Counting absolute number of molecules using unique molecular identifiers

Teemu Kivioja[1,2,3,*], Anna Vähärautio[1,3,*], Kasper Karlsson[4], Martin Bonke[1], Sten Linnarsson[4] and Jussi Taipale[3]

[1]Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Biomedicum, P.O. Box 63 (Haartmaninkatu 8) and [2]Department of Computer Science, P.O. Box 68, FIN-00014 University of Helsinki, Finland, [3]Department of Biosciences and Nutrition, and [4]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Sweden.

[*]These authors contributed equally to this work. Address correspondence to JT (jussi.taipale@ki.se) and ST (sten.linnarsson@ki.se).

Advances in molecular biology have made it easy to identify different DNA or RNA species and to copy them. Identification of nucleic acid species can be accomplished by reading the DNA sequence[1,2]; currently millions of molecules can be sequenced in a single day using massively parallel sequencing[3,4]. Efficient copying of DNA-molecules of arbitrary sequence was made possible by molecular cloning[5], and the polymerase chain reaction[6]. Differences in the relative abundance of a large number of different sequences between two or more samples can in turn be measured using microarray hybridization[7] and/or tag sequencing[8,9]. However, determining the relative abundance of two different species and/or the absolute number of molecules present in a single sample has proven much more challenging. This is because it is hard to detect individual molecules without copying them, and even harder to make defined number of copies of molecules. We show here that this limitation can be overcome by using unique molecular identifiers (umis), which make each molecule in the sample distinct.

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. [10]). This severely limits the specificity of DNA microarrays[7,11], as DNA or RNA molecules with different sequences can hybridize to the same target sequence[12,13]. The wide range of concentrations also makes counting molecules by massively parallel sequencing[3,4,8,9] 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. [14,15]), but this destroys information about the original levels, decreasing accuracy of measurements. Amplifying weak signals by making exact (digital) copies of the molecules[5,6] 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. [16]). 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[17].

We describe here a method to quantify absolute number of molecules in a sample that does not require detecting each individual molecule, or keeping track of the number of copies made from them. In this method, each individual molecule of interest is first made unique (**Fig. 1a**). This can be accomplished for example by taking a small aliquot, by fragmentation or by addition of a random DNA sequence label. Any combination of these manipulations can be used to generate a library of molecules where each molecule has a distinct sequence. We define the resulting sequences that can be used to uniquely identify copies derived from each molecule *unique molecular identifiers* (umis; **Fig. 1a**).

As long as the complexity of the library is maintained, it can be (differentially) amplified, normalized and otherwise processed without loss of information about how many molecules

were originally present in the sample. This is because making each molecule different from each other during the library generation step stores the information about the original number of DNA molecules into a molecular memory consisting of the number of distinct sequences (umis) in the library (**Fig. 1a**). Whereas measuring the number of copies of each sequence is difficult, counting the number of distinct sequences (umis) is trivial, and this information is not lost during amplification or any other complexity-preserving manipulation of the library. Normalization of such a library can be performed without loss of accuracy, allowing a much more even precision of measurement across the dynamic range.

Sequencing of the library is then used to determine the absolute number of DNA molecules of each species in the original sample (**Fig. 1a**). When enough sequences have been obtained, each umi will have been observed multiple times, and the number of original DNA molecules can be determined simply by counting the number of umis. However, long before all umis are observed, increasingly precise estimates of the absolute molecule number can be made. 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. More formally, the number of unobserved umis can be estimated based on the distribution of the copy numbers of the observed umis (see Methods for details). Thus, only a small sample of all of the molecules need to be counted in order to accurately estimate the number of molecules in the original sample.

The umi counting method is very effective when simulated data is used (see example in **Fig. 1b**). To assess whether it can also be used to improve measurement precision in an experimental setting, we used umis to count molecules in two different contexts, digital karyotyping and mRNA-sequencing (mRNA-seq). For digital karyotyping, we mixed equal amounts of genomic DNA from a boy with Down's syndrome and his mother. As cell-free DNA from plasma of pregnant women contains a mixture of parental and fetal DNA, this setting is relevant to non-invasive prenatal diagnostics[18,19]. The mixed DNA was fragmented to generate a library of molecules, after which a sample containing less than a single genome copy was taken. In a sample of this size, each molecule is expected to have a different 5' and 3' ends, either of which can be used as umi. After amplification by PCR and sequencing of 20 million reads, we collected the read counts over 5 Mbp genomic intervals. As shown in **Fig. 2a**, the result did not clearly identify that 50% of the sample was derived from DNA with trisomy 21 and a single copy of X. To see if sequencing depth was limiting, we performed the same analysis on a normal

male genome sequenced to 279 million reads, but the coefficient of variation (CV) decreased only slightly (from 7.8% to 7.5%), showing that standard read counting does not converge on the true copy number (**Fig. 2b**).

