Effects of mapping algorithms on gene selection for RNA-Seq analysis: pulmonary response to acute neonatal hyperoxia

Chin-Yi Chu, Soumyaroop Bhattacharya, Zhongyang Zhou, Min Yee, Ashley Lopez, Valerie A Lunger, Bradley Buczynski, Michael Oreilly, Thomas Mariani

Background: A major goal of RNA-Seq data analysis is to reconstruct the full set of gene transcripts expressed in a biological sample in order to quantify their expression levels. The process typically involves multiple steps including mapping short sequence reads to a reference genome, and estimating expression levels based on these mappings. Multiple algorithms and approaches for each processing step exist, and the impact of different methods on estimation of gene expression is not entirely clear.

Methods: We evaluated the impact of three common mapping algorithms on differential expression analysis in an RNA-Seq dataset describing the lung response to acute neonatal hyperoxia. RNA-Seq data generated using the Illumina platform were mapped and aligned using CASAVA, TopHat, and SHRiMP against the mouse genome. Significance Analysis of Microarrays and Cuffdiff were used to identify differentially expressed genes between hyperoxia-challenged and age matched control mice.

Results: 1403 genes were detected as differentially expressed by least one mapping and gene selection method. A majority of genes (>65%) were identified by all three mapping methods, regardless of the gene selection approach. Expression patterns for 52 genes were examined by quantitative polymerase chain reaction (qPCR). Importantly, we found different validation rates for genes selected by each method; 72% for CASAVA, 69% for TopHat and 63% for SHRiMP. Surprisingly, the validation rate for genes selected by all three mapping methods was no greater than the best single method.

Conclusion: The choice of mapping strategy impacts the reliability of gene selection for RNA-Seq data analysis.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.833v1 | CC-BY 4.0 Open Access | rec: 13 Feb 2015, publ: 13 Feb 2015

Effects of Mapping Algorithms on Gene Selection for RNA-Seq Analysis: Pulmonary Response to Acute Neonatal Hyperoxia.

Chin-Yi Chu[1,2,a], Soumyaroop Bhattacharya[1,2,a], Zhongyang Zhou[1], Min Yee[1], Ashley M Lopez[1], Valerie A Lunger[1], Bradley W Buczynski[1], Michael A O'Reilly[1,3], Thomas J Mariani[1,2]

[a]These authors contributed equally.

[1]Division of Neonatology, [2]Pediatric Molecular and Personalized Medicine (PMPM) Program, and [3]Perinatal and Pediatric Origins of Disease (PPOD) Program, Department of Pediatrics, University of Rochester Medical Center, Rochester NY, United States of America

3 Figures
5 Tables
4 Supplemental Figures
1 Supplemental Table

Author Emails:
Chin-Yi Chu: chinyi_chu@urmc.rochester.edu
Soumyaroop Bhattacharya: soumyaroop_bhattacharya@urmc.rochester.edu
Zhongyang Zhou: zzhou10@ur.rochester.edu
Min Yee: min_yee@urmc.rochester.edu
Ashley M Lopez: ashley_lopez@urmc.rochester.edu
Valerie A Lunger: valerie_lunger@urmc.rochester.edu
Bradley W Buczynski: bradley.buczynski@wilresearch.com
Michael A O'Reilly: michael_oreilly@urmc.rochester.edu
Thomas J Mariani: tom_mariani@urmc.rochester.edu

Address for Correspondence:

Thomas J Mariani, PhD
Division of Neonatology and
Center for Pediatric Biomedical Research
University of Rochester Medical Center,
601 Elmwood Ave, Box 850,
Rochester, NY 14642, USA

Phone: 585-276-4616
Fax: 585-276-2643
Email: Tom_Mariani@URMC.rochester.edu

Running Title: Comparative mapping of RNA-Seq

Keywords: TopHat, SHRiMP, CASAVA, qPCR
Supported by: HL67392, HL091968, HHSN272201200005C

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.833v1 | CC-BY 4.0 Open Access | rec: 13 Feb 2015, publ: 13 Feb 2015

ABSTRACT

Background: A major goal of RNA-Seq data analysis is to reconstruct the full set of gene transcripts expressed in a biological sample in order to quantify their expression levels. The process typically involves multiple steps including mapping short sequence reads to a reference genome, and estimating expression levels based on these mappings. Multiple algorithms and approaches for each processing step exist, and the impact of different methods on estimation of gene expression is not entirely clear.

Methods: We evaluated the impact of three common mapping algorithms on differential expression analysis in an RNA-Seq dataset describing the lung response to acute neonatal hyperoxia. RNA-Seq data generated using the Illumina platform were mapped and aligned using CASAVA, TopHat, and SHRiMP against the mouse genome. Significance Analysis of Microarrays and Cuffdiff were used to identify differentially expressed genes between hyperoxia-challenged and age matched control mice.

Results: 1403 genes were detected as differentially expressed by least one mapping and gene selection method. A majority of genes (>65%) were identified by all three mapping methods, regardless of the gene selection approach. Expression patterns for 52 genes were examined by quantitative polymerase chain reaction (qPCR). Importantly, we found different validation rates for genes selected by each method; 72% for CASAVA, 69% for TopHat and 63% for SHRiMP. Surprisingly, the validation rate for genes selected by all three mapping methods was no greater than the best single method.

Conclusion: The choice of mapping strategy impacts the reliability of gene selection for RNA-Seq data analysis.

