# Counting absolute numbers of molecules using unique molecular identifiers

Teemu Kivioja[1–3,5], Anna Vähärautio[1,3,5], Kasper Karlsson[4], Martin Bonke[1], Martin Enge[3], Sten Linnarsson[4] & Jussi Taipale[3]

**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 PCR[4], digital microarray profiling[5] 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 microarrays[9] 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).

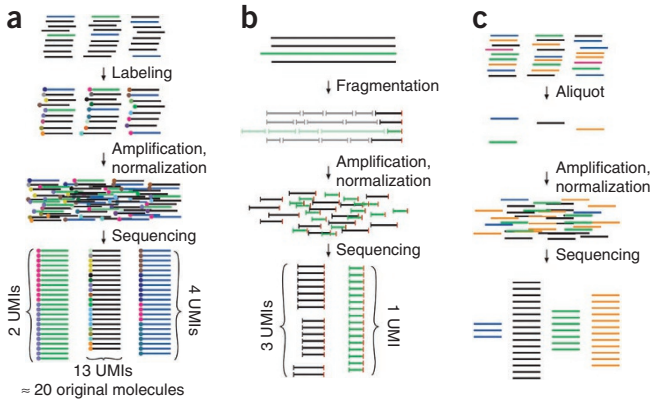


**Figure 1** | UMIs can be generated by adding oligonucleotide labels, fragmenting, taking a small enough aliquot or a combination thereof. (**a**) Three different DNA species (green, blue and black lines) are labeled with a collection of random labels (colored filled circles). Two green molecules are 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.

[1]Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Helsinki, Finland. [2]Department of Computer Science, University of Helsinki, Helsinki, Finland. [3]Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. [4]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [5]These authors contributed equally to this work. Correspondence should be addressed to J.T. (jussi.taipale@ki.se) or S.L. (sten.linnarsson@ki.se).

RECEIVED 10 MARCH; ACCEPTED 27 SEPTEMBER; PUBLISHED ONLINE 20 NOVEMBER 2011; DOI:10.1038/NMETH.1778

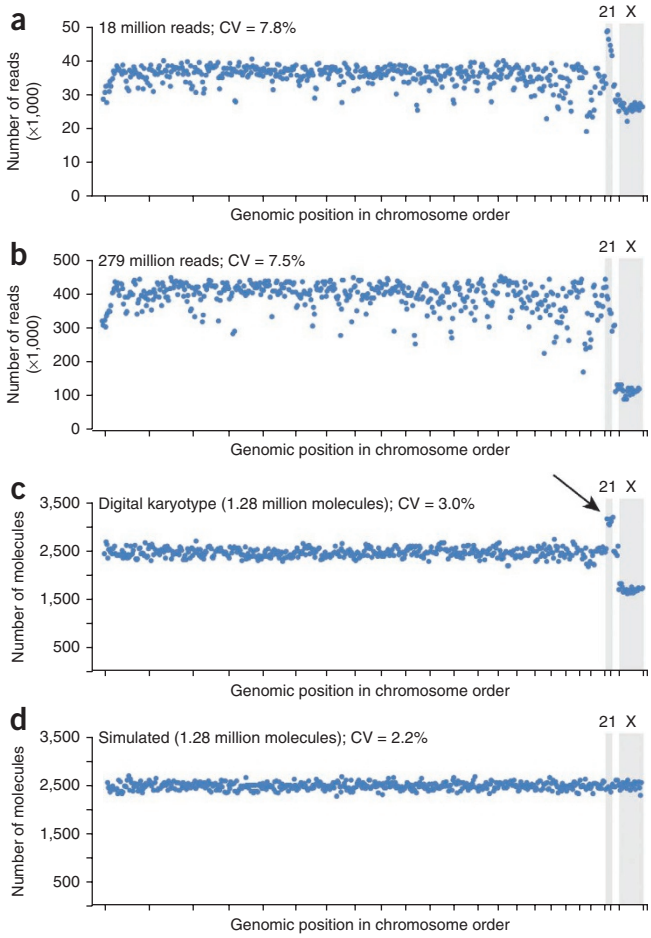


**Figure 2** | Digital karyotyping by counting the absolute number of molecules. (**a**) Standard digital karyotype based on genomic DNA from a boy with trisomy 21 and from his mother, mixed 1:1. (**b**) Standard digital karyotype of a sample from a male with a normal chromosome count. (**c**) The same 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

