

Counting absolute numbers of molecules using unique molecular identifiers

Teemu Kivioja^{1–3,5}, Anna Vähärautio^{1,3,5},
Kasper Karlsson⁴, Martin Bonke¹, Martin Enge³,
Sten Linnarsson⁴ & Jussi Taipale³

Counting individual RNA or DNA molecules is difficult because they are hard to copy quantitatively for detection. To overcome this limitation, we applied unique molecular identifiers (UMIs), which make each molecule in a population distinct, to genome-scale human karyotyping and mRNA sequencing in *Drosophila melanogaster*. Use of this method can improve accuracy of almost any next-generation sequencing method, including chromatin immunoprecipitation–sequencing, genome assembly, diagnostics and manufacturing-process control and monitoring.

Determining the relative abundance of two different molecular species or the absolute number of molecules in a single sample is challenging. We describe an absolute counting method that can use amplification but does not require detecting each original molecule or keeping track of the number of copies made. In this method, each molecule in a population is first made unique. This can be accomplished by adding a random DNA sequence label, by fragmenting or by taking an aliquot of a complex mixture that is small enough to contain only distinct molecules (Fig. 1a–c). Any combination of these manipulations can be used to generate a library in which each molecule has a distinct identifying

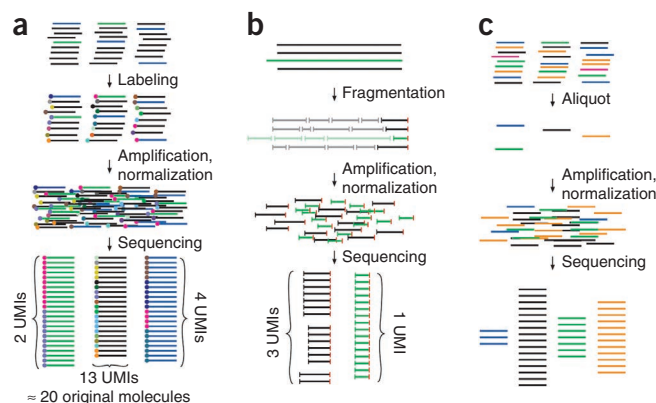
sequence. We designate the resulting sequences that can be used to uniquely identify copies derived from each molecule unique molecular identifiers (UMIs; Fig. 1).

As long as the complexity of the library of molecules is maintained, the library can be amplified, normalized or otherwise processed without loss of information about the original molecule count because the number of UMIs in the library acts as a molecular memory of the number of molecules in the starting sample (Supplementary Fig. 1). Upon deep sequencing, each UMI will be observed multiple times, and the number of original DNA molecules can be determined simply by counting each UMI only once. However, long before all UMIs are observed, increasingly precise estimates of the absolute molecule number can be made (Online Methods). This is in contrast to other counting methods, such as direct single-molecule sequencing^{1–3}, which require that all counted molecules are observed directly. In addition, many existing digital molecule counting methods such as digital PCR⁴, digital microarray profiling⁵ and direct single-molecule sequencing^{1–3} cannot be effectively multiplexed and are thus generally only applicable to measuring one or few molecular species from many samples, or many species from a single sample. Counting methods that introduce random tags to make molecules unique before amplification have been suggested^{5,6} and applied to analysis of RNA–protein interactions^{7,8}. In addition, three recent publications applied such labeling methods to the analysis of selected target genes by using either microarrays⁹ or sequencing^{9–11}. Here we apply the idea more generally and show that it can be used for absolute quantification.

The UMI method is very effective on simulated data (Supplementary Fig. 1). To assess whether it can be used to improve experimental measurements, we applied UMI counting to digital karyotyping and mRNA sequencing (mRNA-seq).

Figure 1 | UMIs can be generated by adding oligonucleotide labels, fragmenting, taking a small enough aliquot or a combination thereof.

(a) Three different DNA species (green, blue and black lines) are labeled with a collection of random labels (colored filled circles). Two green molecules are originally present (top), corresponding to two different UMIs (red, blue) among the sequenced molecules (green; bottom). Information about the original number of molecules (top) is preserved in the number of different UMIs detected by sequencing a sample of the amplified and normalized library (bottom). Even if some UMIs are not observed, the original number of molecules can be estimated using count statistics. (b) The original molecule is randomly fragmented, and a short unique sequence from the resulting fragments constitutes each UMI; here only the fragment adjacent to the poly(A) sequence (red vertical bars) is amplified. (c) An aliquot is taken from a sample that has many identical molecules such that on average, less than one copy of each molecule remains.



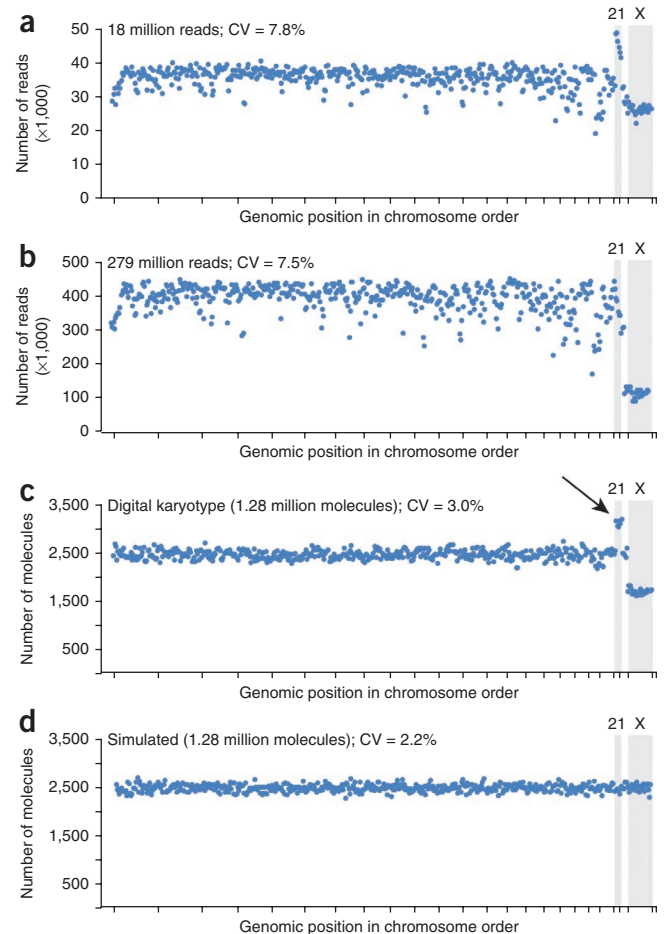
¹Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Helsinki, Finland. ²Department of Computer Science, University of Helsinki, Helsinki, Finland.

³Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. ⁴Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. ⁵These authors contributed equally to this work. Correspondence should be addressed to J.T. (jussi.taipale@ki.se) or S.L. (sten.linnarsson@ki.se).

Figure 2 | Digital karyotyping by counting the absolute number of molecules. (a) Standard digital karyotype based on genomic DNA from a boy with trisomy 21 and from his mother, mixed 1:1. (b) Standard digital karyotype of a sample from a male with a normal chromosome count. (c) The same sample as in a was analyzed by UMI counting (CV = 3.0%). Arrow highlights uniformly elevated copy number of regions in chromosome 21. (d) Simulated sample by uniform random sampling of 1.28 million molecules *in silico* from the NCBI human genome build 37 (CV = 2.2%). Number of reads and of molecules aligned to each 5-megabase-pair window is indicated. Chromosomes 21 and X are indicated by shading (the Y chromosome was excluded because its sequence was too repetitive for reliable alignments).

For digital karyotyping, we mixed equal amounts of genomic DNA from a boy with Down's syndrome and his mother. We fragmented the mixed DNA to generate a library of molecules, after which we took a sample containing less than a single genome copy. In combination with fragmentation, the use of a small aliquot reduces complexity such that each molecule is expected to have unique ends (Fig. 1b,c). The mapped genomic position of either end can be used as an UMI. After amplification by PCR and sequencing of 20 million reads, we binned read counts in 5-megabase-pair intervals. Total counts did not clearly show that half of the sample was derived from DNA with trisomy 21 and a single copy of the X chromosome (Fig. 2). In contrast, reanalyzing the sample by counting UMIs clearly revealed increased and decreased copy numbers of all 5-megabase-pair intervals in chromosomes 21 and X, respectively. As cell-free DNA from the plasma of pregnant women contains a mixture of parental and fetal DNA, detection of aberrant chromosome copy numbers is relevant to noninvasive prenatal diagnostics^{12,13}, although in clinical samples the ratio of fetal to parental DNA is generally lower than what we used here. The accuracy of the UMI method was close to the theoretical limit imposed by the sample size (Fig. 2), suggesting that development of the UMI method for diagnostic use appears feasible.

