

RESEARCH ARTICLE

Network-Based Isoform Quantification with RNA-Seq Data for Cancer Transcriptome Analysis

Wei Zhang¹, Jae-Woong Chang², Lilong Lin³, Kay Minn⁴, Baolin Wu⁵, Jeremy Chien⁴, Jeongsik Yong², Hui Zheng³, Rui Kuang^{1*}

1 Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States of America, **2** Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States of America, **3** Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, Peoples Republic of China, **4** Department of Cancer Biology, University of Kansas Medical Center, Kansas City, Kansas, United States of America, **5** Division of Biostatistics, School of Public Health, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States of America

* kuang@cs.umn.edu



OPEN ACCESS

Citation: Zhang W, Chang J-W, Lin L, Minn K, Wu B, Chien J, et al. (2015) Network-Based Isoform Quantification with RNA-Seq Data for Cancer Transcriptome Analysis. *PLoS Comput Biol* 11(12): e1004465. doi:10.1371/journal.pcbi.1004465

Editor: Xianghong Jasmine Zhou, University of Southern California, UNITED STATES

Received: March 26, 2015

Accepted: July 11, 2015

Published: December 23, 2015

Copyright: © 2015 Zhang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The matlab source code is available at <http://compbio.cs.umn.edu/Net-RSTQ/>. The list of TCGA patient samples and GEO cell line samples used in the experiments are also provided through the URL.

Funding: WZ, JC and RK were supported by NSF grant III 1117153. <http://www.nsf.gov/>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

High-throughput mRNA sequencing (RNA-Seq) is widely used for transcript quantification of gene isoforms. Since RNA-Seq data alone is often not sufficient to accurately identify the read origins from the isoforms for quantification, we propose to explore protein domain-domain interactions as prior knowledge for integrative analysis with RNA-Seq data. We introduce a Network-based method for RNA-Seq-based Transcript Quantification (Net-RSTQ) to integrate protein domain-domain interaction network with short read alignments for transcript abundance estimation. Based on our observation that the abundances of the neighboring isoforms by domain-domain interactions in the network are positively correlated, Net-RSTQ models the expression of the neighboring transcripts as Dirichlet priors on the likelihood of the observed read alignments against the transcripts in one gene. The transcript abundances of all the genes are then jointly estimated with alternating optimization of multiple EM problems. In simulation Net-RSTQ effectively improved isoform transcript quantifications when isoform co-expressions correlate with their interactions. qRT-PCR results on 25 multi-isoform genes in a stem cell line, an ovarian cancer cell line, and a breast cancer cell line also showed that Net-RSTQ estimated more consistent isoform proportions with RNA-Seq data. In the experiments on the RNA-Seq data in The Cancer Genome Atlas (TCGA), the transcript abundances estimated by Net-RSTQ are more informative for patient sample classification of ovarian cancer, breast cancer and lung cancer. All experimental results collectively support that Net-RSTQ is a promising approach for isoform quantification. Net-RSTQ toolbox is available at <http://compbio.cs.umn.edu/Net-RSTQ/>.

Author Summary

New sequencing technologies for transcriptome-wide profiling of RNAs have greatly promoted the interest in isoform-based functional characterizations of a cellular system. Elucidation of gene expressions at the isoform resolution could lead to new molecular mechanisms such as gene-regulations and alternative splicings, and potentially better molecular signals for phenotype predictions. However, it could be overly optimistic to derive the proportion of the isoforms of a gene solely based on short read alignments. Inherently, systematical sampling biases from RNA library preparation and ambiguity of read origins in overlapping isoforms pose a problem in reliability. The work in this paper exams the possibility of using protein domain-domain interactions as prior knowledge in isoform transcript quantification. We first made the observation that protein domain-domain interactions positively correlate with isoform co-expressions in TCGA data and then designed a probabilistic EM approach to integrate domain-domain interactions with short read alignments for estimation of isoform proportions. Validated by qRT-PCR experiments on three cell lines, simulations and classifications of TCGA patient samples in several cancer types, Net-RSTQ is proven a useful tool for isoform-based analysis in functional genomes and systems biology.

Introduction

Application of next generation sequencing technologies to mRNA sequencing (RNA-Seq) is a widely used approach in transcriptome study [1–3]. Compared with microarray technologies, RNA-Seq provides information for expression analysis at transcript level and avoids the limitations of cross-hybridization and restricted range of the measured expression levels. Thus, RNA-Seq is particularly useful for quantification of isoform transcript expressions and identification of novel isoforms. Accurate RNA-Seq-based transcript quantification is a crucial step in other downstream transcriptome analyses such as isoform function prediction in the pioneer work in [4], and differential gene expression analysis [5] or transcript expression analysis [6]. Detecting biomarkers from transcript quantifications by RNA-Seq is also a frequent common practice in biomedical research. However, transcript quantification is challenging since a variety of systematical sampling biases have been observed in RNA-Seq data as a result of library preparation protocols [7–10]. Moreover, in the aligned RNA-Seq short reads, most reads mapped to a gene are potentially originated by more than one transcript. The ambiguous mapping could result in hardly identifiable patterns of transcript variants [10, 11].

A useful prior knowledge that has been largely ignored in RNA-Seq transcriptome quantification is the relation among the isoform transcripts by the interactions between their protein products. The protein products of different isoforms coded by the same gene may contain different domains interacting with the protein products of the transcripts in other genes. Previous studies suggested that alternative splicing events tend to insert or delete complete protein domains/functional motifs [12] to mediate key linkages in protein interaction networks by removal of protein domain-domain interactions [13]. The work in [4, 14] also suggested unique patterns in isoform co-expressions. Thus, the abundance of an isoform transcript in a gene can significantly impact the quantification of the transcripts in other genes when their protein products interact with each other to accomplish a common function as illustrated by a real subnetwork in Fig 1, which is constructed based on domain-domain interaction databases [15, 16] and Pfam [17]. Motivated by our observation that the protein products of highly co-expressed transcripts are more likely to interact with each other by protein domain-domain

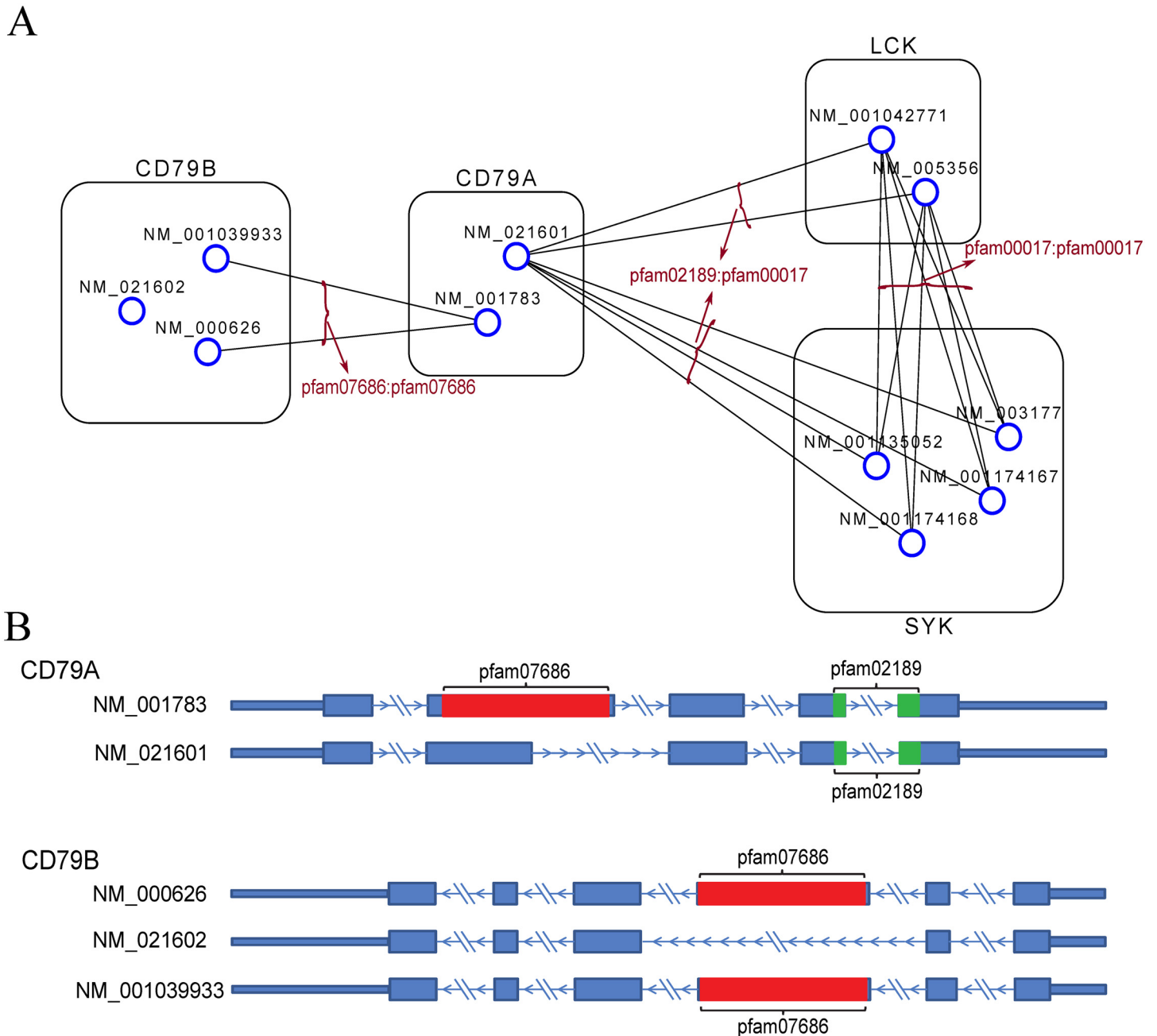


Fig 1. An isoform transcript network based on protein domain-domain interactions. (A) The subnetwork shows the domain-domain interactions among transcripts from four human genes, CD79B, CD79A, LCK and SYK. In the network, the nodes represent isoform transcripts, which are further grouped and annotated by their gene name; and the edges represent domain-domain interactions between two transcripts. Each edge is also annotated by the interacting domains in the two transcripts. (B) RefSeq transcript annotations of CD79A and CD79B are shown with Pfam domain marked in color. The Pfam domains were detected with Pfam-Scan software. Note that no interaction is included between transcripts NM_001039933 and NM_000626 of gene CD79B without assuming self-interactions for modeling simplicity. For better visualization, only the interactions coincide with PPI are shown in the figure.

doi:10.1371/journal.pcbi.1004465.g001

binding in four TCGA RNA-Seq datasets (see the section [Results](#)), we constructed two human transcript interaction networks of different sizes based on protein domain-domain interactions to improve transcript quantification. Based on the constructed transcript network, we propose a network-based transcript quantification model called Net-RSTQ to explore domain-domain

interaction information for estimating transcript abundance. In the Net-RSTQ model, Dirichlet prior representing prior information in the transcript interaction network is introduced into the likelihood function of observing the short read alignments. The new likelihood function of Net-RSTQ can be alternating-optimized over each gene with expectation maximization (EM). It is important to note that the Dirichlet prior from the neighboring isoforms play two possible roles. On one hand, for the isoforms in the same gene but with different interacting partners, the different prior information will help differentiate their expressions to reflect their different functional roles. On the other hand, for the isoforms in the same gene with the same interacting partners, the uniform prior assumes no difference in their functional roles and thus, promotes a smoother expression patterns across the isoforms. In both cases, the Dirichlet prior captures the functional variations/similarities across the isoforms in each gene as prior information for estimation of their abundance.

