G&T-seq: Separation and parallel sequencing of the genomes and transcriptomes of single cells

Iain C. Macaulay[1], Wilfried Haerty[2]*, Parveen Kumar[3]*, Yang I. Li[2†], 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‡], Magdalena Zernicka-Goetz[5], Frederick J. Livesey[6], Chris P. Ponting[1,2]^, Thierry Voet[1,3]^

* equal contribution

^ equal contribution

Affiliations:

[1]Sanger Institute-EBI Single-Cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

[2]MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX, UK.

[3]Department of Human Genetics, University of Leuven, KU Leuven, Leuven, 3000 Belgium

[4]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

[5]Department of Physiology, Development and Neuroscience, Downing Site, University of Cambridge, Cambridge, CB2 3DY, UK.

[6]Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, CB2 1QN, UK

[7]Sequencing R&D, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

[8]Cytogenetics core facility, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

[†]Present address: Department of Genetics, Stanford University, Stanford, CA, 94305, USA

[‡]Present address: New York Genome Center, 101 Ave. of the Americas 7th Fl., New York, NY, 10013, USA

Abstract:

The ability to simultaneously sequence the genome and transcriptome of the same single cell offers a powerful means to dissect functional genetic heterogeneity at the cellular level. 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 we reveal cellular properties that cannot be inferred from DNA or RNA sequencing alone, including associations between DNA copy number variation and gene expression dosage. We further demonstrate the detection of coding inter-chromosomal fusions and single nucleotide variants in both the genomes and transcriptomes of individual cells. G&T-seq enables the study of genotype-phenotype associations in single cells, and the investigation of DNA cell lineage trees of healthy and diseased tissues with transcriptome-inferred cell types and states.

Genome sequencing of single cells is crucial for revealing the genetic heterogeneity and cell lineage relationships within an organism or tissue in health and diseased states[1-5]. Single cell transcriptome sequencing is equally important for defining cell type and cell status using gene expression profiles[6-13]. However, new methods for integrated DNA and RNA analyses of the same cell are needed to enable the study of genotype-phenotype associations within single cells. This will reveal the diverse consequences of genetic variation and the degrees of gene expression heterogeneity−in terms of transcript levels and isoforms−arising from genetic variation, and will allow the annotation of DNA cell lineage trees with cell types and states inferred from full transcriptome information from the same cells.

Here we introduce G&T-seq (Genome & Transcriptome sequencing), in which a single cell's polyadenylated (polyA) RNA is first separated from genomic DNA using a biotinylated oligo-dT primer in an adaptation of the method applied by Klein *et al.*[14,15], then both genome and transcriptome are amplified in parallel and sequenced (Figure 1A). Prior to separation, External RNA Controls Consortium (ERCC) spike-in RNAs can be added to the lysis buffer in order to assess the technical performance of RNA capture and amplification. The method has been automated on a conventional liquid handling robotic platform, but may also be performed manually. The method allows full-length whole transcriptome analysis with an on-bead initiated modification of the Smart-seq2 protocol[16,17] and separate whole genome amplification (WGA) using an amplification method of choice, in this instance PicoPlex[18] or Multiple Displacement Amplification (MDA)[19].

The breast cancer cell line HCC38 and B-lymphoblastoid cell line HCC38-BL, derived from the same patient[20] and previously characterized by genome sequencing[21], were used to benchmark G&T-seq and to demonstrate its power for comparing genomic and transcriptional variation between these two different, yet related, cell lines. For both HCC38 and HCC38-BL lines, 86 single cells as well as 8 multi-cell samples (duplicates of 5, 10, 20 and 50 cells) of each cell type were FACS sorted into 96 well plates and processed using G&T-seq. Negative control wells (n = 2 per plate) into which no cell was sorted, were processed in parallel. The genomes of half of the samples were amplified using MDA, while those of the remaining half were amplified using PicoPlex. In total, 192 DNA and 192 RNA sequencing libraries were generated from single cells, multi-cell samples and negative controls. Of the 172 single cells analysed in this experiment 130 (75.6%) passed quality control (QC, see Methods) criteria for both WGA and WTA; in 61.9% of the cells failing QC, both WTA and WGA criteria were unmet (Supplementary Table 1), most likely indicating that no cell was sorted into the lysis buffer or that cell lysis was incomplete.

First pass sequencing of the single cell genomes at low coverage (0.036x ± 0.022x SD for PicoPlex; 0.13x ± 0.06x SD for MDA; Supplementary File 1) and subsequent focal sequence read depth analyses delivered copy number profiles of the single cells and multi-cell controls that were highly concordant with those observed from the bulk (non-WGA) DNA even for a highly rearranged genome such as HCC38 (Figure 1B, Supplementary Figure 1A). DNA copy number profiles derived using G&T-seq showed similar accuracy to those produced using conventional WGA performed in isolation (Supplementary Figure 1A). Similarly, the normal female HCC38-BL cells generally displayed a copy number of two across the autosomes and X-chromosome (Figure 1B). As previously observed[2,22] PicoPlex amplification outperformed MDA in preserving copy number concordance (Figure 1B, Supplementary Figure 1A,B), and was selected as the method of choice for all further experiments in which copy number was assessed, while MDA was preferred for full-genome sequencing and single nucleotide variant detection. To investigate the breadth of genome coverage attainable by the method, we performed deep DNA sequencing on 4 single HCC38 and 4 single HCC38-BL cells (MDA amplified) using the HiSeq X platform, and attained a mean sequencing depth of 33.3x per cell (± 0.9x SD). With this coverage the method captured up to 78.3% of genomic bases (mean 67.2% ± 8.1% SD) per single cell (Supplementary Table 2). Although the method reached a similar breadth of genome coverage as sequences of conventional single-cell MDA performed in isolation, the coverage is less evenly distributed across the genome (Supplementary Figure 1C). Additionally, G&T-seq PicoPlex and MDA sequences showed similar performance to conventional single-cell PicoPlex and MDA analyses across regions of different GC content (Supplementary Figure 1D).