Unlike read-counting methods that are inherently limited by errors introduced during amplification, the UMI method can be made more accurate by increasing sample size and sequencing depth. Accuracy can be increased additionally by considering the number of unique consecutive and unique overlapping fragments. This is because consecutive fragments are likely to be derived from a single chromosome molecule, whereas overlapping fragments



must all be derived from different copies of the same chromosome. In simulated experiments, we found that this information can be used to accurately estimate the original number of molecules even when only a small fraction of the fragments derived from them are detected (Supplementary Fig. 2 and Supplementary Note).

Generating UMIs from both high- and low-abundance species in a single reaction requires the use of sequence labels (Fig. 1a and Online Methods). We tested a labeling protocol for counting cellular

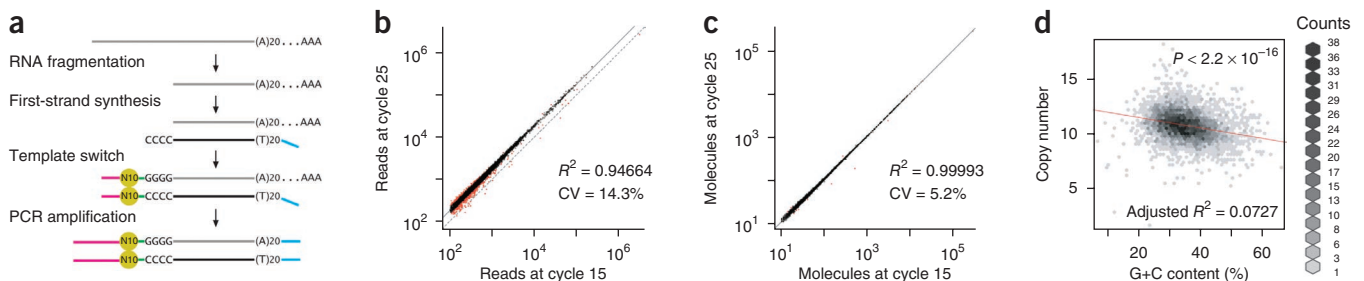


Figure 3 | Accuracy of mRNA-seq can be improved by the UMI method. (a) mRNA-seq libraries were generated by fragmenting total RNA and reverse-transcribing it to cDNA using an oligo(dT) primer with an Illumina linker (blue) and a 5' template switch adaptor containing another Illumina linker (magenta), a 10-base-pair random label (N10) and an index sequence (green). The combination of label sequence and the 5' mapped position of the RNA fragment forms the UMI. (b,c) Measurements of expression of the same set of genes after 15 (x axes) and 25 (y axes) PCR amplification cycles were obtained using total read counts (b) or the UMI method (c). Most individual transcript total read counts obtained in the two measurements are far from each other and from diagonal (dashed gray line), and this effect is corrected by the UMI method (c). Genes whose mean in the two measurements deviated more than 5% from the fitted line (gray, solid) are in red. (d) A density plot shows average copy number of UMIs after 15 PCR cycles as a function of the average G+C content of the fragments for each measured gene from b and c. Red line in d indicates a least-squares fit, for which a *P* value and adjusted *R*² value are given.

mRNA molecules. We fragmented *Drosophila melanogaster* S2 cell RNA, converted it to cDNA using oligo(dT)-primed reverse transcription and incorporated an oligonucleotide with a 10-base-pair random label by template switching (Fig. 3). We amplified the resulting cDNA fragments directly by PCR and sequenced them. In this method, only one fragment is derived from each mRNA. The sequence of the label and the 5' mapped position of the fragment together define the UMI (Supplementary Fig. 3).

Counting the total reads after 15 or 25 PCR cycles revealed an amplification bias that resulted in loss of accuracy, with 418 of the 5,097 measured genes differing more than 5% between the samples (Fig. 3b). Using UMIs to estimate the number of cDNAs in the original sample resulted in much higher correlation between the samples ($R^2 = 0.99993$), and the number of genes differing by 5% or more was only ten (Fig. 3c). Whereas robust measurement of gene expression by read counting requires normalization¹⁴ that renders all measurements dependent on each other, with the UMI method one can reproducibly detect the number of molecules after different numbers of PCR cycles without normalization. Analysis of the average copy number of UMIs mapping to each gene revealed a clear bias in G+C content in the raw read counts (Fig. 3d), presumably owing to preferential amplification of sequences with low G+C content¹⁵. However, the G+C content explained only a small fraction of the variance, indicating that a simple correction cannot be used to substantially improve the accuracy of the total read counting method.

Analysis of replicate samples revealed that the precision of the total read counting and UMI methods were similar (Supplementary Fig. 4). This was at least partly due to a reproducible bias in the total read counting method (Supplementary Fig. 5). Read counting biases introduced by PCR (Fig. 3d) or *in silico* (Supplementary Fig. 6) could be identified using the UMI method. Furthermore, when we used small amounts of input RNA, with the UMI method we could infer the relative sizes of the different samples. Also, estimates based on the UMI method were better correlated with smaller coefficients of variation (CV) between different aliquots than the total read counts (Supplementary Fig. 7).

The UMI method is compatible with sample indexing using separate DNA barcodes, allowing parallel analysis of samples. In addition to the applications described above, the UMI method could be used to monitor mixing of complex solutions and to trace flow patterns. In principle, the method can be applied to count all types of molecules or particles such as proteins or viruses that can be stoichiometrically labeled with DNA and subsequently purified from free label.

In contrast to previous approaches, the UMI method also can be used to accurately estimate the number of molecules without actually observing all of them. Use of overlapping and consecutive fragments can extend the method to fragments that were lost during sample preparation (Supplementary Fig. 2). Furthermore, nonrandom UMI labels can provide information about relationships or interactions between molecules. For example, UMIs can contain information about fragments that were consecutive in the original molecule ('junction labels') or that were linked together as one macromolecular complex ('correlative labels'). Junction

labels could be introduced by inserting a DNA label using viral integrase or recombinase; this method is likely to have utility in shotgun genome assembly of repetitive sequences. Correlative labels, in turn, can be introduced by a 'split-and-pool' method, whereby the sample is split into a large number of wells, labeled with different labels and then pooled. Fragments from a macromolecular complex are more likely to contain the same labels than unconnected fragments. Correlative labels can be used to analyze chromatin structure.

The UMI method and its variations are thus likely to improve a large number of next-generation sequencing-based molecule-counting applications and also enable new methods for tracking relationships between molecules.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Accession codes. ArrayExpress E-MTAB-816 and European Nucleotide Archive ERA063165 (sequences derived from RNA from the S2 cell line).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank M. Taipale, H. Secher Lindroos, E. Ukkonen and T. Whittington for critical review of the manuscript, and E. Iwarsson (Karolinska University Hospital) for the trisomy-21 DNA. This work was supported by European Research Council project Growth Control, Academy of Finland postdoctoral researcher's projects 122197 and 134073, and the Swedish Foundation for Strategic Research grant MDB09-0052.

