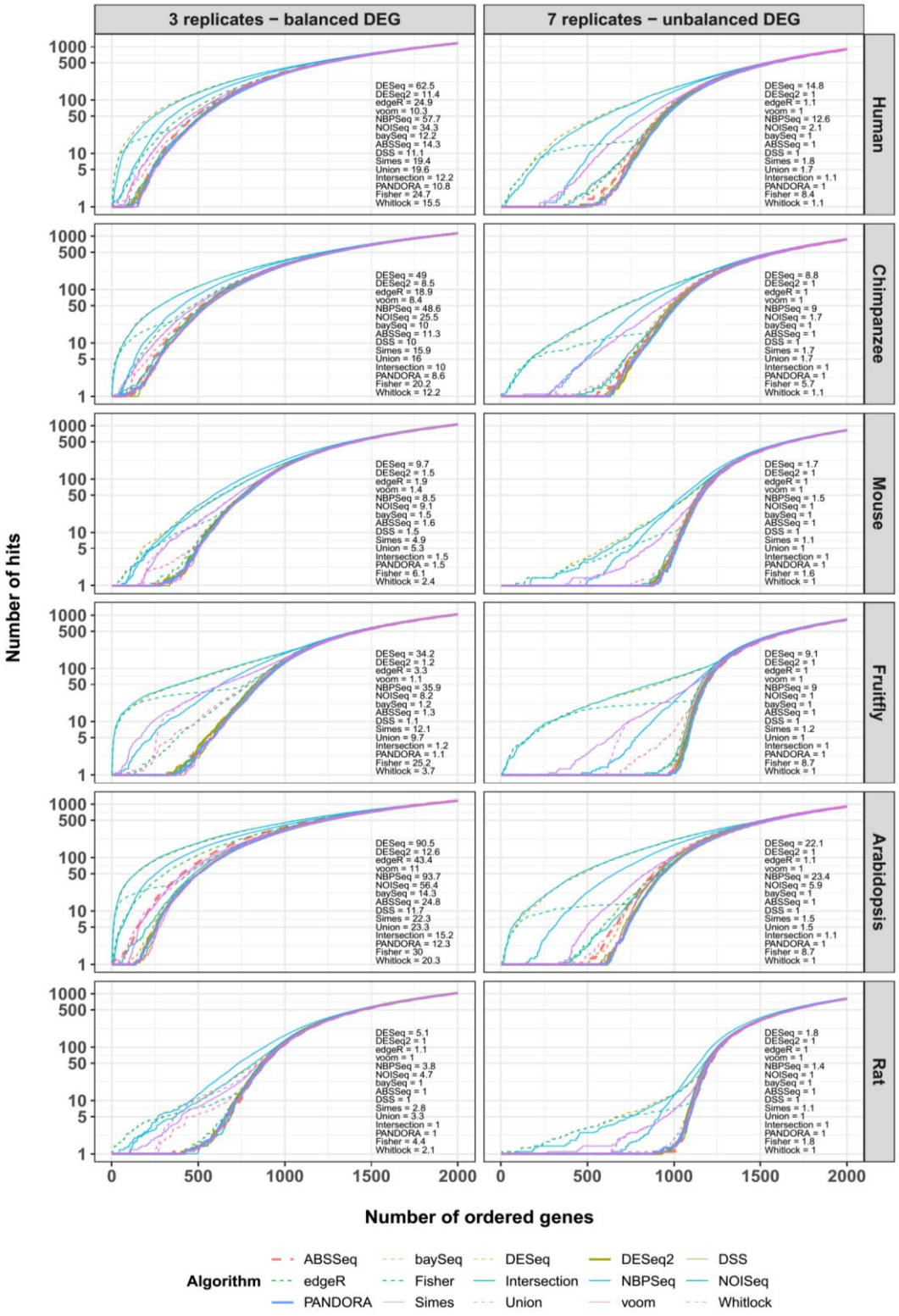
Supplementary Figure 18: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. mRNA biotype scheme is here examined. NOISeq and Fisher are not included for reasons explained in Sup. Figure 14 legend.
(BH p-value threshold: 0.05, EDASeq normalization)



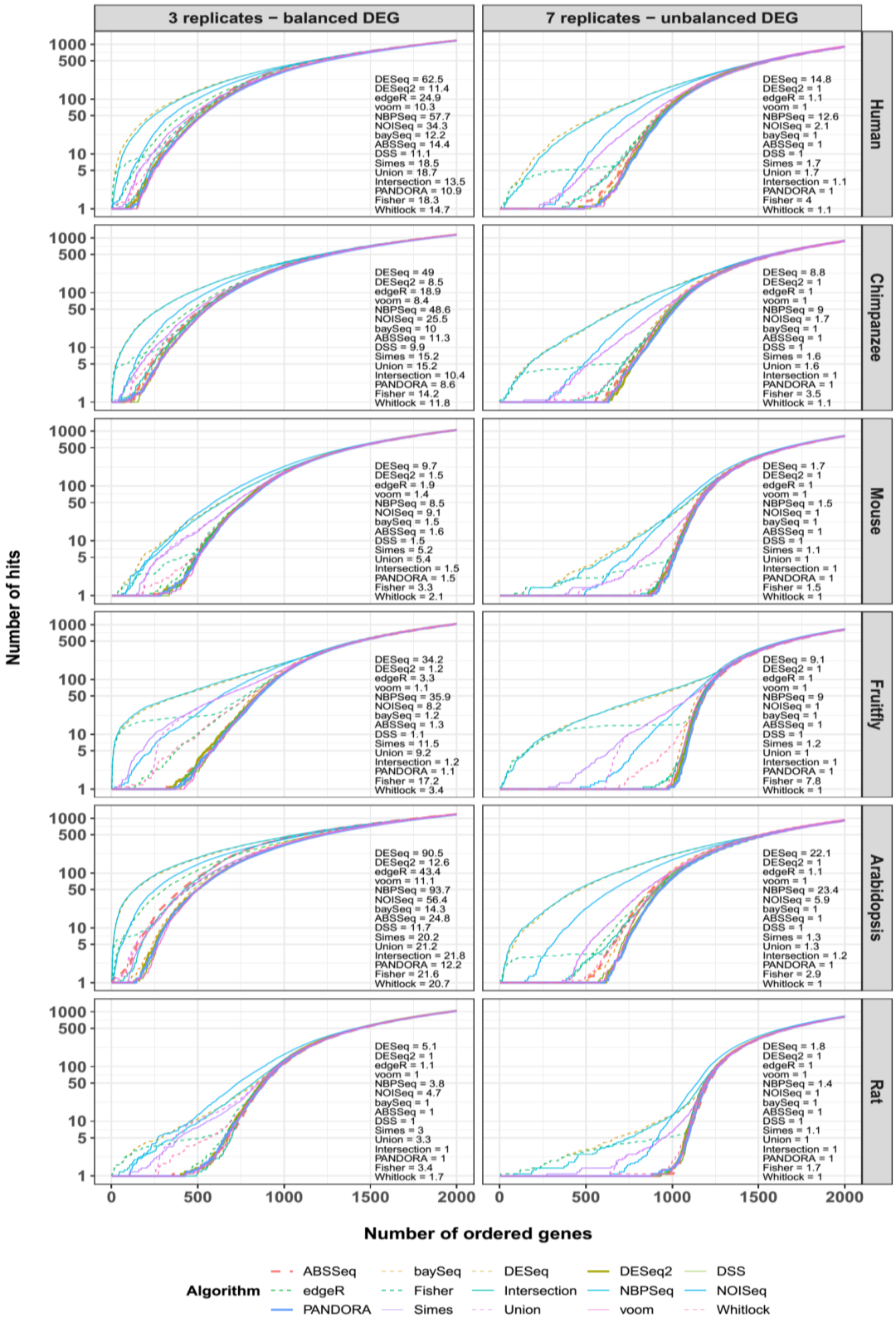
Supplementary Figure 19: Hierarchical clustering of tools' concordance analysis using z-scores calculated from individual metrics applied. lncRNA biotype scheme is here examined. NOISeq and Fisher are not included for reasons explained in Sup. Figure 14 legend.
(BH p-value threshold: 0.05, EDASeq normalization)