Parallel analysis of the transcriptomes of the same cells detected the expression of 4,000-11,000 transcripts per cell with a transcript per million (TPM) count greater than 1 (Supplementary Figure 2A), with HCC38 cells expressing substantially higher numbers of genes (mean 9725 ± 729 SD) than HCC38-BL (mean 6126 ± 1659 SD). Both populations are readily distinguishable by Principal Component Analysis (Supplementary Figure 2B) and when clustering cells by gene expression correlation (Supplementary Figure 2C). The method faithfully preserves the distinct transcriptional profiles of these two cell types (Figure 1C, Supplementary Figure 2D). Read coverage is observed across the full transcript length, even up to 15 kb from the polyA tail (Supplementary Figure 3).

In a direct comparison of G&T-seq to conventional single-cell Smart-seq2 performed in isolation, G&T-seq captured the expression of more genes than Smart-seq2 in both HCC38 and HCC38-BL (25.8% and 24.4% more, respectively; Supplementary Figure 4A), showed similar detection of ERCC spike-ins confirming that the relative abundance of transcripts is preserved in both methods (Supplementary Figure 4B, C) and improved relative coverage over transcript length (Supplementary Figure 4D). There was no discernible difference in the GC content distribution of transcripts detected by both methods (Supplementary Figure 4E). It is possible that the different lysis and reverse transcription conditions in G&T-seq contribute to improved stability of the mRNA before reverse transcription, and improvements in single cell cDNA synthesis have previously been observed following bead capture of mRNA[23].

Interestingly, following G&T-seq we observed a subclonal population of single cells in the HCC38-BL line containing a trisomy of chromosome 11 (10% of HCC38-BL cells [4/40], Figure 2A), which was confirmed subsequently by interphase FISH analysis on separate cells of the same cell line (Supplementary Figure 5). Furthermore, a loss (cells 56 and 79) and gain (cell 82) of the entire q-arm of chromosome 16 was observed by DNA sequencing in other HCC38-BL cells (Figure 2B). The overall expression of genes on chromosome 11 in HCC38-BL cells carrying the trisomy 11 was higher relative to the expression of the genes on the same chromosome in the diploid cells (Figure 2C). Also, the sub-chromosomal genomic imbalances of chromosome 16 were generally corroborated by the expected gene expression changes in the transcriptomes of the same cells (cells 56, 79 and 82; Figure 2C) although a further 16p "gain" was observed from the transcriptome for cell 91. These data show that (sub)-chromosomal copy number in single cells are mostly positively correlated with gene expression in the same cell.

To investigate whether chromosome-wide expression dosage is established after a chromosomal missegregation in a single cell division, we applied G&T-seq to all blastomeres of seven 8-cell cleavage stage embryos, five of which were treated at the 4-cell stage of *in vitro* culture with reversine, an inhibitor of Aurora kinase[24] and monopolar spindle 1 (MPS1)[25], to induce chromosome missegregation. Following G&T-seq of individual blastomeres, DNA copy number profiling revealed a diploid karyotype in untreated embryos (Figure 2D) whereas reciprocal aneuploidies were observed in sister blastomeres of reversine treated embryos (Figure 2E). In those cells where chromosomal gains or losses (either reciprocal or non-reciprocal) were seen at the genomic level, we observed concomitant increases and decreases in chromosome-wide relative gene expression levels following G&T-seq analysis (Figure 2F, Supplementary Figures 6-9), which establishes for the first time that gene expression dosage effects can be rapidly established following acquisition of aneuploidies during a single cell division.

To further explore the power of G&T-seq to detect chromosomal aneuploidies in parallel with whole transcriptome expression profiling, we analysed neurons derived from anisogenic induced pluripotent stem cells (iPSCs) carrying a trisomy 21 (n = 19, Figure 2H) or not (n = 22, Figure 2G). From the DNA sequencing data, the trisomy 21 was detected in 95% of cells tested (18/19 cells), and in one of the 22 control cells which, however, manifested further chromosomal anomalies. Parallel RNA sequencing revealed elevated expression of chromosome 21 genes in the trisomic over the disomic cells (Figure 2I). However, consistent chromosome-wide transcriptomic variation was also observed on other autosomes. This variation may reflect genome-wide consequences of the trisomy 21 in the regulation of gene expression[26], the different genetic backgrounds of the cell lines or marked alterations in chromatin organisation in trisomy 21 neurons. In line with the genomically unstable nature of iPSC derived neurons[27], further numerical and structural chromosomal aberrations were observed (Figure 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 towards unbalanced expression between the chromosomal arms (Supplementary Figure 10).

Fusion transcripts arising from chromosomal translocations are often implicated as driver mutations or serve as diagnostic markers in cancer[28,29]. The parallel identification of fusion transcripts and their causative genomic rearrangements in single cells would be a powerful tool when interpreting the functional consequences of genomic rearrangements during the emergence and clonal evolution of cancer. Out of the 4 previously annotated fusions in this cell line we detected 3 (*RRP15-ACBD6*, *MBOAT2-PRKCE*, *SLC26A6-PRKAR2A*)[21] across the low-depth single-cell transcriptomes. Interestingly, we identified in addition a novel fusion transcript, *MTAP-PCDH7* (Figure 3A), in 21% (9/42) of the HCC38 single cells by RNA-sequencing and confirmed expression by qPCR in 81% (35/42) of the cells (Supplementary Figure 11). This fusion has previously been characterised in another breast cancer cell line[30] but not in HCC38[21]. We next took advantage of the availability of full-length cDNA from these single cells and by long-read sequencing on the Pacific Biosciences RSII obtained the complete *MTAP-PCDH7* fusion transcript in 3 of the 4 single cells tested, indicating that the transcript is a protein-coding fusion of exons 1-6 of *MTAP* and 3, 4, and 6 of *PCDH7* (Figure 3B). Deep sequencing, paired-end mapping and split-read analysis of the genomes of four HCC38 cells identified also the causative chromosomal rearrangement underlying the *MTAP-PCDH7* fusion in three cells (Figure 3C), which was further confirmed by qPCR in 60% of the HCC38 cells (or 71% of the *MTAP-PCDH7* expressing cells, Supplementary Figure 11). Parallel genome and transcriptome sequencing thus offers a means by which the transcriptional consequences of genomic rearrangements can be observed in single cells.

