Massive and parallel expression profiling using microarrayed single-cell sequencing

Sanja Vickovic1, Patrik L. Ståhl2,*, Fredrik Salmén1,*, Sarantis Giatrellis2, Jakub Orzechowski Westholm3, Annelie Mollbrink4, José Fernández Navarro2, Joaquin Custodio4, Magda Bienko4, Lesley-Ann Sutton5, Richard Rosenquist5, Jonas Frisén2 & Joakim Lundeberg1

Single-cell transcriptome analysis overcomes problems inherently associated with averaging gene expression measurements in bulk analysis. However, single-cell analysis is currently challenging in terms of cost, throughput and robustness. Here, we present a method enabling massive microarray-based barcoding of expression patterns in single cells, termed MASC-seq. This technology enables both imaging and high-throughput single-cell analysis, characterizing thousands of single-cell transcriptomes per day at a low cost (0.13 USD/cell), which is two orders of magnitude less than commercially available systems. Our novel approach provides data in a rapid and simple way. Therefore, MASC-seq has the potential to accelerate the study of subtle clonal dynamics and help provide critical insights into disease development and other biological processes.
RNA sequencing has been an invaluable tool for gene expression analysis that has recently progressed from bulk analysis and averaging multiple cells’ transcriptome profiles to single-cell profiling. We have advanced from studying group-specific or condition-dependent fold-changes using microarrays to transcript counting and isoform analysis. This has afforded the potential to unravel both variations among individual cells and stochastic changes across the gene body.

Averaging gene expression levels in a population of cells is beneficial when comparing states of particular tissues in different conditions or developmental stages, and this approach has provided numerous advances and biomarkers for diverse pathological, and other conditions. However, it cannot clarify the discrete roles of individual cells nor the transcriptomic triggers responsible for changes in their phenotypes. In addition, scarcity of biological material often precludes the profiling of rare cell populations by conventional RNA sequencing methods.

There have been major recent technological breakthroughs in the ability to analyse single cells, using methods including cell encapsulation in droplets, solid-surface complementarity DNA (cDNA) analysis, and in situ messenger RNA (mRNA) hybridizations. These methods enable quantitative analysis of gene expression in single cells and have been applied, for example, to study of mouse embryogenesis and expression bimodality. Nevertheless, these methods do not provide any possibilities in combining cell imaging and transcriptome profiling, exhibit low-throughput by analysing a single cell at a time or require expensive droplet instrumentation when available at high-throughput.

In this paper, we describe a novel method, termed micro-arrayed single-cell sequencing (MASC-seq), a single tube approach for analysis of single cells using a barcoded microarray, and demonstrate its ability to profile single cells, in both model cell lines and primary chronic lymphocytic leukaemia (CLL) patient cells. MASC-seq can both image cells to provide qualitative information on cells’ morphology and profile the expression of hundreds to thousands of single cells daily, far more than current standard procedures based on fluorescence-activated cell sorting (FACS) into plates or single-cell picking into individual reaction volumes. MASC-seq could be compared to commercially available systems such as the Fluidigm C1 (ref. 21), which also provides an imaging system before library preparation. However, MASC-seq is improved in terms of daily throughput, not limited by cell size and also is the first system that enables cDNA synthesis of single cells to run in parallel in a single-reaction lowering chances of technical variation in library preparation. MASC-seq is based on commercially available products and reagents and requires only an extra imaging system when compared with standard RNA-sequencing.

**Results**

**Principles of MASC-seq technology.** With MASC-seq, single cells can either simply be smeared and randomly positioned or FACS sorted onto a 6.5 × 6.8 mm² microarray of barcoded DNA oligonucleotides printed in a 33 × 35 matrix with 200 µm centre-to-centre pitch (Fig. 1). The matrix contains 1,007 unique DNA barcodes surrounded by a frame used for orientation during positioning. After attachment, a high-resolution image is taken, which links the position of each barcode sequence with each individual cell, and provides information concerning cell morphology. The image also gives information about the number of cells present on top of each barcoded oligonucleotide spot. In MASC-seq the cDNA is synthesized in a hybridization cassette from ~ 500 single (given 47% occupancy) cells simultaneously in a single well, thereby reducing possibilities of technical variation in the single-cell cDNA synthesis and library preparation steps. This not only increases robustness, but also lowers time and labour costs. After cDNA synthesis, the cells are removed from the microarray surface by proteinase K digestion and the probes are cleaved from the surface with a uracil-specific excision reagent enzyme, which targets the uracil sequence located at the 5’ end of the microarray barcodes. Each cell barcode consists of a uniquely designed 18 nt sequence followed by a unique molecular identifier (UMI), for individual transcript counting, and an oligo-dTVN sequence, thus the method involves 3’ tagging (Fig. 1). The cleaved material is ready for in vitro amplification and library preparation following a procedure similar to the cell expression by linear amplification and sequencing (CEL-seq) protocol. Around 10,000 single-cell libraries can be prepared, for subsequent sequencing, in 2 days.