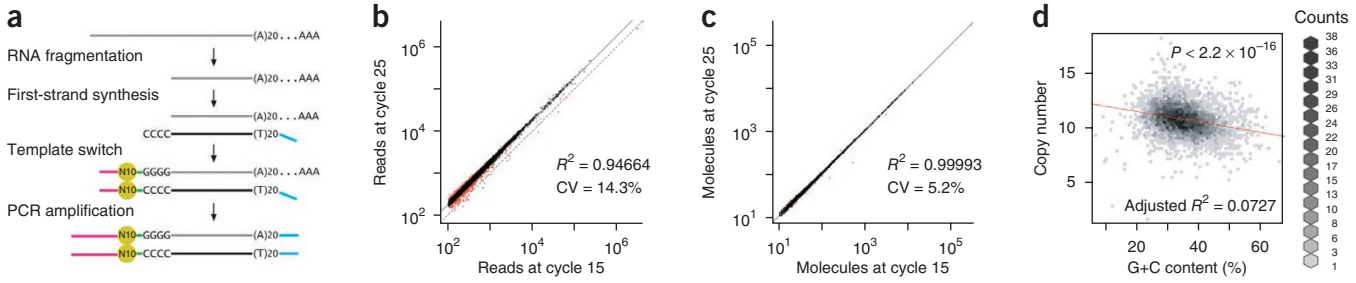


**Figure 3** | Accuracy of mRNA-seq can be improved by the UMI method. (**a**) mRNA-seq libraries were generated by fragmenting total RNA and reverse-transcribing it to cDNA using an oligo(dT) primer with an Illumina linker (blue) and a 5′ template switch adaptor containing another Illumina linker (magenta), 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 normalization[14] 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 content[15]. 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. Whitington 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.**

1. Ozsolak, F. *et al. Nat. Methods* **7**, 619–621 (2010).
2. Lipson, D. *et al. Nat. Biotechnol.* **27**, 652–658 (2009).
3. Ozsolak, F. *et al. Nature* **461**, 814–818 (2009).
4. Vogelstein, B. & Kinzler, K.W. *Proc. Natl. Acad. Sci. USA* **96**, 9236–9241 (1999).
5. Macevicz, S.C. US patent application 11/125,043 (2005).
6. Hug, H. & Schuler, R. *J. Theor. Biol.* **221**, 615–624 (2003).
7. Konig, J. *et al. Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
8. Wang, Z. *et al. PLoS Biol.* **8**, e1000530 (2010).
9. Fu, G.K., Hu, J., Wang, P.H. & Fodor, S.P. *Proc. Natl. Acad. Sci. USA* **108**, 9026–9031 (2011).
10. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
11. Casbon, J.A., Osborne, R.J., Brenner, S. & Lichtenstein, C.P. *Nucleic Acids Res.* **39**, e81 (2011).
12. Chiu, R.W. *et al. Proc. Natl. Acad. Sci. USA* **105**, 20458–20463 (2008).
13. Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L. & Quake, S.R. *Proc. Natl. Acad. Sci. USA* **105**, 16266–16271 (2008).
14. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
15. Benita, Y., Oosting, R.S., Lok, M.C., Wise, M.J. & Humphery-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 described[16], 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 Bowtie[17].

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 unmappable 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 protocol[18] 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 $MgCl_2$ (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 $MgCl_2$ (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 package[21] 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 values[22]. 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 $d = (1 / N) \Sigma_{i = [1..N]}(x_i - y_i) / (x_i + y_i)$, 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).






doi:10/103/nmeth.1778