7. Appendix II

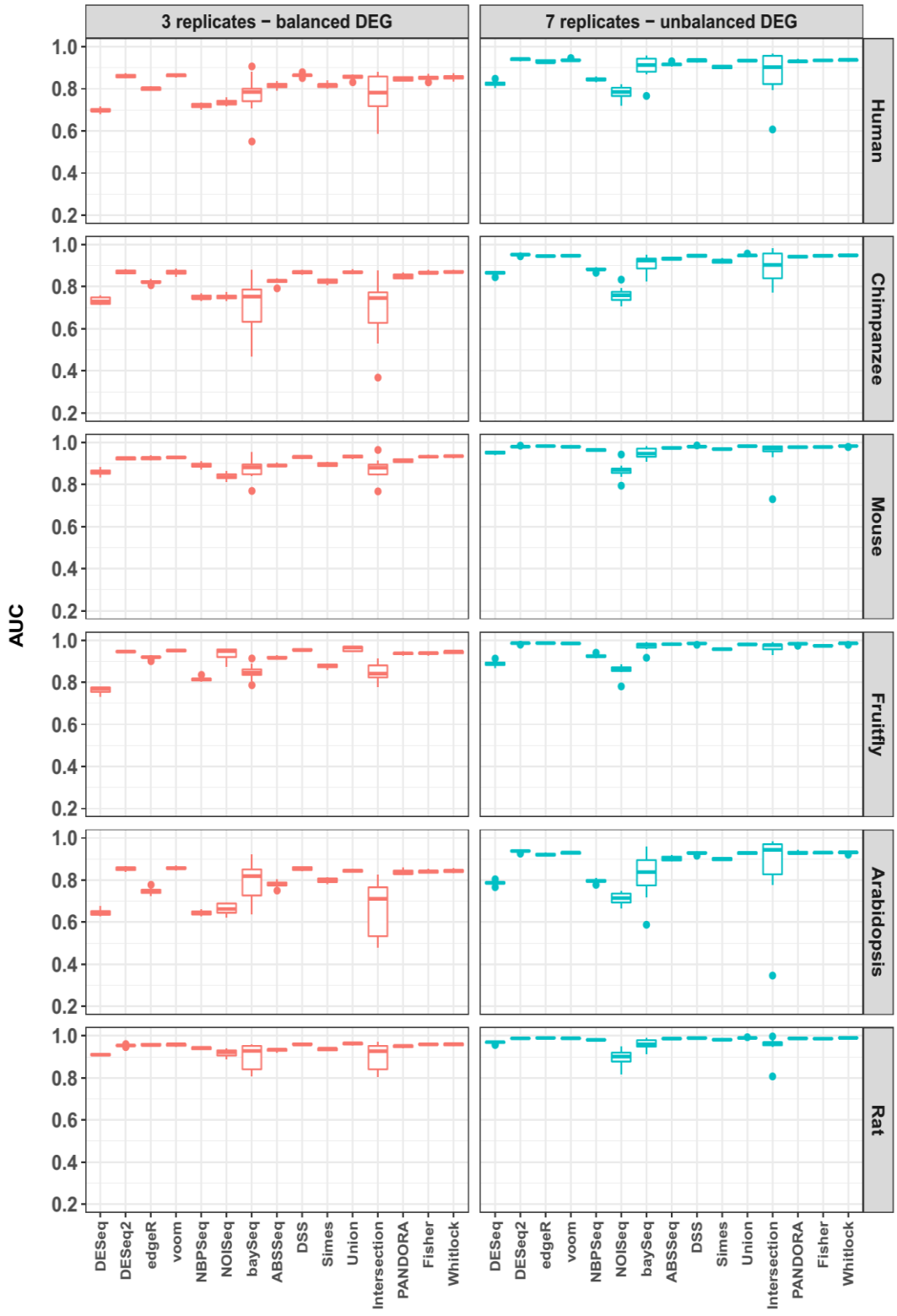
Simulated and real data evaluation figures using each tools' suggested normalization method.



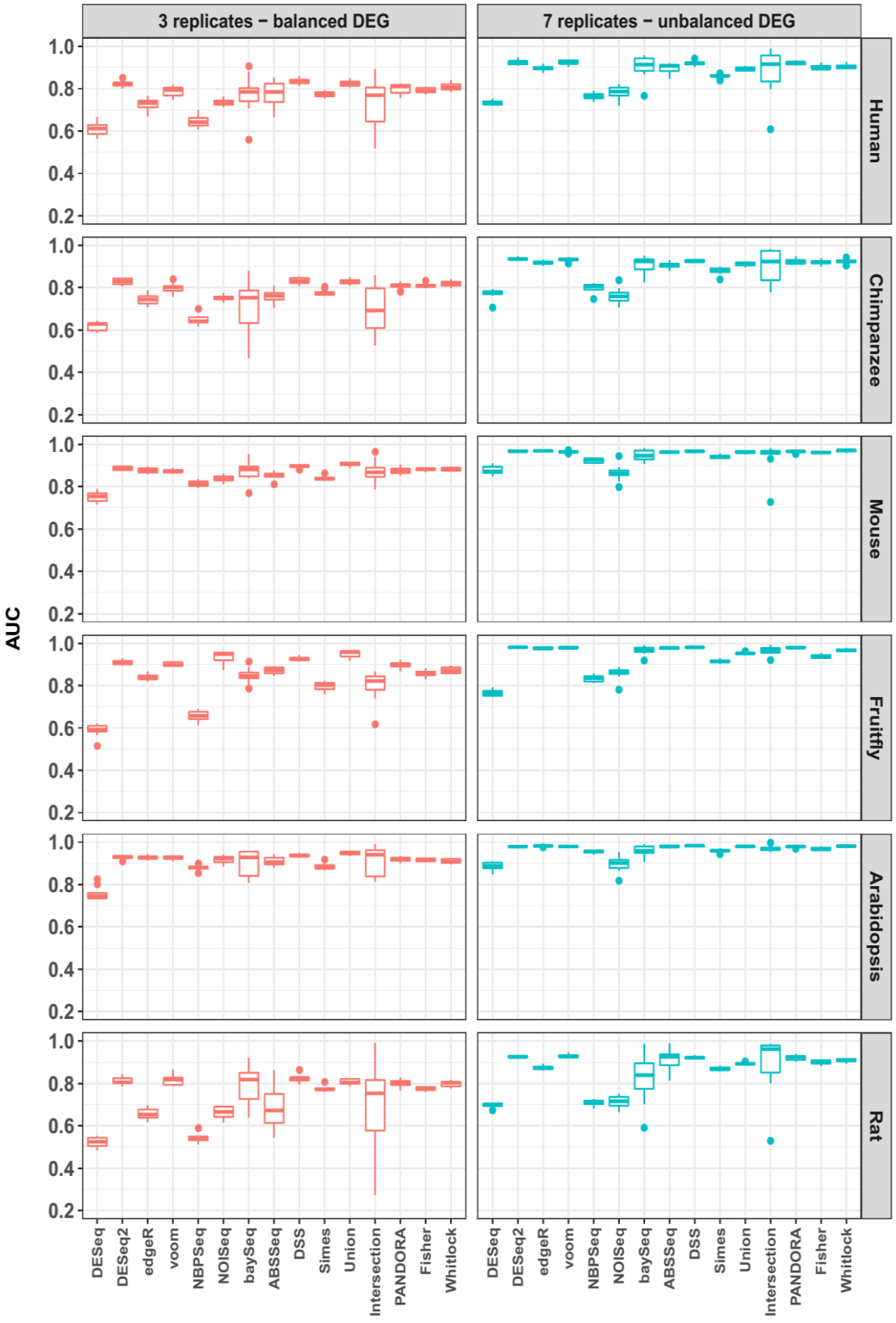
Supplementary Figure 20: False Discovery Curves (FDC) using raw p-values after each tool normalization FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. (significant p-value threshold: 0.05, each tool normalization)



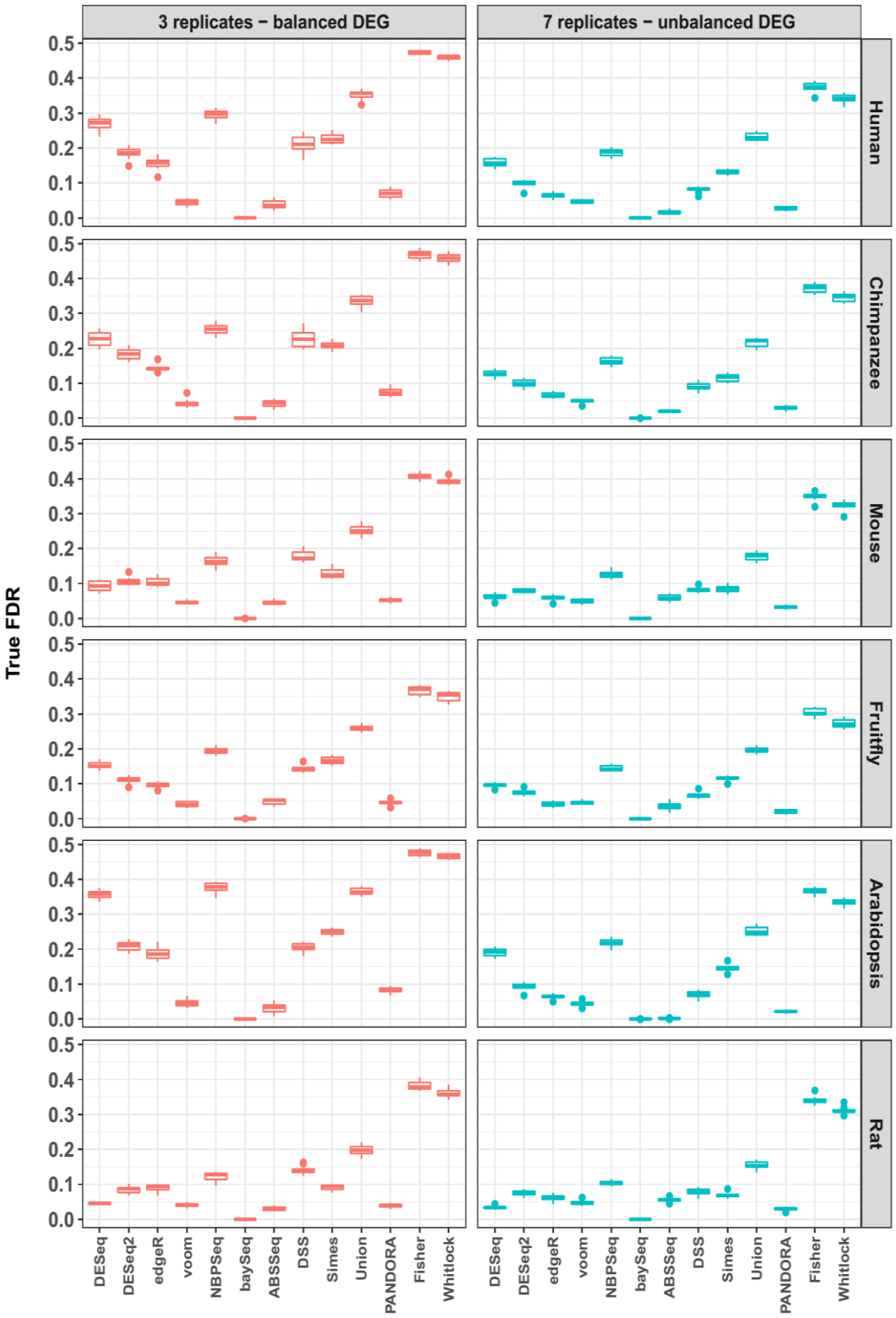
Supplementary Figure 21: False Discovery Curves (FDC) using adjusted p-values after each tool normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. (BH p-value threshold: 0.05, each tool normalization)



