nature biotechnology

Integrated genome and transcriptome sequencing of the same cell

Siddharth S Dey[1,2,4], Lennart Kester[1,2,4], Bastiaan Spanjaard[1,2], Magda Bienko[1–3] & Alexander van Oudenaarden[1,2]

Single-cell genomics and single-cell transcriptomics have emerged as powerful tools to study the biology of single cells at a genome-wide scale. However, a major challenge is to sequence both genomic DNA and mRNA from the same cell, which would allow direct comparison of genomic variation and transcriptome heterogeneity. We describe a quasilinear amplification strategy to quantify genomic DNA and mRNA from the same cell without physically separating the nucleic acids before amplification. We show that the efficiency of our integrated approach is similar to existing methods for single-cell sequencing of either genomic DNA or mRNA. Further, we find that genes with high cell-to-cell variability in transcript numbers generally have lower genomic copy numbers, and vice versa, suggesting that copy number variations may drive variability in gene expression among individual cells. Applications of our integrated sequencing approach could range from gaining insights into cancer evolution and heterogeneity to understanding the transcriptional consequences of copy number variations in healthy and diseased tissues.

One of the central questions in biology is to understand how genotype influences phenotype. Over the past decade, advances in microarrays and, more recently, next-generation sequencing have started to provide glimpses of this correlation at the genome-wide level[1–4]. However, these studies make measurements starting from a large population of cells or complex tissues, thus providing only an average measurement over the entire population. This obscures direct quantification of how genetic variability may affect the transcriptome at the single-cell level. Furthermore, as cell populations exposed to the same environment can also exhibit dramatic cell-to-cell variability in gene expression[5], the ability to understand the correlation between genotype and gene expression will require direct measurement of the transcriptome and the genome of the same cell. Recently, single-cell genome sequencing[6–11] and single-cell transcriptome sequencing[12–21] have emerged as promising tools for quantifying genetic and expression variability between individual cells[22,23]. However, as these single-cell technologies are limited to quantification of either the transcriptome or the genome, it is currently not possible to explore the relation between genetic and expression variability in single cells. Here we describe a method to simultaneously quantify both the genome and transcriptome of the same cell.

To successfully amplify small quantities of genomic DNA (gDNA) and mRNA from single cells in a way that reduces handling, transfer and separation steps, we devised gDNA-mRNA sequencing (DR-Seq), a method that does not involve physical separation of the nucleic acids before amplification, thereby minimizing losses and chances of contamination. First, hand-picked single cells are lysed and reverse transcribed using a poly-T primer (called adaptor-1x (Ad-1x)) including cell-specific barcodes, a 5′ Illumina adaptor and a T7 promoter overhang to convert mRNA to single-stranded cDNA[13] (**Fig. 1a**). The gDNA and single-stranded cDNA are then subjected to quasilinear whole-genome amplification with an adaptor that has a defined 27-nt sequence at the 5′ end followed by eight random nucleotides[7] (Ad-2) (**Fig. 1a**). After seven rounds of amplification, the gDNA and cDNA are copied to generate a variety of different short (0.5–2.5 kb) amplicon species, with a majority of amplicons containing Ad-2 at both ends and a small fraction of cDNA-derived amplicons containing Ad-2 at one end and Ad-1x at the other (**Fig. 1a**).

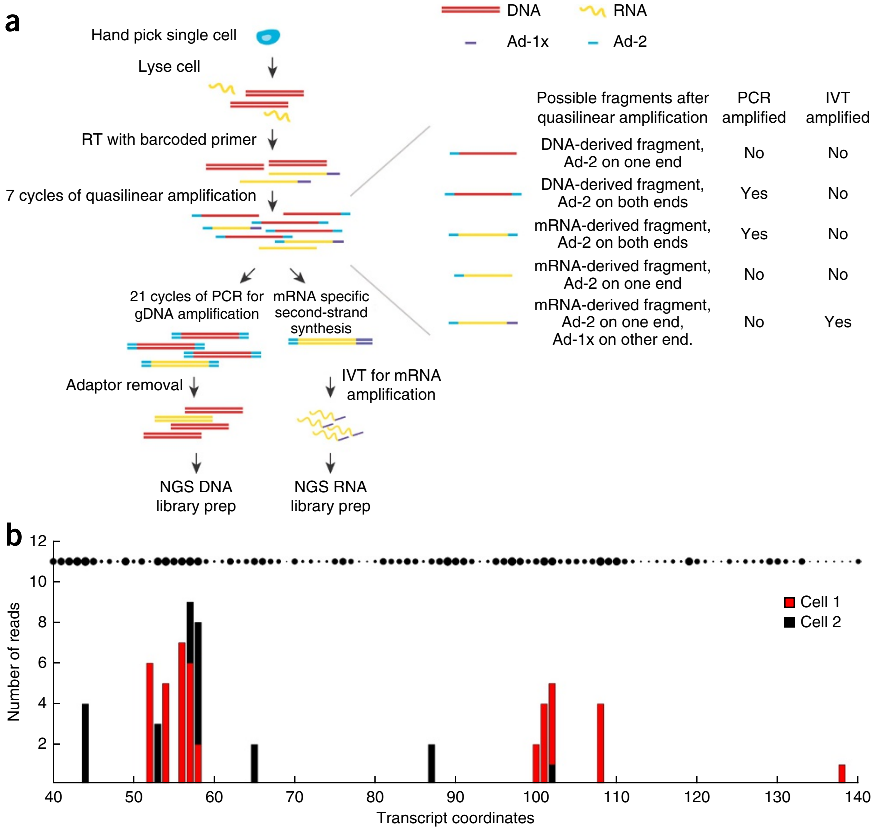
Next, the sample is divided to further amplify gDNA and cDNA (**Fig. 1a**). The half used to sequence gDNA is first amplified by PCR. After sonication, removal of Ad-2 and preparation of a cell-specific indexed Illumina library, this half allows quantification of gDNA. The other half used for cDNA sequencing is converted to double-stranded cDNA and amplified using *in vitro* transcription such that the amplified RNA (aRNA) is uniquely produced from cDNA but not gDNA (**Fig. 1a**). 3′ Illumina adaptors are then ligated to the aRNA and subjected to reverse transcription and PCR, allowing quantification of mRNA.