Finally, we explored the potential of G&T-seq to enable the detection of single nucleotide variants (SNVs) in genomic DNA and mRNA from the same single cell. By targeted re-sequencing of 365 cancer genes in the DNA of HCC38-BL single cells (n = 36) and HCC38 single cells (n = 32), all amplified with MDA, we identified 3849 and 4273 SNV calls, respectively. Of these, 3314 (86.1%) and 3832 (89.6%) were concordant with the expected call of bulk HCC38-BL and HCC38 DNA sequencing. For those concordant DNA-variants across HCC38-BL and HCC38 cells, we subsequently investigated the matching, but low coverage, RNA-sequencing and detected 213 variant calls (or 88.7% out of 240 single-cell HCC38-BL DNA-variants covered in the RNA) and 528 (or 96.8% out of 545 single-cell HCC38 DNA-variants covered in the RNA) in the RNA that are identical to the DNA variant, respectively.

G&T-seq complements the recently published DR-seq approach[31], which offers a methodologically different approach to analyse the genome and transcriptome of a single cell in parallel. DR-seq first uses a pre-amplification of the DNA and polyadenylated mRNA of a cell within a single tube, which is subsequently split to allow further amplification of both the genome using a PCR-based assay and the cDNA using *in vitro* transcription (IVT) followed by a second reverse transcription and PCR. DR-seq thus initiates the amplification of the DNA and mRNA of a cell without their physical separation. Consequently, it 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 strongly biased to the 3' end, as expected from the modified CEL-Seq method[10], and thus largely prohibit the detection of splicing isoforms, fusion genes and expressed coding SNVs. In contrast, G&T-seq investigates the genome of a cell using a WGA-method of choice without having to mask coding sequences during analysis, and additionally provides access to full-length transcripts from the same cell.

In conclusion, by parallel sequencing of the genome and transcriptome of a single cell, G&T-seq enables multiparameter sequencing of single cells. The method is compatible with automation for high-throughput processing. We have demonstrated that the method can readily distinguish the transcriptional consequences of chromosomal aneuploidies and inter-chromosomal fusions and offers potential to characterise coding SNVs at the single cell level, opening up novel avenues for the exploration of somatic genomic variation. Beyond the shallow sequencing required for the detection of full chromosome aneuploidies and their transcriptional consequences, we show that 'deep' single-cell genome sequences can also be obtained using Illumina's HiSeq X Ten platform, enabling the detection of SNVs and chromosomal rearrangements in a cell at a cost approaching that of current human exome targeted DNA-sequencing. The integrated analysis of the cell's transcriptome, genome - and eventually epigenome - will allow a more complete understanding of the extent, function and evolution of cellular heterogeneity in normal development and disease processes.

Contributions:

ICM developed the method, performed experiments, analysed data and wrote the paper. WH, PK, YL, TXH analysed data and prepared figures and text for the paper. MJT performed experiments and assisted with method development. MG and MZG provided mouse blastomeres. NS and FJL provided iPSC derived neurons. PC, LMS, MS, PDE, MAQ and HPS assisted with library preparation for targeted, HiSeq X and PacBio sequencing. RB performed cytogenetic analysis of cell lines. CPP and TV acquired the funding, oversaw the research, designed the method, analysed data and wrote the paper. All authors have read and approved the manuscript for submission.

Competing Financial Interests:

The authors declare no competing financial interests.

Corresponding Authors:

Correspondence to Iain Macaulay (im2@sanger.ac.uk), Chris P. Ponting (cp11@sanger.ac.uk), Thierry Voet (Thierry.Voet@med.kuleuven.be).

Acknowledgements:

We thank the Wellcome Trust Sanger Institute (WTSI) sequencing pipelines for processing of sequencing samples. This work was supported by the Wellcome Trust (to TV & CPP) and funding from the Research Foundation Flanders (FWO) and the KU Leuven to TV [FWO-G.0687.12; KU Leuven SymBioSys, PFV/10/016]. NVdA is supported via FWO scholarship [FWO-1.1.H28.12]. WH and CPP are also funded by the Medical Research Council. LMS was funded by the Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 262055. MZG and the work in the lab is funded by the Wellcome Trust and MG is supported by Mary Gray Studentship from St. John's College, Cambridge. NS was supported by the Woolf-Fisher Trust (NZ), FJL is supported by a Wellcome Trust Senior Investigator award.

Methods:

Cell culture:

HCC38 (derived from subclone B8FF4C) cells were cultured as described previously[2]. HCC38-BL cells were cultured in RPMI-1640 (Life Technologies) supplemented with 10% Fetal Bovine Serum (Life Technologies).

Murine embryo collection and culture

Animals were maintained in the Gurdon Institute Animal Facility (Cambridge, UK). All experiments were conducted in compliance with Home Office regulations. F1 (C57BL6xCBA) females were superovulated by injection with 10IU of pregnant mare's serum gonadotropin (PMS, Intervet, USA) followed by 10 IU of human chorionic gonadotropin (hCG, Intervet, USA) 48 hours later. These females were then mated with F1 males. 2-cell embryos, collected 48 hours after hCG injection, were dissected out of oviducts in M2 medium supplemented with 4mg/ml BSA. Embryos were cultured in drops of KSOM supplemented with 4mg/ml BSA under paraffin oil at 37.5°C in 5% CO_2.

Reversine treatment

Embryos were cultured in KSOM until the late 4-cell stage (56 hours post hCG) when they were treated with reversine (Cayman chemicals, USA) for 8 hours during the 4-8 cell transition. Reversine was dissolved in DMSO (final concentration of DMSO was 0.005%) and used at a concentration of 1μM in KSOM. Embryos were incubated under paraffin oil at 37.5°C in a 5% CO_2 during the period of treatment. Control embryos were incubated in the equivalent DMSO concentration but in the absence of reversine under the same conditions.

