Bioinformatics Advance Access published August 8, 2011

FDM: A Graph-based Statistical Method to Detect Differential Transcription using RNA-seq data

Darshan Singh [1]*, Christian F. Orellana [1], Yin Hu [5], Corbin D. Jones [2], Yufeng Liu [3], Derek Y. Chiang [4], Jinze Liu [5], Jan F. Prins [1]

[1]Department of Computer Science,[2]Department of Biology, [3]Department of Statistics and Operations Research,[4]Department of Genetics, University of North Carolina at Chapel Hill
[5]Department of Computer Science, University of Kentucky

Associate Editor: Dr. Alex Bateman

ABSTRACT

Motivation: In eukaryotic cells, alternative splicing expands the diversity of RNA transcripts and plays an important role in tissue-specific differentiation, and can be misregulated in disease. To understand these processes, there is a great need for methods to detect differential transcription between samples. Our focus is on samples observed using short-read RNA sequencing (RNA-seq).

Methods: We characterize *differential transcription* between two samples as the difference in the relative abundance of the transcript isoforms present in the samples. The magnitude of differential transcription of a gene between two samples can be measured by the square root of the Jensen Shannon Divergence (JSD*) between the gene's transcript abundance vectors in each sample. We define a weighted splice-graph representation of RNA-seq data, summarizing in compact form the alignment of RNA-seq reads to a reference genome. The Flow Difference Metric (FDM) identifies regions of differential RNA-transcript expression between pairs of splice graphs, without need for an underlying gene model or catalog of transcripts. We present a novel non-parametric statistical test between splice graphs to assess the significance of differential transcription, and extend it to group-wise comparison incorporating sample replicates.

Results: Using simulated RNA-seq data consisting of four technical replicates of two samples with varying transcription between genes, we show that (1) the FDM is highly correlated with JSD* ($r = 0.82$) when average RNA-seq coverage of the transcripts is sufficiently deep, (2) the FDM is able to identify 90% of genes with differential transcription when JSD* > 0.28, and coverage > 7. This represents higher sensitivity than Cufflinks (without annotations), and rDiff (MMD), which respectively identified 69% and 49% of the genes in this region as differential transcribed. Using annotations identifying the transcripts, Cufflinks was able to identify 86% of the genes in this region as differentially transcribed. Using experimental data consisting of four replicates each for two cancer cell lines (MCF7 and SUM102), FDM identified 1425 genes as significantly different in transcription. Subsequent study of the samples using qRT-PCR of several differential transcription sites identified by FDM, confirmed significant differences at these sites.

Availability: http://csbio-linux001.cs.unc.edu/nextgen/software/FDM

Contact: darshan@email.unc.edu

*to whom correspondence should be addressed

1 INTRODUCTION

The transcriptome is a key vantage point for a molecular biologist's study of phenotypic differences between cells that result from environmental factors, cell specialization, or disease. Classically this study has been conducted largely by observing differential gene expression levels using microarrays or high-throughput RNA sequencing technologies. However, detailed analysis of the transcriptome has shown that significant variation is also encoded in the diversity and relative abundance of a gene's constituent transcripts (Wang *et al.*, 2008; Sultan *et al.*, 2008; Kwan *et al.*, 2008). Consequently, beyond measuring differences in overall expression of genes between samples, there is a need to measure differences in expression at the transcript level.

We define *differential transcription* of a gene between samples as a difference in the relative abundance of the gene's transcript isoforms in the samples. In this manner, differential transcription is independent of the overall gene expression in the samples.

Short-read RNA sequencing technologies (RNA-seq) have evolved rapidly to sample the transcriptome at increasing depth and accuracy (Wang *et al.*, 2009). Using RNA-seq datasets obtained from samples, the locus and depth of coverage by reads aligned to a reference genome provide the starting point for the detection of differential transcription (Pan *et al.*, 2008).

Recently, two approaches have emerged to detect differential transcription between samples. The first approach is based on transcript inference and abundance estimation of the transcripts, as performed by tools like Cufflinks (Trapnell *et al.*, 2010), rQuant (Bohnert and Rätsch, 2010), Trans-Abyss (Robertson *et al.*, 2010), and Scripture (Guttman *et al.*, 2010). Applying these methods to each of two samples, differential transcription can be determined directly for each gene using the estimated relative abundances of the gene's transcripts in the two samples. However, transcript inference algorithms rely on heuristics to resolve the transcript structure because the inference problem is, in general, underdetermined. As a result, some transcripts may be missed or inferred incorrectly. Abundance estimation, in turn, is not able to correctly explicate the observed distribution of read alignments when starting from an incomplete or incorrect transcript model. Thus differential transcription measured in this fashion may be inaccurate.

The second approach to detect differential transcription is based on observing loci in the reference genome at which reads from

the two datasets align with different depth of coverage (after appropriate normalization for differing gene expression). The idea is that differential transcription should be revealed by different utilization of some exons. (Stegle *et al.*, 2010) describe two methods along these lines. The first is based on a priori analysis of annotated transcripts to identify regions that could reveal differential transcription. In each region a Poisson statistical test is applied. The second method is without dependence on known transcript structure, and uses a non-parametric kernel-based statistical test called Maximum Mean Discrepancy. Using synthetic data, both methods are shown by (Stegle *et al.*, 2010) to give accurate detection of differential transcription.

In this paper we introduce an approach that does not depend on annotations and instead leverages the splicing structure of a gene uncovered by spliced read alignments using tools like TopHat (Trapnell *et al.*, 2009), MapSplice (Wang *et al.*, 2010), or PALMapper (Jean *et al.*, 2010). Using the read alignments from these tools, a splice graph is constructed with edges corresponding to transcribed intervals or splices, weighted by read coverage. We introduce the *flow difference metric* (FDM) to measure the difference between two graphs in the relative utilization of edges at splicing points. Using synthetic samples, for which we know the transcripts and their relative abundances, we show the FDM between two samples is highly correlated with the JSD*, provided coverage of the edges is sufficient. Hence the FDM can serve as a metric of differential transcription, without need to infer the underlying transcripts, or need for any annotation.

To interpret the significance of the FDM we define a permutation test that can be efficiently implemented on the splice graph representation of the RNA-seq data. Since pairwise comparison of two samples is often insufficient to draw robust conclusions about differential transcription between two biological conditions, we extend the statistical test to incorporate replicates in each condition, when they are available. The test identifies differential transcription that is significant between conditions more often than it is significant within replicates.

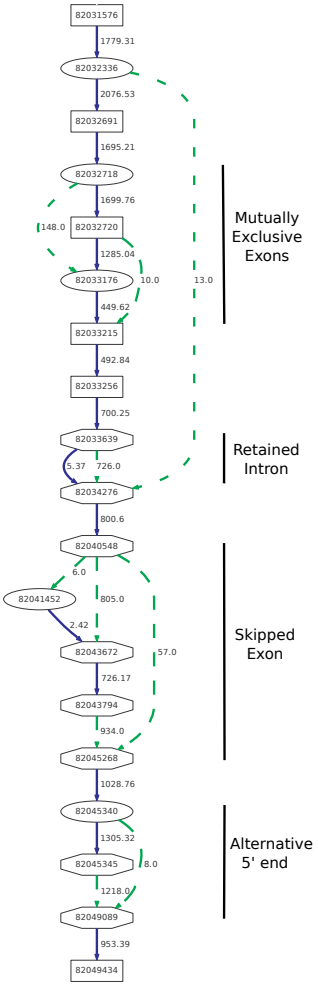

Fig. 1. ACT-Graph: The nodes are genome coordinates. A solid (blue) edge represents an exon or part of an exon labeled with the average depth of read coverage along the interval. A dashed (green) edge is a splice edge and is labeled by the number of reads that include the splice. Alternative splicing features such as mutually exclusive exons, a retained intron, and a skipped exon are illustrated. Nodes drawn as boxes, circles, and hexagons, respectively represent annotated-only positions, novel-only splice positions and both annotated and novel positions.

2 METHODS

2.1 Jensen-Shannon Divergence as a Measure of Differential Transcription

Let G be a gene with n different transcripts. In a given sample, the *transcript abundance vector* for G gives the relative abundance of each transcript isoform, i.e. the fraction of each isoform among all isoforms of G. One measure of differential transcription between two samples A and B, with transcript abundance vectors V_A and V_B, is the Jensen-Shannon Divergence

$$JSD(V_A, V_B) = H\left(\frac{V_A + V_B}{2}\right) - \frac{H(V_A) + H(V_B)}{2}$$

where $H(V)$ is the Shannon entropy. The JSD itself is not a metric, but JSD* $= \sqrt{\text{JSD}}$ does satisfy the properties of a metric.