AUTHOR CONTRIBUTIONS

S.L., J.T., A.V., T.K. and M.E. conceived and designed experiments. A.V., K.K. and M.B. performed biological experiments. S.L., J.T., A.V. and T.K. analyzed data. J.T., A.V., T.K., M.E. and S.L. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Ozsolak, F. *et al. Nat. Methods* **7**, 619–621 (2010).
- Lipson, D. *et al. Nat. Biotechnol.* **27**, 652–658 (2009).
- Ozsolak, F. *et al. Nature* **461**, 814–818 (2009).
- Vogelstein, B. & Kinzler, K.W. *Proc. Natl. Acad. Sci. USA* **96**, 9236–9241 (1999).
- Macevicz, S.C. US patent application 11/125,043 (2005).
- Hug, H. & Schuler, R. *J. Theor. Biol.* **221**, 615–624 (2003).
- König, J. *et al. Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
- Wang, Z. *et al. PLoS Biol.* **8**, e1000530 (2010).
- Fu, G.K., Hu, J., Wang, P.H. & Fodor, S.P. *Proc. Natl. Acad. Sci. USA* **108**, 9026–9031 (2011).
- Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
- Casbon, J.A., Osborne, R.J., Brenner, S. & Lichtenstein, C.P. *Nucleic Acids Res.* **39**, e81 (2011).
- Chiu, R.W. *et al. Proc. Natl. Acad. Sci. USA* **105**, 20458–20463 (2008).
- Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L. & Quake, S.R. *Proc. Natl. Acad. Sci. USA* **105**, 16266–16271 (2008).
- Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
- Benita, Y., Oosting, R.S., Lok, M.C., Wise, M.J. & Humphrey-Smith, I. *Nucleic Acids Res.* **31**, e99 (2003).

ONLINE METHODS

Digital karyotyping. Genomic DNA was obtained by informed consent from three individuals, a boy with diagnosed trisomy 21, his mother and an unrelated adult male (here referred to as ‘adult male’ sample). This study was conducted with the approval of the Regional Ethical Review Board in Stockholm (Regionala etikprövningsnämnden i Stockholm). The samples from the boy and his mother were mixed 1:1 (‘mixed’ sample). Samples were prepared by enzymatic fragmentation, adaptor ligation and PCR as previously described¹⁶, except that the mixed sample was aliquoted before PCR, aiming to obtain ~20 million molecules, and was ligated with a mixture of eight adapters carrying distinct 6-base-pair (bp) barcodes. Thus, before amplification, the adult male sample was expected to contain billions of molecules, whereas the mixed sample was expected to contain 20 million molecules (the actual number of UMIs was 1.28 million; we attribute the difference to losses in sample preparation). Sequences were generated on an Illumina Genome Analyzer with 76-bp single reads for the mixed sample and 100-bp paired-end reads for the adult male sample. Reads were mapped to the genome using Bowtie¹⁷.

We analyzed the genome in non-overlapping 5-megabase-pair windows. To obtain a reliable estimate of the effective size of each window, accounting for repeats and other un-mappable sequences, we generated a simulated dataset with 34 million reads and mapped this to the genome. The number of hits per window was taken as the effective size of that window, and windows having more than 10% repeats were discarded; this eliminated all of the chromosome Y data.

To ensure that sequencing depth did not limit accuracy of the total read count method, we performed the same analysis on an adult male genome sequenced to 279 million reads. The CV (determined from the 5-megabase intervals) decreased only slightly (from 7.8% to 7.5%), showing that standard read counting does not converge on the true copy number.

For absolute molecule counting, we used the 5′ genomic position of each mapped read as the UMI. Among the 20 million reads, we observed 1.28 million UMIs. The library was intentionally made from a small aliquot and sequenced until all molecules were observed multiple times to improve precision and accuracy of the UMI method. Expected copy numbers per genome for chromosomes 21 and X were 1.25 and 0.75, respectively, as the sample should contain five copies of chromosome 21, three copies of chromosome X and four copies of all other chromosomes. We observed 1.26 copies of chromosome 21 and 0.75 copies of chromosome X. To verify that UMIs did in fact identify single molecules, we searched for instances in which copies of a UMI (that is, multiple reads aligned to the same position) had different barcodes. We found only 25 such instances in the entire genome, demonstrating that UMIs indeed represented single molecules.

To determine the best accuracy theoretically obtainable with 1.28 million UMIs, we generated a simulated sample with this number of molecules and analyzed it along with the real samples. The CV for the UMI method was 3.0%, close to the theoretically maximal accuracy of 2.2% obtained by uniform random sampling of 1.28 million molecules.

mRNA-seq. Application of the UMI method to mRNA sequencing requires the use of labels. This is because different mRNA

species are present in very different concentrations in cells. Taking a small aliquot can result in loss of low-abundance species. Fragmentation of high-abundance species, in turn, can result in generation of multiple fragments with the same start and/or end positions.

For mRNA-seq, total RNA from *Drosophila* S2 cells transfected with GFP dsRNA was hydrolyzed via a 3-min incubation at 70 °C in 1× RNA fragmentation buffer (Ambion). The reaction was terminated as instructed by the manufacturer. cDNA synthesis was performed according to the SMART protocol¹⁸ with addition of adapters for massively parallel sequencing^{19,20} using an oligo(dT)-containing adaptor (see **Supplementary Table 1** for sequence; Eurofins MWG Operon). For absolute molecule counting, a random 10-base DNA sequence label (N) was added to the 5′ adaptor (see **Supplementary Table 1** for sequence; Integrated DNA Technologies). To denature RNA and anneal the 3′ adaptor, 12 pmol of both adapters were incubated at 72 °C for 2 min with 3 µl of the unpurified solution containing 50 ng of fragmented total RNA. RNA was then reverse-transcribed with 200 U SuperScriptIII (Invitrogen) in a 15 µl volume with the provided buffer, 1 mM dNTPs, 2 mM dithiothreitol (DTT) and excess MgCl₂ (added to 15 mM). The reaction was carried out at 55 °C for 1 h and enzyme was inactivated by incubation at 70 °C for 15 min. Uracil-specific excision reagent was used to degrade the random label sequence in the template-switch oligonucleotide (5 U of USER per 50 ng of total RNA at 37 °C for 30 min; New England Biolabs).

The libraries were amplified using Phusion High-Fidelity DNA polymerase (Finnzymes) from 2 µl of unpurified cDNA reaction mixture with 300 nM Illumina single-read sequencing library primers. PCR was performed according to manufacturers’ instructions. In the 50 µl reactions 20% trehalose was included and the following cycle settings were used: denaturation, 1 min at 98 °C, followed by 15 to 25 cycles of 10 s at 98 °C, 30 s at 64 °C and 1 min at 72 °C. Final extension was 11 min. In the PCR-cycle experiment, half of the reaction volume was extracted at cycle 15 and replaced with fresh master mix. PCR products were purified with one volume of Agencourt XP beads (Beckman) and subjected to Illumina GAIIx massively parallel sequencing according to manufacturer’s instructions (54-bp reads).

mRNA-seq with low input. Samples with RNA amounts corresponding to 10, 100 or 1,000 *Drosophila* S2 cells (**Supplementary Fig. 7**) were prepared with about 0.07, 0.7 or 7 ng of fragmented RNA from the same total RNA from GFP dsRNA-transfected cells used in the mRNA-seq experiments (**Fig. 3**). RNA was converted to cDNA using a paired-end-compatible oligo(dT) adaptor (12 pmol; see **Supplementary Table 1** for sequence; Integrated DNA Technologies), the labeled 5′ adaptor described in the section above (12 pmol) and 300 U SuperScriptIII in 30 µl volume with the provided buffer, 0.8 mM dNTPs, 2.9 mM DTT and excess MgCl₂ (added to 17 mM). The cDNA was treated with 60 U of exonuclease I (New England Biolabs) before uracil-specific excision reagent. The whole volume of treated cDNA was included in a 100 µl PCR; samples were amplified as described above and sequenced (Illumina HiSeq 2000 instrument, 54-nucleotide reads from both ends).

Estimation of unobserved molecules. The principle of the UMI method is that the original number of molecules in a sample

can be estimated as the sum of observed and unobserved UMIs. The number of unobserved UMIs can be estimated based on the distribution of the copy numbers of the observed UMIs. For example, if one observes UMIs on average ten times (average copy number = 10), it is likely that very few UMIs have been missed. However, if the average copy number is two, a substantial fraction of all UMIs have not yet been observed. In general, we assumed that all of the UMIs of a gene had an equal probability of being observed. Thus, the number of molecules from each gene was estimated by fitting a zero-truncated Poisson distribution to the UMI copy number distribution using the generalized additive models for location, scale and shape (GAMLSS) R package²¹ and adding the predicted number of unobserved UMIs to the observed UMI count.

mRNA-seq data analysis. The sequencing reads were analyzed as follows: after removal of the label and index sequences and the following two bases, the sequencing reads were mapped to reference sequences from Ensembl version 52 using Burrows-Wheeler aligner (BWA) software version 0.5.8 with default parameter values²². Two bases were removed from the 5' end of the reads after the index and label sequences to exclude additional guanines occasionally added by the template-switch method.

