SUPPLEMENTARY INFORMATION

Supplementary Methods

Models for the simulation of fragmentation

Nebulization. Already early reports on results from RNA-Seq experiments based on nebulization observed reads accumulating at the 5'-end of transcripts and around the center, especially of shorter transcript forms (28). These observations coincide well with breakpoint distributions obtained by a theoretical model of mechanical breakage that considers molecules as rigid stiffs (24), in which breakpoints recursively accumulate around the midpoint of iteratively broken fragments. According to this model, the average expected fragment size depends on the length of the nebulized DNA molecule: comparatively short molecules accumulate higher breaking probabilities during the time it takes to fragment the longer molecules in the transcript population.

In the light of these preliminary studies, we simulate nebulization by an iterative two-step process: first, a random orientation of the molecule in the shear field—i.e., the point (q) where the shearing stress is applied—is determined by random sampling under a Gaussian function centered at a molecule's midpoint. Subsequently, the breaking probability p_b is deduced from the exponential:

$$p_b = 1\text{- } \exp\text{-}((\min(q, len\text{-}q) + c) / \lambda)^M, \qquad (1)$$

where len is the molecule length, λ is a parameter that describes the limiting size below which molecules are very unlikely broken by the shearing field; M is a parameter describing the force of the shear field and determines the steepness of the slope in the resulting fragment size distribution; c finally is a constant that adjusts p_b to be 0.5 for a size exactly between λ ($p_b \rightarrow 0$) and 2λ ($p_b \rightarrow 1$). In our model, a Bernoulli trial on p_b determines whether a simulated break incurs at a given position. Recursive breaking continues until thermodynamics equilibrium as assumed by convergence of the fraction of breaks per time unit in the library falling below a given threshold (t=1%).

Hydrolysis. Frequencies $f(d)$ of fragment sizes d produced by a uniform random fragmentation process have demonstrated to fall along Weibull distributions (δ,η), if the fragmentation thermodynamics depends on the molecule size (25):

$$f(d) = \delta/\eta \ (d/\ \eta)^{\delta-1} \exp\text{—}(d/\eta)^{\delta} \qquad (2)$$

Scale parameter η represents the intensity of fragmentation (i.e., breaks per unit length), and—as a determinant of the mean expected fragment size—is assumed to be constant across molecules of different lengths for fragmentation protocols where the number of produced fragments depends on the molecule length. Shape parameter δ reflects the geometric relation in which random fragmentation is breaking a molecule (e.g., δ=1 corresponds to uniform fragmentation on the linear chain of nucleotides, δ=2 splits uniformly the surface, and δ=3 the volume, etc.).

Employing empirical data from spike-in sequences, we evaluated the fitting obtained by weighted subsampling from Weibull distributions with varying shape parameters. Weights for the subsampling (Fig. 2B, solid line) were derived by separating the characteristics of the combined Weibull distributions before filtering (dashed line in Fig. 2B and 2C) from the observed insert size distribution (Fig. 2B, dashed-dotted line). The quality of fit was measured as the *p*-value computed by a Kolgomorov-Smirnov test, comparing the *in silico* produced insert size distribution (Fig.2A, dashed lines) for each of the spike-in sequences under investigation with its experimental couterpart (Fig.2A, solid lines) under the null hypothesis that both samples were drawn from the same distribution. By this, we empirically found that the observed differences can be qualitatively reproduced under a constant decay rate (η=200nt), when shape parameter δ depends logarithmically on the molecule length (Supplementary Fig.4).

In our uniform random fragmentation model, we use a 3-step algorithm to tokenize a molecule; first, geometry δ and the number *n* of fragments that are obtained from the molecule are determined. We found empirically that parameter δ depends logarithmically on *len*, the length of the molecule that is fragmented $\delta=\log(len)$. The number of fragments produced from a specific RNA molecule is determined by $n=len/E(d_{max})$, where $E(d_{max})$ is the expectancy of the most abundant fragment size, computed from η and the gamma-function Γ of δ:

$$E(d_{max})= \eta\Gamma(1/\delta + 1) \qquad (3)$$

Second, $(n-1)$ breakpoints are sampled uniformly from the interval [0;1[, resulting in relative length fractions $x_1, \ldots, x_n$ for all *n* fragments. Third, relative fragment sizes x_i are transformed from unit space to sizes d_i that follow a Weibull distribution of shape δ by:

$$d_i = \left(\frac{x_i}{C}\right)^{\frac{1}{\delta}} \qquad (4)$$

where C= $(len\ /\ \Sigma(x_i^{1/\ \delta}))^{-\delta}$ is a constant of the transformation to ensure that the sizes of the *n* fragments sum up exactly to the given molecule length *len*. Latter transformation—other than the original model (25)—produces a slightly distorted Weibull distribution for the sizes d_i, however the deviation is sufficiently small to be neglected in our applications. The original work (25) also shows that the theory on fragment sizes falling along Weibull distributions holds for d_i<<*len*, which in RNA-Seq is not always satisfied because transcripts with lengths close to $E(d_{max})$ can exist.

A model for simulated reverse transcription

Our algorithm determines start point and extension separately for first and for second strand synthesis.

1. During first strand synthesis, poly-*d*T primers induce priming events in the poly-A tail, whereas random primers provoke successful initiation events along the entire molecule, and anchored primers trigger exactly one priming event at the 3-end of the respective fragment. In sequencing protocols without sequence-specific biases, each priming event is assigned a random location uniformly sampled along the corresponding stretch. Optionally, start points of first strand synthesis are determined by importance sampling according to weights of an optional PWM capturing sequence-biases (7,8).

2. The point where second strand synthesis initiates is simulated by the length of the first DNA strand, which can be between RT_{min} and RT_{max} nucleotides, but maximally the distance of the first strand synthesis priming event from the 5'-end of the RNA template. The point of priming second strand synthesis in the presence of sequence biases is drawn from a distribution according to the PWM capturing the bias, or from a uniform distribution otherwise.

In the case of multiple priming events with random primers, several enzymes concurrently transcribe parts of the RNA molecule, and collisions with downstream DNA-RNA hybrids are resolved by displacement according to a Bernoulli trial. Standard literature about molecular techniques (21) provides values for RT_{min}~500nt and for RT_{max} ~5,000nt, under which the model reproduces fairly well experimental characteristics (Results).