We adopt JSD* to measure differential transcription in this paper, because it defines an objective measure of difference in transcript populations that is independent of the computational methods we examine. It has also been used to report differential transcription in other methods, e.g. CuffDiff (Trapnell *et al.*, 2010).

2.2 Aligned Cumulative Transcript Graph (ACT-Graph)

The alignment of RNA-seq reads to a reference genome provides (1) the genomic coordinates of transcribed bases and (2) the start and end coordinates of splices. As a consequence of alternative splicing, transcribed bases and splices may be part of multiple RNA transcripts and hence their coverage by aligned reads reflects their total utilization by all transcripts.

In the literature, transcripts have been mostly represented as paths in an acyclic directed graph with exons as nodes and splices as edges, e.g. (Heber *et al.*, 2002) and (Sammeth (2009) http://flux.sammeth.net/capacitor.html). Analyzing the read coverage information with this data structure has limitations. Firstly, this representation can only be used if all exons are known beforehand, which is usually not the case. Secondly, if two or more exons overlap in a region (e.g. in the case of alternative 5' donor sites or 3'

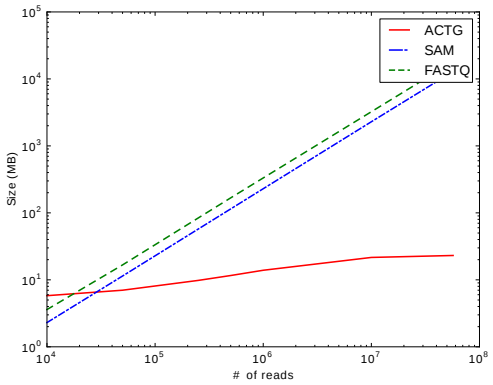

Fig. 2. ACT-Graph Compression (Section 2.2.2): Plot of file sizes of ACT-Graph (ACTG), FastQ file (FASTQ) and the alignment file (SAM). As the number of reads increases, the storage used by ACT-Graph increases orders of magnitude more slowly than other representations.

acceptor sites), the read coverage needs to be determined separately for each of those exons. Our graph representation addresses these limitations.

The ACT-Graph is a weighted directed acyclic multi-graph in which nodes correspond to genomic coordinates of splice start or end sites or to transcription start or end sites. Edges correspond to transcribed intervals (exonic edge) or to spliced-out intervals (splice edge). The weight of an exonic edge is its average coverage over the genomic interval it spans and the weight of a splice edge is the number of reads that include the splice. The direction of the edges is the direction of transcription. Each exonic edge is transcribed as whole, i.e. it is included in its entirety in a transcript or not at all.

In principle an ACT-Graph is the sum of weighted paths (flows), each of which is a transcript with some specific abundance. Therefore, we named the graph the Aligned Cumulative Transcript Graph (ACT-Graph). Figure 1 shows an example ACT-Graph. In practice, since reads are sampled non-uniformly from transcripts due to various biases, we use average coverage as an approximation of the total abundance.

2.2.1 ACT-Graph Construction. The following describes the step-by-step construction of an ACT-Graph from RNA-seq data:

1. Spliced Alignment: RNA-seq reads are aligned to the reference genome using a gapped aligner such as MapSplice (Wang *et al.*, 2010).
2. ACT-Graph nodes: The ACT-Graph nodes are created using one of the following: a) Splices: genomic coordinates of splice start and end locations are obtained from spliced alignments b) Interpreting start and end sites of transcripts: We can use inference or annotations to identify these sites. We can infer the start of a transcript based on the observation that the first (ℓ-1) bases following the start coordinate, where ℓ is the RNA-seq read length, show a characteristic ramp of increasing coverage as there are increasingly many ways for a read to sample bases further away from the start of a transcript. A transcription end site is inferred similarly. Alternatively transcript start and end coordinates can be taken from gene annotations, if available. Nodes introduced in this fashion are not harmful if the transcripts happen not to be expressed.
3. ACT-Graph edges and weights: a) A *splice edge* is inferred from a spliced alignment. The weight of the splice edge is the number of reads that support the splice. The direction of the edge is inferred from the orientation of the flanking bases in the intron for canonical splices or it can be inferred from the direction of other splices in the gene. b) An *exonic edge* connects two adjacent nodes (from the sorted list of nodes) if the genomic interval is fully covered or nearly fully covered and has an average coverage above threshold. We use a threshold of 1. The weight of an exonic edge is the average coverage of that genomic region. Averaging over the genomic region gives a better estimate of the number of transcripts that use that genomic region.
4. ACT-Graph genes and transcribed regions: a transcribed region is a connected component in the ACT-Graph when edges are considered as undirected, and typically would correspond to genes. If gene annotations are available, the regions can be restricted to known genes. The coverage of a gene is defined as average base coverage over all the bases of the exonic regions in the gene.

2.2.2 ACT-Graph Compressed Representation. The ACT-Graph is stored in the standard GFF format. The field TYPE tells if the line describes a node, a splice edge, or an exonic edge. The field SCORE is used for weight of the edges. The ACT-Graph format is a concise summary of alignments; and is powerful representation for quantitative analysis of alternative splicing. Figure 2 shows the compression achieved by the ACT-Graph representation as a function of the number of reads. The ACT-Graph is typically two to three orders of magnitude smaller than the SAM file or the raw reads, depending on the number of reads in the dataset and can be used for a number of downstream analyses, such as differential transcription.

2.2.3 Alternative Splicing Features in ACT-Graph. The ACT-Graph can be used to identify various alternative splicing features in a gene. Each alternative splicing feature can be represented by a subgraph which can be searched in the ACT-Graph. Figure 1 shows examples of various such features in a gene.

2.3 Flow Difference Metric

In this section, we describe the Flow Difference Metric, which uses the ACT-Graph to find genes with differential transcription. As stated earlier, the ACT-Graph can be viewed as the sum of weighted paths or flows, each of which corresponds to a transcript with some abundance. ACT-Graph nodes that have $m > 1$ incoming or outgoing edges indicate that at least m transcripts use that node. These nodes are called *divergence nodes*. Divergence nodes imply alternative splicing. The m incoming/outgoing edges are called the divergence edges. The weights of divergence edges signify the relative abundances of alternative transcripts passing through the divergence node. The normalized weights of all the divergence edges of a node are grouped together in a vector called the flow vector for the node. The difference between flow vectors in ACT-Graphs constructed from different samples indicates the magnitude of differential transcription between the two samples.

We measure the difference in flow vectors using a metric called the Flow Difference Metric (FDM) which is defined as follows. Assume an ACT-Graph has n divergence nodes. The flow vector for divergence node i of sample A is defined as $V_i^A = [e(a,i)_1, ..., e(a,i)_m]$ where m is the number of edges at node i and $e(a,i)_j$ is the normalized coverage at edge j, such that $\sum_{j=1}^{m} e(a,i)_j = 1$. The flow difference between samples A and B at divergence node i is

$$FD_i(A,B) = \sum_{j=1}^{m} |e(a,i)_j - e(b,i)_j|$$

The Flow Difference Metric (FDM) is computed as

$$FDM(A,B) = \frac{1}{2n} \sum_{i=1}^{n} (FD_i(A,B))$$

as illustrated in Figure 3.

It is important that ACT-Graphs of both samples have identical nodes and edges. If a node or edge is present in only one ACT-Graph, it is added to the other one with weight zero. The weights of exonic edges split by added nodes are re-computed using the alignments.