The paper is organized as following. In the section [Materials and Methods](#), we describe the procedure to construct protein domain-domain interaction networks, the mathematic description of the probabilistic model and the Net-RSTQ algorithm, qRT-PCR experiment design, and RNA-Seq data preparation. In the section [Results](#), we first demonstrate the correlation between protein domain-domain interactions and isoform transcript co-expressions across samples in four cancer RNA-Seq datasets from The Cancer Genome Atlas (TCGA) to justify using domain-domain interactions as prior knowledge. We then compared the predicted isoform proportions with qRT-PCR experiments on 25 multi-isoform genes in three cell lines, H9 stem cell line, OVCAR8 ovarian cancer cell line and MCF7 breast cancer cell line. Net-RSTQ was also applied to four cancer RNA-Seq datasets to quantify isoform expressions to classify patient samples by the survival or relapse outcomes. In addition, simulations were also performed to measure the statistical robustness of Net-RSTQ over randomized networks.

Materials and Methods

In this section, we first describe the construction of the transcript interaction network and review the base probabilistic model for transcript quantification with RNA-Seq data. We then introduce the network-based transcript quantification model (Net-RSTQ) by applying the protein domain-domain interaction information as prior knowledge. The notations used in the equations are summarized in [Table 1](#). At last, qRT-PCR experiment design and RNA-Seq data preparation are explained.

Transcript network construction

Two binary transcript networks were constructed by measuring the protein domain-domain interactions (DDI) between the domains in each pair of transcripts in four steps. First, the translated transcript sequences of all human genes were obtained from RefSeq [18]. Second, Pfam-Scan was used to search Pfam databases for the matched Pfam domains on each transcript with $1e-5$ e-value cutoff [17]. Note that only high quality, manually curated Pfam-A entries in the database were used in the search. Third, domain-domain interactions were obtained from several domain-domain interaction databases, and if any domain-domain interaction exists between a pair of transcripts, the two transcripts are connected in the transcript network. Specifically, 6634 interactions between 4346 Pfam domain families from two 3D structure-based DDI datasets (iPfam [15] and 3did [16]) inferred from the protein structures in Protein Data Bank (PDB) [19] were used in the experiments. Besides these highly confident structure-based DDIs, transcript interactions constructed from 2989 predicted high-confidence DDIs and 2537 predicted medium-confidence DDIs in DOMINE [20] were also included if the transcript interaction agrees with protein-protein interactions (PPI) in HPRD [21].

Table 1. Notations.

Notation	Description
N	total # of genes
\mathcal{T}	set of transcripts; T_{ik} is the k^{th} transcript of the i^{th} gene; \mathcal{T}_i denotes the transcripts of the i^{th} gene
l_{ik}	length of transcript T_{ik}
\mathcal{r}	set of reads; r_{ij} is the j^{th} read aligned to the i^{th} gene; \mathcal{r}_i is the read set aligned to the i^{th} gene
ρ_{ik}	the probability of a read generated by transcript T_{ik} in the i^{th} gene
\mathbf{P}_i	the probability of a read generated by transcript T_i in the i^{th} gene, specifically, $[\rho_{i1}, \dots, \rho_{i, \mathcal{T}_i }]$
\mathbf{P}	concatenate of all \mathbf{P}_i , specifically, $[\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_N]$
ρ_{ik}	relative abundance of the transcript T_{ik} in the i^{th} gene
$\boldsymbol{\pi}$	transcript expression; π_{ik} is the expression of the k^{th} transcript of the i^{th} gene
ϕ_{ik}	average expressions (normalized) of transcript T_{ik} 's neighbors in the transcript network
$\boldsymbol{\alpha}$	parameters of Dirichlet distribution; $\alpha_{ik} = \lambda\phi_{ik} + 1$ is the parameter of the Dirichlet distribution of ρ_{ik}
q_{ijk}	read sampling probability, $q_{ijk} = \frac{1}{l_{ik}-l_r+1}$ if read r_{ij} is aligned to transcript T_{ik} , otherwise $q_{ijk} = 0$
\mathbf{S}	binary matrix for transcript interaction network

doi:10.1371/journal.pcbi.1004465.t001

In the experiments, we focused on the transcripts from two cancer gene lists from the literature for better reliability in annotations. The first smaller transcript network consists of 11736 interactions constructed from the 3D structure-based DDIs and 421 interactions constructed from the predicted DDIs among the 898 transcripts in 397 genes from the first gene list [22]. The second larger transcript network contains 711,516 interactions constructed from the 3D structure-based DDIs among 5599 transcripts in 2551 genes in a larger gene list [23]. Since inclusion of the predicted DDIs results in a much higher density in the large network, the large network does not include predicted DDIs to prevent too many potential false positive interactions. The characteristics of the two transcript networks are summarized in Table 2. The density of the two networks are 3.02% and 4.54% respectively, which are in similar scale with the PPI network. Both networks show high clustering coefficients, suggesting modularity of sub-networks. Note that self-interactions (interactions between transcript(s) in the same gene) are not considered since Net-RSTQ only utilizes positive correlation between the expressions of neighboring transcripts in different genes. For simplicity, Net-RSTQ assumes that self-interactions will not change the transcript quantification of an individual gene in the model.

In Fig 1(A) a subnetwork of the transcripts in gene CD79A and CD79B with their direct neighbors in the small transcript network is shown. The RefSeq transcript annotations of CD79A and CD79B are shown in Fig 1(B). In CD79A transcript NM_001783 contains an extra domain pfam07686 while transcript NM_021601 only contains a shorter hit pfam02189. Note that pfam02189 also has the same hit in NM_001783 with an e-value larger than 1e-5. In CD79B transcripts NM_001039933 and NM_000626 contain a domain pfam07686, which is removed in alternative splicing of NM_021602. In the transcript subnetwork shown in Fig 1(A), the transcripts in CD79A or CD79B have different interaction partners in the network. In the transcripts in CD79A, the expression of NM_021601 will correlate with the transcripts in LCK

Table 2. Network characteristics.

	# of Gene	# of Transcripts	# of Interactions	Density	Diameter	Avg. # of Neighbors	Avg. Cluster Coefficients
Small Network	397	898	12157	3.02%	9	27.08	0.3578
Large Network	2551	5599	711516	4.54%	9	254.16	0.5255

doi:10.1371/journal.pcbi.1004465.t002

and SYK, and NM_001783 will correlate with two transcripts in CD79B. The isoform transcripts in LCK and SYK show no different DDIs suggesting there is no functional variation by protein bindings and more similar expression patterns are potentially expected as prior knowledge.

Base model for transcript quantification

We first consider the method proposed in [24, 25] as the base model for quantification of the transcripts in a single gene. Let T_i denote the set of the transcripts in the i th gene and T_{ik} be the k th transcript in T_i . The probability of a read being generated by the transcripts in T_i is modeled by a categorical distribution specified by parameters p_{ik} , where $\sum_{k=1}^{|T_i|} p_{ik} = 1$ and $0 \leq p_{ik} \leq 1$. For the set of the reads r_i aligned to gene i , we consider the likelihood of that each of the $|r_i|$ short reads is sampled from one of the transcripts to which the read aligns. Specifically, for each read r_{ij} aligned to transcript T_{ik} , the probability of obtaining r_{ij} by sampling from T_{ik} , namely $Pr(r_{ij}|T_{ik})$ is $q_{ijk} = \frac{1}{l_{ik}-l_r+1}$ [8, 26, 27], where l_r is the length of the read. Assuming each read is independently sampled from one transcript, the uncommitted likelihood function [24] to estimate the parameters P_i from the observed read alignments against gene i is

$$\mathcal{L}(P_i; r_i) = Pr(r_i|P_i) = \prod_{j=1}^{|r_i|} Pr(r_{ij}|P_i) = \prod_{j=1}^{|r_i|} \sum_{k=1}^{|T_i|} Pr(T_{ik}|P_i) Pr(r_{ij}|T_{ik}) = \prod_{j=1}^{|r_i|} \sum_{k=1}^{|T_i|} p_{ik} q_{ijk}. \quad (1)$$

This likelihood function is concave but it may contain plateau in the likelihood surface. Therefore, Expectation Maximization (EM) is then applied to obtain the optimal P_i . In the EM algorithm, the expectation of read assignments to transcripts were estimated in the E-step and the likelihood function with the expected assignments can be maximized in the M-step to estimate P_i . The relative abundance of the transcript T_{ik} in gene i , ρ_{ik} , can be derived from

$$\rho_{ik} = \frac{p_{ik}}{\sum_{k=1}^{|T_i|} \frac{p_{ik}}{l_{ik}}}, \quad (2)$$

and the transcript expressions in gene i , π_{ik} , can be calculated by

$$\pi_{ik} = \frac{|r_i| p_{ik}}{l_{ik}}. \quad (3)$$

The base model is applied independently to each individual gene and no relation among the transcripts is considered.

Network-based transcript quantification model

In the Net-RSTQ model, the transcript interaction network S based on protein domain-domain interactions is introduced to calculate a prior distribution for estimating P jointly across all the genes and all the transcripts. The model assumes that the prior distribution of P_i is a Dirichlet distribution specified by parameters α_i and each α_{ik} is proportional to the read count by average expression of the transcript T_{ik} 's neighbors in the transcript network S . The prior read count ϕ_{ik} is defined as follows,

$$\phi_{ik} = l_{ik} \left(\pi' \frac{S_{*,(i,k)}}{\sum (S_{*,(i,k)})} \right), \quad (4)$$

where $S_{*,(i,k)}$ is a binary vector represents the neighborhood of transcript T_{ik} in transcript network S and $\sum (S_{*,(i,k)})$ is the size of the neighborhood. The calculation of each ϕ_{ik} is illustrated in

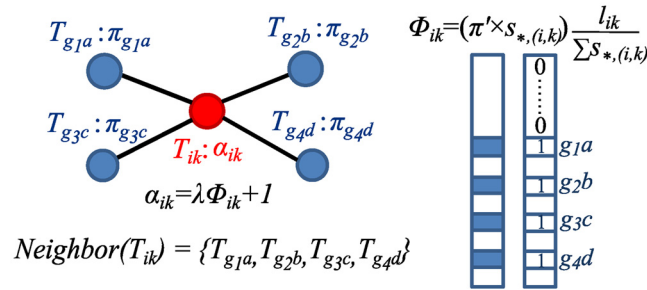


Fig 2. Transcript interaction neighborhood. In this toy example, transcript T_{ik} has four neighbor transcripts $\{T_{g_1a}, T_{g_2b}, T_{g_3c}, T_{g_4d}\}$, which are transcripts from g_1, g_2, g_3 and g_4 , respectively. The neighborhood expression ϕ_{ik} of T_{ik} is then calculated as the average of its neighbor transcripts' expressions and further normalized by transcript length, represented as the vector product between π and $S_{*,(i,k)}$ normalized by the number of neighbors $\sum S_{*,(i,k)}$ and the transcript length l_{ik} in the figure.

doi:10.1371/journal.pcbi.1004465.g002

Fig 2. The Dirichlet parameter α_i is defined as a function of ϕ_{ik} as

$$\alpha_{ik} = \lambda \phi_{ik} + 1, \tag{5}$$

where $\lambda > 0$ is a tuning parameter balancing the belief between the prior-read count and the aligned-read count.