The Model for the simulated PCR Amplification

The efficiency of PCR amplification is either specified by an universal success rate *p*, or, by a normal distribution $p=f(mean_{GC},SD_{GC})$ parameterized to capture GC preferential biases (default $mean_{GC}=0.5$ and $SD_{GC}=0.1$). Given *p*, the number of copies produced from a certain fragment is determined by random sampling under the cumulative binomial:

$$P_S(N) = \sum_{k=0}^{\lceil \frac{N}{2} \rceil} P_{S-1}(N-k) \binom{N-k}{k} p^k (1-p)^{n-2k} \qquad (5)$$

with *S* denoting the PCR cycle and *N* the number of molecules. As default, we assume 15 PCR cycles (*S*=15), and sample randomly the number of duplicates yielded by PCR amplification under the corresponding probability distribution $P_{15}(N)$ for all possible values of $N=[1;2^{15}]$. The recursion terminates by

$$P_0(N) = 0$$
$$P_1(N) = \begin{cases} 1 : N = 1 \\ 0 : else \end{cases}$$

A quality-based model for Illumina Sequencing Errors

General models for simulating sequencing errors have been proposed for the Roche and the Illumina platform (22). Herein, we extend the proposed Illumina model to take into account quality-dependend crosstalk, i.e., the preference of substitution according to a certain quality value assigned (Supplementary Fig.2):

1. A quality value is randomly drawn from an empirical distribution, depending exclusively on the position within the read, i.e, the Illumina sequencing cycle. This correlation varies even between different sequencing chemistries of the Illumina platform as can be seen by comparison of 35nt reads (9) from 2008 to the 76nt reads (29) produced in 2010 (grey charts in Supplementary Fig.2 A and B).
2. According to the error probability intrinsic to the quality value, a Bernoulli trial decides whether the genomic base is mutated in the read sequence.
3. Our in-depth study of substitution rates reveals that preferences of miss-called nucleotides correlate stronger with quality values than with the read position (Supplementary Fig.2C). Therefore, we implemented in our model a first order Markov process that determines the mutated nucleotide based on the quality values of a read.

A general model for transcript expression in a cell population

Although gene expression generally follows Zipf's law, it also has been noted that in lowly expressed genes this simple model does not fit and a gradually increasing deviation from the log-log linear regression towards the least abundant genes is observed (Supplementary Fig.3). For all of the investigated datasets, this deviation decays even faster than exponential, and from the analyzed data we deduced by non-linear fitting a general correlation between a gene's expression rank *x* and its normalized expression level *y*

$$y = y_0 x^k exp^{\frac{x}{a}(\frac{x}{b})^2}, \qquad (6)$$

where y_0 is the—expression level of the most abundant gene, *k* is the exponent of the underlying Zipf's law which governs the slope of the log-log plot. In the experiments we investigated, this coefficient varied between -0.6 and -0.9 (Supplementary Tab.2). The parameters *a* and *b* control the exponential decay, and we empirically found that a=b~10^4 in our datasets. A previous study on gene expression derived from ESTs and SAGE (serial analysis of gene expression) tags demonstrated that the similarity between observed distributions and Zipf's law gets stronger when accumulating more data (2), which translates to our model as by changing parameters *a* and *b*.

In order to simulate realistic transcript expression levels as observed in cellular transcriptomes, we implemented a transcriptome simulator into the Flux Simulator pipeline (Methods). For the simulation, we modified Zipf's law that has been reported earlier as a basic rule for expressed genes (2,5) to observations of from RNA-Seq experiments (Supplementary Fig.3).

To estimate realistic ranges for the parameters of the model, we have compared theoretic results with previously published reports on the cellular distribution of gene expression levels, three of these are in human tissues (36-38), and one is in mouse (36). These traditional expression studies commonly cluster the cellular RNA complement into three domains—i.e., "super-prevalent", "intermediate" and "complex/rare" transcripts—and in order to distinguish between these classes we assumed a concentration of ~10^5 RNA molecules per cell (11) and adopted the thresholds of >500 copies per cell for "super-prevalent" transcripts and of >15 copies/cell for "intermediate" transcripts (38).

Supplementary Table 2 summarizes the comparison of values reported in the literature with our transcriptome simulations. Note that while the comparison is sensitive to the thresholds that we enforce to separate the different fractions, with a suitable range of parameters, Supplementary Formula 6 is able to reproduce quite appropriately the spectrum of expression profiles observed in mammalian cells.

Models for variation in transcription start sites and poly-A tails

Literature knows different types of transcription start site distributions depending on the specific promoter of a gene (39), however, a major class of genes exhibits an about exponential probability peak centered at the main transcription start site (40). Motivated by these observations we realized variation in transcription starts as exponentially distributed around the annotated start site (default mean 10nt)—obviously in an abstraction of the processes that determine the formation of RNA-polymerase initiation complexes, such as transcription factor concentrations, promoter-configuration and –occupancy, etc.

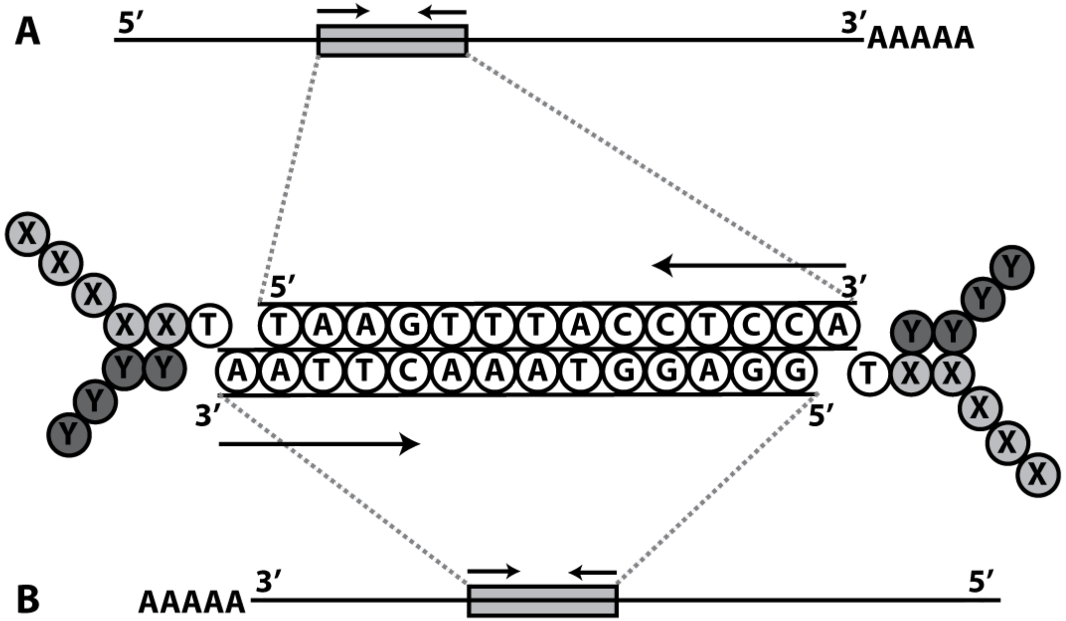
During poly-adenylation in the nucleus it is generally assumed that about 200-250 adenine residues are added to the primary transcript sequence (41). However, accurate estimates about the nature of poly-A tail sizes are currently still unavailable. Due to its flexibility, we therefore have choosen a Weibull-approximation of the normal distribution (shape=2, scale=300) to sample random lengths of poly-A tails.