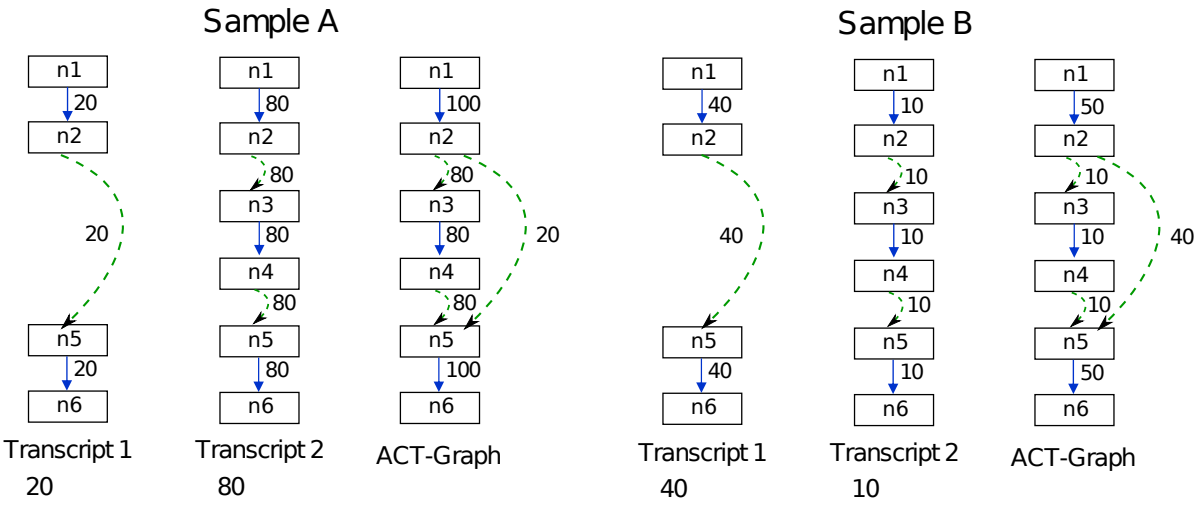

Fig. 3. FDM and JSD illustration: An example for a gene in two samples A and B is shown. The gene has two transcripts with expression ratio of 1:4 and 4:1 in the two samples, respectively. The FDM is computed using the two ACT-Graphs . The ACT-Graphs have 2 divergence nodes- node n2 has outdegree 2, and node n5 has indegree 2. $FDM(A,B) = \frac{1}{2n}(FD_{n2}(A,B) + FD_{n5}(A,B)) = \frac{1}{2 \cdot 2}((|0.8-0.2|+|0.2-0.8|)+(|0.8-0.2|+|0.2-0.8|)) = 0.6$. The JSD is computed using the ground truth knowledge of the transcript abundance vectors. V_A = [0.2,0.8] and V_B = [0.8,0.2]. $JSD(V_A, V_B) = H\left(\frac{V_A+V_B}{2}\right) - \frac{H(V_A)+H(V_B)}{2}$ = 0.28. Thus JSD*= 0.53 is the magnitude of differential transcription representing ground truth.

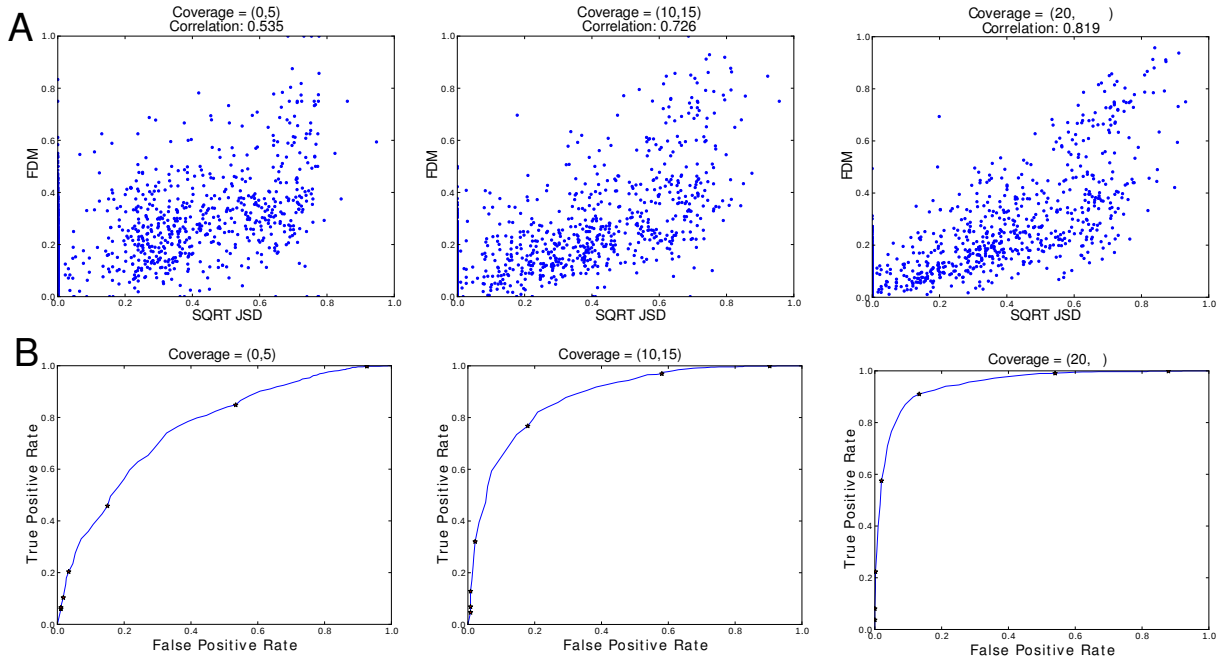

Fig. 4. Sensitivity and specificity of the FDM as a function of read coverage (Section 3.1.1 & 3.1.2) : Synthetic data of three sample pairs of 1500 genes each is analyzed. The first sample pairs have low gene coverage (coverage = [0,5]), the second sample pairs have medium gene coverage (coverage = [10,15]), and the third sample pairs have high gene coverage (coverage of 20 or higher). **(A)** JSD* - FDM Correlation: The points in the scatter plots correspond to (JSD*, FDM) values for a gene, where JSD* is ground truth and FDM is computed from ACT-Graphs. When the average gene coverage is high, the correlation between JSD* and FDM is high. For average coverage higher than 20, the correlation is 0.819. **(B)** FDM as a classifier for JSD*: a gene is marked positive for differential transcription if JSD* is more than 0.22 and negative otherwise. FDM is used to classify genes as positive or negative. Thus for each value of FDM, we get some true positives and some false positives. By varying FDM, the complete curve is plotted. The FDM values of (0.01,0.02,0.04,0.08,0.16,0.32.0.64) are marked on the curve. With coverage of 20 or higher, 90% of true positives can be identified with about 10% false positives.

2.3.1 FDM Properties

Lemma: The FDM is between 0 and 1

Lemma: FDM is a metric

1. FDM(A,B) ≥ 0
2. FDM(A,B) = 0 if and only if A = B
3. FDM(A,B) = FDM(B,A)
4. FDM(A,B) $\leq$ FDM(A,C) + FDM(B,C)

The proofs of both lemmas are in the supplementary materials.

2.3.2 FDM Usage. FDM may be applied between ACT-Graphs without need for normalization by the number of reads or read length, because the FDM is based on ratios of coverage, and these factors scale coverages linearly. Using synthetic data, we show that FDM has a high correlation with JSD*. The details of this are in section 3.1. Since we do not know the transcripts or their relative abundance, we use the FDM as a metric for differential transcription.

2.4 Statistical Tests for Differential Transcription

2.4.1 Statistical test to find genes with significant differential transcription. We use the FDM as a test statistic to find genes with significant differential transcription between two samples. The ACT-Graph of each gene is different, so the range of FDM values differs from one gene to another. Thus the FDM value for a gene is in itself not sufficient to tell if the differential transcription is significant. Instead, we devised a non-parametric test to determine whether differential transcription is significant. We create the null distribution of FDM for a gene, and test if the FDM value for the two samples has a significant p-value. The null hypothesis is that the gene has no differential transcription in two samples. The process of creating the FDM null distribution is illustrated in Figure 6 in the supplementary materials. Assume that there are N aligned reads in both the sample datasets. Create ACT-Graphs for the two samples such that nodes and edges are identical. The reads are partitioned into p equal-sized groups in both samples, and an ACT-Graph is created from the alignments of each group of N/p reads. Thus for each sample we have p ACT-Graphs. The $2p$ ACT-Graphs are randomly shuffled into two groups of p partitions each and a composite ACT-Graph for each group is created by simply adding the edge weights of the p ACT-Graphs in the group. Now the FDM is computed between ACT-Graphs of these two groups. This gives a value of the random variable which follows the null FDM distribution. By shuffling partitions a sufficient number of times, we get a null distribution of the FDM. In this fashion, the FDM null distribution is created for each gene, and the p-value for the specific partition that corresponds to the reads of the two samples can be computed. Section 1.5 in the supplementary materials provides details on sensitivity to the choice of p and the number of permutations.

2.4.2 Statistical test for multiple replicates. A single pairwise comparison is often insufficient to draw robust conclusions about differential transcription. Due to several uncontrolled factors, RNA-seq replicates may vary considerably more than predicted from sampling error alone. Thus, pairwise comparison between replicates may yield false positives. If we have multiple replicates of the two samples, we can apply one more level of permutation test to further filter the false positives. Let us assume that there are r replicates each of the two samples. Replicates from first sample are called group 1, and replicates from other sample are called group 2. The FDM pairwise statistical test can be applied to all $\binom{2r}{2}$ pairs. Out of those, r^2 pairs are between replicates in different groups, and the rest are between replicates in the same group. Now, if a gene has significant differential transcription *between groups* more often than *within groups*, it is likely to be true positive. The difference *between groups* and *within groups* is used as the test statistic. By permuting the group label of the replicates, we get the null distribution of the test statistic. The p-value of the statistic is computed for the original labeling and tested for significance.

