G&T-seq: parallel sequencing of single-cell genomes and transcriptomes

Iain C Macaulay[1], Wilfried Haerty[2,10], Parveen Kumar[3,10], Yang I Li[2,9], Tim Xiaoming Hu[2], Mabel J Teng[4], Mubeen Goolam[5], Nathalie Saurat[6], Paul Coupland[7], Lesley M Shirley[7], Miriam Smith[7], Niels Van der Aa[3], Ruby Banerjee[8], Peter D Ellis[7], Michael A Quail[7], Harold P Swerdlow[7,9], Magdalena Zernicka-Goetz[5], Frederick J Livesey[6], Chris P Ponting[1,2,11] & Thierry Voet[1,3,11]

The simultaneous sequencing of a single cell's genome and transcriptome offers a powerful means to dissect genetic variation and its effect on gene expression. Here we describe G&T-seq, a method for separating and sequencing genomic DNA and full-length mRNA from single cells. By applying G&T-seq to over 220 single cells from mice and humans, we discovered cellular properties that could not be inferred from DNA or RNA sequencing alone.

Single-cell genome sequencing is crucial for revealing genetic heterogeneity and cell-lineage relationships in normal and diseased tissue[1–4]. Single-cell transcriptome sequencing is equally important for defining cell types and states[5–10]. However, new methods for integrated DNA and RNA analyses are needed for studies of genotype-phenotype associations within single cells. Integrated methods can expose the diverse effects of genetic variation on transcript levels and isoforms and allow for the annotation of DNA-based cell-lineage trees with information on cell type and state from the same cells.

Here we introduce G&T-seq (genome and transcriptome sequencing), in which a single cell's polyadenylated (poly(A)) RNA is separated from its genomic DNA using a biotinylated oligo-dT primer in an adaptation of a previous method[11,12] and both the genome and the transcriptome are then amplified in parallel and sequenced (**Fig. 1a**). Prior to separation, External RNA Controls Consortium (ERCC) spike-in RNAs can be added to the lysis buffer to allow for assessment of the technical performance of RNA capture and amplification. The method has been automated on a conventional liquid-handling robotic platform, but it may also be performed manually. G&T-seq enables full-length whole-transcriptome analysis using a modified Smart-seq2 protocol[13,14] with an on-bead initiated first-strand cDNA synthesis (Online Methods) and separate whole-genome amplification (WGA) using a method of choice (Online Methods).

To benchmark G&T-seq, we used the breast cancer line HCC38 and the B lymphoblastoid line HCC38-BL, which are derived from the same patient[15] and were previously characterized by genome sequencing[16]. For both lines, we used flow cytometry to sort 86 single cells and eight multicell samples (duplicates of 5, 10, 20 and 50 cells) into 96-well plates and processed them using G&T-seq. Two wells per plate containing no cells were processed in parallel. We amplified the genomes of half of the samples using multiple-displacement amplification (MDA)[17] and those of the remaining samples using PicoPlex[18]. In total, 192 DNA and 192 RNA sequencing libraries were generated from single cells, multicell samples and negative controls. Of the 172 single cells analyzed in this experiment, 130 (75.6%) passed quality-control (QC; Online Methods) criteria for both WGA and whole-transcriptome amplification (WTA); in 61.9% of the samples that failed QC, both WTA and WGA criteria were unmet (**Supplementary Table 1**), which most likely indicates that no cell was sorted into the lysis buffer or that cell lysis was incomplete in those samples.

Low-coverage genome sequencing (0.036× ± 0.022× (mean ± s.d.) for PicoPlex, 0.13× ± 0.06× for MDA; **Supplementary Data 1**) and subsequent focal analyses of sequence-read depth yielded copy-number profiles from single cells and multicell controls that were highly concordant with those from unamplified bulk DNA (**Fig. 1b** and **Supplementary Fig. 1a**). DNA copy-number profiles derived using G&T-seq showed accuracy similar to that of profiles produced using conventional WGA performed in isolation (**Supplementary Fig. 1a**). As previously observed[3,19], PicoPlex amplification outperformed MDA in preserving copy-number concordance (**Fig. 1b** and **Supplementary Fig. 1a,b**) and was our method of choice for all further experiments in which copy number was assessed, whereas we preferred MDA for full-genome sequencing and detection of single-nucleotide variants (SNVs).

To investigate the breadth of genome coverage attainable by G&T-seq with MDA, we performed deep DNA sequencing on four single HCC38 cells and four single HCC38-BL cells using

[1]Sanger Institute–EBI Single-Cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, UK. [2]MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK. [3]Department of Human Genetics, University of Leuven, Leuven, Belgium. [4]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [5]Department of Physiology, Development and Neuroscience, Downing Site, University of Cambridge, Cambridge, UK. [6]Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. [7]Sequencing R&D, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [8]Cytogenetics Core Facility, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [9]Present addresses: Department of Genetics, Stanford University, Stanford, California, USA (Y.I.L.); New York Genome Center, New York, New York, USA (H.P.S.). [10]These authors contributed equally to this work. [11]These authors jointly directed this work. Correspondence should be addressed to I.C.M. (im2@sanger.ac.uk), C.P.P. (cp11@sanger.ac.uk) or T.V. (Thierry.Voet@med.kuleuven.be).

