bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene-expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ

Joseph W. Foley[1,2*], Chunfang Zhu[1], Philippe Jolivet[2,3], Shirley X. Zhu[1], Peipei Lu[1], Michael J. Meaney[2], and Robert B. West[1*]

[1]Department of Pathology, Stanford University School of Medicine, Stanford, California, USA
[2]Ludmer Centre for Neuroinformatics and Mental Health, Douglas Hospital Research Centre, McGill University, Montreal, Quebec, Canada
[3]Department of Biochemistry, Université de Montréal, Montreal, Quebec, Canada
[*]Correponding author

January 10, 2018

Abstract

RNA sequencing (RNA-seq) is a sensitive and accurate method for quantifying gene expression. Small samples or those whose RNA is degraded, such as formalin-fixed, paraffin-embedded (FFPE) tissue, remain challenging to study with nonspecialized RNA-seq protocols. Here we present a new method, Smart-3SEQ, that accurately quantifies transcript abundance even with small amounts of total RNA and effectively characterizes small samples extracted by laser-capture microdissection (LCM) from FFPE tissue. We also obtain distinct biological profiles from FFPE single cells, which have been impossible to study with previous RNA-seq protocols, and we use these data to identify possible new macrophage phenotypes associated with the tumor microenvironment. We propose Smart-3SEQ as a highly scalable method to enable large gene-expression profiling experiments unconstrained by sample size and tissue availability. In particular, Smart-3SEQ's compatibility with FFPE tissue unlocks an enormous number of archived clinical samples, and combined with LCM it allows unprecedented studies of small cell populations and single cells isolated by their *in situ* context.

Introduction

Omnis cellula e cellula (every cell from a cell): this idea promoted by Rudolf Virchow, the grandfather of modern pathology, rings no less true today than 160 years ago [1]. Cell-centered concepts are emerging that suggest it is not just the genomic alterations of a cell but its developmental state that are critical to our understanding of disease. Technologies to study individual single cells, not complex aggregates of multiple cell types, will give us a greater understanding of the biology of many diseases. Multiple large-scale collaborative projects are underway to understand individual cell biology in both normal and pathological tissues [2, 3]

Single-cell gene-expression profiling is essential for these endeavors. Prominent new methods rely on fresh tissue to generate disaggregated intact single cells that can be profiled using microfluidic technologies [4, 5]. These approaches are generating extensive information on new cell types and cell-specific transcriptome profiles. Such information will have a significant impact on the clinical management of disease. However, this impact is diminished by the limitations of fresh clinical material. Fresh clinical material is difficult to obtain for research, cannot be banked, and the selection of material for analysis is based on *a priori* clinical knowledge and gross examination with no cellular context. Moreover, the isolation process often leads to selective degradation of specific cell types leading to a bias in the profiled cell populations. For these reasons, it is not suitable for many clinical study designs.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We have developed an RNA-seq method that addresses these problems. Our method, Smart-3SEQ, is robust in archival formalin-fixed, paraffin-embedded (FFPE) material and allows microscopic selection of specific cells. It combines the template-switching SMART method [6, 7] and protocol optimizations of Smart-seq2 [8] with the streamlined 3′ end–targeting approach of 3SEQ [9], yielding a simpler procedure than any of those. In addition, Smart-3SEQ incorporates unique molecular identifiers to increase the accuracy of transcript counting with low input amounts [10]. This method has conceptual similarities to the prototype RNA-seq method used to demonstrate UMIs, but with a streamlined protocol and optimizations for laser capture microdissection and FFPE material. Thus our protocol is cost-effective in reagent usage and working time, sensitive to single-cell amounts of RNA, and robust to degraded samples. Here we demonstrate that Smart-3SEQ quantifies transcript abundance accurately with a wide range of amounts of input, and in particular it can create libraries from single cells dissected out of tissue with degraded RNA, enabling the study of cells in the tumor microenvironment.

Results

The Smart-3SEQ method

Previous single-cell RNA-seq protocols, Smart-seq [7] and Smart-seq2 [8], generate cDNA libraries by the SMART method [6]: they take advantage of Moloney murine leukemia virus–derived reverse transcriptase's "template switching" ability to generate full-length, double-stranded cDNA molecules from RNA templates in a single incubation. However, when the original RNA is degraded, there may not be full-length molecules available to reverse-transcribe, so RNA integrity is very important with these methods. In contrast, the 3SEQ method [9] is optimal for degraded RNA such as that from FFPE tissue. In 3SEQ, the RNA is fragmented before reverse transcription, eliminating the difference between intact and degraded samples. Reverse transcription is still primed by an anchored oligo(dT) primer, so from each transcript only the fragment containing the beginning of the poly(A) tail is sequenced; this approach, also called digital gene expression, generates a short identifying "tag" sequence rather than read the full length of the transcript. Smart-3SEQ combines this 3′-targeted approach with template-switching in a streamlined and sensitive protocol (Figure 1; see Supplementary File 1 for a technical description).

