SUPPLEMENTARY METHODS FOR THE PAPER TRANSCRIPT ASSEMBLY AND QUANTIFICATION BY RNA-SEQ REVEALS UNANNOTATED TRANSCRIPTS AND ISOFORM SWITCHING DURING CELL DIFFERENTIATION

COLE TRAPNELL, BRIAN A WILLIAMS, GEO PERTEA, ALI MORTAZAVI, GORDON KWAN, MARIJKE J VAN BAREN, STEVEN L SALZBERG, BARBARA J WOLD, AND LIOR PACHTER

This document is a companion to the paper "Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms". It describes the experiment performed, the details of the Cufflinks assembler and abundance estimation methods, and provides supplementary tables, figures and supporting evidence for claims in the paper.

Contents

Nature Biotechnology: doi:10.1038/nbt.1621

List of Figures

List of Tables

1. Sequencing experiment

The data analyzed in the paper consisted of 430,467,018 paired 75bp reads sequenced from the transcriptome of mouse skeletal muscle C2C12 cells induced to undergo myogenic differentiation. Total RNA was extracted from these cells, and subsequently mRNA was isolated at four different time points (-24 hours, 60 hours, 120 hours, 168 hours). cDNA was prepared following a similar procedure to the one described in [17]. Fragmentation of the mRNA followed by size selection resulted in fragment lengths approximately 200nt long for all of the time-points. The distribution of fragment lengths is shown in Supplementary Figure 1 (in Section 3 this distribution of fragment lengths is referred to as F). These estimates are based on alignments of the spiked-in sequences using `Bowtie` 0.12 [12].

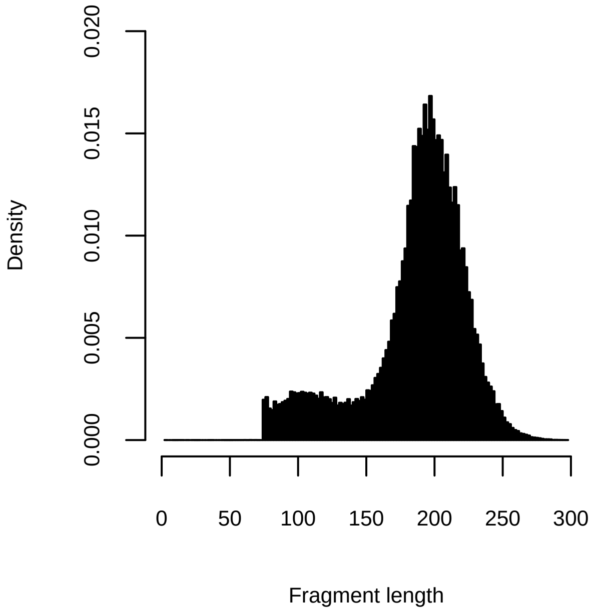


Figure 1. Fragment length distributions of C2C12 time-course libraries.

2. Mapping fragments to the genome

In principle, an algorithm that infers individual transcript abundances by measuring the fraction of fragments originating from each of a set of known transcripts would begin by computing alignments between fragments and the set of known transcripts that may be contained in the sample. However, because the transcriptome for mouse is incompletely annotated, such an analysis requires mapping of fragments to the genome as a proxy for mapping directly to transcripts. This means that alignments of short sequencing reads must be allowed to span exon-exon splice junctions in genomic coordinate space. We previously developed a program called `TopHat` to map RNA-Seq reads to the genome. TopHat does not require a reference transcriptome and can therefore be used to discover novel splice junctions. [24]

Fragments were mapped to build 37.1 of the mouse genome with TopHat version 1.0.13. We extended our previous algorithms described in [24] to exploit the longer paired reads used in the study. The original TopHat program used a seed-and-extend alignment strategy to find spliced alignments of unpaired RNA-Seq experiments. However, due to computational limitations, our original method reported only alignments across GT-AG introns shorter than 20Kb by default. This strategy also could not align reads that spanned multiple splice junctions. However, as sequencing technology has improved and longer (paired end) reads have become available, we have modified the software to employ new strategies to align reads across splice junctions. TopHat version 1.0.7 and later splits a read 75bp or longer in three or more segments of approximately equal size (25bp), and maps them independently. Reads with segments that can be mapped to the genome only non-contiguously are marked as possible intron-spanning reads. These "contiguously unmappable" reads are used to build a set of possible introns in the transcriptome.

2.1. **Discovering splice junctions.** Suppose S is a read of length l that crosses a splice junction. TopHat splits S into $n = \lfloor l/k \rfloor$ segments where $k = 25$bp (k is a parameter that can be adjusted by the user in TopHat). At most one of these segments must cross the splice junction, and junctions can be discovered if they lie in any of the segments. TopHat maps the segments $s_1, ..., s_n$ with Bowtie to the genome, and checks for internal segments $s_2, ..., s_{n-1}$ that do not map anywhere to the genome, as well as for pairs of successive segments s_i, s_{i+1} that both align to the genome, but not adjacently. When a segment s_i fails to align because it crosses a splice junction, but s_{i-1} and s_{i+1} are aligned (say at starting at positions x and y, respectively), TopHat looks for the donor and acceptor sites for the junction near x and y. Assuming (without loss of generality) that the transcript is on Crick strand of the genome the donor must fall within k bases upstream of position $x + k$, and the acceptor must be within k bases downstream of y, a total of k possible exon-exon splice junctions. Similarly, when successive segments s_i and s_{i+1} align to the genome non-adjacently at positions x and y, the junction spanned by the read must be from positions $x + k$ to y in the genome.

TopHat accumulates an index of potential splice junctions by examining segment mapping for all contiguously unmappable reads. For each junction the program then concatenates $k$bp upstream of the donor to $k$bp downstream of the acceptor to form a synthetic spliced sequence around the junction. The segments of the contiguously unmappable reads are then aligned against these synthetic sequences with Bowtie. The resulting contiguous and spliced segment alignments for these reads are merged to form complete alignments to the genome, each spanning one or more splice junctions.

2.2. **Resolving multiple alignments for fragments.** The alignments for both reads from a mate pair are examined together to produce a set of alignments for the corresponding library fragment as a whole, reported in SAM format [14]. These fragment alignments are ranked according to the procedure described below, and only highest ranking alignments are reported. The ranks are designed to incorporate very loose assumptions on intron and gene length, namely that introns longer than 20kb are rare. Let x and y be fragment alignments. Then $x < y$ if *any* of the following (applied in order) are true:

(1) x is a singleton, and y has both ends mapped,
(2) x crosses more splice junctions than y,
(3) the reads from x map significantly farther apart in the genome than expected according to the library's fragment length distribution (≥ 3 s.d.), and the reads from y do not,
(4) the reads from x are significantly closer together than expected according to the library's fragment length distribution, and the reads from y are not,
(5) The reads from x map more than 100bp farther apart than the reads from y,
(6) x and y both span an intron, and x spans a longer one,
(7) x has more mismatches than y to the genome.

Fragments that have multiple equally good alignments according to the above rules are all reported. If there are n alignments for a fragment, each is assigned a probability of only $1/n$ of being correct. The SAM format encodes this probability in the mapping quality field, which is later used by `Cufflinks` to reduce the contribution of multiply mapping fragments (to $1/n$ of a uniquely mappable read) in FPKM calculations (FPKM is a measurement of expression, and is formally defined in Section 3). The recent work of Li et al. [13] addresses the problem of probabilistically assigning multi-reads, and it should be possible to incorporate the ideas of that paper into future versions of `Tophat` and `Cufflinks`.

Sample	**Sequenced fragments**	**Aligned fragments**	**Singleton fragments**	**Spliced fragments**	**Multi-mapping fragments**	**Total alignments**
-24 hours	42,184,539	35,852,366	11,031,886	8,824,825	1,768,041	41,663,170
60 hours	70,192,031	57,071,494	18,104,211	15,778,114	2,265,378	64,637,511
120 hours	41,069,106	27,914,989	14,431,734	7,711,026	1,881,772	33,929,133
168 hours	61,787,833	50,705,080	20,396,250	14,585,287	2,458,292	58,797,912
Total	215,233,509	171,543,929	63,964,081	46,899,252	8,373,483	199,027,726

TABLE 1. Number of fragments sequenced, aligned and mapped with `TopHat`. Singleton fragments are fragments for which only one end could be mapped. Spliced fragments include at least one end that maps across a junction. The numbers in the total alignment column may not be the sum of the entries in each row, because some fragments fall into multiple classes.

3. Transcript abundance estimation

3.1. **Definitions.** A *transcript* is an RNA molecule that has been transcribed from DNA. A *primary transcript* is an RNA molecule that has yet to undergo modification. The *genomic location* of a primary transcript consists of a pair of coordinates in the genome representing the $5'$ transcription start site and the $3'$ polyadenylation cleavage site. We denote the set of all transcripts in a transcriptome by T. We partition transcripts into *transcription loci* (for simplicity we refer to these as loci) so that every locus contains a set of transcripts all of whose genomic locations do not overlap the genomic location of any transcript in any other locus. Formally, we consider a maximal partition of transcripts into loci, a partition denoted by G, where the genomic location of a transcript $t \in g \in G$ does not overlap the genomic location of any transcript u where $u \in h \in G$ and $h \neq g$. We emphasize that the definition of a transcription locus is not biological; transcripts in the same locus may be regulated via different promoters, and may differ completely in sequence (for example if one transcript is in the intron of another) or have different functions. The reason for defining loci is that we will see that they are computationally convenient.

We assume that at the time of an experiment, a transcriptome consists of an ensemble of transcripts T where the proportion of transcript $t \in T$ is ρ_t, so that $\sum_{t \in T} \rho_t = 1$ and $0 \leq \rho_t \leq 1$ for all $t \in T$. Formally, a *transcriptome* is a set of transcripts T together with the abundances $\rho = \{\rho_t\}_{t \in T}$. For convenience we also introduce notation for the proportion of transcripts in each locus. We let $\sigma_g = \sum_{t \in g} \rho_t$. Similarly, within a locus g, we denote the proportion of each transcript $t \in g$ by $\tau_t = \frac{\rho_t}{\sigma_g}$. We refer to ρ, σ and τ as *transcript abundances.*

Transcripts have lengths, which we denote by $l(t)$. For a collection of transcripts $S \subset T$ in a transcriptome, we define the length of S using the weighted mean:

$$l(S) = \frac{\sum_{t \in S} \rho_t l(t)}{\sum_{t \in S} \rho_t}. \tag{1}$$

It is important to note that the length of a set of transcripts depends on their relative abundances; the reason for this will be clear later.

One grouping of transcripts that we will focus on is the set of transcripts within a locus that share the same transcription start site (TSS). Unlike the concept of a locus, grouping by TSS has a biological basis. Transcripts within such a group are by definition alternatively spliced, and if they have different expression levels, this is most likely due to the spliceosome and not due to differences in transcriptional regulation.

3.2. **A statistical model for RNA-Seq.** In order to analyze expression levels of transcripts with RNA-Seq data, it is necessary to have a model for the (stochastic) process of sequencing. A *sequencing experiment* consists of selecting a total of M fragments of transcripts uniformly at random from the transcriptome. Each fragment is identified by sequencing from its ends, resulting in two reads called *mate pairs.* The length of a fragment is a random variable, with a distribution we will denote by F. That is, the probability that a fragment has length i is $F(i)$ and $\sum_{i=1}^{\infty} F(i) = 1$. In this paper we assume that F is normal, however in principle F can be estimated using data from the experiment (e.g. spike-in sequences). We decided to use the normal approximation to F (allowing for user specified parameters of the normal distribution) in order to simplify the requirements for running `Cufflinks` at this time.

The assumption of random fragment selection is known to oversimplify the complexities of a sequencing experiment, however without rigorous ways to normalize we decided to work with the uniform at random assumption. It is easy to adapt the model to include more complex models that address sequencing bias as RNA-Seq experiments mature and the technologies are better understood.

The transcript abundance estimation problem in paired-end RNA-Seq is to estimate ρ given a set of transcripts T and a set of reads sequenced from the ends of fragments. In `Cufflinks`, the transcripts T can be specified by the user, or alternatively T can be estimated directly from the reads. The latter problem is the transcript assembly problem which we discuss in Section 4. We ran `Cufflinks` in the latter "discovery" mode where we assembled the transcripts without using the reference annotation.

The fact that fragments have different lengths has bearing on the calculation of the probability of selecting a fragment from a transcript. Consider a transcript t with length $l(t)$. The probability of selecting a fragment of length k from t at one of the positions in t assuming that it is selected uniformly at random, is $\frac{1}{l(t)-k}$. For this reason, we will define an adjusted length for transcripts as

$$\tilde{l}(t) = \sum_{i=1}^{l(t)} F(i)(l(t) - i + 1). \tag{2}$$

We also revisit the definition of length for a group of transcripts, and define

$$\tilde{l}(S) = \frac{\sum_{t \in S} \rho_t \tilde{l}(t)}{\sum_{t \in S} \rho_t}. \tag{3}$$

It is important to note that given a read it may not be obvious from which transcript the fragment it was sequenced from originated. The consistency of fragments with transcripts is important and we define the *fragment-transcript matrix* $A_{R,T}$ to be the $M \times |T|$ matrix with $A(r,t) = 1$ if the fragment alignment r is completely contained in the genomic interval spanned by t, and all the implied introns in r match introns in t (in order), and with $A(r,t) = 0$ otherwise. Note that the reads in Figure 1c in the main text are colored according to the matrix $A_{R,T}$, with each column of the matrix corresponding to one of the three colors (yellow, blue, red) and reads colored according to the mixture of colors corresponding to the transcripts their fragments are contained in.

Even given the read alignment to a reference genome, it may not be obvious what the length of the fragment was. Formally, in the case that $A_{R,T}(r,t) = 1$ we denote by $I_t(r)$ the fragment length from within a transcript t implied by the (presumably unique) sequences corresponding to the mate pairs of a fragment r. If $A_{R,T}(r,t) = 0$ then $I_t(r)$ is set to be infinite and $F(I_t(r)) = 0$.

Given a set of reads, we assume that we can identify for each of them the set of transcripts with which the fragments the reads belonged to are consistent. The rationale for this assumption is the following: we map the reads to a reference genome, and we assume that the read lengths are sufficiently long so that every mate-pair can be uniquely mapped to the genome. We refer to this mapping as the *fragment alignment*. We also assume that we know all the possible transcripts and their alignments to the genome. Therefore, we can identify for each read the possible transcripts from which the fragment it belonged to originated.

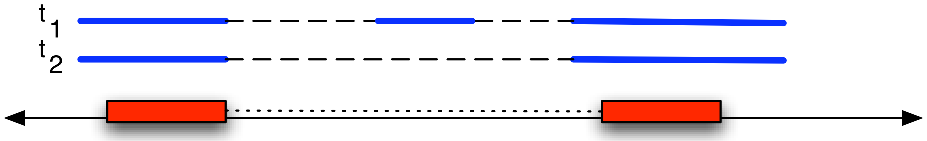


FIGURE 2. Alignments of reads to the genome (rectangles) may be consistent with multiple transcripts (in this case both t_1 and t_2). The transcripts t_1 and t_2 differ by an internal exon; introns are indicated by long dashed lines. If we denote the fragment alignment by r, this means that $A_{R,T}(r, t_1) = 1$ and $A_{R,T}(r, t_2) = 1$. It is apparent that the implied length $I_{t_1}(r) > I_{t_2}(r)$ due to the presence of the extra internal exon in t_1.