**Human adenocarcinoma cell line as a model system.** A human breast adenocarcinoma cell line, MCF-7, was used as a model system to evaluate the quality and quantity of data produced using the MASC-seq method. An experiment was first performed to establish where the cDNA was labelled during the reverse transcription reaction (Fig. 2a). This generated a high-resolution fluorescent (Cy3) print, which could be superimposed with the image of either haematoxylin-stained (Fig. 2b) or fluorescently labelled cells taken before cDNA synthesis, providing a convenient way to assess the cells’ quality and visually colocalize the

---

**Figure 1 | MASC-Seq overview.** A FACS machine sorts single cells onto a barcoded microarray, printed with six replicates on an activated glass slide. The throughput of the method and microarray design as a 33 × 35 ID matrix is illustrated. An alternative is to pipette and smear cells which then distribute randomly across the array. Positions of the cells and IDs are noted in a high-resolution image and cDNA is only transcribed when an individual cell lands on top of the barcoded oligo-dTVN primer (ID).

---
Figure 2 | Schematic illustration and results of cDNA and cell colocalization. (a) cDNA is labelled with Cy3 nucleotides during reverse transcription. (b) A haematoxylin image is taken before cDNA synthesis, and a Cy3 image after cDNA synthesis and removal of cells from the microarray. (c) Cell-cDNA colocalization prints for cells used in the study.

cDNAs with single cells. This colocalization confirmed that each cell would produce decoded reads with the correct cell-barcode combination (Fig. 2b, black and white arrow). Furthermore, the cDNA print could only be created when cells were positioned on top of the cell barcode containing the oligo-dTVN sequence (Fig. 2b, red arrows). The results were visually inspected for all cell types used in the study (Fig. 2c). Diffusion of the cDNA signal from the cells’ borders was estimated at 0.81 ± 0.13 μm per cell—(as expected)24 (Supplementary Fig. 2a).

More than 99% of the genes in the single-cell libraries were also found in the bulk RNA sequencing experiments (Fig. 3e). Two types of bulk RNA sequencing libraries were obtained, by polyA-selected cDNA synthesis from 300 ng total RNA via reverse transcription either in solution or anchored on the barcoded microarray surface. In both cases, the single-cell average (n = 156) correlated well (Fig. 3f and Supplementary Fig. 3a) with data obtained from the bulk population-average experiments (R = 0.92 and R = 0.90, respectively). A total 15,937 genes were identified in the single-cell and bulk experiments, of which 1,970 also had DORs of 3.3 and 7.4% compared with each other. Additionally, we compared the acquired gene expression profiles of the single cells between each other (Supplementary Fig. 3b,c) and noted that the expression profiles can vary significantly between single cells (DOR = 35.46%, n = 136). In summary, these findings demonstrate the quality and reproducibility of the data obtained using our MASC-seq technique.

Efficiency and sensitivity. MASC-seq efficiency as well as the variation in MCF-7 cells were compared with single FACS sorted MCF-7 cells prepared with the CEL-Seq protocol11. To assure a fair comparison, cells were taken from the same culturing plate on the same day for both experiments. Also, a UMI-based primer was used when preparing libraries with the CEL-Seq approach to ensure possibility for UMI filtering. Randomly selected cells from the Masc-seq protocol and the same number of cells created with the CEL-Seq protocol (n = 36 for each) shared 64% of the genes (Fig. 4a), with MASC-seq detecting 1.4 times more protein-coding genes in total. On average, Masc-seq captures 17.5 times more unique transcripts and 6.5 times unique protein-coding genes per cell (Fig. 4b). However, most importantly, Masc-seq achieves lower variation of gene expression between cells, as...
compared with CEL-Seq, due to a single-tube reaction principle (Fig. 4c). Furthermore, single-molecule fluorescent in situ hybridization (smFISH) was performed on MCF-7 cells to determine the sensitivity. Absolute numbers of transcripts for seven well characterized genes present in a single cell were compared between the platforms. The sensitivity for the MASC-seq technique was determined to 17.3% (Fig. 4d).