We first applied DR-Seq to a mouse embryonic stem cell line (E14) to validate how the method compares to existing single-cell gDNA or mRNA sequencing techniques. We performed DR-Seq on E14 cells and sequenced the mRNA from 13 single cells together with the gDNA from three of the 13 cells. Recently a few single-cell transcriptomics methods have employed random sequence–based barcodes to identify unique mRNA molecules, thereby reducing PCR and other amplification biases[17,20,21]. Because cDNA molecules in DR-Seq are randomly primed by Ad-2 during quasilinear amplification, we used

[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, the Netherlands. [2]University Medical Center Utrecht, Cancer Genomics Netherlands, Utrecht, the Netherlands. [3]Present address: Science for Life Laboratory, Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm, Sweden. [4]These authors contributed equally to this work. Correspondence should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu).

Received 25 April 2014; accepted 12 December 2014; published online 19 January 2015; doi:10.1038/nbt.3129

Figure 1 Schematic of DR-Seq. (**a**) After single-cell lysis and reverse transcription (RT) with adaptor Ad-1x (purple), gDNA (red) and single-stranded cDNA (yellow) are amplified by Ad-2 (blue) using a quasilinear amplification strategy. The majority of the short amplicons contain Ad-2 at both ends, and cDNA-derived amplicons contain Ad-2 at one end and Ad-1x at the other. The sample is split into two halves and processed separately to amplify and sequence gDNA or cDNA. IVT, *In vitro* transcription; NGS, next-generation sequencing. (**b**) Distribution of reads within 100 nucleotides of *Dppa5a* for two cells as a function of the random priming location by Ad-2. The unique length-based identifiers found in the two cells can be used to count the original number of cDNA molecules within each cell and minimize amplification biases. Distinct positions are randomly primed within each cell, with high-affinity binding sites being preferentially primed. Dot size indicates the binding propensity of each location.



Possible fragments after quasilinear amplification	PCR amplified	IVT amplified
DNA-derived fragment, Ad-2 on one end	No	No
DNA-derived fragment, Ad-2 on both ends	Yes	No
mRNA-derived fragment, Ad-2 on both ends	Yes	No
mRNA-derived fragment, Ad-2 on one end	No	No
mRNA-derived fragment, Ad-2 on one end, Ad-1x on other end.	No	Yes

the genomic position of such priming events to minimize amplification biases and achieve resolution close to that of identifying unique mRNA molecules (**Supplementary Fig. 1**). Because all amplification products that are generated downstream (during quasilinear amplification, *in vitro* transcription and PCR) from the first randomly primed cDNA-derived amplicon retain the same genomic priming location, amplification-derived duplicates could be removed to identify unique cDNA molecules. The genomic priming location of the first randomly primed cDNA-derived amplicon was called its length-based identifier (**Supplementary Fig. 1**). For example, although hundreds of reads were detected for the *Dppa5a* gene, only 34 and 27 unique length-based identifiers were detected in the two cells shown (**Fig. 1b**). A zoomed-in view of 100 nucleotides within this transcript shows that only a few distinct positions are randomly primed in the two cells, with several reads at each genomic coordinate (**Fig. 1b**). Thus, unique length-based identifiers have the potential to reduce amplification biases and technical noise to enable quantification of the original number of cDNA molecules. To demonstrate that length-based identifiers can be used to achieve resolution close to identifying unique transcripts in single cells, we showed that the original cDNA molecules were primed only once on average during the quasilinear amplification steps, thereby enabling length-based identifiers to uniquely tag each original cDNA molecule (**Supplementary Fig. 2** and **Supplementary Note**). Next, we identified the theoretical number of unique binding sites (and, therefore, length-based identifiers) available for adaptor Ad-2 for each gene in the transcriptome to ensure that the original cDNA molecules from each gene could be counted accurately without reaching saturation (**Supplementary Fig. 3** and **Supplementary Note**). For a majority of the genes, we found between 50 and 250 theoretical binding sites, similar to the resolution of 4-bp random barcodes that have been used as unique molecule identifiers (UMIs) to quantify single-cell transcriptomes[20,21] (**Supplementary Fig. 4** and **Supplementary Note**). Finally, we found that for a majority of expressed genes (>95%), the number of detected length-based identifiers was much smaller than the theoretical number of binding sites, thereby implying that the length-based identifiers do not undercount the number of original cDNA molecules (**Supplementary Fig. 5** and **Supplementary Note**). Together these results show that length-based identifiers in DR-Seq can be used to minimize amplification biases and accurately estimate the underlying distribution of original cDNA molecules.