3 RESULTS

3.1 Experiments with Simulated Data

In biological data we typically do not know the exact set of transcripts and their relative abundance in a sample, using which we could calculate the JSD*. Hence we use synthetic data, for which we know the exact transcript expression vectors for each gene, to determine (1) the correlation of the FDM and the JSD* metrics, (2) the power of the FDM method when used as a classifier for a particular value of JSD* under various levels of read coverage, and (3) the advantage of the groupwise significance test.

The RNA-seq dataset is simulated as follows. We use the annotated transcripts for human genome as a reference. Genes which have at least two transcripts are selected. Each of the genes is assigned an expression level sampled from an empirical distribution of observed expression levels in human genes. The individual transcripts of the genes are each assigned a relative abundance so that their sum is 1. The vector of relative abundances is called the *transcript expression vector*. For example, a gene with two transcripts T_1 and T_2 and a transcript expression vector of $[p_1, p_2]$ indicates that $p_1\%$ of transcripts are T_1 and $p_2\%$ are T_2. A read of size ℓ from a transcript is a random segment of size ℓ taken from the transcript sequence generated using the reference DNA. The number of reads generated from a transcript is proportional to the product of gene coverage, transcript expression and the length of the transcript. The alignment for every read is known, and hence the input SAM datasets consist of reads that are uniquely and perfectly aligned. Additional details on the datasets created can be found in the supplementary materials.

3.1.1 FDM correlation with JSD.* We create three pairs of simulated RNA-seq datasets each with different gene coverages. The three pairs of datasets have 1500 genes each. They are generated by varying gene coverages over three ranges - [0,5] , [10-15] and 20 or higher. The JSD* for the genes is varied over the range 0.0 to 1.0.

The ACT-Graph is created for all the genes for both the samples in the pair. The FDM is computed for each gene in the pair. From the transcript expression vectors of the genes, the JSD*, which represents the ground truth of differential transcription, is computed.

In Figure 4 we see that the correlation of FDM and JSD* increases as read coverage of the gene increases. This is as expected; when gene coverage is lower, the ACT-Graph edges will have lower weights. Since ratios are used, a small change in edge weight caused by random effects would affect the FDM considerably.

3.1.2 FDM as a classifier for JSD.* We tested if FDM can classify genes as high JSD* genes and low JSD* genes. We call a gene positive for high JSD* if the JSD* is greater than 0.22, and negative otherwise. This threshold is arbitrary; we obtained similar results for other values. For each gene, we create ACT-Graphs for two samples and compute the FDM. For a constant c, if FDM $> c$, we classify the gene as positive. Some of the positives are true positives (using JSD definition) and some false positives. For each c, we get true positives and false positives. By varying c from 0.01 to 0.99 over a step of 0.01, we get the complete ROC. Figure 4 shows that with high coverage, 90% of true positives can be identified with about 10% of false positives.

3.1.3 FDM method over synthetic replicates. We created two synthetic tissues over 2100 genes with at least two transcripts. The JSD* between genes in the two tissues varies randomly over the range 0.01 to 1.00. The distribution of JSD* and log(Coverage) are in Figure 1(e) and 1(f) respectively in supplementary materials. Four replicates were created for each of the tissues resulting in eight samples. FDM method was applied over all the $\binom{8}{2}$ pairs of which 16 pairs were *between group* and 12 were

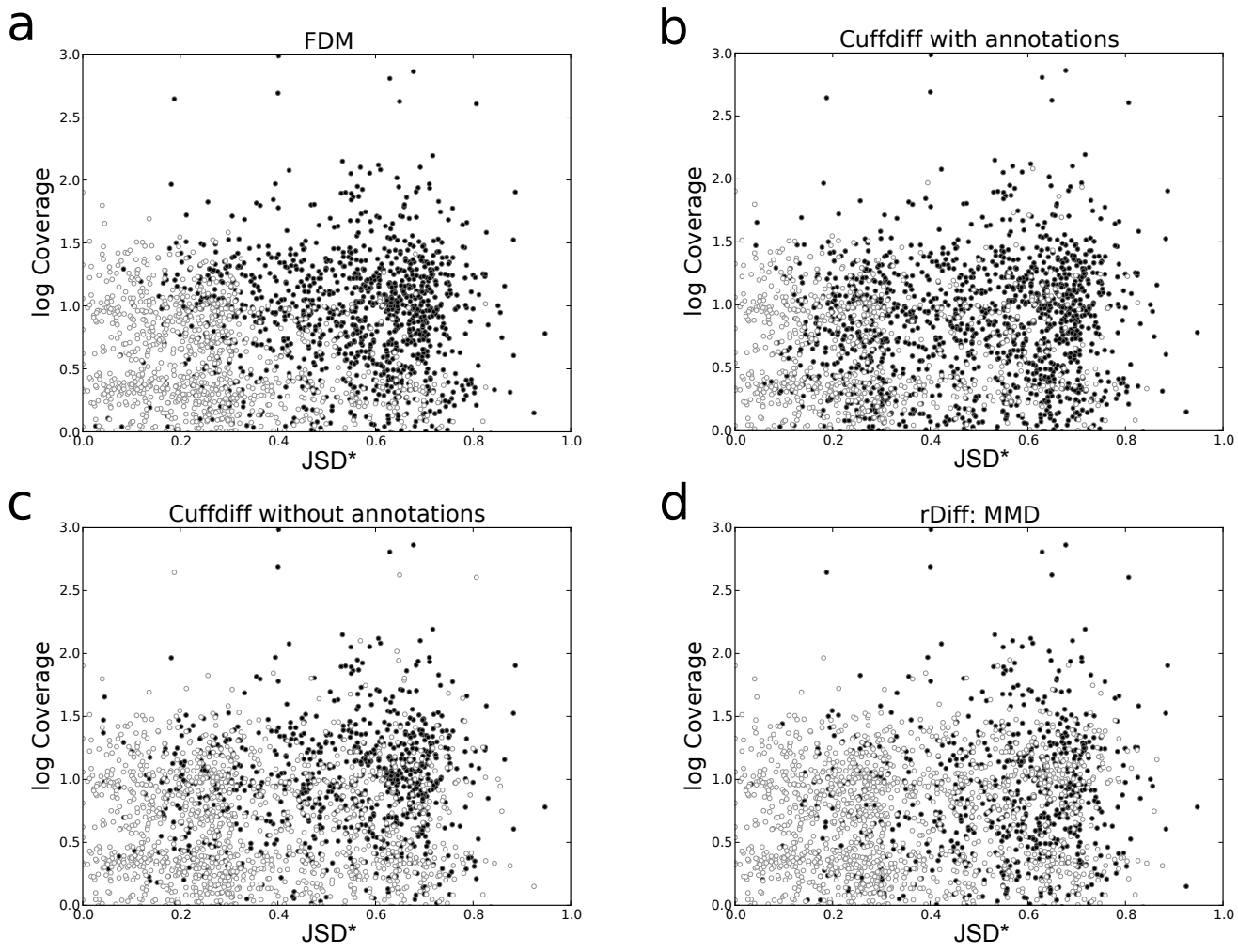

Fig. 5. Detection of differential transcription by different methods. The circles in scatterplots (a - d) represent 2100 genes in two samples with varying differential transcription (measured by JSD*) and varying depth of RNA-seq sampling (measured by the average coverage per transcribed nucleotide). Filled circles correspond to genes with significant differential transcription according to each of the methods. (a) FDM consistently identifies differential transcription when coverage is high or JSD* is high. For example, for genes with JSD* > 0.28 and log(coverage) > 0.85 (coverage > 7), FDM was able to identify 90% of the genes as differentially transcribed. Two other methods not using annotations, (c) Cuffdiff (without annotations), and (d) rDiff (MMD), had lower sensitivity, identifying differential transcription in 68%, and 49% of the genes in this region, respectively. (b) For comparison, when Cuffdiff with gene annotations identifies 86% of the genes in this region as differentially transcribed.

within group comparisons. We used p-value ≥ 0.05 as significant. For creating FDM null distribution, the number of partitions we used was 30 and the number of permutations was 1000. Section 1.5 in the supplementary materials shows that increasing the number of partitions and permutations has little effect on the results. The method finds 90 % of the genes which have JSD* > 0.28 and coverage > 7 as significant.

