Simultaneous generation of many RNA-seq libraries in a single reaction

Alexander A Shishkin[1], Georgia Giannoukos[2], Alper Kucukural[3], Dawn Ciulla[2], Michele Busby[2], Christine Surka[1], Jenny Chen[2,4], Roby P Bhattacharyya[2], Robert F Rudy[2], Milesh M Patel[2], Nathaniel Novod[2], Deborah T Hung[2], Andreas Gnirke[2], Manuel Garber[3,5], Mitchell Guttman[1,6] & Jonathan Livny[2,6]

Although RNA-seq is a powerful tool, the considerable time and cost associated with library construction has limited its utilization for various applications. RNAtag-Seq, an approach to generate multiple RNA-seq libraries in a single reaction, lowers time and cost per sample, and it produces data on prokaryotic and eukaryotic samples that are comparable to those generated by traditional strand-specific RNA-seq approaches.

RNA-seq has become the gold standard for mapping transcriptomes[1,2], profiling changes in splicing[3,4] and measuring gene expression levels[5,6]. The most widely used method for RNA-seq library construction is the dUTP approach[7]. Although this approach provides high-quality strand-specific RNA-seq profiles, it involves generation of a single library for a single sample[7]. As such, this method is time consuming and expensive to perform on many samples, limiting its utility for applications that require profiling hundreds or thousands of individual samples, such as whole-transcriptome profiling of cancer samples[8,9] or screening the effects of genetic perturbations on gene expression[10–12].

One approach to increase the throughput of library construction is to attach a unique barcode to individual samples and pool these samples to generate a single RNA-seq library. The advantage of this approach is that the cost and time for library generation per sample is reduced as the total number of samples increases. Recently, several protocols have been developed to generate cDNA libraries from pools of barcoded RNA samples[13–17]. Although these protocols represent an important conceptual advance, they are currently limited because either they introduce barcodes through an oligo(dT) primer and therefore can only profile the 3′ ends of eukaryotic mRNAs[13–15] or they introduce barcodes at later stages of library construction, usually just before PCR amplification, thereby limiting the cost and time saved by multiplexing[16,17].

Here we report RNAtag-Seq, a method for generating a single RNA-seq library containing large numbers of RNA samples that are barcoded and pooled before library construction (Online Methods, **Fig. 1** and **Supplementary Protocol**). Barcoding in RNAtag-Seq is achieved through direct ligation of adaptors to RNA, enabling strand-specific, quantitative sequencing of full-length transcripts in diverse prokaryotic and eukaryotic species at a quality highly comparable to that of the well-established dUTP method[7] for single-sample library construction (**Supplementary Figs. 1** and **2** and **Supplementary Table 1**).

To enable pooled library construction of large numbers of samples by RNAtag-Seq, we empirically identified sets of barcoded adaptors that provide uniform read counts across many samples. First, we designed 96 random barcoded adaptors (see Online Methods) and tagged a single *Escherichia coli* RNA sample with each barcode. We identified a set of 32 barcoded adaptors that, when individually ligated to replicate samples before pooled library construction, produced uniform read counts across these samples (less than threefold variation; **Supplementary Fig. 3a** and **Supplementary Table 2**). This variation is sequence intrinsic as independent synthesis of these barcodes produced similar read distributions (**Supplementary Table 2**). To define sets of barcodes of arbitrary size that provide similarly uniform read counts, we synthesized a pool of RNA adaptors that contained a degenerate 6-nt barcode (N_6) (Online Methods), ligated this pool to mouse RNA, quantified the number of reads obtained for each barcode, and identified several large cohorts of barcodes that (i) yielded uniform read counts and (ii) were divergent enough to allow for correct sample assignment even in the case of sequencing errors (**Supplementary Table 2** and Online Methods). We selected one cohort of 54 barcodes and individually synthesized and ligated them to mouse RNA. This was followed by pooling, library construction and sequencing. The resulting distribution of reads per barcode was highly uniform (less than twofold variation; **Supplementary Fig. 3b**), demonstrating that this approach enables the identification of large sets of barcoded adaptors that perform well in pooled generation of RNA-seq libraries.

