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).

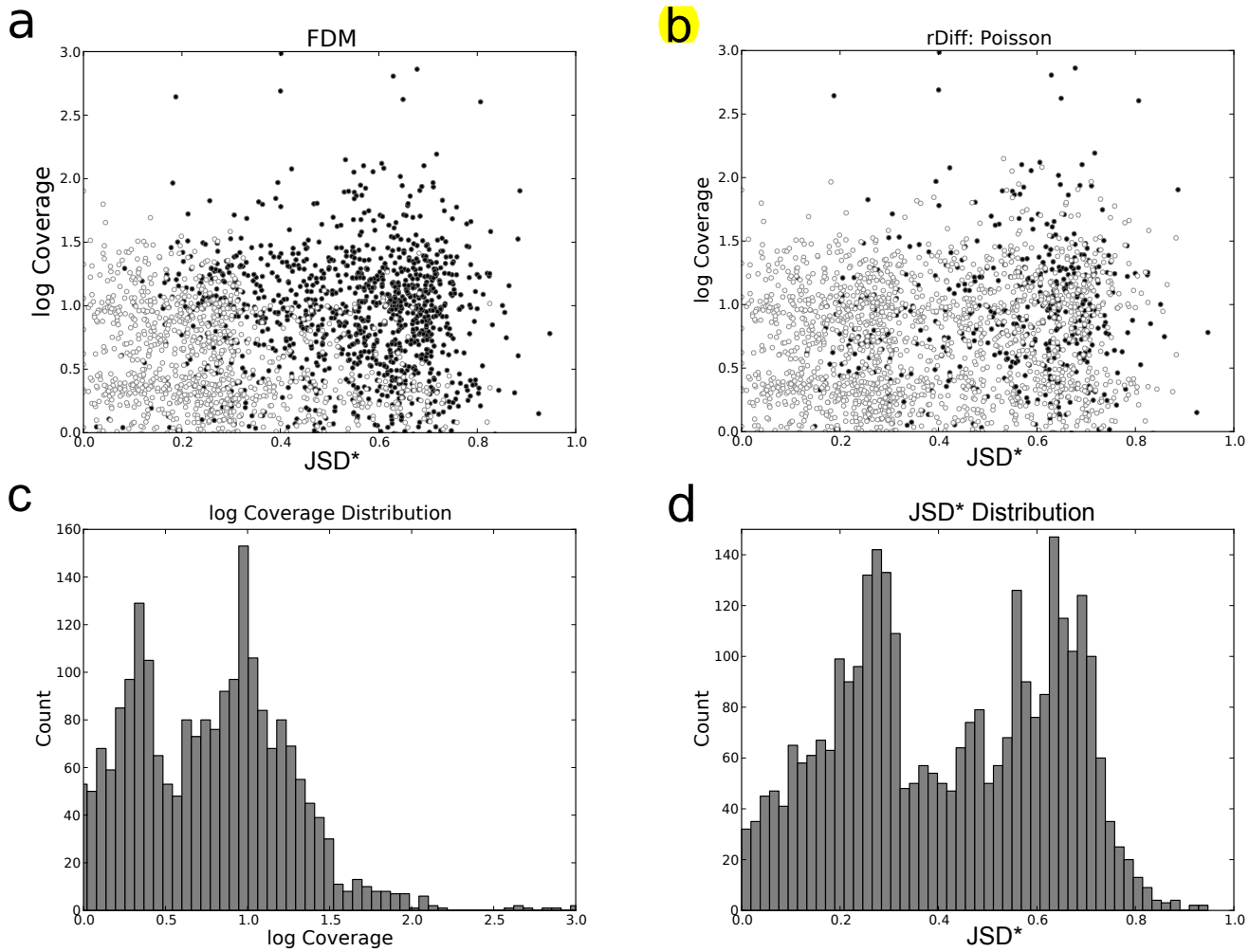

Fig. 1. The rDiff (Poisson) method using gene annotations is compared with FDM on the detection of differential transcription on our synthetic dataset with 2100 genes. For genes with JSD* $>$ 0.28 and log(coverage) $>$ 0.85, rDiff (Poisson) identified differential transcription between 34% of the genes.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