INTRODUCTION

Genome-wide expression profiling is used to assess the expression of thousands of genes simultaneously, with the ultimate goal of creating a comprehensive description of expression at the mRNA level. Historically DNA microarray technology has been used to measure genome-wide expression. With the advent of next-generation sequencing, high-throughput sequence-based approaches for expression analysis are becoming an increasingly popular alternative to microarrays. RNA-Seq, otherwise known as Whole Transcriptome Shotgun Sequencing (WTSS), refers to the use of high-throughput technologies to sequence cDNA in order to get information about a sample's RNA content. Applications of RNA-Seq may include analyzing the genome for coding and non-coding RNA [1]. RNA-Seq has been used in studies of transcription in yeast [2], Arabidopsis [3], mouse [4, 5], and human [6, 7]. RNA-Seq is a promising replacement for microarrays as initial studies have shown that RNA-Seq expression estimates are highly reproducible[6] and are often more accurate, based on assessments by either quantitative PCR (qPCR) or spike-in experiments [2, 5]. The primary advantages of RNA-Seq are its large dynamic range (spanning five orders of magnitude), low background noise, reduced input sample (RNA) requirement and ability to detect novel transcripts, when studied at appropriate coverage depth [8]. However, similar to the early days of microarray analysis, there are still a number of experimental and computational issues remaining to be resolved for RNA-Seq.

The process of RNA-Seq involves four major steps: (i) generation of cDNA libraries from the RNA samples, (ii) generating millions of short sequence “reads” from these cDNA libraries (that are proportional to the mRNA diversity of the initial RNA sample), (iii) mapping the sequence reads to a reference genome, and (iv) summarizing the mapped reads into raw expression level measures, for subsequent normalization and analysis. Methods and computational tools are being rapidly developed to meet the challenges of these steps. Presently, more than 60 mappers are available to accomplish this task [9, 10]. Our choices of mappers represent a spliced (TopHat) and a contemporary unspliced (SHRiMP) freeware aligner, as well as the vendor recommended default aligner. While both CASAVA and TopHat are spliced read aligner which is a critical feature when aligning RNA-seq reads to a reference genome, we chose to compare them with SHRiMP, which is an unspliced read aligner[11]. Consensus Assessment of Sequence And Variation, (CASAVA) is the part of Illumina's sequencing analysis software that performs alignment of a sequencing run to a reference genome and subsequent variant analysis and read counting [12]. Its underlying alignment algorithm is ELAND (Efficient Large-Scale Alignment of Nucleotide Databases). ELAND is very fast and can perform multi-seed and gapped alignment. TopHat is a fast and popular spliced aligner for RNA-Seq reads [13]. It aligns RNA-Seq reads to mammalian-sized genomes by using its underlying mapping engine, either Bowtie or Bowtie2. TopHat consists of three mapping steps: transcriptome mapping, genome mapping, and spliced mapping. This three-step pipeline assures the best and unique alignment for each read. It can identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes. Short Read Mapping Program (SHRiMP) is an unspliced aligner and it aligns reads against a target genome using an algorithm distinct from that TopHat uses [14]. It was initially developed with the multitudinous short reads of next generation sequencers and Applied Biosystem's colorspace genomic representation. The algorithm can align reads with extensive polymorphism and sequencing errors. It can also perform the multi-seed alignment to speed up the alignment.

Here we studied the biological impact of these three mapping algorithms on identification of differentially expressed genes in a typical experimental model dataset. We included molecular validation for genes identified in individually mapped datasets as a means of determining the estimated reliability of each method. Similar studies in the past have used empirical and simulated datasets [15-18], however, these studies did not have further validation results.

METHODS

RNA-Seq Data

This data set represents high throughput sequencing of seven cDNA libraries generated from whole lung tissue RNA recovered from mice treated under hyperoxic conditions in the newborn period (n=3) or age-matched controls (n=4). Mice were exposed to 100% oxygen, 40-70% humidity between birth and postnatal day 10 as shown previously [19]. Mice were housed in sterile microisolator cages in a specified pathogen-free environment and exposed to super-physiological levels of oxygen according to a protocol (Protocol No. 2007-121R) approved by the University Committee on Animal Resources at the University of Rochester. The University Committee on Animal Resources (UCAR) at the University of Rochester reviewed and approved these studies. The data were generated on the Illumina Genome Analyzer II platform at an average sequence depth of 20 million reads (65 bases in length) per sample as reported previously [20].

Short-Read Alignment

Sequences were mapped to the mouse genome (mm10), containing 21,718 unique genes, using multiple alignment algorithms. CASAVA (v 1.7) pipeline software was implemented with the manufacturer default settings, using the ELAND aligner program in Multiplexer setting, for separating bar-coded reads into different bins. Tophat (v 2.0.9) was implemented using Bowtie (v 2.1.0) and SAMtools (v 0.1.18) with default parameters (b2-sensitive, report-secondary-alignments, library-type fr-unstranded). SHRiMP (v 2.2.3) was implemented using SAMtools (v 0.1.18) with default parameters (gmapper – ls –qv-offset 33 single_en_reads-). For TopHat and SHRiMP mapping, HTSeq (v 0.5.3p3) was used to generate the count matrices with the following parameters: 'htseq-count -m intersection-strict -s no'. For both SHRiMP and TopHat alignments, mouse gene annotation file (GTF) for mouse genome build 10 obtained from UCSC was used for alignment. Codes for mapping and generating the raw counts are available in shown in Supplemental Table 1.