Notably, normalized gene expression values were highly correlated among replicate samples barcoded with different adaptors (**Supplementary Figs. 4** and **5a** and **Supplementary Table 3**). In addition, the frequencies of dinucleotide pairs following the adaptor sequence closely mirrored those present in all annotated genes (**Supplementary Fig. 5b**), highlighting that the sequence of RNA fragments did not markedly affect their ligation frequency.

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Bioinformatics Core, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [4]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [6]These authors jointly directed this work. Correspondence should be addressed to M. Guttman (mguttman@caltech.edu) or J.L. (livny@broadinstitute.org).

RECEIVED 27 OCTOBER 2014; ACCEPTED 8 FEBRUARY 2015; PUBLISHED ONLINE 2 MARCH 2015; DOI:10.1038/NMETH.3313

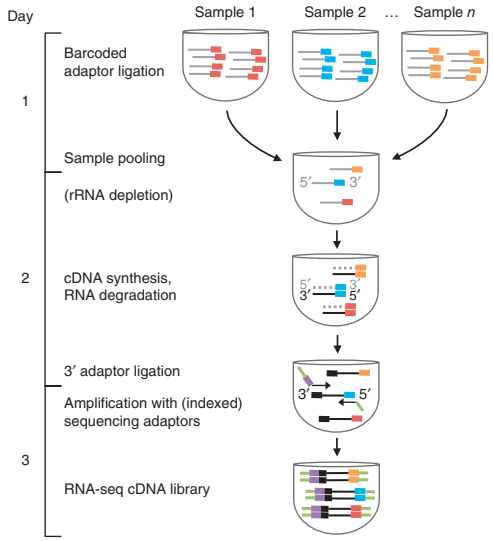

Figure 1 | Schematic of RNAtag-Seq method. Gray and black lines correspond to RNA and cDNA, respectively. Colored blocks represent unique sequence barcodes. Light green lines and purple bars represent Illumina sequencing adaptors and Illumina index barcodes, respectively.

Together, these data demonstrate that ligation of different barcoded adaptors does not introduce quantification biases. Although the total number of reads across barcodes varied by as much as threefold, this variation does not affect gene quantification or differential expression analysis because these measures normalize gene values by the total number of reads generated for a sample[18].

To demonstrate the utility of RNAtag-Seq for identifying differential gene expression in mammalian samples, we profiled the transcriptomes of 11 different tissues and three developmental stages in the mouse (**Supplementary Table 4**). In total, we identified >4,500 differentially expressed genes across all samples (fold change >2, adjusted *P* value (P_{adj}) < 0.01). Notably, the differentially expressed genes that were detected recapitulated the known gene expression differences that mark these different tissue samples (**Fig. 2a**). For example, 649 genes whose expression was significantly different in the brain and spinal cord compared to the remainder of the samples (fold change >2, $P_{adj} < 0.01$) were highly enriched for various neural functions including generation of neurons ($P_{adj} < 0.01$), regulation of synaptic transmission ($P_{adj} < 0.001$) and ion transport ($P_{adj} < 0.001$) (**Fig. 2b** and **Supplementary Table 5**). Similarly, we identified 150 genes that were differentially expressed between the eye and all remaining samples. These genes were highly enriched for visual perception ($P_{adj} < 10^{-60}$), sensory perception of light ($P_{adj} < 10^{-60}$), photoreceptor cell development ($P_{adj} < 10^{-15}$) and eye development ($P_{adj} < 10^{-15}$) (**Fig. 2b**). These results highlight that RNAtag-Seq is able to pick out the well-characterized differences in gene expression across multiple samples in mammalian specimens.

We also profiled and compared the transcriptomes of multiple individual bacterial samples processed in a single pool. Recent work has shown that transcriptional responses to antibiotic exposure can be used to distinguish drug-susceptible and drug-resistant bacteria in clinical samples[19]. However, such signatures have been determined for few clinically relevant pathogens, owing in part to the large number of individual samples needed for these analyses. To determine a transcriptional signature for ciprofloxacin susceptibility in *E. coli*, we profiled samples from ciprofloxacin-susceptible (CipS) and ciprofloxacin-resistant (CipR) clinical isolates, two of each exposed to ciprofloxacin and two of each not, at four time points of growth (**Supplementary Table 6**). A total of 665 and 334 genes were significantly up- and downregulated, respectively (fold change > 3, $P_{adj} < 0.05$), after drug exposure of CipS strains (**Fig. 2c** and **Supplementary Table 7**), including all 20 genes in the well-characterized SOS pathway induced by DNA-damaging agents such as fluoroquinolones[20] (**Fig. 2c** and **Supplementary Fig. 6**). Sixty-seven genes were differentially expressed in only CipS strains and at all time points following ciprofloxacin exposure (**Fig. 2c** and **Supplementary Fig. 6**), suggesting their expression provides a