3.1.4 Comparison with other methods. The results of FDM were compared against other methods not using annotations, namely Cuffdiff (without annotations) and rDiff (MMD), using synthetic RNA-seq datasets defined in the previous section. We ran Cuffdiff as included in release 1.0.3 of the Cufflinks software. Since the data is synthetic and without sampling bias, we deactivated the bias correction module. We used the upper quartile normalization option in order to improve the accuracy of the abundance estimation. All genes with p-value ≤ 0.05 were marked as significant. We ran rDiff.web as provided in http://galaxy.tuebingen.mpg.de/. The only option available for the software is which method to use: we used the "MMD-based" method. All the genes with p-value ≤ 0.05 were marked as significant. The scatter plots in 5 show the results. For genes with JSD* > 0.28 and coverage > 7, FDM was able to identify 90% of the genes as differentially transcribed. This represents higher sensitivity than Cuffdiff (without annotation) and rDiff (MMD), which identified differential transcription between 68%, and 49% of the genes in this region, respectively. For comparison, we also ran Cuffdiff with gene annotations, which identified differential transcription in 86% of the genes in this region.

3.2 Experiments with Biological Data

We used RNA-seq data for 4 replicates each of the cancer cell lines MCF7 and SUM102. Each dataset has about 80 million single-ended reads of length 100 nucleotides.

We used the FDM method to find genes with differential transcription between SUM102 and MCF7. We used MapSplice to align the RNA-seq datasets. Using these alignments, we created ACT-Graphs for all the known genes. We applied the FDM statistical test to all the $\binom{8}{2}$ pairs of replicates. Out of these 28 pairs, 6 pairs were of MCF7-MCF7, another 6 for SUM102-SUM102, and 16 were MCF7-SUM102. The number of significantly different genes in single pair comparison are:

- MCF7-MCF7 : 1949 (average over 6 pairs)
- SUM102-SUM102: 1966 (average over 6 pairs)
- MCF7-SUM102: 2727 (average over 16 pairs)

Next we applied the statistical test for replicates to get the most significant genes. After applying the replicates statistical test, 1425 genes were judged to have significant differential transcription between MCF7 and SUM102. CD46 is one of the genes found to be significantly different. The UCSC browser bedgraph tracks for gene CD46 (Figure 6) shows that the middle exon has a different skipping ratio in MCF7 and SUM102. Additional examples can be found in the supplementary materials.

We performed qRT-PCR on three genes to validate the FDM results. Details for the method can be found in Section 1.4 of the supplementary materials. For CD46, the skipped exon (chr1:207963598-207963690) was found to be expressed more than two fold higher in SUM102 than in MCF7 as measured by qRT-PCR. Working from the ACT-Graphs, average skipping ratios in the MCF7 samples were 0.16 and in the SUM102 samples were 0.5 predicting an average 3.1 fold change. For NPC2 (shown in the supplementary materials), the retained intron (chr14:74946991-74947405) was expressed at least ten fold more in MCF7 than in SUM102 as measured by qRT-PCR. Working from the ACT-Graphs, an average fold change of 25 was predicted. Both experimental results were in congruence with the FDM results. Using Cuffdiff with annotations on our dataset, NPC2 was judged to have significant differential transcription, but the test for CD46 failed and thus was inconclusive.

A third gene ZNF408 (shown in the supplementary materials) gave a different result in the biological experiment than predicted by the FDM method. We directly resequenced cDNA derived from the mRNA from both cell lines and genomic DNA from both cell lines. The region of interest (chr11:46724721-46724734) has a high number of mutations in MCF7 compared to the reference genome, a common observation for cancer cell lines and cell lines that have been propagated extensively. This caused reads from a region of MCF7 to not align to the reference genome, and present a difference in the ACT-Graphs. Thus the incorrect result is due to alignment limitations, rather than to FDM.

4 DISCUSSION

4.1 FDM - JSD* correlation

Although Figure 4 shows a high correlation between FDM and JSD*, there still are genes with high FDM and low JSD*. These genes are artifacts of low coverage at some divergence nodes and could be filtered out. Since FDM uses ratios, a variance in small edge weights can cause high variance in the flow difference.

There are also some genes with high JSD* but low FDM. These can be due to complex gene models with many transcripts giving rise to many divergence nodes. When most transcripts have low abundance and are unchanged between samples and just a few similar transcripts have larger abundance changes, then JSD* can be large, yet only a few divergence nodes observe large flow changes, and these are attenuated by the remaining unchanged nodes to create an FDM value that is not exceptional under permutation testing. Focusing on divergence nodes with flow differences could improve detection of these cases.

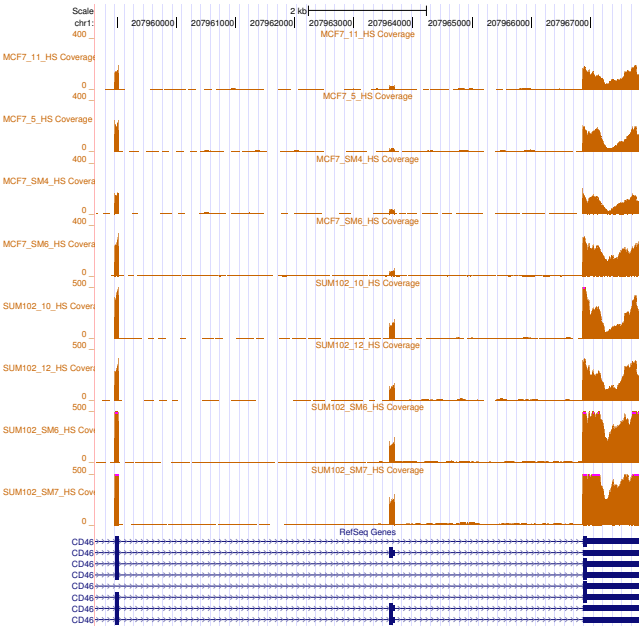

Fig. 6. UCSC browser: Gene CD46 in MCF7 and SUM102 (Section 3.2). The first 4 samples are from MCF7 and next 4 samples are from SUM102. This gene was identified as a gene with differential expression using FDM methodology. Note that the middle exon is skipped in different ratios in MCF7 and SUM102. This result was verifed by qRT-PCR. Additional figures in supplement.

4.2 FDM and sequencing bias

Sample preparation protocols can introduce significant deviations from the assumption of uniform sampling of reads along transcript isoforms, in ways which are not fully understood. It is useful to consider how such sampling bias would affect FDM. Roberts *et al.* (2011) cite two types of sampling bias.

Sequence-specific bias (Hansen *et al.*, 2010) is related to the underlying sequence of nucleotides in a transcript, resulting in preferential locations for read starts. Sequence-specific bias affects the count of reads whose alignment starts within an exonic edge in the ACT-Graph the same way for all transcripts utilizing the exonic edge. Associating average coverage with such an edge both smooths local variation due to sequence-specific bias, and is independent of the underlying transcripts involved. In effect, sequence-specific bias is minimized in this fashion.

Position-specific bias (Bohnert and Rätsch, 2010) is related to position in the transcript, and results in increased sampling at transcript starts and ends. Position-specific bias affects both exonic and spliced edge coverage according to the specific transcript utilizing the edge, and this will change as the relative abundance of transcripts changes, which will alter the magnitude of the flow difference in a divergence node. However, we have indicated that the magnitude of a gene's FDM signal varies by gene, and for this reason a non-parametric test is used to determine significance. Thus we believe the effect of position-specific bias will not substantially affect the determination of significance. In summary, while further investigation and validation is needed, we expect FDM to be largely insensitive to sequence-specific and position-specific sampling bias.

4.3 FDM and read length

The FDM method is specifically designed to detect differential transcription with short reads (35 - 100 nt), for which transcript reconstruction can be unreliable and, we would argue, is not needed. As we increase read length, read alignments become more accurate and the coverage on ACT-Graph edges increases, both of which improve the accuracy of the method. At the same time, if increased read length comes at the expense of deep sampling (under a fixed throughput assumption), then sensitivity would be expected to decrease.

Paired end reads can improve FDM accuracy depending on the operation of the underlying RNA-seq aligner. At the least, paired-end reads yield higher quality alignments because of the extra constraints on mate pair distance and alignment orientation. MapSplice aligns paired end reads using these constraints and also incorporates a maximum likelihood method operating on the splice graph to infer the alignment of the complete insert, including the unsequenced fragment, given the distribution of insert lengths (Hu *et al.*, 2010). This results in an effective increase in read length and coverage and hence can improve the accuracy of FDM.

5 CONCLUSION

While splice graphs were introduced nearly a decade ago (Heber *et al.*, 2002), our definition is intended to record RNA-seq read coverage in such a graph (this is also the approach taken in the Flux Capacitor). To make such graphs efficient to analyze, we choose a specific representation that differs from classic splice graphs. Nodes are labeled with genomic coordinates which are unique and help address the ambiguities caused by overlapping exons and unannotated genomic regions. The node labels are also well suited for computing the union of graphs from which the edge set for comparison of coverages is easy to determine. The ACT-Graph representation can dramatically decrease the data storage requirement for RNA-seq data. It is not a lossless compression as the underlying reads cannot be recovered from the ACT-Graph, but it does suffice for the analysis of differential expression and transcription.

