

# STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array

Hannah Hochgerner<sup>1,2</sup>, Peter Lönnerberg<sup>1,2</sup>, Rebecca Hodge<sup>3</sup>, Jaromir Mikes<sup>2</sup>, Abeer Heskol<sup>1</sup>, Hermann Hubschle<sup>4</sup>, Philip Lin<sup>4</sup>, Simone Picelli<sup>1,2</sup>, Gioele La Manno<sup>1,2</sup>, Michael Ratz<sup>5</sup>, Jude Dunne<sup>4</sup>, Syed Husain<sup>4</sup>, Ed Lein<sup>3</sup>, Maithreyan Srinivasan<sup>4</sup>, Amit Zeisel<sup>1,2</sup>† and Sten Linnarsson<sup>1,2</sup>†

<sup>1</sup> Division of Molecular Neurobiology, Dept. of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

<sup>2</sup> Science for Life Laboratory, Solna, Sweden

<sup>3</sup> Allen Institute for Brain Science, Seattle, Washington, USA

<sup>4</sup> WaferGen Biosystems Inc., Fremont, California, USA

<sup>5</sup> Dept. of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden

† Corresponding authors. Email: [amit.zeisel@ki.se](mailto:amit.zeisel@ki.se) (A.Z.) [sten.linnarsson@ki.se](mailto:sten.linnarsson@ki.se) (S.L.)

## Abstract

Single-cell RNA-seq has become routine for discovering cell types and revealing cellular diversity, but currently no high-throughput platform has been used successfully on archived human brain samples. We present STRT-seq-2i, an addressable 9600-microwell array platform, combining sampling by limiting dilution or FACS, with imaging and high throughput at competitive cost. We applied the platform to fresh single mouse cortical cells and to frozen post-mortem human cortical nuclei, matching the performance of a previous lower-throughput platform.

## 1 Introduction

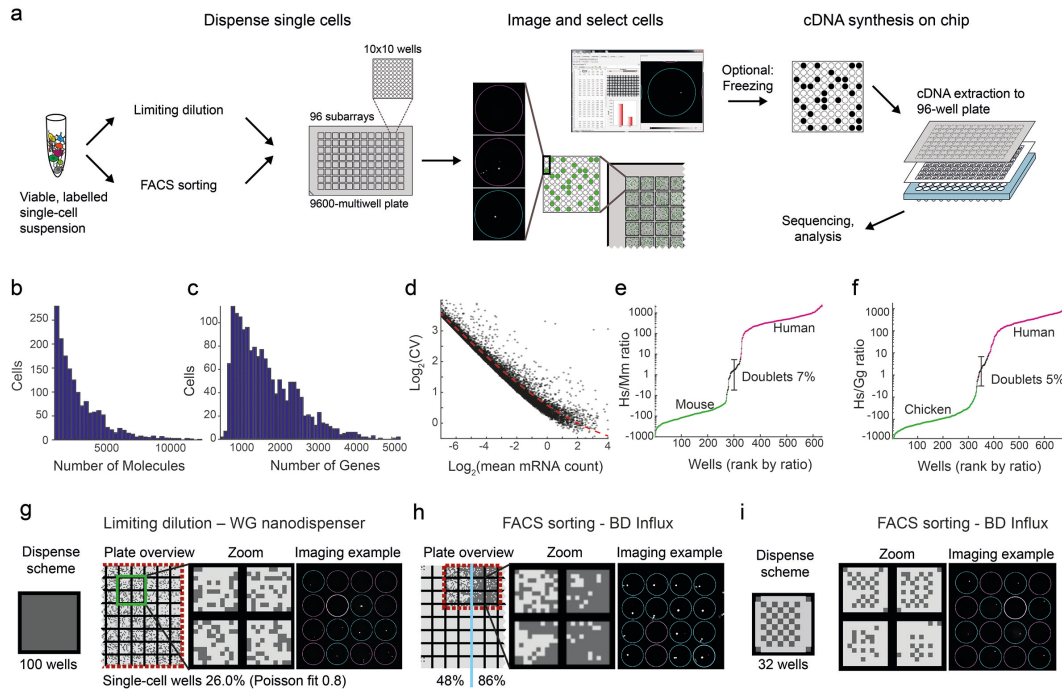
2 Single-cell RNA sequencing has become the method of choice for discovering cell types<sup>1,2</sup>  
3 and lineages<sup>3-5</sup>, and for characterizing the heterogeneity of tumors<sup>6,7</sup> and normal tissues such  
4 as lung<sup>8</sup> and the nervous system<sup>9</sup>. Protocols with high levels of accuracy, sensitivity and  
5 throughput are now available commercially and from academia. Commonly used platforms  
6 include valve microfluidic devices<sup>10,11</sup>, microtiter plate formats such as SMART-seq2,  
7 MARS-seq, CEL-seq2 and STRT-seq<sup>11-14</sup>, as well as droplet microfluidics<sup>15-17</sup>.

8  
9 An ideal platform should combine high throughput, low cost and flexibility, while  
10 maintaining the highest sensitivity and accuracy. Desirable features include imaging of each  
11 individual cell (e.g. to identify doublets and to measure fluorescent reporters), flexibility to  
12 sort cells (e.g. by FACS) and to combine multiple samples in a single run. While current  
13 valve microfluidics and microtiter plate-based formats meet most of these requirements, they  
14 are often expensive and low throughput. In contrast, droplet microfluidics achieve very high  
15 throughput and low cost per cell, but at the expense of flexibility. In particular, multistep  
16 protocols present a challenge to droplet-based systems, do not permit imaging and typically  
17 do not scale well to a large number of samples (as opposed to cells).