Culture of Trisomy 21 and control iPSCs

Induced pluripotent stem cells were cultured on mitomycin treated mouse embryonic fibroblasts according to standard protocols[32]. Trisomy 21 iPSCs were obtained from the Harvard Stem Cell

Institute[33,34] and control iPSCs were a gift from Y. Takashima (Cambridge Stem Cell Institute)[35] Pluripotent stem cells were differentiated into cortical neurons by dual SMAD inhibition in the presence of retinoids according to previously described methods[35,36]. Following differentiation, cortical cultures were maintained for eighty days to allow the full complement of mature neurons to be generated. Cultures were dissociated using Trypsin and washed once in pre-warmed 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 where specified) were manually picked or FACS sorted into 2.5 µl of RLT Plus buffer (Qiagen) and 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 analysed by the Smart-seq2 were processed as described in Picelli *et al.*[16] with the same final dilution of ERCCs added to each reaction.

The separation of genomic DNA and mRNA can be performed manually or using 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- AAG CAG TGG TAT CAA CGC AGA GTA $CT_{30}VN$-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, the conjugated beads (10 µl) were then added directly to the cell lysate and incubated for 20 min with mixing to prevent 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), transferred to a fresh plate. To maximise gDNA capture, the beads were then washed 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 x RNAse inhibitor (SUPERasin, Life Technologies) at room temperature. After each wash, the buffer was pooled with the original supernatant. To minimise sample loss, the same tips were used for all wash steps, and tips were washed with 10 µl wash buffer after supernatant collection; this wash buffer was also transferred to the supernatant/wash pool.

Immediately following 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 x, 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' -AAG CAG TGG TAT CAA CGC AGA GTA CrGrG+G-3', where the prefix "r" indicates a ribonucleic acid base, while the prefix "+" indicates a locked nucleic acid (LNA) base ; 10 µM, Exiqon), 1 µl dNTP mix (10mM, Thermo Scientific) and 3.6 µl nuclease-free water (Life Technologies)) was added to each well. Reverse transcription was performed with mixing on a Thermomixer (Eppendorf) at 42 °C for 60 min followed by 30 min at 50 °C and 10 min at 60 °C.

PCR was then performed immediately by adding PCR mastermix (12.5 ul KAPA HiFi HotStart ReadyMix with 0.25 µl PCR primer (5' -AAG CAG TGG TAT CAA CGC AGA GT-3', 10 mM)) to the 10 uL RT reaction mixture. The sample was then mixed and the following thermal cycling protocol used: 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 (EB, Qiagen).

Genomic DNA precipitation and amplification:

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

Library preparation and sequencing:

Between 1 and 5 ng of amplified cDNA or gDNA were used as input for library preparation using the Nextera XT kit (Illumina), as per the manufacturer's instructions. Samples were individually barcoded during library preparation, and subsequently pooled for multiplexed sequencing on a HiSeq 2500 (Illumina) run in fast mode or a MiSeq.

For deep sequencing WGA from 4 HCC38 and 4 HCC38-BL single cells was subjected to standard Illumina paired end library construction and sequenced on the Illumina HiSeq X platform according to the manufacturer's instructions. For targeted sequencing, WGA product from single cells and multi-cell controls were sheared to 100-400 bp, subjected to standard Illumina paired-end library preparation and enriched using SureSelect target enrichment (Agilent) using a custom panel of 365 cancer associated genes. Enriched libraries were pooled and sequenced on a Hiseq 2500 (Illumina) according to the manufacturer's protocol.

Genomic read alignments

Sequencing reads were first investigated at the 5'end for adaptor contamination. PicoPlex and MDA reads resulting from Nextera library preparation were trimmed for 23 bases to remove adapter sequence contamination, and were subsequently aligned to GRCh37 human reference genome (or mm10 for mouse) using BWA (version 0.6.2)[37]. 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 further processed by removal of PCR duplicates and genomic coverage was calculated using Picard (http://picard.sourceforge.net/) and Bedtools respectively.

Estimation of genomic copy number variation

Single-cell copy number analysis was performed as described previously[38]. Briefly, for focal read depth analysis, genomic bins were first defined by generating artificial reads of a length equal to the single-cell reads after trimming from every base in the human genome and mapping them back to the reference genome using BWA[39]. Reads mapping at multiple loci were discarded resulting in 'uniquely' mappable positions. Subsequently, the human genome was divided into non-overlapping bins of 500kb unique positions resulting in physical bin sizes of 514 kb on average (28 kb SD, when 1% of the top bins were removed). The uniquely mapped reads of the cell with minimum quality of 30 were counted in these bins and to each bin's single-cell read-count a value of one was added, and bins with a %GC-content of less than 28% were discarded. The $\log_2$ ratio (LogR) values for these non-overlapping variable bins were subsequently computed 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 PCF (the penalty parameter, γ, was set to 15 for HCC38 and HCC38-BL samples or 25 for iPSC and murine induced aneuploidy cells). Integer DNA-copy number was estimated as $2^{\mathrm{logR}}\times\Psi$, where the average ploidy, Ψ, of the cell 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; average bin size 546kb (39 kb SD, when 1% of the top bins were 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 below 0.6 for PicoPlex, and below 2 for MDA samples. The higher cutoff for MDA was chosen because of the higher noise in single-cell MDA data in general[2]. High MAPD values result from greater noise, characteristic of poor quality samples. For mouse cells we retained PicoPlex samples with a MAPD of 0.8 or below. Furthermore, samples having less than 2% mapped reads were excluded from further analysis. Samples with less than 3,500 transcripts detected (TPM > 1) were also excluded from downstream analysis.

Identification of genomic SNVs

MDA reads resulting from Truseq library preparation were trimmed for 6 bases, and were subsequently aligned to the human reference genome (GRCh37) using BWA. Following duplicate read removal using Picard, the BAM files were recalibrated and variants were called using GATK 3.1.1[40] and a minimum read coverage of 2.

Transcriptome read alignment

Reads were checked for the presence of adapters and trimmed using "Trim Galore" (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Adaptors were removed using Cutadapt[41] and Tophat2[42] (using default settings) was used to align all RNA-seq libraries onto the human genome assembly hg19 (mm10 for mouse), including the ERCC sequences. Expression measurements, expressed as transcript per million (TPM) were then calculated for every annotated protein coding gene using RSEM[43]. Uniquely mapping reads were counted for each gene using HTSeq and normalization across libraries was performed using DESeq2[44].