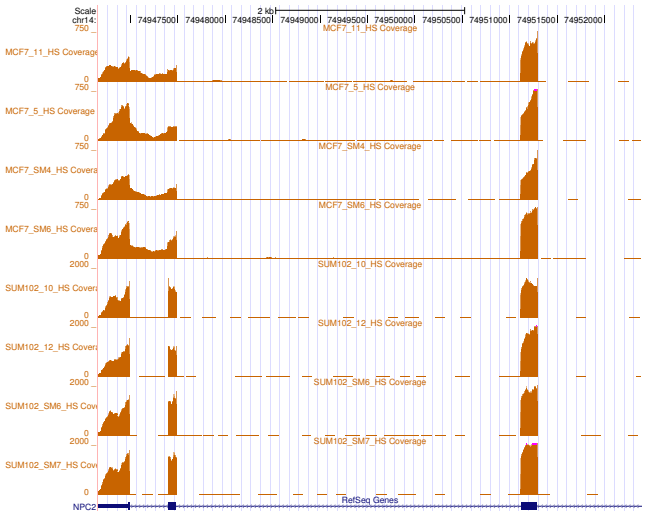

Fig. 2. NPC2: MCF7 shows evidence of first intron retention and second exon skipping. The first exon retention was confirmed by qRT-PCR

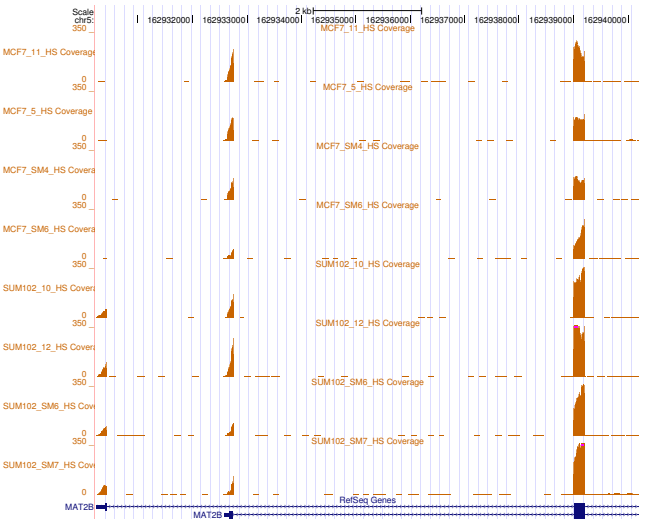

Fig. 4. MAT2B: First exon is different in SUM102 transcripts

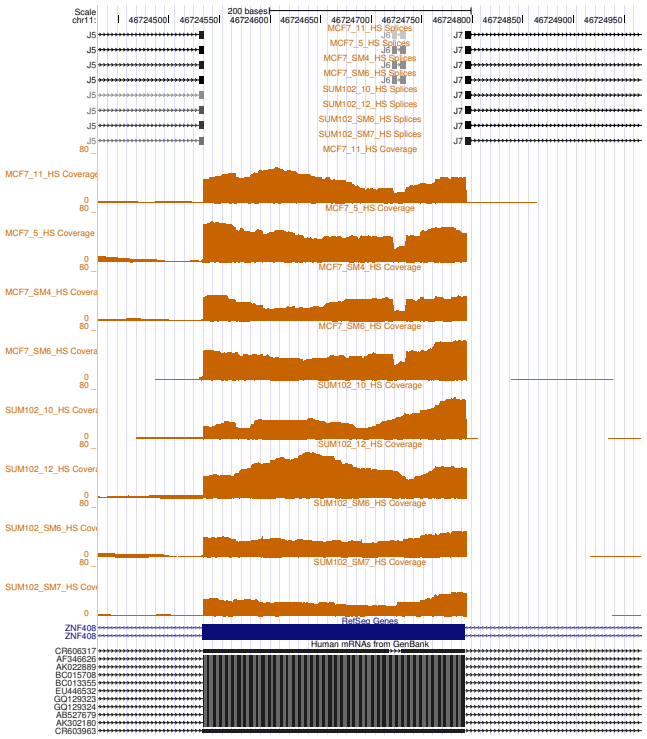

Fig. 3. ZNF408: MCF7 shows evidence of a transcript which doesn't occur in SUM102. This transcript uses the splice occuring only in MCF7. qRT-PCR could not confirm this result. We directly resequenced cDNA derived from the mRNA from both cell lines and genomic DNA from both cell lines. The region of 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