18  
19 The adult human brain poses a particular challenge for single-cell genomics. With few  
20 exceptions, samples from human brain are only available in the form of frozen post-mortem  
21 specimens. Although good human brain banks exist, where the postmortem interval has been  
22 minimized and RNA of high quality can be extracted, it is not possible to obtain intact whole  
23 cells from such materials. Somewhat surprisingly, it has been shown that nuclei can be  
24 sufficient to derive accurate cell type information<sup>18</sup>, including from frozen human brain  
25 specimens<sup>19</sup>. However, nuclei have not yet been successfully analyzed on high-throughput  
26 platforms such as droplets or microwell arrays.

27  
28 To meet these challenges, we developed an addressable nanoliter-volume microwell array  
29 platform compatible with our previously described STRT-seq chemistry, which is sufficiently  
30 sensitive to analyze both whole cells and nuclei. We designed a custom aluminum plate with  
31 outside dimensions conforming to standard microtiter plates, but with 9600 wells arranged in  
32 96 subarrays of 100 wells each (Fig. 1a). The wells were designed with a diameter and  
33 spacing large enough to be addressable by a microsolenoid nanodispenser capable of  
34 depositing as little as 35 nL per well. With a maximum well volume of 1  $\mu$ L, this facilitates  
35 efficient multi-step protocols that include separate lysis, reverse transcription and PCR steps  
36 with sufficient dilutions to avoid inhibition of later steps by the reagents used in previous  
37 steps. We modified and extensively reoptimized our 5' STRT-seq method (Fig. S1) by  
38 introducing an additional level of indexing ('dual index'), to allow multiplexing first within  
39 each subarray and then across the whole plate. Sequencing libraries were designed for single  
40 rather than paired-end reads, contributing to a competitive per-cell cost of the method.

41



**Figure 1** | Technical performance. (a) STRT-seq-2i workflow overview. (b-c) Distribution of molecule (b) and gene counts (c) for cortex data (Fig. 2). (d) Coefficient of variation (CV) as a function of mean number of molecules  $m$  expressed in cortex cells. The fitted line represents an offset Poisson,  $\text{Log}_2 \text{CV} = \text{Log}_2(m^{-0.5} + 0.5)$ . (e-f) Doublet rates as estimated by the ratio of species-specific molecules, per well, in mouse-human (e) and chicken-human (f) two-species experiments. (g-h) Single-cell well success rate when addressing 100 wells per unit by (g) limiting dilution or (h) FACS with 200 nL (left) or 50 nl PBS (right) predisposed. (i) Accuracy of FACS demonstrated by checkerboard pattern sort to 32 wells per unit.

The addressable microwell array format allows the user to process multiple samples per plate in parallel, and includes an imaging checkpoint for single-cell positive wells. Cells can be deposited by limiting dilution or by FACS, yielding up to ~3000 or ~7500 single cells per plate, respectively, and multiple plates can be prepared at once and frozen for later processing.

To adapt 5' STRT-seq (also known as C1-STRT<sup>11</sup>) for dual indexing (STRT-seq-2i), we optimized all key steps of the protocol, including cell lysis (Fig. S2a-c), reverse transcription (Fig. S2d-e), PCR (Fig. S2f-g) and sequencing library preparation (Fig. S2h) to increase yield and quality.

In order to validate the method, we first measured its technical performance (Fig. 1b-f and Fig. S3). Using cells freshly isolated from mouse somatosensory cortex (as previously described<sup>1</sup>) we generated an average of 41,000 mapped mRNA reads per cell. We observed an average of 3686 detected genes, and 8706 detected mRNA molecules (Fig. S3a, distribution Fig. 1b-c) for the typical cortical pyramidal cell, comparable to previously published data<sup>1</sup> with an average read depth of 500,000 mapped mRNA reads per cell (4550 genes, 17530 molecules). The number of mRNA molecules and genes detected varied greatly by cell type, indicating that this variation was dominated by biological, not technical, factors (Suppl. Fig. S3a). Noise followed an overdispersed Poisson distribution, as expected (Fig. 1d).

73

74 Next, to assess possible cross-contamination between wells and subarrays, we performed  
75 mixed two-species experiments with human (Hek293) and mouse (mES) or human and  
76 chicken (DF-1) cells. Approximately 7% (mouse) or 5% (chicken) of wells contained  
77 molecules stemming from both species at roughly equal ratios, indicating true doublet wells  
78 (Fig 1e-f, Fig. S3b). These doublets were likely due to inefficient detection of poorly stained  
79 cells by imaging, since post-analysis manual inspection of putative doublet wells could not  
80 confirm the doublets. In contrast, background reads from the other species in single-cell wells  
81 was low (average 37 molecules). Therefore, ambient RNAs in the suspensions or cross-  
82 contamination occurring further downstream (e.g. during barcode indexing steps or library  
83 preparation), all contributed little to final mRNA counts.