We are now ready to write down the likelihood equation for the model. We will write $L(\rho|R)$ for the likelihood of a set of fragment alignments R constructed from M reads. The notation $Pr(trans. = t)$ means "the probability that a fragment selected at random originates from transcript t".

$$L(\rho|R) = \prod_{r\in R} Pr(rd.\ aln. = r) \tag{4}$$

$$= \prod_{r\in R}\sum_{t\in T} Pr(rd.\ aln. = r|trans. = t)Pr(trans. = t) \tag{5}$$

$$= \prod_{r\in R}\sum_{t\in T} \frac{\rho_t \tilde{l}(t)}{\sum_{u\in T}\rho_u \tilde{l}(u)} Pr(rd.\ aln. = r|trans. = t) \tag{6}$$

$$= \prod_{r\in R}\sum_{t\in T} \frac{\rho_t \tilde{l}(t)}{\sum_{u\in T}\rho_u \tilde{l}(u)} \left(\frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right) \tag{7}$$

$$= \prod_{r\in R}\sum_{t\in T} \alpha_t \left(\frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right), \tag{8}$$

where

$$\alpha_t = \frac{\rho_t \tilde{l}(t)}{\sum_{u\in T}\rho_u \tilde{l}(u)}. \tag{9}$$

Observe that α_t is exactly the probability that a fragment selected at random comes from transcript t, and we have that $\sum_{t \in T} \alpha_t = 1$. In light of the probabilistic meaning of the $\alpha = \{\alpha_t\}_{t \in T}$, we refer to them as *fragment abundances.*

It is evident that the likelihood function is that of a linear model and that the likelihood function is concave (Proposition 15) so a numerical method can be used to find the α. It is then possible, in principle, to recover the ρ using Lemma 14. However the number of parameters is in the tens of thousands, and in practice this form of the likelihood function is unwieldy. Instead, we re-write the likelihood utilizing the fact that transcripts in distinct loci do not overlap in genomic location.

We first calculate the probability that a fragment originates from a transcript within a given locus g:

$$\beta_g := \sum_{t \in g} \alpha_t \tag{10}$$

$$= \frac{\sum_{t \in g} \rho_t \tilde{l}(t)}{\sum_{u \in T} \rho_u \tilde{l}(u)} \tag{11}$$

$$= \frac{\sum_{t \in g} \sigma_g \tau_t \tilde{l}(t)}{\sum_{h \in G} \sum_{u \in h} \sigma_h \tau_u \tilde{l}(u)} \tag{12}$$

$$= \frac{\sigma_g \sum_{t \in g} \tau_t \tilde{l}(t)}{\sum_{h \in G} \sigma_h \sum_{u \in h} \tau_u \tilde{l}(u)} \tag{13}$$

$$= \frac{\sigma_g \tilde{l}(g)}{\sum_{h \in G} \sigma_h \tilde{l}(h)}. \tag{14}$$

Recall that $\sigma_g = \sum_{t \in g} \rho_t$ and that $\tau_t = \frac{\rho_t}{\sigma_g}$ for a locus g.

Similarly, the probability of selecting a fragment from a single transcript t conditioned on selecting a transcript from the locus g in which t is contained is

$$\gamma_t = \frac{\tau_t \tilde{l}(t)}{\sum_{u \in g} \tau_u \tilde{l}(u)}. \tag{15}$$

The parameters $\gamma = \{\gamma_t\}_{t \in g}$ are conditional fragment abundances, and they are the parameters we estimate from the data in the next Section. Note that for a transcript $t \in g$, $\alpha_t = \beta_g \cdot \gamma_t$ and it is easy to convert between fragment abundances and transcript abundances using Lemma 14.

We denote the fragment counts by X; specifically, we denote the number of alignments in locus g by X_g. Note that $\sum_{g \in G} X_g = M$. We also use the notation g_r to denote the (unique) locus from which a read alignment r can be obtained.

The likelihood function is given by

$$L(\rho|R) = \prod_{r\in R} Pr(rd.\ aln. = r) \tag{16}$$

$$= \prod_{r\in R}\sum_{g\in G} Pr(rd.\ aln. = r|locus = g)Pr(locus = g) \tag{17}$$

$$= \prod_{r\in R}\frac{\sigma_{g_r}\tilde{l}(g_r)}{\sum_{g\in G}\sigma_g\tilde{l}(g)} Pr(rd.\ aln. = r|locus = g_r) \tag{18}$$

$$= \prod_{r\in R}\beta_{g_r}\sum_{t\in g_r} Pr(rd.\ aln. = r|locus = g_r, trans. = t)Pr(trans. = t|locus = g_r) \tag{19}$$

$$= \prod_{r\in R}\beta_{g_r}\sum_{t\in g_r}\frac{\tau_t\tilde{l}(t)}{\sum_{u\in g_r}\tau_u\tilde{l}(u)} Pr(rd.\ aln. = r|locus = g_r, trans. = t) \tag{20}$$

$$= \left(\prod_{r\in R}\beta_{g_r}\right)\left(\prod_{r\in R}\sum_{t\in g}\gamma_t\cdot Pr(rd.\ aln. = r|locus = g_r, trans. = t)\right) \tag{21}$$

$$= \left(\prod_{r\in R}\beta_{g_r}\right)\left(\prod_{r\in R}\sum_{t\in g}\gamma_t\cdot\frac{F(I_t(r))}{l(t)-I_t(r)+1}\right) \tag{22}$$

$$= \left(\prod_{g\in G}\beta_g^{X_g}\right)\left(\prod_{g\in G}\left(\prod_{r\in R:r\in g}\sum_{t\in g}\gamma_t\cdot\frac{F(I_t(r))}{l(t)-I_t(r)+1}\right)\right). \tag{23}$$

Explicitly, in terms of the parameters ρ, Equation (23) simplifies to Equation (8) but we will see in the next section how the maximum likelihood estimates $\hat{\rho}$ are most conveniently obtained by first finding $\hat{\beta}$ and $\hat{\gamma}$ using Equation (23).

We note that it is biologically meaningful to include prior distributions on σ and τ that reflect the inherent stochasticity and resulting variability of transcription in a cell. This will be an interesting direction for further research as more RNA-Seq data (with replicates) becomes available allowing for the determination of biologically meaningful priors. In particular, it seems plausible that specific isoform abundances may vary considerably and randomly within cells from a single tissue and that this may be important in studying differential splicing. We mention to this to clarify that in this paper, the confidence intervals we report represent the variability in the maximum likelihood estimates $\hat{\sigma}_j$ and $\hat{\tau}_j^k$, and are not the variances of prior distributions.

3.3. **Estimation of parameters.** We begin with a discussion of identifiability of our model. Identifiability refers to the injectivity of the model, i.e.,

$$\text{if } Pr_{\rho_1}(r) = Pr_{\rho_2}(r)\ \forall r\in R,\quad \text{then } \rho_1 = \rho_2. \tag{24}$$

The identifiability of RNA-Seq models was discussed in [9], where a standard analysis for linear models is applied to RNA-Seq (for another related biological example, see [20] which discusses identifiability of haplotypes in mixed populations from genotype data). The results in these papers apply to our model. For completeness we review the conditions for identifiability. Recall that $A_{R,T}$ is the fragment-transcript matrix that specifies which transcripts each fragment is compatible with. The following theorem provides a simple characterization of identifiability:

Theorem 1. *The RNA-Seq model is identifiable iff $A_{R,T}$ is full rank.*

Therefore, for a given set of transcripts and a read set R, we can test whether the model is identifiable using elementary linear algebra. For the results in this paper, when estimating expression with given annotations, when the model was not identifiable we picked *a* maximum likelihood solution, although in principle it is possible to bound the total expression of the locus and/or report identifiability problems to the user.

Returning to the likelihood function

$$\left(\prod_{g\in G}\beta_g^{X_g}\right)\left(\prod_{g\in G}\left(\prod_{r\in R:r\in g}\sum_{t\in g}\gamma_t\cdot\frac{F(I_t(r))}{l(t)-I_t(r)+1}\right)\right), \tag{25}$$

we note that both the β and γ parameters depend on the ρ parameters. However, we will see that if we maximize the β separately from the γ, and also each of the sets $\{\gamma_t : t \in g\}$ separately, then it is always possible to find ρ that match both the maximal β and γ. In other words, the problem of finding $\hat{\rho}$ is equivalent to finding $\hat{\beta}$ that maximizes $\prod_{g\in G}\beta_g^{X_g}$ and separately, for each locus g, the $\hat{\gamma}_t$ that maximize

$$\prod_{r\in R:r\in g}\sum_{t\in g}\gamma_t\frac{F(I_t(r))}{l(t)-I_t(r)+1}. \tag{26}$$

We begin by solving for the $\hat{\beta}$ and $\hat{\gamma}$ and the variances of the maximum likelihood estimates, and then explain how these are used to report expression levels.

We can solve for the $\hat{\gamma}$ using the fact that the model is linear. That is, the probability of each individual read is linear in the read abundances γ_t. It is a standard result in statistics (see, e.g., Proposition 1.4 in [19]) that the log likelihood function of a linear model is concave. Thus, a hill climbing method can be used to find the $\hat{\gamma}$. We used the EM algorithm for this purpose.

Rather than using the direct ML estimates, we obtained a regularized estimate by importance sampling from the posterior distribution with a proposal distribution we explain below. The samples were also used to estimate variances for our estimates.

It follows from standard MLE asymptotic theory that the $\hat{\gamma}$ are asymptotically multivariate normal with variance-covariance matrix given by the inverse of the observed Fisher information matrix. This matrix is defined as follows:

Definition 2 (Observed Fisher information matrix)**.** The observed Fisher information matrix is the negative of the Hessian of the log likelihood function evaluated at the maximum likelihood estimate. That is, for parameters $\Theta = (\theta_1,\ldots,\theta_n)$, the $n \times n$ matrix is

$$\mathcal{F}_{k,l}(\hat{\Theta}) \;=\; -\frac{\partial^2 log(\mathcal{L}(\Theta|R))}{\partial\theta_k\theta_l}|_{\theta=\hat{\theta}}. \tag{27}$$

In our case, considering a single locus g, the parameters are $\Theta = (\gamma_{t_1}, \dots, \gamma_{t_{|g|}})$, and as expected from Proposition 15:

$$(28)\, \mathcal{F}_{t_k,t_l}(\hat{\Theta}) = \sum_{r \in R : r \in g} \left[\frac{1}{\left(\sum_{h \in g} \hat{\gamma}_h \frac{F(I_h(r))}{l(h) - I_h(r) + 1} \right)^2} \frac{F(I_{t_k}(r))F(I_{t_l}(r))}{(l(t_k) - I_{t_k} + 1)(l(t_l) - I_{t_l} + 1)} \right].$$

Because some of the transcript abundances may be close to zero, we adopted the Bayesian approach of [11] and instead sampled from the joint posterior distribution of Θ using the proposal distribution consisting of the multivariate normal with mean given by the MLE, and variance-covariance matrix given by the inverse of (28). If the Observed Fisher Information Matrix is singular then the user is warned and the confidence intervals of all transcripts are set to $[0, 1]$ (meaning that there is no information about relative abundances).

The method used for sampling was importance sampling. The samples were used to obtain a maximum-a-posterior estimate for $\hat{\gamma}_t$ for each t and for the variance-covariance matrix which we denote by Ψ^g (where $g \in G$ denotes the locus). Note that Ψ^g is a $|g| \times |g|$ matrix. The covariance between $\hat{\gamma}_{t_k}$ and $\hat{\gamma}_{t_l}$ for $t_k, t_l \in g$ is given by $\Psi^g_{t_k,t_l}$.

Turning to the maximum likelihood estimates $\hat{\beta}$, we use the fact that the model is the log-linear. Therefore,

$$(29) \qquad \hat{\beta}_g = \frac{X_g}{M}.$$

Viewed as a random variable, the counts X_g are approximately Poisson and therefore the variance of the MLE $\hat{\beta}_g$ is approximately X_g. We note that for the tests in this paper we directly used the total counts M and the proportional counts X_g, however it is easy to incorporate recent suggestions for total count normalization, such as [3] into `Cufflinks`.

The favored units for reporting expression in RNA-Seq studies to date is not using the transcript abundances directly, but rather using a measure abbreviated as FPKM, which means "expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced". These units are equivalent to measuring transcript abundances (multiplied by a scalar). The computational advantage of FPKM, is that the normalization constants conveniently simplify some of the formulas for the variances of transcript abundance estimates.

For example, the abundance of a transcript $t \in g$ in FPKM units is

$$(30) \qquad \frac{10^6 \cdot 10^3 \cdot \alpha_t}{\tilde{l}(t)} = \frac{10^6 \cdot 10^3 \cdot \beta_g \cdot \gamma_t}{\tilde{l}(t)}.$$

Equation (30) makes it clear that although the abundance of each transcript $t \in g$ in FPKM units is proportional to the transcript abundance ρ_t it is given in terms of the read abundances β_g and γ_t which are the parameters estimated from the likelihood function.

The maximum likelihood estimates of β_g and γ_t are random variables, and we denote their scaled product (in FPKM units) by A_t. That is $Pr(A_t = a)$ is the probability that for a random set of fragment alignments from a sequencing experiment, the maximum likelihood estimate of the transcript abundance for t in FPKM units is a.

Using the fact that the expectation of a product of independent random variables is the product of the expectations, for a transcript $t \in g$ we have

$$E[A_t] = \frac{10^9 X_g \hat{\gamma}_t}{\tilde{l}(t)M}. \tag{31}$$

Given the variance estimates for the $\hat{\gamma}_t$ we turn to the problem of estimating $Var[A_t]$ for a transcript $t \in g$. We use Lemma 13 to obtain

$$\begin{aligned} Var[A_t] &= \left(\frac{10^9}{\tilde{l}(t)M}\right)^2 \left(\Psi^g_{t,t} X_g + \Psi^g_{t,t} X_g^2 + (\hat{\gamma}_t)^2 X_g\right) && (32) \\ &= X_g \left(\frac{10^9}{\tilde{l}(t)M)}\right)^2 \left(\Psi^g_{t,t}(1 + X_g) + (\hat{\gamma}_t)^2\right). && (33) \end{aligned}$$

This variance calculation can be used to estimate a confidence interval by utilizing the fact [1] that when the expectation divided by the standard deviation of at least one of two random variables is large, their product is approximately normal.

Next we turn to the problem of estimating expression levels (and variances of these estimates) for groups of transcripts. Let $S \subset T$ be a group of transcripts located in a single locus g, e.g. a collection of transcripts sharing a common TSS.