To obtain the optimal \mathbf{P} jointly for all genes, we introduce a pseudo-likelihood model to estimate \mathbf{P} iteratively in each iteration. Assuming uniform $Pr(r_i)$, the pseudo-likelihood function is defined as,

$$\mathcal{L}(\mathbf{P}, \boldsymbol{\alpha}; \mathbf{r}) = \prod_{i=1}^N \mathcal{L}(\mathbf{P}_i, \boldsymbol{\alpha}_i; r_i) = \prod_{i=1}^N \frac{Pr(\mathbf{P}_i | \boldsymbol{\alpha}_i) Pr(r_i | \mathbf{P}_i)}{Pr(r_i)} \propto \prod_{i=1}^N Pr(\mathbf{P}_i | \boldsymbol{\alpha}_i) Pr(r_i | \mathbf{P}_i). \tag{6}$$

Note that the pseudo-likelihood model relies on the independence assumption among the likelihood functions of each individual gene when the $\boldsymbol{\alpha}$ parameters of the Dirichlet priors are pre-computed. Thus, the model simply takes the product of the likelihood function from each gene. Each prior distribution $Pr(\mathbf{P}_i | \boldsymbol{\alpha}_i)$ follows the Dirichlet distribution,

$$Pr(\mathbf{P}_i | \boldsymbol{\alpha}_i) = C(\boldsymbol{\alpha}_i) \prod_{k=1}^{|\mathbf{T}_i|} p_{ik}^{\alpha_{ik}-1}, \text{ where } C(\boldsymbol{\alpha}_i) = \frac{\Gamma(\sum_k \alpha_{ik})}{\prod_k \Gamma(\alpha_{ik})}. \tag{7}$$

Integrating eqs (1) and (7), the pseudo-likelihood function in eq (6) can be rewritten with Dirichlet prior as

$$\begin{aligned} \mathcal{L}(\mathbf{P}; \mathbf{r}) &= \prod_{i=1}^N \left[C(\boldsymbol{\alpha}_i) \prod_{k=1}^{|\mathbf{T}_i|} p_{ik}^{\alpha_{ik}-1} \right] \left[\prod_{j=1}^{|\mathbf{r}_i|} \sum_{k=1}^{|\mathbf{T}_i|} p_{ik} q_{ijk} \right] \\ &= \prod_{i=1}^N \left[C(\lambda \phi_i + 1) \prod_{k=1}^{|\mathbf{T}_i|} p_{ik}^{\lambda \phi_{ik}} \right] \left[\prod_{j=1}^{|\mathbf{r}_i|} \sum_{k=1}^{|\mathbf{T}_i|} p_{ik} q_{ijk} \right]. \end{aligned} \tag{8}$$

In the pseudo-likelihood function in eq (8), the only hyper-parameter λ balances the proportion between the Dirichlet priors and the observed read counts of each transcript. The larger the λ , the more belief put on the priors.

The Net-RSTQ algorithm

The Net-RSTQ algorithm optimizes eq (8) by dividing the optimization into sub-optimization problems of sequentially estimating each P_i . Specifically, we fix all P_c , $c \neq i$, and thus ϕ_i when estimating P_i with EM in each iteration and repeat the process multiple rounds throughout all the genes. In each step, the neighborhood expression ϕ is recomputed with new P_i for computing the quantification of the next gene. For each sub-optimization problem, we estimate P_i with a fixed ϕ , the part of the likelihood function in eq (8) involved with the current variables P_i is

$$\bar{\mathcal{L}}(P_i; r_i) = \left[\prod_{g \in nb(i)} C(\lambda \phi_g + 1) \prod_{k=1}^{|T_g|} p_{gk}^{\lambda \phi_{gk}} \right] \left[C(\lambda \phi_i + 1) \prod_{k=1}^{|T_i|} p_{ik}^{\lambda \phi_{ik}} \right] \left[\prod_{j=1}^{|r_i|} \sum_{k=1}^{|T_i|} p_{ik} q_{ijk} \right], \quad (9)$$

where $nb(i)$ is the set of the genes containing transcripts that are neighbors of the transcripts in gene i in the transcript network. Eq (9) consists of three terms separated by the braces. The second and the third terms are the Dirichlet prior and the likelihood of the observed counts in the data for gene i . The first term is the Dirichlet priors of the neighbor transcripts of each T_{ik} . These prior probabilities are involved since ϕ_g are functions of the current variable P_i (eqs (3)–(5)). Eq (9) cannot be easily solved with standard techniques. We adopt a heuristic approach to only take steps that will increase the whole pseudo-likelihood function in eq (8). The Net-RSTQ algorithm is outlined below

Algorithm 1 Net-RSTQ

```

1: Initialization: random initialization or base EM (eq (1)) estimation of  $P^{(0)}$ 
2: for round  $t = 1, \dots$  do
3:    $P^{(t)} = P^{(t-1)}$ 
4:   for gene  $i = 1, \dots, N$  do
5:     compute  $\phi_i$  based on  $P^{(t)}$  with eqs (3) and (4)
6:     estimate  $P_i$  with EM algorithm (see next section)
7:     if  $\bar{\mathcal{L}}(P_i) > \bar{\mathcal{L}}(P_i^{(t)})$  then
8:        $P_i^{(t)} = P_i$ 
9:     end if
10:  end for
11:  if  $\max(abs(P^{(t)} - P^{(t-1)})) < 1e-6$  then
12:    break
13:  end if
14: end for
15: return  $P$ 

```

In the algorithm, the outer for-loop between line 2–14 performs multiple passes of updating P . The inner for-loop between line 4–10 scans through each gene to update each P_i . Line 7 checks the the difference in the likelihood $\bar{\mathcal{L}}$ of gene i before and after the estimated P_i is applied. The newly estimated P_i is kept in line 8 only if the likelihood $\bar{\mathcal{L}}$ in eq (9) is higher. The convergence of P is checked at line 11. In each sub-optimization problem, EM algorithm (described in the next section) is applied to estimate P_i . After convergence, the transcripts expression π can be learned by eq (3) with the optimal P .

Estimating P_i given ϕ_i

In line 6 of Algorithm 1, we maximize the likelihood function of the sub-optimization problem in eq (9) to learn P_i as

$$\mathcal{L}(P_i; r_i) = \left[C(\lambda \phi_i + 1) \prod_{k=1}^{|T_i|} p_{ik}^{\lambda \phi_{ik}} \right] \left[\prod_{j=1}^{|r_i|} \sum_{k=1}^{|T_i|} p_{ik} q_{ijk} \right]. \quad (10)$$

Note that eq (10) is the part of eq (9) without the Dirichlet priors of the neighboring genes. In line 7 of Algorithm 1, the ignored Dirichlet priors are combined with the likelihood in eq (10), when $\bar{\mathcal{L}}(\mathbf{P}_i)$ is computed, to evaluate the whole likelihood in eq (9). The likelihood function in eq (10) is defined on a categorical variable with Dirichlet prior, which can be solved with EM algorithm. Following EM formulation in [26], the expectation a_{ijk} , a soft assignment of read j to transcript k in gene i , is first estimated in the expectation step and \mathbf{P}_i is then learned in the maximization step. When ϕ_i is given, by taking log of eq (10) we can write the EM steps to find \mathbf{P}_i below.

E step. Letting *Match* signify a matching between reads and transcripts, and *Match*(j) be the transcript from which read j originates, we get:

$$\log [\mathcal{L}(\mathbf{P}_i; \mathbf{r}_i, \mathbf{Match})] = \log C(\lambda\phi_i + 1) + \sum_{k=1}^{|\mathbf{T}_i|} \lambda\phi_{ik} \log(p_{ik}) + \sum_{j=1}^{|\mathbf{r}_i|} \log(p_{i\text{Match}(j)} q_{ij\text{Match}(j)}), \quad (11)$$

which leads to

$$\begin{aligned} \mathcal{Q}(\mathbf{P}_i | \mathbf{P}_i^{(it)}) &= E_{\text{Match}|\mathbf{r}_i, \mathbf{P}_i^{(it)}}[\log(\mathcal{L}(\mathbf{P}_i; \mathbf{r}_i))] \\ &= \log C(\lambda\phi_i + 1) + \sum_{k=1}^{|\mathbf{T}_i|} \lambda\phi_{ik} \log(p_{ik}^{(it)}) + \sum_{j=1}^{|\mathbf{r}_i|} \sum_{k=1}^{|\mathbf{T}_i|} (\log p_{ik}^{(it)} + \log q_{ijk}) * \frac{p_{ik}^{(it)} q_{ijk}}{\sum_{k=1}^{|\mathbf{T}_i|} p_{ik}^{(it)} q_{ijk}} \quad (12) \\ &= \log C(\lambda\phi_i + 1) + \sum_{k=1}^{|\mathbf{T}_i|} \lambda\phi_{ik} \log(p_{ik}^{(it)}) + \sum_{j=1}^{|\mathbf{r}_i|} \sum_{k=1}^{|\mathbf{T}_i|} a_{ijk} \log(p_{ik}^{(it)}) + \sum_{j=1}^{|\mathbf{r}_i|} \sum_{k=1}^{|\mathbf{T}_i|} a_{ijk} \log(q_{ijk}) \end{aligned}$$

where it is the it^{th} iteration in EM and

$$a_{ijk} = \frac{p_{ik}^{(it)} q_{ijk}}{\sum_{k=1}^{|\mathbf{T}_i|} p_{ik}^{(it)} q_{ijk}}. \quad (13)$$

M step. Given that q_{ijk} and ϕ_i are known, the above reduces to maximizing

$$\mathbf{P}_i^{(it+1)} = \arg \max_{\mathbf{P}_i} \left[\sum_{k=1}^{|\mathbf{T}_i|} \lambda\phi_{ik} \log(p_{ik}) + \sum_{j=1}^{|\mathbf{r}_i|} \sum_{k=1}^{|\mathbf{T}_i|} a_{ijk} \log(p_{ik}) \right]. \quad (14)$$

Using Lagrange multipliers and differentiating, eq (14) is maximized when

$$p_{ik}^{(it+1)} = \frac{\lambda\phi_{ik} + \sum_{j=1}^{|\mathbf{r}_i|} a_{ijk}}{\sum_{k=1}^{|\mathbf{T}_i|} (\lambda\phi_{ik} + \sum_{j=1}^{|\mathbf{r}_i|} a_{ijk})}. \quad (15)$$

After EM algorithm converges, we update \mathbf{P} with the newly estimated \mathbf{P}_i only if the update leads to increase of eq (9). It can be seen from eq (15) that the role of λ is a parameter controlling the balance between the prior-read count and the aligned-read count. To see that, recall ϕ_{ik} is the prior-read count of transcript T_{ik} by the average expression of its neighbors (eq (4)) and $\sum_{j=1}^{|\mathbf{r}_i|} a_{ijk}$ is the expected aligned-read count of transcript T_{ik} . λ directly balances the contributions from the two terms. Therefore, a reasonable choice of λ should apply to RNA-Seq data with similar level of noise or bias in general.

qRT-PCR experiment design

Three qRT-PCR experiments are designed to measure the isoform proportions of 25 multi-isoform genes in three cell lines, H9 stem cell line, OVCAR8 ovarian cancer cell line and MCF7

breast cancer cell line. The cell lines were selected based on the available of both RNA-Seq data and cell culture in our labs. The qRT-PCR experiments focused on the gene with most different quantification results reported by Net-RSTQ and other compared methods. Due to the limitations in time and cost of running qRT-PCR experiments, only the 25 genes in the three cell lines were tested with all the results reported in the experiments. Quantitation of the real-time PCR results was done on the data from H9 human embryonic stem cells to obtain the absolute expressions for comparing more than two transcripts and comparative Ct method was done on the data from OVCAR8 ovarian cancer cells and MCF7 breast cancer cells to obtain the ratio between a pair of transcripts.

H9 Stem cell line. Total RNA was extracted from human embryonic stem (ES) H9 cells by using TRIzol (Invitrogen). To repeat the experiments of triplicate three times, 5 μ g RNA was used to synthesize complementary DNA with ReverTra Ace (Toyobo) and oligo-dT (Takara) according to the manufacturer's instructions. Transcript levels of genes were determined by using Premix Ex Taq (Takara) and analysed with a CFX-96 Real Time system (Bio-Rad). The templates for different transcripts were generated with PCR by using the template primers in [S1 Table](#). After isolation and purification, the templates were used to generate the standard curves with qRT-PCR by using the qRT-PCR primers for different transcripts. The generated standard curves have coefficient of determination (R²) over 0.999. The qRT-PCR primers were then applied to determine the expression levels of different transcripts in H9 ES cells by calculating with the standard curves. The expressions were carried out in three independent replications and the standard deviations were provided after the average.

Ovarian cancer cell line. 1 μ g of total RNAs were isolated from untreated OVCAR8 cells using Trizol (Invitrogen). RNA was reverse-transcribed using Superscript II reverse transcriptase (Invitrogen) according to manufacture protocol. Real-time PCR was performed on CFX384 Real-time system (Bio-Rad) with FastStart SYBR Green Master (Roche) with the primer sets in [S2 Table](#). PCR conditions are 10 min at 95°C and 40 cycles of 95°C for 45 sec and 60°C for 45 sec. Quantitation of the real-time PCR results was done using comparative Ct method. Two replicates of qRT-PCR were performed using total RNAs isolated.