RECEIVED 18 NOVEMBER 2014; ACCEPTED 27 MARCH 2015; PUBLISHED ONLINE 27 APRIL 2015; DOI:10.1038/NMETH.3370

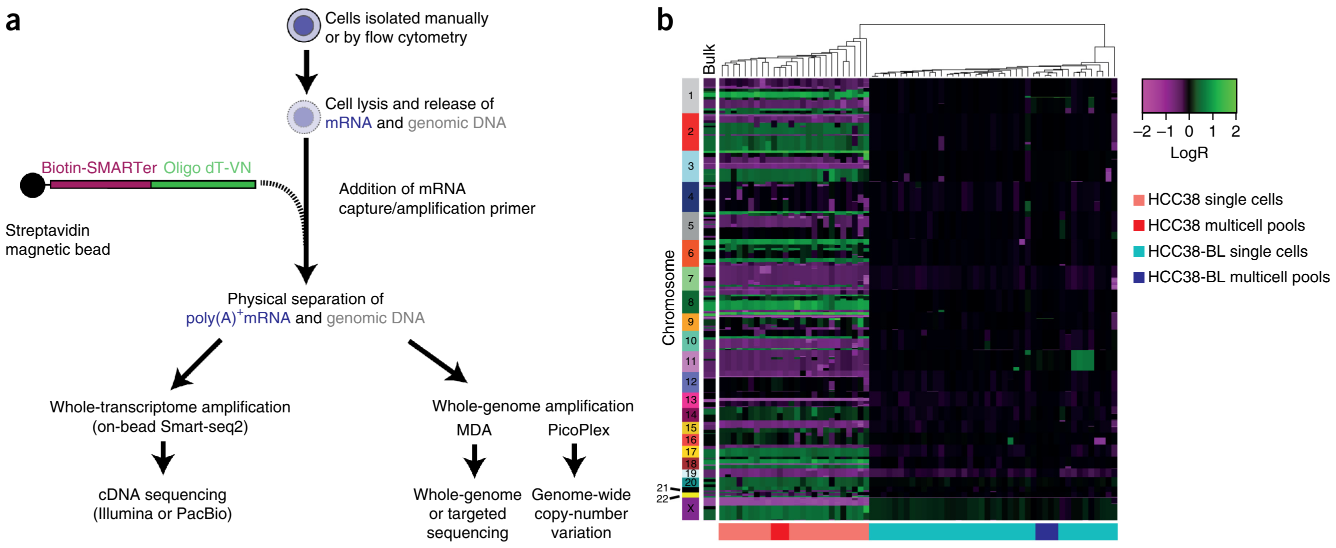

Figure 1 | G&T-seq enables integrated analysis of the genome and transcriptome of a single cell. (**a**) In G&T-seq, genomic DNA and poly(A)$^+$ mRNA are physically separated after cell lysis and then undergo separate amplification, library preparation and sequencing. Amplification of poly(A)$^+$ mRNA is done with a modified Smart-seq2 method, whereas DNA can be amplified by any method; here, MDA and PicoPlex were chosen to suit the downstream application. (**b**) Heat map of the genome-wide DNA copy numbers ($\log_2$ ratio (LogR)) in single cells and multicell controls isolated from HCC38 (*n* = 3 multicell controls, *n* = 23 single cells) and HCC38-BL (*n* = 4 multicell controls, *n* = 39 single cells) cell lines and amplified using PicoPlex. For reference, the copy-number profile derived from bulk HCC38 DNA (not subjected to WGA) is shown on the left. To cluster copy-number profiles, we applied the hclust R package using default parameters.

the HiSeq X platform. With this technique we captured up to 78.3% of genomic bases (67.2% ± 8.1% (mean ± s.d.)) at a depth of 33.3× per single cell (±0.9×; **Supplementary Table 2**). Although the method provided a similar breadth of genome coverage as sequences of conventional single-cell MDA performed in isolation, the coverage was less evenly distributed across the genome (**Supplementary Fig. 1c**). Additionally, G&T-seq showed GC-bias effects similar to those seen in conventional single-cell PicoPlex and MDA analyses (**Supplementary Fig. 1d**).

In parallel transcriptome analysis of the same 130 cells, we detected the expression of 4,000–11,000 transcripts per cell with a transcripts-per-million count greater than 1 (**Supplementary Fig. 2a**), with HCC38 cells expressing substantially more genes (9,725 ± 729) than HCC38-BL cells (6,126 ± 1,659). Both populations could be readily distinguished by principal-component analysis (**Supplementary Fig. 2b**) and by clustering cells by gene expression correlation (**Supplementary Fig. 2c**). The method faithfully preserved the distinct transcriptional profiles of these two cell types (**Supplementary Fig. 2d**). Read coverage was observed across the full transcript length, even up to 15 kb from the poly(A) tail (**Supplementary Fig. 3**).

In a direct comparison of G&T-seq to conventional single-cell Smart-seq2 performed in isolation, the detection of ERCC spike-ins, the number of genes expressed and coverage over transcript length were similar (**Supplementary Fig. 4a–d**). There was no discernible difference in the GC-content distributions of transcripts detected by the two methods (**Supplementary Fig. 4e**).

Using G&T-seq, we discovered a subclonal population of cells in the HCC38-BL line that contained a trisomy of chromosome 11 (10% of HCC38-BL cells; **Fig. 2a**), which was confirmed by FISH on separate cells from the same cell line (**Supplementary Fig. 5**). Furthermore, a loss (cells 56 and 79) or gain (cell 82) of the entire q-arm of chromosome 16 was observed by DNA sequencing in other HCC38-BL cells (**Fig. 2b**). The overall expression of genes on chromosome 11 in HCC38-BL cells carrying trisomy 11 was higher relative to the expression of the genes on the same chromosome in diploid cells (**Fig. 2c**). Also, the subchromosomal genomic imbalances of chromosome 16 were generally corroborated by the expected changes in gene expression in the transcriptomes of the same cells (cells 56, 79 and 82; **Fig. 2c**), although a further 16q 'gain' was observed in the transcriptome for cell 91. These data show that (sub)chromosomal copy number in a single cell is mostly positively correlated with gene expression in that cell.