Normalization

Alignment counts obtained from CASAVA, SHRiMP and TopHat were further normalized using reads per million bases (RPM) or trimmed-mean (TM), independently. RPM normalization involves dividing the raw count for each individual gene in a particular sample by the sum total of counts for all the genes in that sample (Equation 1). It is a modification of the RPKM (reads per kilobase transcript per million reads) approach, excluding normalization for transcript length [5].

$$Count_{RPM} = \frac{Count_{Raw}}{\sum_{sampleA} Count} \quad \text{…..(1)}$$

TM normalization involves calculation of the mean of 5th-95th percentile of raw counts for each sample, which is used to determine a sample-specific normalization factor. The raw counts for all genes for a sample are then multiplied by this factor (Equation 2).

$$Count_{TrimMEAN} = Count_{Raw}\left(\frac{1}{\mu_{\text{Trim (5-95\%)}}}\right) \quad \text{……(2)}$$

Gene Selection

Genes that were not detected in all samples of at least one experimental group (hyperoxia-treated or control) were excluded from analysis. Multiple gene selection approaches were applied on each normalized and filtered dataset, from each of the three mapping algorithms. The filtered and normalized data, were however not log-transformed. Significance Analysis for Microarrays (SAM) is a frequentist approach used for identification of differentially expressed genes that uses a modified t-statistic with permutations [21]. SAM was applied using the minimum median False Discovery Rate (FDR) possible, in MultiExperiment Viewer (MeV) v4.8.1 (http://www.tm4.org/mev.html). For the purposes of this study, SAM threshold of median FDR of 0 was applied. Cuffdiff uses the Cufflinks transcript quantification to calculate gene expression levels in different conditions, and tests them for significant differences [22]. It uses FPKM (fragments per kilobase of exon per million fragments mapped) normalization. Cuffdiff was applied using mouse gene annotation file (GTF) for mouse genome build 10 obtained from UCSC with a significance threshold of FDR adjusted q <0.05.

Molecular Validation

cDNA were synthesised from RNA samples isolated from individual lungs using the iScript Reverse Transcription Kit (BioRad, Hercules, CA). qPCR was performed on a Viia7 (Applied Biosystems, Santa Clara, CA) using SYBR green chemistry as previously described [23]. Gene-specific assays primer sequences were retrieved from the MGH Primer Bank (http://pga.mgh.harvard.edu/primerbank). Gene expression levels (dCt) were calculated relative to the measured Ct value of PPIA (peptidyl prolyl isomerase A or cyclophilin A) as an internal, endogenous control and were analyzed for relative expression changes by the ddCt method as previously described [23]. QPCR data were assessed using both the students T-test and the Mann-Whitney U test at a p<0.05.

RESULTS

The main objective here is to compare the effect of different, state-of-the-art alignment and mapping procedures upon the accuracy of gene selection. The data set we used for this analysis consists of 3 experimental treatment samples (separate pools of RNA isolated from lungs of neonatal mice challenged with hyperoxia) and 4 matched, normoxic controls [20]. As indicated in Figure 1, in order to complete the analysis, Fastq files were independently mapped using 3 different alignment methods; CASAVA, TopHat and SHRiMP. The mapped data were normalized and filtered, using tools appropriate for the mapping methods, and analyzed for differential expression using Significance Analysis for Microarrays (SAM) or Cuffdiff. In particular, we used reads per million (RPM) or trimmed mean (TM) normalization for SAM, and "fragments per kilobase of exon per million fragments mapped" (FPKM) normalization for Cuffdiff. Differential expression was estimated using SAM, for all three mapping methods, or Cuffdiff for TopHat and SHRiMP mapping. qPCR was used to assess the reliability of differential expression estimation.

Mapping Statistics

For each mapping algorithm we calculated the total number of reads, percentage of genes detected, alignment rate, number of aligned reads, and the number of reads assigned to genes. Table 1 shows a summary of these mapping statistics. SHRiMP gave highest total number of reads followed by those of CASAVA and TopHat. On an average, SHRiMP gave highest percent genes, detected (80%) which was slightly above those of CASAVA (77%) and TopHat (76%). In addition, the total number of reads assigned to mapped genes was highest in SHRiMP (12.75 x 10^6), followed by TopHat (12.56 x 10^6) and CASAVA (11.18 x 10^6).

The correlation analysis confirmed strong general concordance on the gene expression measurements across mappers. Pearson correlation coefficients between the raw or either of normalized counts generated by the TopHat and CASAVA was found to be well above 0.8 indicating the data were of comparable quality (Figure 1). While the pearson correlation coefficients between the raw or either of normalized counts generated by the SHRiMP and CASAVA were found to be 0.46 or more, the correlation coefficients between the raw or either of normalized counts generated by the SHRiMP and TopHat were 0.58 or more. Figure 2 shows representative plots (from one control and one hyperoxia sample) for same-sample correlations among the normalized counts obtained using three mappers. Correlation plots for raw and both normalized counts can be seen in Supplemental Figure 1.

CASAVA (v 1.7)

On average, expression of 77% of the genes in the genome was detected in the samples using CASAVA. After removing signal from genes not present in all samples of at least one experimental group, the expression of 16,079 genes were assessed for differential expression (Table 2). As described elsewhere [20] SAM identified 1020 genes in the TM-normalized data and 813 genes in the RPM-normalized data. A total of 798 genes were common, and 300 of these had a fold-change greater than or equal to 2. There were a greater number of genes showing significant decreases in expression in response to treatment (56%) than were increased.

SHRiMP (v 2.2.3)

On average, expression of 80% of the genes in the genome was detected in the lung tissue samples using SHRiMP. After removing signal from genes not present in all samples of at least one experimental group, the expression of 17,814 genes were assessed for differential expression (Table 3). SAM identified 879 genes in the TM-normalized data and 937 genes in the RPM-normalized data. A total of 857 genes were common, and 386 of these had a fold-change greater than or equal to 2. Similar to CASAVA mapped data, there were more genes showing significant decreases in expression in response to treatment (60%) than were increased. Cuffdiff identified 3886 genes as differentially expressed, and 1087 of those had a magnitude of change greater than or equal to 2. In addition to selecting a greater number of genes as significantly affected, Cuffdiff identified a greater number of genes showing significant increases in expression in response to treatment (57%) than were decreased.