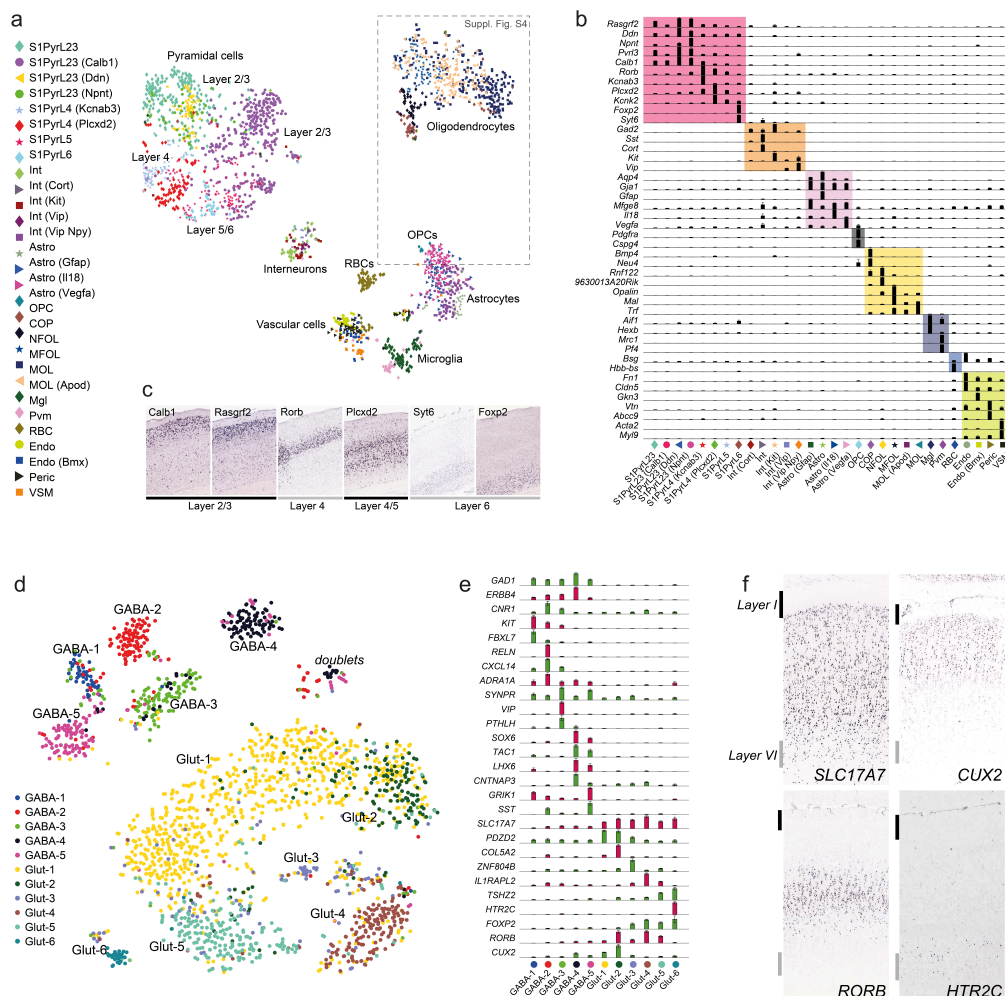
84

85 In order to assess the performance of different cell deposition strategies, we first dispensed  
86 cells using limiting dilution, i.e. loading an average of one cell per well. We designed 32  
87 barcodes, to allow recovery of up to 32 wells per subarray or 3072 total (slightly below the  
88 Poisson limit of 3552 cells). In practice, we observed an average single-cell fill rate of almost  
89 2500 cells per plate (Fig. 1g, Table S1). In order to improve yield per plate, we used FACS to  
90 sort cells directly into the wells. In this mode, with optimal sorting parameters, we were able  
91 to get 86% single cells (Fig. 1h), or more than 8,000 cells per plate, although sorting that  
92 many cells was a slow process (see Methods). FACS also has other advantages, e.g. it can  
93 reduce the incidence of doublets, can be used to focus on desired rare subpopulations, and to  
94 link molecular surface properties to each individual cell by index sorting. To ensure the  
95 accuracy of FACS dispensing, we sorted cells in a checkerboard pattern, showing a  
96 deposition error rate of 4.1% of total addressed wells (Fig. 1i).

97

98 Applying the method to mouse somatosensory cortex (S1 region), in five independent  
99 experiments, we selected approximately 2200 cells (Fig. 2a). Biclustering with BackSPINv2  
100 algorithm<sup>9</sup> resolved the structure of subclasses to a similar level as reported previously (Fig  
101 2a, Fig. S4). We detected all major cell types, including excitatory and inhibitory neurons,  
102 oligodendrocytes, astrocytes, endothelial cells, microglia and ependymal cells. We also  
103 detected known subtypes. For instance, pyramidal neurons formed distinct clusters that  
104 showed layer-specific expression profiles (Fig 2b-c)<sup>20</sup>. Importantly, the method showed  
105 reduced bias against cell size compared with the Fluidigm C1, demonstrated by the presence  
106 of the small oligodendrocyte precursor cells (OPC) in this dataset, which were not detected in  
107 our earlier results (Zeisel et al. 2015<sup>1</sup>; but see also Marques et al. 2016<sup>9</sup>, where OPCs were  
108 detected in a much larger dataset). Further, the full oligodendrocyte lineage was present and  
109 previously described markers (eg. *Pdgfra*, *Itpr2* and *Apod*)<sup>9</sup> could be related to the maturation  
110 process from OPC to myelin-forming oligodendrocyte (MFOL) (Fig S4).





111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124

**Figure 2 | Heterogeneity of cell-types in the mouse somatosensory cortex and human temporal cortex.** (a) tSNE visualization for clustering of 2192 single-cells, colored by BackSPINv2 clusters. (b) Top marker genes of each cell type presented as normalized average expression by cluster, with major cell classes overlaid by colored boxes. (c) Genes specific to pyramidal neuron subclasses by layer specificity, confirmed by *in situ* hybridization from Allen Mouse Brain Atlas. Image credit: Allen Institute. (d) tSNE visualization for clustering of 2028 post-mortem isolated neuronal nuclei from the middle temporal gyrus, colored by BackSPINv2 clusters. (e) Top marker genes of each neuronal subtype presented as normalized average expression by cluster. (f) Validation of pyramidal neuron (Glut) gene expression layer specificity, by *in situ* hybridization from Allen Human Brain Atlas. The outermost layers I and VI are indicated by strokes. Image credit: Allen Institute.

125 To test the versatility and sensitivity of the platform, we next used neuronal (NeuN+ FACS-  
126 sorted) nuclei isolated from a frozen post-mortem human middle temporal gyrus specimen. In  
127 a single experiment, we obtained 2028 nuclei. Despite shallow sequencing (mean <62 000  
128 reads per cell, Fig. S5), BackSPINv2 hierarchical clustering revealed eleven distinct  
129 glutamatergic and GABAergic cell types (Fig. 2d). These were characterized by exclusive or  
130 combinatorial expression of genes (Fig. 2e), and validated by comparison with Allen Human  
131 Brain Atlas<sup>21</sup> (Fig. 2f). To our knowledge, this makes STRT-seq-2i the only current high-  
132 throughput platform amenable to single-nuclei RNA-seq in human postmortem tissue.