Read coverage profile over gene body

All genes having total exonic lengths greater than 2kb, 10kb and 15kb were used for each of the 3 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 (i.e. TTS)) were obtained for all genes with sufficient total exonic length at single nucleotide resolution, and the read coverage profiles of each gene were aligned (including inverting the profile for genes on the reverse strand) precisely at the polyadenylation tail for the exonic region profile. Likewise, profiles for “upstream of TSS” were aligned at the TSS, the profiles for the “intronic region” were aligned at the intronic nucleotide that is nearest to the polyadenylation tail, and the profiles for “downstream of TTS” were aligned at the Transcription Termination Site. After alignment, the read coverage profiles were truncated to only the plotting length. To ensure the aggregated profile is not dominated by a handful of extremely highly expressed genes, for each gene, a single maximum across all four profiles was obtained and all four profiles were normalized by dividing by that same number to obtain the relative coverage. This was also to ensure that the relative height of the four profiles for each gene was preserved and remained comparable both before and after normalization and aggregation. The coverage profiles over the four regions respectively were aggregated across all genes and all HCC38 single cells to form the final read coverage profile over genes.

Differential expression using single-cell RNA-seq data

To identify genes appropriate for sample clustering, several TPM cutoffs on expression levels were considered. A TPM cutoff of 1 in at least 16 samples was found to be appropriate for clustering by assessing the number of protein coding genes exceeding 0.1, 0.5, 1, 5, and 10 TPM (Supplementary Figure 2A). The union of these genes were then used to compute the Spearman correlation between all sample pairs. To cluster samples, the command “heatmap.2” in the R package “gplots” was used, 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, a Bayesian approach to single-cell differential expression analysis[45] was used. All genes were then ranked in terms of the maximum likelihood estimates of their difference in expression levels. The TPM of each gene is normalized by the median of the TPM of this gene across all samples and presented in heatmaps as log2-fold differences from this median.

Whole-chromosome expression dosing

In order to assess the transcriptional consequences of copy number variation, for each chromosome we calculated a chromosomal RPKM value 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 using the median expression for the same chromosome in control cells (human HCC38-BL cells, human iPSC-derived neurons disomic for chromosome 21, or mouse blastomeres of control embryos for the induced aneuploidy).

Identification of fusion transcripts

For each cell, candidate gene fusions were identified using TopHat-Fusion[46] and Defuse[47] independently. Only fusions identified in multiple single cells and by both algorithms in the same cell were further considered.

Full length transcript sequencing

The cDNA from each of four single cells were converted into SMRTbell libraries for sequencing on the PacBio RS II (Pacific Bioscience). Briefly, the double stranded cDNA molecules were ligated with hairpin adapters 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[48]. After removal of chimeric reads, only the best scoring alignments for each read were further considered.

SNV calling from single-cell RNA-seq data

To identify SNVs from HCC38 single-cell RNA-Seq data, a pipeline that uses SNiPR[49] was implemented. SNVs were called in each sample separately. To estimate the number of false positive calls, variants called from bulk DNA sequencing of HCC38[21] were used as a gold standard reference and compared to variants called in single cells.

References

1 Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell reports* **8**, 1280-1289, doi:10.1016/j.celrep.2014.07.043 (2014).

2 Voet, T. *et al.* Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic acids research* **41**, 6119-6138, doi:10.1093/nar/gkt345 (2013).

3 A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. *Science* **258**, 67-86 (1992).

4 Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics* **14**, 618-630, doi:10.1038/nrg3542 (2013).

5 Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886-895, doi:10.1016/j.cell.2012.02.025 (2012).

6 Macaulay, I. C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS genetics* **10**, e1004126, doi:10.1371/journal.pgen.1004126 (2014).

7 Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* **32**, 1053-1058, doi:10.1038/nbt.2967 (2014).

8 Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology* **20**, 1131-1139, doi:10.1038/nsmb.2660 (2013).

9 Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology* **14**, R31, doi:10.1186/gb-2013-14-4-r31 (2013).

10 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**, 666-673, doi:10.1016/j.celrep.2012.08.003 (2012).

11 Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology* **30**, 777-782, doi:10.1038/nbt.2282 (2012).

12 Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779, doi:10.1126/science.1247651 (2014).

13 Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369, doi:10.1038/nature13437 (2014).

14 Klein, C. A. *et al.* Combined transcriptome and genome analysis of single micrometastatic cells. *Nature biotechnology* **20**, 387-392, doi:10.1038/nbt0402-387 (2002).

15 Guzvic, M. *et al.* Combined genome and transcriptome analysis of single disseminated cancer cells from bone marrow of prostate cancer patients reveals unexpected transcriptomes. *Cancer research*, doi:10.1158/0008-5472.CAN-14-0934 (2014).

16 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).

17 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).

18 Langmore, J. P. Rubicon Genomics, Inc. *Pharmacogenomics* **3**, 557-560, doi:10.1517/14622416.3.4.557 (2002).

19 Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5261-5266, doi:10.1073/pnas.082089499 (2002).

20 Gazdar, A. F. *et al.* Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *International journal of cancer. Journal international du cancer* **78**, 766-774 (1998).

21 Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-1010, doi:10.1038/nature08645 (2009).

22 de Bourcy, C. F. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PloS one* **9**, e105585, doi:10.1371/journal.pone.0105585 (2014).

23 Huang, H. *et al.* Non-biased and efficient global amplification of a single-cell cDNA library. *Nucleic acids research* **42**, e12, doi:10.1093/nar/gkt965 (2014).

24 D'Alise, A. M. *et al.* Reversine, a novel Aurora kinases inhibitor, inhibits colony formation of human acute myeloid leukemia cells. *Molecular cancer therapeutics* **7**, 1140-1149, doi:10.1158/1535-7163.MCT-07-2051 (2008).