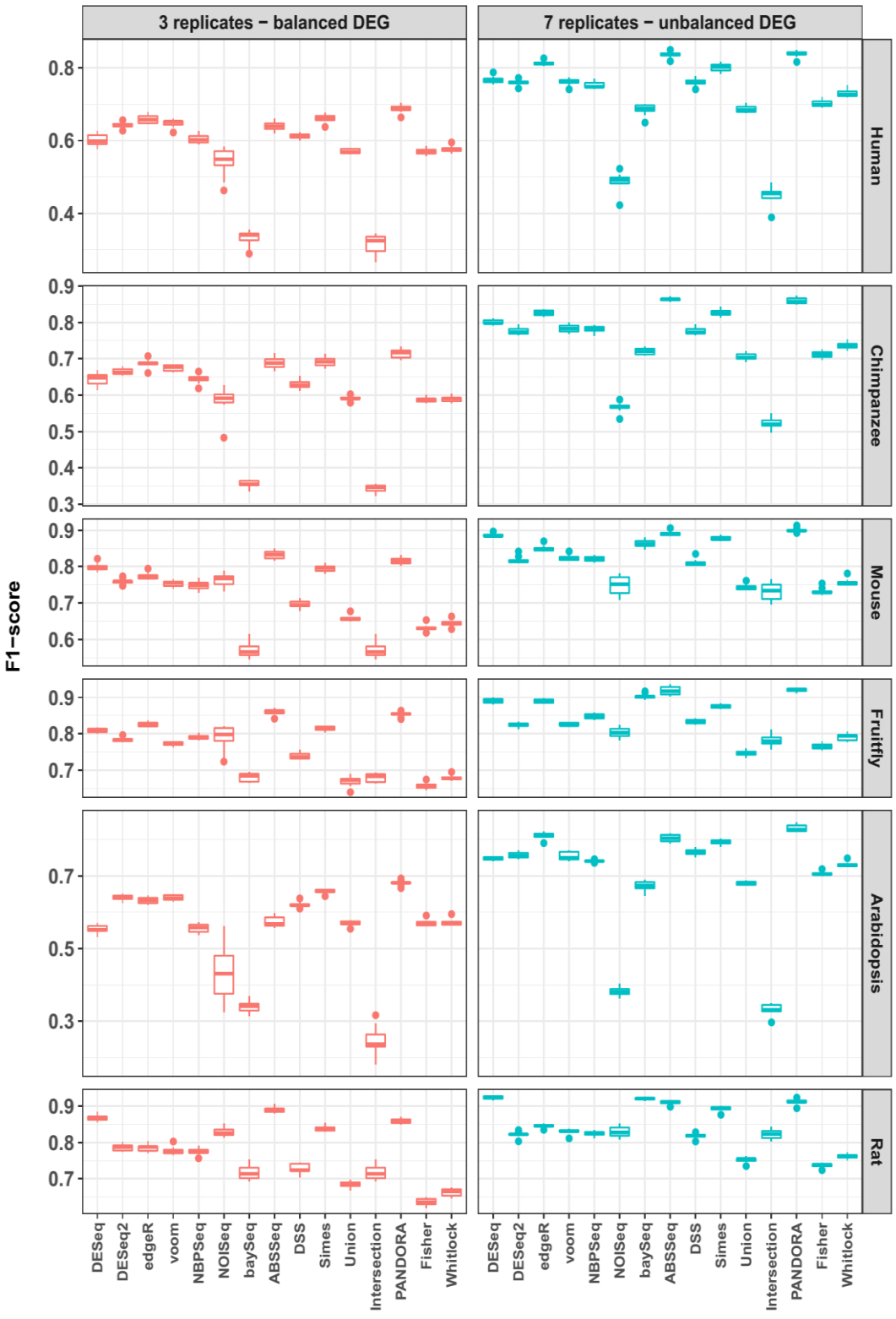
Supplementary Figure 22: Area under the ROC curve (AUC) using raw p-values after each tool normalization. AUC are summarized across ten iterations for each tool and simulation design using unadjusted p-values. (significant p-value threshold: 0.05, each tool normalization)



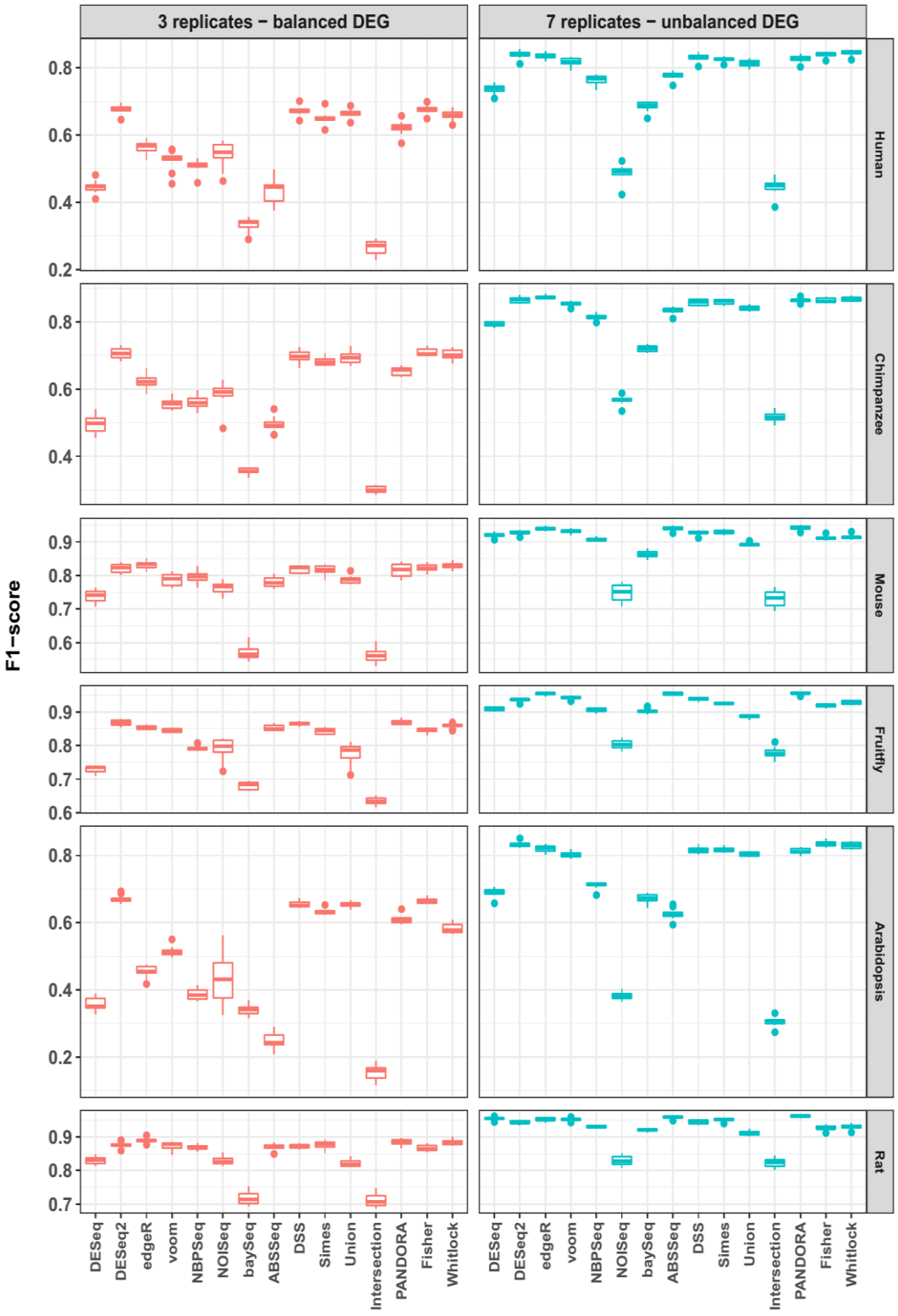
Supplementary Figure 23: Area under the ROC curve (AUC) using adjusted p-values after each tool normalization. AUC are summarized across ten iterations for each tool and simulation design, using adjusted p-values. (BH p-value threshold: 0.05, each tool normalization)