**Barcode crosstalk.** To estimate the degree of barcode crosstalk (to ensure that the data decoded for each cell were accurate and contained within the corresponding cell-barcode combination), we mixed human MCF-7 and mouse 3T3 fibroblasts and smeared them on the barcoded microarray. We then estimated species-specific transcript counts for each of the barcodes that had received a single cell (Supplementary Fig. 4a) based on the high-resolution image. In the single-cell libraries, which produced over 2,000 reads and revealed 1,000 uniquely expressed protein-coding genes per cell (Supplementary Fig. 4b,c), only 30–40 reads were misassigned per barcode, thus only 1.42–1.68% of the total species-specific reads were misassigned to human and mouse barcodes, respectively. Furthermore, t-SNE separated species even at the orthologous gene level (Supplementary Fig. 4d), generating two distinct clusters with all cells from each species clustered together. Population averages correlated well with those of pure samples ($R > 0.88$; Supplementary Fig. 4e,f), further confirming the quality of the data.

**Differential expression in single leukaemic cells.** To assess the applicability of our method for studying a disease state, we analysed primary single-sorted neoplastic B cells obtained from three patients diagnosed with CLL, assigned to different major CLL subsets, with distinct clinical and biological characteristics (clinically classified poor-prognostic subsets #1 and #2, and the good-prognostic subset #4)$^{26,27}$. CD5$^+$CD19$^+$ cells from patients were FACS sorted onto MASC-seq arrays ($n = 1,189 \pm 186$ cells per case). As expected, these small-sized cells (7–10 μm) showed a proportionally lower amount of labelled cDNA in each cell (Fig. 2c, left panel).

Average gene expression levels in cells of the three CLL subsets clearly differed (Supplementary Fig. 5a), with only 43% of the expressed genes shared (Supplementary Fig. 5b), and the strongest differences between the subsets were among the most abundantly expressed genes (Supplementary Fig. 5c). Notably, comparing the 500 genes with highest expression in each subset, only 30 genes showed high expression in two subset pairs and no genes were shared in all three subsets (Supplementary Fig. 5d).
Further analysis through t-SNE and hierarchical clustering revealed subtle differences between single cells within each CLL subset. A number of major and minor clusters were observed in connection to each subset (Fig. 5a–c).

Differential expression analysis based on the hierarchical clustering results revealed unique expression signatures for each of the clusters. For example, in subset #1, the two minor clusters were defined by strong expression of a number of distinct genes, downregulated in the major cluster (Fig. 5a, Supplementary Fig. 5a). Subset #2 exhibited the strongest expression levels per cell (Supplementary Fig. 5a), possibly related to the proliferative drive and very poor prognosis for patients assigned to this subset. For all clusters, the cells exhibited strong cell cycle signatures, mostly indicative of commitment to genome replication, in accordance with previous microarray analysis. For all subsets, the cycling cells were contained within the major cluster.

Hierarchical clustering revealed that, as expected, most single cells within each subset had similar expression profiles, and clustered with other cells of their subset, but a few cells from both subsets #2 and #4 clustered with the poor-prognosis subset #1 cells (Fig. 5e, color-coded by subset and by cluster). This cluster containing cells from all three subsets was marked by a differential expression signature of inhibin beta A (INHBA) (Fig. 5f), a gene associated with cancer progression and poor survival.

Discussion

The importance of heterogeneity between, and within, tumours for therapeutic responses and patient outcomes is well known. However, detecting unique markers within cells of a tumour, predicting associated phenotypic changes and linking the heterogeneity to disease progression is far from straightforward, requiring single-cell analysis. Previous advances in single-cell analysis have enabled measurement of gene expression in a large number of individual cells. This has addressed the drawback of population-averaging in bulk analysis. Further enabling rapid and simple analysis of thousands of cells would help accelerate experimental throughput in both academic and clinical research, potentially resulting in improved patient monitoring, particularly in relation to given therapy.