25 Santaguida, S., Tighe, A., D'Alise, A. M., Taylor, S. S. & Musacchio, A. Dissecting the role of MPS1 in chromosome biorientation and the spindle checkpoint through the small molecule inhibitor reversine. *The Journal of cell biology* **190**, 73-87, doi:10.1083/jcb.201001036 (2010).

26 Letourneau, A. *et al.* Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature* **508**, 345-350, doi:10.1038/nature13200 (2014).

27 McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science* **342**, 632-637, doi:10.1126/science.1243472 (2013).

28 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724 (2009).

29 Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nature reviews. Cancer* **7**, 233-245, doi:10.1038/nrc2091 (2007).

30 Ha, K. C. *et al.* Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC medical genomics* **4**, 75, doi:10.1186/1755-8794-4-75 (2011).

31 Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nature biotechnology*, doi:10.1038/nbt.3129 (2015).

32 Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature biotechnology* **27**, 275-280, doi:10.1038/nbt.1529 (2009).

33 Park, I. H. *et al.* Disease-specific induced pluripotent stem cells. *Cell* **134**, 877-886, doi:10.1016/j.cell.2008.07.041 (2008).

34 Shi, Y. *et al.* A human stem cell model of early Alzheimer's disease pathology in Down syndrome. *Science translational medicine* **4**, 124ra129, doi:10.1126/scitranslmed.3003771 (2012).

35 Shi, Y., Kirwan, P., Smith, J., Robinson, H. P. & Livesey, F. J. Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nature neuroscience* **15**, 477-486, S471, doi:10.1038/nn.3041 (2012).

36 Shi, Y., Kirwan, P. & Livesey, F. J. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nature protocols* **7**, 1836-1846, doi:10.1038/nprot.2012.116 (2012).

37 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

38 Moller, E. K. *et al.* Next-generation sequencing of disseminated tumor cells. *Frontiers in oncology* **3**, 320, doi:10.3389/fonc.2013.00320 (2013).

39 Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nature protocols* **7**, 1024-1041, doi:10.1038/nprot.2012.039 (2012).

40 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

41 Marcel, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 3 (2011).

42 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

43 Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500, doi:10.1093/bioinformatics/btp692 (2010).

44 Love, M., Huber, W and Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, doi: http://dx.doi.org/10.1101/002832 (2014).

45 Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**, 740-742, doi:10.1038/nmeth.2967 (2014).

46 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

47 McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS computational biology* **7**, e1001138, doi:10.1371/journal.pcbi.1001138 (2011).

48 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

49 Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *American journal of human genetics* **93**, 641-651, doi:10.1016/j.ajhg.2013.08.008 (2013).