The Flow Difference Metric captures the signal of differential transcription directly from a pair of ACT-Graphs, without knowledge or inference of the underlying transcripts, or need for normalization. The FDM has high correlation with JSD*, which is an independent measure of differential transcription. We showed that FDM can be used as classifier for differential transcription. We presented a statistical method using a permutation test on ACT-Graphs to find genes with significant differential transcription between pairs of samples or between groups of replicates.

ACKNOWLEDGEMENTS

We thank the referees for their insightful questions and comments, and thank Charles Perou for RNA samples from the MCF-7 and SUM-102 cell lines and Anais Monroy for qRT-PCR validation.

Funding: This work was supported by the National Science Foundation [ABI/EF grant number 0850237 to J.L. and J.F.P.]. Additional support was provided by the National Institutes of Health: NCI TCGA [grant number CA143848 to Charles Perou], and NCI GI SPORE Developmental Project Award [P50CA106991 to D.Y.C.], and an Alfred P. Sloan Foundation fellowship (D.Y.C.).

REFERENCES

Bohnert, R. and Rätsch, G. (2010). rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Research*, **38**(suppl 2), W348–W351.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, **28**(5), 503–510.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, **38**(12), e131.

Heber, S., Alekseyev, M., Sze, S.-H., Tang, H., and Pevzner, P. A. (2002). Splicing graphs and EST assembly problem. *Bioinformatics*, **18**(suppl 1), S181–S188.

Hu, Y., Wang, K., He, X., Chiang, D. Y., Prins, J. F., and Liu, J. (2010). A probabilistic framework for aligning paired-end RNA-seq data. *Bioinformatics*, **26**(16), 1950–1957.

Jean, G., Kahles, A., Sreedharan, V. T., Bona, F. D., and Rätsch, G. (2010). *RNA-Seq Read Alignments with PALMapper*, pages 32:11.6.1–32:11.6.37. Current Protocols in Bioinformatics. John Wiley & Sons, Ltd.

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T. J., Sladek, R., and Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. *Nature Genetics*, **40**(2), 225–231.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, **40**(12), 1413–1415.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J., and Pachter, L. (2011). Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**(3), R22.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A., and Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, **7**(11), 909–912.

Stegle, O., Drewe, P., Bohnert, R., Borgwardt, K., and Rätsch, G. (2010). Statistical tests for detecting differential rna-transcript expression from read counts. *Nature Precedings*.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, **321**(5891), 956–960.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9), 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5), 511–515.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221), 470–476.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, **38**(18), e178.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1), 57–63.

1 SUPPLEMENTARY MATERIALS

1.1 FDM Properties and Proofs

Lemma: The FDM is between 0 and 1
Proof: Let A and B be the two samples. For a given gene, assume that there are n divergence positions in the ACT-Graphs. Let V_i^A and V_i^B be the flow vectors for divergence node i for samples A and B respectively. Let $V_i^A = [e(a,i)_1, ..., e(a,i)_m]$.

Let $FD_i(A,B)$ be the flow difference at the divergence node i:

$$FD_i(A,B) = \sum_{j=1}^{m} |e(a,i)_j - e(b,i)_j|$$

Since absolute value is non-negative:

$$FD_i(A,B) = \sum_{j=1}^{m} |e(a,i)_j - e(b,i)_j| \geq 0 \quad (1)$$

Mathematically,

$$|e(a,i)_j - e(b,i)_j| \leq |(e(a,i)_j)| + |(e(b,i)_j)|$$

Thus,

$$\sum_{j=1}^{m} |e(a,i)_j - e(b,i)_j| \leq \sum_{j=1}^{m} |e(a,i)_j| + \sum_{j=1}^{m} |e(b,i)_j|$$

By definition,

$$\sum_{j=1}^{m} e(a,i)_j = 1; \sum_{j=1}^{m} e(b,i)_j = 1.$$

Also, since $e(a,i)_j$ and $e(b,i)_j$ are positive numbers,

$$FD_i(A,B) \leq 1 + 1 = 2 \quad (2)$$

By definition,

$$FDM(A,B) = \frac{1}{2n} \sum_{i=1}^{n} (FD_i(A,B))$$

From equations 1 and 2,

$$0 \leq FD_i(A,B) \leq 2$$

$$\frac{1}{2n} \cdot n \cdot 0 \leq \frac{1}{2n} \cdot \sum_{i=1}^{n} (FD_i(A,B) \leq \frac{1}{2n} \cdot n \cdot 2$$

$$0 \leq FDM(A,B) \leq 1$$

The FDM always lies between 0 and 1 irrespective of gene's size or number of constituent transcripts.

Lemma: FDM is a metric
Proof:

1. FDM(A,B) ≥ 0
2. FDM(A,B) = 0 if and only if A = B
   Proof: FDM will be zero if and only if $FD_i = 0$ at all the i divergence nodes. $FD_i = 0$ if and only if percent flow at each of the paths is exactly same. Please note that FDM will also be zero if one ACT-Graph has all the edge weights of the other ACT-Graph scaled up by the same factor. In that case also, the ACT-Graphs would represent the same transcripts with same relative abundances, though with different overall gene expression.
3. FDM(A,B) = FDM(B,A)
   Proof: FDM is sum of absolute differences, and absolute difference is commutative.
4. FDM(A,B) $\leq$ FDM(A,C) + FDM(B,C)
   Proof: For a divergence node i, let V_i^A be flow vector for A, V_i^B be flow vector for B and V_i^C be flow vector for C. Let $V_i^A = [e(a,i)_1, ..., e(a,i)_m]$. V_i^B and V_i^C also are similarly defined.

   $$FD_i(A,B) = \sum_{j}^{m} |e(a,i)_j - e(b,i)_j|$$

   Mathematically,

   $|e(a,i)_j - e(b,i)_j| \leq |e(a,i)_j - e(c,i)_j| + |e(b,i)_j - e(c,i)_j|$

   Thus

   $$FD_i(A,B) \leq FD_i(A,C) + FD_i(B,C).$$

   Summation over all divergence nodes gives

   $$FDM(A,B) \leq FDM(A,C) + FDM(B,C)$$

   Here, we assume that all the three ACT-Graph have same nodes and edges.

1.2 Simulated Data Results

Secns 3.1.1 and 3.1.3 in the main document describe two different experiments with different purposes. The synthetic data for the experiments was generated from a large space of potential inputs that can be tested for differential transcription. An input consists of a gene (selected from genes annotated with two or more transcript isoforms), a gene expression level (selected from an empirical distribution of gene expression levels) for each sample, and a relative abundance profile for the isoforms for each sample (also selected from an empirical distribution of profiles).

For the two experiments, different conditions determined the number of inputs (i.e. genes) to be tested. In the first experiment, the 3 intervals of coverage had different numbers of genes falling into each interval, and the goal was to have the same number of genes in each interval for fairness of comparison. Thus the number of genes in each interval was limited to 1500, approximately the fewest number in any interval. The total number of reads in this experiment was 100 million. Since these reads were generated in one run and the genes were separated according to interval of coverage, it is difficult to tell how many reads pertain to each of the three categories.

In the second experiment the goal was to limit the space of inputs to cover 3 orders of magnitude in gene expression levels (again, empirically determined). This resulted in 2100 genes for this experiment, and about 2.75 million 100 bp reads in each sample. The distribution of coverage values and JSD* values in the set of inputs is shown in 1 (c) and (d).