TopHat (v 2.0.9)

On average, expression of 76% of the genes in the genome was detected in the lung tissue samples using TopHat. After removing signal from genes not present in all samples of at least one experimental group, the expression of 16,892 genes were assessed for differential expression (Table 4). SAM identified 880 genes in the TM-normalized data and 951 genes in the RPM-normalized data. A total of 860 genes were common, and 396 of these had a fold-change greater than or equal to 2. Cuffdiff identified 2831 genes as differentially expressed, and 1044 of those had a fold-change greater than or equal to 2. Again (similar to SHRiMP mapped data) Cuffdiff identified a greater number of genes showing significant increases in expression in response to treatment (57%) than were decreased.

Consistency of Gene Selection

We tested the consistency of individual gene selection tools. SAM identified 699 genes that were selected as differentially expressed by all three mapping methods, of which 251 had a fold change greater than or equal to 2 (Supplemental Figure 2). CASAVA tended to be more conservative in the number of genes identified, with a substantial majority of these genes also selected by SHRiMP and TopHat.

Cuffdiff identified 2719 genes that were selected as differentially expressed using both SHRiMP and TopHat, of which 919 had a fold change greater than or equal to 2 (Supplemental Figure 3). Again, there was a high degree of consistency between SHRiMP and TopHat, with nearly 90% of genes identified using data mapped by both methods.

There were a total of 240 genes identified by all these analyses (Supplemental Figure 4). SAM, implemented as described, was much more highly conservative in gene selection. Most genes identified by SAM were also identified by Cuffdiff, while a majority of genes identified by Cuffdiff (74%) were not identified by SAM.

In order to compare gene selection estimation with other mapping algorithms (TopHat and SHRiMP), we subsequently, focused on Cuffdiff as well. A total of 267 genes were identified as differentially expressed by SAM on CASAVA, and by Cuffdiff on both SHRiMP and TopHat , which had a fold change greater than or equal to 2 (Figure 3).

Molecular Validation

Predicted changes were evaluated by qPCR for 52 of the genes (Table 5). Out of these 32 genes were identified using CASAVA mapped data separately as reported previously [20]. We chose 10 additional genes each that were uniquely selected by SHRiMP or TopHat for qPCR validation. These genes were chosen based on prior knowledge of their relevance to oxidative stress response or lung biology in general. Genes, that had a significant difference between the two groups (hyperoxia and controls) by either t-test or Mann-Whitney U test at p-value less than 0.05, were designated as successfully validated. We report detailed qPCR results for 32 genes identified using CASAVA mapped data separately [20]. We validated differential expression for 23 of 32 (72%) of these genes selected by CASAVA. Of these 32 genes, 29 were selected as differentially expressed by TopHat, and 20 of those genes (69%) showed significant differences in expression by qPCR. Of these 32 genes, 28 were selected as differentially expressed by SHRiMP, and 19 of those genes (63%) showed significant differences in expression by qPCR. Of these 32 genes, 27 were selected as differentially expressed by all three mapping methods, and 18 of those genes (67%) showed significant differences in expression by qPCR.

Two of the 32 genes were uniquely selected by CASAVA, and both of those genes (100%) showed significant differences in expression by qPCR. For TopHat, 7 of the 10 (70%) additional unique genes, and a total of 27 of 39 (69%) genes tested showed significant differences in expression by qPCR. For SHRiMP, 5 of the 10 (50%) additional unique genes, and a total of 24 of 38 (63%) genes tested showed significant differences in expression by qPCR.

DISCUSSION

RNA-Seq is becoming the method of choice for genome-wide transcriptomics analysis, and has been used to identify post-transcriptional changes and other modifications in human diseases such as cancer [24]. Many methods for data processing and analysis are available, but the best approaches to estimate differential expression for specific types of data sets are not at all clear. Here, we describe an empirical assessment of the impact three different mapping algorithms upon the reliability of differential expression estimation, in a typical genome-wide expression data set from an animal model of disease. Our initial analysis of this data set indicated a substantial gene expression response associated with the experimental challenge, such that it was appropriate for the current studies. The rationale behind choosing these three included multiple factors, such as manufacturer recommendation, usage among research community, cost, efficiency and ease of usage, among others.

Researchers can consider multiple mapping algorithms in their RNA-Seq analysis, which brings up issues of interoperability. To achieve interoperability, input and output formats need to be standardized. Currently the level of interoperability is high since most of the mapping algorithms accept FASTQ format input files and generate SAM/BAM files as output. Prior publications comparing various combinations of mapping algorithms and tests for differential gene expression have found high level of consistency among the results [15, 16]. However, these prior studies did not include subsequent attempts of molecular validation, which is a critical step in any differential expression analysis.

It is important to point out that, of the three different mapping algorithms compared in this study, two of them (SHRiMP and TopHat) were developed at academic institutions, and are freely available, while the third (CASAVA) is commercial software that is provided by the manufacturer of the sequencing instrument. We observed that SHRiMP mapping resulted in higher rates of genes detected as expressed in the samples (80% for SHRiMP, 77% for CASAVA and 76% for TopHat), when compared to either CASAVA or TopHat. This is likely due to the fact that SHRiMP allows reads to be mapped to multiple loci, unlike TopHat and CASAVA, which require reads are mapped to a single locus. Interestingly, a higher number of genes detected as expressed led to higher estimation of differential expression for SHRiMP, but a somewhat lower level of accuracy as defined by qPCR.

We report here only a subset of the possible permutations of analysis that could be completed with the mapping, normalization and gene selection methods we have included. In addition to comparing the mappers, we also looked at the effects of methods of count generation by running correlation analysis among the raw and normalized counts of the same samples, and found that even among methods using different count generation approaches (CASAVA and TopHat), there was a high level of correlation among the counts (Supplemental Figure 1). This indicated to us that the counting methods, independent of mapping algorithm, may not have a big impact. Our analysis also revealed that there was a high level of consistency among the two normalization methods (RPM and TM) on each mapped version of data, when it comes to gene selection, irrespective of the test for differential expression used. We, however did notice higher number of genes being identified by Cuffdiff when using FPKM normalized counts. Even though SAM appeared to be effective in the current data set, applying this analytical approach to other RNA-Seq datasets identified a number of limitations. For this, and other reasons (e.g., free access to software, difficulties in generating bam files from CASAVA), we decided to use SAM analysis of CASAVA data as our benchmark, and focus on the efficiency of Cuffdiff selection using SHRiMP and TopHat mapping.