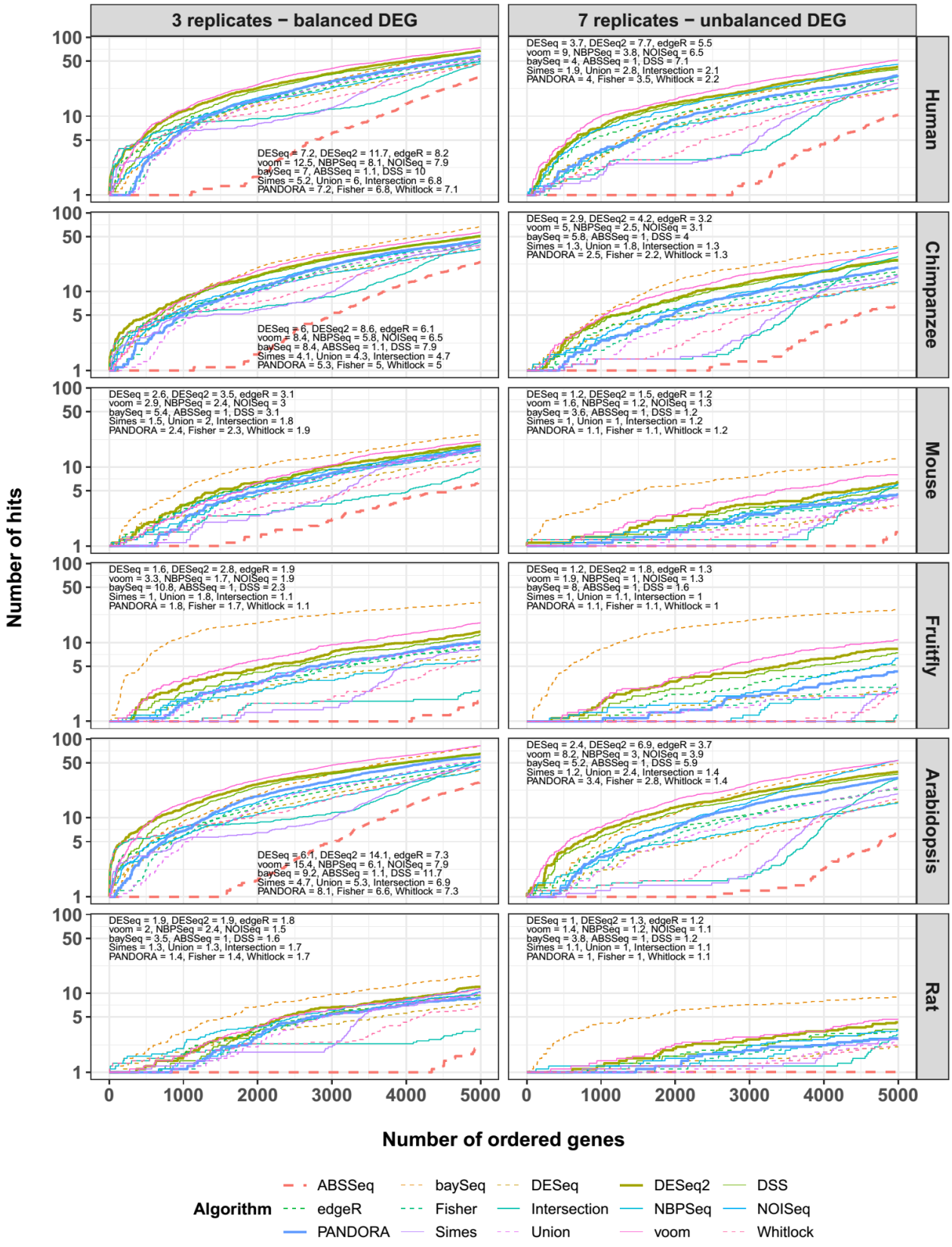
Supplementary Figure 24: False Discovery Rate (FDR) summarized across ten iterations for each tool and simulation design at a BH adjusted p-value threshold of 0.05. Intersection had no discoveries after p-value adjustment and thus it was not included. NOISeq was excluded for reasons explained further above (each tool normalization)



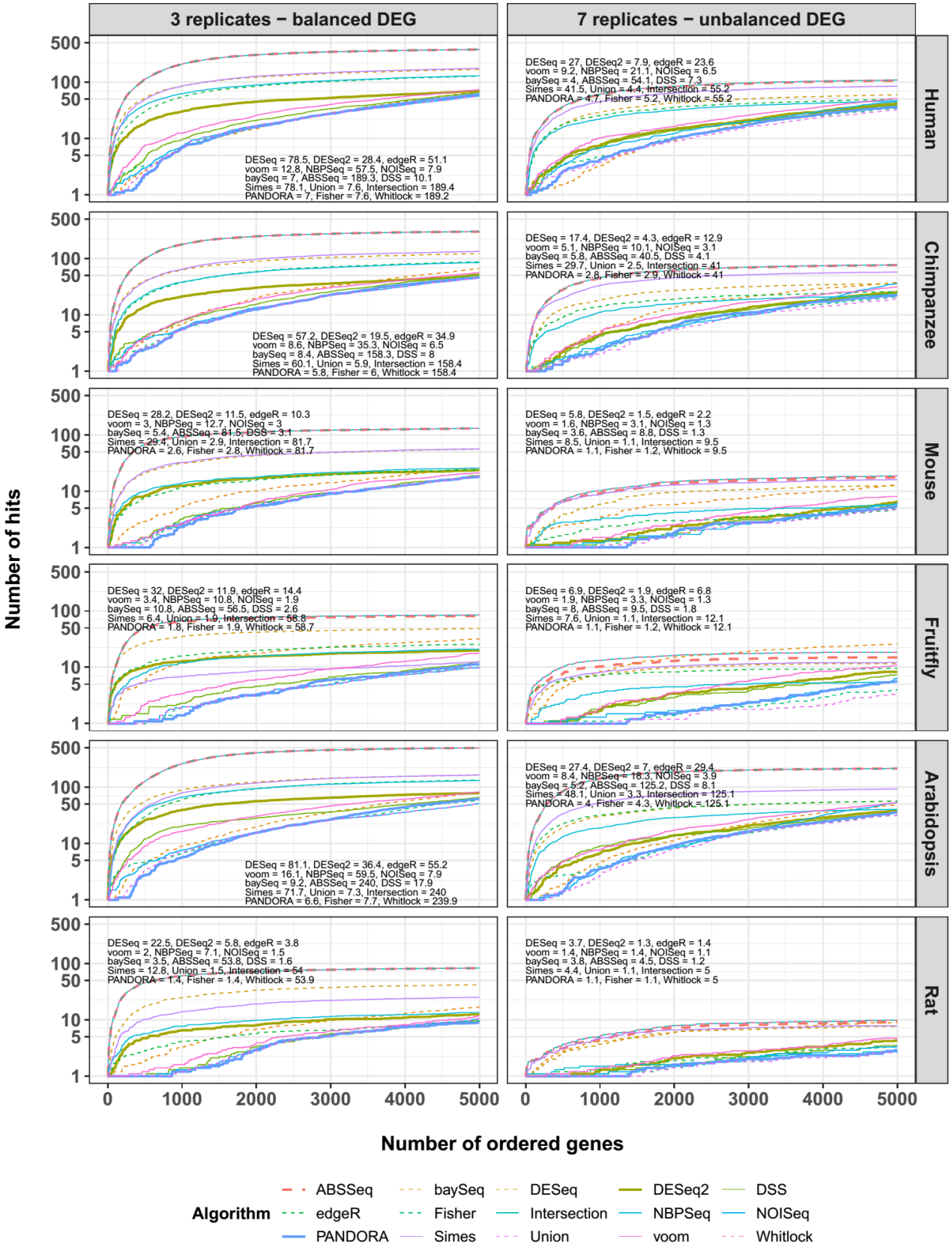
Supplementary Figure 25: F_1-score (precision-sensitivity tradeoff) using raw p-values after each tool normalization. F_1-score is summarized across ten iterations for each tool and simulation design, using unadjusted p-values. (significant p-value threshold: 0.05, each tool normalization)