Figure 1

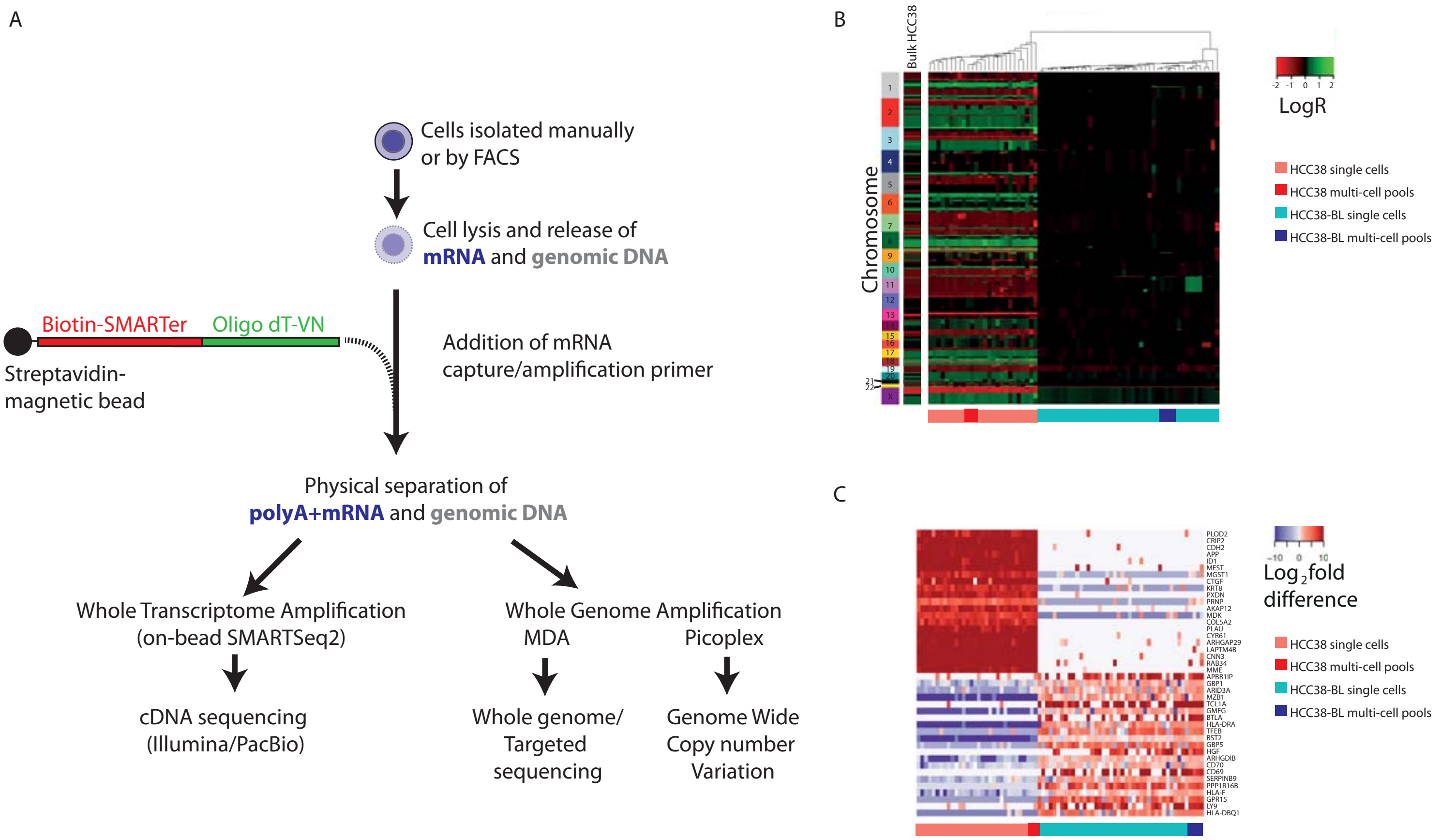

A
Cells isolated manually or by FACS
Cell lysis and release of mRNA and genomic DNA
Biotin-SMARTer
Oligo dT-VN
Streptavidin-magnetic bead
Addition of mRNA capture/amplification primer
Physical separation of polyA+mRNA and genomic DNA
Whole Transcriptome Amplification (on-bead SMARTSeq2)
cDNA sequencing (Illumina/PacBio)
Whole Genome Amplification
MDA
Picoplex
Whole genome/ Targeted sequencing
Genome Wide Copy number Variation
B
Bulk HCC38
Chromosome
1
2
3
4
5
6
7
9
10
11
12
13
15
16
17
19
21
22
X
-2 -1 0 1 2
LogR
HCC38 single cells
HCC38 multi-cell pools
HCC38-BL single cells
HCC38-BL multi-cell pools
C
PLOD2
CRIP2
CDH2
APP
ID1
MEST
MGST1
CTGF
KRT8
PXDN
PRNP
AKAP12
MDK
COL5A2
PLAU
CYR61
ARHGAP29
LAPTM4B
CNN3
RAB34
MME
APBB1IP
GBP1
ARID3A
MZB1
TCL1A
GMFG
BTLA
HLA-DRA
TFEB
BST2
GBP5
HGF
ARHGDIB
CD70
CD69
SERPINB9
PPP1R16B
HLA-F
GPR15
LY9
HLA-DBQ1
−10 0 10
Log2fold difference
HCC38 single cells
HCC38 multi-cell pools
HCC38-BL single cells
HCC38-BL multi-cell pools

Figure 2

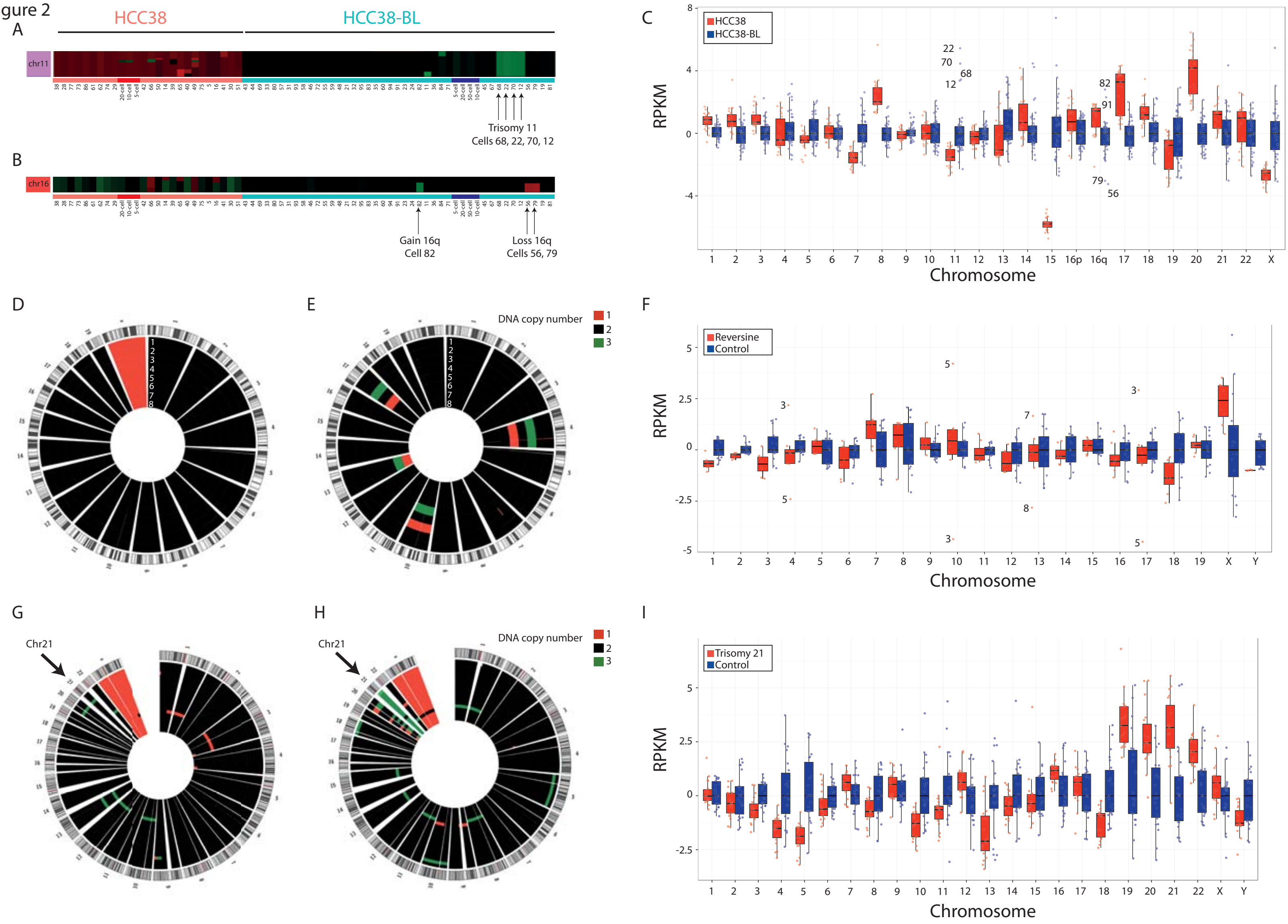
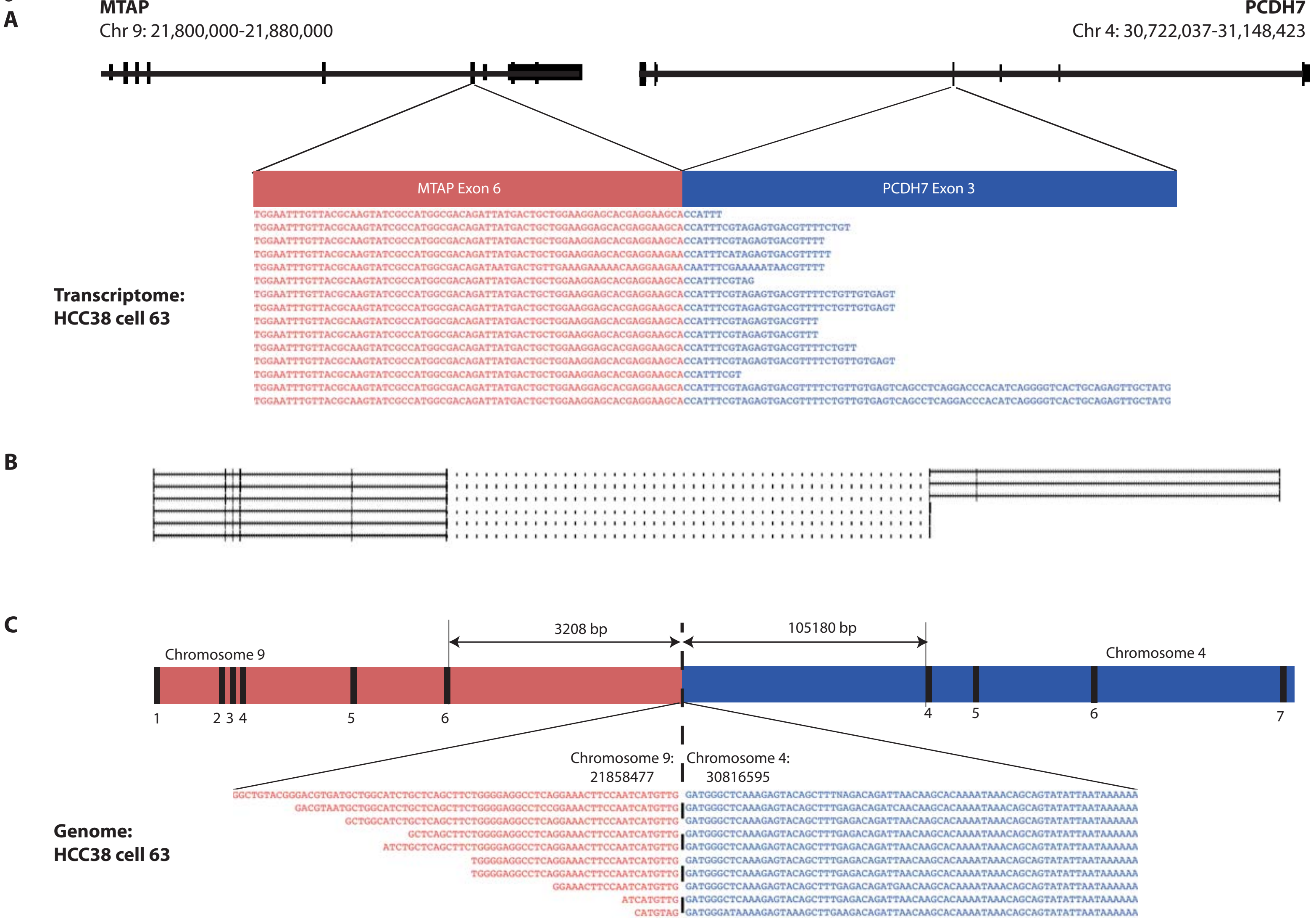