The presented MASC-seq method for analysing expression profiles in single cells by combining cell imaging with high-throughput single-cell RNA sequencing has been thoroughly validated using both human and mouse cell lines, and primary patient samples. Experiments to evaluate bulk RNA-sequencing
CLL data have allowed characterization of CLL patients into subsets, but averaging gene expression cannot reveal intratumour heterogeneity. Using MASC-seq we observed differences between and within different CLL patient subsets. Also, in each of the subsets, we found a major clone that supports the idea of clonal cooperation and long-lasting clonal equilibrium encouraging overall cancer progression in CLL. Whether these differences are related to the functional pathology of CLL requires further investigation and more extensive studies. Nevertheless, these findings illustrate the heterogeneity within each patient and underscore the importance of analysing transcriptomes at the single-cell level.

The greatest advantage of MACS-seq is generating hundreds of single-cell expression profiles in a single reaction, thus simultaneously lowering technical variability, costs and labour. It can generate expression data from ~10,000 single cells in only 2 days at a cost of just USD 0.13 per cell (Supplementary Data 4), an approach at least 200–300 times lower in cost than the currently most widely applied methods by researchers or commercially. High-throughput approaches such as MASC-seq will greatly facilitate investigations of the biological processes involved in diseases like cancer, and will help to improve our understanding of complex biological phenomena at the single-cell level.

**Methods**

**Array production.** Six of the microarrays were printed per Codelink glass slide. Each microarray could be used in an individual experiment, with a specific Illumina indexing primer, and each glass slide could accommodate up to 16 microarrays, but slides with six replicates were prepared and used in the reported experiments to facilitate daily use. The printing process was performed by ArrayJet LTD (Scotland, UK) using the ArrayJet Spider system. The DNA oligonucleotides were spotted in 200 μm centre-to-centre vertical and horizontal pitch fashion.

**Figure 5 | Results of analysis of differential expression in single leukaemic cells.** (a–c) Clusters in subsets #1, #2 and #4 (respectively) detected by t-SNE and hierarchical clustering displayed in three-dimensional plots. (d) Barplot of single CLL B cells ordered by average cell cycle score and subset. (e) t-SNE plot of single CLL cells expressing most abundant genes found in each patient subset (colour-coded by subset in the left panel and by cluster in the right panel). (f) Violin plot INHBA expression for all three clusters.
arrays’ frame was labelled to enable visualization during FACS as described in Supplementary Information.

Cell handling and total RNA extraction. MCF-7 (human breast metastatic adenocarcinoma) cells were cultured at 37 °C in a 5% CO₂ environment. The breast cancer cells were grown in Eagle’s Minimum Essential Medium supplemented with 10% FBS (both from Thermo Fisher Scientific, Life Technologies, Paisley, UK), harvested at 70% plate confluency by trypsinization, and RNA was extracted from ~1 million of the cells using a RNeasy Mini Kit (Qiagen, Limburg, The Netherlands). NIH/3T3 (mouse embryonic fibroblast, hereafter 3T3) cells were cultured in Dulbecco’s Modified Eagle’s Medium supplemented with 10% FBS (both from Thermo Fisher Scientific, Life Technologies, Paisley, UK). These 3T3 cells were harvested at 70% plate confluency for RNA extraction, as described above. After trypsinization, the cells were washed twice with 1 × PBS. Total MCF-7 RNA (300 ng) was fragmented to 350 nt fragments, on average, using a magnesium fragmentation protocol involving 5 min incubation at 95 °C and used in all bulk reference experiments, for CDNNA synthesis in both solution and on the microarray surface. MCF-7 and 3T3 cells were used in further single-cell experiments by smearing them on the microarray surface after fixation. Before attachment to the surface, the buffer was exchanged to 0.1 × saline citrate (SSC) and the cells kept on ice until the protocol was started. Smearing was performed by taking 3 μl of cells at a concentration ~2,000 cells μl⁻¹ and first slowly pipetting the cells on the array surface with taking care not to actually touch the surface with the tip. Then, the cell solution could be smeared with a side of a pipette tip (careful not to touch the surface of the array) followed by attachment at 37 °C for 5 min to the slide. MCF-7 was obtained from the German collection of microorganisms and cells (DSMZ) and authenticated with RNA sequencing. The 3T3 cell line was obtained from ATCC. Both cell lines were cultured for mycoplasma contamination (Minerva Biolabs Gmbh, Berlin, Germany).