The analogy of Equation (30) for the FPKM of the group is

$$\begin{aligned} & \frac{10^6 \cdot 10^3 \cdot \beta_g \cdot \left(\sum_{t \in S} \gamma_t\right)}{\tilde{l}(S)} && (34) \\ = \; & 10^6 \cdot 10^3 \cdot \beta_g \cdot \sum_{t \in S} \frac{\gamma_t}{\tilde{l}(t)}. && (35) \end{aligned}$$

As before, we denote by B_S the random variables for which $Pr(B_S = b)$ is the probability that for a random set of fragment alignments from a sequencing experiment, the maximum likelihood estimate of the transcript abundance for all the transcripts in S in FPKM units is b. We note that the B_S are products and sums of random variables (Equation (35)). This makes Equation (35) more useful than the equivalent unsimplified Equation (34), especially because $\tilde{l}(S)$ is, in general, a ratio of two random variables.

We again use the fact that the expectation of independent random variables is the product of the expectation, in addition to the fact that expectation is a linear operator to conclude that for a group of transcripts S,

$$E[B_S] = \frac{10^9 \cdot X_g \cdot \sum_{t \in S} \frac{\hat{\gamma}_t}{\tilde{l}(t)}}{M}. \tag{36}$$

In order to compute the variance of B_S, we first note that

$$Var\left[\sum_{t\in S}\frac{\hat{\gamma}_t}{\tilde{l}(t)}\right] = \sum_{t\in S}\frac{1}{\tilde{l}(t)^2}\Psi^g_{t,t} + \sum_{t,u\in S}\frac{1}{\tilde{l}(t)\tilde{l}(u)}\Psi^g_{t,u}. \tag{37}$$

Therefore,

$$Var[B_S] = X_g\left(\frac{10^9}{M}\right)^2\left((1+X_g)\left(\sum_{t\in S}\frac{1}{\tilde{l}(t)^2}\Psi^g_{t,t} + \sum_{t,u\in S}\frac{1}{\tilde{l}(t)\tilde{l}(u)}\Psi^g_{t,u}\right) + \left(\sum_{t\in S}\frac{\hat{\gamma}_t}{\tilde{l}(t)}\right)^2\right). \tag{38}$$

We can again estimate a confidence interval by utilizing the fact that B_S is approximately normal [1].

3.4. **Assessment of abundance estimation.** We evaluated the accuracy of `Cufflinks`' transcript abundance estimates by first comparing the estimated FPKM values for the spiked-in sequences in each sample against their intended concentrations. Spike FPKMs were highly correlated across a 5-log dynamic range in all four samples (Supplementary Figure 3). However, because sequenced spike fragments were unambiguously mappable, we performed additional simulation to measure the accuracy of the software in alternatively spliced loci.

To assess the accuracy of `Cufflinks`' estimates, we simulated an RNA-Seq experiment using the FluxSimulator, a freely available software package that models whole-transcriptome sequencing experiments with the Illumina Genome Analyzer. [23] The software works by first randomly assigning expression values to the transcripts provided by the user, constructing an amplified, size-selected library, and sequencing it. Mouse UCSC transcripts were supplied to the software, along with build 37.1 of the genome. FluxSimulator then randomly assigned expression ranks to 18,935 transcripts, with the expression value y computed from the rank x according to the formula

$$y = \left(\frac{x}{5.0\times 10^7}\right)^{-0.6} e^{-\left(\frac{x}{9.5\times 10^3}\right)-\left(\frac{x}{9.5\times 10^3}\right)^2}. \tag{39}$$

From these relative expression levels, the software constructed an *in silico* RNA sample, with each transcript assigned a number of molecules according to its abundances. The software modeled the polyadenylation of each transcript by adding a poly-A tail (of mean length 125nt) after the terminal exon. FluxSimulator then simulated reverse transcription of *in silico* mRNAs by random hexamer priming, followed by size selection of RT products to between 175 and 225 nt. The resulting "library" of 6,601,805 cDNA fragments was then sampled uniformly at random for simulated sequencing, where the initial and terminal 75bp of each selected fragment were reported as reads. FluxSimulator does not allow precise control over the number of reads generated (Michael Sammeth, personal communication), but nevertheless generated 13,203,516 75nt paired-end RNA-Seq reads. These reads included sequencing errors; FluxSimulator includes a position-specific sequencing error model.

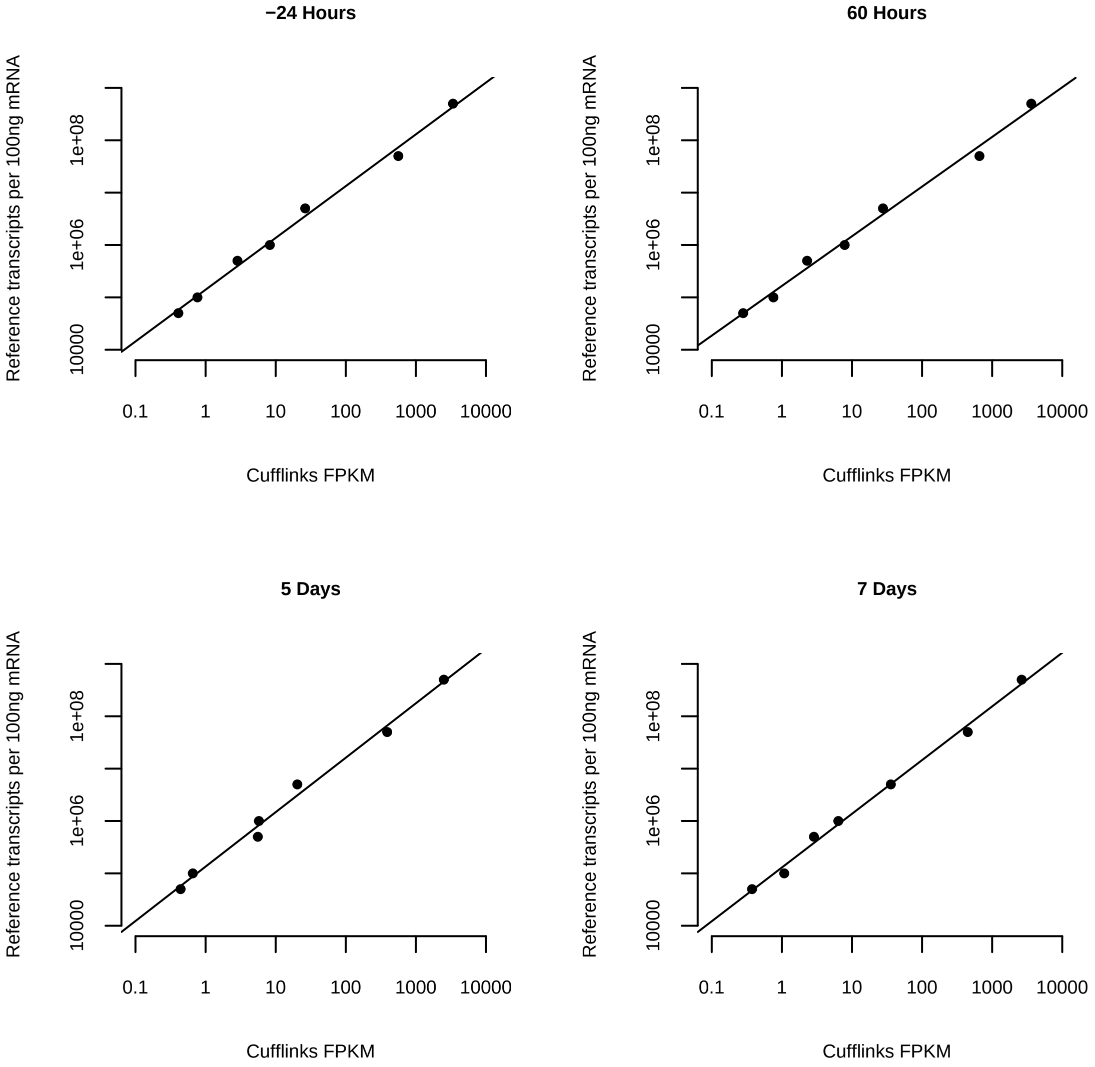


FIGURE 3. `Cufflinks`' abundance estimates of spiked-in sequences.

Fragments were mapped with `TopHat` to the mouse genome using identical parameters to those used to map the C2C12 reads, mapping a total of 6,176,961 (93% of the library). These alignments were supplied along with the exact set of expressed transcripts to `Cufflinks`, to measure `Cufflinks`' abundance estimation accuracy when working with a "perfect" assembly (Supplementary Figure 4). Estimated FPKM was very close to true *in silico* FPKM across a dynamic range of expression of nearly six orders of magnitude ($R^2 = 0.95$).

Estimation of transcript abundances by assigning fragments to them may be inaccurate if one is working with an incomplete set of transcripts for a particular sample. To

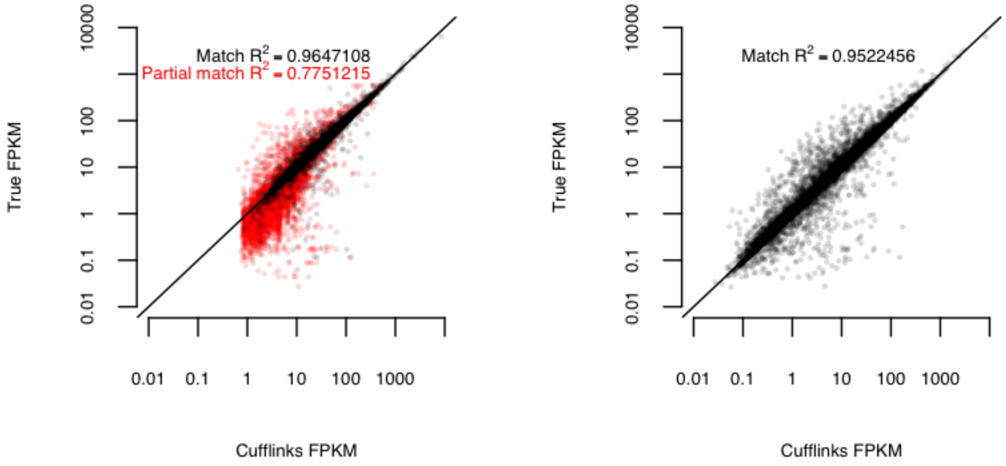


FIGURE 4. *In silico* assessment of the accuracy of `Cufflinks` abundance estimation when provided with a perfect assembly (a) and after *de novo* comparative assembly (b). Red points indicate in silico transcripts that were only partially recovered, where black points were fully reconstructed by `Cufflinks`. Simulated reads were aligned with `TopHat` and the alignments were provided to `Cufflinks` along with the structures of the transcripts in the simulated sample.

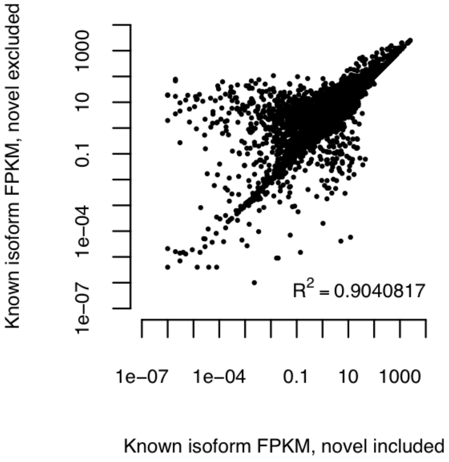


FIGURE 5. Excluding novel C2C12 transcripts from abundance estimation results in inaccurate estimates for known transcripts.

evaluate the impact of missing transcripts, we removed the newly discovered transcripts from our high-confidence set and re-estimated the abundances of known transcripts, and then compared them to those obtained when working with the complete high-confidence set. While estimates of known transcripts were overall similar or identical when working with both sets, reflecting single-isoform or fully annotated genes, isoforms of some alternatively spliced genes differed greatly. (Supplementary Figure 5)

As a final note, we point out that a naïve, yet popular, current approach to expression estimation is to sum the reads mapping to a gene (where the sum is taken across all exons appearing in all possible isoforms), and then to normalize the count by either the total number of exonic bases, or by the average length of the transcripts. We call the former method the "projective normalization" method, and the latter the "average length" method.

Proposition 3. *The totally projective normalization method is correct only for single isoform genes. If a gene has two or more isoforms the expression is underestimated.*

Proof: The effective length of the gene is overestimated, hence the expression level is underestimated. To see this, first note that the length of some transcript in a gene is less than the total number of exonic bases among all transcripts. Then, if $a_1, \ldots, a_n$ are real numbers all greater than zero and $b_1, \ldots, b_n$ are not all equal, we have

$$\frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i} < max_i(b_i), \tag{40}$$

so that the effective length in equation (1) is always less than the total number of exonic bases among all transcripts.

Stated differently, the projective normalization method has the problem that it produces numbers that are not proportional to the ρ, so that it is not additive. The average length method is flawed for the same reason. The transcript abundances are not taken into account in computing the effective lengths. In some cases the method might produce the correct answer (for the wrong reasons), but it is bound to be incorrect on most examples, especially in genes with transcripts of variable lengths and non-uniform abundances.

4. Transcript assembly

4.1. **Overview.** `Cufflinks` takes as input alignments of RNA-Seq fragments to a reference genome and, in the absence of an (optional) user provided annotation, initially assembles transcripts from the alignments. Transcripts in each of the loci are assembled independently. The assembly algorithm is designed to aim for the following:

(1) Every fragment is consistent with at least one assembled transcript.
(2) Every transcript is tiled by reads.
(3) The number of transcripts is the smallest required to satisfy requirement (1).
(4) The resulting RNA-Seq models (in the sense of Section 3) are identifiable.

In other words, we seek an assembly that parsimoniously explains the fragments from the RNA-Seq experiment; every fragment in the experiment (except those filtered out during a preliminary error-control step) should have come from a `Cufflinks` transcript, and `Cufflinks` should produce as few transcripts as possible with that property. Thus, `Cufflinks` seeks to optimize the criterion suggested in [27], however, unlike the method in that paper, `Cufflinks` leverages Dilworth's Theorem [4] to solve the problem by reducing it to a matching problem via the equivalence of Dilworth's and König's theorems (Theorem 19 in Appendix A). Our approach to isoform reconstruction is inspired by a similar approach used for haplotype reconstruction from HIV quasi-species [5].

4.2. **A partial order on fragment alignments.** The `Cufflinks` program loads a set of alignments in SAM format sorted by reference position and assembles non-overlapping sets of alignments independently. After filtering out any erroneous spliced alignments or reads from incompletely spliced RNAs, `Cufflinks` constructs a partial order (Definition 16), or equivalently a directed acyclic graph (DAG), with one node for each fragment that in turn consists of an aligned pair of mated reads. First, we note that fragment alignments are of two types: those where reads align in their entirety to the genome, and reads which have a split alignment (due to an implied intron).

In the case of single reads, the partial order can be simply constructed by checking the reads for *compatibility*. Two reads are *compatible* if their overlap contains the exact same implied introns (or none). If two reads are not compatible they are *incompatible*. The reads can be partially ordered by defining, for two reads x, y, that $x \leq y$ if the starting coordinate of x is at or before the starting coordinate of y, and if they are compatible.

In the case of paired-end RNA-Seq the situation is more complicated because the unknown sequence between mate pairs. To understand this, we first note that pairs of fragments can still be determined to be incompatible if they cannot have originated from the same transcript. As with single reads, this happens when there is disagreement on implied introns in the overlap. However compatibility is more subtle. We would like to define a pair of fragments x, y to be compatible if they do not overlap, or if every implied intron in one fragment overlaps an identical implied intron in the other fragment.