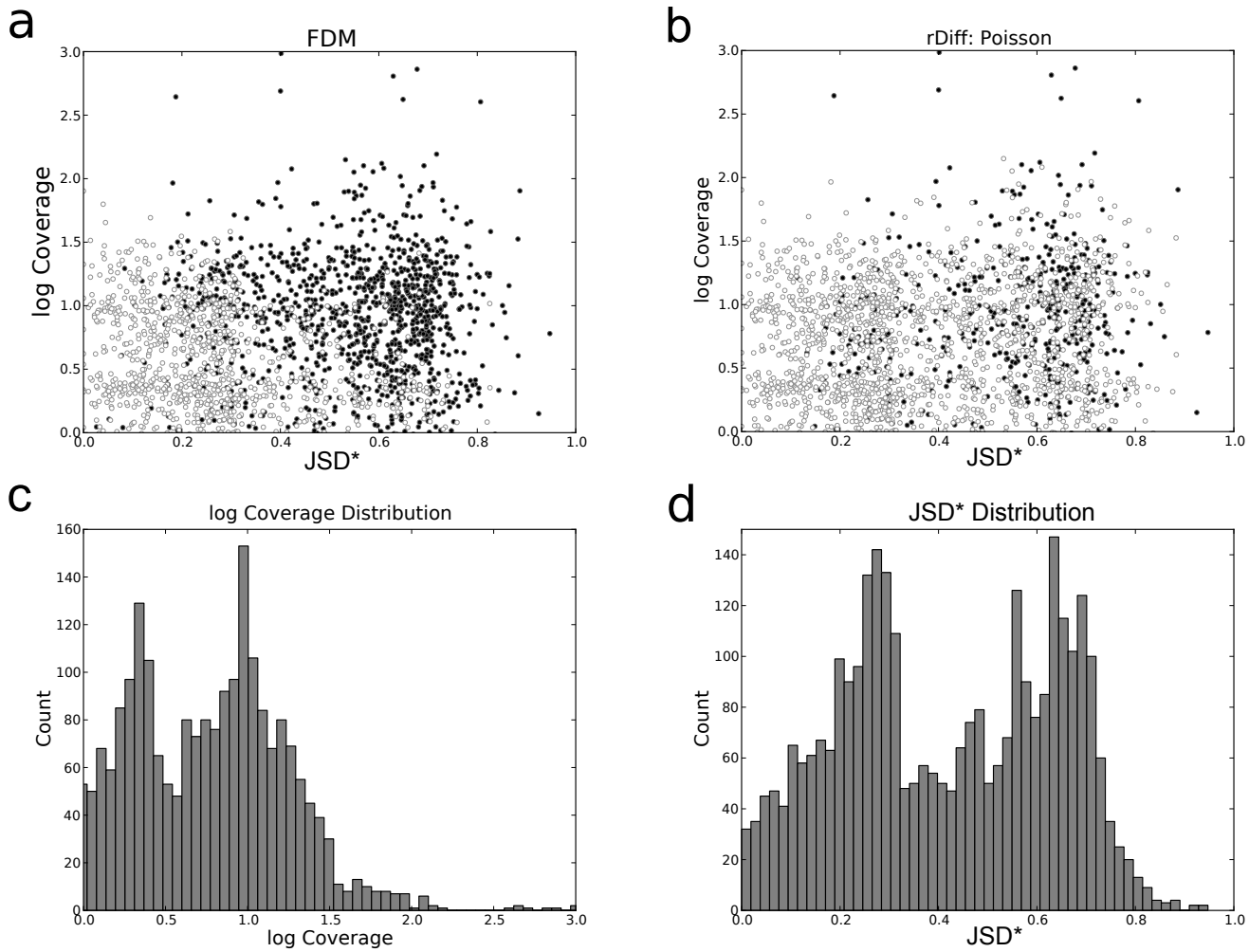

Fig. 1. The rDiff (Poisson) method using gene annotations is compared with FDM on the detection of differential transcription on our synthetic dataset with 2100 genes. For genes with JSD* $>$ 0.28 and log(coverage) $>$ 0.85, rDiff (Poisson) identified differential transcription between 34% of the genes. The histograms (c,d) are the distributions in our dataset of average coverage of the genes and JSD* respectively.

1.3 Biological data results

1.3.1 Examples of genes which are differentially transcribed in MCF7 and SUM102 Figures 2, 3 and 4 provide examples of differential transcription between two groups of samples. In each of the figures, the first four samples are from MCF7 cancer cell line MCF7 and the next four are from cancer cell line SUM102.

1.3.2 Example of gene where within-group differential transcription is also significant We observed that some genes have variation within replicates. The replicates statistical test filtered off such genes. Figure 5 gives example of one such gene.

1.4 qRT-PCR validation

RNA was isolated from the cell lines using standard Trizol protocol (Invitrogen, Inc.). Genomic DNA was isolated using PureGene DNA isolation kit (Qiagen, Inc). cDNA was made from the RNA with SuperScript cDNA synthesis kit (Invitrogen, Inc.) and oligo-dT primers (Bioneer, Inc). PCR was performed using reagents from New England Biolabs on an Eppendorf epGradient Mastercycler; qRT-PCR was performed with Bio-Rad Syber Green reagents on a C1000 five color thermocycler (Tm 54-55 C).

CD46 forward and reverse primers:
TACCTAACTGATGAGACCCACAGA and
AAGCAAACCTTTCTCTCATCTCTC.

NPC2 forward and reverse primers:
TAACCCTAGGGCAAGTTATCAGAC and
GGTTGAAGGAAAGAAGAGAGAGTG.

Sequencing of PCR products from cDNA and DNA was performed at the UNC Genomic Analysis Facility. Sequence cleanup was performed using 4peaks software (http://www.mekentosj.com/).

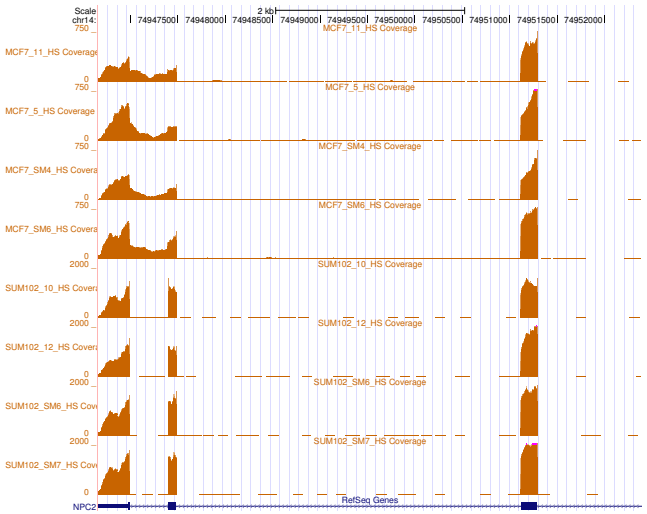

Fig. 2. NPC2: MCF7 shows evidence of first intron retention and second exon skipping. The first exon retention was confirmed by qRT-PCR

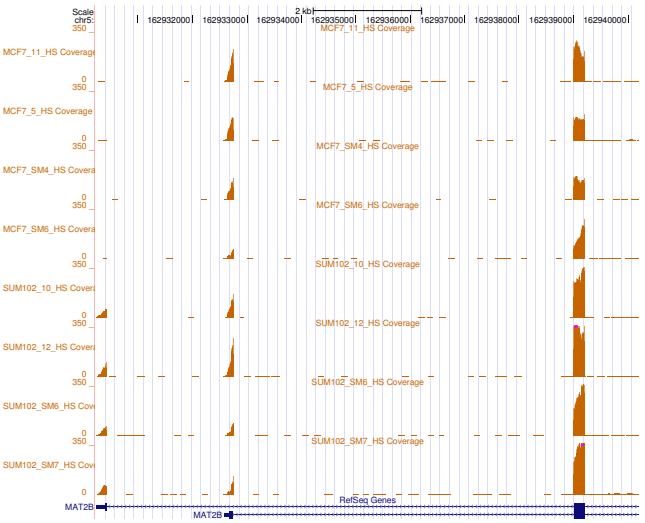

Fig. 4. MAT2B: First exon is different in SUM102 transcripts

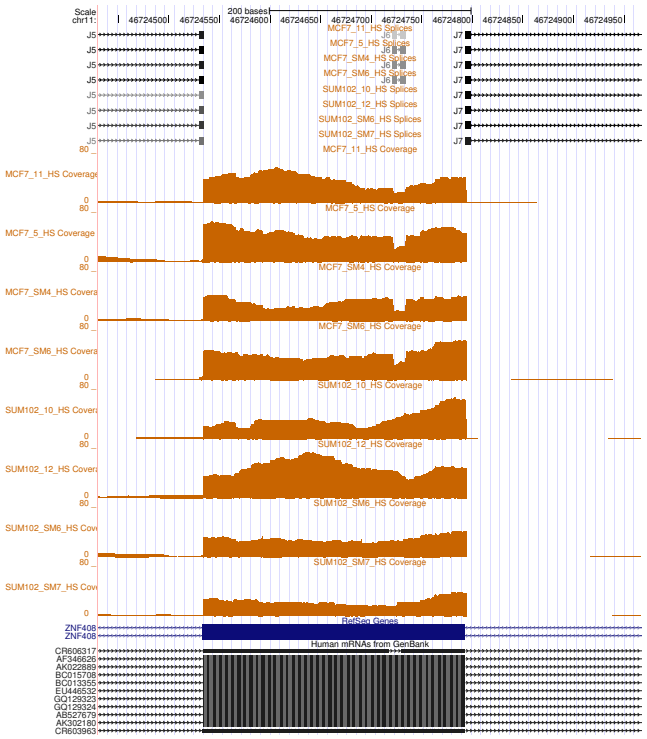

Fig. 3. ZNF408: MCF7 shows evidence of a transcript which doesn't occur in SUM102. This transcript uses the splice occuring only in MCF7. qRT-PCR could not confirm this result. We directly resequenced cDNA derived from the mRNA from both cell lines and genomic DNA from both cell lines. The region of interest (chr11:46724721-46724734) has a high number of mutations in MCF7 and SUM102 compared to the reference genome, a common observation for cell lines that have been propagated extensively. This caused errors in read alignments. FDM method uses read alignments as input. Incorrect input caused FDM method to give incorrect results