Patient sample collection. Cryopreserved peripheral blood mononuclear samples derived from 6 CLL patients were included from the Biobank at Uppsala University Hospital, Sweden. All cases were diagnosed and classified according to recently revised iwCLL criteria with a typical CLL immunophenotype. Cases were selected for CD19/CD5 co-expression and a poor prognosis. Informed consent was obtained according to the Helsinki declaration and the study was approved by the Regional Ethics Review Committee in Uppsala (2014/33).

FACS sorting. The patient samples were collected as described. Before sorting, the cells were fixed in a way to be compatible with staining and immunohistochemical methods. The CD5+ CD19+ population (for antibody staining see Supplementary Information) was sorted on the labelled array with position adjustments to increase the sorting accuracy on the correct barcoded DNA oligonucleotide position. The FACS sorter utilized for analyses and single-cell sorting was a BD Influx by Becton Dickinson. Cells analyses and sorting were performed using a 70 μm nozzle. The FACS sorting stage spatial resolution is <100 μm but the droplet size is the one setting the spatial resolution during the sorting stage. FACS setup: 70 μm nozzle created droplets with a trace diameter of 340 ± 40 μm on the collection slide. To increase the positional precision in single-cell sorting we used Cell Precision, a kit that integrates an indication laser and a camera beneath the sorting slide to the FACS sorter. The laser trace size on the slide was 330 ± 20 μm and the effective positioning resolution was <100 μm.

The single-cell sorting was performed in successive sorting cycles giving adequate space between neighbouring matrix positions for avoiding droplet fusion. The FACS sorting matrix was 11-by-12 with a matrix unit length of 600 μm. The successive sorting cycles differed in the initialization position, which was displaced by 200 μm for every successive cycle at the X dimension (3 cycles) and 200 μm at the Y dimension (3 cycles). In total, we performed 9 sorting cycles to achieve maximum coverage of the oligonucleotides microarrays. Adequate time between successive sorting cycles was given in order for the sorted droplet to evaporate and the encapsulated cell to end up on the oligonucleotides base. The total sorting time per microarray was 8 min. The average efficiency for indexed sorting was 47% (Supplementary Data 5). After sorting, the slide was heated to 37 °C for 5 min.

Visualization of cell positions. Images of sorted and stained cells (for staining protocol see Supplementary Information) on barcoded microarrays were recorded using a Metafer Vslide scanning system (MetaSystems, Mannheim, Germany) installed at the Arbeitsgemeinschaft zur Erforschung der Leber (Carl Zeiss, Jena, Germany). All images were taken with the x20 Plan-Apochromat objective lens, and stitched with the VSlide software (v1.0.0). Before scanning the Cy3 emission range (560–610 nm), the glass slide was mounted with SlowFade Gold Antifade reagent (Life Technologies, Paisley, UK). In case a +40 object was used, the scanning was performed on the Zeiss LSM780 system and the images stitched with the ZEN (Zeiss) software. Photoshop CS6 software (Adobe Systems) was used to merge images.

cDNA synthesis and library preparation. Cell permeabilization. To permeabilize cells on the slide, the slide was placed in an ArrayIT hybridization cassette diving the individual arrays into 8 × 10 × 0.5 cm1 × 1 × 0.05 cm (Sigma-Aldrich, MO) solution (pH 1) warmed to 37 °C was added to each of the sample arrays for 30 s, carefully pipetted out, and the surface carefully washed with 100 μl 0.1 × SSC. cDNA synthesis. The cDNA synthesis mixture contained 1,280 U of Superscript II Reverse Transcriptase, 256 U of RNaseOut, 3.2 μl of 0.1 M dithiorethiol (DTT) and 0.8 × First Strand buffer (all from Invitrogen, Life Technologies, Paisley, UK), 3.2 μg actinomycin-c (Sigma-Aldrich, St. Louis, MO), 0.4 mM dNTPs mix (Thermo Fisher Scientific, Life Technologies, Paisley, UK) and 1.5 × BSA (NEP, Ipswich, MA, USA). The reaction volume was set to 70 μl by adding water. The mix was then added to the microarray surface with the attached ArrayIT hybridization cassette, and then incubated for 16 h. After cDNA synthesis, the microarray wash was performed with 100 μl W3.

Cell removal. To remove cells, proteinase K (Qiagen, Limburg, Netherlands) was mixed with proteinase K digestion (PKD) buffer 1:4 then added to the microarray surface with the attached ArrayIT hybridization cassette. After incubation for 1 h at 56 °C, the glass slide was sequentially washed in 2 × SSC supplemented with 0.1% sodium dodecyl sulphate at 50 °C for 600 s, 0.2 × SSC on 60 s and 0.1 × SSC for 60 s at room temperature, then spin-dried.