However it is important to note that it may be impossible to determine the compatibility (as defined above) or incompatibility of a pair of fragments. For example, an unknown region internal to a fragment may overlap two different introns (that are incompatible with each other). The fragment may be compatible with one of the introns (and the fragment from which it originates) in which case it is incompatible with the other. Since the opposite situation is also feasible, compatibility (or incompatibility) cannot be assigned. Fragments for which the compatibility/incompatibility cannot be determined with respect to every other fragment are called *uncertain*. Finally, two fragments are called *nested* if one is contained within the other.

Before constructing a partial order, fragments are extended to include their nested fragments and uncertain fragments are discarded. These discarded fragments are used in the abundance estimation. In theory, this may result in suboptimal (i.e. non-minimal assemblies) but we determined empirically that after assembly uncertain fragments are almost always consistent with one of the transcripts. When they are not, there was no completely tiled transcript that contained them. Thus, we employ a heuristic that substantially speeds up the program, and that works in practice.

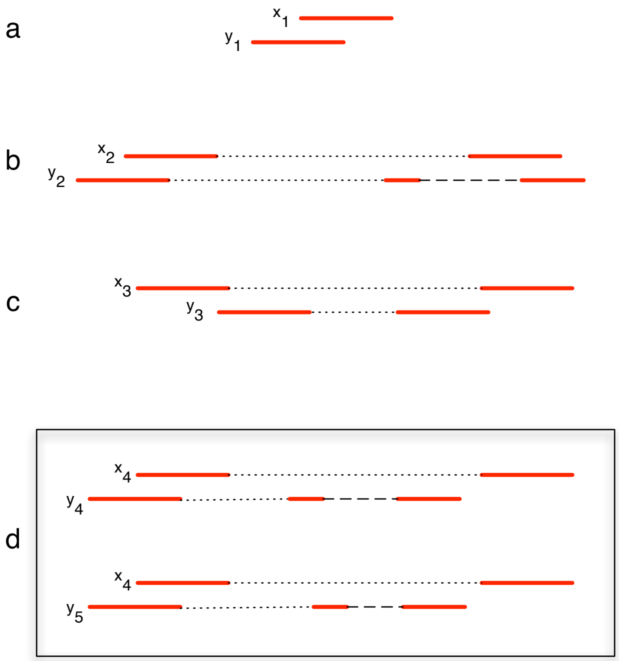


FIGURE 6. Compatibility and incompatibility of fragments. End-reads are solid lines, unknown sequences within fragments are shown by dotted lines and implied introns are dashed lines. The reads in (a) are compatible, whereas the fragments in (b) are incompatible. The fragments in (c) are nested. Fragment x_4 in (d) is uncertain, because y_4 and y_5 are incompatible with each other.

A partial order P is then constructed from the remaining fragments by declaring that $x \leq y$ whenever the fragment corresponding to x begins at, or before, the location of the fragment corresponding to y and x and y are compatible. In what follows we identify P with its Hasse diagram (or covering relation), equivalently a directed acyclic graph (DAG) that is the transitive reduction.

Proposition 4. *P is a partial order.*

Proof: The fragments can be totally ordered according to the locations where they begin. It therefore suffices to check that if x, y, z are fragments with x compatible with y and y compatible with z then x is compatible with z. Since x is not uncertain, it must be either compatible or incompatible with z. The latter case can occur only if x and/or z contain implied introns that overlap and are not identical. Since y is not nested within z and x is not nested within y, it must be that y contains an implied intron that is not identical with an implied intron in either x or z. Therefore y cannot be compatible with both x and z. □

4.3. **Assembling a parsimonious set of transcripts.** In order to assemble a set of transcripts, `Cufflinks` finds a (minimum) partition of P into chains (see Definition 16).

A partition of P into chains yields an assembly because every chain is a totally ordered set of compatible fragments $x_1, \ldots, x_l$ and therefore there is a set of overlapping fragments that connects them. By Dilworth's theorem (Theorem 17), the problem of finding a minimum partition P into chains is equivalent to finding a maximum antichain in P (an antichain is a set of mutually incompatible fragments). Subsequently, by Theorem 19, the problem of finding a maximum antichain in P can be reduced to finding a maximum matching in a certain bipartite graph that emerges naturally in deducing Dilworth's theorem from König's theorem 18. We call the key bipartite graph the "reachability" graph. It is the transitive closure of the DAG, i.e. it is the graph where each fragment x has nodes L_x and R_x in the left and right partitions of the reachability graph respectively, and where there is an edge between L_x and R_y when $x \leq y$ in P. The maximum matching problem is a classic problem that admits a polynomial time algorithm. The Hopcroft-Karp algorithm [10] has a run time of $O(\sqrt{V}E)$ where in our case V is the number of fragments and E depends on the extent of overlap, but is bounded by a constant times the coverage depth. We note that our parsimony approach to assembly therefore has a better complexity than the $O(V^3)$ PASA algorithm [8].

The minimum cardinality chain decomposition computed using the approach above may not be unique. For example, a locus may contain two putative distinct initial exons (defined by overlapping incompatible fragments), and one of two distinct terminal and a constitutive exon in between that is longer than any read or insert in the RNA-Seq experiment. In such a case, the parsimonious assembly will consist of two transcripts, but there are four possible solutions that are all minimal. In order to "phase" distant exons, we leverage the fact that abundance inhomogeneities can link distant exons via their coverage. We therefore weight the edges of the bipartite reachability graph based on the percent-spliced-in metric introduced by Wang *et al.* in [26]. In our setting, the percent-spliced-in ψ_x for an alignment x is computed by counting the alignments overlapping x in the genome that are compatible with x and dividing by the total number of alignments that overlap x, and normalizing for the length of the x. The cost $C(y, z)$ assigned to an edge between alignments y and z reflects the belief that they originate from different transcripts:

$$C(y, z) = -\log(1 - |\psi_y - \psi_z|). \tag{41}$$

Rather than using the Hopcroft-Karp algorithm, a modified version of the `LEMON` [7] and `Boost` [15] graph libraries are used to compute a *min-cost* maximum cardinality matching on the bipartite compatibility graph. Even with the presence of weighted edges, our algorithm is very fast. The best known algorithm for weighted matching is $O(V^2 logV + VE)$.

Because we isolated total RNA, we expected that a small fraction of our reads would come from the intronic regions of incompletely processed primary transcripts. Moreover, transcribed repetitive elements and low-complexity sequence result in "shadow" transfrags that we wished to discard as artifacts. Thus, `Cufflinks` heuristically identifies artifact transfrags and suppresses them in its output. We also filter extremely low-abundance minor isoforms of alternatively spliced genes, using the model described in Section 3 as a means of reducing the variance of estimates for more abundant transcripts. A transcript x meeting any of the following criteria is suppressed:

(1) x aligns to the genome entirely within an intronic region of the alignment for a transcript y, and the abundance of x is less than 15% of y's abundance.
(2) x is supported by only a single fragment alignment to the genome.
(3) More than 75% of the fragment alignments supporting x, are mappable to multiple genomic loci.
(4) x is an isoform of an alternatively spliced gene, and has an estimated abundance less than 5% of the major isoform of the gene.

Prior to transcript assembly, `Cufflinks` also filters out some of the alignments for fragments that are likely to originate from incompletely spliced nuclear RNA, as these can reduce the accuracy abundance estimates for fully spliced mRNAs. These filters and the output filters above are detailed in the source file `filters.cpp` of the source code for `Cufflinks`.

In the overview of this Section, we mentioned that our assembly algorithm has the property that the resulting models are identifiable. This is a convenient property that emerges naturally from the parsimony criterion for a "minimal explanation" of the fragment alignments. Formally, it is a corollary of Dilworth's theorem:

Proposition 5. *The assembly produced by the* `Cufflinks` *algorithm always results in an identifiable RNA-Seq model.*

Proof: By Dilworth's theorem, the minimum chain decomposition (assembly) we obtain has the same size as the maximum antichain in the partially ordered set we construct from the reads. An antichain consists of reads that are pairwise incompatible, and therefore those reads must form a permutation sub-matrix in the fragment-transcript matrix $A_{R,T}$ with columns corresponding to the transcripts in a locus, and with rows corresponding to the fragments in the antichain. The matrix $A_{R,T}$ therefore contains permutation sub-matrices that together span all the columns, and the matrix is full-rank.

4.4. **Assessment of assembly quality.** To compare `Cufflinks` transfrags against annotated transcriptomes, and also to find transfrags common to multiple assemblies, we developed a tool called `Cuffcompare` that builds structural equivalence classes of transcripts. We ran `Cuffcompare` on each the assembly from each time point against the combined annotated transcriptomes of the UCSC known genes, `Ensembl`, and `Vega`. Because of the stochastic nature of sequencing, *ab initio* assembly of the same transcript in two different samples may result in transfrags of slightly different lengths. A `Cufflinks` transfrag was considered a complete match when there was a transcript with an identical chain of introns in the combined annotation.

When no complete match is found between a `Cufflinks` transfrag and the transcripts in the combined annotation, `Cuffcompare` determines and reports if another substantial relationship exists with any of the annotation transcripts that can be found in or around the same genomic locus. For example, when all the introns of a transfrag match perfectly a part of the intron chain (sub-chain) of an annotation transcript, a "containment" relationship is reported. For single-exon transfrags, containment is also reported when the exon appears fully overlapped by any of the exons of an annotation transcript. If there is no perfect match for the intron chain of a transfrag but only some exons overlap and there is at least one intron-exon junction match, `Cuffcompare` classifies the transfrag as a putative "novel" isoform of an annotated gene. When a transfrag is unspliced (single-exon) and it overlaps the intronic genomic space of a reference annotation transcript, the transfrag is classified as potential pre-mRNA fragment. Finally, when no other relationship is found between a `Cufflinks` transfrag and an annotation transcript, `Cuffcompare` can check the repeat content of the transfrag's genomic region (assuming the soft-masked genomic sequence was also provided) and it would classify the transfrag as "repeat" if most of its bases are found to be repeat-masked.

When provided multiple time point assemblies, `Cuffcompare` matches transcripts between samples that have an identical intron structure, placing all mutually matching transcripts in the same equivalence class. The program reports a non-redundant set of transcript structures, choosing the longest transcript from each equivalence class as the representative transcript. `Cuffcompare` also reports the relationships found between each equivalence class (transcripts that have a complete match across time points) and reference transcripts from the combined annotation set, where applicable.

Table 2 includes the classifications of the transfrags reported by `Cufflinks` after assembling the C2C12 reads. While only 13.5% of assembled transfrags represent known transcripts, `Cufflinks` assigns more than 76% of reads to these, reflecting the fact that moderate and highly-abundant transfrags generate most of the library fragments in the experiment. Less abundant transcripts receive less complete sequencing coverage, resulting in numerous transfrags that partially but compatibly match known transcripts. Supplementary Figure 7 shows the categories of `Cufflinks` transfrags as estimated depth of sequencing coverage increasing.

Category	**Transfrags**	**% of total transfrags**	**Assembled reads** (%)
Match to known isoform	39,857	13.5	76.7
Novel isoform of known gene	18,565	6.3	11.3
Contained in known isoform	71,029	24.1	4.6
Repeat	41,906	14.2	0.6
Intronic	32,658	11.1	0.6
Polymerase run-on	18,522	6.3	0.5
Intergenic	48,604	16.5	1.2
Other artifacts	22,483	7.7	4.5
Total transfrags	293,624	100.0	100.0

Table 2. Classification of all transfrags produced at any time point with respect to annotated gene models and masked repeats in the mouse genome. Transfrags that are present in multiple time point assemblies are multiply counted to preserve the relative distribution of transfrags among the categories across the full experiment.

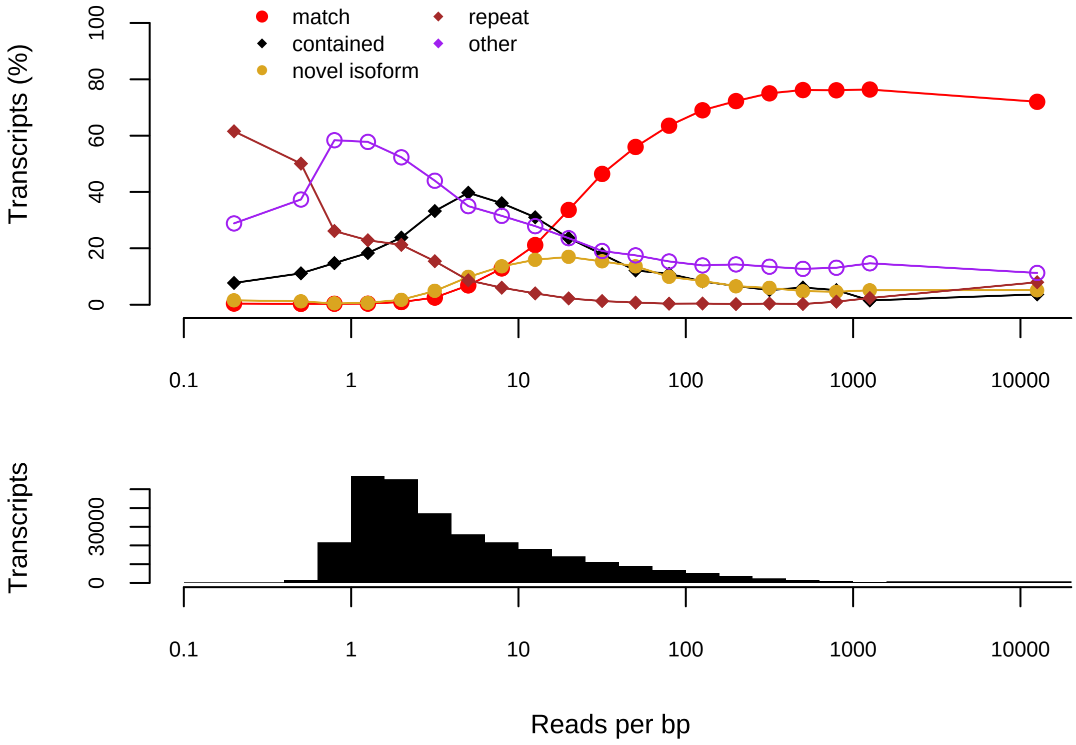


FIGURE 7. Categorization of `Cufflinks` transcripts by estimated depth of read coverage.

We selected the `Cufflinks` transfrags that did not have a complete match or "containment" relationship with a known annotation transcript, but were classified by `Cuffcompare` as putative "novel isoforms" of known genes. We explored the sequence similarity between these transfrags and two sets of mRNA sequences: one set representing the mouse transcriptome and consisting of all mouse ESTs in dbEST plus all reviewed or validated RefSeq mouse mRNAs, and the other consisting of all reviewed or validated RefSeq mRNAs from other mammalian species.