Encouragingly, we find that all mapping approaches performed similarly. Overall validation rates for estimation of differential expression were comparable. CASAVA was most conservative with gene selection, consistently estimating fewer genes as differentially expressed as compared to SHRiMP and TopHat. However, this was associated with a higher frequency of validation; 72% for all genes, 100% for unique genes. Conversely, SHRiMP tended to be slightly more liberal than TopHat, but had a reduced validation rate than the other mapping approaches (63% for all genes, 50% for unique genes). It is likely that this is at least partially due to its tolerance for including non-unique mapping for individual reads. Interestingly, TopHat selected genes as differentially expressed at nearly the same rate as SHRiMP, but its validation rate was much closer to that of CASAVA (69% for all genes, 70% for unique genes). It is significant to note that individual genes selected by all mapping methods did not demonstrate a higher validation rate than those identified by a single mapping method alone.

CONCLUSIONS

In summary, these data describe the differences that can be expected in the performance of these three common mapping strategies, when applied to a typical genome-wide expression data set comparing biological paradigms, such as an animal- or cell model response. Our data highlight the importance of considering the analytical goals when choosing a data analysis approach. For instance, focused analyses may want to consider a more conservative approach with a slightly higher validation rate, while discovery approaches may be more tolerant to, and benefit from, more liberal approaches with a slightly lower validation rate.

ACKNOWLEDMENTS

We would like to thank University of Rochester Center for Integrated Research Computing (CIRC) for providing computational resources. We would also like to thank University of Rochester Medical Center Genomics Research Center (GRC) for their support and Stephen L. Welle, Ph.D. for his guidance and suggestions in preparation of the manuscript.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

CC carried out the sequence alignment, differential expression analysis, and drafted the manuscript. SB carried out the differential expression analysis and drafted the manuscript. ZZ carried out qPCR validation. AML, VAL participated in cDNA synthesis and qPCR. MY, BWB participated in generating the animal data. MAO conceived of the study, and participated in its design. TJM conceived of the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS: **The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments**. *Nature protocols* 2012, **7**(8):1534-1550.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science* 2008, **320**(5881):1344-1349.
3. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis**. *Cell* 2008, **133**(3):523-536.
4. Cloonan N, Grimmond SM: **Transcriptome content and dynamics at single-nucleotide resolution**. *Genome biology* 2008, **9**(9):234.
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods* 2008, **5**(7):621-628.
6. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome research* 2008, **18**(9):1509-1517.
7. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing**. *BioTechniques* 2008, **45**(1):81-94.
8. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews Genetics* 2009, **10**(1):57-63.
9. Fonseca NA, Rung J, Brazma A, Marioni JC: **Tools for mapping high-throughput sequencing data**. *Bioinformatics* 2012, **28**(24):3169-3177.
10. Lindner R, Friedel CC: **A comprehensive evaluation of alignment algorithms in the context of RNA-seq**. *PloS one* 2012, **7**(12):e52403.
11. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nature methods* 2011, **8**(6):469-477.
12. Gambera D, Carta S, Crainz E, Fortina M, Maniscalco P, Ferrata P: **Metallosis due to impingement between the socket and the femoral head in a total hip prosthesis. A case report**. *Acta bio-medica : Atenei Parmensis* 2002, **73**(5-6):85-91.
13. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome biology* 2013, **14**(4):R36.
14. David M, Dzamba M, Lister D, Ilie L, Brudno M: **SHRiMP2: sensitive yet practical SHort Read Mapping**. *Bioinformatics* 2011, **27**(7):1011-1012.
15. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, Nielsen J: **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae**. *Nucleic acids research* 2012, **40**(20):10084-10097.
16. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC bioinformatics* 2013, **14**:91.
17. Benjamin AM, Nichols M, Burke TW, Ginsburg GS, Lucas JE: **Comparing reference-based RNA-Seq mapping methods for non-human primate data**. *BMC genomics* 2014, **15**(1):570.

18. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR *et al*: **A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data**. *PloS one* 2014, **9**(8):e103207.

19. Yee M, Vitiello PF, Roper JM, Staversky RJ, Wright TW, McGrath-Morrow SA, Maniscalco WM, Finkelstein JN, O'Reilly MA: **Type II epithelial cells are critical target for hyperoxia-mediated impairment of postnatal lung development**. *American journal of physiology Lung cellular and molecular physiology* 2006, **291**(5):L1101-1111.

20. Bhattacharya S, Zhou Z, Yee M, Chu CY, Lopez AM, Lunger VA, Solleti SK, Resseguie E, Buczynski B, Mariani TJ *et al*: **The genome-wide transcriptional response to neonatal hyperoxia identifies Ahr as a key regulator**. *American journal of physiology Lung cellular and molecular physiology* 2014, **307**(7):L516-523.

21. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(9):5116-5121.

22. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nature biotechnology* 2013, **31**(1):46-53.

23. Bhattacharya S, Go D, Krenitsky DL, Huyck HL, Solleti SK, Lunger VA, Metlay L, Srisuma S, Wert SE, Mariani TJ *et al*: **Genome-wide transcriptional profiling reveals connective tissue mast cell accumulation in bronchopulmonary dysplasia**. *American journal of respiratory and critical care medicine* 2012, **186**(4):349-358.

24. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer**. *Nature* 2009, **458**(7234):97-101.