To investigate whether chromosome-wide expression dosage is established after a chromosomal mis-segregation in a single cell division, we applied G&T-seq to all blastomeres of seven eight-cell cleavage-stage mouse embryos, five of which were treated with reversine[20,21] at the four-cell stage of *in vitro* culture to induce chromosome mis-segregation. After G&T-seq of individual blastomeres, DNA copy-number profiling revealed a diploid karyotype in untreated embryos (**Fig. 2d**), whereas reciprocal aneuploidies were observed in sister blastomeres of reversine-treated embryos (**Fig. 2e**). In those cells where chromosomal gains or losses (either reciprocal or nonreciprocal) were seen at the genomic level, we observed concomitant increases and decreases in chromosome-wide relative gene expression levels after G&T-seq analysis (**Fig. 2f** and **Supplementary Figs. 6–9**), which established for the first time (to our knowledge) that the effects of gene expression dosage can be rapidly established after the acquisition of aneuploidies during a single cell division.

We also analyzed neurons derived from anisogenic induced pluripotent stem cells (iPSCs) carrying trisomy 21 (*n* = 19; **Fig. 2g**) or not (*n* = 22; **Fig. 2h**). From the DNA-sequencing data, the trisomy was detected in 95% of cells tested (18 of 19 cells) and in 1 of the 22 control cells, although this cell manifested further