Breast cancer cell line. 0.5 μ g of total RNAs purified from MCF7 cells was used for oligo d(T)₂₀-primed reverse transcription (Superscript III; Life Technologies). SYBR Green was used to detect and quantitate PCR products in real-time reactions with the primer sets in [S3 Table](#). PCR conditions for qRT-PCR analysis are 2 min 94°C and 40 cycles of 94°C for 30 sec, 60°C for 20 sec and 72°C for 30 sec. Quantitation of the real-time PCR results was done using comparative Ct method. GAPDH mRNA was used as a normalization control for quantitation. Three replicates of qRT-PCR were performed using total RNAs isolated.

RNA-Seq data preparation

Three cell line RNA-Seq datasets were used for evaluating the accuracy of transcript quantification by comparison with qRT-PCR results. The first dataset is the H9 embryonic stem cell line data from [\[28\]](#), downloaded from SRA. The second dataset is an in-house dataset from the ovarian cancer cell line OVCAR8 prepared at University of Kansas Medical Center. The third dataset is the MCF7 breast cancer cell line data from [\[29\]](#), downloaded from SRA. There are 23,397,325 single-end 34bp reads in the stem cell line dataset, 19,892,473 paired-end 100bp reads in the OVCAR8, and 21,855,632 paired-end 76bp reads in the MCF7 mapped to the human hg19 reference genome by TopHat2.0.9 [\[30\]](#) with up to 2 mismatches allowed. Exon coverages and read counts of exon-exon junctions were generated by SAMtools [\[31\]](#) to be utilized with Net-RSTQ and base EM ([eq \(1\)](#)). Cufflinks [\[32\]](#) directly infers transcript expressions based on the alignment by TopHat with the min isoform fraction set to 0 for better sensitivity.

Table 3. Summary of patient samples in TCGA datasets. The samples are classified by cutoffs on survival and relapse time based on the available clinical information in each dataset.

Cancer Type	Event	# of Patients by years
Ovarian serous cystadenocarcinoma(OV)	Survival	76(<3 ys) vs 62(>4 ys)
	Relapse	79(<1.5 ys) vs 68(>2 ys)
Breast invasive carcinoma(BRCA)	Survival	66(<5 ys) vs 57(>8 ys)
	Relapse	42(<5 ys) vs 38(>8 ys)
Lung adenocarcinoma(LUAD)	Survival	47(<2 ys) vs 56(>3 ys)
Lung squamous cell carcinoma(LUSC)	Survival	67(<2 ys) vs 77 (>3 ys)

doi:10.1371/journal.pcbi.1004465.t003

TCGA RNA-Seq datasets of Ovarian serous cystadenocarcinoma (OV), Breast invasive carcinoma (BRCA), Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC) were analyzed for patient outcome prediction with transcript expressions estimated by Net-RSTQ, base EM (eq (1)), RSEM [33] and Cufflinks [32]. Both the gene expression and transcript expression data reported by RSEM [33] in TCGA (level 3 data) were utilized as two baselines for cancer outcome prediction. The raw RNA-Seq fastq files (level 1 data) were downloaded from Cancer Genomics Hub (CGHub) and processed by TopHat for use with Net-RSTQ, base EM and Cufflinks. The patient samples in each dataset were classified into cases and controls based on the survival and relapse outcomes as shown in Table 3. The command lines for preparing the data with RSEM and Cufflinks are available in the S3 Text.

Results

There are six major results in this section, 1) isoform co-expression analysis on TCGA data to show the correlation with protein domain-domain interactions; 2) overlapping the DDIs and KEGG pathways to understand the transcript networks; 3) simulations for model validation and statistical analysis; 4) qRT-PCR experiments to measure the performance of transcript quantification; 5) cancer outcome prediction on TCGA data to measure the quality of transcript quantification as molecular markers; and 6) running time of Net-RSTQ.

Net-RSTQ was compared with base EM (the base model in eq (1)), Cufflinks [32] and RSEM (isoform expression or gene expression) [33]. The accuracy of transcript quantification was directly measured on the simulated data with ground-truth expressions and qRT-PCR data from the three cell lines. Cancer outcome prediction on four TCGA cancer datasets evaluates the potential of using isoform expressions as predictive biomarkers in clinical settings. Statistical assessment was also performed on randomized transcript networks to evaluate the significance of the results.

Isoform co-expressions correlate with protein domain-domain interactions

To investigate the correlation between protein domain-domain interactions and isoform transcript co-expressions, we calculated the number of transcript pairs that are both nearby (being neighbors or having a distance up to 2) in the transcript network and highly co-expressed in the TCGA samples. The transcript co-expressions were calculated by Pearson's correlation coefficients of each pair of transcripts across all the samples in each dataset with the isoform transcript quantification by Cufflinks. The transcript pairs were then sorted by the correlation coefficients from the largest to the smallest and grouped into bins of size 1000. The number of transcript pairs that are nearby in the transcript networks out of 1000 pairs are calculated within each bin and plotted in Fig 3(A) and 3(B) for the two cancer gene lists, respectively. In

both [Fig 3\(A\)](#) and [3\(B\)](#), the left column shows the plots of the number of pairs that are neighbors in the transcript network, and the right column shows the plots of the number of transcript pairs with a distance up to 2 in the transcript network, among the 1000 pairs in each bin. In all the plots, similar trends are observed in all the four cancer datasets: there are more interacting isoform pairs in the bins with higher co-expressions. For example, among the 1000 transcript pairs with the highest correlation coefficients, there are 73 interactions in the transcript network in OV dataset and thus, 73 interactions (y-axis) for bin index 1 (x-axis) is plotted in the left column of [Fig 3\(A\)](#). In all the plots, there is a clear pattern that the numbers of matched nearby transcripts in the transcript network among the 1000 pairs in the first few bins are higher than the expected average of 30 in the small network of density 3.02%, 114 in the small network of density 11.41% (with distance up to 2), 45 in the larger network of density 4.54%, and 203 in the larger network of density 20.33% (with distance up to 2). Moreover, the 2-step walk clearly promoted the number of overlaps with the pairs of higher co-expressions in the small network. For example, the significant overlap is extended from the first 25 bins to approximately the first 50 bins or more in the four datasets. The observation suggests that higher co-expressions exist not only in the direct neighbors in the transcript network but also the nearby nodes by a small distance. By exploring the network structure with prior information through neighbors by many steps in iterations, Net-RSTQ model is expected to propagate the expression values from each transcript to its nearby nodes in the network to capture the co-expressions. Note that considering the neighboring pairs with distance up to 2 in the larger network will result in a graph of density 20.33%, which is likely to contain too many false relations by the two-step walk. Thus, the plots of the larger network of distance-2 pairs are only included for the completeness of the analysis.

The canonical 2x2 chi-square test was also applied to compare the number of the domain-domain interactions in the first 10,000 transcript pairs (first 10 bins) with the number in the rest of the pairs. In all the four datasets in both [Fig 3\(A\)](#) and [3\(B\)](#) with one exception in the LUSC dataset on the large network of distance-2 relation, there is a significant difference that the highly co-expressed transcripts are more likely to interact with each other in the transcript network, confirmed by the significant *p*-values. As explained previously, the exception is likely due to the large number of false-positive pairs in the dense network. The observation further support the hypothesis that protein domain-domain interactions correlate transcript co-expressions reported in previous studies [[12](#), [13](#)].

To further understand the specificity of the domain-domain interactions in the highly co-expressed transcripts, we calculated the number of domain-domain pairs that construct the DDIs in the top 10,000 co-expressed transcript pairs. The statistics suggest high diversity of the type of DDIs. For example, there are 547 interacting transcript pairs among the 201 out of 898 transcripts in the top 10,000 co-expressed transcript pairs in OV dataset for small network. The 547 interacting transcript pairs represent 770 different domain-domain interactions (There might be more than one DDIs between a pair of transcripts). There are 739 interacting transcript pairs among the 538 out of 5599 transcripts in the top 10,000 co-expressed transcript pairs in OV dataset for large network. The 739 interacting transcript pairs represent 1277 different domain-domain interactions. The statistics suggest that the correlation between protein domain-domain interactions and transcript co-expressions is not a bias due to a few highly spurious DDIs. It is a general correlation in many different DDIs and co-expressed transcripts. Very similar statistics were observed in all the datasets and both networks.

To further demonstrate the co-expression relations in the transcript network, two examples are shown in [S1 Fig](#). In [S1\(A\) Fig](#), WHSC1L1 contains two isoforms connected with different interactions in the transcript network. Isoform NM_017778 interacts with 12 transcripts with average correlation coefficients 0.22 and the other isoform NM_023034 interacts with 13 more

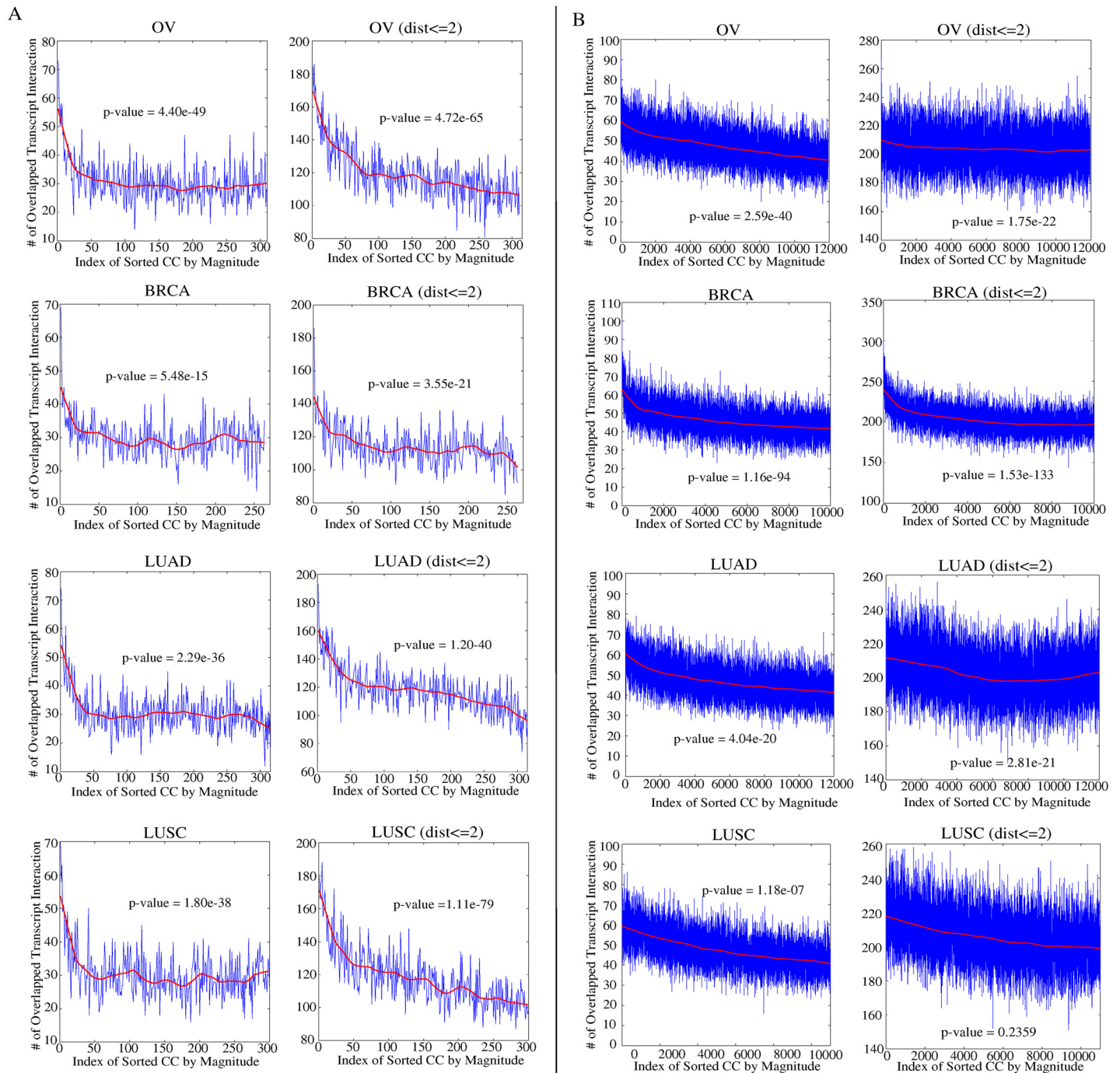


Fig 3. Correlation between transcript co-expression and protein domain-domain interaction in TCGA datasets. The correlation coefficients between transcript expressions across all patient samples are first calculated in each dataset for each pair of transcripts by Cufflinks. The correlation coefficients are then sorted from largest to smallest and grouped into bins of size 1000 each. The x-axis is the index of the bins with lower index indicating larger correlation coefficients. The y-axis is the number of the pairs among the 1000 pairs of transcripts in each bin that coincide with protein domain-domain interaction between the transcript pair. The red line is the smooth plot by fitting local linear regression method with weighted linear least squares (LOWESS) to the curves. *p*-value is reported by chi-square test. (A) Co-expressions are calculated based on the small gene list. (B) Co-expressions are calculated based on the large gene list. In both (A) and (B), the left column shows the plots based on the connected transcript pairs in the transcript network and the right column shows the plots based on the transcript pairs with distance up to 2 in the network.