In contrast, reanalyzing the mixed trisomy-21 sample by counting the umis instead of the reads allowed accurate determination of the DNA copy numbers, clearly revealing increased and decreased copy numbers of 21 and X, respectively (**Fig. 2c**). Among the 20 million reads, we observed 1.28 million umis. On the chromosome level, we observed copy numbers of 1.26 and 0.75 (expected 1.25 and 0.75, respectively for 21 and X). The coefficient of variation for the umi method was 3.0%. This was close to the theoretically maximal accuracy of 2.2% obtained by uniform random sampling of 1.28 million molecules (**Fig. 2d**). Furthermore, unlike the read count method that is inherently limited by the errors introduced during the copying process (compare **Fig. 2** panels **a, b** and **c**), the umi method can be made arbitrarily more accurate by increasing the sample size and sequencing depth. When coverage increases, the number of unique consecutive fragments and the number of unique overlapping fragments can be used to further increase the accuracy of the absolute molecule counting method. This is because consecutive fragments are likely to be derived from a single chromosome molecule, whereas the overlapping fragments must all be derived from different copies of the same chromosome.

If a larger sample is used, the fragments need to be labeled with tags to make all fragments unique. We next tested such a protocol applied to another biologically relevant problem, counting messenger RNA molecules expressed in cells[17,20]. For this, we used a strategy where RNA is randomly fragmented, and converted to cDNA using oligo-dT primed reverse transcription and a template-switch (**Fig. 3**). The template-switch oligonucleotide contained a standard Illumina sequencing primer with or without a 10 base pair random label sequence. The resulting single-stranded cDNA fragments were directly amplified by PCR and sequenced using Illumina Genome Analyzer. In this method, only one fragment is derived from each mRNA, and the combination of the sequences of the label and 5' of the fragment can be used as the umi. Thus, the approximately one million random labels used are sufficient to generate umis from mRNA amount that corresponds to the amount found in ~ 1000 *Drosophila* S2 cells.

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 (not shown). Counting the reads after 15 or 25 PCR amplification cycles from the same

reaction revealed a bias in the PCR that resulted in loss of accuracy of the read counting method, with 418 of the 5097 genes measured differing more than 5% between the samples (**Fig. 3b**, red dots). Using the umis to estimate the absolute number of molecules in the original cDNA 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 10 (**Fig. 3d**). Analysis of the average copy number of the umis mapping to each gene revealed that there was a clear GC-bias in the raw read counts (**Fig. 3c**), presumably due to preferential amplification of sequences with low GC content during the PCR[16]. However, the CG content explained only a small fraction of the copy number variance, indicating that a simple correction cannot be used to significantly improve the accuracy of the read counting method.

In summary, we describe here a method that allows efficient counting of the absolute number of individual molecules in a sample. The method is compatible with sample indexing using separate DNA barcodes, allowing parallel analysis of samples. Existing digital molecule counting methods such as digital PCR[21], digital microarray profiling[22] and single molecule sequencing[23] 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. Furthermore, the presented method can be used to estimate the number of molecules without actually observing all of them. In contrast, deriving accurate estimates based on the previously described methods requires that all molecules are observed, at least in an aliquot of the sample.

In addition to the two applications described here, the presented method could be used to monitor mixing of complex solutions and in tracing flow patterns. Encoding of the concentration information in the number of distinct label sequences permits very extensive amplification and/or 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). In principle, the method can be used 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. The method is likely to have wide applicability in mRNA tag sequencing, ChIP-sequencing; diagnostic applications such as karyotyping and DNA copy number analysis; and manufacturing process control and monitoring.

METHODS SUMMARY

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. Samples were prepared as previously described[24] except that the mixed sample was aliquoted before PCR and ligated with a mixture of eight adapters carrying distinct 6 bp barcodes. The boy/mother samples were mixed 1:1. The 5' positions of mapped[25] reads were used as umis.