To demonstrate that length-based identifiers reduce technical noise by minimizing amplification biases and perform similar to random sequence-based UMIs[17,20,21], we compared cell-to-cell variability in the expression of endogenous genes before and after correction using length-based identifiers. We found that the coefficient of variation (CV) in expression reduced for a majority of genes (80%) in DR-Seq after correcting the expression using length-based identifiers, similar to the reduction observed in CEL-Seq (cell expression by linear amplification and sequencing) after correcting the expression with UMIs (**Fig. 2a** and **Supplementary Fig. 6**)[13,21]. This suggests that length-based identifiers in DR-Seq reduce technical noise, thereby allowing quantification of the underlying biological variability in gene expression between single cells. Further, as single cells contain the same amount of External RNA Controls Consortium (ERCC) spike-in molecules, any cell-to-cell variability detected in these molecules represents technical noise and would be expected to display the lowest CV when compared to endogenous genes with similar mean expression levels. We found that the spike-in molecules typically showed the lowest CV for the entire range of mean expressions only after correction of the read-based DR-Seq data with the length-based identifiers (**Fig. 2b**,**c**). We found a similar trend in CEL-Seq after correcting the read-based data with UMIs (**Supplementary Fig. 7**). As a consequence of this reduction in technical noise, length-based identifiers in DR-Seq and UMIs in CEL-Seq improved cell-to-cell pairwise Pearson correlations in the expression of endogenous genes (**Supplementary Fig. 8**). Taken together, these data strongly suggest

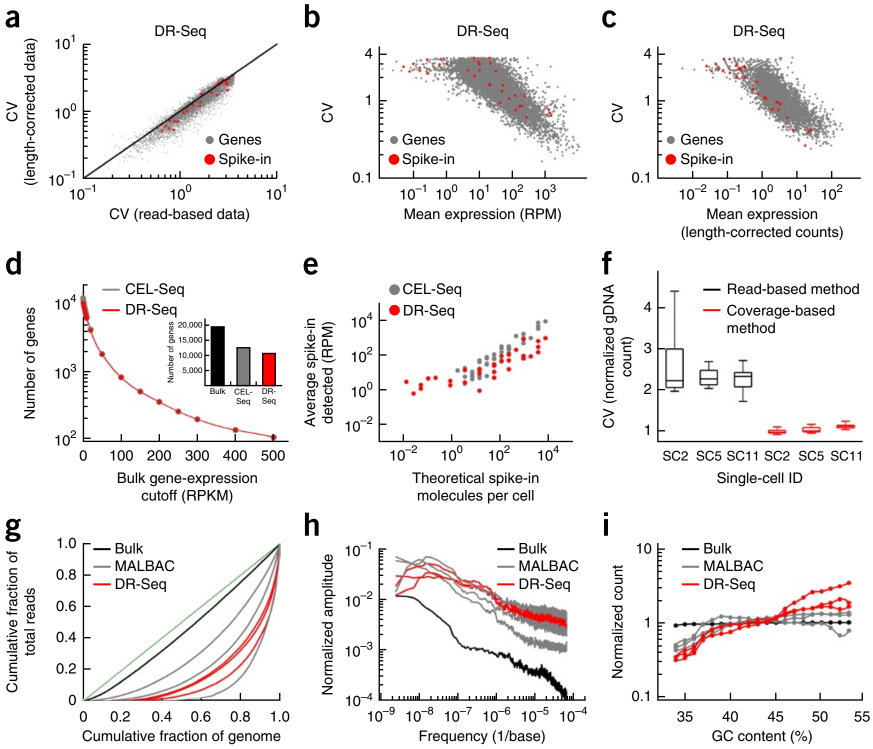


Figure 2 Development of a computational technique to reduce technical noise in DR-Seq data and comparison of DR-Seq to existing single-cell gDNA or mRNA sequencing methods in E14 cells. (**a**) Reduction of cell-to-cell variability in the expression of genes after correction of raw read-based data using length-based identifiers implies reduction in technical noise in DR-Seq data (**Supplementary Fig. 6**). (**b**) CV versus mean expression of genes for read-based data. Compared to endogenous genes (gray), spike-in molecules (red) typically do not display the lowest CV for a given mean expression level, implying that read-based data contain technical noise that obscures biological variability between single cells. (**c**) CV versus mean expression of genes after correcting read-based data using length-based identifiers shows reduced technical variability between single cells (**Supplementary Fig. 7**). (**d**) Comparison of mRNA sequencing results between DR-Seq and CEL-Seq in detecting genes above different expression thresholds obtained from bulk mRNA sequencing data. Inset, total number of genes detected by bulk mRNA sequencing, CEL-Seq and DR-Seq (**Supplementary Fig. 10**). (**e**) Detection of ERCC spike-in molecules in both methods increased monotonically with the expected number of molecules per cell. Shown are spike-ins found in at least two single cells. (**f**) Box plot comparing bin-to-bin variability in gDNA read counts using two different methods for three cells amplified by DR-Seq. The box plots show the CV of read distribution over all the autosomes in the mouse genome. Central mark indicates median, lower and upper edges of the box indicate the 25th and 75th percentiles, respectively, and whiskers extend 1.5 times of the interquartile range beyond the edges of the box. (**g**) Comparison of single-cell gDNA sequencing results between DR-Seq and MALBAC. The green line indicates the theoretical limit, with reads distributed uniformly across the whole genome using Lorenz plots. (**h**) Power spectrum of read distribution over different genomic length scales for bulk sequencing and single cells processed by DR-Seq and MALBAC. (**i**) Read distribution for regions of the genome with different GC contents. RPM, reads per million; RPKM, reads per kilobase per million.

that length-based identifiers in DR-Seq substantially reduce technical noise, thereby allowing us to accurately count the number of original cDNA molecules and capture the underlying biological variability between single cells.