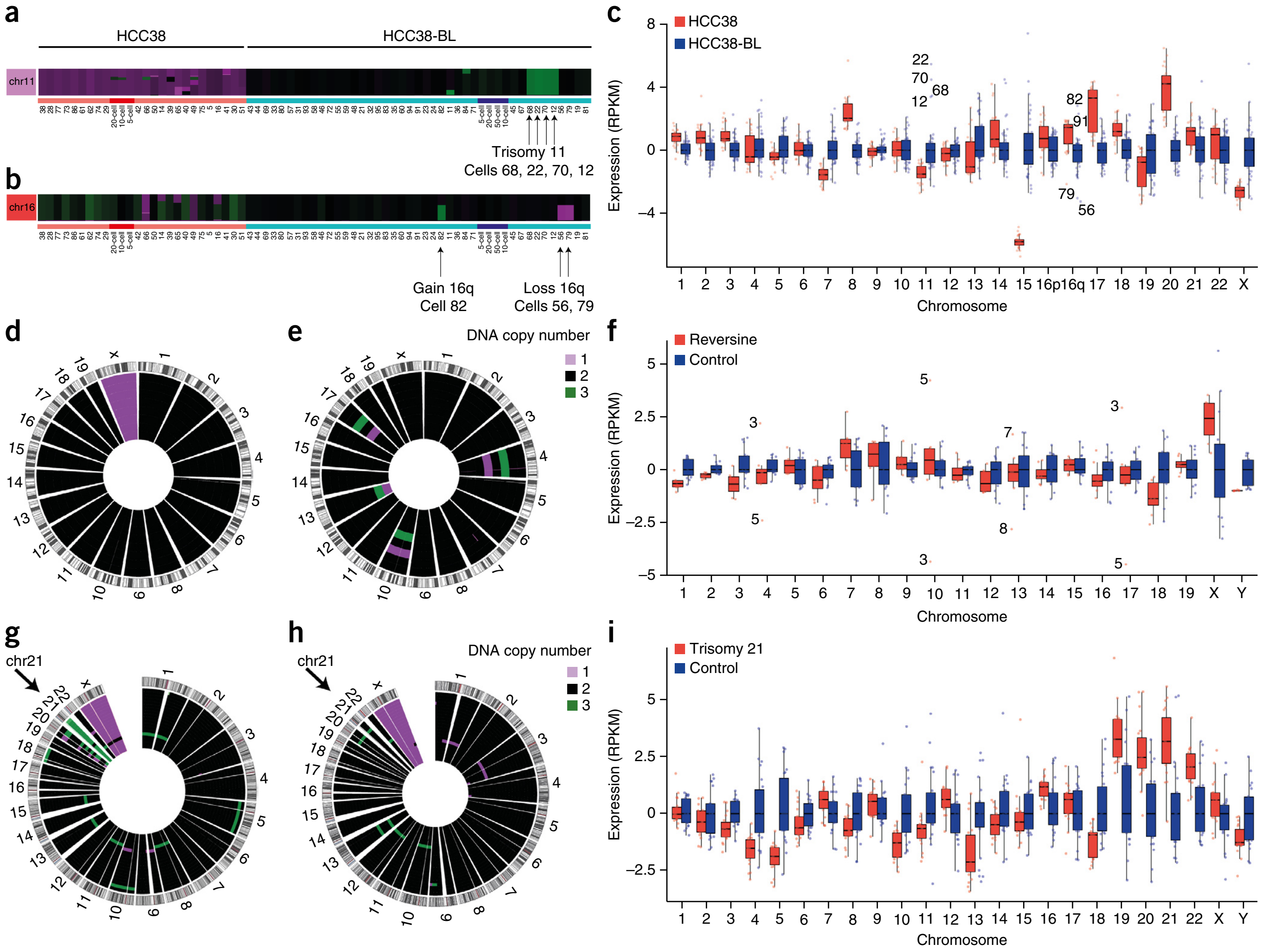

Figure 2 | Simultaneous detection of chromosomal aneuploidy and gene expression dosing in single cells. (**a**,**b**) Zoomed-in views of chromosomes 11 (**a**) and 16 (**b**) from **Figure 1b**. Cells containing trisomy 11 or partial loss or gain of the long arm of chromosome 16 are indicated with arrows. (**c**) Genome-wide gene expression binned per chromosome in single cells from the HCC38 (red, *n* = 23) and HCC38-BL (blue, *n* = 39) lines. Chromosomal RPKM values (chromosomal reads per kilobase of transcript per million reads mapped; Online Methods) for HCC38-BL cells were median centered, and HCC38 values were plotted relative to these values. Numbers adjacent to individual data points denote cell numbers mentioned in **a** and **b** or in the main text. (**d**,**e**) Single-cell DNA copy-number landscape of all cells from a control mouse embryo at the eight-cell stage (**d**) and a reversine-treated mouse embryo with reciprocal aneuploidies in sister blastomeres 3 + 5 and 7 + 8 (**e**). (**f**) Genome-wide expression per chromosome in the control (blue, *n* = 16 cells) and reversine-treated (red, *n* = 8 cells) embryos (RPKMs of the latter are relative to the median-centered control RPKMs). Numbers adjacent to individual data points denote blastomere numbers mentioned in **d** and **e**. (**g**,**h**) Single-cell DNA copy-number landscape of *in vitro*–differentiated neurons derived from trisomy 21 iPSCs (*n* = 19) (**g**) or control disomy 21 iPSCs (*n* = 22) (**h**). (**i**) Genome-wide expression per chromosome in the control (blue, *n* = 22) and trisomy 21 (red, *n* = 19) neurons. In **c**, **f** and **i**, the lower and upper boundaries of the boxes represent, respectively, the 25th and 75th percentiles, and bars denote the median. The whiskers represent the 5th and 95th percentiles.

chromosomal anomalies. Parallel RNA sequencing revealed elevated expression of chromosome 21 genes in the trisomic cells relative to the disomic cells (**Fig. 2i**). However, consistent chromosome-wide transcriptomic variation was also observed on other autosomes. This variation might reflect genome-wide effects of trisomy 21 on the regulation of gene expression[22], the different genetic backgrounds of the cell lines or marked alterations in chromatin organization in trisomy 21 neurons. In line with the genomically unstable nature of iPSC-derived neurons[23], further numerical and structural chromosomal aberrations were observed (**Fig. 2g,h**), including a recurrent chromosome 20p loss coupled with a chromosome 20q gain in the trisomy 21 line, for which we observed a concordant trend toward unbalanced expression between the chromosomal arms (**Supplementary Fig. 10**).

Fusion transcripts arising from chromosomal translocations often are implicated as driver mutations or serve as diagnostic markers in cancer[24,25]. We identified a fusion transcript, *MTAP-PCDH7* (**Supplementary Fig. 11a**), in 21% (9 out of 42) of the single HCC38 cells by RNA sequencing and confirmed expression by quantitative PCR (qPCR) in 81% (35 out of 42) of the same cells (**Supplementary Fig. 12**). This fusion has been characterized in another breast cancer cell line[26], but not in HCC38 cells[16]. Long-read sequencing on the Pacific Biosciences RSII showed the complete *MTAP-PCDH7* fusion transcript in three of the four single cells tested, which indicated that the transcript is a protein-coding fusion of exons 1–6 of *MTAP* and exons 3, 4 and 6 of *PCDH7* (**Supplementary Fig. 11b**). Deep sequencing, paired-end mapping and split-read analysis of the genomes of four HCC38 cells also identified the causative chromosomal rearrangement underlying the *MTAP-PCDH7* fusion in three cells (**Supplementary Fig. 11c**), which was further confirmed by qPCR in 60% of the HCC38 cells, or 71% of the *MTAP-PCDH7*–expressing cells (**Supplementary Fig. 12**).

Finally, we explored the ability of G&T-seq to enable the detection of SNVs in genomic DNA and mRNA from the same cell. By targeted resequencing of 365 cancer genes in the DNA of single HCC38-BL cells (n = 36) and single HCC38 cells (n = 32) amplified by MDA, we called 3,849 and 4,273 SNVs, respectively. Of these, 3,314 (86.1%) and 3,832 (89.6%), respectively, were concordant with the expected calls of bulk HCC38-BL and HCC38 DNA sequencing. For those concordant DNA variants across HCC38-BL and HCC38 cells, we detected 213 and 528 identical variants, respectively, in matching low-coverage single-cell RNA-sequencing data, representing 88.7% and 96.8% of all the concordant DNA variants that are present in transcribed regions.

G&T-seq complements the recently published DR-seq method (genomic DNA–mRNA sequencing)[27], which offers a different approach for analyzing the genome and transcriptome of a single cell in parallel. DR-seq begins with the pre-amplification of single-cell DNA and mRNA within a single tube, which is subsequently split for further independent amplification of the genomic DNA and cDNA. Because it amplifies DNA and mRNA without physical separation, DR-seq requires *in silico* masking of the exonic regions of the genome to determine DNA copy-number variation. Furthermore, the RNA sequence reads obtained from DR-seq are biased toward the 3′ end. In contrast, G&T-seq can be used to investigate the genome of a cell with any WGA method of choice, without the need to mask coding sequences during analysis, and it also provides access to full-length transcripts from the same cell.