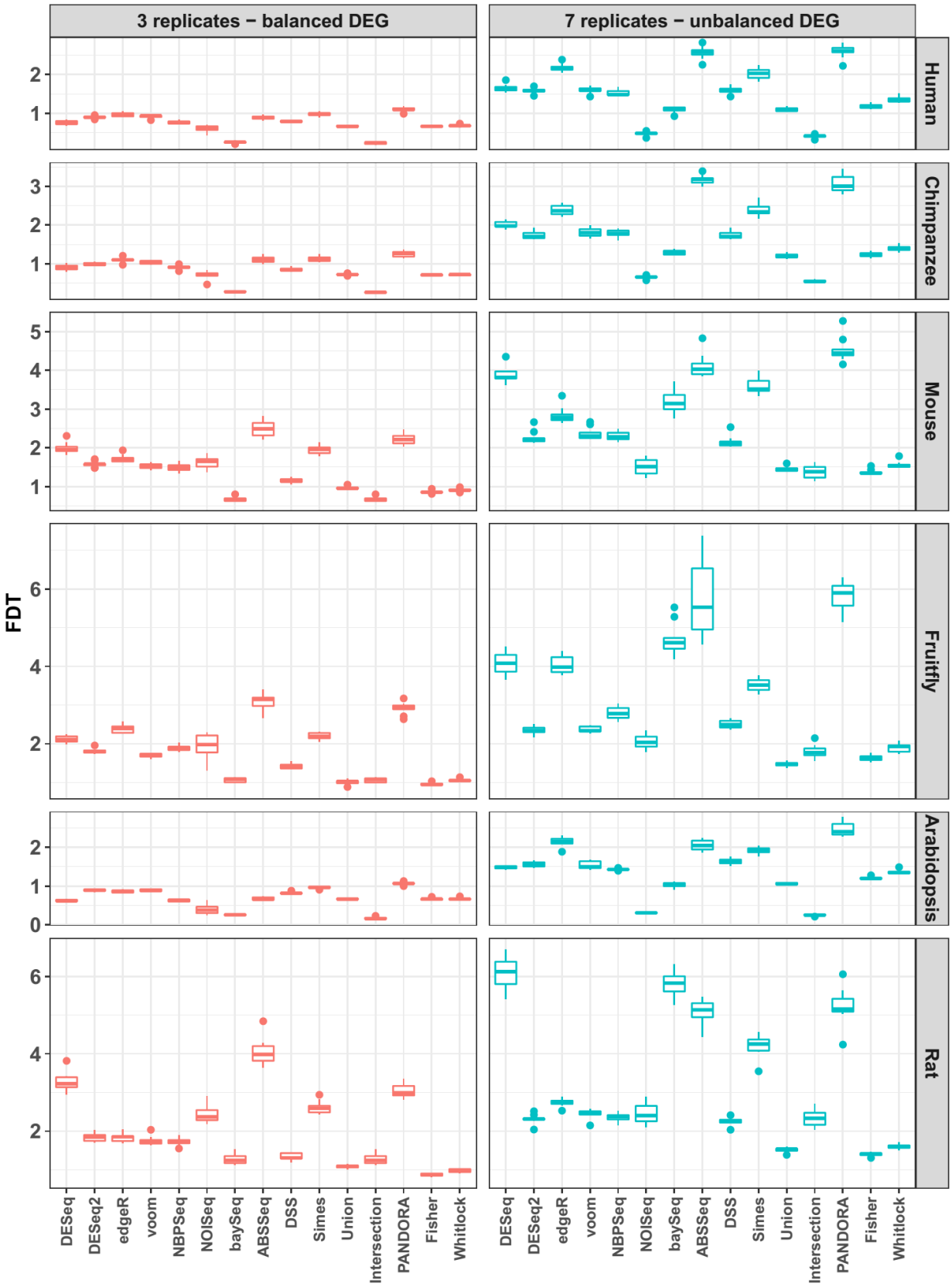
Supplementary Figure 26: F_1-score (precision-sensitivity tradeoff) using raw p-values after each tool normalization. F_1-score is summarized across ten simulations for each tool and simulation design, using adjusted p-values. (BH p-value threshold: 0.05, each tool normalization)



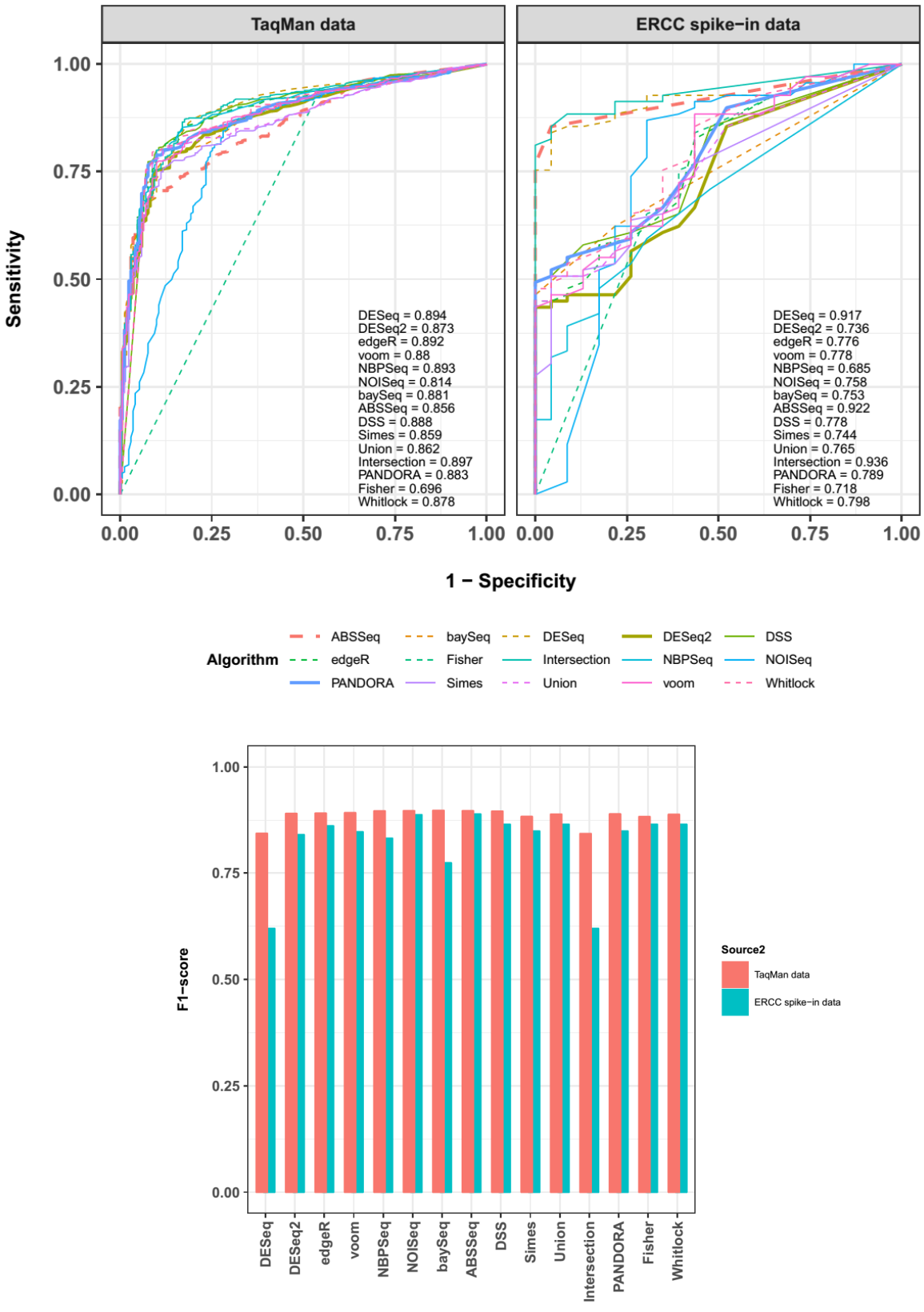
Supplementary Figure 27: False Negative Curves (FNC) using raw p-values after each tool normalization. FNCs are summarized across ten iterations for each tool and simulation design examining the last 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the top left or bottom right corner of each plot (significant p-value threshold: 0.05, each tool normalization)