doi:10.1371/journal.pcbi.1004465.g003

transcripts with average correlation coefficients 0.30 compared with the average correlation coefficient 0.188 against the other unconnected isoforms across the samples in the OV dataset. In [S1 \(B\) Fig](#), gene BRD4 contains two isoforms both of which are connected with the same 14 neighbors in the network. The average correlation coefficients between these two isoforms and the 14 neighboring isoforms are both above 0.26 compared with the average correlation coefficient less than 0.15 against the other unconnected isoforms across the samples on the BRCA dataset. In both examples, we observed high degree of agreement between co-expressions and DDIs.

Protein domain-domain interactions enrich KEGG pathways

To further understand the transcript networks, we overlapped the DDIs between genes in the two networks with the 294 human KEGG pathways [\[34\]](#). Among the 397 genes in the small network, 10.97%(17284) of the pairs are co-members in at least one KEGG pathway. The 10.97% KEGG co-member pairs covers 42.70%(2122) of the DDIs among the genes while the other 89.03%(140352) non-co-member pairs covers 57.30%(2748) of the DDIs. By these numbers, there is about 6-fold enrichment of DDIs in the KEGG co-member genes in the small network. Among the 2551 genes in the large network, the 5.15%(335372) KEGG co-member pairs covers 12.45%(40812) of the DDIs among genes while the other 94.85%(6172229) non-co-member pairs covers 87.55%(287090) of the DDIs. By these numbers, there is about 2.6-fold enrichment of DDIs in the KEGG co-member genes in the large network. We also list the KEGG pathways that are highly enriched with DDIs in the large network in [S4 Table](#). Specifically, we consider the subnetwork of genes that are members of one KEGG pathway and calculated the density of DDIs in the subnetwork to compare to the overall density of 5.04% in the whole network. Interestingly, most of the enriched pathways are signaling pathways and disease pathways with very high DDI densities.

Net-RSTQ captures network prior in simulations

In the simulations, we applied flux-simulator [\[35\]](#) to generate paired-end short reads simulating real RNA-Seq experiment *in silico* based on a ground truth transcript expression profile, using hg19 reference human genome and RefSeq annotations downloaded from UCSC Genome Browser. To generate the ground-truth expression profiles, the gene expressions were sampled from a poisson distribution and the proportions of the isoforms in each gene were derived based on a neighbor average expression in the small transcript network and an initial mixed power law expression profile with gaussian noise. A sequential updating was used to compute the proportion of each isoform by adding the neighbors' average expressions to the initial expression. The update procedure can be found in the [S2 Text](#). At last, flux-simulator was applied to simulate the short reads based on the ground truth transcript expression file. 15 million 76-bp paired reads were generated by Flux Simulator and mapped to the reference genome by TopHat [\[30\]](#) with up to two mismatches allowed. To account for the large dynamic range of abundances, the expressions were normalized by $\log_2(\text{expression}+1)$.

The correlation coefficients between the transcript abundances estimated by Net-RSTQ under various λ , base EM ([eq \(1\)](#)), Cufflinks and RSEM, and the ground truth transcript abundances are reported in [Fig 4](#). Furthermore, Net-RSTQ was also tested with 100 randomized networks with permuted indexes of transcripts in the transcript network. To assess the impact of the network prior, two cases are shown. [Fig 4\(A\)](#) reports the correlation between the transcripts in which isoforms coded by the same gene are connected with different neighbors (109 out of 898 transcripts in 29 genes). [Fig 4\(B\)](#) reports the results from all the genes with more than one isoform (712 out of 898 transcripts in 211 genes). In both comparisons, the transcript expressions estimated by Net-RSTQ achieve higher correlation with the ground truth compared with

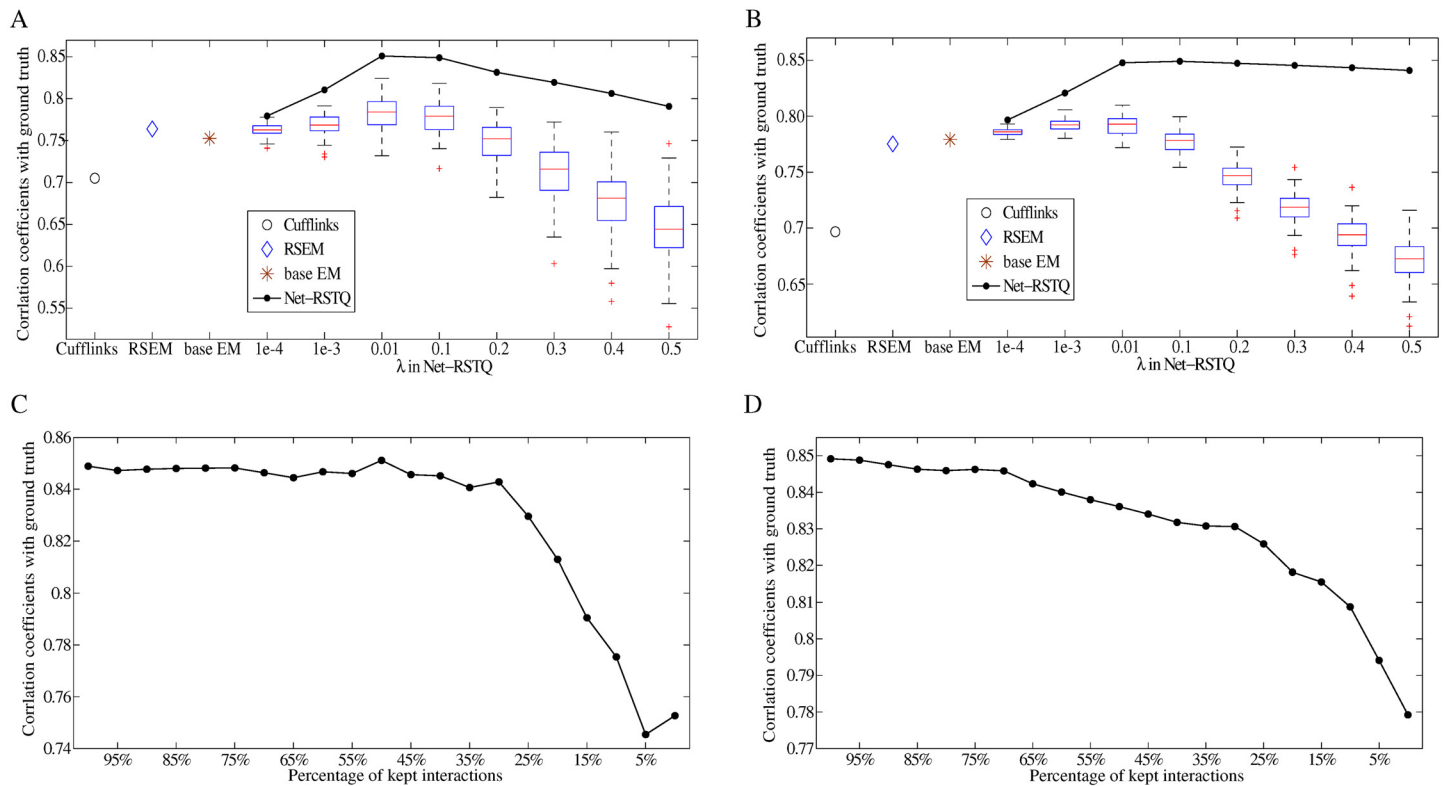


Fig 4. Correlation between estimated transcript expressions and ground truth in simulation. In (A) and (B) x-axis are labeled by the compared methods and different λ parameters of Net-RSTQ. The bar plots show the results of running Net-RSTQ with 100 randomized networks. In (C) and (D), x-axis are the percentage of edges that are removed from the networks. The plots show the results of running Net-RSTQ with the incomplete networks. (A) and (C) report the results of 109 transcripts of the isoforms in the same gene with different domain-domain interactions. (B) and (D) report the results of 712 isoforms in genes with multiple isoforms.

doi:10.1371/journal.pcbi.1004465.g004

base EM, Cufflinks and RSEM. Slightly higher improvement was observed in the first case than in the second case since the network prior plays more significant role in differentiating the isoform expressions by their different neighbors. When randomized networks are used, Net-RSTQ leads to similar or worse results due to the wrong prior information. Note that since the datasets were generated to partially conform to the network prior, the isoform expressions are relatively “smooth” among the neighboring isoforms. Net-RSTQ tends to generate smoother expressions than base EM, Cufflinks and RSEM. When applying Net-RSTQ with small λ s and randomized network priors, slight improvement was also observed due to the smoothness assumption on the data.

To evaluate the effect of missing edges in the transcript network due to the undetected protein domain-domain interactions, we randomly removed certain percentages of the edges in the transcript network and then run Net-RSTQ with $\lambda = 0.1$ on the incomplete networks. The results are shown in Fig 4(C) and 4(D) for the 109 transcripts with different neighbors and the 712 transcripts in the gene with more than one transcript, respectively. It is intriguing to observe that only when a large percentage of the edges are removed, the performance of Net-RSTQ is affected. Intuitively, the observation can be explained by the fact that the Dirichlet prior parameter is proportional to the average of the neighbors’ expressions. As long as some of the neighbors are still connected to the target transcript in the network, the prior information is still useful. The result suggests that Net-RSTQ is relatively robust to utilize transcript networks potentially constructed with a large percentage of undetected protein domain-domain interactions.