Description of experimental datasets employed in the study

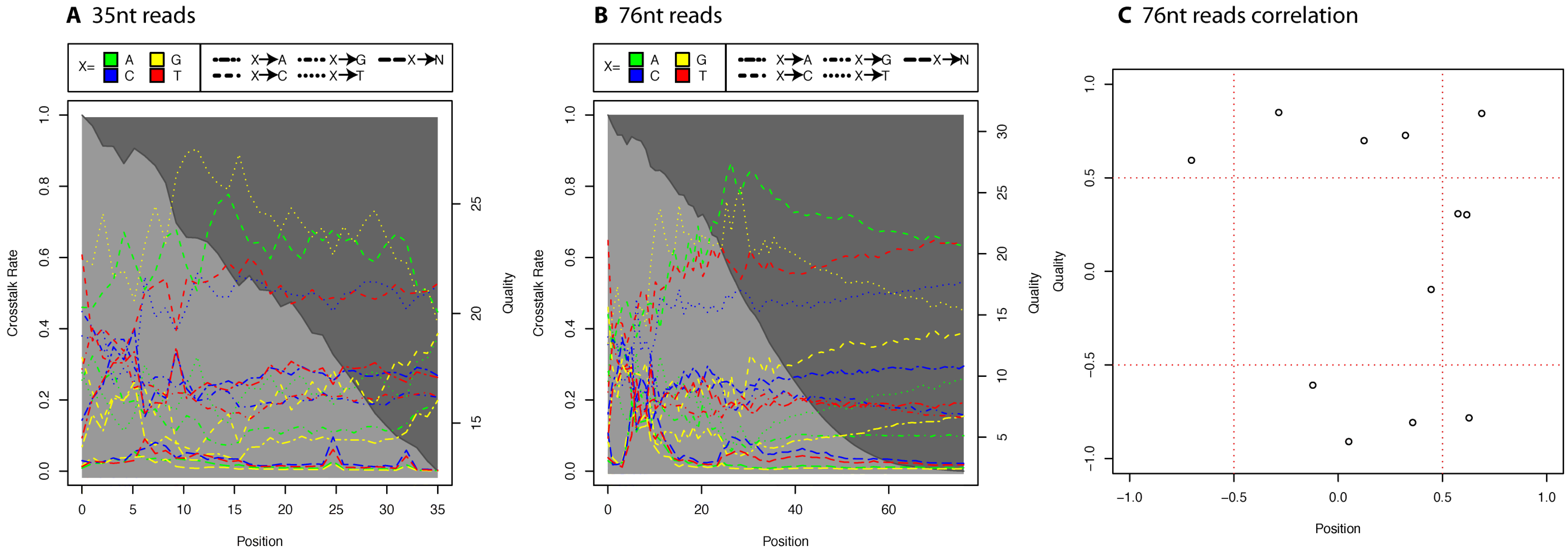
The paired-end dataset containing the spike-in control sequences (GSE20846) and the yeast dataset (GSE11209) have been obtained from the Gene Expression Omnibus database, the murine data (SRA001030) from the National Center for Biotechnology Information's (NCBI) short-read archive, the huan datasets from the European Nucleotide Archive (ERA000183), and the cress dataset from GenBank by accession numbers EH795234 through EH995233 and EL000001 through EL341852.

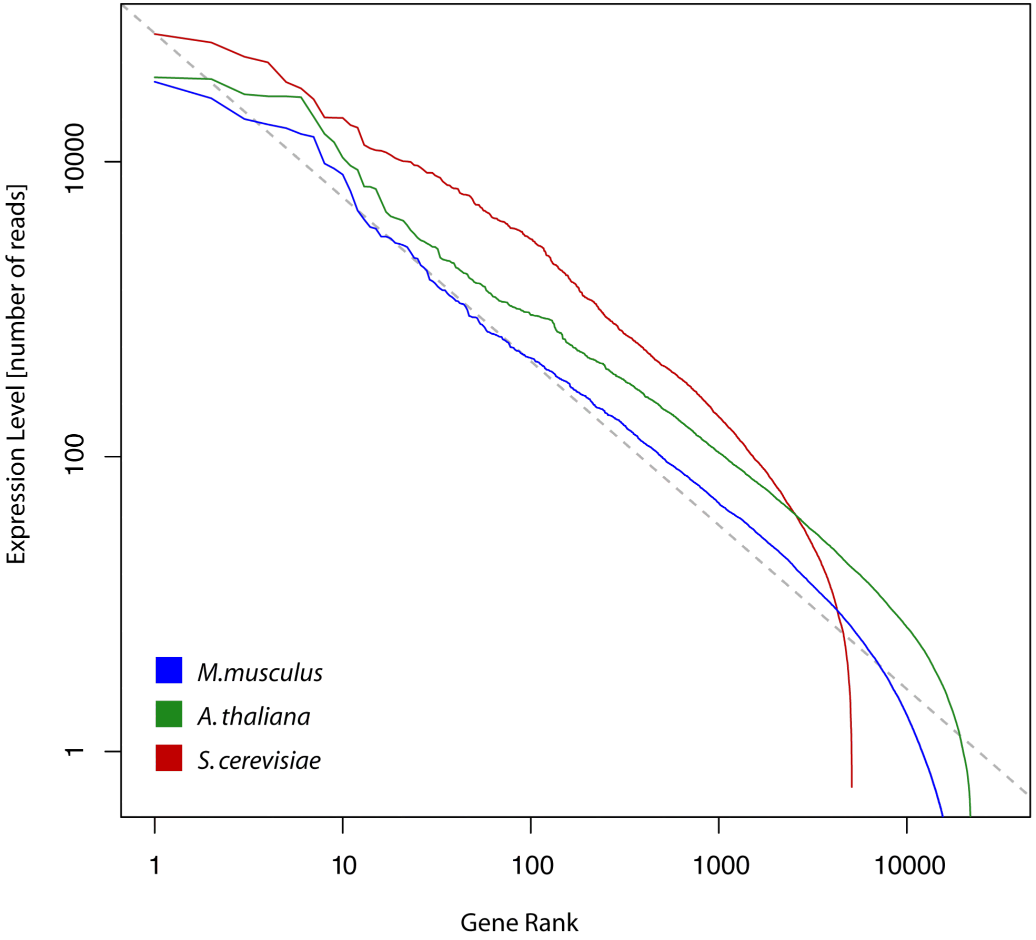
All these RNA-Seq experiments have targeted the poly-A+ RNAome, but substantially differ in the following points (Supplementary Tab.1): the mouse and the spike-in dataset has been produced by a protocol using exclusively random priming for first strand synthesis of RT, whereas the cress and the yeast dataset involved poly-*d*T primers. In the mouse/spike-in experiment RNA was hydrolzed prior to RT, in the cress experiment the cDNA library has been nebulized or fragmented by DNAseI in the yeast experiment, respectively. For the mouse and spike-in dataset 200nt +/- 25nt fragments, for the yeast experiment 100-300nt fragments have been selected from the cDNA library before sequencing, wehereas the cress experiment did not involve any size selection step. The mouse, spike-in, and the yeast dataset have been sequenced on the Illumina platform, whereas cress reads were obtained from a Roche 454 pyrosequencer (Supplementary Tab.1).