133

134 In summary, STRT-seq-2i is a flexible and high-throughput platform for single-cell RNA-seq.  
135 It retains many of the advantages of STRT-seq, such as the use of unique molecular  
136 identifiers (UMIs) for absolute quantification, and single-read rather than paired-end  
137 sequencing for lower cost. But the transition to an addressable microwell format confers  
138 additional benefits. First, we gained the flexibility to deposit cells by dilution or by FACS,  
139 including by index sorting to track molecular surface properties of each cell and link them to  
140 the final data. We can freely deposit multiple samples (up to eight, currently) on each plate,  
141 and thereby optimize the allocation of sequencing power in complex experimental designs,  
142 something which is difficult using droplet microfluidics. Second, the plate is compatible with  
143 imaging, so that single-cell wells could be verified (albeit currently with a significant false  
144 negative rate due to imperfect cell staining), and fluorescent reporters can be linked to the  
145 final expression profile of each single cell. Third, plates can be filled, frozen – and optionally  
146 shipped – for processing at a later time (Fig. S1a-b, Alternative B). This should prove useful  
147 e.g. as single-cell RNA-seq enters clinical settings, where sample procurement and sample  
148 processing are often performed at distinct sites. Fourth, the low reaction volume and the use  
149 of single read sequencing keep costs low, and we estimate a current cost of approximately  
150 \$1/cell, including 100,000 raw reads. Finally, we note that the open addressable microwell  
151 format, in contrast to droplet microfluidics, could easily be adapted to perform any multistep  
152 protocol currently implemented in regular microtiter plates (as long as they are strictly  
153 additive). This should enable a similar flexibility and throughput for other applications, such  
154 as full-length mRNA-seq (e.g. SMART-seq<sup>12</sup>), whole-genome amplification, and the  
155 detection of chromatin modification<sup>22</sup> and conformation<sup>23</sup>.

156

157

## 158 **Methods**

### 159 *Cell culture*

160 Human Hek293 and chicken DF-1 cells were cultured in complete DMEM medium. Mouse  
161 ES cells<sup>24</sup> were maintained under feeder-free conditions in LIF-2i medium on 0.1% gelatin-  
162 coated culture plates<sup>25</sup>. The cells were trypsinized, washed, counted and assessed for cell  
163 viability.

### 164 *Animals*

165 Male and female wild type CD-1 mice (Charles River) between postnatal days 21-37 were  
166 used. All experimental procedures followed the guidelines and recommendations of Swedish  
167 animal protection legislation and were approved by the local ethical committee for  
168 experiments on laboratory animals (Stockholms Norra Djurförsöksetiska nämnd, Sweden).

169

### 170 *Human post mortem tissue*

171 Postmortem human brain tissue was provided to the Allen Institute for Brain Science by the  
172 San Diego Medical Examiner's (SDME) office after obtaining permission for tissue  
173 collection from decedent next-of-kin. Tissue specimens were de-identified and assigned a  
174 numerical ID, and the Allen Institute for Brain Science obtained the tissue under a legal  
175 agreement that prevents SDME from sharing the key to the code or any identifying  
176 information about tissue donors. The collection and use of postmortem human brain tissue for  
177 research purposes was reviewed by the Western Institutional Review Board (WIRB). WIRB  
178 determined that, in accordance with federal regulation 45 CFR 46 and associated guidance,

179 the use of and generation of data from de-identified specimens from deceased individuals  
180 does not constitute human subjects research requiring IRB review. All tissue collection was  
181 performed in accordance with the provisions of the Uniform Anatomical Gift Act described in  
182 Health and Safety Code §§ 7150, et seq., and other applicable state and federal laws and  
183 regulations. The tissue specimen used in this study was pre-screened for known  
184 neuropsychiatric or neuropathological history, and underwent routine serological testing and  
185 screening for RNA quality (RNA integrity number  $\geq 7$ ).

186

#### 187 *Single cell suspension from mouse cortex*

188 Single cell suspensions from adolescent mouse cortex were generated as described before<sup>9</sup>.  
189 Briefly, mice were anesthetized with isoflurane, perfused with ice-cold aCSF and brains  
190 collected. Brains were then sectioned using a vibratome or brain matrix and the  
191 somatosensory cortex was microdissected. Single cell suspensions were generated using the  
192 Worthington Papain dissociation system, with modifications as described<sup>9</sup>.

193

#### 194 *Isolation, sorting and processing post-mortem adult human neuronal nuclei.*

195 Nuclei were isolated from a -80°C frozen tissue piece taken from the middle temporal gyrus  
196 of the cerebral cortex using Dounce homogenization, as described before<sup>26</sup>. Briefly, the tissue  
197 piece was thawed in homogenization buffer (10mM Tris pH 8.0, 250mM sucrose, 25mM  
198 KCl, 5mM MgCl<sub>2</sub>, 0.1mM DTT, 1x Protease Inhibitor (Promega), 0.4U/μl RNasin Plus  
199 RNase inhibitor (Promega) 0.1% Triton X-100) and gently homogenized with 5-10 gentle  
200 strokes using a loose pestle, followed by 5-10 strokes with a tight pestle. The homogenate  
201 was filtered through a 30μm cell strainer and nuclei pelleted by centrifugation, 10min at 900g.  
202 Nuclei were resuspended and incubated for 15min at 4°C in blocking buffer (1x PBS with  
203 0.5% BSA and 0.2U/μl RNasin Plus RNase inhibitor), an aliquot quality assessed under the  
204 microscope, and stained with conjugated primary mouse-anti-NeuN<sub>PE</sub> antibody, 1:500  
205 (Millipore) rotating at 4°C for 30min. Stained suspensions were washed in blocking buffer,  
206 centrifuged 5min at 450g, transferred to FACS tubes, supplemented with 1μg/μl DAPI and  
207 sorted (FSC/SSC singlets, DAPI+, PE+). Sorted nuclei were frozen at -80°C in PBS with 10%  
208 DMSO and 0.8% BSA.