We then compared the length-corrected mRNA sequencing results obtained from DR-Seq to results obtained from sequencing 33 single E14 cells by CEL-Seq[13,21] (**Supplementary Fig. 9**). To avoid sampling bias in comparing DR-Seq to CEL-Seq, we chose the 13 cells with the highest read counts out of 33 that were sequenced using CEL-Seq. Despite these stringent criteria, DR-Seq and CEL-Seq detected similar numbers of genes, with 9,735 genes in common between the two methods (**Fig. 2d** and **Supplementary Fig. 10**). The numbers of genes detected above different expression thresholds were also similar between the methods (**Fig. 2d**). In addition, gene expression correlations for each method compared to bulk sequencing were similar (**Supplementary Table 1**). Finally, analysis of synthetic ERCC spike-in RNA molecules showed that 66 different spike-in species from a total of 92 were detected by DR-Seq, compared to 51 species detected by CEL-Seq[24] (**Fig. 2e** and **Supplementary Fig. 11**). This increased sensitivity in detecting low-abundance spike-ins is probably due to the exponential amplification of cDNA-derived amplicons for the remainder of the quasilinear amplification steps in DR-Seq. For the higher range of concentrations, the number of spike-in molecules detected correlated well with the expected number of molecules (**Fig. 2e**). Further, we detected linear correlation over three orders of magnitude between the detected and theoretical number of spike-in molecules for both CEL-Seq and DR-Seq, which suggests that the methods have similar dynamic ranges in detecting transcripts (**Fig. 2e**). Expression of endogenous genes also spanned three orders of magnitude in both CEL-Seq and DR-Seq, implying that the methods have similar sensitivities in amplifying reverse-transcribed cDNA molecules (**Fig. 2d**). Taken together, these analyses of the two methods across different metrics suggest that DR-Seq performs similarly to CEL-Seq and the additional steps involved in amplifying gDNA do not adversely affect the mRNA sequencing results (**Supplementary Fig. 12** and **Supplementary Note**).

To analyze gDNA sequencing results in DR-Seq, we mapped reads from the gDNA fraction to the genome after masking out the coding sequences. This masking is performed because the fraction that is used to sequence gDNA also contains reads that originate from cDNA molecules within coding sequences (**Fig. 1a**). By masking the coding sequences within the genome, such ambiguous reads that might arise from either gDNA or cDNA within coding regions are discarded computationally, leaving only reads that arise from gDNA (**Supplementary Fig. 13**). Because the coding regions make up a small portion of the genome, such a strategy does not influence copy-number calling over large genomic regions (Online Methods). Further, gDNA reads in DR-Seq were distributed into unequal bins of variable size to account for the masking of the genome (**Supplementary Fig. 14** and Online Methods). Further, to reduce amplification biases introduced during the quasilinear amplification steps, we developed a computational technique to reduce bin-to-bin technical noise in gDNA read counts. During quasilinear amplification, the first amplicons that are generated from the gDNA template do not loop out of the reaction pool and remain templates for the remaining cycles. Thus, differences in the cycle in which gDNA regions are first amplified can introduce bin-to-

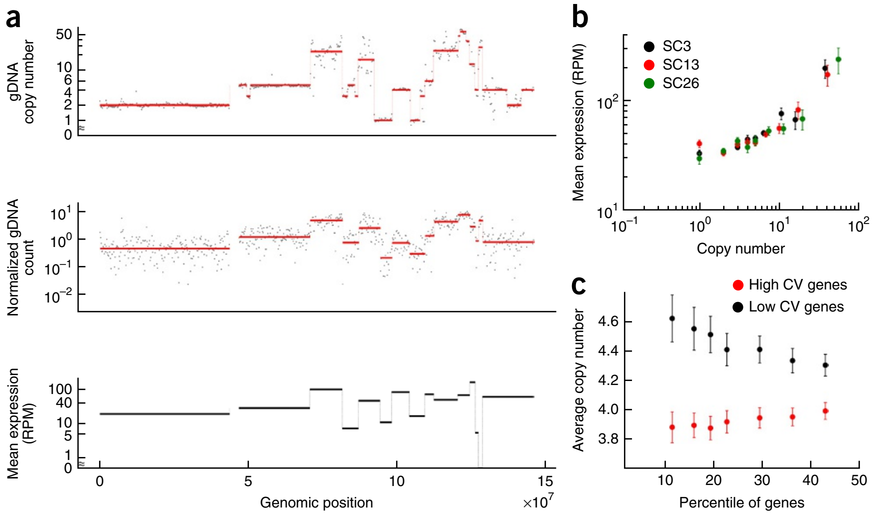


Figure 3 DR-Seq on SK-BR-3 cells. (**a**) Top, raw gDNA data (dots) and different copy numbers (red lines) identified using the CBS algorithm[26] for chromosome 8 in bulk sequencing data. Middle, raw data (dots) and median read counts (red lines) identified using CBS for one cell (SC13). Median read depths for each segment in single cells and the bulk copy numbers are used to estimate copy number variations in single cells (**Supplementary Note**). For each median level identified from the single-cell gDNA data (middle), mean expression of genes within each segment was calculated (bottom). (**b**) Genome-wide quantification of mean expression of genes within different copy number regions (**Supplementary Fig. 25**). Data for three single cells SC3, SC13 and SC26 are shown in black, red and green, respectively. (**c**) For a range of mean expression levels (5–400 reads per million (RPM)), genes showing the highest and lowest noise (quantified as CV) were identified. The *x* axis shows the percentage of most noisy and least noisy genes from all the genes considered in the analysis. The noisiest genes are associated with low copy number regions and vice versa (**Supplementary Fig. 27**). Error bars (**b**,**c**) represent s.e.m. (bootstrap).