Supplementary Figure 1: Mapping directionality of reads with respect to the reference transcript in sequencing-by-synthesis technologies. Arrows mark the both of the four possible ends in double stranded DNA molecules which can serve as templates for the sequencing-by-synthesis reaction, giving rise in this cartoon to the read sequences ATTCA... and GGAGG... respectively. One of the possible reads naturally is a subsequence of the transcript and the other possible read is reverse-complemented compared to the transcript's sequence, regardless of which of the two cDNA strands represents the orientation of the underlying transcript (both possible scenarios are depicted as A and B). By rotational symmetry of the ends that can be sequenced, sense reads always represent the upstream end of the underlying RNA fragment, whereas anti-sense reads stem from downstream ends.

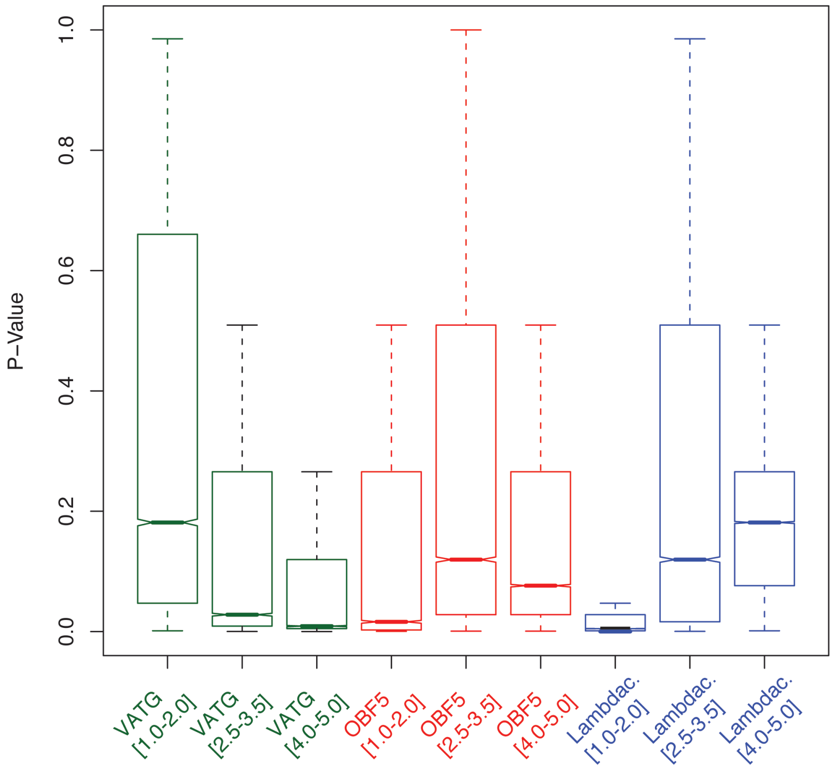


Supplementary Figure 2: Illumina read error models. Mismatches observed after mapping reads obtained by different Illumina sequencing chemistries (35nt respectively 76nt long) to the corresponding genomic reference have been clustered according to the observed nucleotide transition (i.e., the crosstalk). Panel A and B depict these transition probabilities (colored curves, scale on the left axis) as well as the median quality level (grey chart, right axis) along the read sequence (x-axis). For each of the crosstalk transition types we measured the correlation of the observed rate with the position in the read (i.e., cycle number) and with the quality value. Panel C shows the scatter plot between the rate-position and the rate-quality correlation coefficients, pinpointing quality values as a determinant of the transition rate by strongly polarized correlation coefficient values (>0.5 and <-0.5).