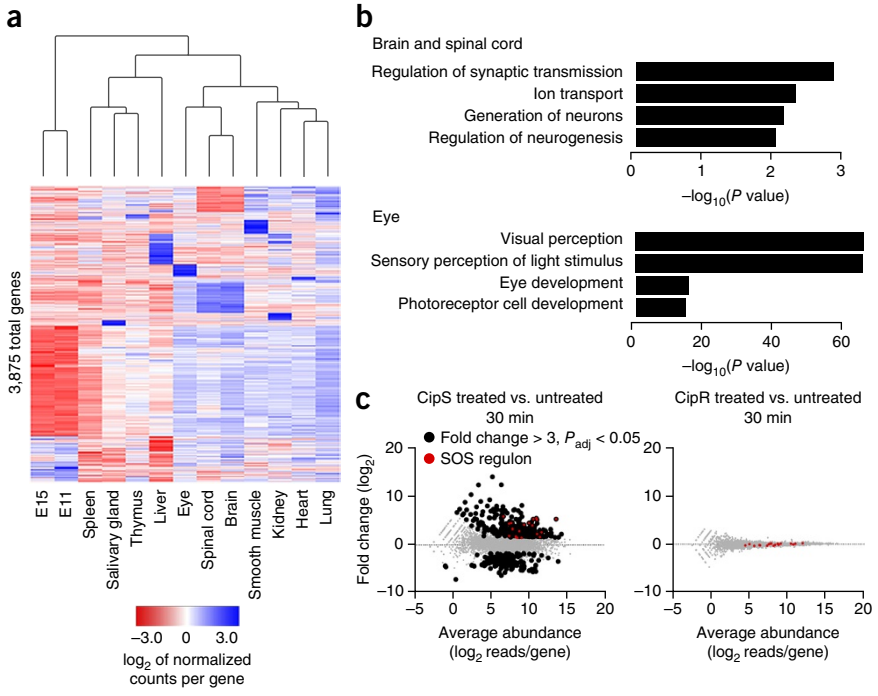

Figure 2 | Differential gene expression analysis using RNAtag-Seq. (**a**) Heat map of all 3,875 differentially expressed genes (fold change >2, $P_{adj} < 0.01$ with P_{adj} corresponding to the *P* value adjusted for multiple testing using the Benjamini-Hochberg procedure) across adult mouse tissues and mouse embryos at developmental stages E11 and E15. (**b**) Selected Gene Ontology categories and their enrichment for specific tissues: brain and spinal cord samples (top) and eye samples (bottom) relative to all samples. The enrichment is plotted as the $-\log_{10}$ of the enrichment *P* value. (**c**) *MA* plots of 2 ciprofloxacin-susceptible (CipS, left) and 2 ciprofloxacin-resistant (CipR, right) *E. coli* clinical isolates 30 min after exposure to ciprofloxacin versus untreated. Genes found to be significantly up- and downregulated (greater than threefold, $P_{adj} < 0.05$) by RNAtag-Seq are colored black. Genes in the SOS regulon are colored red.

specific and stable transcriptional signature for identifying ciprofloxacin susceptible *E. coli* strains. Taken together, our findings demonstrate that RNAtag-Seq is a robust method for identifying differentially expressed genes across numerous libraries generated in a single pool in both eukaryotes and bacteria.