We have shown that by sequencing the genome and transcriptome of a single cell in parallel, G&T-seq can readily distinguish the transcriptional consequences of chromosomal aneuploidies and interchromosomal fusions, and it has the potential to characterize coding SNVs at the single-cell level. The method is compatible with automation for high-throughput processing and, in combination with Illumina's HiSeq X Ten platform, allows for deep single-cell genome sequencing, enabling the detection of SNVs and chromosomal rearrangements in a cell at a cost approaching that of human-exome sequencing. The integrated analysis of a cell's transcriptome, genome and—eventually—epigenome will enable a more complete understanding of the extent, function and evolution of cellular heterogeneity in normal development and disease processes.

METHODS

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

Accession codes. Human data are available from the European Genome-phenome Archive (EGA) with accession number EGAS00001001204. Mouse data are available from ArrayExpress with accession number E-ERAD-381.

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

ACKNOWLEDGMENTS

We thank the Wellcome Trust Sanger Institute (UK) sequencing pipelines and F. Yang of the Cytogenetics Core Facility. This work was supported by the UK Wellcome Trust (to T.V. and C.P.P.) and funding from the Belgian Research Foundation Flanders (FWO) and the University of Leuven (KU Leuven, Belgium) to T.V. (FWO-G.0687.12; KU Leuven SymBioSys, PFV/10/016). N.V.d.A. is supported by an FWO scholarship (FWO-1.1.H28.12). W.H. and C.P.P. are funded by the UK Medical Research Council. L.M.S. was funded by the EU Seventh Framework Programme (FP7/2007-2013) under grant 262055. M.Z.-G. and the work in the lab are funded by the UK Wellcome Trust. M.G. is supported by a UK Mary Gray Studentship from St. John's College, Cambridge, UK. N.S. was supported by the New Zealand Woolf-Fisher Trust. F.J.L. is supported by a UK Wellcome Trust Senior Investigator award. M.J.T. is supported by a Wellcome Trust Sanger Institute Clinical Ph.D. Fellowship (UK). Y.I.L. was supported by a University of Oxford Nuffield Department of Medicine Prize Studentship, UK. Trisomy 21 iPSCs were obtained from the Harvard Stem Cell Institute (Cambridge, Massachusetts, USA), and control iPSCs were a gift from Y. Takashima (Cambridge Stem Cell Institute, Cambridge, UK).

AUTHOR CONTRIBUTIONS

I.C.M. developed the method, performed experiments, analyzed data and wrote the paper. W.H., P.K., Y.I.L. and T.X.H. analyzed data and prepared figures and text for the paper. M.J.T. performed experiments and assisted with method development. N.V.d.A. provided cells and assisted with method development. M.G. and M.Z.-G. provided mouse blastomeres. N.S. and F.J.L. provided iPSC-derived neurons. P.C., L.M.S., M.S., P.D.E., M.A.Q. and H.P.S. assisted with library preparation for targeted, HiSeq X and PacBio sequencing. R.B. performed cytogenetic analysis of cell lines. C.P.P. and T.V. acquired funding, oversaw the research, designed the method, analyzed data and wrote the paper. All authors read and approved the manuscript for submission.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Xu, X. *et al. Cell* **148**, 886–895 (2012).
2. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
3. Voet, T. *et al. Nucleic Acids Res.* **41**, 6119–6138 (2013).
4. Cai, X. *et al. Cell Rep.* **8**, 1280–1289 (2014).
5. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
6. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
7. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
8. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
9. Pollen, A.A. *et al. Nat. Biotechnol.* **32**, 1053–1058 (2014).
10. Shalek, A.K. *et al. Nature* **510**, 363–369 (2014).
11. Klein, C.A. *et al. Nat. Biotechnol.* **20**, 387–392 (2002).
12. Gužvić, M. *et al. Cancer Res.* **74**, 7383–7394 (2014).
13. Picelli, S. *et al. Nat. Methods* **10**, 1096–1098 (2013).
14. Picelli, S. *et al. Nat. Protoc.* **9**, 171–181 (2014).
15. Gazdar, A.F. *et al. Int. J. Cancer* **78**, 766–774 (1998).
16. Stephens, P.J. *et al. Nature* **462**, 1005–1010 (2009).
17. Dean, F.B. *et al. Proc. Natl. Acad. Sci. USA* **99**, 5261–5266 (2002).
18. Langmore, J.P. *Pharmacogenomics* **3**, 557–560 (2002).
19. de Bourcy, C.F. *et al. PLoS One* **9**, e105585 (2014).
20. D'Alise, A.M. *et al. Mol. Cancer Ther.* **7**, 1140–1149 (2008).
21. Santaguida, S., Tighe, A., D'Alise, A.M., Taylor, S.S. & Musacchio, A. *J. Cell Biol.* **190**, 73–87 (2010).
22. Letourneau, A. *et al. Nature* **508**, 345–350 (2014).
23. McConnell, M.J. *et al. Science* **342**, 632–637 (2013).
24. Mitelman, F., Johansson, B. & Mertens, F. *Nat. Rev. Cancer* **7**, 233–245 (2007).
25. Stratton, M.R., Campbell, P.J. & Futreal, P.A. *Nature* **458**, 719–724 (2009).
26. Ha, K.C. *et al. BMC Med. Genomics* **4**, 75 (2011).
27. Dey, S.S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. *Nat. Biotechnol.* **33**, 285–289 (2015).

ONLINE METHODS

Cell culture. HCC38 breast cancer cells (derived from subclone B8FF4C) were cultured as described[3]. HCC38-BL lymphoblastoid cells were cultured in RPMI-1640 (Life Technologies) supplemented with 10% fetal bovine serum (Life Technologies). The HCC38 and HCC38-BL cell lines are commercially available, were authenticated by genome sequencing and have been tested for mycoplasma.