RNA-seq. Fragmented total RNA from *Drosophila melanogaster* S2 cells was synthesized to cDNA library with a modified SMART protocol[20,26,27] using an oligo-dT containing adapter that targets fragments containing a polyA border. For absolute molecule counting, a random ten base DNA sequence label was added to the 5' adapter containing Illumina adapter sequence and a barcode. To ensure that the label incorporation occurs only once, the label was designed to contain deoxyuridine bases, which were excised after reverse transcription. Libraries were amplified with PCR and 54 base pair sequence reads were obtained using Illumina GAIIx.

RNA-seq data analysis. The sequencing reads excluding the label and index sequences and the following two bases were mapped to longest transcript of each gene in the *Drosophila* genome. 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. The number of molecules from each gene was estimated by fitting a zero-truncated Poisson distribution to the umi copy number distribution[28] and adding the predicted number of unobserved umis to the observed umi count.

Normalized cDNA library simulation. 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 [15]). 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 [15]. 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.

Acknowledgments

We thank Drs. Minna Taipale, Hanna Secher Lindroos and Esko Ukkonen for critical review of the manuscript. We thank Magnus Nordenskjld and Erik Iwarsson forthe trisomy-21 DNA.

Author contributions

Experiments were conceived and designed by S.L., J.T., A.V. and T.K. Biological experiments were performed by A.V., K.K. and M.B. Data was analyzed by S.L., J.T., A.V. and T.K. The paper was written by J.T., A.V., T.K. and S.L.

REFERENCES

1 Maxam, A. M. and Gilbert, W., A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74** (2), 560 (1977).

2 Sanger, F. and Coulson, A. R., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94** (3), 441 (1975).

3 Margulies, M. et al., Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437** (7057), 376 (2005).

4 Shendure, J. et al., Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309** (5741), 1728 (2005).

5 Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B., Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70** (11), 3240 (1973).

6 Saiki, R. K. et al., Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230** (4732), 1350 (1985).

7 Schena, M., Shalon, D., Davis, R. W., and Brown, P. O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270** (5235), 467 (1995).

8 Okubo, K. et al., Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* **2** (3), 173 (1992).

9 Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W., Serial analysis of gene expression. *Science* **270** (5235), 484 (1995).

10 Holland, M. J., Transcript abundance in yeast varies over six orders of magnitude. *J Biol Chem* **277** (17), 14363 (2002).

11 Hoheisel, J. D., Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7** (3), 200 (2006).

12 Kane, M. D. et al., Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28** (22), 4552 (2000).

13 Koltai, H. and Weingarten-Baror, C., Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic Acids Res* **36** (7), 2395 (2008).

14 Patanjali, S. R., Parimoo, S., and Weissman, S. M., Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci U S A* **88** (5), 1943 (1991).

15 Zhulidov, P. A. et al., Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res* **32** (3), e37 (2004).

16 Benita, Y. et al., Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res* **31** (16), e99 (2003).

17 Ozsolak, F. and Milos, P. M., RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12** (2), 87 (2011).

18 Chiu, R. W. et al., Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A* **105** (51), 20458 (2008).

19 Fan, H. C. et al., Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A* **105** (42), 16266 (2008).

20 Levin, J. Z. et al., Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7** (9), 709 (2010).

21 Vogelstein, B. and Kinzler, K. W., Digital PCR. *Proc Natl Acad Sci U S A* **96** (16), 9236 (1999).

22 Macevicz, S. C., USA Patent No. 11/125,043 (application) (2005).

23 Ozsolak, F. et al., Amplification-free digital gene expression profiling from minute cell quantities. *Nat Methods* **7** (8), 619 (2010).

24 Linnarsson, S., Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp Cell Res* **316** (8), 1339 (2010).

25 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10** (3), R25 (2009).

26 Cloonan, N. et al., Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5** (7), 613 (2008).

27 Zhu, Y. Y. et al., Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30** (4), 892 (2001).

28 Stasinopoulos, D M and Rigby, R A, Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software* **23** (7), 1 (2007).

29 Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25** (14), 1754 (2009).

FIGURE LEGENDS

Figure 1. Counting absolute number of molecules using unique molecular identifiers (umis). **a,** Schematic description of the molecule counting method. Three different DNA-species (top; green, blue and black lines) are labeled with a collection of random labels (middle; colored filled circles). The number of labels used is larger than the number of molecules of each species, making all molecules in the labeled sample different from each other. Conceptually, each molecule then contains a unique molecular identifier (umi). After amplification and normalization, the information about the original number of molecules (top) is preserved in the number of different umis detected by sequencing of a sample (bottom) of the amplified and normalized library. For example, two green molecules are originally present (top), and two different umis (red, blue) are present in the green DNA-molecules sequenced (bottom). If only some umis are observed multiple times, the original number of molecules can still be estimated using count statistics ('Poisson'; bottom middle). **b,** 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 Methods 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).