209 An aliquot of 100 000 NeuN+ sorted nuclei was thawed from -80°C in a 37°C water bath and  
210 quickly transferred to ice. The nuclei were diluted in 3 volumes of dilution buffer (1x PBS  
211 with 0.5% BSA and 0.5U/μl TaKaRa RNase Inhibitor) and centrifuged 5min at 1000g.  
212 Supernatant was carefully removed and the pellet was resuspended in dilution buffer.

213

#### 214 *Cell dispense using Nanodispenser MSND*

215 Viable single cell suspensions were stained with CellTracker Green CMFDA dye (Life  
216 Technologies) according to the manufacturer's instructions, except incubation was 10min on  
217 ice. Suspensions were washed twice (cortex) or three times (cell lines) in respective medium  
218 and cells counted. Human nuclei were stained with Propidium Iodide ReadyProbes (Life  
219 Technologies), according to the manufacturer's instructions. All suspensions were diluted to  
220 20 cells or nuclei/μl in PBS (cell lines), Ca<sup>2+</sup>/Mg<sup>2+</sup>-free aCSF (mouse cortex) or dilution  
221 buffer (human nuclei). 50nl of the suspension were dispensed to all wells.

222

#### 223 *FACS to wells using BD Influx*

224 Before FACS, wells to be used were dispensed with 50-200nl PBS or 50nl lysis buffer and  
225 the plate was kept on ice. Cells were stained with CellTracker Green (as above) and

226 Propidium Iodide ReadyProbes (Life Technologies), according to the manufacturer's  
227 instructions, to discriminate dead cells. For sorting to single wells, we used BD Influx  
228 instrument (for configuration see Table S2). For a higher stability of sorting streams 140µm  
229 and 200µm nozzles were tested for efficiency. As two independent sort experiments with a  
230 200µm nozzle setup demonstrated decreased efficiency (data not shown) the 140µm nozzle  
231 setup was used for further experiments. The gating strategy was set as follows: (1) Population  
232 of cells based on FSC-H x SSC-H profile, (2) Singlets based on FSC-H x FSC-W, (3) Singlets  
233 based on FSC-H x FSC-A. (4) One of the following options: (a) Cell-Tracker Green positive  
234 (530/40 [488nm]) or (b) Cell-Tracker Green positive (530/40[488nm]) and Propidium iodide  
235 negative (585/29[561nm]). Due to software memory limitations, only a quarter of the full  
236 layout (2400 wells) could be set at a time. Initially, two such quarter layouts, covering half  
237 the plate were created. Using the symmetric plate design, it was then turned 180° to fill the  
238 full 9600-well plate. Layouts were aligned prior to each particular experiment using a dumb  
239 plate covered with thin film, to monitor the position of 3-5 drops of Accudrop fluorescent  
240 beads. Given these limitations, we estimate a total plate fill time of below 1.5 hours.

241

#### 242 *Imaging and Cell selection*

243 For correct imaging positioning, fiducial fluorescent stain or highly concentrated stained cell  
244 suspension was dispensed to corner wells during cell dispense (MSND) or before FACS sort.  
245 The dispensed plate was sealed with MicroAmp Optical Adhesive Film (Applied  
246 Biosystems), centrifuged 3min at 200g and mounted upside-down on automated Nikon  
247 ECLIPSE Ti. All wells were imaged in FITC channel, using a 4x objective (4-by-4 wells per  
248 frame). Imaging took less than 15 minutes, during which the plate was cooled using ice packs,  
249 and then immediately placed on ice during image analysis (Fig. S1a Alt A), or frozen on -  
250 80°C (Fig. S1a Alt B). Imaging files were loaded to the CellSelect Software (WaferGen) with  
251 single cell containing wells selected using varying parameters, depending on the cells used. A  
252 quick manual inspection of included and excluded wells was carried out, and if needed,  
253 analysis parameters (such as Expected Cell Size, Circularity, Brightness) were adapted. If the  
254 plate was held on ice for further processing, a maximum of 7 minutes was allowed per  
255 analysis. A final list of single cell well candidates was saved as a Filter File for dispense of all  
256 downstream reagents.

257

#### 258 *Lysis and reverse transcription*

259 If the plate was immediately processed (Fig. S1a Alt A), 50nl lysis mix (500nM STRT-P1-  
260 T31, 4.5nM dNTP, 2% Triton-X-100, 20mM DTT, 1.5U/µl TaKaRa RNase Inhibitor) was  
261 dispensed, followed by 3min lysis at 72°C. Then, 85nl reverse transcription (RT) mix (2.1X  
262 SuperScript II First-Strand Buffer, 12.6mM MgCl<sub>2</sub>, 1.79M betaine, 14.7U/µl SuperScript II,  
263 1.58U/µl TaKaRa RNase Inhibitor, 10.5µM P1B-UMI-RNA-TSO) were dispensed and RT  
264 carried out 42°C for 90 minutes.

265 If the plate had been stored on dry ice or at -80°C for later processing (Fig. S1a Alt B), it first  
266 was thawed to room temperature, followed by dispense of 70nl Lysis-RT mix (1.62X  
267 SuperScript II First-Strand Buffer, 10.2mM MgCl<sub>2</sub>, 1.36M betaine, 425nM STRT-P1-T31,  
268 3.4mM dNTP, 3.4% Triton-X-100, 11.9mM DTT, 8.5U/µl SuperScript II, 1.28U/µl TaKaRa  
269 RNase Inhibitor, 5.1µM P1B-UMI-RNA-TSO) and reverse transcription at 42°C for 90  
270 minutes.

271 After each dispense and incubation step the plate was centrifuged for 1 minute at maximum  
272 speed (>2000g) to ensure proper collection and mixing of the reagents. For all array sealing,



273 except during imaging, MicroSeal A film (BioRad) was used.