We used megablast to map all mouse ESTs onto this set of `Cufflinks` transfrags, only keeping EST alignments where at least 80% of the EST length was aligned with at least 95% identity. We calculated transfrag coverage by tiling overlapping EST mappings on each transfrag and counted only those transfrags that are covered by ESTs for at least 80% of the transfrag length without any coverage gaps, and with coverage discontinuities only allowed at no more than 10% distance from either end. For the mouse mRNAs alignments we also used megablast with the same basic coverage cutoffs (minimum 80% covered with no more than 10% unaligned on either side of the overlap) but applied to each pairwise alignment independently (i.e. as opposed to EST alignments, no coverage tiling was considered for mRNA alignments). For alignments with the non-mouse mRNAs we used discontiguous megablast with a dual (combined) discontiguous word template (option -N 2), with the same coverage assessment protocol as in the case of mouse mRNA alignments but with the percent identity cutoff lowered to 80%.

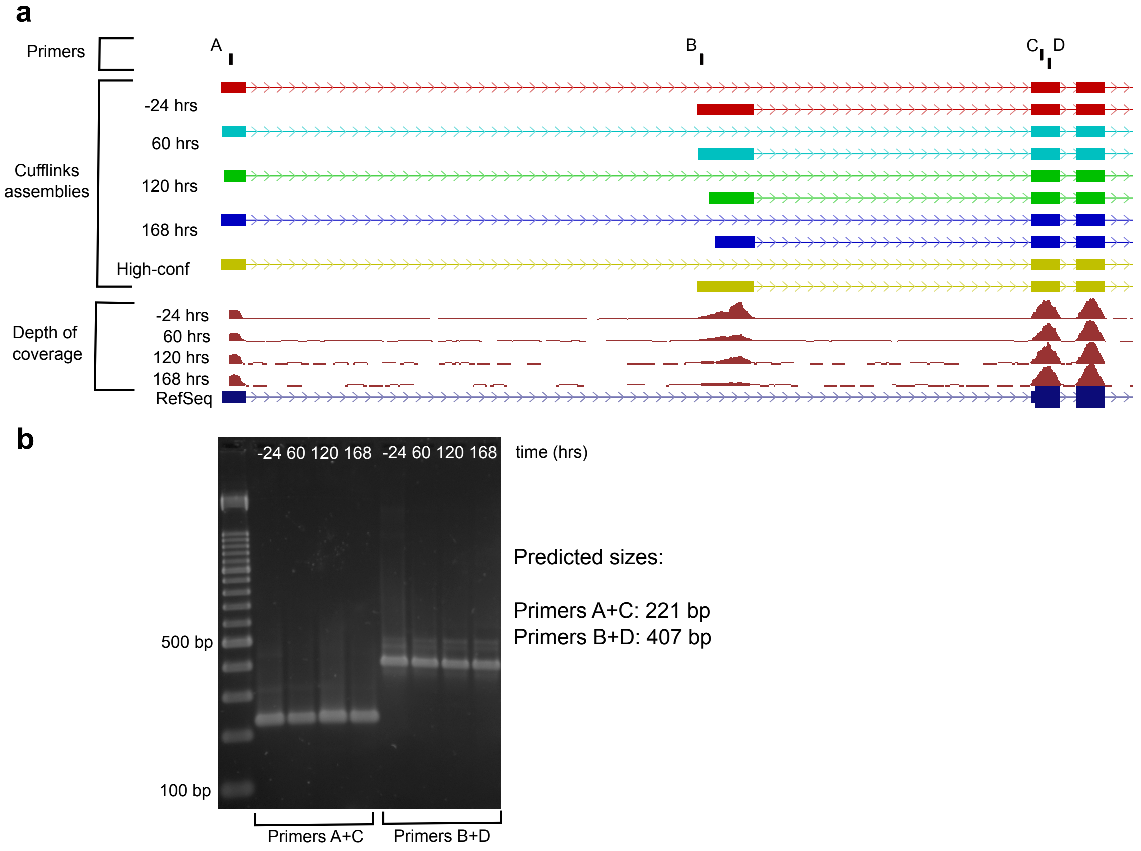


Figure 8. New and known isoforms of Fhl3 recovered by `Cufflinks` at each time point (a) were confirmed by form-specific RT-PCR (b).

To assess the dependence of assembly quality on the depth of sequencing, we mapped and assembled subsets of our reads at the 60 hour time point. We partitioned the three Illumina lanes' worth of data (a total of 140 million reads) into 64 subsets. We then processed a single subset with `TopHat` and `Cufflinks`, as above, and compared the resulting transfrags to the output of `Cufflinks` on all three lanes using `Cuffcompare`. We repeated the mapping and assembly with two subsets, four subsets, eight, and so on. Figure 4 in the main text shows the fraction of reference transcripts captured by `Cufflinks` using all three lanes that are still captured when less data is available. For transcripts with low abundance (<15 FPKM), increased sequencing yields more full-length transcripts. However, for even moderately abundant transcripts (≥15 FPKM),

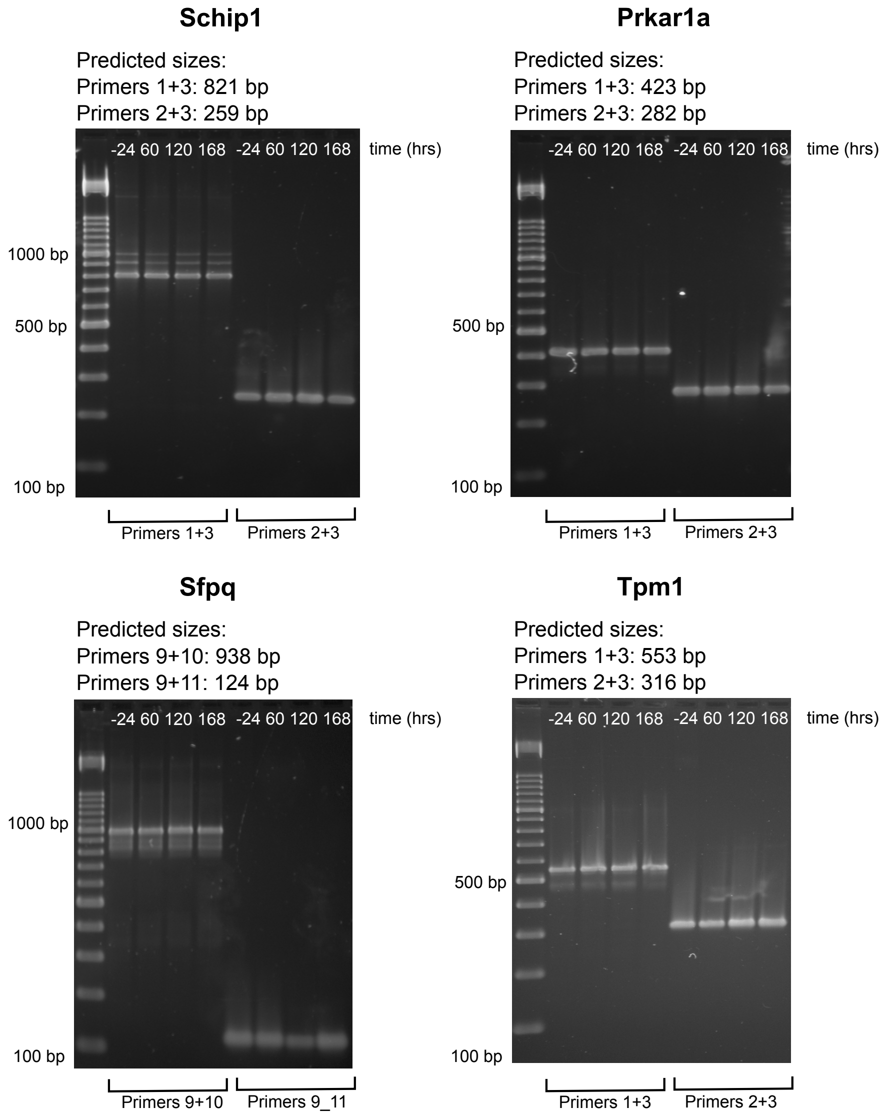


FIGURE 9. RT-PCR of selected genes. For Schip1, `Cufflinks` assembled a known and a novel isoform (with a new TSS), both of which are detected by RT-PCR. Prkar1a is annotated with two alternate first exons and start sites in `UCSC known genes`, both of which were detected. `Cufflinks` assembles the known isoform of the splicing factor Sfpq, along with a novel variant that contains most of RIKEN clone. Tpm1, a gene known to have muscle- and non-muscle-specific isoforms displays previously observed alternative first and last exons.

75% or more of the transcripts are recovered with only 40 million reads, or a lane's worth of Illumina GA II sequencing.

Primer name	sequence	product length	endpoint gel score
FHL3 Ex1Ex3			
Left	CTCGCCGCTGCTCTCTCG	221	+++
Right	GTGTTGTCATAGCACGGAACG		
FHL3 Ex2Ex3			
Left	AGGAAGGGCTCACAAGTGG	407	+++
Right	ATAGCACGGAACGCAGTAGG		
Sfpq Ex9Ex10			
Left	GTGGTGGCATAGGTTATGAAGC	936	+++
Right	CCATTTTCAAAAGCTTTCAAGG		
Sfpq Ex9Ex11			
Left	GTGGTGGCATAGGTTATGAAGC	172	+++
Right	CTCAAGTAAATAAGACTCCAAAATCAGC		
Prkar1aEx1Ex3			
Left	ACAGCAGGGATCTCCTTGTCC	418	+++
Right	CCTCTCAAAGTATTCCCGAAGG		
Prkar1aEx2Ex3			
Left	GCTATCGCAGAGTGGTAGTGAGG	279	+++
Right	CCTCTCAAAGTATTCCCGAAGG		
Schip1Ex1Ex3			
Left	GGCTATGAGGGTGAAAAGTGC	1050	+++
Right	GTATAGATTCCTGGGCCATCG		
Schip1Ex2Ex3			
Left	CAGCATGAGTGGTAACCAAGG	269	+++
Right	GTATAGATTCCTGGGCCATCG		
Tpm1Ex1Ex3			
Left	TGAACAAAAGACCCCAGAGG	565	+++
Right	CTGAAGTACAAGGCCATCAGC		
Tpm1Ex2Ex3			
Left	AGTTTTATTGAGCGTTGAGACG	318	+++
Right	CTGAAGTACAAGGCCATCAGC		

Table 3. Form-specific RT-PCR primers for selected genes, designed with Primer3 [22].

5. Analysis of gene expression dynamics

Expression dynamics of genes are composed of absolute changes in overall transcript abundances, as well as relative changes in transcript abundances over time. Moreover, select groups of transcripts, for example transcripts grouped by TSS, may exhibit specific dynamics due to the underlying biological mechanisms that drive expression.

In this section we describe statistical tests we developed in the multiple hypothesis testing framework for examining absolute and relative changes in arbitrary groups of transcripts.

5.1. **Selection of high-confidence transcripts for expression tracking.** We first restricted our analysis of expression dynamics over the time-course to a set of transcripts we believed were fully sequenced and correctly assembled, and we focused only on known and reliable novel isoforms of annotated genes. This set consisted of transcripts that either were present in the `UCSC genome browser`, `Ensembl`, or `Vega` annotated transcriptomes, or were found in multiple C2C12 time point assemblies. We ignored transfrags classified as intronic pre-mRNA or polymerase run-on, as well as intergenic repeats to focus on coding genes and long non-coding RNAs. This high-confidence set contained a total of 17,416 transcripts, 13,692 of which were in `UCSC known genes`, `Ensembl` or `VEGA` annotation and 3,724 of which are novel. Running `Cufflinks`' abundance estimation algorithm on this high-confidence set of transcripts at each time point allowed us to scan for differentially expressed transcripts, differentially spliced pre-mRNAs, and genes with shifts in promoter preference.

5.2. **Testing for changes in absolute expression.** Between any two consecutive time points, we tested whether a transcript was significantly (after FDR control [2]) up or down regulated with respect to the null hypothesis of no change, with variability in expression due solely to the uncertainties resulting from our abundance estimation procedure. This was done using the following testing procedure for absolute differential expression:

We employed the standard method used in microarray-based expression analysis and proposed for RNA-Seq in [3], which is to compute the logarithm of the ratio of intensities (in our case FPKM), and to then use the delta method to estimate the variance of the log odds. We describe this for testing differential expression of individual transcripts and also groups of transcripts (e.g. grouped by TSS).

We recall that the MLE FPKM for a transcript t in a locus g is given by

$$\frac{10^9 X_g \hat{\gamma}_t}{\tilde{l}(t)M}. \tag{42}$$

Given two different experiments resulting in X_g^a, M^a and X_g^b, M^b respectively, as well as $\hat{\gamma}_t^a$ and $\hat{\gamma}_t^b$, we would like to test the significance of departures from unity of the ratio of MLE FPKMS, i.e.

$$\left(\frac{10^9 X_g^a \hat{\gamma}_t^a}{\tilde{l}(t)M^a}\right) / \left(\frac{10^9 X_g^b \hat{\gamma}_t^b}{\tilde{l}(t)M^b}\right) \tag{43}$$

$$= \frac{X_g^a \hat{\gamma}_t^a M^b}{X_g^b \hat{\gamma}_t^b M^a}. \tag{44}$$

This can be turned into a test statistic that is approximately normal by taking the logarithm, and normalizing by the variance. We recall that using the delta method, if X is a random variable then $Var[log(X)] \approx \frac{Var[X]}{E[X]^2}$.

Therefore, our test statistic is

$$\frac{log(X_g^a) + log(\hat{\gamma}_t^a) + log(M^b) - log(X_g^b) - log(\hat{\gamma}_t^b) - log(M^a)}{\sqrt{\frac{\left(\Psi_{t,t}^{g,a}(1+X_g^a)+(\hat{\gamma}_t^a)^2\right)}{X_g^a(\hat{\gamma}_t^a)^2} + \frac{\left(\Psi_{t,t}^{g,b}(1+X_g^b)+(\hat{\gamma}_t^b)^2\right)}{X_g^b\left(\hat{\gamma}_t^b\right)^2}}}. \tag{45}$$

In order to test for differential expression of a group of transcripts, we replace the numerator and denominator above by those from Equations (36) and (38).

It is has been noted that the power of differential expression tests in RNA-Seq depend on the length of the transcripts being tested, because longer transcripts accumulate more reads [18]. This means that the results we report are biased towards discovering longer differentially expressed transcripts and genes.

5.3. **Quantifying transcriptional and post-transcriptional overloading.** In order to infer the extent of differential promoter usage, we decided to measure changes in relative abundances of primary transcripts of single genes. Similarly, we investigated changes in relative abundances of transcripts grouped by TSS in order to infer differential splicing. These inferences required two ingredients:

(1) A metric on probability distributions (derived from relative abundances).
(2) A test statistic for assessing significant changes in differential promoter usage and splicing as measured using the metric referred to above.

In order to address the first requirement, namely a metric on probability distributions, we turned to an entropy-based metric. This was motivated by the methods in [21] where tests for differences in relative isoform abundances were performed to distinguish cancer cells from normal cells. We extend this approach to be able to test for relative isoform abundance changes among multiple experiments in RNA-Seq.

Definition 6 (Entropy)**.** The entropy of a discrete probability distribution $p = (p_1, \ldots, p_n)$ $(0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1)$ is

$$H(p) = -\sum_{i=1}^{n} p_i log p_i. \tag{46}$$

If $p_i = 0$ for some i the value of $p_i log p_i$ is taken to be 0.