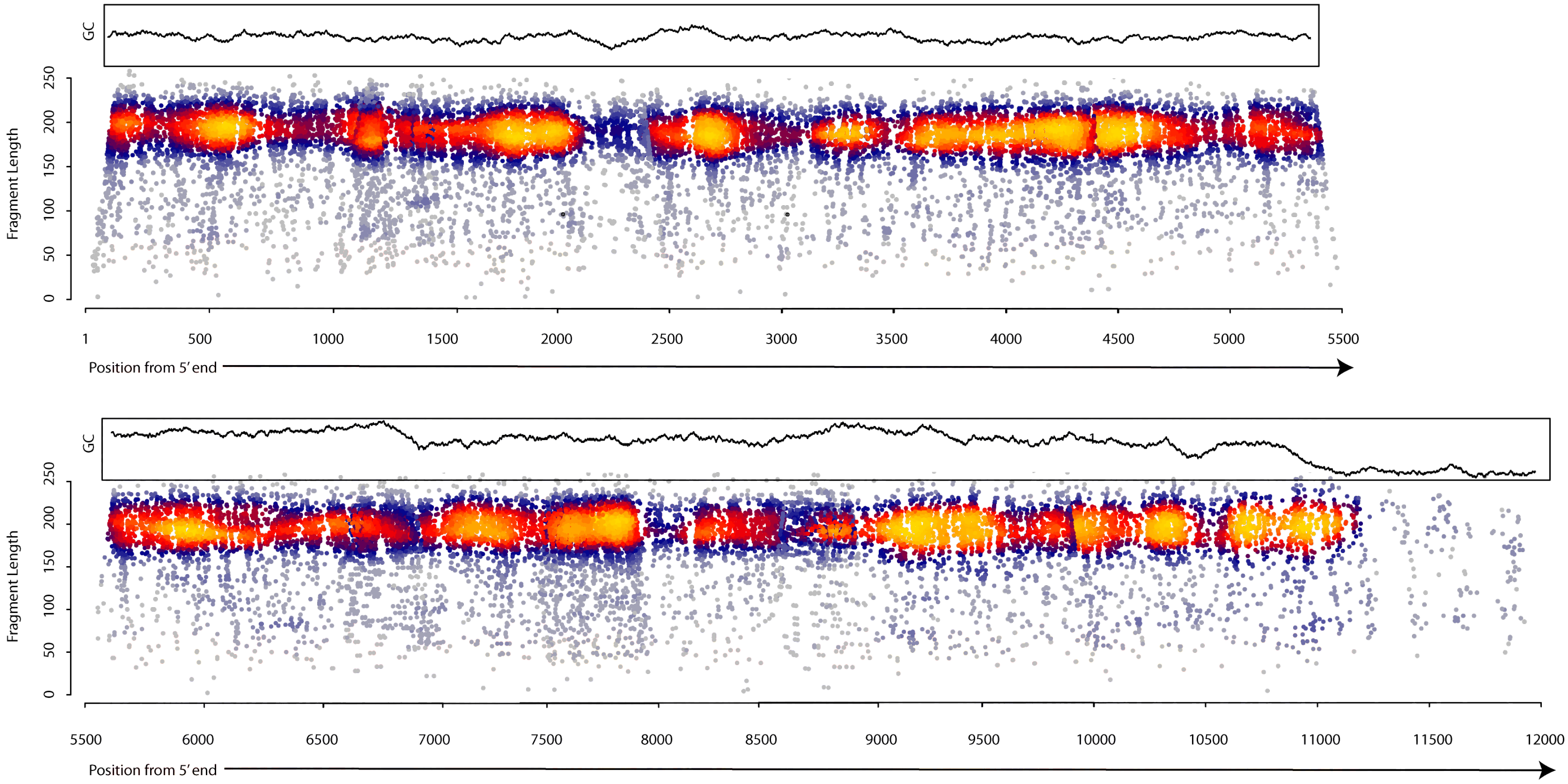


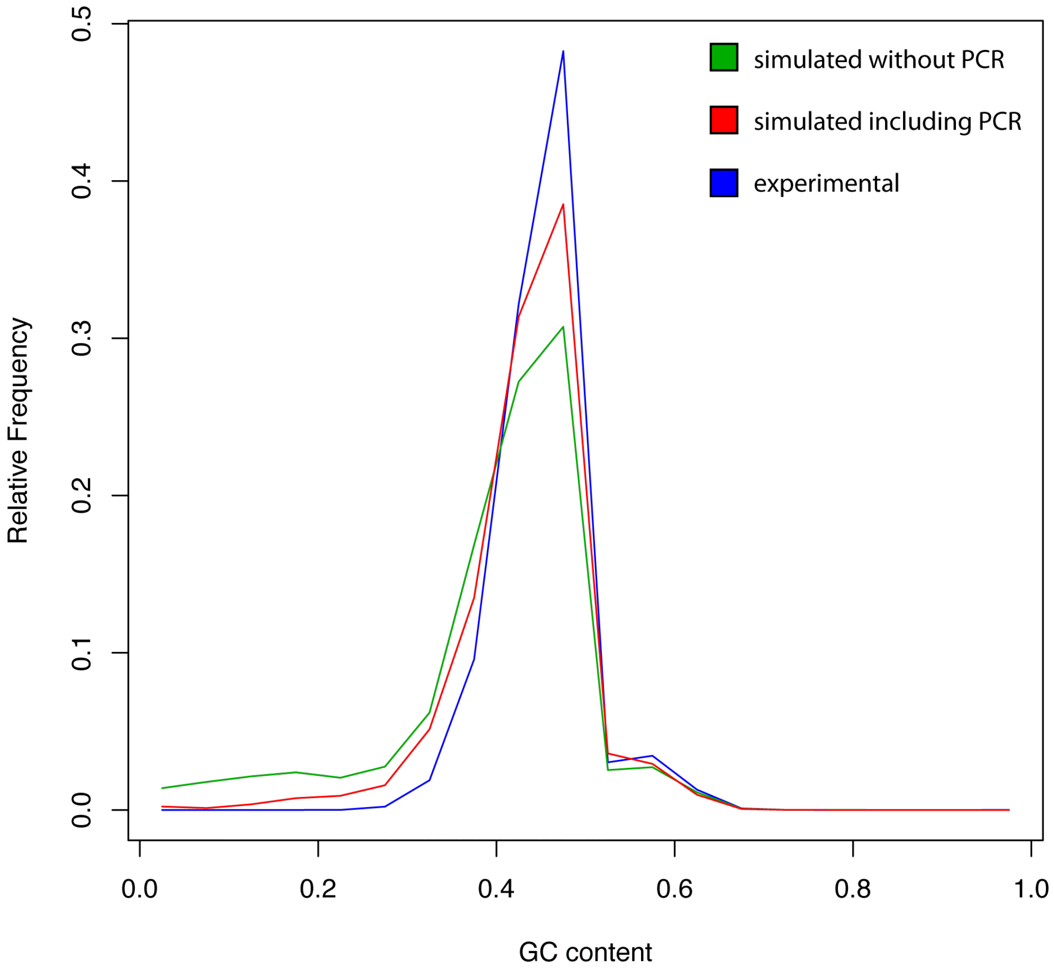
Supplementary Figure 3: Expression profiles observed in RNA-Seq experiments. The curves show the log-log behaviour of transcript expression in RNA-Seq experiments conducted on cellular transcriptomes of the species *M.musculus* (blue), *A.thaliana* (green) and *S.cerevisiae* (red). Expression values for every gene in a corresponding reference annotation (i.e., the murine RefSeq, the TAIR9 annotation of cress, and the SGD yeast annotation) have been estimated by the number of reads mapping to it, and expression levels have been ranked from high to low (x-axis). Although target cells and RNA-Seq experiment protocols differ substantially, all datasets show highly similar characteristics in their transcript abundance distribution: the nature of Zipf's Law underlying gene expression can be noted by the largely linear behaviour in logarithmic scale. However, especially for lowly abundant forms, an exponential decay is notable.



Supplementary Figure 4: Evaluation of Weibull parameters to fit experimentally observed insert size distributions. Within the parameter space of η=[150;250] for scale and δ=[1.0;5.0] for shape, 91,125 combinations of different Weibull distributions were computed, representing the theoretic fragment sizes of the spike-in sequences VATG, OBF5 and Lambdaclone1-1 before filtering. For each tuple of simulated distributions, the combined probability distribution is employed for correction of experimentally observed insert size frequencies. The corrected frequencies subsequently are applied as filtering weights to produce *in silico* predictions, which are compared to the experimental results by p-values computed by Kolgomorov-Smirnov tests, for each of the spike-in sequences separately. The boxplot shows the distribution of these p-values grouped in bins of shape parameters used to simulate the underlying Weibull distribution (i.e., [1.0;2.0], [2.5;3.5], [4.0; 5.0]) for VATG (green), OBF5 (blue) and Lambdaclone1-1 (red). The quality of the fit is best, i.e. p-values for rejecting the null hypothesis of simulated and experimental data following the same distribution are closest to 1, with simulated Weibull curves of shape [1.0;2.0] for short transcripts as VATG (376nt), shape [2.5;3.5] for messenger-sized RNAs as OBF5 (1,429nt), and shape [4.0;5.0] for very long transcripts as Lambdaclone1-1 (11,934nt).