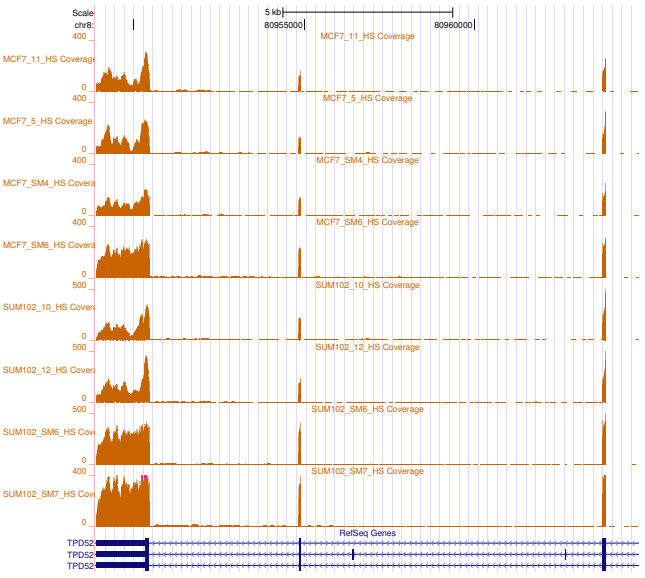

Fig. 5. TPD52: The middle exon is skipped in different ratios within MCF7 replicates and within SUM102 replicates also. FDM replicates statistical test rejected this gene as significant

Table 1. Parameters for FDM Runs

FDM Run	num partitions	num permutations	num output genes
Run 1	30	1000	1010
Run 2	30	1000	999
Run 3	30	2000	998
Run 4	30	2000	1007
Run 5	30	4000	1004
Run 6	30	4000	1001
Run 7	60	1000	1013
Run 8	120	1000	999

Table 2. Results by varying number of partitions

	Run 1 (1010)	Run 7 (1013)	Run 8 (999)
Run 1 (1010)		963 (95.3%)	962 (95.2%)
Run 7 (1013)	963 (95.0%)		957 (94.5%)
Run 8 (999)	962 (96.3%)	957 (95.8%)	

Each item in the cross tab shows the number of genes, and the percentage of genes common between the runs indicated by row and column headers. The parameters used in all the runs are given in Table 1.

Table 3. Results by varying number of permutations

	Run 1 (1010)	Run 3 (998)	Run 5 (1004)
Run 1 (1010)		956 (94.7%)	958 (94.9%)
Run 3 (998)	956 (95.8%)		957 (95.9%)
Run 5 (1004)	958 (95.4%)	957 (95.3%)	

Each item in the cross tab shows the number of genes, and the percentage of genes common between the runs indicated by row and column headers.The parameters used in all the runs are given in Table 1.

Table 4. Results by not varying any parameters

First Run	Second Run	Common Genes
Run 1 (1010)	Run 2 (999)	955 (94.6%)
Run 3 (998)	Run 4 (1007)	955 (95.7%)
Run 5 (1004)	Run 6 (1001)	952 (94.8%)

Each item in the cross tab shows the number of genes, and the percentage of genes common between the runs indicated by row and column headers.The parameters used in all the runs are given in Table 1.

1.5 Results by varying parameters for statistical test

We ran the FDM method on synthetic data for two tissues each having four replicates. All the samples had same set of 2600 genes. The FDM method was run multiple times by varying the two parameters - number of partitions and number of permutations. Table 1 describes the parameters used in the runs.

Table 2 shows that increasing the number of partitions beyond 30 had little effect on the results. The number of common genes in all pairs of runs with different number of partitions was around 95%. Since, the *p-value* was set to 5%, we expect to have 5% false positives in each run. Similarly, table 3 shows that increasing permutationsbeyond 1000 has little effect on the results. Running the FDM without varyingparameters gives similar results as shown in table 4.

1.6 FDM Statistical Test

The process of creating the FDM null distribution is illustrated in figure 6. Assume that there are N aligned reads in both the sample datasets. Create ACT-Graphs for the two samples such that nodes and edges are identical. The reads are partitioned into p equal-sized groups in both samples, and an ACT-Graph is created from the alignments of each group of N/p reads. Thus for each sample we have p ACT-Graphs. The $2p$ ACT-Graphs are randomly shuffled into two groups of p partitions each and a composite ACT-Graph for each group is created by simply adding the edge weights of the p ACT-Graphs in the group. Now the FDM is computed between ACT-Graphs of these two groups. This gives a value of the random variable which follows the null FDM distribution. By shuffling partitions a sufficient number of times, we get a null distribution of the FDM. In this fashion, the FDM null distribution is created for each gene, and the *p-value* for the specific partition that corresponds to the reads of the two samples can be computed.

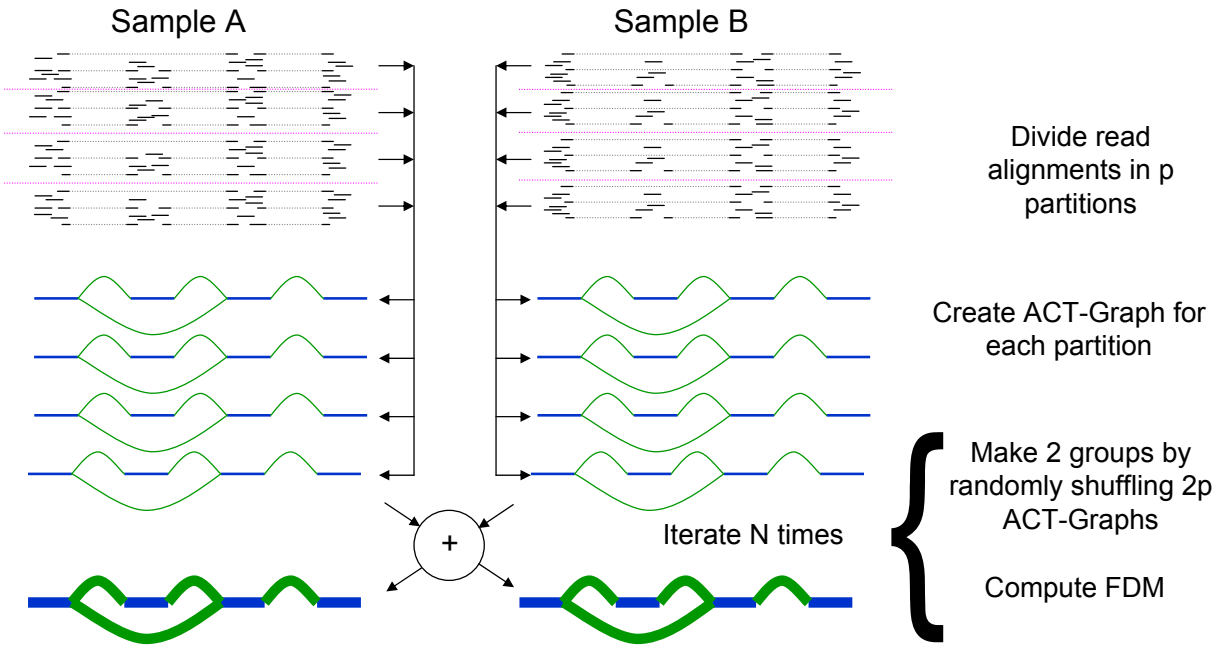

Fig. 6. FDM Statistical Test for a pair: The aligned reads for a gene are divided in p equal-sized partitions for both the samples. ACT-Graphs are created for each of the $2p$ partition that are randomly shuffled to make two groups of p partitions. The ACT-Graphs of each group is created by directly adding the edge weights of p ACT-Graphs. The FDM is computed for two ACT-Graphs. The last two steps are performed N times to get a null distribution for FDM for the gene. If the FDM of the original samples is significant over the null distribution, the gene as significant differential transcription in the pair. This process is performed for all the genes to find all the genes with significant differential transcription in the pair.