For each gene, the sequence of its longest transcript was used as the reference sequence. Reads were discarded from further analysis if they did not contain the constant sequences expected based on oligonucleotide design, mapped to the wrong strand, had a BWA mapping quality score lower than 20 or a base in the label sequence with an Illumina base call quality score lower than 20. A total of 14.8 million reads and 23.9 million reads passed these criteria in *Drosophila* S2 cell samples taken after 15 and 25 PCR amplification cycles, respectively. Total read count of a gene is the number of accepted reads mapped to its reference sequence.

The mapped reads with the same gene, position and label were collected to one UMI and the number of such reads was recorded as the copy number of that UMI. Average copy numbers were 10.7 and 17.0 for samples taken after 15 and 25 PCR cycles, respectively. Sequence errors introduced by library preparation, amplification and sequencing can produce false UMIs with a low copy number. To limit the effect of such errors, two UMIs were merged if they either had identical positions and one mismatch in the label sequences (probable substitution) or consecutive positions, identical label sequences and the UMI closer to the 3' end of the mRNA had a copy number of one and the UMI closer to 5' end had at least a copy number of two (probable deletion).

In addition, all UMIs from positions where UMI average mapping quality was less than 30 were discarded.

The approximately one million random labels used were sufficient to generate UMIs from an mRNA amount that corresponds to the amount found in ~1000 *Drosophila* S2 cells. On average, only 0.0005% of all labels were observed per position, and even in the position with the highest number of labels, less than 2% of over one million labels were observed. In addition, a large fraction of all labels (>70%) were observed, and less than 0.12% overlap of UMIs (position label pairs) was observed between replicate experiments, indicating that the UMIs were not exhausted in the experiments and that the labels were incorporated into cDNAs effectively randomly, independent of position- or label-specific factors. Moreover, the incorporation of the random label sequence did not interfere with the mRNA-seq process; similar counts of reads mapping to each gene were observed in labeled and unlabeled samples (data not shown).

The expression level of a gene was considered to be measured if its total read count was at least 100 and the estimate of the number of molecules was at least 10, and at least one of the UMIs had two or more copies. These cutoffs correspond to approximately 0.2–1 mRNA molecules per cell based on yield estimates from RNA quantification of total RNA and spike controls (data not shown).

The G+C content of the sequenced gene fragments was calculated as the average G+C content of the subsequences from the position of the mapped read to the 3' end of the reference sequence. The solid gray line in **Figure 3b** indicating the size factor (1.79) needed to render the total read counts from different samples comparable was fitted analogously to the method described in reference 14. First, the mean relative difference between the samples was calculated from the Equation $d = (1/N) \sum_{i=1}^N (x_i - y_i) / (x_i + y_i)$, where x_i and y_i are the individual total read counts for transcript i , and N is the total number of transcripts. Then, the size factor s was calculated from the Equation $s = (1 - d) / (1 + d)$, and its logarithm used as the intercept for the solid fit line shown. The corresponding size factor for the absolute molecule counts was 1.02.

16. Linnarsson, S. *Exp. Cell Res.* **316**, 1339–1343 (2010).

17. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).

18. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. *Biotechniques* **30**, 892–897 (2001).

19. Cloonan, N. et al. *Nat. Methods* **5**, 613–619 (2008).

20. Levin, J.Z. et al. *Nat. Methods* **7**, 709–715 (2010).

21. Stasinopoulos, D.M. & Rigby, R.A. *J. Stat. Softw.* **23**, 1–46 (2007).

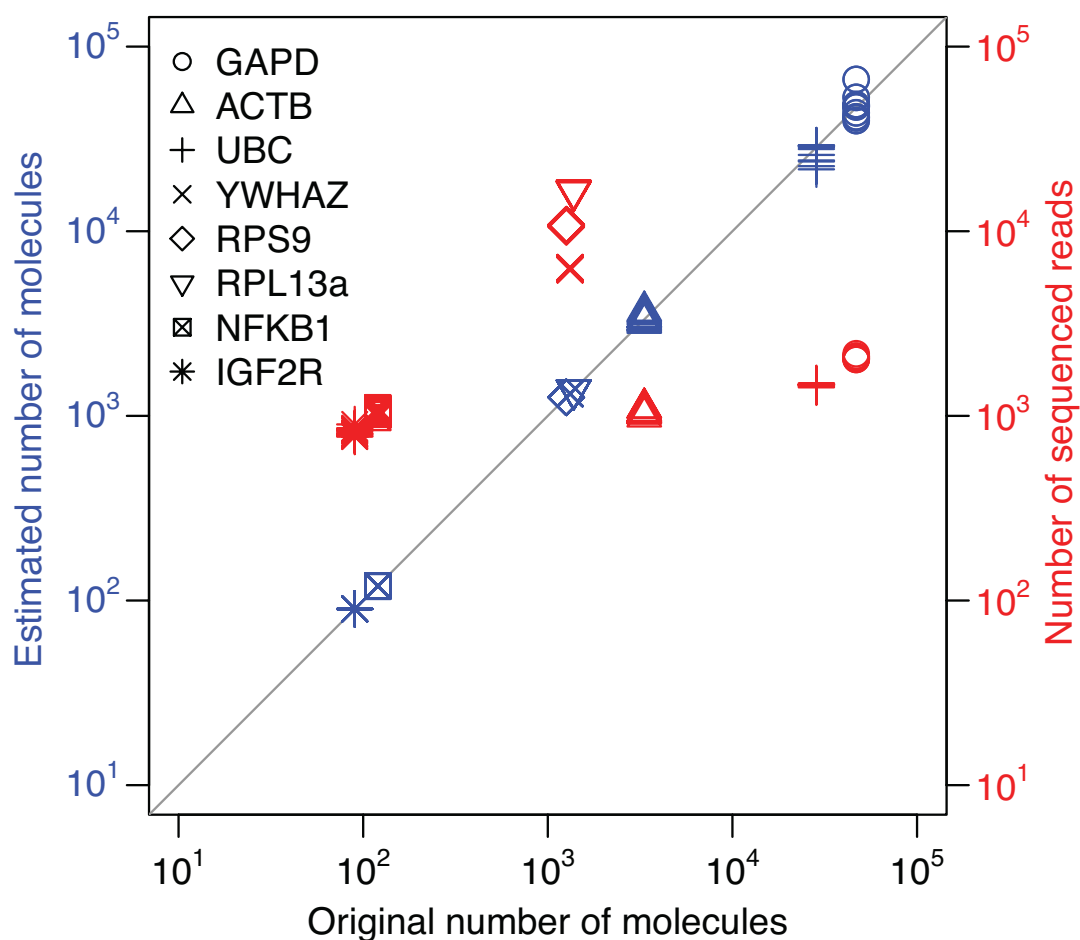
22. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).