Smart-3SEQ has numerous practical advantages for many different applications of gene-expression profiling independent of its use with FFPE tissue (Table S1). Because of the method's efficiency, with no lossy dsDNA ligation and only one lossy cleanup step, it is sensitive at very small amounts of starting RNA. The short number of steps and small reaction volumes allow a batch of libraries to be prepared in half a day with high throughput in a 96-well plate, at a reagent cost of about 5 USD per sample (Supplementary File 6). Library yields are predictable enough that the optimal number of PCR cycles can be reliably estimated from the amount of input material over five orders of magnitude. Furthermore, because Smart-3SEQ reads only a single fragment of each transcript, the one containing the polyadenylation site, it is not sensitive to transcript length, which is a confounder for whole-transcript RNA-seq [11]. Thus Smart-3SEQ allows simple and accurate quantification of transcript abundance; however, it cannot report information about splicing or genotypes, unless the splice junction or polymorphism occurs near the end of the gene, and it cannot detect non-polyadenylated transcripts. Like 3SEQ, because RNA fragmentation is the first step, it is also robust to pre-fragmented RNA, including damaged RNA from FFPE samples.

Validation with reference samples

We validated the accuracy and sensitivity of the Smart-3SEQ method by using it to quantify the reference RNAs used in the SEQC Consortium's benchmarking study [12], though with lower amounts of RNA input. In the first experiment, we used ERCC Mixes 1 and 2, which contain 92 *in vitro*–transcribed, polyadenylated transcripts at known concentrations, which span 6 orders of magnitude and differ between the two mixes [13]. We prepared tenfold serial dilutions at $1/10$–$1/100,000$ the stock concentration, then created a library from 1 µL of each (10.4 fmol–1.04 amol, 6.23 billion–623,000 molecules); at each dilution level we tested two different numbers of PCR cycles (Figure S4).

The proportion of reads aligned to the reference sequences decreased mainly with RNA input, but the number of PCR cycles also made a large difference (Figure S5). 99.99% of aligned reads were in the expected

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

forward-strand orientation. The abundance of each transcript as measured by Smart-3SEQ corresponded linearly with its expected copy number in the sample (Figure 2A; average $r' = 0.990$ in the highest-input, lowest-PCR samples; see Methods for definition of r'). As the amount of RNA input decreased, the accuracy of the measurements worsened for low-abundance transcripts (Figure S6; for 1 amol RNA, average $r' = 0.926$). Likewise the replicate correlations were very high for high inputs but decreased for low inputs (Figure S7), and the signal loss compared to the first dilution increased with subsequent dilutions (Figure S8). However, in the lowest-input samples the expected copy number of several transcripts was less than 1, so we expect considerable sampling error even in the true copy numbers of low-abundance transcripts, which is magnified by the recursive sampling of the serial-dilution process.

In a second validation experiment, we used the Human Brain Reference RNA (HBRR) and the Universal Human Reference RNA (UHRR), which were also used in the SEQC benchmarks [12]. We mixed in the ERCC standards at the same relative concentration used by SEQC (1/500 ERCC stock concentration with 100 ng/µL human RNA) to make our RNA samples comparable with theirs (Figure S9). As in the previous experiment, we saw a high proportion of alignable reads in the high-input samples, which gradually declined as the number of PCR cycles approached 20 and then declined more steeply (Figure 2C). The true abundances of the transcripts in these human samples are not known *a priori*, but it is still possible to validate Smart-3SEQ's measurements by comparing them with those of an alternative method, qPCR, on the same reference RNA samples. Smart-3SEQ showed a strong correlation (average Spearman's $\rho = 0.845$ for high-input samples; Figure S10) with TaqMan-based qPCR measurements of 999 genes [14] and with SYBR-based qPCR of 20,801 genes [12] (average high-input $\rho = -0.827$; Figure S11).

We compared the accuracy of Smart-3SEQ side-by-side with the SEQC data from standard RNA-seq performed on the same ERCC transcripts and human reference RNAs. Smart-3SEQ showed equivalent accuracy to RNA-seq in the ERCC standard curves (Figures S12, S13) and equivalent correspondence with both SYBR- and TaqMan-based qPCR (Figures S14, S15). Though the direct correlation between Smart-3SEQ and RNA-seq was lower (Figure S16), the qPCR platforms were equally correlated with each other as they were with the sequencing platforms (Figure S17). Thus we conclude that although Smart-3SEQ and RNA-seq may have different error profiles, they quantify biological signal equally well.

Demonstration with microdissected cells

To take advantage of Smart-3SEQ's compatibility with small amounts of degraded RNA, we used the method to study single-cell gene expression in the context of the tumor microenvironment, which is a mix of benign and malignant cells. We compared the gene expression measured in single cells to the gene expression obtained from bulk samples (100–500 cells) of homogeneous cell populations as identified by cytology and location within the tumor microenvironment. This is similar to the approach that others have used to assess the quality of single-cell RNA-seq libraries [15]. The previous 3SEQ method showed a strong correlation between gene-expression data obtained from matching fresh-frozen and FFPE material [9]. As such, we used bulk samples from FFPE as the reference point to compare with single-cell FFPE gene expression.