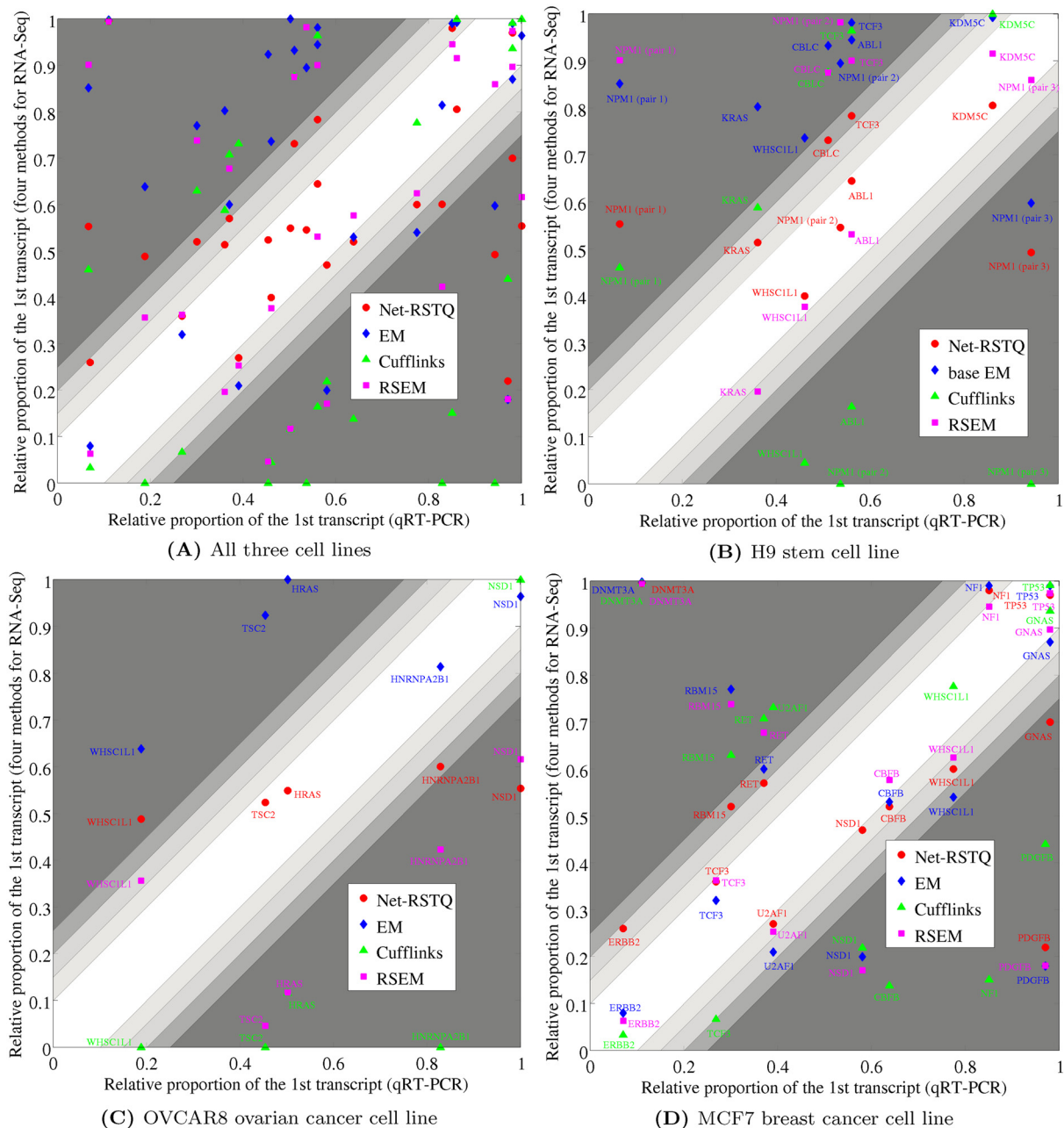
Three qRT-PCR experiments confirmed overall improved transcript quantification

The isoform proportions estimated by Net-RSTQ, base EM, RSEM, and Cufflinks were compared to the qRT-PCR results on the three cell lines. Parameter $\lambda = 0.1$ was fixed in all the Net-RSTQ experiments. Among the genes that Net-RSTQ, base EM, RSEM, and Cufflinks report most different quantification results, qRT-PCR experiments were performed to test the genes with relatively higher coverage of RNA-Seq data, coding two to three isoforms, and the feasibility of designing isoform-specific primers in the qRT-PCR products (see [S1](#), [S2](#) and [S3](#) Tables). Twenty-five genes in total were tested in the three cell lines: seven in H9 stem cell line, five in OVCAR8 ovarian cancer cell line, and thirteen in MCF7 breast cancer cell line. The scatter plots of the relative abundance of the first transcript in each gene estimated by Net-RSTQ, base EM, Cufflinks and RSEM were compared to the qRT-PCR results in [Fig 5\(A\) and 5\(E\)](#). In the scatter plot, the estimated relative abundance by Net-RSTQ were closer to qRT-PCR results measured by the accuracy of various thresholds and Root Mean Square Errors. Net-RSTQ achieved the lowest Root Mean Square Error of 0.291, which is more than 0.05 less than 0.3435, the second best achieved by RSEM. In the 20% confidence region, Net-RSTQ puts 59.3% of the pairs in the region compared with 37%, 29.6%, and 51.9% by base EM, Cufflink, and RSEM, respectively. RSEM performed well by putting 37.0% of the pairs within 10% confidence regions but performed poorly in about half of the pairs with more than 25% error.

The relative abundance of the seven genes in H9 stem cell line is shown in [Figs 5\(B\) and S2\(A\)](#) and [S5 Table](#). In all seven genes tested, the relative abundance estimated by Net-RSTQ is closer to the qRT-PCR results compare to that by base EM and Cufflinks. RSEM performed similarly well on four genes and worse on the other three genes, CBLC, TCF3 and NPM1. The same comparison on the five selected genes in OVCAR8 ovarian cancer cell line is shown in [Figs 5\(C\) and S2\(B\) and S6 Table](#). Cufflinks reports very low expressions in the first transcript in four genes, three of which do not agree with the highly expressed transcript in the qRT-PCR results. While base EM performed better for two genes (NSD1 and HNRNPA2B1), Net-RSTQ performed better on the other three genes (HRAS, TSC2, and WHSC1L1). Net-RSTQ correctly predicted the overall enrichment of isoforms of HNRNPA2B1 and NSD1 (NM_031243 > NM_002137 in HNRNPA2B1 and NM_022455 > NM_172349 in NSD1). It is possible that the expressions of NM_002137 transcript in gene HNRNPA2B1 and NM_172349 in gene NSD1 were slightly over-smoothed by network information in Net-RSTQ with the fixed λ parameter. RSEM performed slightly better on WHSC1L1 and NSD1 but much worse in the other three genes. The same comparison on the thirteen genes in MCF7 breast cancer cell line is shown in [Figs 5\(D\) and S2\(C\) and S7 Table](#). Cufflinks performed poorly on 8 genes with more than 25% error while RSEM, base EM and Net-RSTQ performed poorly on 5, 4 and 3 genes, respectively. Overall, Net-RSTQ performed better than base EM and Cufflinks and slightly better than RSEM. In summary, Net-RSTQ improved the overall isoform quantification significantly in the H9 stem cell data and predicted more consistent cases in OVCAR8 and MCF7 cancer cell lines data. Note that there could be more uncertainties in primer designs due to somatic DNA variations and cell differentiation and proliferation in cancer cell lines, potentially a larger variation in the qRT-PCR experiments on the cancer cell lines is expected than H9 stem cell line.

Net-RSTQ improved overall cancer outcome predictions

To provide an additional evaluation of the quality of transcript quantification, we designed six cancer outcome prediction tasks by the assumption that better transcript quantification always leads to better isoform markers for cancer outcome prediction. Net-RSTQ was compared with base EM, RSEM [[33](#)], and Cufflinks [[32](#)] by classification with the quantification of isoform



Accuracy	Cut-off	Net-RSTQ	base EM	Cufflink	RSEM
	0.1	29.6%	18.5%	18.5%	37.0%
	0.15	44.4%	33.3%	22.2%	40.7%
	0.2	59.3%	37.0%	29.6%	51.9%
	0.25	74.4%	44.4%	33.3%	51.9%
Root Mean Square Error		0.2915	0.3701	0.4111	0.3435

(E) Accuracy and Root Mean Square Error

Fig 5. Validation by comparison with qRT-PCR results. (A) The scatter plots compare the reported relative proportion of each pair of the isoforms of each gene between the computational methods (Net-RSTQ, base EM, Cufflinks, and RSEM) and qRT-PCR experiments. The proportions of the two compared isoforms in a pair are normalized to adding to 1. The x-axis and y-axis are the relative proportion of one of the two isoform (the other is 1 minus the proportion) reported by qRT-PCR and the computational methods, respectively. The scatter points aligning closer to the diagonal line indicate better estimations by a computational method matching to the qRT-PCR results. The unshaded gradient around the diagonal line shows the regions with scatter differences less

than 0.1, 0.15, 0.2 and 0.25, within which the estimations are more similar to the qRT-PCR results. (B)-(D) The scatter plots on each individual dataset. (E) The table shows the percentage of predictions by each method within the unshaded regions and the overall Root Mean Square Error of the predictions by each method compared to the qRT-PCR results.

doi:10.1371/journal.pcbi.1004465.g005

transcripts in two cancer gene lists (397 and 2551 genes) on four cancer datasets. Each dataset is divided into four folds with two folds for training, one fold for validation (parameter tuning), and one fold for test in a four-fold cross-validation. Support Vector Machine (SVM) with RBF kernel [36] were chosen as the classifier. We repeated the four-fold cross-validation 100 times by each method in each dataset.

The average area under the curve (AUC) of receiver operating characteristic of the 100 repeats are reported in Table 4 when the small gene list was used and Table 5 when the large gene list was used. The transcript expressions estimated by Net-RSTQ consistently achieved better average classification results than those by the base EM. To evaluate the statistical significance of the differences between the AUCs generated by Net-RSTQ and the base EM in the 100 repeats, we also report the *p*-values by a binomial test on the number of wins/loses in all the experiments between Net-RSTQ and the base EM in Tables 4 and 5. When the small gene list was tested, three cases were significant with low *p*-values less than 0.001 and two cases were significant with *p*-values just below 0.02 while in the BRCA (survival) data, the *p*-value is only moderately significant even though the average by Net-RSTQ is higher. Overall, Net-RSTQ outperformed the base EM significantly. When the larger gene list was tested, the improvements are not as significant. The improvement was only significant in one dataset, BRCA (survival), and slightly significant in two datasets, OV (relapse) and LUSC (survival). In the other three datasets, the improvements are not significant. Net-RSTQ also outperformed Cufflinks and RSEM (transcript or gene) in five cases except the experiment on BRCA (relapse) dataset in Table 4. In Table 5, the improvements are less obvious. Moreover, the isoform expression features are not more informative than gene expression features. Overall, the classification performance with the small gene list in Table 4 is generally better than or similar to the large gene list in Table 5 possibly suggesting less relevance to survival and relapse in the large gene list.

The parameter λ was tuned by the AUC on the validation set and the optimal λ was used to train the Net-RSTQ model to be tested on the test set. The process is repeated for each fold in 100 repeats. To show the effect of varying the λ on the classification performance in Net-RSTQ, we plotted the average AUC on the validation set across the 100 repeats on the BRCA (survival) dataset with small gene list in S3(A) Fig. The optimal λ was 0.1 in this experiment. The local gradient around the optimal λ suggesting that the transcript network is playing an important role in inferring better transcript quantification from the RNA-Seq data. In S3(B) Fig, the convergence of Net-RSTQ is also illustrated by each update through all the genes in each iteration. After less than 10 overall iterations across 397 genes, Net-RSTQ converged well to a local optimum. Similar convergence patterns were observed in all other TCGA samples.

To understand the role of the transcript network in the transcript expression estimation, we used 100 randomized networks to learn the transcript proportion in each experiment with λ fixed to be 0.1. In each randomization, the edges were shuffled among all the transcripts in the small gene list. For transcript expressions learned by each randomized network, we conducted the same four-fold cross validation to compute the average AUCs among 100 repeats. The box-plot of the AUCs learned with the 100 randomized networks is shown in Fig 6. Compared with the classification results from the true transcript network, the result with randomized networks is always worse. Another important observation is that, the median value of the AUCs across the 100 randomized networks is lower or close to the result by the base EM, which suggests

Table 4. Classification performance of estimated transcript expressions and gene expression on the small cancer gene list. The mean AUC scores of classifying patients by estimated transcript (gene) expression in four-fold cross-validation for each dataset are reported. The best AUCs across the five models using isoforms as features are bold.