Definition 7 (The Jensen-Shannon divergence)**.** The Jensen-Shannon divergence of m discrete probability distributions $p^1, \ldots, p^m$ is defined to be:

$$JS(p^1, \ldots, p^m) = H\left(\frac{p^1 + \cdots + p^m}{m}\right) - \frac{\sum_{j=1}^{m} H(p^j)}{m}. \tag{47}$$

In other words, the Jensen-Shannon divergence of a set of probability distributions is the entropy of their average minus the average of their entropies.

In the case where $m = 2$, we remark that the Jensen-Shannon divergence can also be described in terms of the Kullback-Leibler divergence of two discrete probability distributions. If we denote Kullback-Leibler divergence by

$$D(p^1 \| p^2) = \sum_i p_i^1 log \frac{p_i^1}{p_i^2}, \tag{48}$$

then

$$JS(p^1, p^2) = \frac{1}{2} D(p^1 \| m) + \frac{1}{2} D(p^2 \| m) \tag{49}$$

where $m = \frac{1}{2}(p^1 + p^2)$. In other words the Jensen-Shannon divergence is a symmetrized variant of the Kullback-Leibler divergence.

The Jensen-Shannon divergence has a number of useful properties: for example it is symmetric and non-negative. However it is *not* a metric. The following theorem shows how to construct a metric from the Jensen-Shannon divergence:

Theorem 8 (Fuglede and Topsøe 2004 [6])**.** *The square root of the Jensen-Shannon divergence is a metric.*

The proof of this result is based on a harmonic analysis argument that is the basis for the remark in the main paper that "transcript abundances move in time along a logarithmic spiral in Hilbert space". We therefore call the square root of the Jensen-Shannon divergence the *Jensen-Shannon metric.* We employed this metric in order to quantify relative changes in expression in (groups of) transcripts.

In order to test for significance, we introduce a bit of notation. Suppose that S is a collection of transcripts (for example, they may share a common TSS). We define

$$\kappa_t = \frac{\frac{\gamma_t}{\tilde{l}(t)}}{\sum_{u \in S} \frac{\gamma_u}{\tilde{l}(u)}} \tag{50}$$

to be the proportion of transcript t among all the transcripts in a group S. We let $Z = \sum_{u \in S} \hat{\gamma}_u / \tilde{l}(u)$ so that $\hat{\kappa}_t = \frac{\gamma_t}{\tilde{l}(t)Z}$. We therefore have that

$$Var[\hat{\kappa}_t] = \frac{Var[\hat{\gamma}_t]}{\tilde{l}(t)^2 Z^2}, \tag{51}$$

$$Cov[\hat{\kappa}_t, \hat{\kappa}_u] = \frac{Cov[\hat{\gamma}_t, \hat{\gamma}_u]}{\tilde{l}(t)\tilde{l}(u)Z^2}. \tag{52}$$

Our test statistic for divergent relative expression was the Jensen-Shannon metric. The test could be applied to multiple time points simultaneously, but we focused on pairwise tests (involving consecutive time points). Under the null hypothesis of no change in relative expression, the Jensen-Shannon metric should be zero. We tested for this using a one-sided t-test, based on an asymptotic derivation of the distribution of the Jensen-Shannon metric under the null hypothesis. This asymptotic distribution is normal by applying the delta method approximation, which involves computing the linear component of the Taylor expansion of the variance of $\sqrt{JS}$.

In order to simplify notation, we let $f(p^1, \ldots, p^m)$ be the Jensen-Shannon metric for m probability distributions $p^1, \ldots, p^m$.

Lemma 9. *The partial derivatives of the Jensen-Shannon metric are give by*

$$\frac{\partial f}{\partial p_l^k} = \frac{1}{2m\sqrt{f(p^1, \ldots, p^m)}} log\left(\frac{p_l^k}{\frac{1}{m}\sum_{j=1}^m p_l^j}\right). \tag{53}$$

Let $\hat{\kappa}^1, \ldots, \hat{\kappa}^m$ denote m probability distributions on the set of transcripts S, for example the MLE for the transcript abundances in a time course. Then from the delta method we have that $\sqrt{JS(\hat{\kappa}^1, \ldots, \hat{\kappa}^m)}$ is approximately normally distributed with variance given by

$$Var[\sqrt{JS(\hat{\kappa}^1, \ldots, \hat{\kappa}^m)}] \approx (\bigtriangledown f)^T \Sigma (\bigtriangledown f), \tag{54}$$

where Σ is the variance-covariance matrix for the $\kappa^1, \ldots, \kappa^m$, i.e., it is a block diagonal matrix where the ith block is the variance-covariance matrix for the κ^i_t given by Equations (51,52).

There are two biologically meaningful groupings of transcripts whose relative abundances are interesting to track in a time course. Transcripts that share a TSS are likely to be regulated by the same promoter, and therefore tracking the change in relative abundances of groups of transcripts sharing a TSS may reveal how transcriptional regulation is affecting expression over time. Similarly, transcripts that share a TSS and exhibit changes in expression relative to each other are likely to be affected by splicing or other post-transcriptional regulation. We therefore grouped transcripts by TSS and compared relative abundance changes within and between groups.

We define "overloading" to be a significant change in relative abundances for a set of transcripts (as determined by the Jensen-Shannon metric, see below). The term is intended to generalize the simple notion of "isoform switching" that is well-defined in the case of two transcripts, to multiple transcripts. It is complementary to absolute differential changes in expression: the overall expression of a gene may remain constant while individual transcripts change drastically in relative abundances resulting in overloading. The term is borrowed from computer science, where in some statically-typed programming languages, a function may be used in multiple, specialized instances via "method overloading".

We tested for overloaded genes by performing a one-sided t-test based on the asymptotics of the Jensen-Shannon metric under the null hypothesis of no change in relative abundnaces of isoforms (either grouped by shared TSS for for post-transcriptional overloading, or by comparison of groups of isoforms with shared TSS for transcriptional overloading). Type I errors were controlled with the Benjamini-Hochberg [2] correction for multiple testing. A selection of overloaded genes are displayed in Supplemental Figs. 10 and 11.

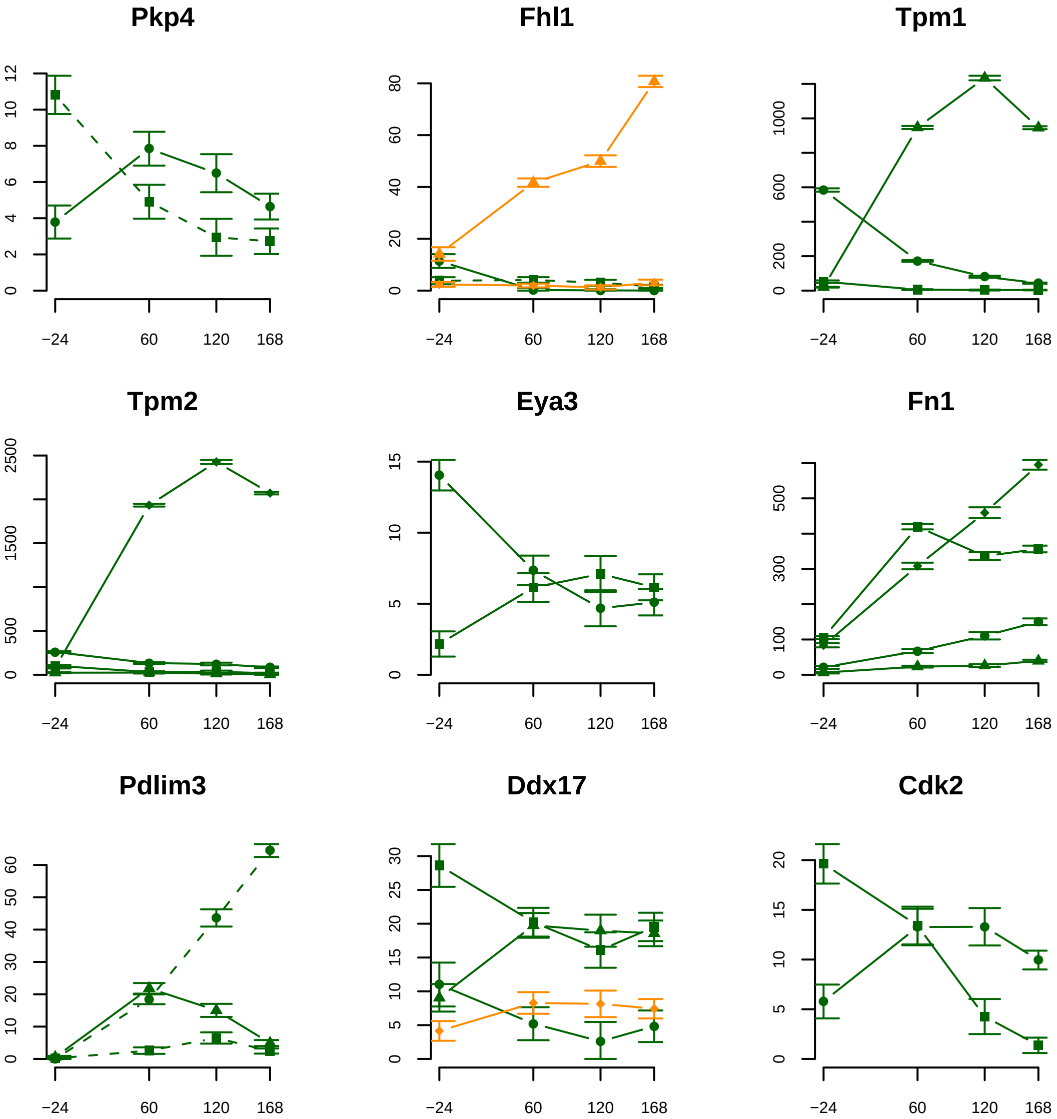


Figure 10. Selected genes with post-transcriptional overloading. Trajectories indicate the expression of individual isoforms in FPKM (y axis) over time in hours (x axis). Dashed isoforms have not been previously annotated. Isoform trajectories are colored by TSS, so isoforms with the same color presumably share a common promoter and are processed from the same primary transcript. It is evident that total gene expression may remain constant during isoforms switching (Eya3) while in other cases changes in relative abundance are accompanied by changes in absolute expression. The Jensen-Shannon metric generalizes the notion of "isoform switching" and is useful in cases with multiple isoforms (e.g. Ddx17).

Nature Biotechnology: doi:10.1038/nbt.1621

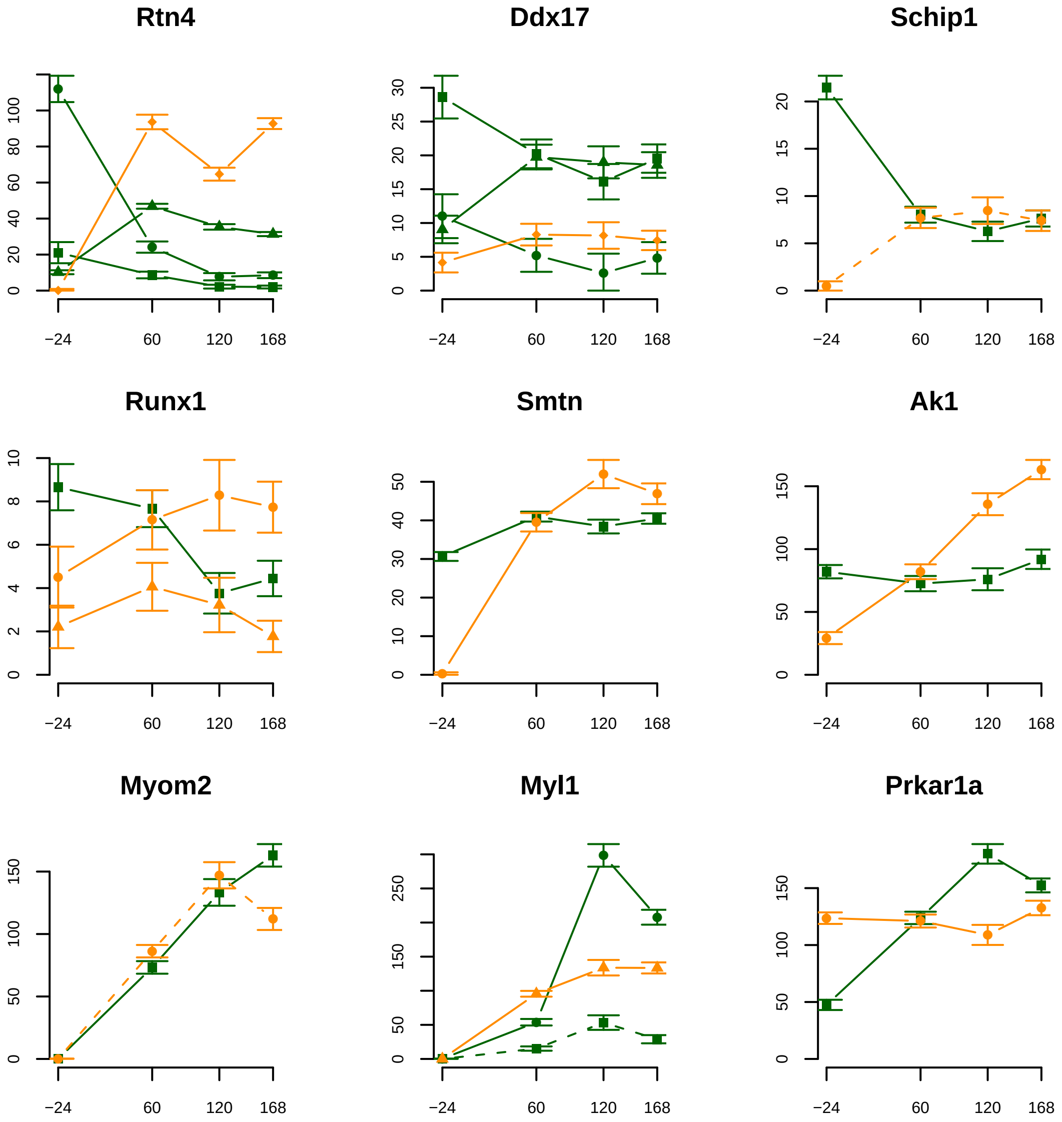


Figure 11. Selected genes with transcriptional overloading. Trajectories indicate the expression of individual isoforms in FPKM (y axis) over time in hours (x axis). Dashed isoforms have not been previously annotated. Isoform trajectories are colored by TSS, so that isoforms with different colors presumably vary in their promoter and are processed from different primary transcripts.

We can visualize overloading and expression dynamics with a plot that superimposes transcriptional and post-transcriptional overloading and gene-level expression over the time course. We refer to these as “Minard plots”, after Charles Joseph Minard’s famous depiction of the advance and retreat of Napoleon’s armies in the campaign against

Russia in 1812 [25]. Minard made use of multiple visual cues to display numerous varying quantities in one diagram. An example of a Minard plot for the gene Myc is shown in Figure 3c, and others are given in Appendix B. The dotted line indicates gene-level FPKM, with measured FPKM indicated by black circles. Grey circles indicate the arithmetic mean of gene-level FPKM between consecutive measured time points, interpolating FPKM at intermediate time points. The total gene expression overloading is visualized as a swatch centered around the interpolated expression curve. The width of the swatch encodes the amount of expression overloading between successive time points. The color of the swatch indicates the relative contributions of transcriptional and post-transcriptional expression overloading.