bin variability and pileup of reads for certain regions of the genome. To correct for this amplification bias, we developed a coverage-based model to more accurately count the original amplicons that are generated from the gDNA template rather than those that are repeatedly amplified from the quasilinear amplification–generated products. Because the coverage-based method is not influenced by amplification duplicates, it reduced technical noise in estimations of gDNA counts over the entire genome (**Supplementary Note**). For the E14 cell line, we found that bin-to-bin technical variability for all the autosomes was twofold lower in our coverage-based method than in the conventional read-based method (**Fig. 2f**). Further analyses quantifying bin-to-bin technical variability and correlations between single cells revealed that the coverage-based method reduced amplification biases and technical noise, thereby improving copy-number calling in cancer genomes (**Supplementary Figs. 15–18**, **Supplementary Table 2** and **Supplementary Note**).

After making these improvements to reduce technical noise, we compared gDNA sequencing results from DR-Seq to results obtained from sequencing three E14 cells using MALBAC (multiple annealing and looping-based amplification cycles)[7], with all single cells sequenced at depths of 0.6–2.5×. To identify sequencing biases and differences in coverage, we used Lorenz plots to compare cumulative read depth to cumulative fraction of the genome covered, ordered by increasing coverage (**Fig. 2g**). Bulk gDNA sequencing, without the need for whole-genome amplification, achieves a read distribution close to the theoretical limit. DR-Seq and MALBAC, relying on quasilinear amplification using random primers to amplify the genome, show greater coverage biases than does bulk sequencing, but they perform similarly to each other (**Fig. 2g**). Furthermore, in assessments of systematic biases and drifts in read distribution along the length of the genome, power spectra showed that DR-Seq and MALBAC showed more bias over large genomic scales (i.e., low frequencies), with both methods performing similarly across the entire range of genomic scales (**Fig. 2h**). Finally, analysis of GC sequencing bias showed that regions of the genome with high and low GC content deviated from the expected normalized counts[25] (**Fig. 2i**). DR-Seq and MALBAC showed similar trends in GC bias, with DR-Seq showing modestly higher bias, possibly owing to the extra round of quasilinear amplification in DR-Seq. However, as the GC bias is easily corrected for, it does not influence the final gDNA analysis (**Supplementary Fig. 18**). Taken together, these results suggest that combined gDNA and mRNA sequencing from the same cell by DR-Seq performs similarly to existing methods for sequencing either the genome or transcriptome of single cells (**Fig. 2**).

We next applied DR-Seq to a breast cancer cell line (SK-BR-3) to understand how copy-number variations in single cancer cells influence gene expression programs. We applied DR-Seq to 21 SK-BR-3 cells and sequenced mRNA from 21 and gDNA from 7 of these cells. We detected 12,205 genes and, as with the E14 data set, found similar correlation between average expression of genes from these single cells and bulk mRNA sequencing (Pearson $r = 0.66$, Spearman $r = 0.69$) (**Supplementary Fig. 19a**,**b** and **Supplementary Table 1**). Similarly, detection of spike-ins correlated well with the expected numbers of molecules (**Supplementary Fig. 19c**,**d**). gDNA from the seven cells was sequenced at a depth of 0.6–1.6× (**Supplementary Table 3**). Sequencing coverage and GC bias were similar to that observed in single E14 cells (**Supplementary Fig. 20a–c**).

After correcting for GC bias, we used the circular binary segmentation (CBS) algorithm to detect breakpoints[26]. Raw data and breakpoint detection for chromosome 8 from one cell correlated well with copy-number changes detected in bulk sequencing (**Fig. 3a**). Similarly, breakpoint detection over the entire genome for all the single cells correlated well with the bulk sequencing results (**Supplementary Fig. 21**). The median read counts for each of the segments were used to estimate copy numbers in single cells (**Supplementary Figs. 18a** and **21**, **Supplementary Table 2** and **Supplementary Note**). We also developed a model to estimate confidence intervals for the copy numbers called by our algorithm (**Supplementary Fig. 22** and **Supplementary Note**). Further, the mean copy numbers over all single cells correlated well with the bulk sequencing copy numbers over the entire genome (**Supplementary Figs. 17** and **18b**,**c**). We also detected considerable cell-to-cell variability in copy numbers over certain regions of the genome (**Supplementary Figs. 17** and **23**). We performed DNA fluorescence *in situ* hybridization (FISH) over four genomic loci that span a large spectrum of copy numbers and found that the mean copy numbers detected by DR-Seq and DNA FISH were in agreement[27] (**Supplementary Fig. 24**). Notably, we also found that the distribution of copy numbers for these four loci in single cells amplified by DR-Seq were not statistically different from distributions obtained by DNA FISH ($P > 0.01$, Kolmogorov-Smirnov test; **Supplementary Table 4**). These results showed that DR-Seq has the sensitivity to capture heterogeneity in copy numbers across single cells.