A
HCC38
HCC38-BL
chr11
Trisomy 11
Cells 68, 22, 70, 12
B
chr16
Gain 16q
Cell 82
Loss 16q
Cells 56, 79
C
HCC38
HCC38-BL
RPKM
Chromosome
22
70
68
12
82
91
79
56
D
E
DNA copy number
1
2
3
F
Reversine
Control
RPKM
Chromosome
G
Chr21
H
Chr21
DNA copy number
1
2
3
I
Trisomy 21
Control
RPKM
Chromosome

Figure 3


A
MTAP
Chr 9: 21,800,000-21,880,000
PCDH7
Chr 4: 30,722,037-31,148,423
MTAP Exon 6
PCDH7 Exon 3
Transcriptome:
HCC38 cell 63
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGAACCATTTCATAGAGTGACGTTTTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATAATGACTGTTGAAAGAAAAACAAGGAAGAACAATTTCGAAAAATAACGTTTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAG
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGTTGTGAGT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGTTGTGAGT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGTT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGTTGTGAGT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGT
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGTTGTGAGTCAGCCTCAGGACCCACATCAGGGGTCACTGCAGAGTTGCTATG
TGGAATTTGTTACGCAAGTATCGCCATGGCGACAGATTATGACTGCTGGAAGGAGCACGAGGAAGCACCATTTCGTAGAGTGACGTTTTCTGTTGTGAGTCAGCCTCAGGACCCACATCAGGGGTCACTGCAGAGTTGCTATG
B
C
3208 bp
105180 bp
Chromosome 9
Chromosome 4
1
2 3 4
5
6
4
5
6
7
Chromosome 9:
21858477
Chromosome 4:
30816595
Genome:
HCC38 cell 63
GGCTGTACGGGACGTGATGCTGGCATCTGCTCAGCTTCTGGGGAGGCCTCAGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTNAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
GACGTAATGCTGGCATCTGCTCAGCTTCTGGGGAGGCCTCCGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATCAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
GCTGGCATCTGCTCAGCTTCTGGGGAGGCCTCAGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
GCTCAGCTTCTGGGGAGGCCTCAGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
ATCTGCTCAGCTTCTGGGGAGGCCTCAGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
TGGGGAGGCCTCAGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
TGGGGAGGCCTCAGGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
GGAAACTTCCAATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATGAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
ATCATGTTG GATGGGCTCAAAGAGTACAGCTTTGAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA
CATGTAG GATGGGATAAAAGAGTAAAGCTTGAAGACAGATTAACAAGCACAAAATAAACAGCAGTATATTAATAAAAAA