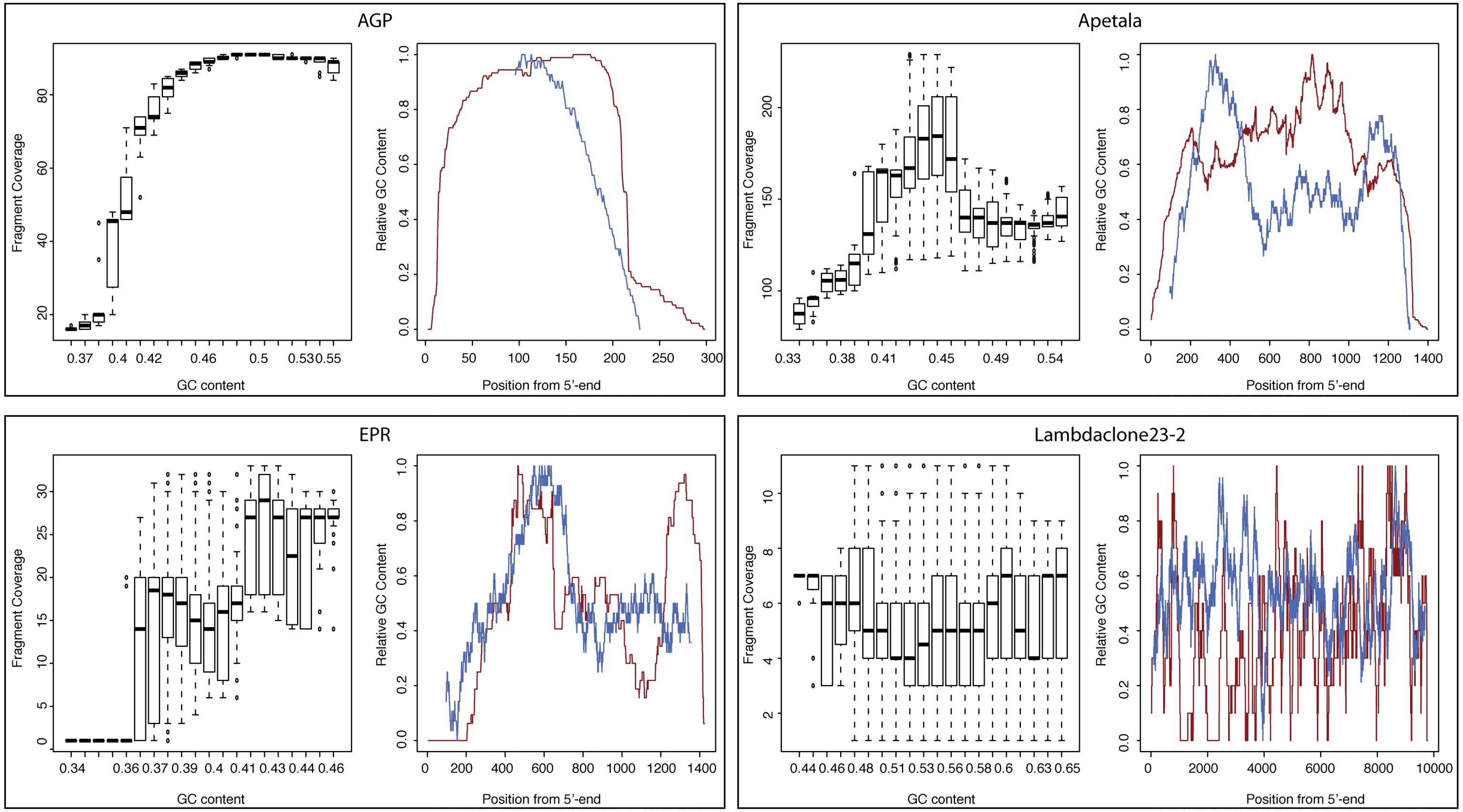
Supplementary Figure 5: Positional biases of fragments in the Lambdaclone1-1 spike-in control. The density scatter plot shows the concentration of fragment centers along the Lambdaclone1-1 control sequence (x-axis), segregated by their lengths (y-axis). The top curve summarizes the relative GC content in windows corresponding to the average fragment size (192nt) centered at the corresponding position.