Figure 2. Digital karyotyping by counting absolute number of molecules. The figure shows the copy number of all 5 Mbp windows on the human genome, normalized to the average of the autosomes. Chromosomes 21 and X are indicated by shading (the Y chromosome was excluded because it was too repetitive). **a,** Standard digital karyotype based on genomic DNA from a boy with trisomy 21 and his mother, mixed 1:1. The coefficient of variation (CV) was 7.8%. **b,** Standard digital karyotype of a normal male sample (CV = 7.5%). **c,** The same sample as in (a) analyzed by absolute molecule counting (CV = 3.0%). **d,** Simulated sample by uniform random sampling of 1.28 million reads in silico (CV = 2.2%).

Figure 3. Accuracy of RNA-seq can be improved by absolute molecule counting. **a,** Schematic description of the RNA-seq method. RNA (gray) is fragmented and reverse transcribed to DNA (black) using an oligo-dT primer with a Illumina linker sequence (blue). A 5' adapter containing another Illumina linker (red), 10 bp random label (yellow) and an index sequence (green) is added to the cDNA by template switch. The combination of label sequence and the position of the 5' end of the RNA forms the umi. **b-d,** Correction of PCR bias by absolute molecule counting. Measurements of expression levels of the same set of genes after 15 (x-axes) and 25 (y-axes) PCR amplification cycles obtained using read counts (b) or absolute molecule counts (d). Genes for which the difference between the measurements is 5% or higher are in red. Number of outliers (red), squared Pearson correlation coefficient (R^2) and coefficient of variation (CV) all indicate the greatly improved accuracy of the absolute molecule counting method. Preferential amplification of fragments with low GC content is revealed by density plot (c) showing average copy number of umis after 15 PCR cycles as a function of the average GC content of the fragments for each measured gene from (b, d). Red line in (c) indicates a least squares fit, for which a p-value and adjusted R^2 value are also given.

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. The boy/mother samples were mixed 1:1. Samples were prepared as previously described[24], except that the mixed sample was aliquoted before PCR, aiming to obtain approximately 20 million molecules (the actual number of umis was 1.28 million; we attribute the difference to losses in sample preparation), and was ligated with a mixture of eight adapters carrying distinct 6 bp barcodes. Sequences were generated on an Illumina Genome Analyzer, 76 bp single-read for the mixed sample and 100 bp paired-end for the adult male sample. Reads were mapped to the genome using Bowtie[25].

We analyzed the genome in non-overlapping 5 Mbp 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 chromosome Y. For absolute molecule counting, we used the 5' position of each read as umi. To verify that umis did in fact identify single molecules, we searched for instances where copies of a umi (i.e. multiple reads aligned to the same position) carried different barcodes. We found only 25 such instances. To determine the theoretical best accuracy obtainable with 1.28 million umis, we generated a simulated sample with this number of reads and analyzed it along with the real samples.

RNA-seq

Total RNA from S2 cells transfected with GFP dsRNA was fragmented with hydrolysis 3 min incubation at 70 °C in 1x RNA fragmentation buffer (Ambion). Reaction was terminated as instructed by manufacturer.

The cDNA synthesis was performed according to the SMART protocol[27] with addition of adapters for massively-parallel sequencing[20,26] using an oligo-dT containing adapter (5'-

CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTTTTTTTTTTTTTTTTTTTTVN-3'; Eurofins MWG Operon) with the following modifications: 3 µl of the unpurified solution containing 50 ng of fragmented total RNA was used in 15 µl cDNA synthesis reaction with 12 pmol of the oligo-dT and template switch oligonucleotides and $MgCl_2$ was added to 15 mM. In addition, a more thermostable reverse transcriptase, SuperScriptIII (Invitrogen; 200 U), was used along with the supplied buffer. For absolute molecule counting, a random ten base DNA sequence label (**N**) was added to the 5' adapter (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNN**dU**NNNN**dU**NNN***GACTT*rGrGrGrG-3'; Integrated DNA technologies). Sequence in italic type represent an index sequence that was used to enable multiplexed sequencing. dU and rG represent deoxyuridine and guanine ribonucleotide, respectively. 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. 20% trehalose was included in the 50 µl reactions, 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 1 volume of Agencourt XP beads (Beckman), and subjected to Illumina GAIIx massively parallel sequencing according to manufacturer's instructions (54 base pair reads). Sequences that are derived from RNA from the S2 cell line will be deposited to NCBI short read archive, accession SRA-0xxxxx.