Figure 1: Analysis Workflow. RNA-Seq raw reads were mapped using three separate methods; CASAVA, SHRiMP and TopHat. Mapped data were normalized using RPM or TM for all three mapping methods, or FPKM for SHRiMP and TopHat only. Normalized data were filtered and tested for differential expression using either SAM for all three mapping methods, or CuffDiff for SHRiMP and TopHat. A subset of the differentially expressed genes was assessed by qPCR.

Figure 2: Correlation between mapped reads of the sample across three mappers. RPM normalized counts from individual mappers for the same sample were plotted to identify the correlation between the expression levels. Shown here are dot plots of two samples (Control: RA 11+12 and Hyperoxia: O2 11+12) between TopHat and CASAVA counts (A:Control, D:Hyperoxia), SHRiMP and CASAVA (B:Control, E:Hyperoxia) and TopHat and SHRiMP (C:Control, F:Hyperoxia).

Figure 3: Summary of gene selection comparison. A comparison of genes with a fold-change greater than or equal to 2 identified as differentially expressed by SAM using CASAVA mapped data, and by Cuffdiff on SHRiMP and TopHat mapped data.

Table 1: Summary of mapping statistics.

	CASAVA	SHRiMP	TopHat
No of genes in genome	21718	23278	23278
Median number of genes mapped	16731	18674	17789
Median number of raw reads (x 10^6)	16.53	16.53	16.53
Median number of reads assigned to genes (x 10^6)	11.18	12.75	12.56
Median alignment rate (%)	NA	85.26	74.94
Median gene detection rate (%)	77.03	80.21	76.42

Table 2: Differential expression estimated by CASAVA.

CASAVA	SAM		
	RPM	TM	Overlap
Significant Genes	1017	1019	798
FC > 2	472	446	300
No. of up-regulated	210	180	132
No. of down-regulated	262	266	168
% of up-regulated	*0.44*	*0.40*	*0.44*
% of down-regulated	*0.56*	*0.60*	*0.56*

Table 3: Differential expression estimation by SHRiMP.

SHRiMP	SAM			Cuffdiff
	RPM	TM	Overlap	FPKM
Significant Gene	937	879	857	3886
FC > 2	401	386	379	1087
No. of up-regulated	170	157	153	623
No. of down-regulated	231	229	226	464
% of up-regulated	*0.42*	*0.41*	*0.40*	*0.57*
% of down-regulated	*0.58*	*0.59*	*0.60*	*0.43*

Table 4: Differential expression estimation by TopHat.

TopHat	SAM			Cuffdiff
	RPM	TM	Overlap	FPKM
Significant Gene	951	880	860	2831
FC > 2	418	412	396	1044
No. of up-regulated	179	172	165	593
No. of down-regulated	239	240	231	451
% of up-regulated	*0.43*	*0.42*	*0.42*	*0.57*
% of down-regulated	*0.57*	*0.58*	*0.58*	*0.43*

Table 5: qPCR validation rate. Genes were significantly different by either T-Test or MWU at p<0.05

Mapping Programs	No. of Gene Chosen	No. of Gene Validated	Validation Rate (%)	Validation Rate of Unique Gene (%)
CASAVA	32	23	71.87%	100%
TopHat	39	27	69.23%	70%
SHRiMP	38	24	63.16%	50%
Overlap	27	18	66.67%	

Figure 1(on next page)

Analysis Workflow.

RNA-Seq raw reads were mapped using three separate methods; CASAVA, SHRiMP and TopHat. Mapped data were normalized using RPM or TM for all three mapping methods, or FPKM for SHRiMP and TopHat only. Normalized data were filtered and tested for differential expression using either SAM for all three mapping methods, or CuffDiff for SHRiMP and TopHat. A subset of the differentially expressed genes was assessed by qPCR.