Supplementary information

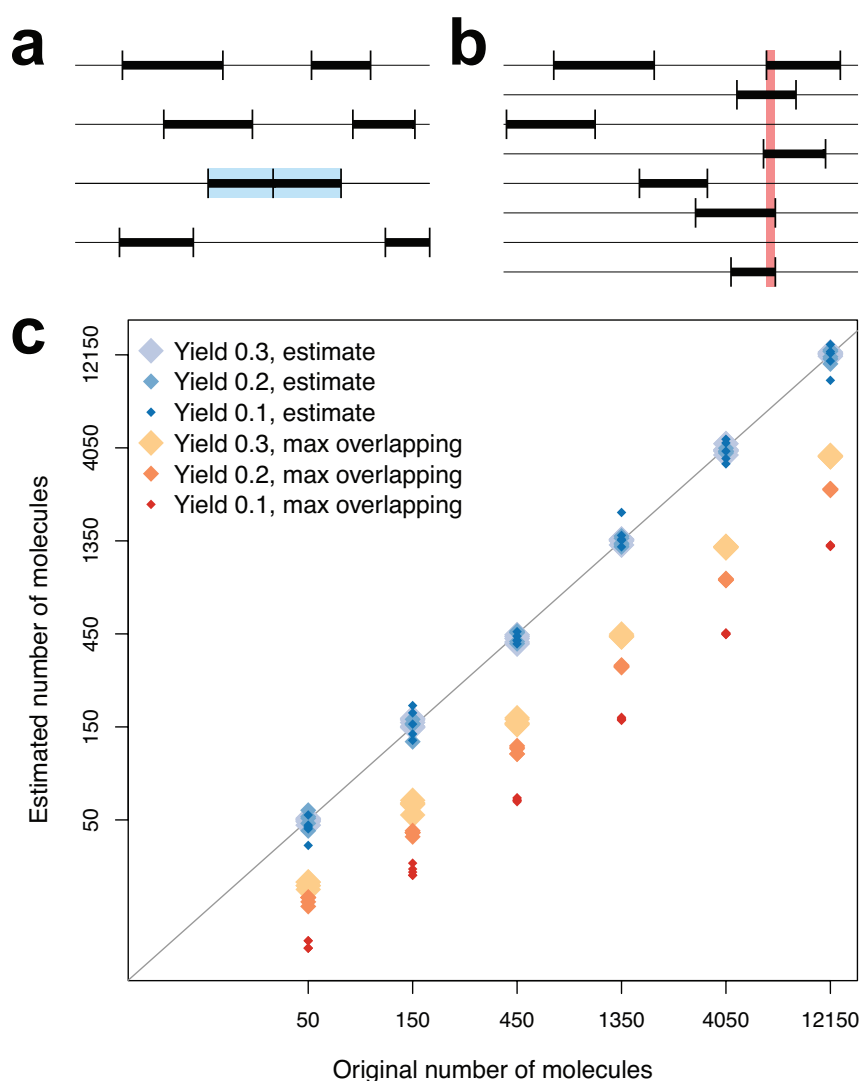
Counting absolute number of molecules using unique molecular identifiers

Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson and Jussi Taipale

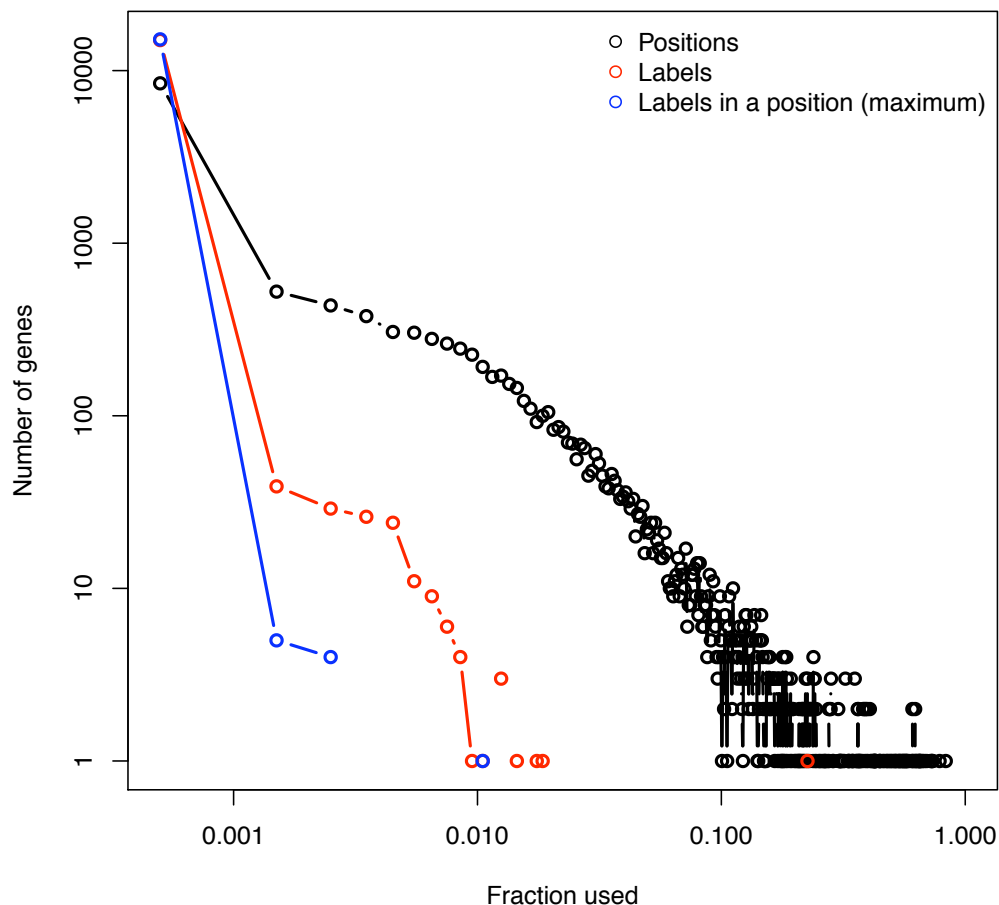
Supplementary Figure 1	Normalized cDNA library simulation
Supplementary Figure 2	Estimating the absolute number of molecules before library preparation
Supplementary Figure 3	The fraction of used positions and labels
Supplementary Figure 4	Reproducibility of the read and umi counting methods
Supplementary Figure 5	Copy number correlation in technical replicates
Supplementary Figure 6	Estimating the number of molecules after removing part of the reads <i>in silico</i>
Supplementary Figure 7	Measuring mRNA abundances from small amount of RNA
Supplementary Table 1	Sequences for oligomers used in mRNA-seq
Supplementary Note	Normalized cDNA library simulation Modeling absolute number of molecules before library preparation Simulation of absolute number of molecules before library preparation



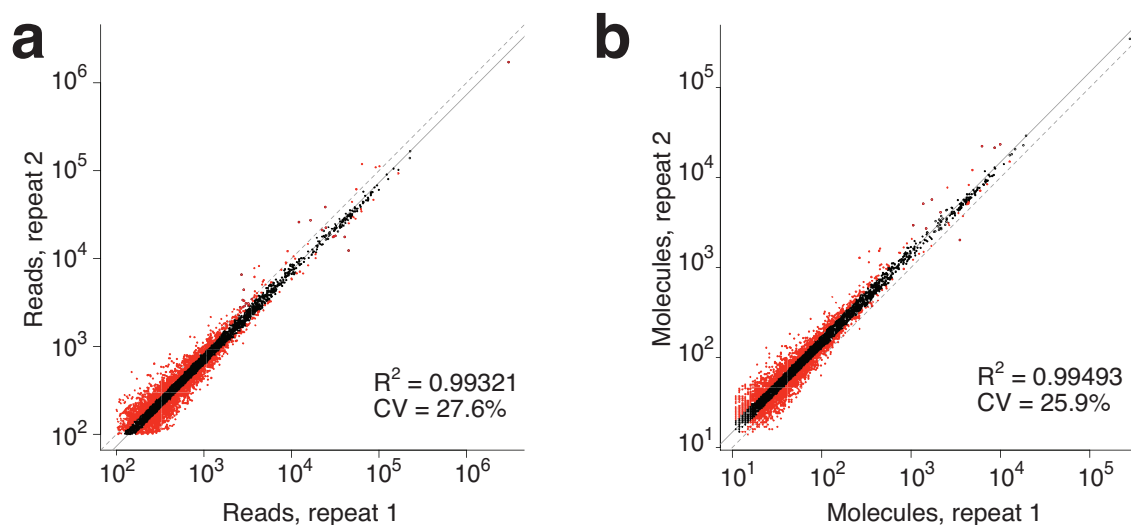
Supplementary Figure 1. Normalized cDNA library simulation. Simulation of an experiment where labels are used to estimate the original number of mRNA species after normalization of a cDNA library. Ten simulations were performed and the original number of cDNA molecules prior to normalization (x-axis) was estimated (y-axis, blue symbols; see **Supplementary Note** for details). The raw number of observed cDNA sequences for each gene is also shown (red symbols). Note that accurate estimates (blue symbols) can be derived even when the normalization decreases high abundance cDNA (GAPD, red circles) to a level that is lower than that of medium-abundance cDNA (RPS9, red diamond).