Probe release. A probe release mix was prepared by mixing 6.4 U of USER enzyme and 1.5 × BSA (both from NEP, Ipswich, MA, USA) 0.8 × second strand buffer (Invitrogen, Life Technologies, Paisley, UK) 70 μM 4NTP mix (Thermo Fisher Scientific, Life Technologies, Paisley, UK) and water to a final volume of 70 μl, and heated to 37 °C. The glass slide (with the attached ArrayIT hybridization cassette) was then incubated with the mixture for 2 h at 37 °C, then 65 μl of the released probe-containing mixture was collected in a 0.2 ml low-binding tube (Eppendorf, Hamburg, Germany). All of the released probe-containing mixtures were performed using low-binding tubes and tips (Biotix, San Diego, CA). Reference material was added as described in Supplementary Information. The array print quality was also examined and used in picture overlay, also described in the Supplementary Information.

Strand synthesis and blunting. Second strands were generated from template cDNA strands using 18.4 U DNA Polymerase I and 0.92 U RNaseH mixed with 0.2 × first strand buffer (all from Invitrogen, Life Technologies, Paisley, UK). The mixtures were incubated for 2 h at 16 °C. To initiate blunting reactions, 15 U of T4 DNA polymerase (NEP, Ipswich, MA) was then added to each mixture for 20 min before the reaction was stopped by adding cold EDTA to a final concentration of 800 μM. Samples were then purified using the Agencourt RNAclean XP system (Beckman Coulter, Pasadena, CA) according to the manufacturer’s instructions, and eluted in 10 μl water. Sample volumes were subsequently reduced to ~5.6 μl using a SpeedVac vacuum centrifuge.

In vitro transcription. In vitro transcription was performed using a MEGAScript T7 Transcription kit (Ambion, Life Technologies, Paisley, UK) with 1.6 μl of the provided enzyme mix in 1 × reaction buffer and 6.4 μl of the provided nuclease 5’-triphosphate (NTP) mix, supplemented with 16 U of the SuperRase In RNAse Inhibitor (Invitrogen, Life Technologies, Paisley, UK). The mixture was added to the purified product and incubated initially for 14 h at 37 °C and subsequently 4 °C until the sample could be processed. The amplified RNA (aRNA) was then purified using the Agencourt RNAclean XP system (Beckman Coulter) according to the manufacturer’s instructions and eluted in 10 μl water. Finally, the purified and aRNA was evaluated using a mRNA Pico Bioanalyzer 2100 system (Agilent, Santa Clara, CA).

Adaptor ligation and second cDNA synthesis. After denaturing secondary strands by incubation for 2 min in 70 °C, aRNA adaptors (5’-AGATCAGGAAGGGCAGCAAGTGATCGACGCTCAG3’-5’ M, InPE2.0 (5’-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3’) was added to the aRNA strands using 300 U of the truncated T4 RNaseH (NEP, Ipswich, MA) in 1 × T4 RNA ligase reaction buffer. The adaptors had each been capped at the 5’ end with dA and adenylated at the 3’ end to mitigate single-strand ligation to the aRNA. 60 U of murine RNase Inhibitor was then added and the mixture was incubated at 25 °C for 1 h. The sample was purified using the Agencourt RNAclean XP system (Beckman Coulter) according to the manufacturer’s instructions and eluted in 10 μl water.

Indexing PCR. Quantitative PCR was first applied to determine the number of cDNA copies for the indexing reaction using the KAPA Library Indexing kit with a mixture supplemented with 1 × Eva Green, with PE1.0. (5’-AGATTGACGCGGCCAC- AGGAAACTATCTCTCAGCGAACGTTCTGCAG3’-5’ M), PE2.0. (5’-GGTACGCTAGGCTACAGGCAGGCCGCTTCAG3’-5’ M) and the Illumina Indexing primer at 5 μM final concentrations. The indexing reaction was then performed using the same amplification conditions and the purified product was used in a second indexing reaction using phosphoramidite building blocks and polyethylene glycol45, then diluted for sequencing on the NextSeq500 instrument. Reference libraries were constructed as described in Supplementary Information and also sequenced on the NextSeq500 instrument.
Establishing fluorescent cDNA signatures from single cells. Cells were routinely attached to the array surface by heating the slide for 5 min at 37 °C, after confirming that they would not detach by placing 121 DAPI-labelled cells on the array (via the FACS procedure), heating, then visually inspecting and counting the cells remaining on the array. The slide was re-examined and the cells were re-counted after washing the glass slide in 0.3% (v/v) sodium dodecyl sulfate in 50 °C for 60 s, 0.2 × SSC for 60 s, and 0.1 × SSC for 60 s at room temperature. Images of the slides were also acquired before and after washing then superimposed. All of the 121 cells remained and did not move on the array surface during the washing steps (data not shown).