Next, comparison of copy-number variations in chromosome 8 to levels of mRNA expression in this single cell showed that the average expression of genes within each segment appeared to be strongly correlated to the copy number of that genomic region (**Fig. 3a**). To quantify this correlation on a genome-wide scale, we calculated the mean expression of genes in different copy-number regions for each cell. We observed a monotonic increase in mean expression with increase in copy number on a genome-wide level across different single cells (**Fig. 3b** and **Supplementary Fig. 25**). This increase in expression with copy number provided additional validation that DR-Seq was sensitive enough to simultaneously detect changes in copy numbers and transcript counts from the same cell (**Supplementary Fig. 26**).

Finally, we investigated whether DNA copy-number variations within the cancer genome could be an important regulator of gene-expression variability. We found that genes that show more cell-to-cell variability in transcript numbers were generally associated with reduced copy-number loci and vice versa, implying that copy number variations could drive variability in gene expression between single cells (**Fig. 3c**, **Supplementary Figs. 27** and **28**, and **Supplementary Note**).

We have developed a method that allows combined gDNA and mRNA sequencing from the same cell using a single-pot strategy. Similar integrated strategies might be used in the future to determine the correlation between DNA methylation and transcription, or nucleosome positioning and transcription, in single cells. Additionally, integrated gDNA and mRNA single-cell sequencing might provide enhanced sensitivity to lineage-tracing studies in tumors and healthy tissue (**Supplementary Table 5**).

METHODS

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

Accession codes. Gene Expression Omnibus: sequencing data have been deposited under accession number GSE62952.

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

ACKNOWLEDGMENTS

We would like to thank C. Zong (Harvard University) for sharing details of MALBAC and D. Grün (Hubrecht Institute) for assistance in data analysis. We would also like to thank N. Crosetto (Hubrecht Institute) for assistance in preparing DNA FISH probes. This work was supported by the European Research Council Advanced grant (ERC-AdG 294325-GeneNoiseControl), the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award and by the Human Frontier Science Program (to M.B.).

AUTHOR CONTRIBUTIONS

S.S.D., L.K. and A.v.O. conceived the method. S.S.D. and L.K. performed experiments. M.B. performed DNA FISH. S.S.D. and B.S. analyzed the data. S.S.D., L.K., B.S. and A.v.O. wrote the manuscript. A.v.O. guided experiments and data analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
2. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
3. Keane, T.M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
4. Sheltzer, J.M., Torres, E.M., Dunham, M.J. & Amon, A. Transcriptional consequences of aneuploidy. *Proc. Natl. Acad. Sci. USA* **109**, 12644–12649 (2012).
5. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
6. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
7. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
8. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
9. Hou, Y. *et al.* Genome analyses of single human oocytes. *Cell* **155**, 1492–1506 (2013).
10. Evrony, G.D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
11. McConnell, M.J. *et al.* Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
12. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* **2**, 666–673 (2012).
14. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
15. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
16. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
17. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
18. Wu, A.R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
19. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
20. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
21. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
22. Junker, J.P. & van Oudenaarden, A. Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell* **157**, 8–11 (2014).
23. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
24. Baker, S.C. *et al.* The External RNA Controls Consortium. A progress report. *Nat. Methods* **2**, 731–734 (2005).
25. Zhang, C. *et al.* A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing. *PLoS ONE* **8**, e54236 (2013).
26. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
27. Bienko, M. *et al.* A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. *Nat. Methods* **10**, 122–124 (2013).

ONLINE METHODS

Tissue culture. E14 cells were cultured in DMEM (Gibco) supplemented with 15% FBS (Gibco), 2 mM GlutaMax (Gibco), 0.1 mM MEM nonessential amino acids, 0.1 mM β-mercaptoethanol (Sigma), 1% Pen/Strep (Gibco) and 1,000 U LIF/ml (ESGRO) on gelatinized petri dishes. SK-BR-3 cells were cultured in McCoy's 5a Medium Modified (ATCC) with 10% FBS and 1% Pen/Strep. Cells were grown at 37 °C and 5% CO_2.

Cell picking. Trypsinized single cells were picked using a mouth pipet with a 30-µm glass capillary under a stereomicroscope. Picked cells were deposited in the center of the lid of a 0.2 ml PCR tube and snap frozen in liquid nitrogen.

DR-Seq. First-strand cDNA synthesis was performed by addition of 2 µL of reaction mix containing 0.2 µL first strand buffer (MessageAmp II, Life technologies), 0.4 µL of dNTP mix (MessageAmp II, Life technologies), 0.1 µL Arrayscript (MessageAmp II, Life technologies), 0.1 µL RNAse inhibitor (MessageAmp II, Life technologies), 0.2 µL RT primer with cell specific barcode (Ad-1x)[13], 0.2 µL 1:500,000 diluted ERCC spike-in mix 1 (Life Technologies) and 0.05% IGEPAL in water. The first strand cDNA synthesis and lysis reaction mix together with the spike-in molecules were added directly to the drop in the lid of the tube containing a single cell. Samples were incubated in a PCR machine with lid and block set to 42 °C for 15 min after which the samples were spun down and incubated for another 105 min. After first strand synthesis, samples were incubated for 10 min at 80 °C. Quasilinear amplification buffer containing 6.0 µL ThermoPol buffer (NEB) 1.0 µL 10 mM dNTP mix, 26 µL water and 0.15 µL 50 µM primer mix (Ad-2)[7] was added to each sample. Samples were incubated for 3 min at 94 °C to denature the DNA. Seven cycles of quasilinear amplification was performed (10 °C for 45 s, 15 °C for 45 s, 20 °C for 45 s, 30 °C for 45 s, 40 °C for 45 s, 50 °C for 45 s, 65 °C for 2 min, 95 °C for 20 s, 58 °C for 40 s and then immediate quenching on ice). Prior to each cycle 0.6 µL polymerase mix containing 2 U Bst large fragment (NEB) and 0.8 U Pyrophage 3173 exo- (Lucigen) was added. The ste, before quenching the reaction on ice (58 °C for 40 s), is not performed for the first quasilinear amplification round. After seven rounds of quasilinear amplification, samples were split in two. One half of the sample was processed for gDNA sequencing, and the other half was processed for mRNA sequencing.