Mouse embryo collection and culture. Animals were housed in the Gurdon Institute Animal Facility (Cambridge, UK). All experiments were conducted in compliance with UK Home Office regulations. F1 (C57BL/6 × CBA) females were superovulated by injection with 10 IU pregnant mare serum gonadotropin (Intervet) followed 48 h later by injection with 10 IU human chorionic gonadotropin (hCG; Intervet). These females were then mated with F1 males. Two-cell embryos, collected 48 h after hCG injection, were dissected out of oviducts in M2 medium supplemented with 4 mg/ml BSA. Embryos were cultured in drops of KSOM culture media supplemented with 4 mg/ml BSA under paraffin oil at 37.5 °C in 5% CO_2.

Reversine treatment. Embryos were cultured in KSOM until the late four-cell stage (56 h after hCG injection) and were treated with reversine (Cayman Chemicals) for 8 h during the transition from four to eight cells. Reversine was dissolved in dimethylsulfoxide (DMSO) (the final concentration of DMSO was 0.005%) and used at a concentration of 1 µM in KSOM. Embryos were incubated under paraffin oil at 37.5 °C in 5% CO_2 during the treatment period. Control embryos were incubated in an equivalent DMSO concentration but in the absence of reversine under otherwise identical conditions.

Culture of iPSCs. Trisomy 21 iPSCs were obtained from the Harvard Stem Cell Institute[28,29], and control iPSCs were a gift from Y. Takashima (Cambridge Stem Cell Institute)[30]. The iPSCs were cultured on mitomycin-treated mouse embryonic fibroblasts using standard protocols[31]. Pluripotent stem cells were differentiated into cortical neurons by dual SMAD inhibition in the presence of retinoids using described methods[30,32]. Following differentiation, cortical cultures were maintained for 80 d for the generation of mature neurons. Cultures were dissociated using trypsin and washed once in prewarmed neural-maintenance media. The cell suspension was diluted in Dulbecco's PBS, and a fine glass needle was used to aspirate individual cells.

Cell lysis, cDNA isolation and amplification. Single cells (or pools of multiple cells) were manually selected or sorted by flow cytometry into 2.5 µl of RLT Plus buffer (Qiagen) and were then processed immediately or stored at −80 °C. Individual wells were supplemented with 1 µl of a 1:250,000 dilution of ERCC spike-in mixture A (Life Technologies). Cells analyzed by conventional Smart-seq2 were processed as described by Picelli *et al.*[14]. Importantly, for this comparison of G&T-seq with conventional Smart-seq2, we used cells from the same cultures (HCC38 and HCC38-BL), sorted at the same time and processed, where possible, with the same batches of reagents. The same final concentration of ERCC spike-ins was added to G&T-seq and conventional Smart-seq2 reactions. All samples in this comparison were sequenced on the same lanes as parts of a multiplexed pool of libraries.

Genomic DNA and mRNA can be separated manually or by conventional liquid-handling robots for parallel processing of multiple single cells. All samples in this study were processed using a Biomek FXP Laboratory Automation Workstation (Beckman Coulter). A modified oligo-dT primer (5′-biotin-triethyleneglycol-AAGCAGTGGTATCAACGCAGAGTACT$_{30}$VN-3′, where V is either A, C or G, and N is any base; IDT) was conjugated to streptavidin-coupled magnetic beads (Dynabeads, Life Technologies) according to the manufacturer's instructions. To capture polyadenylated mRNA, we added the conjugated beads (10 µl) directly to the cell lysate and incubated them for 20 min at room temperature with mixing to prevent the beads from settling. The mRNA was then collected to the side of the well using a magnet, and the supernatant, containing the genomic DNA (gDNA), was transferred to a fresh plate. To maximize gDNA capture, we then washed the beads four times in a wash buffer consisting of 50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 0.5% Tween-20, 0.2× RNAse inhibitor (SUPERasin, Life Technologies) at room temperature. After each wash, the buffer was pooled with the original supernatant. To minimize sample loss, we used the same tips for all wash steps. Tips were washed with 10 µl wash buffer after supernatant collection, and this wash buffer was also transferred to the pooled supernatant and wash buffer.

Immediately after the last wash, 10 µl of a reverse-transcription mastermix (0.50 µl SuperScript II reverse transcriptase (200 U/µl, Life Technologies), 0.25 µl RNAse inhibitor (20 U/µl, Life Technologies), 2 µl Superscript II First-Strand Buffer (5×, Life Technologies), 0.25 µl DTT (100 mM, Life Technologies), 2 µl betaine (5 M, Sigma), 0.9 µl $MgCl_2$ (1 M, Life Technologies), 1 µl Template-Switching Oligo (5′-AAGCAGTGGTATCAACGCAGAGTACrGrG+G-3′, where "r" indicates a ribonucleic acid base and "+" indicates a locked nucleic acid base; 10 µM, Exiqon), 1 µl dNTP mix (10 mM, Thermo Scientific) and 3.6 µl nuclease-free water (Life Technologies)) were added to each well. Reverse transcription was performed with mixing on a Thermomixer (Eppendorf) for 60 min at 42 °C followed by 30 min at 50 °C and 10 min at 60 °C.