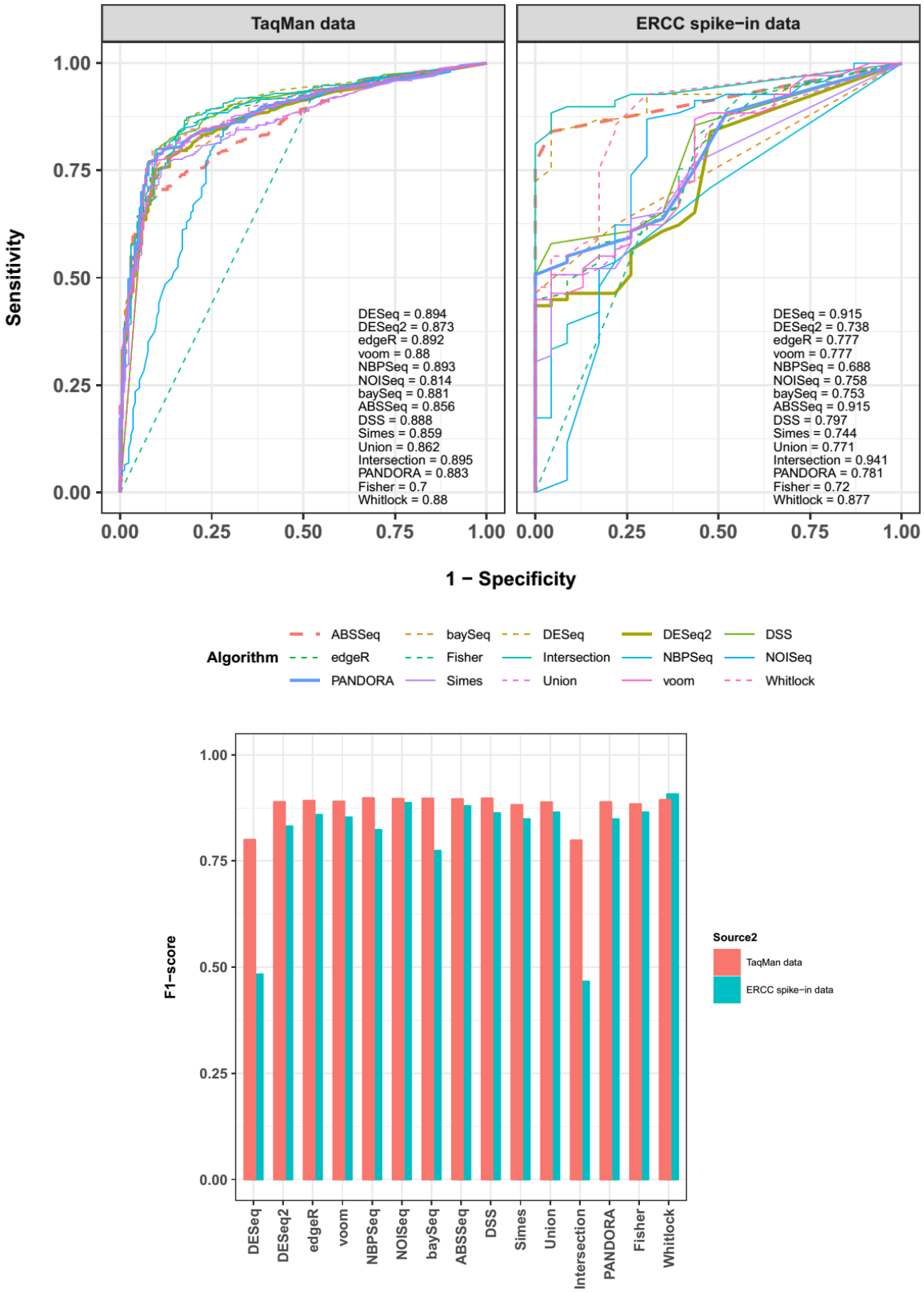
Supplementary Figure 28: False Negative Curves (FNC) using adjusted p-values after each tool normalization. FNCs are summarized across ten iterations for each tool and simulation design examining the last 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the top left or bottom right corner of each plot (BH p-value threshold: 0.05, each tool normalization)



Supplementary Figure 29: False Discovery Tradeoff (FDT) using raw p-values after each tool normalization. FDT is summarized across ten iterations for each tool and simulation design using unadjusted p-values. (statistical significant threshold: 0.05, each tool normalization)



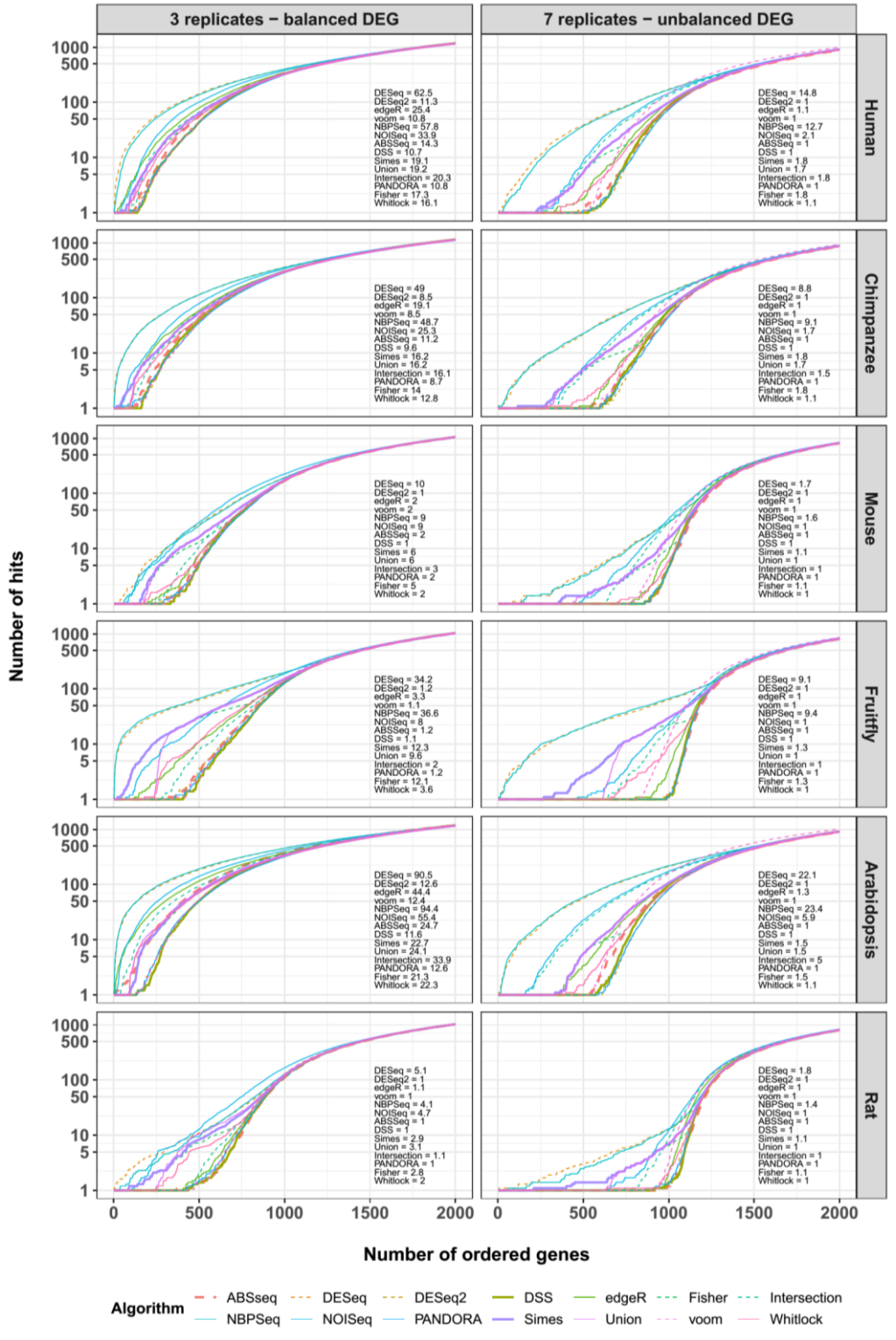
Supplementary Figure 30: ROC and F_1-score analysis of real datasets using raw p-values after each tool normalization (top and bottom respectively). AUC can be found at the bottom right corner of the ROC figures. (significant p-value threshold: 0.05, each tool normalization)



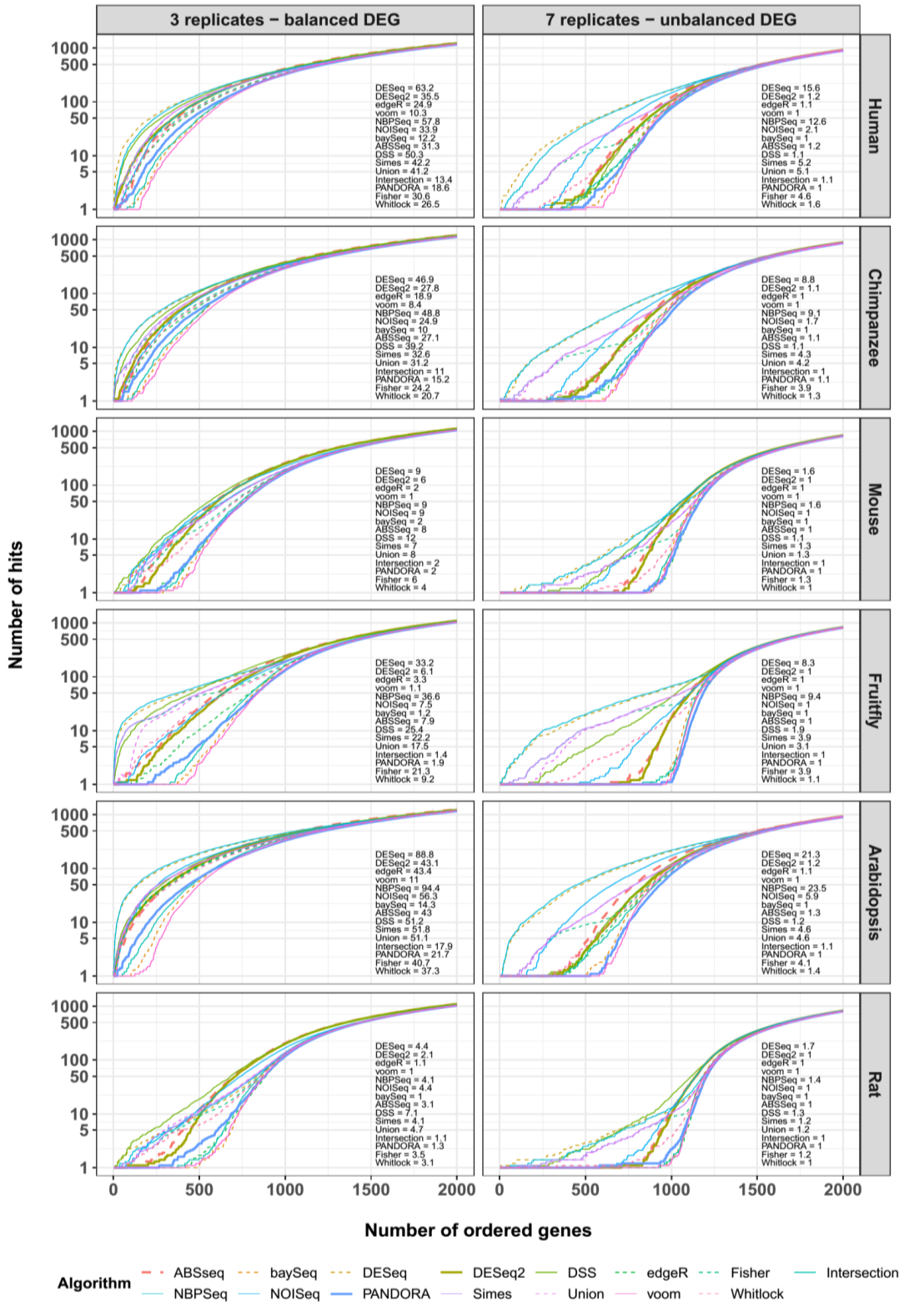
Supplementary Figure 31: ROC and F_1-score analysis of real datasets using adjusted p-values after each tool normalization (top and bottom respectively). AUC can be found at the bottom right corner of the ROC figures. (BH p-value threshold: 0.05, each tool normalization)