RNA-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 bwa software version 0.5.8 with default parameter values[29]. The two bases were removed from the 5' end of the reads after index and label sequences to prevent G bias introduced by the template switch.

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, or either 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 and 23.9 million reads passed these criteria in *Drosophila* S2 cell samples taken after 15 and 25 PCR amplification cycles, respectively.

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.

We assumed that all of the umis of a gene had an equal probability to be 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 GAMLSS R package[28] and adding the predicted number of unobserved umis to the observed umi count. The expression level of a gene was considered to be measured if its 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 cut-offs correspond to approximately 1 to 0.2 mRNA molecules per cell based on yield estimates from RNA quantification of total RNA and spike controls (not shown). The GC content of the sequenced gene fragments were calculated as the average GC content of the subsequences from the position of the mapped read to the 3' end of the reference sequence.

Normalized cDNA library simulation

The simulation example for cDNA normalization corresponds 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. [15]). 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. [15]. 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.

Figure 1 Kivioja et al., 2011

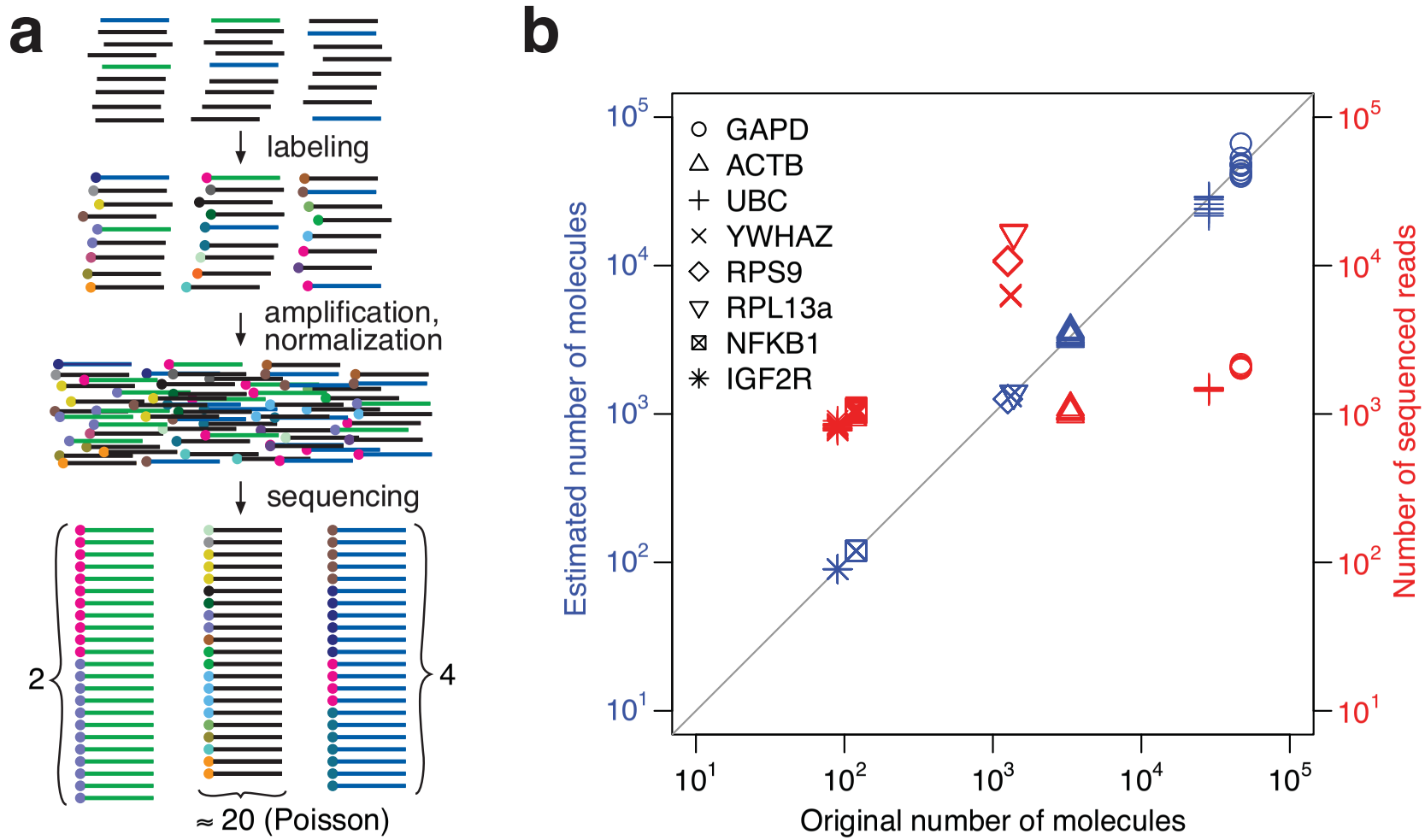

Figure 2 Kivioja et al., 2011

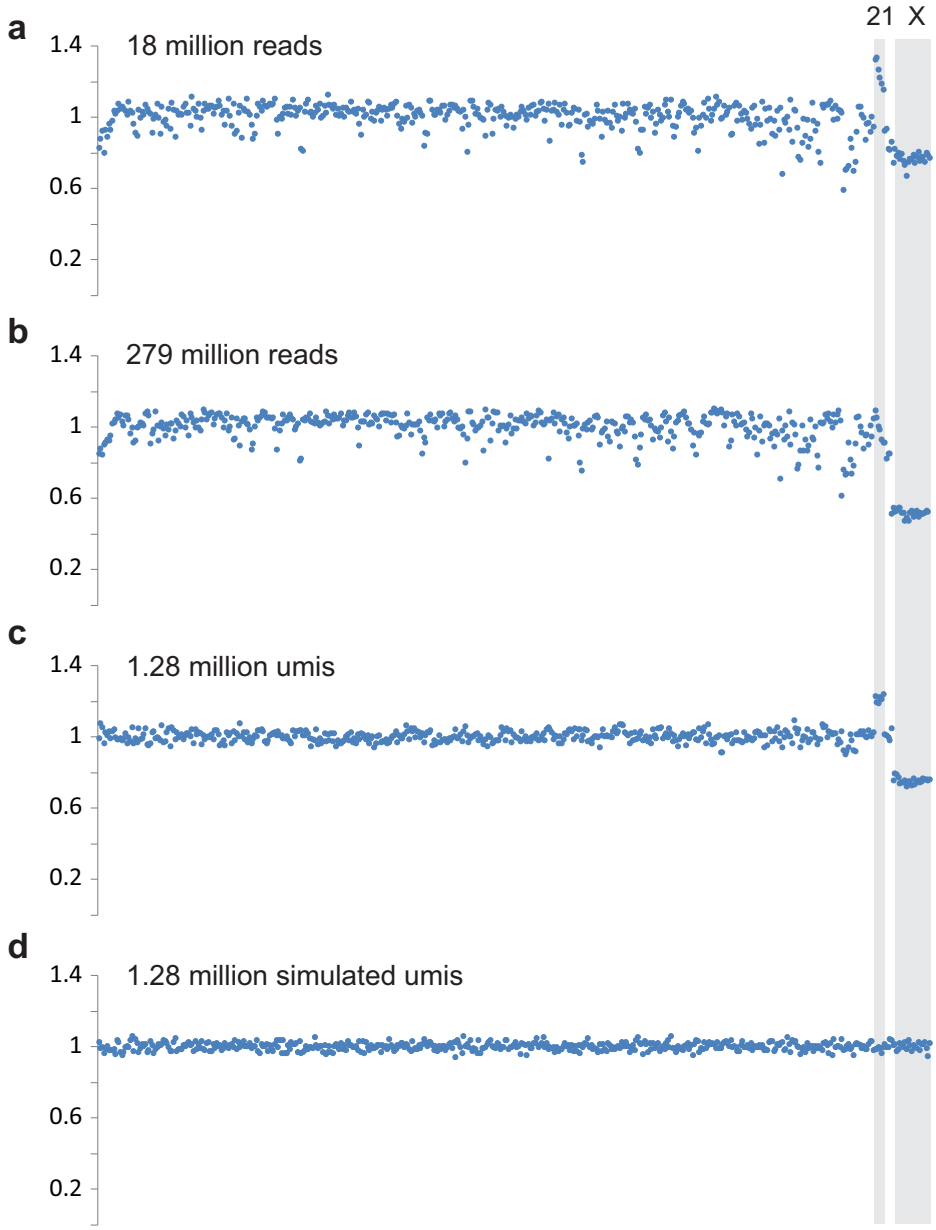

Figure 3 Kivioja et al., 2011