Because many samples can be pooled before library construction, RNAtag-Seq requires lower input RNA amounts per sample than existing protocols. Furthermore, as this method does not require poly(A) capture or enrichment, RNAtag-Seq can also accommodate highly fragmented RNA samples and can be used to profile all transcripts, including short and long as well as coding and noncoding RNAs, in both prokaryotic and eukaryotic samples. RNAtag-Seq can therefore be implemented in a wide variety of biological systems and for diverse applications of RNA-seq such as gene expression profiling in large-scale genetic and chemical screens; human population genetics studies; archived formalin-fixed, paraffin-embedded samples; protein-RNA interaction mapping; small RNA-seq; and simultaneous profiling of pathogen and host transcription during infection.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI Sequence Read Archive: SRP051252.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all members of Lander and Guttman labs for their help, especially J.M. Engreitz, P. McDonel and K. Sirokman; L. Gaffney for assistance with figures; and C. Nusbaum for helpful suggestions on the manuscript. We thank I. Antoshechkin and the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech. This work was supported by a US National Institutes of Health (NIH) Director's Early Independence Award (DP5OD012190 to M. Guttman), funds from the Broad Institute of MIT and Harvard and the California Institute of Technology (M. Guttman) and funds from the US National Institute of Allergy and Infectious Diseases, NIH, Department of Health and Human Services, under contract no. HHSN272200900018C.

AUTHOR CONTRIBUTIONS

A.A.S. and M. Guttman conceived of the approach and developed the initial RNAtag-Seq protocol; A.A.S., C.S., G.G., D.C. and A.G. optimized and streamlined the RNAtag-Seq protocol; A.A.S., G.G., D.C., R.P.B., R.F.R. and M.M.P. generated RNA-seq libraries; A.K., J.C., N.N., M.B., M. Garber, M. Guttman and J.L. analyzed data; R.P.B., D.T.H., M. Guttman and J.L. designed differential expression experiments; M. Guttman and J.L. supervised the project and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Wang, Z., Gerstein, M. & Snyder, M. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
3. Wilhelm, B.T. *et al. Nature* **453**, 1239–1243 (2008).
4. McHugh, C.A., Russell, P. & Guttman, M. *Genome Biol.* **15**, 203 (2014).
5. Garber, M., Grabherr, M.G., Guttman, M. & Trapnell, C. *Nat. Methods* **8**, 469–477 (2011).
6. Nagalakshmi, U. *et al. Science* **320**, 1344–1349 (2008).
7. Levin, J.Z. *et al. Nat. Methods* **7**, 709–715 (2010).
8. Garraway, L.A. & Lander, E.S. *Cell* **153**, 17–37 (2013).
9. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
10. Amit, I. *et al. Science* **326**, 257–263 (2009).
11. Lamb, J. *et al. Science* **313**, 1929–1935 (2006).
12. Ravasi, T. *et al. Cell* **140**, 744–752 (2010).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
14. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
15. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
16. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2012).
17. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2014).
18. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
19. Barczak, A.K. *et al. Proc. Natl. Acad. Sci. USA* **109**, 6217–6222 (2012).
20. Michel, B. *PLoS Biol.* **3**, e255 (2005).

ONLINE METHODS

Culture growth and RNA isolation. Bacterial total RNA was isolated for mid-exponential LB cultures of *Prochlorococcus marinus pastoris* CCMP1986 (31% G+C), *E. coli* K-12 MG1655 (50% G+C) and *Rhodobacter sphaeroides* 2 4 1 uid57653 (67% G+C) as previously described[21]. In comparisons of RNAtag-Seq and dUTP, equal amounts of total RNA from these three organisms were mixed before library construction. Ciprofloxacin-susceptible and ciprofloxacin-resistant *E. coli* strains were provided by Brigham and Women's Hospital under permission from the Institutional Review Board of Partners Health Care, Protocol #2012P001062. In profiling transcriptional responses of *E. coli* to ciprofloxacin, cultures were grown to early log phase in Mueller-Hinton broth and treated with 2.5 mg/L of ciprofloxacin. At the indicated time points, ~5×10^8 cells were harvested by centrifugation, and total RNA was extracted using the Direct-Zol RNA Miniprep protocol (Zymo Research). Mouse tissue for differential expression analyses and K562 RNAs were purchased from Clontech (#636644, Mouse Total RNA Master Panel) and Ambion, respectively.