We then performed PCR immediately by adding PCR mastermix (12.5 µl KAPA HiFi HotStart ReadyMix with 0.25 µl PCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′, 10 mM)) to the 10 µL of reverse-transcription reaction mixture. The sample was then mixed and thermal cycled as follows: 98 °C for 3 min, then 18 cycles of 98 °C for 15 s, 67 °C for 20 s, 72 °C for 6 min and finally 72 °C for 5 min. Amplified cDNA was cleaned up using a 1:1 volumetric ratio of Ampure Beads (Beckman Coulter) and eluted into 25 µl of elution buffer (Buffer EB, Qiagen).

Genomic DNA precipitation and amplification. Genomic DNA present in the pooled supernatant and wash buffer from the mRNA-isolation step was precipitated on Ampure Beads (0.6 volumetric ratio, Beckman Coulter) and eluted directly into the reaction mixtures for amplification by either MDA (Genomiphi V2, GE Healthcare) or PicoPlex (New England BioLabs or Rubicon Genomics).

Amplified gDNA from either protocol was cleaned up using a 1:1 volumetric ratio of Ampure Beads (Beckman Coulter) and eluted into 25 µl of elution buffer (Buffer EB, Qiagen).

Library preparation and sequencing. Between 1 and 5 ng of amplified cDNA or gDNA was used for library preparation using the Nextera XT Kit (Illumina), per the manufacturer's instructions. Samples were barcoded during library preparation and multiplex sequenced on a HiSeq 2500 (Illumina) in fast mode or on a MiSeq.

For deep sequencing, MDA product from four single HCC38 cells and four single HCC38-BL cells was subjected to Illumina paired-end library construction and sequenced on the Illumina HiSeq X platform according to the manufacturer's instructions. For targeted sequencing, MDA products from single cells and multicell controls were sheared to 100–400 base pairs, subjected to standard Illumina paired-end library preparation and enriched using SureSelect target enrichment (Agilent) with a custom panel of 365 cancer-associated genes. Enriched libraries were pooled and sequenced on a HiSeq 2500 (Illumina).

Genomic read alignments. Reads resulting from Nextera library preparation and HiSeq 2500 sequencing were trimmed for 23 bases to remove adaptor sequence contamination and were subsequently aligned to the GRCh37 human reference genome (or mm10 for mouse) using BWA (version 0.6.2)[33]. SAI files were generated using default parameters, and subsequently SAM files were generated with Smith-Waterman for the unmapped mate disabled. The resulting BAM files were deprived of PCR duplicates using Picard (http://broadinstitute.github.io/picard/). HiSeq X data were aligned with BWA-MEM (http://bio-bwa.sourceforge.net/bwa.shtml). The genomic coverage was calculated with Bedtools (version 2.17)[34].

Estimation of genomic copy-number variation. For focal read-depth analysis, we first defined genomic bins by generating artificial reads equal in length to the single-cell trimmed reads from every base in the human genome and mapping them back to the reference genome using BWA (version 0.6.2)[35,36]. Subsequently, the human genome was divided into nonoverlapping bins of 500,000 uniquely mappable positions, resulting in physical bin sizes of 514 kb on average (s.d. = 28 kb when 1% of the top bins was removed). The uniquely mapped reads of the cells with a minimum quality of 30 were counted in these bins, a value of 1 was added to each bin's single-cell read count and bins with a %GC content of less than 28% were discarded. We then computed the $\log_2$ ratio (logR) per bin by dividing the read count of a given bin by the average read count of the bins genome-wide. The logR values were corrected for %GC bias using a Loess fit in R and were normalized according to the median of the genome-wide logR values. Corrected logR values were segmented using piecewise constant fitting (the penalty parameter, γ, was set to 15 for HCC38 and HCC38-BL samples or to 25 for iPSCs and mouse cells). The integer DNA copy number was estimated as $2^{\text{logR}} \times \Psi$, where the average ploidy of the cell, Ψ, was estimated based on the logR value of a large reference region with known DNA copy number without large copy-number aberrations. A similar approach was followed for copy-number profiling of (single-cell) mouse genomes; the average bin size was 546 kb (s.d. = 39 kb when 1% of the top bins was removed). For clustering of the copy-number profiles, we applied the hclust R package using default parameters.

QC filtering. The mean absolute pairwise difference (MAPD) measures the absolute difference between two consecutive %GC-corrected logR values across the genome and then computes the mean of these absolute differences. For human cells, we retained those samples having a MAPD of less than 0.6 for PicoPlex and less than 2 for MDA samples. The higher cutoff for MDA was chosen because of the greater noise in single-cell MDA data in general[3]. High MAPD values result from greater noise, which is characteristic of poor-quality samples. For mouse cells, we retained PicoPlex samples with a MAPD of 0.8 or less. Samples with less than 2% mapped reads or fewer than 3,500 transcripts detected (transcripts per million (TPM) > 1) were also excluded from downstream analysis.

Identification of genomic SNVs. MDA reads resulting from TruSeq library preparation were trimmed for six bases and were aligned to the human reference genome (GRCh37) using BWA. After duplicate-read removal using Picard, the BAM files were recalibrated and variants were called using GATK 3.1.1 software[37] and a minimum read coverage of 2.

Transcriptome read alignment. Adaptor sequences in reads were trimmed using Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Adaptors were removed using Cutadapt[38], and Tophat2 software[39] (using default settings) was used to align the reads onto the human genome assembly hg19 (mm10 for mouse), including the ERCC sequences. Expression measurements, expressed as TPM, were then calculated for every annotated protein-coding gene using RSEM[40]. Uniquely mapped reads were counted for each gene using HTSeq, and normalization across libraries was done using DESeq2 software[41].