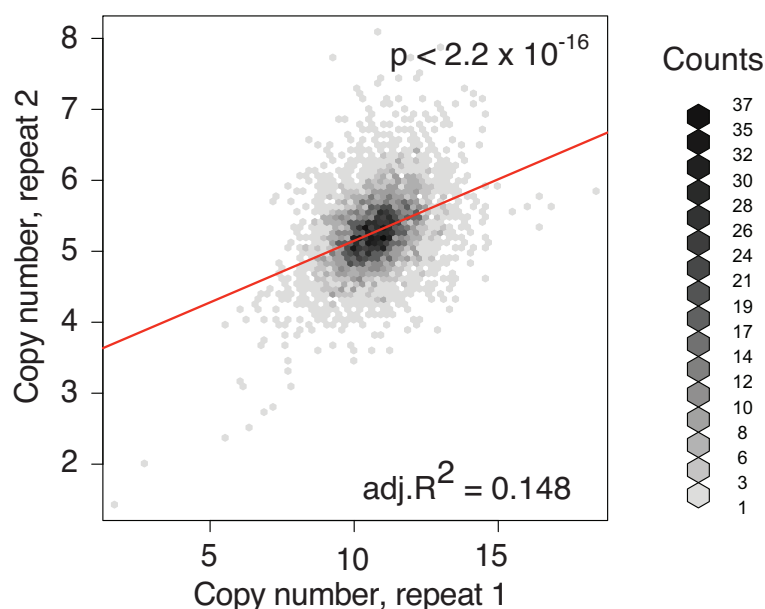
Supplementary Figure 2. Estimating the absolute number of molecules before library preparation. **(a-b)** Schematic representation of two distinctive fragmentation patterns. Complete molecules are thin black lines, observed fragments thick black bars. In both **(a)** and **(b)** the library contains eight fragments. **(a)** The number of original complete molecules was four. **(b)** The number of original molecules was eight but the yield was half of that in **(a)**. Blue in **(a)** highlights consecutive fragments that are frequent when yield is high and red in **(b)** highlights five overlapping fragments which is not possible when the number of molecules is less than five **(a)**. **(c)** Simulation of estimating absolute number of molecules before library preparation. Experiments with different numbers of ten thousand bases long original molecules (x-axis) and three different yields (10%, 20%, and 30%) were simulated (see **Supplementary Note** for details). The molecules were randomly cut into fragments of size 100 bp on average and those of at least size 30 bp were assumed to be mappable. Each combination of number of molecules (x-axis) and yield was simulated five times and the original number of molecules was estimated (y-axis, estimate, blue diamonds) using a probabilistic model (see **Supplementary Note**, note logarithmic scale). For comparison, the maximum number of overlapping fragments (max overlapping, orange and red diamonds) is also shown for each simulated experiment. Note that two simulated experiments in which the first one has yield 30% and the second only 10% but three times more molecules (e.g. 150 vs 450 molecules) have approximately the same number of observed fragments and maximum number of overlapping fragments. Yet, the probabilistic model is able to separate these cases.



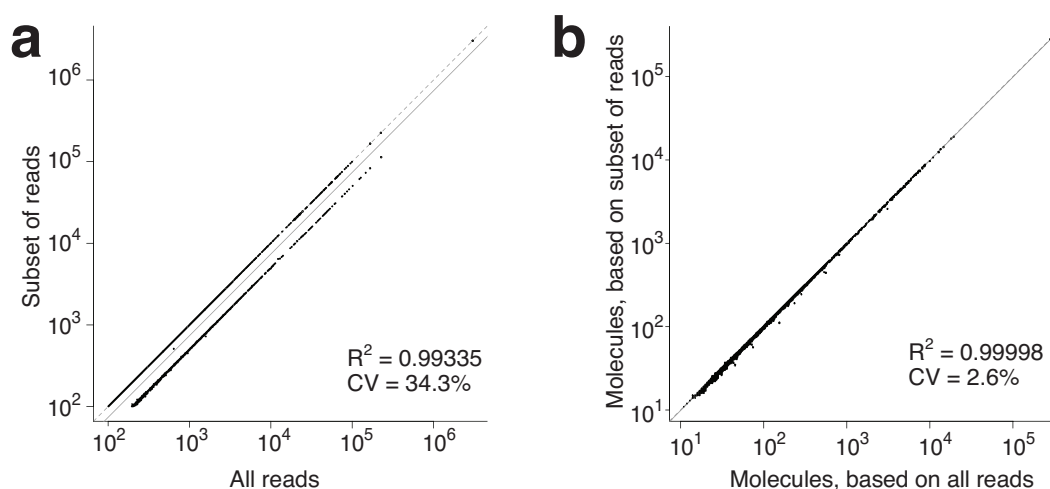
Supplementary Figure 3. The fraction of used positions and labels. The numbers of different types of umis were counted for each gene in the mRNA sample sequenced after 15 PCR amplification cycles and the counts were compared to the total number of umis available. Three different types of potential umis were counted: positions (black), labels (red), and their combinations (blue; the umi used in the mRNA-seq data analysis). In the last case, the maximum number of labels having the same position was recorded to see if counts in any individual position were saturated. Each of these three umi counts was then divided by the theoretical maximum number of umis of that type. Position count was divided by the length of the molecule. The label count and the maximum count of labels in one position were divided by the total number of labels. Each bin of the histogram shows the number of measured genes (y-axis) with a given fraction of used umis (x-axis). There are one thousand bins and both axes are in logarithmic scale. Empty bins are not shown. Note that the positions were not uniformly distributed and thus they were in practice saturated well before reaching one.



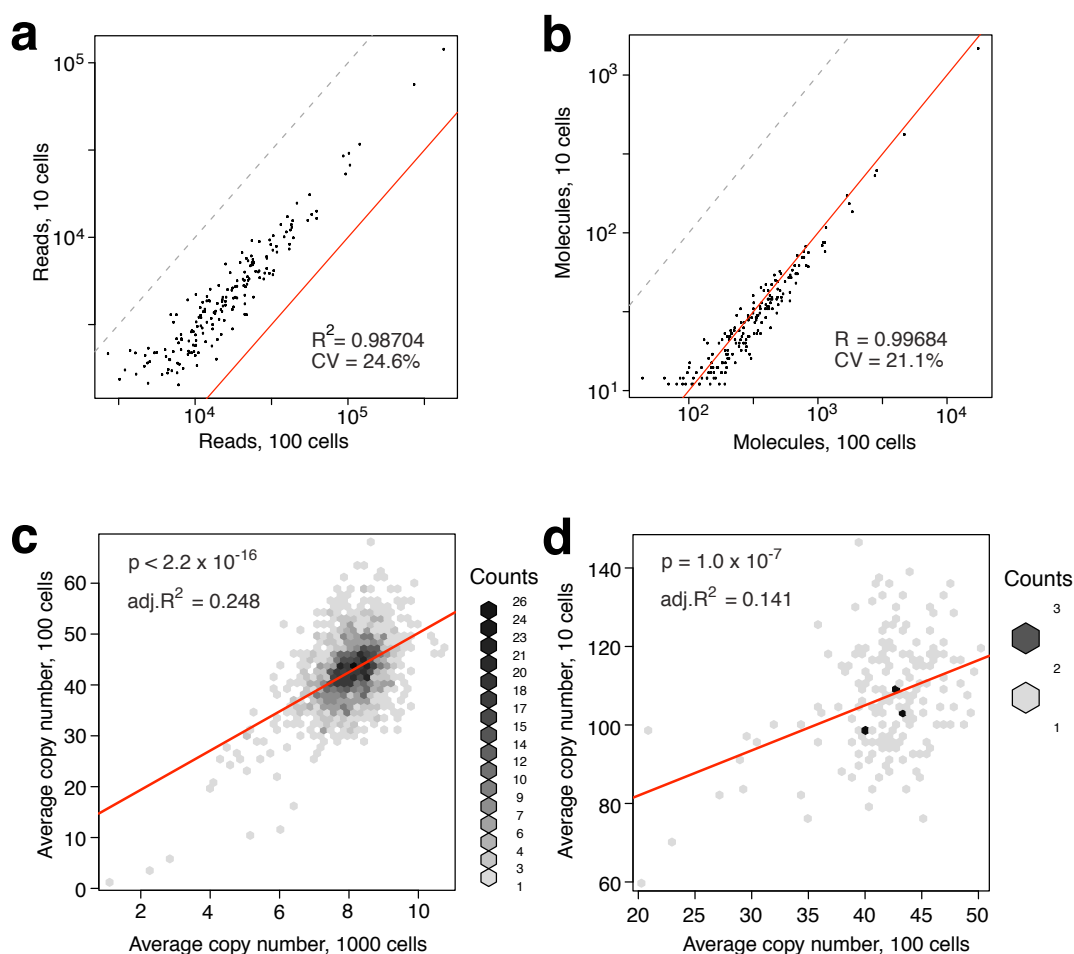
Supplementary Figure 4. Reproducibility of the read and umi counting methods. Measurements of expression levels of the same set of genes after 15 PCR amplification cycles in technical cDNA synthesis replicates (data used in **Fig. 3** on the x axis and a technical replicate on the y axis) obtained using read counts (**a**) or absolute molecule counts (**b**). Dashed grey line indicates the diagonal and solid grey line indicates a fitted line (see Methods for details). The effect of the umi counting method on reproducibility was small, primarily because under the conditions tested, precision of both methods is limited by sampling error (low number of molecules) and variation in the library generation steps. Genes whose mean in the two measurements deviates more than 5% from the fitted line (gray, solid; see Methods for details) are in red. A coefficient of variation (CV) and adjusted R^2 value are also indicated.