Dataset	OV(Survival)	OV(Relapse)	BRCA(Survival)	BRCA(Relapse)	LUAD(Survival)	LUSC(Survival)
Net-RSTQ(Isoform)	0.597	0.607	0.683	0.590	0.635	0.567
base EM(Isoform)	0.570	0.589	0.673	0.542	0.579	0.550
RSEM(Isoform)	0.587	0.550	0.651	0.616	0.613	0.536
Cufflinks(Isoform)	0.563	0.577	0.676	0.593	0.555	0.556
RSEM(Gene)	0.591	0.580	0.651	0.558	0.615	0.559
p-value(Net-RSTQ vs base EM)	0.0011	0.0198	0.1356	2.248e-5	1.948e-8	0.0167

doi:10.1371/journal.pcbi.1004465.t004

Table 5. Classification performance of estimated transcript expressions and gene expression on the large cancer gene list. The mean AUC scores of classifying patients by estimated transcript (gene) expression in four-fold cross-validation for each dataset are reported. The best AUCs across the five models are bold.

Dataset	OV(Survival)	OV(Relapse)	BRCA(Survival)	BRCA(Relapse)	LUAD(Survival)	LUSC(Survival)
Net-RSTQ(Isoform)	0.599	0.585	0.679	0.592	0.604	0.566
base EM(Isoform)	0.590	0.572	0.651	0.571	0.597	0.556
RSEM(Isoform)	0.584	0.569	0.663	0.594	0.587	0.543
Cufflinks(Isoform)	0.562	0.582	0.683	0.580	0.583	0.559
RSEM(Gene)	0.604	0.577	0.675	0.598	0.627	0.554
p-value(Net-RSTQ vs base EM)	0.3798	0.0967	0.0018	0.3822	0.6178	0.1356

doi:10.1371/journal.pcbi.1004465.t005

that the randomized networks play no role in improving classification and even lead to worse result. Overall, the results provide a clear evidence that the transcript network is informative for the transcript expression estimation, and supplies more discriminative features for cancer outcome prediction.

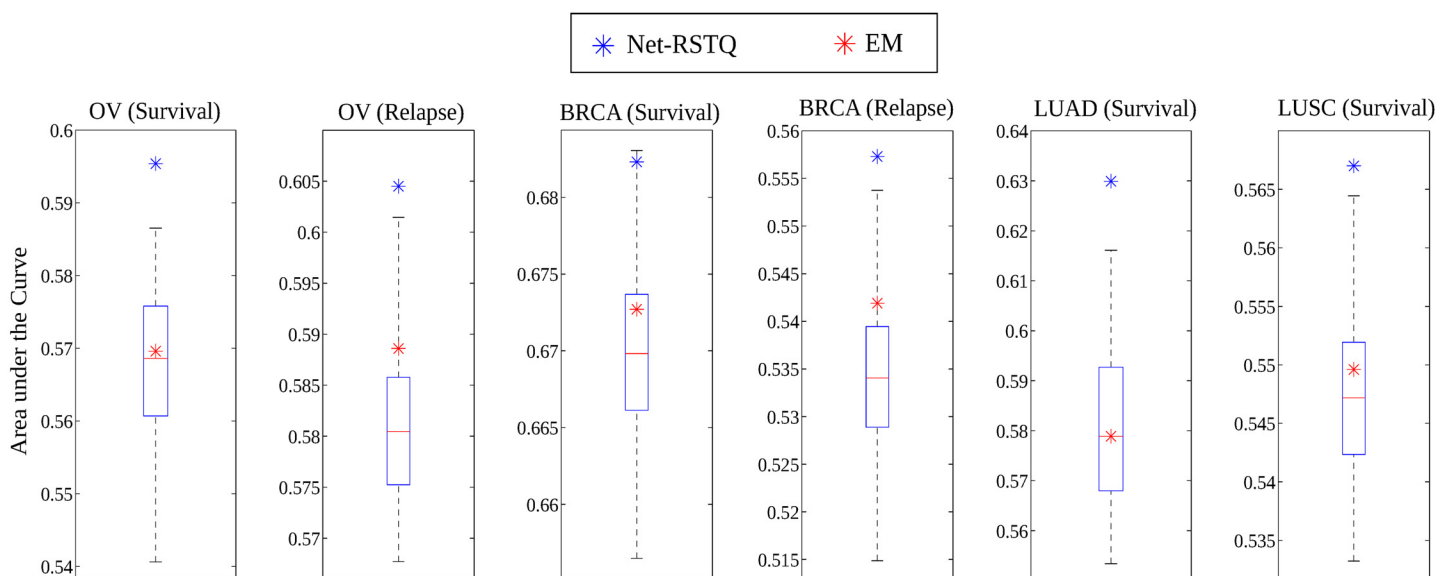


Fig 6. Statistical analysis with randomized networks. Comparison of the classification results by the randomized networks and the true network. The λ parameter was fixed to be 0.1 in all the experiments. The blue star and the red star represent the results with the real network and without network (base EM), respectively. The boxplot shows the results with the randomized networks.

doi:10.1371/journal.pcbi.1004465.g006

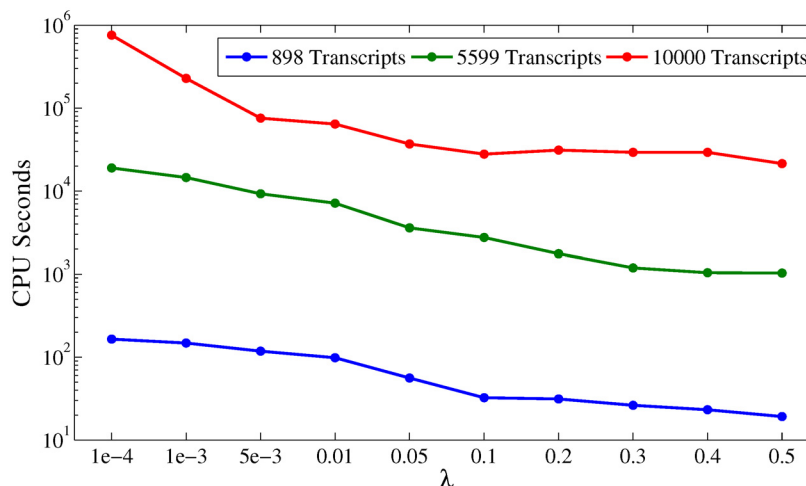


Fig 7. Running time. The plots show the CPU time (Intel Xeon E5-1620 with 3.70GHZ) for running the Net-RSTQ algorithm on three networks, the small transcript network, the large transcript network, and an artificial huge network of 10000 transcripts.

doi:10.1371/journal.pcbi.1004465.g007

Running time

To measure the scalability of Net-RSTQ, we tested the Net-RSTQ algorithm on the data of the MCF7 breast cancer cell line with three different networks, the small network (898 transcripts), the large network (5599 transcripts) and an artificial huge network (10000 transcripts). Fig 7 plots the CPU seconds of running Net-RSTQ on the three networks under different λ s. On the small network, the running time is at most about 100 seconds while on the large network and the huge network, the running time is in the scale of $1-e^3 \sim 1-e^4$ and $1-e^5 \sim 1-e^6$, respectively. When $\lambda = 0.1$, the CPU time for the small network is 32.4 seconds; for the large network is 2755 seconds; and for the artificial large network is 27806 seconds. The results suggest that Net-RSTQ might scale up to about 10000 transcripts, and thus the performance is sufficient for studies focusing on any pathway with up to several thousand genes in the pathway.

Discussion

In the paper, we explored the possibility of improving short-read alignment based transcript quantification with relevant prior knowledge, protein domain-domain interactions. The observation of the correlation between isoform co-expressions and protein domain-domain interactions suggests that the approach is a well-grounded exploration. Different from previously methods [27], Net-RSTQ is a network-based approach that directly incorporates protein domain-domain interaction information for transcript proportion estimation. The experiments suggested a great potential of exploring protein domain-domain interactions to overcome the limitations of short-read alignments and improve transcript quantification for better sample classification.

The Dirichlet prior from the neighboring isoforms play two different roles: differentiating isoform expressions to reflect different functional roles or smoothing isoform expressions to reflect similar functional roles, depending on whether the isoforms of a gene share the same or different interacting partners. This principle in modeling is based on the hypothesis that isoforms playing different functional roles (e.g. containing different protein domains) are more likely to behavior differently than isoforms with the same or similar functional roles (e.g. containing the same protein domains). When the isoforms of a gene interact with different

partners, their expressions correlates with their partners' expressions. And, when the isoforms of a gene interact with the same partners, there is no benefit on differentiating their proportions to drive the functionality. A limitation is that when the functional difference among the isoforms are not captured by domain content, the smoothing role might under-estimate the difference in their proportions. Thus, our future goal is to bring in other type of functional information to distinguish their functional roles in cancer such as preferential adoption of post-transcriptional regulations.

Currently, Net-RSTQ does not directly model multi-hits reads in multiple loci. In the TCGA experiments, around 5–10% of the aligned reads in four datasets have multiple alignments reported by TopHat and only one of the best alignments is considered. To check the effect of the multiple-alignment reads in transcript quantification, we allow up to 20 best alignments by TopHat and normalized the read assignment q_{ijk} by the number of loci that the reads aligned to. The correlation coefficients between the estimated gene expressions before and after the normalization are above 0.98 in all the datasets. A potential rigorous solution is to add iteratively reassignment of the reads to the potential origins based on updated abundance of the involved isoforms. The modification will significantly decrease the computational efficiency and make it impractical on large RNA-Seq datasets.

There is also another alternative of integrating the network information directly as a regularization term on the joint likelihood function of all the genes. We also explored this model in the [S1 Text](#). In the preliminary experiments, we observed very similar outputs between the alternative model and the Net-RSTQ model shown in [S8 Table](#). However, since the alternative model directly works with one large optimization problem across all the genes, the convergence is much slower as shown in [S4 Fig](#) and the optimization package used in the experiments ran into numerical issues. Thus, we believe the Net-RSTQ model is more scalable and robust in comparison.

Currently, Net-RSTQ can scale on transcript network with up to around 5000 transcripts, which is sufficient for more focused analysis of several thousand genes. The running time of Net-RSTQ on such large transcript network is below 2 hours on each TCGA sample, compared with 5–8 hours needed for aligning the short reads. To further scale up Net-RSTQ, we will investigate other faster strategies of utilizing short read information, such as Sailfish [\[37\]](#) which directly estimates isoform expressions by counting k-mer occurrences in reads rather than reads from the alignments. This will be our future direction.

Supporting Information

S1 Text. Alternative model by network-based regularization.

(PDF)

S2 Text. Steps of generating the simulation data.

(PDF)

S3 Text. Cufflinks and RSEM command line.

(PDF)

S1 Fig. Examples of transcript sub-networks with co-expression information. (A) Transcripts in WHSC1L1 with correlation coefficients calculated on the OV dataset. (B) Transcripts in BRD4 with correlation coefficients calculated on the BRCA dataset. Both examples are shown with the neighbors in the small transcript network.

(PDF)

S2 Fig. Evaluation by qRT-PCR experiments. The relative abundance of the transcripts in 7 tested genes in H9 stem cell line (A), 5 tested genes in OVCAR8 ovarian cancer cell line (B),

and 13 tested genes in MCF7 breast cancer cell line (C) estimated by Net-RSTQ, base EM, Cufflinks and RSEM was compared with the qRT-PCR experiments. The total abundance is normalized to 1 over the measured transcripts in each gene.

(PDF)

S3 Fig. Model selection and convergence. The experiment done on BRCA (survival) dataset. (A) Effect of varying λ on the classification performance. The plot shows the average AUC learned from the 100 repeats on validation set for different λ s with the optimal λ in blue. (B) Convergence analysis by the total log-likelihood. The plot shows the change of total log-likelihood in Net-RSTQ with each gene update. Each red cross indicates the end of each round t in line 2 of Algorithm 1.

(PDF)

S4 Fig. (A) Convergence and (B) Running time of the alternative regularized framework with 2000 iterations on MCF7 breast cancer cell line.