8. Appendix III

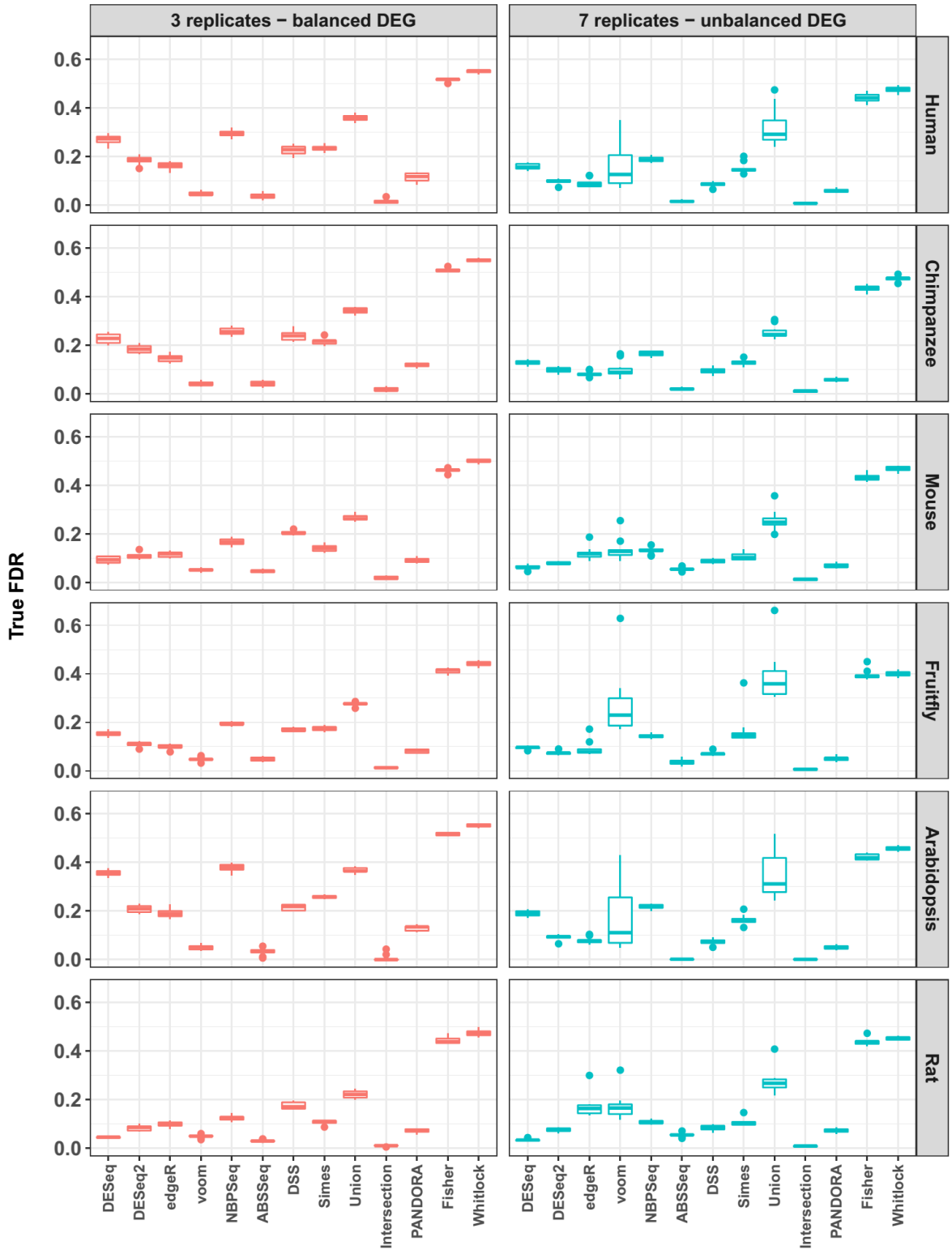
Simulated data evaluation figures using DESeq or TMM normalization.



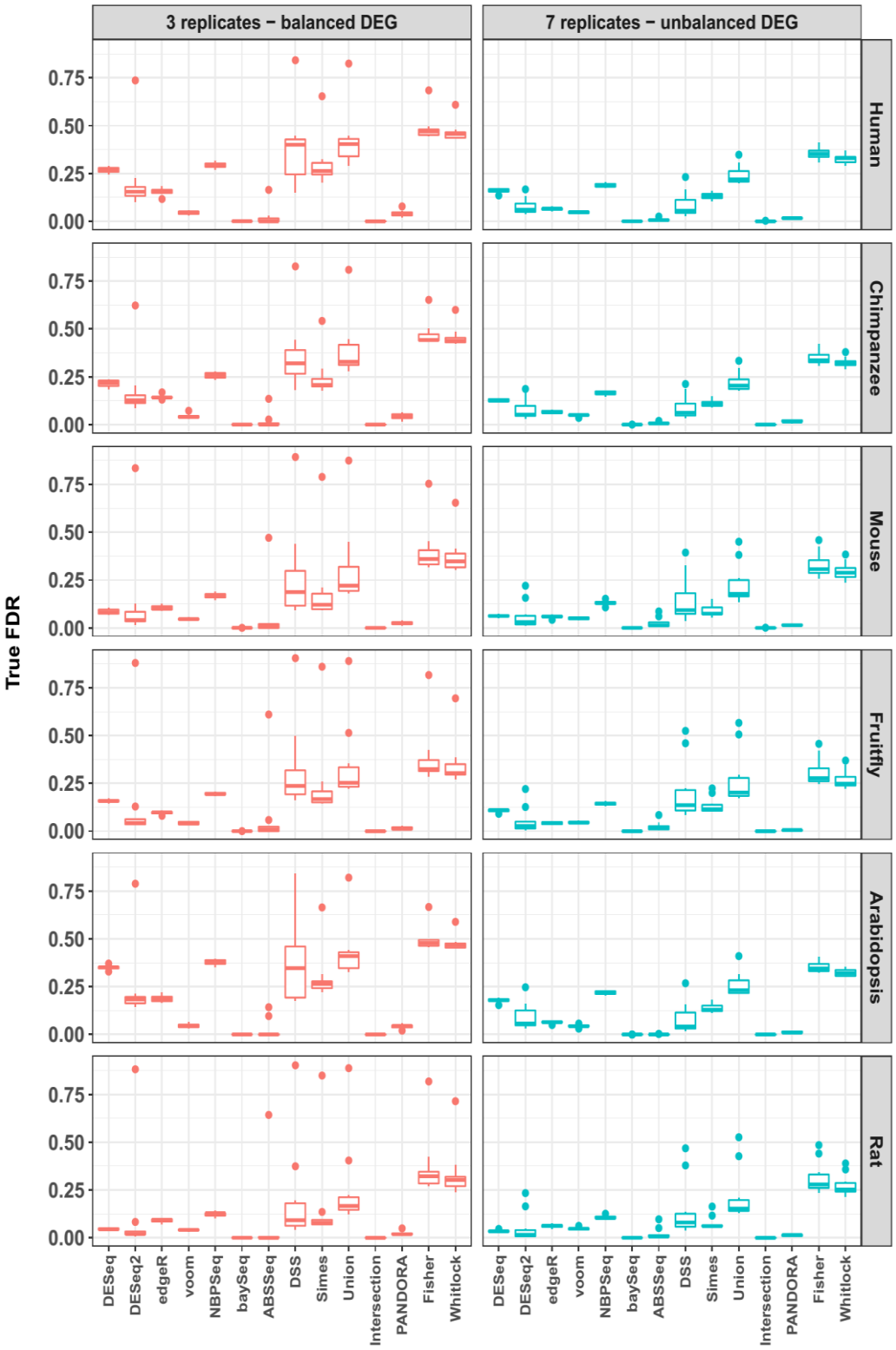
Supplementary Figure 32: False Discovery Curves (FDC) using raw p-values after DESeq normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. (significant p-value threshold: 0.05, DESeq normalization)