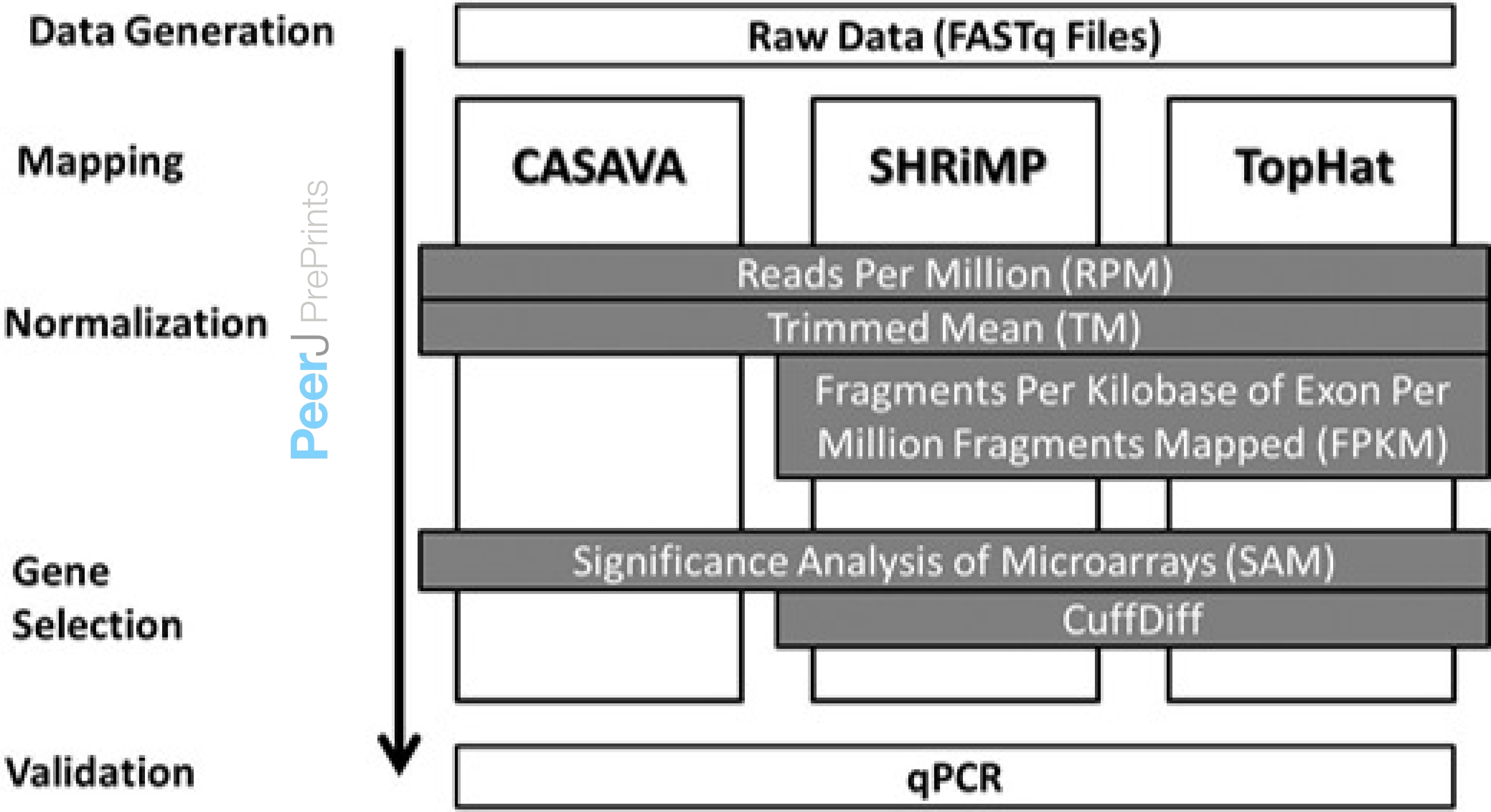

Data Generation
Raw Data (FASTq Files)
Mapping
CASAVA
SHRiMP
TopHat
Normalization
Reads Per Million (RPM)
Trimmed Mean (TM)
Fragments Per Kilobase of Exon Per Million Fragments Mapped (FPKM)
Gene Selection
Significance Analysis of Microarrays (SAM)
CuffDiff
Validation
qPCR

Figure 2(on next page)

Correlation between mapped reads of the sample across three mappers.

RPM normalized counts from individual mappers for the same sample were plotted to identify the correlation between the expression levels. Shown here are dot plots of two samples (Control: RA 11+12 and Hyperoxia: O2 11+12) between TopHat and CASAVA counts (A:Control, D:Hyperoxia), SHRiMP and CASAVA (B:Control, E:Hyperoxia) and TopHat and SHRiMP (C:Control, F:Hyperoxia). @0�

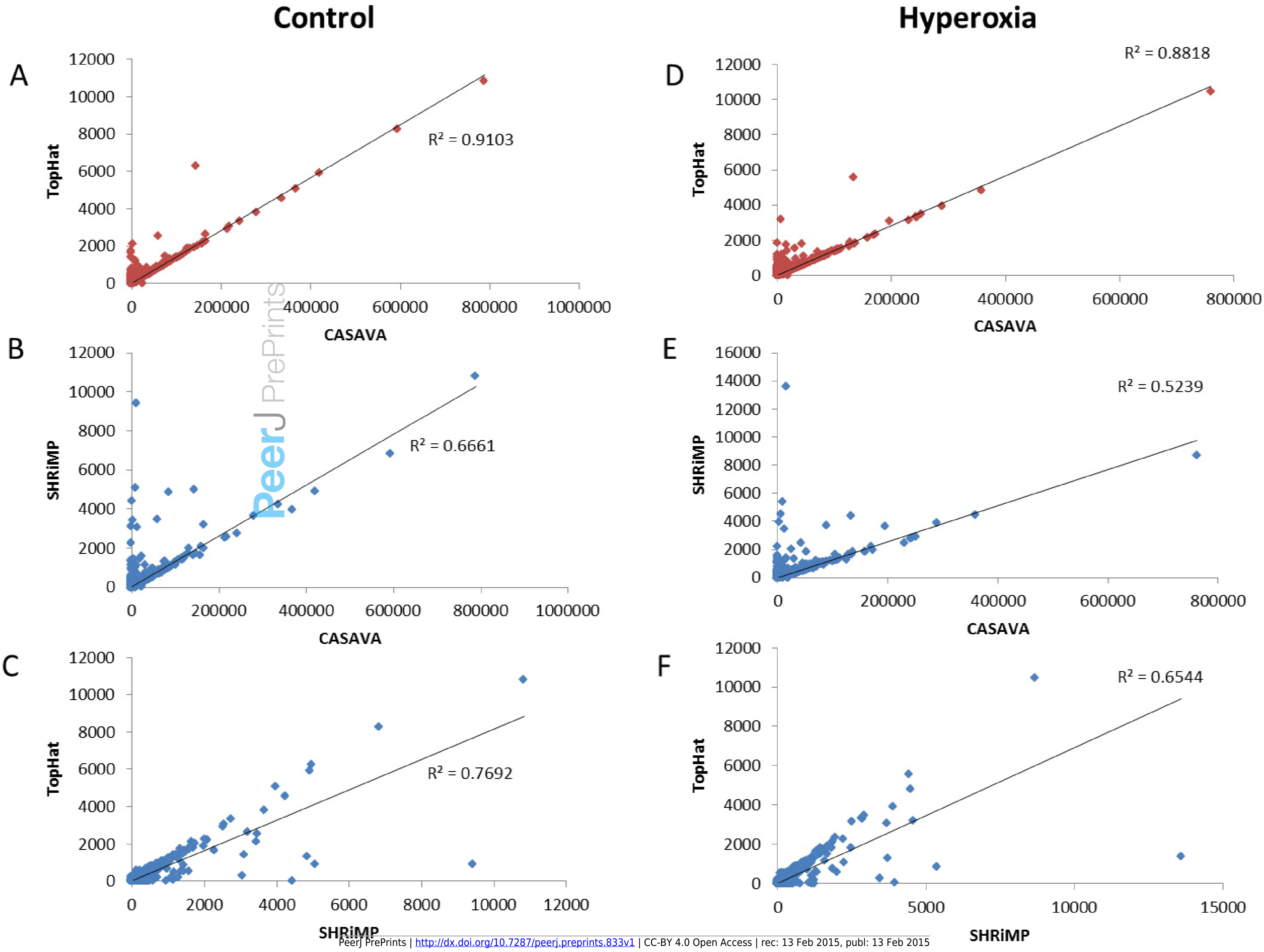
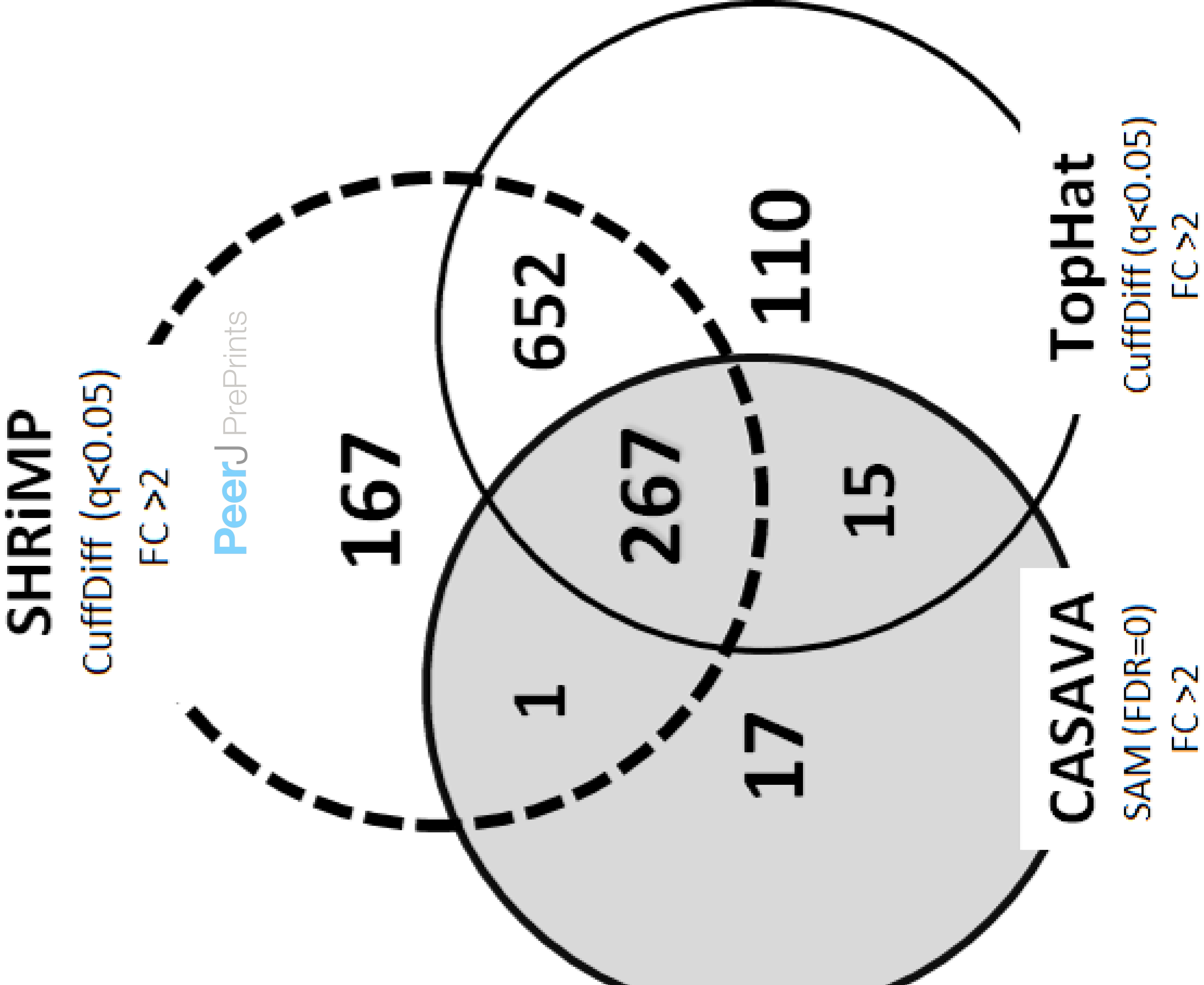

Control
Hyperoxia
A
TopHat
CASAVA
R² = 0.9103
12000
10000
8000
6000
4000
2000
0
0
200000
400000
600000
800000
1000000
B
SHRiMP
CASAVA
R² = 0.6661
12000
10000
8000
6000
4000
2000
0
0
200000
400000
600000
800000
1000000
C
TopHat
SHRiMP
R² = 0.7692
12000
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000
12000
D
TopHat
CASAVA
R² = 0.8818
12000
10000
8000
6000
4000
2000
0
0
200000
400000
600000
800000
E
SHRiMP
CASAVA
R² = 0.5239
16000
14000
12000
10000
8000
6000
4000
2000
0
0
200000
400000
600000
800000
F
TopHat
SHRiMP
R² = 0.6544
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000

Figure 3(on next page)

Summary of gene selection comparison.

A comparison of genes with a fold-change greater than or equal to 2 identified as differentially expressed by SAM using CASAVA mapped data, and by Cuffdiff on SHRiMP and TopHat mapped data.



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.833v1 | CC-BY 4.0 Open Access | rec: 13 Feb 2015, publ: 13 Feb 2015