For mRNA sequencing, second strand synthesis of the quasilinear amplified cDNA was performed using the P1 primer (5′-CGATTGAGGCCGGTAATAC-3′) in a single cycle of PCR (94 °C for 20 s, 51 °C for 20 s, 72 °C for 7 min). After this, samples with nonoverlapping barcodes were pooled and cleaned up on a cDNA purification column (MessageAmp II, Life technologies), and eluted twice with 9 µL of water at 55 °C. Next, the volume of the sample was reduced to 6.4 µL using a SpeedVac. *In vitro* transcription (IVT) mix containing 1.6 µL 10× IVT buffer, 1.6 µL ATP, 1.6 µL GTP, 1.6 µL CTP, 1.6 µL UTP and 1.6 µL enzyme mix (MessageAmp II, Life technologies) was added to the samples and incubated at 37 °C for 13 h. After IVT, the aRNA was immediately cleaned up without fragmentation using the aRNA clean-up columns (MessageAmp II, Life Technologies) and the aRNA was eluted twice in 12 µL of warm water at 55 °C. After clean-up, aRNA quality was assessed on a bioanalyzer (Agilent) Eukaryote Total RNA Pico chip. Library preparation was performed as previously described[13].

For DNA sequencing, the other half of the quasilinear amplification product was amplified further by PCR. PCR mix containing 1.0 µL 10 mM dNTP, 3 µL Thermopol buffer (10×), 0.2 µL 100 µM primer P2 (5′ - GTGAGTGATGGTT GAGGTAGTGTGGAG - 3′) and 1.0 µL Deep Vent$_R$ (exo-) polymerase (NEB) was added to each sample for a final volume of 68 µL. PCR was performed as follows: 21 cycles of 94 °C for 20 s, 59 °C for 20 s, 65 °C for 1 min, 72 °C for 2 min; 72 °C for 5 min at the end. After PCR, the quality of the products was assessed by agarose gel electrophoresis and the samples were cleaned up using a PCR purification column (Qiagen) (**Supplementary Fig. 29**). Next, to remove adaptor Ad-2 from the PCR product before preparing Illumina libraries, another PCR was done starting with 80 ng of product from the previous step. PCR mix containing 0.3 µL of 50 µM primer P3 with a 5′ biotinylated end (5′ - GTGAGCTGGAGTTGAGGTAGTGTGGAG - 3′), 5 µL Thermopol buffer (10×), 1 µL 10mM dNTP and 1 µL Deep Vent$_R$ (exo-) polymerase (NEB) was added to each sample for a final volume of 50 µL. PCR was performed as follows: 94 °C for 2 min, then 4 cycles of 94 °C for 20 s, 46 °C for 20 s, 65 °C for 1 min and 72 °C for 2 min; 9 cycles of 94 °C for 20 s, 59 °C for 20 s, 65 °C for 1 min and 72 °C for 2 min. The PCR product was sheared using a sonicator (Biorupter) on the low power setting with 15 cycles of 1 min (30 s on, 30 s off) with constant cooling at 4 °C. The sheared products were then cleaned up using a PCR purification column (Qiagen) and eluted in 50 µL water. The final product distribution was verified on a bioanalyzer (Agilent) High Sensitivity DNA chip to have an average product size of approximately 300 bp. The DNA products were then added to Dynabeads MyOne Streptavidin C1 beads (Life Technologies) in 50 µL 2× BW buffer (10 mM Tris-HCl, 1mM EDTA and 2mM NaCl). After immobilizing the DNA products on the beads for 15 min, the biotinylated DNA was separated using a magnetic stand and the supernatant was stored. The biotinylated DNA was digested on the magnetic beads and the beads were washed twice with 50 µL 1× BW buffer. These two washes were then combined with the first supernatant and purified using a PCR purification column (Qiagen). Finally, Illumina libraries were prepared with different index primers for each single cell using the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB).

Applying DR-Seq to E14 and SK-BR-3 cells, the typical success rate in amplifying single cells was approximately 70% (21/30 for SK-BR-3 and 13/18 for E14 cells).

Libraries were sequenced on an Illumina Hi-seq 2500. cDNA libraries from DR-Seq were sequenced with 100 bp paired-end sequencing and the gDNA and other cDNA libraries (from CEL-Seq or bulk) were sequenced with 50 bp or 100 bp paired-end sequencing.