To ensure that the cDNA was localized under the cells’ surfaces, we performed another straightforward but informative preliminary experiment, in which cells from patient samples or cell lines were attached to the array surface by the FACS procedure or smearing (respectively), stained with haematoxylin, and visualized. The cDNA was then labelled during the cDNA synthesis reaction by supplementing the mixture with 25 µM Cy3-labelled dCTPs (PerkinElmer, Waltham, MA). In addition, the concentration of Cy5 was reduced to 10 µM while concentrations of the remaining dNTPs remained the same. The mixture was incubated overnight at 37 °C, then the cells were removed from the array surface and the underlying Cy3-cDNA print was visualized. Diffusion signatures were estimated using ImageJ. The cells were fluorescently labelled prior as described in the Supplementary information and the cDNA synthesis reaction supplemented as described here. Distance of the Cy3-cDNA signal compared with the cell border was estimated for 9 3T3 cells and 11 MCF-7 cells.

Data analysis. Mapping and demultiplexing. The samples were sequenced in paired end mode. The forward read was sequenced at 31 nt containing the cell barcode sequence and the UMI, and the reverse read consisted of 121 nt providing the matching transcription information. First, the reads were trimmed from both ends on a Burrows-Wheeler aligner (BWA) quality-based approach, and the adaptor sequences were removed. The reads were then mapped to a NCBI’s reference human transcriptome using Bowtie2 and annotated against the RefSeq transcriptome reference containing sequences annotated as NM_ and NR_ sequences (as of 11 March 2014). The mapped reads were filtered to identify (and discard) ribosomal sequences. HTSeq count, with the setting ‘-intersection-nonempty’, was employed to count the number of reads per gene, marking the results as gene expression values in the downstream analysis. The reads could then be demultiplexed, using the 18 nt cell barcode processed with a kmer-approach with two mismatches permitted during the process.

The reads belonging to each of the cell barcodes were subsequently filtered to remove duplicates based on the UMI. To avoid unnecessary delays in data processing, a minimal hamming distance was set and read clusters were created. Potential duplicate sequences were discarded from further analyses. The standard UMI was a semi-randomized 9 nt sequence, WSNWNWNNW, but when analysing reference bulk libraries in solution, it was combined with a fully randomized 8 nt sequence (NNNNNNNN) embedded in the template switch oligo to generate enough possible UMI combinations. UMI-filtered data were used in all of the following data analyses. The reads mapping to gene MALAT1 were removed from analysis due to problems with self-priming. An illustration of data processing is depicted in Supplementary Fig. 6.

Normalization and data pre-processing. Expression profiles were normalized by adjusting the total number of UMIs per barcode (corresponding to the total number of transcripts per cell) to 200,000 reads (providing so called TP200K) and adding a pseudocount before transforming the data to log2 scale. In case of CLL samples, data was normalized to 10,000 reads per barcode (so called TP10K).

To assess expression signatures associated with empty barcodes (which did not receive cells), a high-resolution first examined expression profile of the 50 most highly expressed genes apart from recognized ‘housekeeping genes’ 19, and removed them from the data set. This resulted in removal of 15 genes from the model MCF-7 and 3T3 data sets (Supplementary Data 7). Visual inspection of the reads confirmed that most of them were cytoplasmic non-coding RNA or binding protein sequences, apart from the apparently most abundant mitochondrial signature. To mitigate these most abundant genes present in the background, we first compared their levels of expression and distributions to those present in the single-cell libraries. All of these highly abundant genes exhibit higher and more even levels of expression continuously in the single-cell libraries as compared with the background libraries with the expressions of ACTB and GAPDH exemplified in Supplementary Fig. 7a. All the background libraries also correlated well to each other (Supplementary Fig. 7b) concluding the background libraries were similar to each other and most probably a result of background cell-free RNA material present in the cell-suspension buffer before smearing cells onto the array. Similarly, when amount of reads mapping to background barcodes was compared between FACS sorted libraries and smeared libraries, the smeared libraries on average exhibited 11% more reads mapping to the background (Supplementary Fig. 8), at the same sequencing depth, further strengthening the fact the background in depended on the material present in the suspension buffer.