274

#### 275 *Indexed PCR and extraction*

276 After reverse transcription, 32 index primers (DI-P1A-idx[1-32]-P1B) for PCR were  
277 dispensed, such that each candidate well per 10x10 subarray received a unique index. Primer  
278 dispense was carried out in a 100nl dispense step with 2% bleach washes between each set of  
279 primers, and achieving 200nM final primer concentration in the PCR reaction. PCR mix (final  
280 concentration 1X KAPA HiFi Ready Mix supplemented with 0.2mM dNTP, 100nM DI-PCR-  
281 P1A) was dispensed in 565nl (Alt A) or 430nl (Alt B), with stock concentrations adapted  
282 accordingly. PCR was run as follows: 95°C 3min. 5 cycles: 98°C 30sec, 67°C 1min, 72°C  
283 6min. 15 cycles: 98°C 30sec, 68°C 30sec, 72°C 6min. 72°C 5min, 10°C hold. After PCR, an  
284 extraction block was mounted on a clean 96-well plate. On top, the plate was mounted, upside  
285 down, to align with the extraction block. The assembly was centrifuged 5min at maximum  
286 speed (>3000g). The 96-well plate containing pooled, index amplified cDNA was assessed  
287 for quality on Bioanalyzer, sealed and stored at -20°C.

288

#### 289 *Tagmentation and isolation of 5' fragments*

290 Amplified cDNA was tagmented using 96 transposomes, each with a different index to target  
291 cDNA from one subarray each. Transposome stocks were assembled (6.25µM barcoded  
292 adapter (STRT-Tn5-Idx[1-96]), 6.25µM Tn5 transposase, 40% glycerol), 37°C 1h, and  
293 concentration adapted according to Tn5 activity, if needed. Tn5 reactions were assembled  
294 with 3µl transposome and 2µl amplified cDNA, in a total 20µl 1x CutSmart buffer (NEB),  
295 and incubated at 55°C 20min. 100µl Dynabeads MyOne Streptavidin C1, washed according  
296 to the manufacturer's instructions, were diluted 1:20 in BB buffer (10mM Tris HCl pH 7.5,  
297 5mM EDTA, 250mM NaCl, 0.5% SDS), added to the tagmentation reaction 1:1, and  
298 incubated at room temperature 15min. After incubation, all samples were pooled to one tube,  
299 washed twice in TNT (20mM Tris HCl pH 7.5, 50mM NaCl, 0.02% Tween-20) and  
300 resuspended in 50µl TNT. Remaining adapters were cleaned by adding 10µl ExoSAP IT  
301 (Affymetrix) and incubating 15min 37°C, followed by two more washes in TNT and one  
302 careful wash in EB. The single-stranded library was eluted in 50µl nuclease-free water by  
303 incubating 10min at 70°C and collecting the supernatant to a new tube. The library was then  
304 bound to 1.5X volumes AMPure beads (Beckman), supernatant removed, the beads  
305 resuspended in Second PCR mix (1X KAPA HiFi Ready Mix supplemented with 200mM  
306 4K-P1\_2ndPCR, 200nM P2\_4K\_2ndPCR), and cycled (95°C 2min. 8 cycles: 98°C 30sec,  
307 65°C 10sec, 72°C 20sec. 72°C 5min, 10°C hold). The supernatant was transferred, cleaned  
308 with 0.7X volumes of AmPure beads and eluted. The eluate was bound to 0.5X volumes of  
309 AmPure beads, the supernatant transferred, again bound in 1X volume of AmPure beads, and  
310 eventually eluted in EB.

311

#### 312 *Illumina high-throughput sequencing*

313 The quality and molar concentration of libraries were assessed by Bioanalyzer. Sequencing  
314 was performed on Illumina HiSeq2000 or 2500 with Single-End 50 cycle kit using the Read1  
315 DI-Read1-Seq, Index 1 STRT-Tn5-U and Index 2 DI-idxP1A-Seq. Reads of 45 bp were  
316 generated, expected to start with a 6 bp UMI, followed by 3 guanidines, and the 5' transcript.  
317 The two index reads of 8 and 5 bp corresponded to Index 1 (subarray barcode) and Index 2  
318 (well barcode), respectively.

319

320 *Data analysis*

321 Reads flagged as invalid by the Illumina HiSeq control software were discarded. In the  
322 remaining reads, any 3' bases with a quality score of B were removed. If any of the six UMI  
323 bases in the 5' end had a Phred score <17 the read was discarded, else the UMI was cut away  
324 and saved. The following three bases had to be G for the read to be kept. These, as well as  
325 any additional up to totally nine (presumably template-switch derived) G:s were cut away.  
326 Reads were discarded if the remaining transcript-derived sequence ended in a poly(A)  
327 sequence leaving <25 bases, or if the remaining sequence consisted of fewer than six non-A  
328 bases or a dinucleotide repeat with fewer than six other bases at either end. The reads were  
329 sorted by the well, as defined by the combined two index read barcodes.

330 Alignment was performed using the Bowtie aligner<sup>27</sup>, allowing for up to three mismatches  
331 and up to 24 alternative mappings. Reads with no alignments were realigned against an  
332 artificial chromosome, containing all possible splice junctions arising from the exons defined  
333 by the UCSC transcript models<sup>28</sup>. The coordinates of aligned splice junctions were translated  
334 back to the corresponding genomic positions.

335 For expression level calculation, the exons of each locus with several transcript variants were  
336 merged to a combined model representing total expression of the locus. To account for  
337 incomplete cap site knowledge, the 5' ends of all models were extended by 100 bases, but not  
338 beyond the 3' end of any upstream nearby exon of another gene of the same orientation.