Generation and sequencing of cDNA libraries. K562 dUTP libraries were generated as described with rRNA depleted using the RNaseH approach[22]. Bacterial dUTP libraries were generated as described[21] with rRNA depleted using with RiboZero (Epicentre). In all RiboZero reactions, the maximal amount recommended by the manufacturer per reaction was used to avoid an additional quantification step during library construction and ensure the RNA did not exceed the capacity of the solution. RNAtag-Seq cDNA libraries were generated according to the detailed protocol in the **Supplementary Protocol**. Briefly, 200–400 ng of total RNA was fragmented, depleted of genomic DNA and dephosphorylated before its ligation to barcoded adaptors with a 5′ phosphate and a 3′ blocking group. DNA adaptors carried 5′-AN_8-3′ barcodes and RNA adaptors 5′-rArN_6-3′ barcodes. Sequences of these barcodes are provided in the **Supplementary Protocol** and **Supplementary Table 2**. Barcoded RNAs were pooled and depleted of rRNA using the appropriate RiboZero rRNA depletion kit (Epicentre) for bacterial and K562 pools (8 samples per pool, **Supplementary Table 2**) and as previously described[23] for mouse pools. These pools of barcoded RNAs were converted to Illumina cDNA libraries in three key steps: (i) reverse transcription of the RNA using a primer designed to the constant region of the barcoded adaptor; (ii) degradation of the RNA and ligation of a second adaptor to the single-stranded cDNA; (iii) PCR amplification using primers that target the constant regions of the 3′ and 5′ ligated adaptors and contain the full sequence of the Illumina sequencing adaptors (**Fig. 1**). Two SPRI cleanup steps are included following adaptor ligations to ensure efficient removal of adaptor dimers (<1% of our sequencing reads represented adaptor dimers). Modifications of the RNAtag-Seq protocol used in generation of mouse libraries are detailed in Appendix A in the **Supplementary Protocol**. cDNA libraries were sequenced on Illumina MiSeq or HiSeq2500

RNA-seq data analysis. For the analysis of RNAtag-Seq data, reads from each sample in the pool were identified on the basis of their associated barcode using in-house scripts. Up to one mismatch in the barcode was allowed with the caveat that it did not enable assignment to more than one barcode. Barcode sequences were removed from reads before alignment. Analysis of bacterial data was conducted as previously described[22,24]. Briefly, reads were aligned to the appropriate RefSeq reference genomes using BWA[25]. Gene annotations were obtained from RefSeq and Rfam[26]. The overall fragment coverage of genomic regions corresponding to features such as ORFs and rRNAs was conducted using an in-house pipeline as described[22,24]. To account for differences in the efficiency of rRNA depletion, we calculated normalized abundance per gene from the number of fragments per kilobase of that gene per million fragments aligned to all ORFs (FPKMO: fragments per kilobase of ORF per million fragments aligned to all ORFs). Sequencing reads from *E. coli* clinical isolates were aligned to the *E. coli* HS reference sequence (RefSeq accession NC_009800). SOS-responsive genes in *E. coli* HS were identified on the basis of their homologs in the SOS regulon of *E. coli* K-12 (ref. 27).

Analyses of K562 data were performed using the computational pipeline developed in Adiconis *et al.*[22]. Briefly, to calculate the number of read pairs per gene, we aligned reads to the human genome version 19 using Bowtie 0.12.7 (ref. 28) and calculated reads per gene using scripts based on the BamTools API[29]. Normalized read counts per gene were calculated using the RSEM package version 1.1.17 (ref. 30).

In the analysis of mouse tissue data, Bowtie 2 (ref. 31) was used to remove reads aligning to rRNA, and the remaining reads were aligned by RSEM[30] to a mouse transcript reference files generated using UCSC annotations. RSEM was then used to calculate total and normalized reads per gene and genes that to which fewer than ten total reads aligned were eliminated from further analysis.

Custom software used to process and analyze RNA-Seq data software is not being made available as part of this publication.

Identifying a set of 32 DNA adaptors that work well together. To design a set of random barcoded DNA adaptors, we enumerated all possible 8-nt barcodes and then selected a random set of 96 barcodes that contained at least 3-nt distances with all other sequences in the selected set. This distance would allow us to assign a read even if with two sequencing errors within the 8-nt barcode. We then synthesized these 96 DNA adaptor with a 5′ adenosine followed by the barcode sequence. We ligated each of the barcodes to the same total RNA and measured the uniformity of coverage. We identified 32 adaptors that showed the lowest across-barcode variation (less than threefold).