Some genes, such as tropomyosin I and II, feature a single primary transcript, and so all overloading is by definition post-transcriptional. Others, like Fhl3, have two primary transcripts, but only a single isoform arises from each, so all overloading is transcriptional. Genes with multiple primary transcripts, one or more of which are alternatively spliced, such as Myc or RTN4, display both forms.

6. The `Cufflinks` software

The transcript assembly and abundance estimation algorithms are implemented in freely available open source software called `Cufflinks` that is available from
`http://cufflinks.cbcb.umd.edu/`
Furthermore methods for comparing annotations across time points, and for performing the differential expression, promoter usage and splicing tests are implemented in the companion programs `Cuffdiff` and `Cuffcompare`. Instructions on how to install and run the software are provided on the website.

The input to `Cufflinks` consists of fragment alignments in the SAM format [14]. These may consist of either single fragment alignments, or alignments of mate-pairs (paired-end reads produce better assemblies and more accurate abundance estimates than single reads). `Cufflinks` will assemble the transcripts using the algorithm in Section 4, and transcript abundances will be estimated using the model in Section 3. Transcript coordinates and abundances are reported in the Gene Transfer Format (GTF). User supplied annotations may be provided to `Cufflinks` (optional input) in which case they form the basis for the transcript abundance estimation.

Some of the algorithms here rely on sufficient depth of sequencing in order to produce reliable output. `Cufflinks` determines that depth is sufficient where possible to check that required assumptions hold. For example, in loci where one or more isoforms have extremely low relative expression, the observed Fisher Information Matrix may not be positive definite after rounding errors. In this case, it is not possible to produce a reliable variance-covariance matrix for isoform fragment abundances. `Cufflinks` will report a numerical exception in this and similar cases. When an exception is reported, the confidence intervals for the isoforms' abundances will be set from 0 FPKM to the FPKM for the whole gene. If such an exception is generated during a `Cuffdiff` run, no differential analysis involving the problematic sample will be performed on that locus.

7. Appendix A: Lemmas and Theorems

The following elementary/classical results are required for our methods and we include them so that the supplement is self-contained.

Lemma 10. *Let $X_1, \ldots, X_n$ be random variables and $a_1, \ldots, a_n$ real numbers with $Y = \sum_{i=1}^n a_i X_i$. Then*

$$Var[Y] = \sum_{i=1}^{n} a_i^2 Var[X_i] + 2 \sum_{i<j} a_i a_j Cov[X_i, X_j]. \tag{55}$$

Lemma 11 (Taylor Series)**.** *If X and Y are random variables then*

$$\begin{aligned} Var[f(X,Y)] \;\approx\; & \left(\frac{\partial f}{\partial X}(E[X], E[Y])\right)^2 Var[X] \\ & + 2\frac{\partial f}{\partial X}(E[X], E[Y])\frac{\partial f}{\partial Y}(E[X], E[Y]) Cov[X, Y] \\ & + \left(\frac{\partial f}{\partial Y}(E[X], E[Y])\right)^2 Var[Y]. \end{aligned} \tag{56}$$

Corollary 12. *If X and Y are independent then*

$$Var\left[log\left(\frac{X}{Y}\right)\right] \;\approx\; \frac{V[X]}{E[X]^2} + \frac{V[Y]}{E[Y]^2}. \tag{57}$$

Corollary 13. *If X and Y are independent random variables then*

$$Var[XY] = Var[X]Var[Y] + E[X]^2 Var[Y] + E[Y]^2 Var[X]. \tag{58}$$

The above result is exact using the 2nd order Taylor expansion (higher derivatives vanish).

Lemma 14 ([13])**.** *Let $a_1, \ldots, a_n, w_1, \ldots, w_n$ be real numbers satisfying: $w_i \neq 0$ and $0 \leq a_i \leq 1$ for all i, $\sum_{i=1}^n a_i = 1$ and $\sum_{i=1}^n a_i w_i \neq 0$. Let $b_j = \frac{a_j w_j}{\sum_{i=1}^n a_i w_i}$. Then $a_j = \frac{b_j \frac{1}{w_j}}{\sum_{i=1}^n b_i \frac{1}{w_i}}$.*

Proof:

$$b_j = \frac{a_j w_j}{\sum_{i=1}^n a_i w_i} \tag{59}$$

$$\Rightarrow \sum_{k=1}^n \frac{b_k}{w_k} = \sum_{k=1}^n \frac{a_k}{\sum_{i=1}^n a_i w_i} \tag{60}$$

$$= \frac{1}{\sum_{i=1}^n a_i w_i} \tag{61}$$

$$= \frac{b_j}{a_j w_j} \tag{62}$$

$$\Rightarrow a_j = \frac{b_j \frac{1}{w_j}}{\sum_{i=1}^n b_i \frac{1}{w_i}}. \tag{63}$$

□

Proposition 15 ([19])**.** *Let $f_i(\theta) = \sum_{j=1}^d a_{ij}\theta_j + b_i$ $(1 \leq i \leq m)$ describe a linear statistical model with $a_{ij} \geq$ for all i, j. That is, $\sum_{i=1}^m f_i(\theta) = 1$. If $u_i \geq 0$ for all i then the log likelihood function*

$$l(\theta) = \sum_{i=1}^m u_i log(f_i(\theta)) \tag{64}$$

is concave.

Proof: It is easy to see that

$$\left(\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}\right) = -A^T diag\left(\frac{u_1}{f_1(\theta)^2}, \dots, \frac{u_m}{f_m(\theta)^2}\right) A, \tag{65}$$

where A is the $m \times d$ matrix whose entry in row i and column j equals a_{ij}. Therefore the Hessian is a symmetric matrix with non-positive eigenvalues, and is therefore negative semi-definite. □

Definition 16. A partially ordered set is a set S with a binary relation $\leq$ satisfying:

(1) $x \leq x$ for all $x \in S$,
(2) If $x \leq y$ and $y \leq z$ then $x \leq z$,
(3) If $x \leq y$ and $y \leq x$ then $x = y$.

A *chain* is a set of elements in $C \subseteq S$ such that for every $x, y \in C$ either $x \leq y$ or $y \leq x$. An *antichain* is a set of elements that are pairwise incompatible.

Partially ordered sets are equivalent to directed acyclic graphs (DAGs). The following min-max theorems relate chain partitions to antichains and are special cases of linear-programming duality. More details and complete proofs can be found in [16].

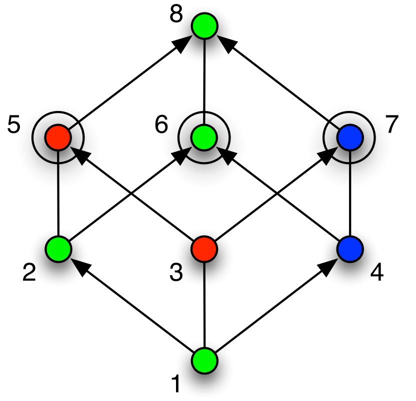
Theorem 17 (Dilworth's theorem)**.** *Let P be a finite partially ordered set. The maximum number of elements in any antichain of P equals the minimum number of chains in any partition of P into chains.*

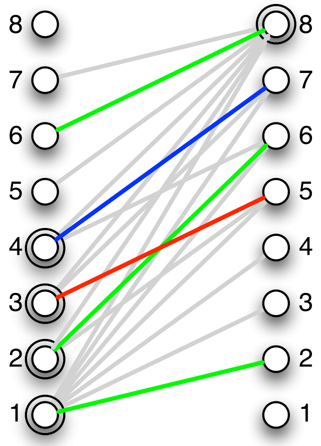
Theorem 18 (König's theorem). *In a bipartite graph, the number of edges in a maximum matching equals the number of vertices in a minimum vertex cover.*

Theorem 19. *Dilworth's theorem is equivalent to König's theorem.*

Proof: We first show that Dilworth's theorem follows from König's theorem. Let P be a partially ordered set with n elements. We define a bipartite graph $G = (U, V, E)$ where $U = V = P$, i.e. each partition in the bipartite graph is equally to P. Two nodes u, v form an edge $(u, v) \in E$ in the graph G iff $u < v$ in P. By König's theorem there exist both a matching M and a a vertex cover C in G of the same cardinality. Let $T \subset S$ be the set of elements not contained in C. Note that T is an antichain in P. We now form a partition W of P into chains by declaring u and v to be in the same chain whenever there is an edge $(u, v) \in M$. Since C and M have the same size, it follows that T and W have the same size.

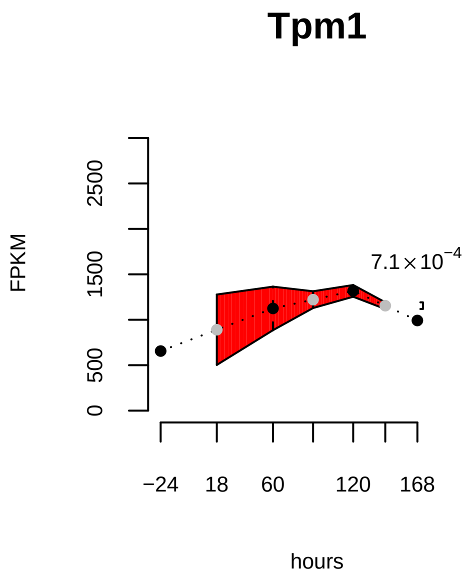
To deduce König's theorem from Dilworth's theorem, we begin with a bipartite graph $G = (U, V, E)$ and form a partial order P on the vertices of G by defining $u < v$ when $u \in U, v \in V$ and $(u, v) \in E$. By Dilworth's theorem, there exists an antichain of P and a partition into chains of the same size. The non-trivial chains in P form a matching in the graph. Similarly, the complement of the vertices corresponding to the anti-chain in P is a vertex cover of G with the same cardinality as the matching. □

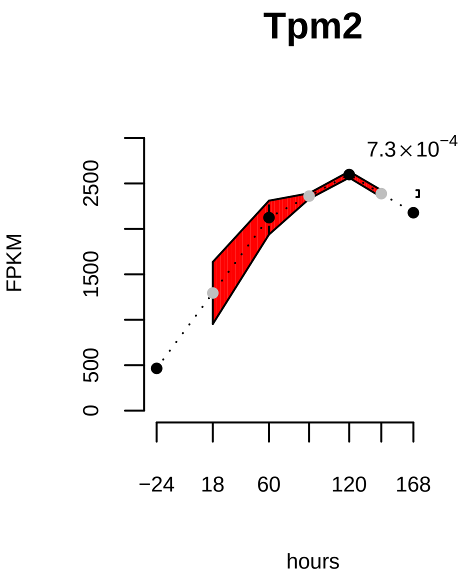


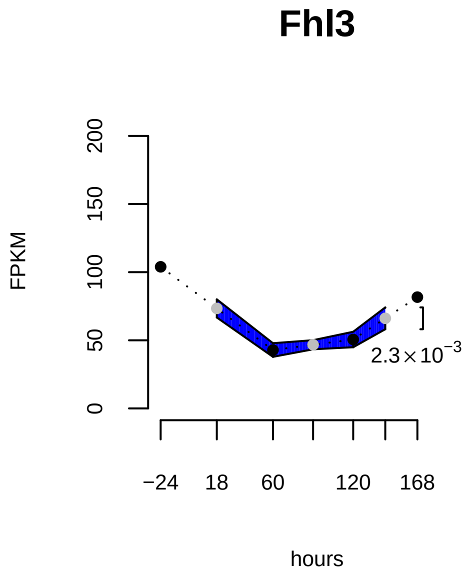


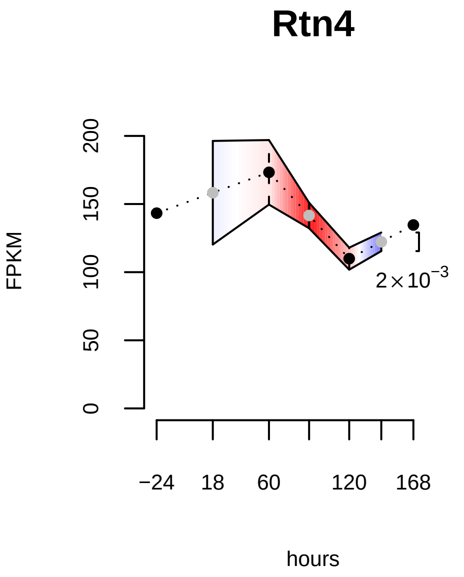
The equivalence of Dilworth's and König's theorems is depicted above. The partially ordered set with 8 elements on the left is partitioned into 3 chains. This is the size of a minimum partition into chains, and is equal to the maximum size of an antichain (Dilworth's theorem). The antichain is shown with double circles. On the right, the reachability graph constructed from the partially ordered set on the left is shown. The maximum matching corresponding to the chain partition consists of 5 edges and is equal in size to the number of vertices in a minimum vertex cover (König's theorem). The vertex cover is shown with double circles. Note that 8=3+5.