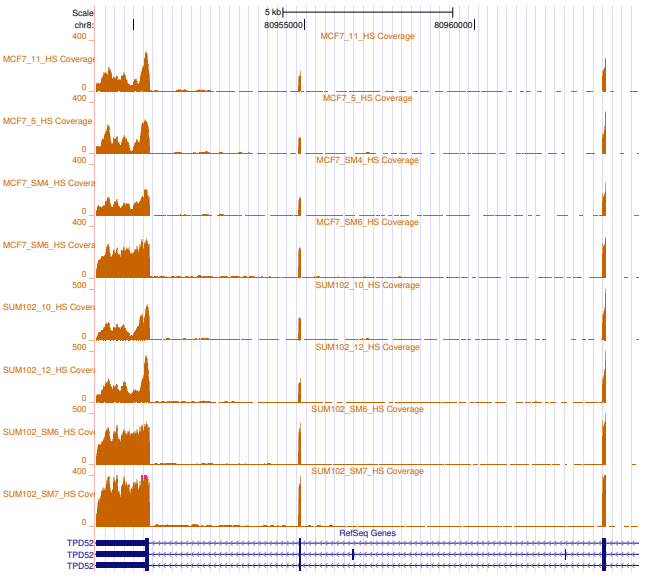

Fig. 5. TPD52: The middle exon is skipped in different ratios within MCF7 replicates and within SUM102 replicates also. FDM replicates statistical test rejected this gene as significant

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.

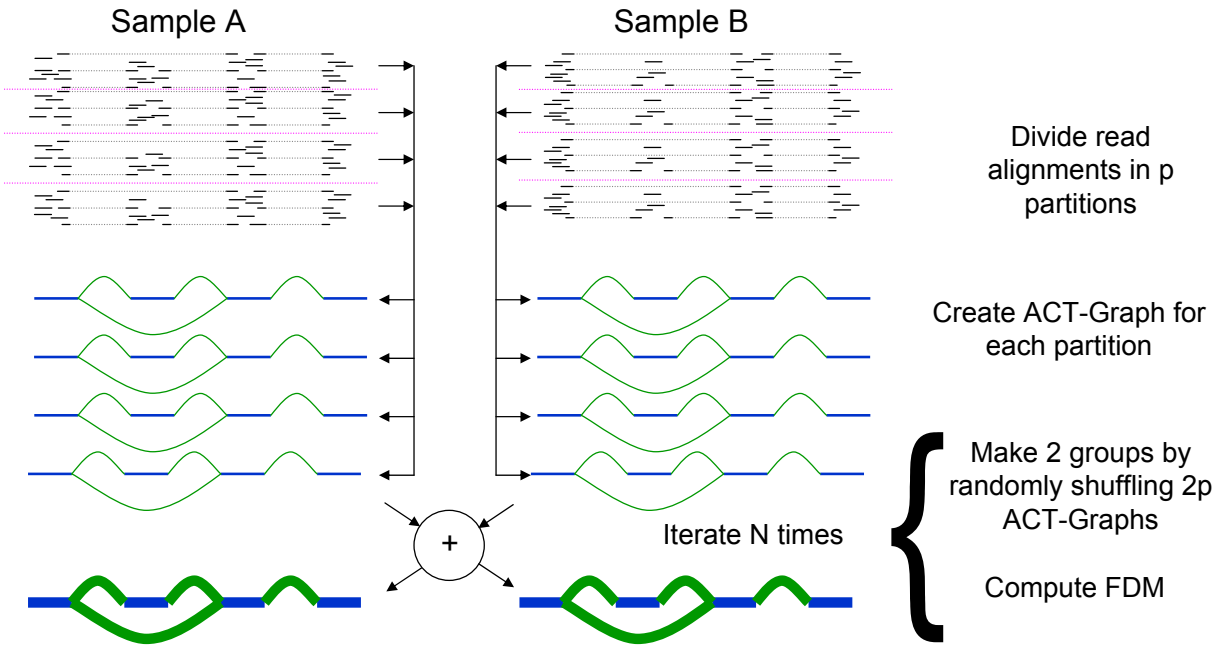

Fig. 6. FDM Statistical Test for a pair: The aligned reads for a gene are divided in p equal-sized partitions for both the samples. ACT-Graphs are created for each of the $2p$ partition that are randomly shuffled to make two groups of p partitions. The ACT-Graphs of each group is created by directly 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.