339 The annotation step was performed separately for each well. For each genomic position and  
340 strand combination the number of reads in each UMI was counted. Any multiread that  
341 mapped to some repeat outside exons was assigned randomly as one of these repeats and did  
342 not contribute to the transcript corresponding to the exon. Else, if the multiread mapped to  
343 some exon, and not to any repeat outside exons, it was assigned to the exon where it was  
344 closest to the transcript model 5' end. If it had no exon mapping, it was assigned randomly at  
345 one of the mappings. The total number of molecules at each mapping position was  
346 determined by the number of distinct UMIs observed. Any UMI represented by only a single  
347 read was excluded, in order to reduce false molecules due to PCR and sequencing errors. The  
348 raw UMI count was corrected for the UMI collision probability as described<sup>29</sup>.

349 The human nuclei samples were analyzed similarly, but all reads mapping anywhere within  
350 the whole locus, defined as the region (including actual introns) from the start of the 100 base  
351 5'end extension of the first exon up to the end of the last exon, were counted as exon-derived.

352

353 *Clustering and analysis mouse cortex cells*

354 Analysis of mouse cortex samples included the following steps: (1) Loaded all 7769 cells. (2)  
355 Filtered on 800-2000 total molecules per cell and ratio total molecules/total genes >1.2,  
356 resulting in 6449 cells. (3) Excluded doublets identified by co-expression of known marker  
357 genes (*Stmn2* – neurons, *Mog* – oligodendrocytes, *Aqp4* – astrocytes, *Fnl* – endothelial,  
358 *Clqc* – microglia), 5514 cells retained. (4) Permuted order of cells and genes. (5) Clustering  
359 by BackSPINv2 with following parameters: splitlev = 7; Nfuture = 300; Nfuture1 = 500;  
360 N\_to\_backspin = 10; N\_to\_cells = 500; mean\_tresh = 0.01; fdr\_th = 0.3; min\_gr\_cells = 5;  
361 min\_gr\_genes = 10; stop\_th = [0.5,0.5]; flag\_val\_stip = 1;. (6) Manual inspection of  
362 clustering results and separation of cells to neurons and non-neurons. Clusters that showed no  
363 specific marker, interpreted as low quality, and cells originating from a Hek293 sample



364 (Table S1) were removed (1882 cells removed, Fig S4a) (7) Separate clustering of neurons  
365 and non-neurons with following parameters: splitlev = 6; Nfuture = 200; Nfuture1 = 800;  
366 N\_to\_backspin = 10; N\_to\_cells = 500; mean\_tresh = 0.01; fdr\_th = 0.3; min\_gr\_cells = 5;  
367 min\_gr\_genes = 20; stop\_th = [0.5,0.5]; flag\_val\_stip = 1;. (a) For neurons, this resulted in 35  
368 clusters, we excluded 12 clusters for which no specific marker was obtained and merged  
369 some of the remaining 23 clusters into final 13 clusters (Fig S4b). (b) For only non-neurons  
370 the same BackSPINv2 parameters resulted in 33 clusters, we excluded 7 clusters without  
371 specific marker and merged some of the remaining 26 clusters into 17 final clusters (Fig S5).  
372 tSNE projection<sup>30</sup> was used for visualization only. We used the following parameters: 410  
373 genes, perplexity 20, PCA components 20, epsilon 100, distance correlation number of  
374 iterations 1000 (Matlab code <https://lvdmaaten.github.io/tsne/>). In the heatmap (Fig S4) we  
375 used the same 410 genes. Color on heatmaps represent normalized expression (log-transform  
376 per gene: mean=0, standard deviation=1) and are saturated 1-99%.

377

### 378 *Clustering and analysis human cortex nuclei*

379 Analysis of human cortex nuclei included the following steps: (1) Loaded all 2842 cells. (2)  
380 Filtered out cells with less than 500 detected molecules (2306 cells retained). (3) Removed  
381 genes expressed in less than 20 cells, or more than 60% of cells. (4) Normalized each cell to  
382 total 3000 molecules. (5) Clustering by BackSPINv2 using the following parameters: splitlev  
383 = 6;Nfuture = 500;Nfuture1 = 500;N\_to\_backspin = 10;N\_to\_cells = 800;  
384 mean\_tresh = 0.01; fdr\_th = 0.3;min\_gr\_cells = 5;min\_gr\_genes = 5;stop\_th =  
385 [0.5,0.5];flag\_val\_stip = 2. Clustering resulted in 31 clusters. (6) One cluster expressing glial  
386 genes was removed, the rest were manually merged into 11 clusters.  
387 For visualization, we selected top enriched genes, as described above. tSNE visualization was  
388 run as above, with the following parameters: initial PCA dimension = 20, perplexity = 2,  
389 epsilon = 100, correlation as distance, maximum iterations = 1000.

390

### 391 *Two-species analysis*

392 A combined two-species bowtie-1<sup>27</sup> alignment index was constructed from transcript models  
393 as defined by the UCSC refFlat table data<sup>31</sup>. In order to obtain equally-sized and similar  
394 representations of the two transcriptomes to compare, we naively restricted the analysis to the  
395 transcripts that had identical names in the two species (disregarding upper/lower case).  
396 Illumina HiSeq reads were processed as follows: Reads ending in a poly(A) sequence leaving  
397 less than 25 alignable 5' bases were discarded. Any 3' bases with a quality score of B were  
398 removed. If the remaining sequence consisted of fewer than six non-A bases or a dinucleotide  
399 repeat with fewer than six other bases at either end, the read was discarded.  
400 The filtered reads were aligned to the bowtie-1 index allowing for up to 3 mismatches. A read  
401 was considered unequivocally belonging to a species and counted if it had a perfect match to  
402 a transcript of that species, but no match in the other species, also when allowing up to 3  
403 mismatches. Molecule counts were obtained as the number of distinct six nucleotide long  
404 unique molecular identifiers (UMIs) of the reads aligned at each position. UMIs represented  
405 by a single read were not counted, since our previous experiments have shown that these to a  
406 large extent are artifacts stemming from PCR and sequencing errors and result in  
407 overestimates of molecule counts.