A pooled strategy to identify large cohorts of adaptors that work well together. To identify a cohort of RNA adaptors that work well together, we synthesized an RNA adaptor containing a 5′ adenosine followed by a degenerate 6-nt barcode (N_6) that should contain large complexity of all possible 6-nt barcodes. We ligated these N_6 adaptors to poly(A)-selected mouse ES cell RNA and generated a library. We then sequenced this library and measured the total number of reads present for each individual N_6 sequence. To account for differences in the number of reads due to differences in the relative abundance of the adaptor rather than its ligation efficiency, we sequenced the adaptor pool directly. We computed an enrichment score that is defined as the coverage of the RNA samples that contain a given barcode divided by the number of reads present for each barcode alone.

doi:10.1038/nmeth.3313

Using this normalized score, we identified several cohorts of sequences that contained uniform coverage within the set (less than twofold variation) and a minimum nucleotide distance of 2 to allow for sequencing error correction. We chose a small cohort that contained 54 barcodes and synthesized these individually to generate a panel of barcodes.

Comparison of RNAtag-Seq and dUTP. To calculate coverage across K562 transcripts, we aligned reads using Bowtie 0.12.7 (ref. 28) to the human transcriptome obtained from the UCSC Genome Browser and calculated the distribution of reads along the length of these transcripts as described[22]. Coverage across bacterial ORFs was calculated using in-house scripts based on the alignment of reads to the genome. The calculation of the number of genes detected as a function of reads sequenced was conducted using code available in the Scotty package[32].

Quantification of dinucleotide frequencies. Dinucleotide frequencies were calculated for the first two bases of reads aligning to *E. coli* from RNA-seq data generated from mixed *P. marinus*, *E. coli* and. *R sphaeroides* RNA. For read 1 in RNAtag-Seq data, sequences of barcodes plus the universal 3′ nucleotide were removed before alignment. Dinucleotide frequencies for all *E. coli* ORFs were generated using sliding two-base windows across the entire lengths of RefSeq-annotated protein-encoding genes.

Identification of differentially expressed genes. For differential expression analysis of *E. coli*, DESeq[18] was used to compare total reads per ORF between relevant time points or conditions. At each time point and condition, the two antibiotic-resistant strains were treated as biological replicates as were the two antibiotic-susceptible strains. For differential expression analysis of mouse tissues, DESeq2 (ref. 33) was used to compare RSEM expected count values per gene. In both DESeq and DESeq2 analyses, adjusted *P* values (P_{adj}) were used as cutoffs for statistical significance. Because RNA level patterns of some of the cell types analyzed are very similar, we compared all possible combinations of one cell type, two cell types and three cell types versus all other cell types and joined all genes sets with more than 40 genes in each comparison to generate the heat map in **Figure 2a**. Assignment and analysis of mouse Gene Ontology groups was conducted in R using the DAVID web service[34].

21. Giannoukos, G. *et al. Genome Biol.* **13**, R23 (2012).
22. Adiconis, X. *et al. Nat. Methods* **10**, 623–629 (2013).
23. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. & Lander, E.S. *Cell* **154**, 240–251 (2013).
24. Haas, B.J., Chin, M., Nusbaum, C., Birren, B.W. & Livny, J. *BMC Genomics* **13**, 734 (2012).
25. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
26. Gardner, P.P. *et al. Nucleic Acids Res.* **37**, D136–D140 (2009).
27. Novichkov, P.S. *et al. BMC Genomics* **14**, 745 (2013).
28. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
29. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P. & Marth, G.T. *Bioinformatics* **27**, 1691–1692 (2011).
30. Li, B. & Dewey, C.N. *BMC Bioinformatics* **12**, 323 (2011).
31. Langmead, B. & Salzberg, S.L. *Nat. Methods* **9**, 357–359 (2012).
32. Busby, M.A., Stewart, C., Miller, C.A., Grzeda, K.R. & Marth, G.T. *Bioinformatics* **29**, 656–657 (2013).
33. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
34. Jiao, X. *et al. Bioinformatics* **28**, 1805–1806 (2012).