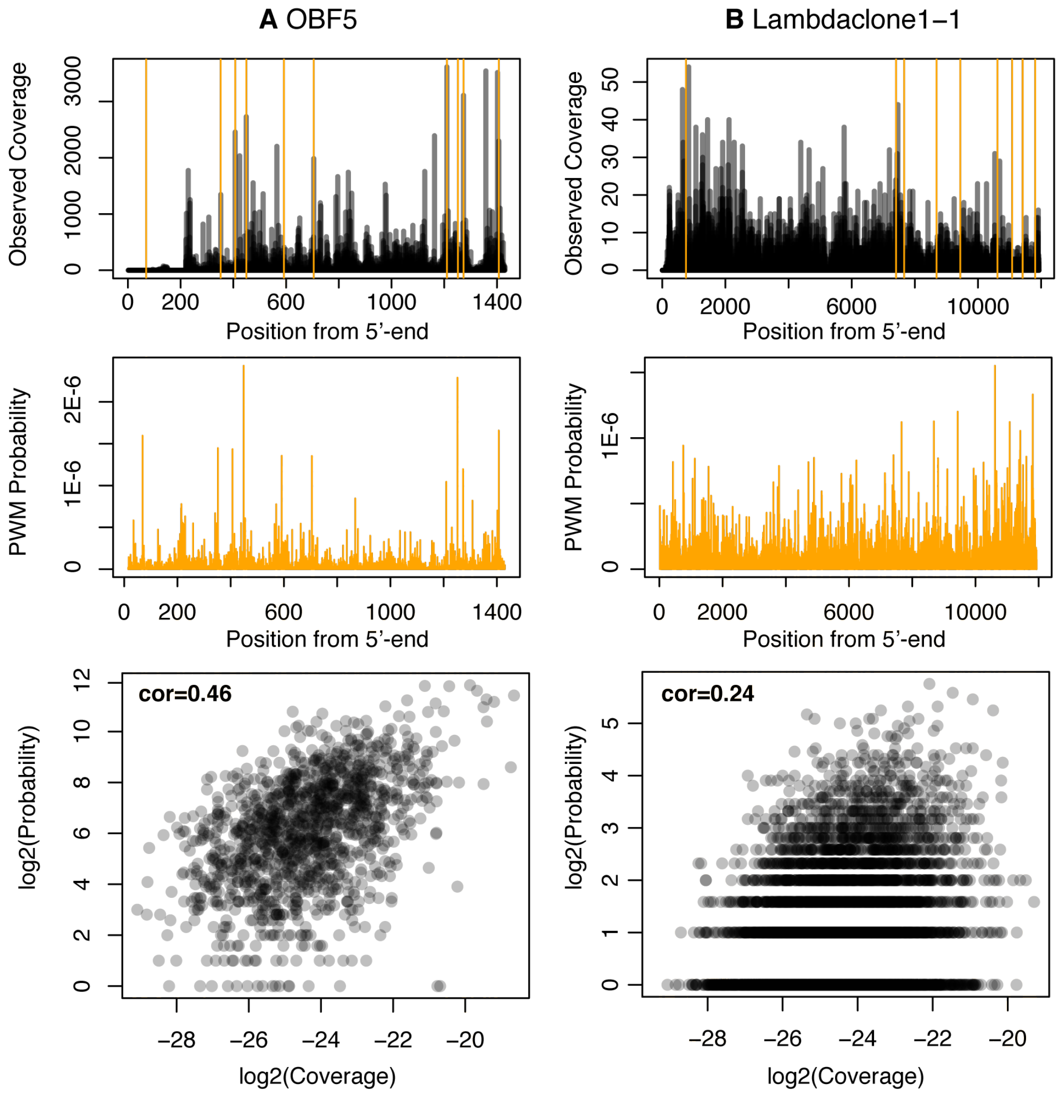




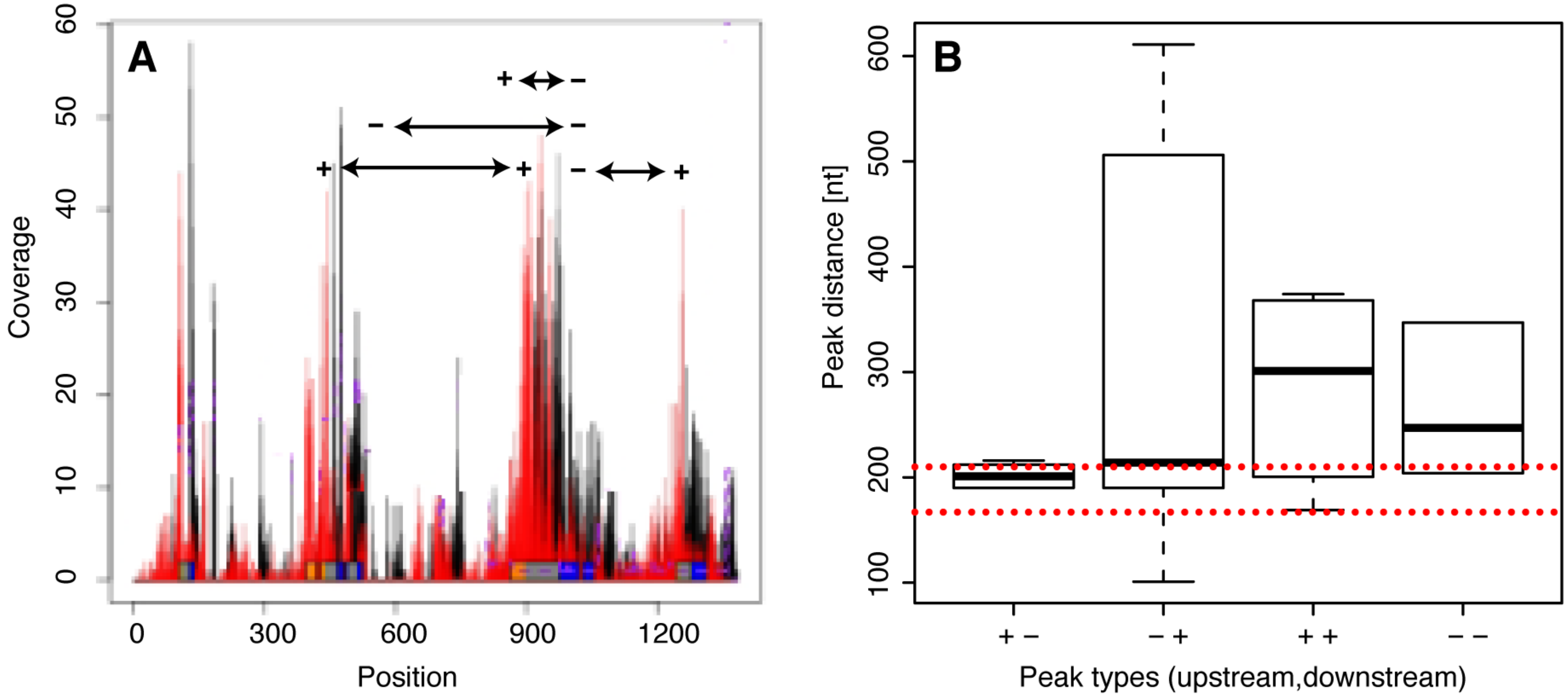
Supplementary Figure 6: GC content histograms before and after PCR amplification. The distribution of GC content in fragments of the simulated primary library (i.e., before PCR-amplification, green curve) differs markedly from the distribution found by experimental results (blue curve). The PCR model in the Flux Simulator assumes a Gaussian distribution (mean=50% GC content and standard deviation 10% GC content) for the amplification efficiency based on a transcript's sequence composition. Applying the *in silico* model changes the composition of fragments in the final library (red curve) and shifts the original GC distribution (green curve) towards the distribution deduced from the experiment (blue curve).

Supplementary Figure 7: GC content and fragment bias of additional spike-in sequences. GC content of the control sequences AGP (top-left panel), Apetala (top-right), EPR (bottom-left) and Lambdaclone23-2 (bottom-right) has been measured along each transcript considering a window of the same size as the average fragment length (192nt). To measure coverage, the number of fragments that include a position is considered. The left panel shows for each of the spike-in sequences the correlation between positions binned by their GC content and coverage, the right panel depicts the normalized coverage (red curve) and GC content (blue curve) along the transcript body.





Supplementary Figure 8: Prediction of sequence biases by position weight matrices. The top panels show points where first strand synthesis initiated for the OBF5 (**A**) and the Lambdaclone1-1 (**B**) control sequences. Sequence biases observed in the experiment have been collected in a position weight matrix, which subsequently was employed to score every position of a sequence (middle panels). The bottom panels summarize the correlation between observation and prediction in log-scale. Although the highest peaks of the predicted profiles (yellow bars in top panels) coincide fairly often with peak observations from experimental data, but probabilities derived from position weight matrices correlate only moderately with the observed frequencies (Pearson coefficient 0.46 and 0.24, respectively).



Supplementary Figure 9: Mapping directionality of stacked reads reveals insert size distribution. (**A**) Reads mapped to the spike-in control sequences are classified according to their alignment directionality into sense (symbol "+") and anti-sense ("–"). Read stacks are identified by 18 or more reads aligning at the same position, and the distance of each read stack to the next downstream stack in the transcript (i.e., the next "+" and the next "–" stack, if present) is determined. (**B**) The boxplot of these distance distributions shows that exclusively distances between "+/–" read stacks describe a distribution that falls largely within the fragment size range of the experiment (red bars marking Q1=167nt and Q3=210nt of the fragment size distribution determined by read pairing).