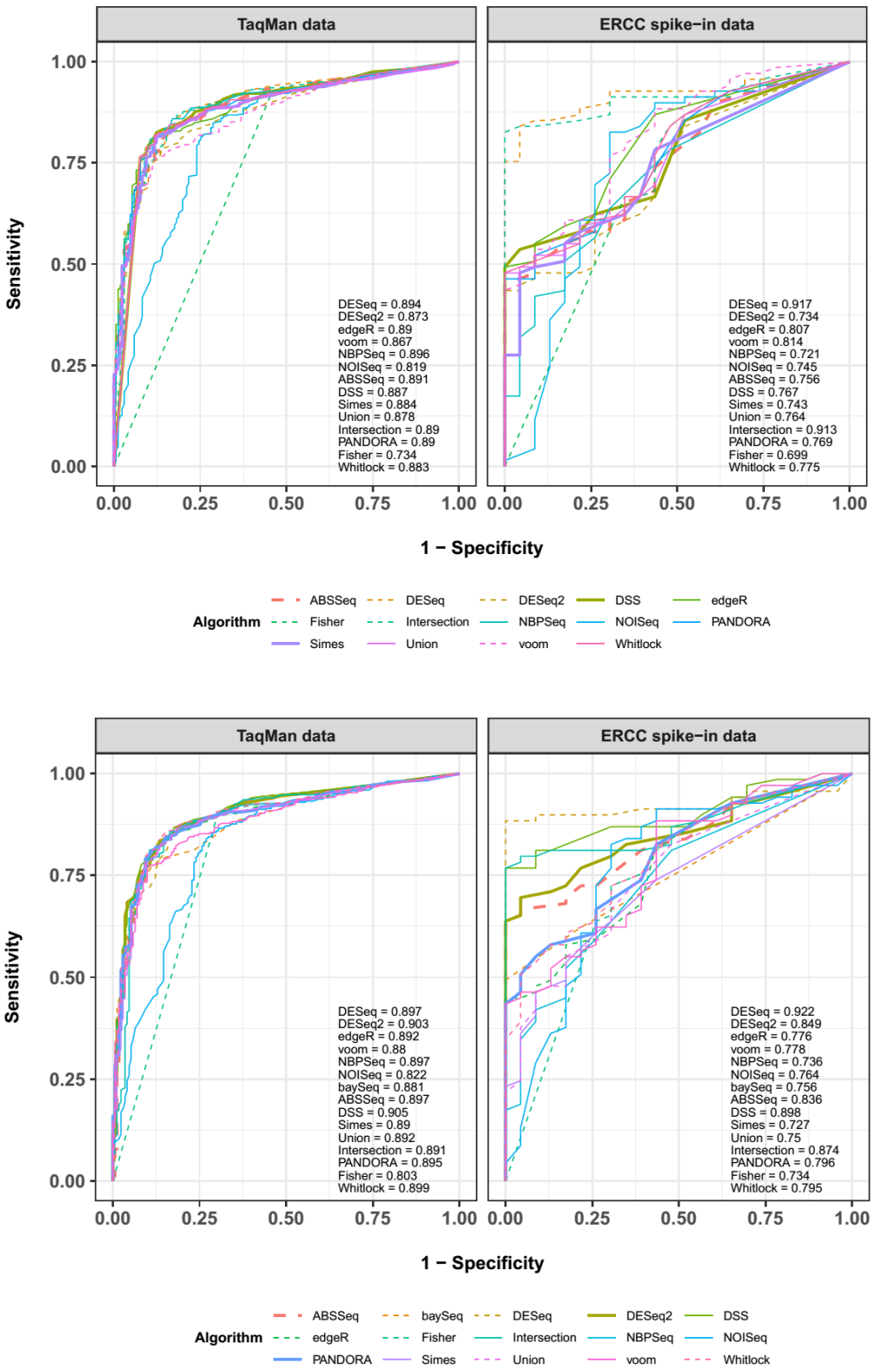
Supplementary Figure 33: False Discovery Curves (FDC) using raw p-values after TMM normalization. FDCs are summarized across ten iterations for each tool and simulation design examining the first 500 DEGs ranked according to statistical significance. The calculated Area Under each FDC (AUFC) can be found at the bottom right corner of each plot. (significant p-value threshold: 0.05, TMM normalization)



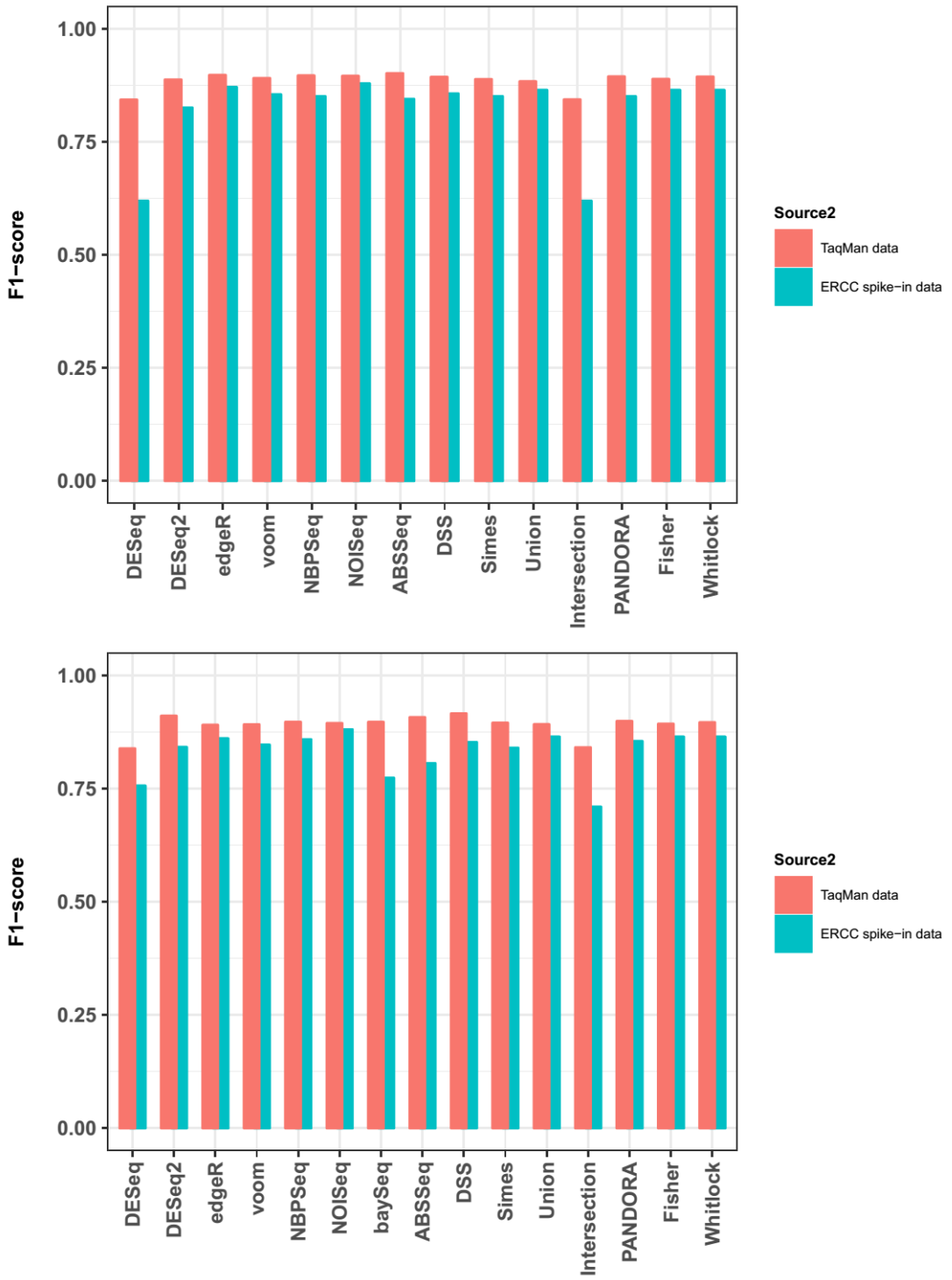
Supplementary Figure 34: False Discovery Rate (FDR) using raw p-values after DESeq normalization. FDR is summarized across ten iterations for each tool and simulation design at a BH adjusted p-value threshold of 0.05. NOISeq was excluded for reasons explained further above (DESeq tool normalization)



Supplementary Figure 35: False Discovery Rate (FDR) using adjusted p-values after TMM normalization. FDR is summarized across ten iterations for each tool and simulation design at a BH adjusted p-value threshold of 0.05. NOISeq was excluded for reasons explained further above (TMM normalization)



Supplementary Figure 36: ROC analysis of real datasets using raw p-values after DESeq and TMM normalization (top and bottom respectively). AUC can be found at the bottom right corner of the ROC figures. (significant p-value threshold: 0.05)



Supplementary Figure 36: F_1-score analysis of real datasets using raw p-values after DESeq and TMM normalization (top and bottom respectively). (significant p-value threshold: 0.05)