8. Appendix B: selected Minard plots

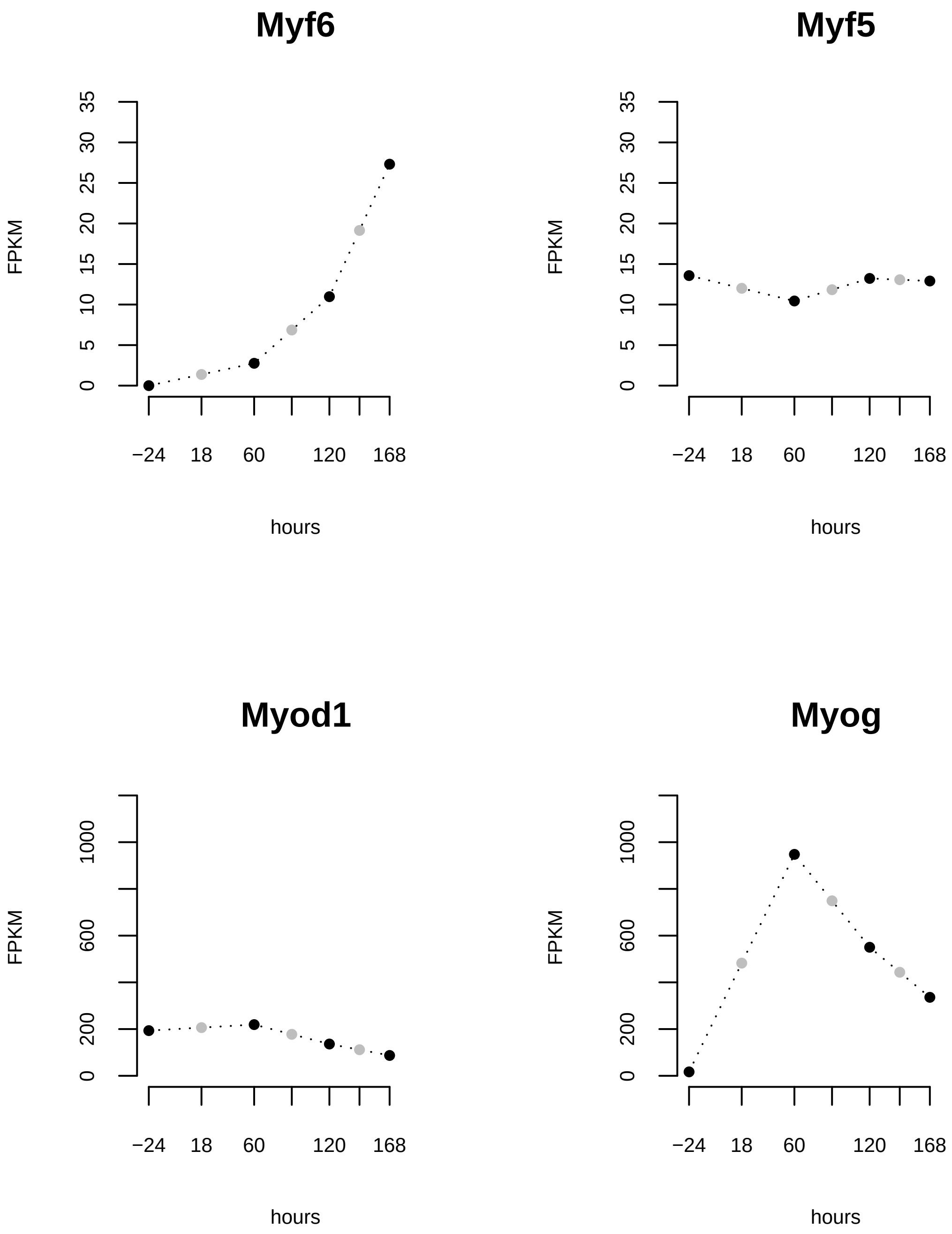


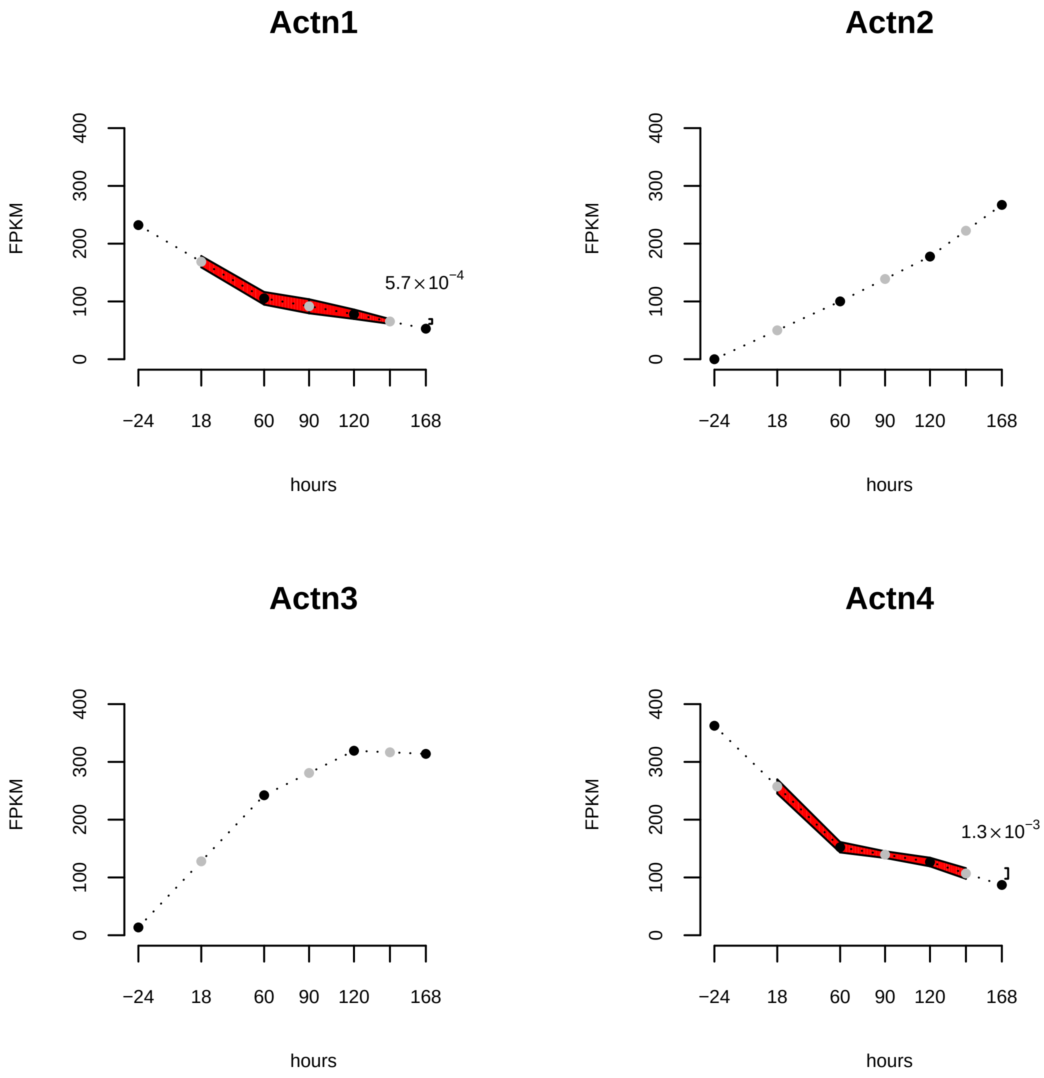


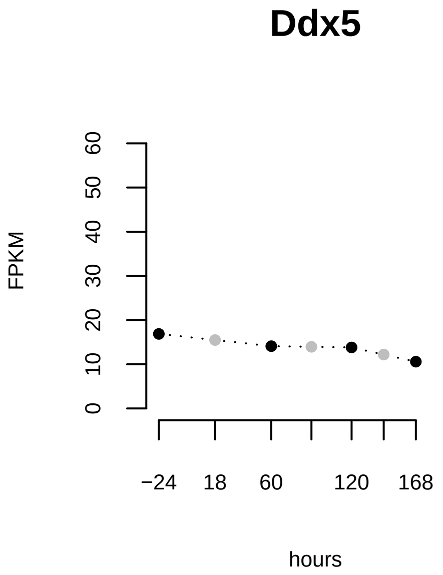


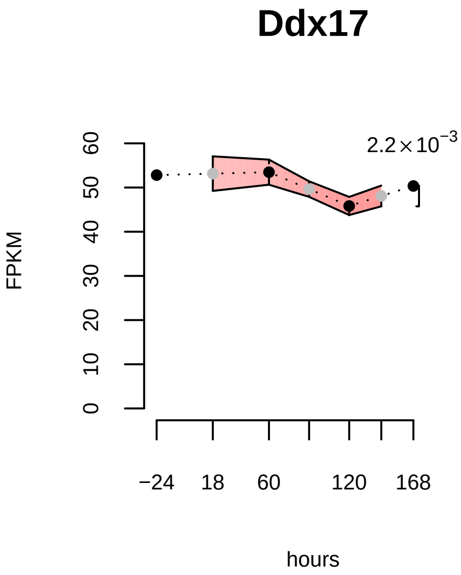


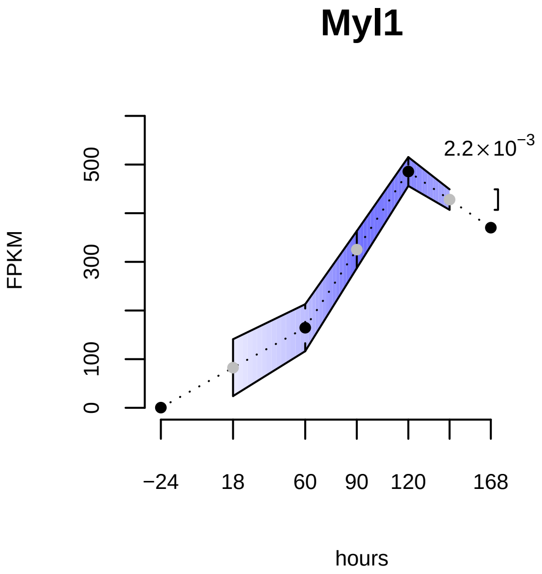
Nature Biotechnology: doi:10.1038/nbt.1621

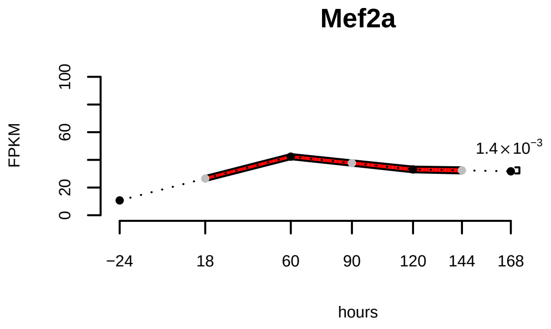

Myf6
FPKM
0
5
10
15
20
25
30
35
−24
18
60
120
168
hours
Myf5
FPKM
0
5
10
15
20
25
30
35
−24
18
60
120
168
hours
Myod1
FPKM
0
200
600
1000
−24
18
60
120
168
hours
Myog
FPKM
0
200
600
1000
−24
18
60
120
168
hours


Actn1
FPKM
0
100
200
300
400
−24
18
60
90
120
168
hours
5.7 × 10^{-4}
Actn2
FPKM
0
100
200
300
400
−24
18
60
90
120
168
hours
Actn3
FPKM
0
100
200
300
400
−24
18
60
90
120
168
hours
Actn4
FPKM
0
100
200
300
400
−24
18
60
90
120
168
hours
1.3 × 10^{-3}

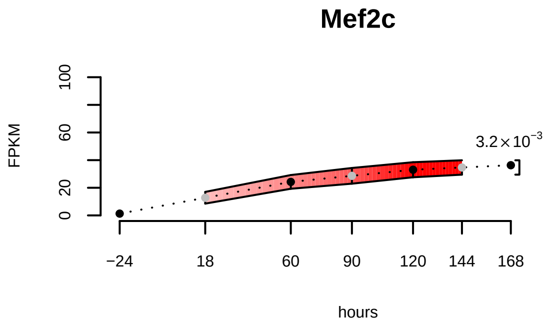

Ddx5
FPKM
0
10
20
30
40
50
60
−24
18
60
120
168
hours


Ddx17
FPKM
0
10
20
30
40
50
60
2.2×10^{-3}
−24
18
60
120
168
hours

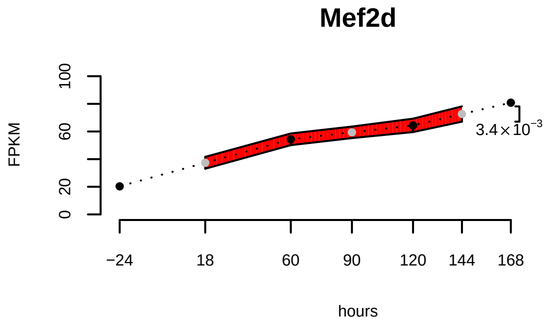

Myl1
FPKM
0
100
300
500
2.2×10^{-3}
−24
18
60
90
120
168
hours



Mef2a



Mef2c



Mef2d

References

1. LA Aroian, VS Taneja, and LW Cornwell, *Mathematical forms of the distribution of the product of two normal variables*, Communications in Statistics: Theory and Methods **A7** (1978), 165–172.
2. Y Benjamini and Y Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, Series B (Methodological) **57** (1995), 289–300.
3. JH Bullard, E Purdom, KD Hansen, and S Dudoit, *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*, BMC Bioinformatics **11** (2010), 94.
4. RP Dilworth, *A decomposition theorem for partially ordered sets*, The Annals of Mathematics **51** (1950), 161–166.

5. N Eriksson, L Pachter, Y Mitsuya, S-Y Rhee, C Wang, B Gharizadeh, M Ronaghi, RW Shafer, and N Beerenwinkel, *Viral population estimation using pyrosequencing*, PLoS Computational Biology **4** (2008), e1000074.
6. B Fuglede and F Topsøe, *Jensen-Shannon divergence and Hilbert space embedding*, Proceedings of the IEEE International Symposium on Information Theory, 2004, p. 31.
7. Lemon graph library, `http://lemon.cs.elte.hu/trac/lemon`.
8. BJ Haas, AL Delcher, SM Mount, JR Wortman, RK Smith, LI Hannick, R Maiti, CM Ronning, DB Rusch, CD Town, SL Salzberg, and O White, *Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies*, Nucleic Acids Research **31** (2003), 5654–5666.
9. D Hiller, H Jiang, W Xu, and WH Wong, *Identifiability of isoform deconvolution from junction arrays and RNA-Seq*, Bioinformatics **25** (2009), 3056–3059.
10. JE Hopcroft and RM Karp, *An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs*, SIAM Journal on Computing **2** (1973), 225–231.
11. H Jiang and WH Wong, *Statistical inferences for isoform expression in RNA-Seq*, Bioinformatics **25** (2009), 1026–1032.
12. B Langmead, C Trapnell, M Pop, and SL Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*, Genome Biology **10** (2009), R25.
13. B Li, V Ruotti, RM Stewart, JA Thomson, and CN Dewey, *RNA-Seq gene expression estimation with read mapping uncertainty*, Bioinformatics **26** (2009), 493–500.
14. H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data Processing Subgroup, *The sequence alignment/map format and SAMtools*, Bioinformatics **25** (2009), 2078–2079.
15. Boost C++ libraries, `http://www.boost.org/`.
16. L Lóvasz and MD Plummer, *Matching Theory*, American Mathematical Society Press, 2009.
17. A Mortazavi, BA Williams, K McCue, L Schaeffer, and B Wold, *Mapping and quantifying mammalian transcriptomes by RNA-Seq*, Nature Methods **5** (2008), 585–587.
18. A Oshlack and MJ Wakefield, *Transcript length bias in RNA-Seq data confounds systems biology*, Biology Direct **4** (2009), 14.
19. L Pachter and B Sturmfels (eds.), *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.
20. I Pe'er and JS Beckmann, *Recovering frequencies of known haplotype blocks from single-nucleotide polymorphism allele frequencies*, Genetics **166** (2004), 2001–2006.
21. W Ritchie, S Granjeaud, D Puthier, and D Gautheret, *Entropy measures quantify global splicing disorders in cancer*, PLoS Computational Biology **4** (2008), e1000011.
22. S Rozen and H J Skaletsky, *Primer3 on the WWW for general users and for biolgist programmers*, Methods and Protocols: Methods in Molecular Biology **4** (2000), 365–386.
23. M Sammeth, V Lacroix, P Ribeca, and R Guigó, *Flux capacitor simulator:* `http://flux.sammeth.net/`, 2009.
24. C Trapnell, L Pachter, and S Salzberg, *TopHat: discovering splice junctions with RNA-Seq*, Bioinformatics **25** (2009), 1105–1111.
25. Edward R. Tufte, *The visual display of quantitative information*, Graphics Press, 2001.
26. ET Wang, R Sandberg S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore, GP Schroth, and CB Burge, *Alternative isoform regulation in human tissue transcriptomes*, Nature **456** (2008), 470–476.
27. Y Xing, A Resch, and C Lee, *The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures*, Genome Research **14** (2004), 426–441.