(PDF)

S1 Table. Primer sets of the transcripts in seven genes of H9 stem cell line. * The numbers refer to the isoforms in the first column.

(PDF)

S2 Table. Primer sets of the transcripts in five genes of OVCAR8 cancer cell line. * Gene contains more transcript(s) which can not be quantified by qRT-PCR.

(PDF)

S3 Table. Primer sets of the transcripts in thirteen genes of MCF7 cancer cell line. * Gene contains more transcript(s) which can not be quantified by qRT-PCR.

(PDF)

S4 Table. Overlapped KEGG pathways with large transcript network. We consider the sub-network of genes that are members of one KEGG pathway and calculated the density of DDIs in the subnetwork.

(PDF)

S5 Table. qRT-PCR results on H9 stem cell line. * Standard deviation of Iso1 + Iso3 is 5.7% and Iso3 is 4.4%

(PDF)

S6 Table. qRT-PCR results on OVCAR8 cancer cell line. * Gene contains more transcript which can not be quantified by qRT-PCR.

(PDF)

S7 Table. qRT-PCR results on MCF7 cancer cell line. * Gene contains more transcript(s) which can not be quantified by qRT-PCR.

(PDF)

S8 Table. Correlation Coefficients between the results of Net-RSTQ and the alternative regularized framework with different λ s. The highest correlation coefficients for each λ in the alternative regularized framework is bold.

(PDF)

Acknowledgments

The results are based upon data generated by The Cancer Genome Atlas established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who

constitute the TCGA research network can be found at <http://cancergenome.nih.gov>. The dbGaP accession number to the specific version of the TCGA dataset is phs000178.v8.p7.

Author Contributions

Conceived and designed the experiments: WZ RK. Performed the experiments: WZ JWC LL KM. Analyzed the data: WZ BW RK. Contributed reagents/materials/analysis tools: JWC LL KM BW JC JY HZ. Wrote the paper: WZ JC JY HZ RK.

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*. 2008; 5(7):621–628. doi: [10.1038/nmeth.1226](https://doi.org/10.1038/nmeth.1226)
2. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456(7221):470–476. doi: [10.1038/nature07509](https://doi.org/10.1038/nature07509) PMID: [18978772](https://pubmed.ncbi.nlm.nih.gov/18978772/)
3. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*. 2011; 108(50):19867–19872. doi: [10.1073/pnas.1113972108](https://doi.org/10.1073/pnas.1113972108)
4. Li W, Kang S, Liu CC, Zhang S, Shi Y, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Research*. 2014; 42(6):e39. doi: [10.1093/nar/gkt1362](https://doi.org/10.1093/nar/gkt1362) PMID: [24369432](https://pubmed.ncbi.nlm.nih.gov/24369432/)
5. Yang EW, Girke T, Jiang T. Differential Gene Expression Analysis Using Coexpression and RNA-Seq Data. *Bioinformatics*. 2013; 29(17):2153–2161. doi: [10.1093/bioinformatics/btt363](https://doi.org/10.1093/bioinformatics/btt363) PMID: [23793751](https://pubmed.ncbi.nlm.nih.gov/23793751/)
6. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, et al. Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology*. 2012;(1):46–53. doi: [10.1038/nbt.2450](https://doi.org/10.1038/nbt.2450) PMID: [23222703](https://pubmed.ncbi.nlm.nih.gov/23222703/)
7. Roberts A, Trapnell C, Donaghey J, Rinn J, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*. 2011; 12(3):R22. doi: [10.1186/gb-2011-12-3-r22](https://doi.org/10.1186/gb-2011-12-3-r22) PMID: [21410973](https://pubmed.ncbi.nlm.nih.gov/21410973/)
8. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2010; 26(4):493–500. doi: [10.1093/bioinformatics/btp692](https://doi.org/10.1093/bioinformatics/btp692) PMID: [20022975](https://pubmed.ncbi.nlm.nih.gov/20022975/)
9. Turro E, Su SY, Goncalves A, Coin L, Richardson S, et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*. 2011; 12(2):R13. doi: [10.1186/gb-2011-12-2-r13](https://doi.org/10.1186/gb-2011-12-2-r13) PMID: [21310039](https://pubmed.ncbi.nlm.nih.gov/21310039/)
10. Huang Y, Hu Y, Jones CD, MacLeod JN, Chiang DY, et al. A Robust Method for Transcript Quantification with RNA-Seq Data. *Journal of Computational Biology*. 2013; 20(3):167–187. doi: [10.1089/cmb.2012.0230](https://doi.org/10.1089/cmb.2012.0230) PMID: [23461570](https://pubmed.ncbi.nlm.nih.gov/23461570/)
11. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*. 2009; 25(8):1026–1032. doi: [10.1093/bioinformatics/btp113](https://doi.org/10.1093/bioinformatics/btp113) PMID: [19244387](https://pubmed.ncbi.nlm.nih.gov/19244387/)
12. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, et al. Increase of functional diversity by alternative splicing. *Trends in Genetics*. 2003; 19(3):124–128. doi: [10.1016/S0168-9525\(03\)00023-4](https://doi.org/10.1016/S0168-9525(03)00023-4) PMID: [12615003](https://pubmed.ncbi.nlm.nih.gov/12615003/)
13. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, et al. Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome. *Journal of Proteome Research*. 2004; 3(1):76–83. doi: [10.1021/pr034064v](https://doi.org/10.1021/pr034064v) PMID: [14998166](https://pubmed.ncbi.nlm.nih.gov/14998166/)
14. Tseng YT, Li W, Chen CH, Zhang S, Chen J, et al. IIIDB: a database for isoform-isoform interactions and isoform network modules. *BMC Genomics*. 2015; 16(Suppl 2):S10. doi: [10.1186/1471-2164-16-S2-S10](https://doi.org/10.1186/1471-2164-16-S2-S10) PMID: [25707505](https://pubmed.ncbi.nlm.nih.gov/25707505/)
15. Finn RD, Marshall M, Bateman A. iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*. 2005; 21(3):410–412. doi: [10.1093/bioinformatics/bti011](https://doi.org/10.1093/bioinformatics/bti011) PMID: [15353450](https://pubmed.ncbi.nlm.nih.gov/15353450/)
16. Stein A, Ceol A, Aloy P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*. 2011; 39(Database-Issue):718–723. doi: [10.1093/nar/gkq962](https://doi.org/10.1093/nar/gkq962)
17. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate JG, et al. The Pfam protein families database. *Nucleic Acids Research*. 2012; 40(Database-Issue):290–301. Available from: <http://dblp.uni-trier.de/db/journals/nar/nar40.html#PuntaCEMTBPFCCHHSEBF12>. doi: [10.1093/nar/gkr1065](https://doi.org/10.1093/nar/gkr1065)

18. Pruitt KD, Tatusova TA, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*. 2012; 40(Database-Issue):130–135. Available from: <http://dblp.uni-trier.de/db/journals/nar/nar40.html#PruittTBM12>. doi: [10.1093/nar/gkr1079](https://doi.org/10.1093/nar/gkr1079)
19. Berman HM, Westbrook JD, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000; 28(1):235–242. Available from: <http://dblp.uni-trier.de/db/journals/nar/nar28.html#BermanWFGBWSB00>. doi: [10.1093/nar/28.1.235](https://doi.org/10.1093/nar/28.1.235) PMID: [10592235](https://pubmed.ncbi.nlm.nih.gov/10592235/)
20. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*. 2011; 39(Database-Issue):730–735. Available from: <http://dblp.uni-trier.de/db/journals/nar/nar39.html#YellaboinaTZRJ11>. doi: [10.1093/nar/gkq1229](https://doi.org/10.1093/nar/gkq1229)
21. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. Human Protein Reference Database—2009 update. *Nucleic Acids Research*. 2009; 37(Database-Issue):767–772. Available from: <http://dblp.uni-trier.de/db/journals/nar/nar37.html#PrasadGKKKMTSRVMBSSRRKKAKMRKRMRRCP09>. doi: [10.1093/nar/gkn892](https://doi.org/10.1093/nar/gkn892)
22. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. A census of human cancer genes. *Nature Reviews Cancer*. 2004 Mar; 4(3):177–183. doi: [10.1038/nrc1299](https://doi.org/10.1038/nrc1299) PMID: [14993899](https://pubmed.ncbi.nlm.nih.gov/14993899/)
23. Higgins ME, Claremont M, Major JE, Sander C, Lash AE. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Research*. 2007; 35(suppl 1):D721–D726. Available from: http://nar.oxfordjournals.org/content/35/suppl_1/D721.abstract. doi: [10.1093/nar/gkl811](https://doi.org/10.1093/nar/gkl811) PMID: [17088289](https://pubmed.ncbi.nlm.nih.gov/17088289/)
24. Xing Y, Yu T, Wu YN, Roy M, Kim J, et al. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic acids research*. 2006; 34(10):3150–3160. doi: [10.1093/nar/gkl396](https://doi.org/10.1093/nar/gkl396) PMID: [16757580](https://pubmed.ncbi.nlm.nih.gov/16757580/)
25. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2010; 26(4):493–500. Available from: <http://bioinformatics.oxfordjournals.org/content/26/4/493.abstract>. doi: [10.1093/bioinformatics/btp692](https://doi.org/10.1093/bioinformatics/btp692) PMID: [20022975](https://pubmed.ncbi.nlm.nih.gov/20022975/)
26. Rozov R, Halperin E, Shamir R. MGMR: leveraging RNA-Seq population data to optimize expression estimation. *BMC Bioinformatics*. 2012; 13(S-6):S2. PMID: [22537041](https://pubmed.ncbi.nlm.nih.gov/22537041/)
27. Pachter L. Models for transcript quantification from RNA-Seq. *ArXiv*. 2011;p. 1104.3889.
28. Tanaka Y, Kim KY, Zhong M, Pan X, Weissman SM, et al. Transcriptional regulation in pluripotent stem cells by methyl CpG-binding protein 2 (MeCP2). *Human Molecular Genetics*. 2014; 23(4):1045–1055. doi: [10.1093/hmg/ddt500](https://doi.org/10.1093/hmg/ddt500) PMID: [24129406](https://pubmed.ncbi.nlm.nih.gov/24129406/)
29. Daemen A, Griffith O, Heiser L, Wang N, Enache O, et al. Modeling precision treatment of breast cancer. *Genome Biology*. 2013; 14(10):R110. doi: [10.1186/gb-2013-14-10-r110](https://doi.org/10.1186/gb-2013-14-10-r110) PMID: [24176112](https://pubmed.ncbi.nlm.nih.gov/24176112/)
30. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. 2013; 14(4):R36. doi: [10.1186/gb-2013-14-4-r36](https://doi.org/10.1186/gb-2013-14-4-r36) PMID: [23618408](https://pubmed.ncbi.nlm.nih.gov/23618408/)
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25(16):2078–2079. doi: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) PMID: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/)
32. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*. 2010; 28(5):511–515. doi: [10.1038/nbt.1621](https://doi.org/10.1038/nbt.1621)
33. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011; 12:323. Available from: <http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi12.html#LiD11>. doi: [10.1186/1471-2105-12-323](https://doi.org/10.1186/1471-2105-12-323) PMID: [21816040](https://pubmed.ncbi.nlm.nih.gov/21816040/)
34. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 2000; 28(1):27–30. Available from: <http://nar.oxfordjournals.org/content/28/1/27.abstract>. doi: [10.1093/nar/28.1.27](https://doi.org/10.1093/nar/28.1.27) PMID: [10592173](https://pubmed.ncbi.nlm.nih.gov/10592173/)
35. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*. 2012; 40(20):10073–10083. doi: [10.1093/nar/gks666](https://doi.org/10.1093/nar/gks666) PMID: [22962361](https://pubmed.ncbi.nlm.nih.gov/22962361/)
36. Vapnik VN. *Statistical Learning Theory*. Wiley-Interscience; 1998.
37. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotech*. 2014 May; 32(5):462–464. doi: [10.1038/nbt.2862](https://doi.org/10.1038/nbt.2862)