Read-coverage profile over gene body. All genes having total exonic lengths greater than 2 kb, 10 kb and 15 kb were used for each of the three panels in **Supplementary Figure 3**, respectively. Read-coverage profiles for each of four regions (upstream of the transcription start site (TSS), concatenated exonic region, concatenated intronic region and downstream of the transcription termination site (TTS)) were obtained for all genes with sufficient total exonic lengths at single-nucleotide resolution, and the read-coverage profiles of each gene were aligned (including by inverting the profile for genes on the reverse strand) precisely at the polyadenylation tail for the exonic-region profile. Similarly, the profiles for 'upstream of TSS' were aligned at the TSS, the profiles for the 'intronic region' were aligned at the intronic nucleotide that was nearest to the polyadenylation tail and the profiles for 'downstream of TTS' were aligned at the TTS. After alignment, the read-coverage profiles were truncated to only the plotting length. To ensure that the aggregated profile was not dominated by a handful of extremely highly expressed genes, we obtained a single maximum for each gene across all four profiles, and we normalized all four profiles by dividing by that number to obtain the relative coverage. This was also done to ensure that the relative height of the four profiles for each gene was preserved and remained similar both before and after normalization and aggregation. The coverage profiles over the four regions were aggregated across all genes and all single HCC38 cells to form the final read-coverage profile over genes.

doi:10.1038/nmeth.3370

Single-cell differential expression analysis. To identify genes appropriate for sample clustering, we considered several TPM cutoffs for expression levels. We determined that a TPM cutoff of 1 was appropriate for clustering in at least 16 samples by assessing the number of protein-coding genes exceeding 0.1, 0.5, 1, 5 and 10 TPM (**Supplementary Fig. 2a**). The union of these genes was then used to compute the Spearman correlation between all sample pairs. To cluster samples, we used the command "heatmap.2" in the R package gplots, which uses the hierarchical clustering function hclust; "average linkage" was the option chosen to perform clustering.

To identify genes differentially expressed between HCC38 and HCC38-BL samples, we used a Bayesian approach to single-cell differential expression analysis[42]. All genes were then ranked in terms of the maximum-likelihood estimates of their differences in expression level. The TPM of each gene was normalized by the median of the TPM of that gene across all samples, and values are presented in heat maps as $\log_2$-fold differences from this median.

Whole-chromosome expression dosing. To assess the transcriptional consequences of copy-number variation, for each chromosome we calculated the number of chromosomal reads per kilobase of transcript per million reads mapped (RPKM) to reflect the number of reads mapping across a single composite coding sequence built using all coding sequences within the chromosome. For each chromosome, RPKM values were normalized according to the median expression for that chromosome in control cells (human HCC38-BL cells, human iPSC-derived neurons disomic for chromosome 21 or mouse blastomeres of control embryos).

Identification of fusion transcripts. For each cell, we identified candidate gene fusions using TopHat-Fusion[43] and Defuse[44] independently. Only fusions identified in multiple single cells and by both algorithms in the same cell were considered further.

Full-length transcript sequencing. The cDNA from four single cells was converted into SMRTbell libraries for PacBio RS II sequencing (Pacific Biosciences). Briefly, the double-stranded cDNA molecules were ligated with hairpin adaptors and loaded into a SMRTcell sequencing chip. Two SMRTcell wells were loaded per single-cell cDNA library. The PacBio reads were processed using the IsoSeq pipeline (Pacific Biosciences) and mapped onto the hg19 version of the human genome using blat[45]. After the removal of chimeric reads, only the best-scoring alignments for each read were considered further.

SNV calling from single-cell RNA-seq data. To identify SNVs from HCC38 single-cell RNA-Seq data, we implemented a pipeline that uses SNiPR[46]. SNVs were called in each sample separately. To estimate the number of false positive calls, we used variants called from bulk DNA sequencing of HCC38 samples as a gold-standard reference.

Code availability. Custom code is available upon request.

28. Park, I.H. *et al. Cell* **134**, 877–886 (2008).
29. Shi, Y. *et al. Sci. Transl. Med.* **4**, 124ra129 (2012).
30. Shi, Y., Kirwan, P., Smith, J., Robinson, H.P. & Livesey, F.J. *Nat. Neurosci.* **15**, 477–486, S471 (2012).
31. Chambers, S.M. *et al. Nat. Biotechnol.* **27**, 275–280 (2009).
32. Shi, Y., Kirwan, P. & Livesey, F.J. *Nat. Protoc.* **7**, 1836–1846 (2012).
33. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
34. Quinlan, A.R. & Hall, I.M. *Bioinformatics* **26**, 841–842 (2010).
35. Baslan, T. *et al. Nat. Protoc.* **7**, 1024–1041 (2012).
36. Møller, E.K. *et al. Front. Oncol.* **3**, 320 (2013).
37. DePristo, M.A. *et al. Nat. Genet.* **43**, 491–498 (2011).
38. Marcel, M. *EMBnet.journal* **17**, 10–12 (2011).
39. Trapnell, C. *et al. Nat. Protoc.* **7**, 562–578 (2012).
40. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. *Bioinformatics* **26**, 493–500 (2010).
41. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
42. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. *Nat. Methods* **11**, 740–742 (2014).
43. Kim, D. *et al. Genome Biol.* **14**, R36 (2013).
44. McPherson, A. *et al. PLoS Comput. Biol.* **7**, e1001138 (2011).
45. Kent, W.J. *Genome Res.* **12**, 656–664 (2002).
46. Piskol, R., Ramaswami, G. & Li, J.B. *Am. J. Hum. Genet.* **93**, 641–651 (2013).

doi:10.1038/nmeth.3370