Bioinformatic analysis. For bulk mRNA and CEL-Seq libraries, paired-end sequencing reads were aligned to the transcriptome using Burrows-Wheeler Aligner (BWA) with default parameters. For single-cell mRNA processed using DR-Seq, the Ad-2 adaptor sequence was trimmed computationally from the right mate and then aligned to the transcriptome using BWA with default parameters. For the E14 cells, we used the RefSeq gene models based on the mouse genome release mm10. For the SK-BR-3 cells, we used the RefSeq gene models based on the human genome release hg19. For bulk mRNA sequencing, both mates of each read were mapped to the transcriptome. For CEL-Seq and DR-Seq, the right mate of each read pair was mapped to the transcriptome and the ERCC spike-ins. The left mate was used to identify the cell from which the transcript came based on the cell-specific barcode. Reads mapping to more than one region were distributed uniformly. For the bulk mRNA sequencing libraries, PCR duplicates were removed to obtain the data set used in all the analysis. The left mate of the CEL-Seq libraries also contained a random 4-bp sequence, introduced during reverse transcription, to count unique cDNA molecules, as previously described[21] (**Supplementary Fig. 6**). Length-based identifiers were determined for each read in the single-cell mRNA libraries processed by DR-Seq using the first coordinate of the right mate after trimming off adaptor Ad-2 (**Fig. 1b**). The length-based identifiers were used to minimize amplification biases and achieve resolution close to identifying unique cDNA molecules (**Fig. 2a–c**, **Supplementary Figs. 6–8** and **Supplementary Note**).

For bulk gDNA and MALBAC libraries, paired-end sequencing reads were aligned to the genome release mm10 for mouse cells (E14) and to the genome release hg19 for human cells (SK-BR-3) using BWA with default parameters. For the single-cell gDNA libraries processed by DR-Seq, paired-end sequencing reads were aligned to a masked genome mm10 for mouse cells and to a masked genome hg19 for human cells using BWA with default parameters. The masked genomes mm10 and hg19 were created by replacing all the coding sequences within the genome with 'N' because the fraction used to sequence gDNA contains sequences that could originate from the cDNA within coding regions (**Fig. 1a** and **Supplementary Fig. 13**). By masking the coding sequences within the genome, such ambiguous reads that might arise from either gDNA or cDNA are discarded computationally, leaving only reads that arise from gDNA. This does not pose a problem for calling copy number variations because the coding region constitutes only approximately 2% of the genome. Therefore, gDNA sequencing results obtained from DR-Seq can be used to quantify copy number variations and single nucleotide variants in the genome (**Fig. 3a**, **Supplementary Fig. 21** and **Supplementary Table 5**). Next, all PCR duplicates within mapped reads from the bulk, MALBAC or DR-Seq libraries are removed. As the first step toward quantifying the gDNA data, the genome is divided into bins. To account for the masking of the genome in the DR-Seq data, the start and end coordinates of each bin are chosen such that the length of all bins are the same after excluding coding regions within each bin. This variable binning strategy provides a more accurate description of the distribution of reads within each bin as reads that map to coding regions are masked from the analysis (**Supplementary Figs. 13** and **14**). Next, to further reduce amplification biases, we developed a coverage-based method to quantify the reads within bins. This coverage-based method significantly reduces bin-to-bin technical noise (**Supplementary Note**, **Fig. 2f** and **Supplementary Figs. 15** and **16**). The reads are then corrected for GC bias[25]. The corrected read distribution is then used to identify breakpoints using the circular binary segmentation (CBS) algorithm[26]. Finally, the median read counts for each segment are used to call copy number variations in single cells (**Supplementary Note**).

DNA FISH. SK-BR-3 cells were grown on coverslips and fixed in methanol/acetic acid solution (3:1 vol) for 10 min at room temperature (RT) upon reaching confluency. They were washed with PBS/0.1% Triton X-100 and treated with 100 µg/mL of RNAseA in PBS for 1 h at 37 °C. They were then washed with PBS and dehydrated with ethanol series (70% ethanol, 85% ethanol, 100% ethanol) followed by overnight air drying. The next day they were denatured in 70% formamide and 2× SSC buffer at 75 °C for 5 min, and then placed in 70% ethanol for 2 min, followed by a 2-min incubation in 85% ethanol and 2-min incubation in 100% ethanol. They were then air dried for 30 min, and during this time the probes were denatured at 75 °C for 5 min. The hybridization was set up with HD FISH probes resuspended in the CEP buffer (Abbott) (100 ng/20 µL). Coverslips were sealed on microscope slides with a rubber cement and incubated at 37 °C overnight. The next day the coverslips were removed from the slides and washed three times in 2× SSC followed by two washes in 0.2× SSC/0.2% Tween 20 at 56 °C for 7 min each. Afterwards they were rinsed with 4× SSC/0.2% Tween 20 and washed once with 2× SSC for 5 min. They were then incubated with 50 ng/mL DAPI and 2× SSC for 5 min at RT and mounted in the mounting solution containing 2× SSC, 10 mM Tris, 0.4% glucose, 100 µg/mL catalase, 37 µg/mL glucose oxidase, 2 mM Trolox. The probes were designed using the www.hdfish.eu database. For CCDC40, HTT and FHIT genes, the HD-FISH probes were prepared by PCR as previously described[27] (**Supplementary Fig. 24** and **Supplementary Table 4**). For *ZMIZ1*, 40-mer oligonucleotides with a 3′ functional amino group were synthesized by Biosearch Technologies Inc., and coupling to Cy5 (GE Healthcare) was performed in house.

doi:10.1038/nbt.3129