Supplementary Figure 5. Copy number correlation in technical replicates. Density plot showing the average copy numbers of umis of the same set of genes after 15 PCR amplification cycles in technical cDNA synthesis replicates (data used in **Figure 3** on the x axis and a technical replicate on the y axis). Bias in the read counting method appeared to be at least in part reproducible, as indicated by correlation between the average copy numbers of the umis for each gene. Red line indicates a least squares fit, for which a *P*-value and adjusted R^2 value are also given.



Supplementary Figure 6. Estimating the number of molecules after removing part of the reads *in silico*. A random half of the genes was chosen from the mRNA sample sequenced after 15 PCR amplification cycles. For each of the chosen genes, each sequencing read was either kept or removed with probability 0.5. **(a)** The original number of reads for each gene (x-axis) compared to the number of reads after removing part of the data (y-axis). **(b)** The number of molecules estimated from all reads (x-axis) compared to number of molecules estimated from the reduced set of reads (y-axis). Dashed grey line indicates the diagonal and solid grey line indicates a fitted line (see Methods for details).



Supplementary Figure 7. Measuring mRNA abundances from small amount of RNA. Expression levels were measured from samples corresponding to RNA from approximately 1000, 100, and 10 *Drosophila* S2 cells (same *GFP*-transfected total RNA sample as in **Figure 3**). **(a)** The number of reads from 10 cells (y-axis) compared to number of reads from 100 cells (x-axis). **(b)** The estimated number of molecules using the umi method from 10 cells (y-axis) compared to estimated number of molecules from 100 cells (x-axis). Dashed grey line indicates diagonal and red solid line indicates the expected ratio of 0.1. **(c, d)** Copy number correlation between samples. Density plots showing the average copy numbers of umis of the same set of genes for 1000 and 100 **(c)** and 100 and 10 cells **(d)**, respectively. Red line indicates a least squares fit, for which a *P*-value and adjusted R^2 value are also given. Expression levels were estimated by fitting to truncated Poisson distribution as described in Methods except that counts smaller than 10% of the mode of the observed count distribution (excluding counts < 4) were not used in the fitting to avoid fitting to background noise. The smallest used count was 5 and 15 for 100 and 10 cells, respectively. Only relatively few genes are measured from the smallest sample because we consider a gene measured only if the estimate is at least 10 molecules (see Methods).

Supplementary Table 1. Sequences for oligomers used in mRNA-seq. Sequences in square brackets and parentheses represent an index sequence that was used to enable multiplexed sequencing, and a random ten base DNA sequence label, respectively. dU and rG represent deoxyuridine and guanine ribonucleotide, respectively.

Oligomer	Sequence (5'-3')
Oligo-dT adapter	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTVN
5'-adapter	ACACTCTTCCCTACACGACGCTCTTCCGATCT(NNN)dU(NNNN)dU(NNN)[GACTT]rGrGrGrG
Paired-end compatible oligo-dT adapter	CTCGGCATTCTGCTGAACCGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTVN

SUPPLEMENTARY NOTE

Normalized cDNA library simulation

Measuring the abundances of multiple different molecular species in a complex mixture is difficult, because individual measurements can interfere with each other, and different species can be present in wildly different concentrations. For example, differences in concentration between high-abundance mRNA and low abundance mRNA in a cell or tissue sample can range over six to ten orders of magnitude (see for example Ref. [1]). This severely limits the specificity of DNA microarrays [2, 3], as DNA or RNA molecules with different but similar sequences can hybridize to the same target sequence [4, 5]. The wide range of concentrations also makes counting molecules by massively parallel sequencing [6, 7, 8, 9] challenging. Hundreds of thousands of sequencing reads matching abundant RNA species need to be counted before even a single read mapping to a rare species is found. This results in very large differences in measurement precision between high and low abundance species.

The range of concentrations can be compressed by subtractive hybridization (i.e. library normalization, Refs. [10, 11]), but this destroys information about the original levels, decreasing accuracy of measurements. Amplifying weak signals by making exact (digital) copies of the molecules [12, 13] similarly leads to decreased accuracy. As exact copies made from identical original molecules are indistinguishable, determining the original number of molecules after copying requires knowledge of the number of copies made. This is very difficult to determine because all known molecular copying processes are stochastic and are affected by DNA sequence, length and experimental conditions (see for example Ref. [14]). In addition, the cumulative nature of the error during copying makes it even more difficult to accurately measure very small numbers of molecules, for example levels of mRNAs in a single cell [15].

As the the umi method encodes the concentration information in the number of distinct sequences, it in principle permits very extensive amplification and normalization of the samples without loss of quantitative information. This should dramatically improve quantitative analysis of molecules that are present in small amounts, either because of their low fractional abundance (e.g. mRNA gene expressed at low level) or due to small size of the analyzed sample (e.g. single cell).

We used simulation to test whether the umi method can recover the absolute original numbers of molecules accross a wide dynamic range after library normalization or other extensive processing. The cDNA normalization simulations correspond to sequencing approximately 20 million reads from a genome-wide cDNA library. Ten simulations were performed for a total of 82,830 molecules representing eight different cDNA species (frequencies obtained in an actual cDNA normalization experiment, Ref. [11]). Each molecule was given a random 10 bp label (1,048,576 labels), and their frequencies adjusted to correspond to those observed in the amplified and normalized library of Ref. [11]. Next a random sample (with replacement) of 40,000 molecules was taken from the pool. The original number of cDNA molecules prior to normalization was estimated from label count distribution as in RNA-seq data analysis (see Methods).

Modeling absolute number of molecules before library preparation

The umi approach can be extended to estimate of the number of molecules present in a sample before library preparation. In the following, it is assumed that labeled fragments of DNA or RNA molecules are used as umis and thus copies derived from the same fragment can be identified. The fragmentation pattern seen in the library enables uncoupling the original number of molecules from the yield of the library preparation. Consecutive fragments are likely to be derived from the same original full-length molecule whereas several overlapping fragments cannot be derived from a single molecule. For example, the observed number of overlapping fragments is always smaller or equal to the number of full-length molecules which gives a lower bound for the number of complete molecules in the sample before the library preparation.

More formally, we assume that both ends of a set of DNA/RNA fragments from a sample have been observed (for example by paired-end sequencing). We describe a mapped fragment f that starts at s_f (the leftmost mapping position) and ends at e_f (the rightmost mapped position) as the closed interval $I_f = [s_f, e_f]$. We say that two fragments f and g are *consecutive* if that $e_f = s_g - 1$ (where s_g is the starting position of g) and *overlapping* if $I_f \cap I_g \neq \emptyset$.

Let

$$F = nk p_o = nk y s p_m \quad (1)$$

be the total number of observed fragments derived from one molecular species where n is the original number of molecules, k is the number of fragments per molecule, and p_o is the overall probability of a fragment being observed. The probability p_o can be described as a product $y s p_m$ where y is the yield of the library preparation, s is the probability that a fragment in the library is sequenced, and p_m is the probability that a fragment can be mapped to a reference sequence (we only consider here that it is long enough for reliable alignment but other mappability information can be added). The sequencing probability s can be estimated from the umi copy number distribution of individual fragments, and thus for the sake of simplicity is assumed here to be one.