Expression profiles of single MCF-7 cells. In experiments which involved smearing MCF-7 cells on the array surface, four sets of libraries were obtained: background (no cells), singles (1 cell), doublets (2 cells) and clusters (> 2 cells). Gene expression profiles of 1,021 background, 136 single-cell, 107 doublet and 858 cluster libraries were acquired from four MCF-7 array experiments performed on the same day under the same conditions. Data were filtered based on mean number of reads present in the single-cell libraries, that is, all cell libraries above this threshold were taken into analysis. After data pre-processing, signals from all of the remaining genes were used in subsequent analyses. These included hierarchical clustering of Pearson’s correlation distances, and t-SNE to visualize the results (although the groups were clearly separated by the first two components).

We examined ranked gene expression by linking it to the mean CV for all four groups in the MCF-7 data. We also evaluated cell-to-cell Pearson’s correlation coefficients. Finally, DOR were defined as percentage of genes that were not expressed in one data set while showing any levels of expression in the other.

Barcode cross-talk experiment. To evaluate barcode crosstalk, 3T3 and MCF-7 cells were mixed in an ~1:1 ratio, smeared on an array, and expression profiles of barcoded single-cell libraries (identified from high-resolution images, 97 in total) were analysed, as described above. A list of gene orthologs was downloaded from NCBI’s non-orthologous sequences (of 2,828 human-specific and 2,348 mouse-specific protein-coding genes) in each of the 97 barcoded libraries. In total 46 MCF-7 and 51 3T3 single cells were correctly identified. We then examined barcode crosstalk by evaluating raw species-specific read counts in each of the demultiplexed single-cell libraries.

We further investigated whether or not the single-cell libraries could be separated using orthologous sequences, by examining cells’ correlations based on the 2,000 most variable (based on CV) gene orthologs. t-SNE was run for 10,000 iterations on two significant principal component loadings (P < 0.01) providing clear species-specific clustering. We also compared the average normalized expression of the decoded species-specific libraries to that of pure species libraries created separately.

Chronic lymphocytic leukaemia single-cell analysis. Each patient sample was processed on a separate MASC-seq slide to avoid any chances of cross-contamination between the pools. Expression profiles were derived by hierarchical clustering on non-orthologous sequences (of 2,828 human-specific and 2,348 mouse-specific protein-coding genes) in each of the 97 barcoded libraries. In total 46 MCF-7 and 51 3T3 single cells were correctly identified. We then examined barcode crosstalk by evaluating raw species-specific read counts in each of the demultiplexed single-cell libraries.

To inspect major differences between the subsets, we examined expression of the 500 most strongly expressed genes in each of the subsets. Subsets #1, #2 and #4 had 469, 474 and 495 abundantly expressed genes not present at high levels in any other subset, respectively. Pearson’s correlations between cells expressing these genes were obtained and used in hierarchical clustering and t-SNE, the latter based on 47 significant principal component loadings (P < 0.01). A likelihood ratio test 49 was used to determine whether the expression patterns of the three clusters determined in the previous step significantly differed (P < 0.001).

Additionally, to evaluate intra-subset differences, we further examined the 2,000 most variable genes (based on CV) in each of the subsets, after reducing differences among the distances of Pearson correlation matrices and the significant principal components (P < 0.01). t-SNE plots were created in three dimensions and hierarchical clustering was applied to each of the clusters identified in the subsets. The significance of differences in expression patterns between the clusters was then tested as above, using the likelihood ratio test and a P < 0.001 significance threshold.

Cell cycle analysis. A list of cell-cycle-specific genes representing five stages of the cell cycle (G1/S, S, G2/M, M and G0/G1) was taken from a published source 49, and genes that passed a certain threshold (R > 0.3) in each of the respective cell cycle phases were selected for further analysis. The data were normalized in two steps: one stage phase-specific barcode was generated for each of the five phases, to determine its current cell cycle stage 31. When analysing the CLL samples, a cell-cycle-specific score was further calculated for each patient’s subset, thereby creating subset-specific scores.