Supplementary Table 1: Summary of reference datasets. The table summarizes experimental data sets employed in our studies.

Species	Tissue	Platform	RT	Fragmentation	Size Selection	Read Length	Mappings
Spike-in (29)	–	Illumina	random	hydrolysis	200nt±25nt	76nt x 2	530,995
H.sapiens FRT (17)	placenta	Illumina	anchored	hydrolysis	(> adapter)	37nt x 2	9,184,734
H.sapiens STD (17)	placenta	Illumina	anchored	hydrolysis	(> adapter)	37nt x 2	9, 529,116
H.sapiens [1]	16 tissues	Illumina	anchored	hydrolysis	mean 350nt	100nt	204,768,395
M.musculus (11)	liver	Illumina	random	hydrolysis	200nt±25nt	25nt	31,577,110
A.thaliana (28)	seedling	Roche454	poly-*d*T	nebulization	–	94nt±20nt	541,852
S.cerevisiae (9)	–	Illumina	poly-*d*T	DNaseI	100nt-300nt	35nt	14,125,182

[1] ERP000546 in the European Nucleotide Archive

Supplementary Table 2: Evaluation of the transcriptome simulator. We compared expression values yielded by the transcriptome simulation implemented in the Flux Simulator to literature values for human and murine cell types (rows). For super-prevalent, intermediate and complex/rare transcripts, the number of distinct spliceforms (column "transcripts") and the respective fraction of RNA mass they constitute (column "%RNA") are presented. Note that the resolution of abundance classification is general and often not clearly delineated, which especially affects the sensitive threshold that separates "super-prevalent" from "intermediate" forms.

Reference cell type	**expressed**	**super-prevalent**		**intermediate**		**complex/rare**	
	transcripts	transcripts	%RNA	transcripts	%RNA	transcripts	%RNA
typical mammalian cell (37)	10,000 – 30,000	NA		NA		NA	
human fibroblast (42)	12,000	NA		NA		11,000	~30[1]
typical somatic cell (38)	16,000 –22,000	10 – 15	10-20	1,000 – 2,000	40– 45	15,000 – 20,000	40 – 45[1]
Simulation[2] (EnsEMBL hg18)	14,457 – 34,230	4 – 18	3– 11	985 – 2,528	39 – 58	13,547 – 31,684	38 – 51
typical murine cell (36)	10,500 – 22,000	5 – 10	>20%	500 – 2,000	40 – 60	10,000 – 20,000	<20 – 45
Simulation[3] (EnsEmbl mm9)	11,295 – 28,092	8 – 22	7%– 24%	1,188 – 1,457	46 – 57	9,551 – 26,934	19 – 43

[1] Wheras Williams (42) reports transcripts with <14 copies/cell as "complex/rare", the fraction specified by Martin and Pardee (38) refers to transcripts with < 15 copies/cell. [2] Parameter space k=(-0.6) to (-0.7), y_0=10^4, a=b=9,000 – 25,000. [3] Parameter space k=(-0.7) to (-0.8), y_0=10^4, a=b=5,500 – 15,500.

Supplementary Table 3: Flux Simulator parameters employed for the simulations. The table shows name-value pairs for all parameters to produce the simulation results presented in Fig. 6. Default parameter values are in italics.

M.musculus		***H.sapiens***		***S.cerevisiae***		***A.thaliana***	
NB_MOLECULES	5,000,000	NB_MOLECULES	5,000,000	NB_MOLECULES	5,000,000	NB_MOLECULES	5,000,000
TSS_MEAN	*25*	TSS_MEAN	50	*TSS_MEAN*	*25*	TSS_MEAN	100
POLYA_SCALE	*300*	POLYA_SCALE	NaN	POLYA_SCALE	80	POLYA_SCALE	200
POLYA_SHAPE	*2*	POLYA_SHAPE	NaN	POLYA_SHAPE	2	POLYA_SHAPE	1.5
FRAG_SUBSTRATE	RNA	FRAG_SUBSTRATE	RNA	*FRAG_SUBSTRATE*	*DNA*	*FRAG_SUBSTRATE*	*DNA*
FRAG_METHOD	UR	FRAG_METHOD	UR	FRAG_METHOD	EZ	FRAG_METHOD	NB
FRAG_UR_ETA	170	FRAG_UR_ETA	350	FRAG_EZ_MOTIF	DNAsel.pwm	FRAG_NB_LAMBDA	600
FRAG_UR_D0	*1*	*FRAG_UR_D0*	*1*			FRAG_NB_M	5
RTRANSCRIPTION	*YES*	*RTRANSCRIPTION*	*YES*	*RTRANSCRIPTION*	*YES*	*RTRANSCRIPTION*	*YES*
RT_PRIMER	*RH*	*RT_PRIMER*	*RH*	RT_PRIMER	PDT	RT_PRIMER	PDT
RT_LOSSLESS	*YES*	*RT_LOSSLESS*	*YES*	*RT_LOSSLESS*	*YES*	*RT_LOSSLESS*	*YES*
RT_MIN	*500*	*RT_MIN*	*500*	RT_MIN	500	RT_MIN	400
RT_MAX	*5,500*	*RT_MAX*	*5,500*	RT_MAX	2,500	RT_MAX	2,600
GC_MEAN	0.5	GC_MEAN	NaN	GC_MEAN	0.5	PCR_ROUNDS	13
FILTERING	YES	PCR_PROBABILITY	0.05	FILTERING	YES	*FILTERING*	*NO*
SIZE_SAMPLING	MH	FILTERING	NO	SIZE_SAMPLING	MH		
READ_NUMBER	15,000,000	READ_NUMBER	150,000,000	READ_NUMBER	1,000,000	READ_NUMBER	2,000,000
READ_LENGTH	75	READ_LENGTH	75	*READ_LENGTH*	*36*	READ_LENGTH	100
PAIRED_END	YES	PAIRED_END	YES	PAIRED_END	NO	PAIRED_END	NO

SUPPLEMENTARY REFERENCES

36. Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res*, **10**, 1617-1630.
37. Davidson, E.H. (1976) *Gene Activity in Early Development*. Acad. Press, New York.
38. Martin, K.J. and Pardee, A.B. (2000) Identifying expressed genes. *Proc Natl Acad Sci U S A*, **97**, 3789-3791.
39. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, **38**, 626-635.
40. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**, 1509-1517.
41. Bienroth, S., Keller, W. and Wahle, E. (1993) Assembly of a processive messenger RNA polyadenylation complex. *Embo J*, **12**, 585-594.
42. Williams, J.G. (1981) In R.Williamson (ed.), *Genetic engineering*. Acad.Press, Vol. 1, pp. 2.