408

409

410 *Primer sequences*

411

<b>Lysis / reverse transcription</b>	STRT-P1-T31	5' Bio-AATGATACGGCGACCACCGATCG- TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
	P1B-UMI-RNA-TSO	5' Bio-rCrTrArCrArCrGrArCrGrCrTrCrTrTrCrCrGrArTrCrT- rNrNrNrNrNrN-rGrGrG
<b>PCR 1</b>	DI-PCR-P1A	5' Bio-AATGATACGGCGACCACCGA
	DI-P1A-idx[1-32]-P1B	5' Bio-AATGATACGGCGACCACCGAGATCTACAC-XXXXX- CTACACGACGCTCTCCGATC
<b>Tagmentation</b>	STRT-Tn5-Idx[1-96]	CAAGCAGAAGACGGCATAACGA-YYYYYYYYY- GCGTCAGATGTGTATAAGAGACAG
	STRT-TN5-U	5' PHO-CTGTCTCTTATACACATCTGACGC
<b>PCR 2</b>	P1_2nd_PCR	AATGATACGGCGACCACCGAGATC
	P2_2nd_PCR	CAAGCAGAAGACGGCATAACGAGAT
<b>Sequencing</b>	DI-Read1-Seq	ATGATACGGCGACCACCGAGATCTACAC-NNNNNN- CTACACGACGCTCTCCGATCT
	STRT-Tn5-U	5' PHO-CTGTCTCTTATACACATCTGACGC
	DI_idxP1A-Seq	AATGATACGGCGACCACCGAGATCTACAC

412

413 **Author contributions**

414 A.Z, M.S, S.L, H.Ho, H.Hu, G.L.M and J.D conceived and designed the method. H.Hu, P.Li,  
415 J.D, S.H, M.S and A.Z designed and engineered the microarray. H.Ho, R.H, J.M, A.H and  
416 A.Z performed experiments. M.R, R.H and E.L provided materials. H.Ho, H.Hu, P.Lö, A.Z  
417 and S.L analyzed data. H.Ho, A.Z and S.L drafted the manuscript, with input from all authors.  
418

419 **Acknowledgements**

420 We thank Feng Zhang for mouse ES cells. We thank Anna Johnsson for lab management and  
421 Anna Juréus for technical assistance and sequencing.  
422

423 **Competing Financial Interests**

424 S.L, H.Ho, P.Lö, S.P, G.L.M. and A.Z. are co-inventors of the method, for which a patent  
425 application has been submitted by WaferGen Inc., and may receive license or royalty  
426 payments. M.S, S.H, J.D, P.Li and H.Hu are employees of WaferGen Inc. R.H, J.M, A.H,  
427 M.R and E.L declare no competing financial interests.  
428

428

429

430

## 431 References

432

- 433 1. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell  
434 RNA-seq. *Science (80-. )*. **347**, (2015).
- 435 2. Tasic, B. *et al.* Adult cortical cell taxonomy by single cell transcriptomics. *Nat.*  
436 *Neurosci.* (2016). doi:10.1038/nn.4216
- 437 3. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human,  
438 and Stem Cells. *Cell* **167**, 566–580.e19 (2016).
- 439 4. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-  
440 cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
- 441 5. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid  
442 Progenitors. *Cell* **163**, 1663–1677 (2015).
- 443 6. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human  
444 oligodendrogloma. *Nature* **539**, 309–313 (2016).
- 445 7. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-  
446 cell RNA-seq. *Science (80-. )*. **352**, (2016).
- 447 8. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using  
448 single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
- 449 9. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central  
450 nervous system. *Science (80-. )*. **352**, (2016).
- 451 10. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular  
452 heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat.*  
453 *Biotechnol.* **32**, 1053–1058 (2014).
- 454 11. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat.*  
455 *Methods* **11**, 163–166 (2013).
- 456 12. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single  
457 cells. *Nat. Methods* **10**, 1096–1098 (2013).
- 458 13. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free  
459 Decomposition of Tissues into Cell Types. *Science (80-. )*. **343**, (2014).
- 460 14. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by  
461 Multiplexed Linear Amplification. *Cell Rep.* **2**, 666–673 (2012).
- 462 15. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual  
463 Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
- 464 16. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to  
465 Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).
- 466 17. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells.  
467 *bioRxiv* (2016).
- 468 18. Habib, N. *et al.* Div-Seq: A single nucleus RNA-Seq method reveals dynamics of rare  
469 adult newborn neurons in the CNS. *bioRxiv* 1–20 (2016). doi:10.1101/045989
- 470 19. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA  
471 sequencing of the human brain. *Science* **352**, 1586–90 (2016).
- 472 20. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature*  
473 **445**, 168–176 (2007).
- 474 21. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain  
475 transcriptome. *Nature* **489**, 391–399 (2012).
- 476 22. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of  
477 regulatory variation. *Nature* **523**, 486–490 (2015).

- 478 23. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome  
479 structure. *Nature* **502**, 59–64 (2013).
- 480 24. Platt, R. J. *et al.* CRISPR-Cas9 knockin mice for genome editing and cancer modeling.  
481 *Cell* **159**, 440–55 (2014).
- 482 25. Koehler, K. R. & Hashino, E. 3D mouse embryonic stem cell culture for generating inner  
483 ear organoids. *Nat. Protoc.* **9**, 1229–1244 (2014).
- 484 26. Krishnaswami, S. R. *et al.* Using single nuclei for RNA-seq to capture the transcriptome  
485 of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).
- 486 27. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient  
487 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- 488 28. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013.  
489 *Nucleic Acids Res.* **41**, D64-9 (2013).
- 490 29. Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. A. Counting individual DNA molecules by  
491 the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 9026–31  
492 (2011).
- 493 30. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**,  
494 2579–2605 (2008).
- 495 31. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic*  
496 *Acids Res.* **43**, D670-81 (2015).
- 497
- 498