We assume that DNA or RNA is fragmented so that the break points are independently distributed with a common uniform distribution with a break probability p_b and that fragments are short compared to the full-length molecules (so that edge effects can be avoided). In addition, we assume that the average length of the fragments originally in the sample a_l can be accurately estimated which also gives an estimate of the number of fragments per molecule k . As fragment lengths have a geometric distribution and the probability of a break between two bases is $p_b = 1/a_l$, the probability p_a of a fragment having at least the smallest mappable length r is $P(l \geq r) = (1 - 1/a_l)^{r-1}$. Thus, given these assumptions and an estimate of the number of molecules $\hat{n}$, all unknowns in Equation 1 including yield can be estimated.

The goal is to derive a model that gives the likelihood of the observed fragments given a number of molecules and then use that to estimate the number molecules by choosing the number that maximizes this likelihood. First note that given the assumptions, the numbers of consecutive fragments derived from one molecule, i.e. how many consecutive fragments are retained before the first missed one, have

geometric distribution. However, it is difficult to directly utilize this fact as we cannot observe which fragment pairs actually originate from the same molecule. Alternatively, we can divide the reference sequence of the molecular species into such windows that a start of at most one fragment from one molecule can be observed in a window i.e. each molecule is observed at most once in one window. Then observing a start of a fragment from a molecule in a given window is a Bernoulli trial with an unknown probability of success. Thus, the numbers of observed fragment starts in the windows have binomial distribution with unknown probability of success and number of trials n . Estimation of the binomial parameters when both of them are unknown is a classic problem that has been studied for over 50 years [16]. It seems to be a fundamentally difficult problem as common techniques such as maximum likelihood lead to very unstable estimators and even the best known estimators perform poorly when the probability of success is small. Also, in our case the information about consecutive fragments would not be utilized at all.

Given the difficulties with the above approaches, we develop a model that gives a likelihood for a given number molecules depending on how well the starting and ending fragments at each position can be explained as originating from same (low number of molecules, high yield) or different molecules (high number of molecules, low yield). Let random variables E_i and S_{i+1} be the number of observed fragments ending and starting at consecutive positions i and $i + 1$, respectively. We define the likelihood of the number molecules n given the observations as

$$L(n \mid \mathbf{E}, \mathbf{S}) = \prod_{i=\lceil a_l \rceil}^{L-\lceil a_l \rceil-1} P(E_i, S_i \mid n) \quad (2)$$

The very first and last positions are ignored to avoid edge effects. The model treats breaks observed in different positions as independent. This is not strictly true despite the random breakage assumption. For example, it is not possible observe ends (or starts) of two fragments originating from the same molecule closer than r bases from each other due to the mapping limitations. We omit these dependencies in the model as their effect seems small with realistic fragment lengths and yields (see below) and as they are difficult to account for since it is possible to observe both an end of a fragment and a start of another from the same molecule within r .

At each position, the model needs to explain the observed fragment starts and ends as originating from the same or different molecules. The smaller the number molecules and the higher the yield, the higher is the probability to observe both the ending and starting fragment originating from a break. Conversely, if the number of molecules is high and the yield is low, the starts and ends are more likely to originate from different molecules. More formally, let C_i be the number such molecules that have a break between i and $i + 1$ and both the ending and the starting fragment are observed. As we cannot know which, if any, of the (ending, starting) fragment pairs observed actually originate from the same molecule, we sum the probability of the observed data over all possible values of C_i given n . Thus, the probability of the observed values $E_i = e$ and $S_i = s$ between positions i and $i + 1$ given n can be written

as

$$P(E_i = e, S_{i+1} = s | n) = \sum_{c=0}^n P(C_i = c, E_i = e, S_{i+1} = s | n). \quad (3)$$

The joint probability of C_i, E_i , and S_{i+1} given n has a multinomial distribution as there are for each of the n molecules independently the following possible outcomes between i and $i + 1$ that affect the observations: A molecule either has

- a break and both the starting and ending fragments are observed, probability $p_b p_o^2$, number of occurrences c
- a break and only the ending fragment is observed, probability $p_b p_o (1 - p_o)$, number of occurrences $e - c$
- a break and only the starting fragment is observed, probability $p_b p_o (1 - p_o)$, number of occurrences $s - c$
- no break or has a break but neither of the fragments is observed, probability $(1 - p_b) + p_b (1 - p_o)^2$, number of occurrences $n - e - s + c$.

Naturally the joint probability in Equation (3) is 0 if $c > e$ or $c > s$. The probability to observe a fragment given fixed n is $p_o = F/nk$ (Equation (1)).

To get an estimate the number of molecules, we calculate the likelihood of Equation (2) for a range of different values of n and choose the one that maximizes the likelihood as the maximum likelihood estimator $\hat{n}$. An appropriate range can be chosen using the maximum number of overlapping fragments as the lower bound of the number of molecules.

Simulation of absolute number of molecules before library preparation

The inputs of the simulation were the number of molecules n and the length of the molecules or their mappable regions, the average (100 bp) and the minimum mappable (30 bp) size of the fragments, and the yield y . The simulator

1. constructs the given number of complete molecules of the given length
2. fragments the molecules randomly according to the uniform break probability determined by the given fragment average size
3. samples the fragments randomly according to the given yield
4. selects fragments that are at least as long as the given minimum length
5. estimates the number of molecules based on the starting end ending positions of the fragments in the final sample and the average size of all fragments.

The estimation part utilizes the maximum observed number of overlapping fragments t as the lower bound of the number of molecules and calculates the likelihood given in Equation (2) for all values between t and $30t$ and chooses as the estimate $\hat{n}$ the one giving the highest likelihood. The simulation ignores the effects of labeling, amplification, sequencing, and mapping (every fragment in the library is sequenced at least once, labels and end points always uniquely identify fragments, and mapping is perfect if a fragment is long enough, and length distribution of fragments can be accurately estimated).

SUPPLEMENTARY REFERENCES

- [1] Holland, M. Transcript abundance in yeast varies over six orders of magnitude. *Journal of Biological Chemistry* **277**(17), 14363–14366 (2002).
- [2] Schena, M., Shalon, D., Davis, R., and Brown, P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235), 467–470 (1995).
- [3] Hoheisel, J. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics* **7**(3), 200–210 (2006).
- [4] Kane, M., Jatkoe, T., Stumpf, C., Lu, J., Thomas, J., and Madore, S. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic acids research* **28**(22), 4552–4557 (2000).
- [5] Koltai, H. and Weingarten-Baror, C. Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic Acids Research* **36**(7), 2395–2405 (2008).
- [6] Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., et al. Genome sequencing in open microfabricated high density picoliter reactors. *Nature* **437**(7057), 376–380 (2005).
- [7] Shendure, J., Porreca, G., Reppas, N., Lin, X., McCutcheon, J., Rosenbaum, A., Wang, M., Zhang, K., Mitra, R., and Church, G. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**(5741), 1728–1732 (2005).
- [8] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature genetics* **2**(3), 173–179 (1992).
- [9] Velculescu, V., Zhang, L., Vogelstein, B., and Kinzler, K. Serial analysis of gene expression. *Science* **270**(5235), 484–487 (1995).
- [10] Patanjali, S., Parimoo, S., and Weissman, S. Construction of a uniform-abundance (normalized) cDNA library. *Proceedings of the National Academy of Sciences* **88**(5), 1943–1947 (1991).
- [11] Zhulidov, P. et al. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research* **32**(3), e37 (2004).
- [12] Cohen, S., Chang, A., Boyer, H., and Helling, R. Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences* **70**(11), 3240 (1973).
- [13] Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H., and Arnheim, N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**(4732), 1350–1354 (1985).
- [14] Benita, Y., Oosting, R., Lok, M., Wise, M., and Humphery-Smith, I. Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic acids research* **31**(16), e99 (2003).
- [15] Ozsolak, F. and Milos, P. RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics* **12**(2), 87–98 (2011).
- [16] DasGupta, A. and Rubin, H. Estimation of binomial parameters when both n, p are unknown. *Journal of statistical planning and inference* **130**(1–2), 391–404 (2005).