We reviewed FFPE tissue from a mastectomy specimen obtained from a patient with ductal carcinoma *in situ* (DCIS). We microdissected single-cell and bulk samples from ducts involved by DCIS and from an adjacent focus of stromal macrophages, 6 bulk tissue samples and 10 single cells of each type plus 6 no-cell controls (Figure 3; see Supplementary File 1 for validation of the dissection method). Unlike fresh-tissue single-cell RNA-seq, in which the tissue is processed without first visualizing the cells or their context within the tissue, the use of LCM allowed us to microscopically identify specific areas and cells to capture within the mastectomy specimens. This effectively allowed us to pick specific single cells of interest from the millions of cells present in the clinical archival specimen. Based on the above serial-dilution experiments with the human reference RNAs, a sample of 500 cells is well beyond enough input for high-quality data from a bulk library. This number of cells can also be obtained from a sections of a single breast duct involved by DCIS or a nearby collection of macrophages.

There are minor differences between the quality of the single-cell and bulk libraries. 27% of reads from the bulk libraries were uniquely alignable (Figure 4A); this is presumably lower than the yield from intact reference RNA because the RNA in FFPE tissue is fragmented below the desired length. The percentage of uniquely alignable reads was 17% in the single-cell libraries, likely lower because of the increased abundance of PCR primer dimers in samples with very low input. However, only 2% of reads from the negative-control

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

libraries were uniquely alignable.

To assess the degree to which Smart-3SEQ's gene-expression profiles of dissected single cell can recapitulate the information from bulk cell populations, we first determined which genes were significantly differentially expressed (multiple testing–adjusted $p_{\mathrm{adj}} < 0.05$) between bulk stromal macrophage and ductal tissue samples. As expected, the gene-expression profiles between ductal cells and stromal macrophages are considerably different and identify biomarkers indicative of DCIS and macrophage populations. The top 100 significantly differentially expressed genes, according to only the bulk samples, are shown as a heatmap to illustrate the differences (Figure 4B). The significantly differentially expressed genes include common clinical biomarkers for macrophages (*CD68* and *CD163*) and DCIS cells (*EPCAM*, *KRT7*, *KRT18*, and *ERBB2* (*HER2*)) (Figure 4C, Supplementary File 3). The single-cell samples generally have gene-expression profiles similar to the cell type–matched bulk samples. All the gene-expression patterns of single stromal macrophages match the bulk stromal macrophages. No stromal macrophages match ductal-cell signal. Remarkably, five ductal cells demonstrate a gene-expression pattern similar to the bulk stromal macrophages.

To explore whether we can use the single-cell data to identify distinct populations of cells, we performed t-SNE, a machine-learning technique that visualizes multidimensional data in two dimensions [16], on the entire dataset including negative controls (Figure 4D). t-SNE shows that the five ductal cells whose gene-expression profiles resemble those of macrophages also cluster with the stromal macrophage cells and the bulk stromal macrophage samples. The other five ductal cells cluster with the bulk ductal tissue samples while the control samples cluster distinctly from all others. These findings suggest that the bulk ductal tissue samples are heterogeneous with at least two different cell types, DCIS cells and intraductal macrophage cells, which are resolved with single-cell dissection but not bulk tissue homogenization. Immunohistochemistry for a macrophage marker on the duct used for dissection (*CD163*) confirms that there are macrophages alongside DCIS cells (Figure 3C).

To further clarify the cell status, we note that the DCIS cells from this mastectomy have an amplification of a region including the *ERBB2* (*HER2*) locus, which encodes an important therapeutic target. Examination of gene expression from the single and bulk DCIS cells demonstrates a coordinated increase in signal at a possible amplicon surrounding the *ERBB2* locus (Figure 5). This amplification appears in all of the bulk DCIS samples as well as the single ductal cells whose expression profiles match the bulk DCIS, but the amplification is absent in the bulk macrophage samples and in the single ductal cells that match the macrophage profile.

There are 52 genes that were significantly differentially expressed ($p_{\mathrm{adj}} < 0.05$) between single stromal macrophages and single ductal cells inferred to be macrophages (Supplementary File 3). Many of the most significant genes are related to extracellular matrix interactions and leukocyte signaling. These data raise the possibility that intraductal macrophages have a phenotype that is distinct from stromal macrophages, and further highlight the ability of single-cell LCM and Smart-3SEQ to dissect complex cell populations that would otherwise be missed in bulk analysis.

Discussion

Smart-3SEQ is a streamlined, sensitive, and robust method for gene-expression profiling that compares favorably with previous methods. The protocol has fewer steps and takes less time than any other RNA-seq method to date, and achieves a very low cost by using minimal volumes of common reagents. Here we show that Smart-3SEQ quantifies transcript abundance accurately across at least five orders of magnitude of RNA input amounts, from common working concentrations down to single cells, including those with degraded RNA. This accuracy is maintained even with inexpensive sequencing conditions (Figures S19, S20). These attributes make Smart-3SEQ useful for a variety of biological research.

Combining Smart-3SEQ with laser-capture microdissection (LCM Smart-3SEQ) enables a new kind of single-cell genomics research on clinical FFPE samples. In most "single-cell" genomics experiments, a piece of tissue is homogenized and thousands of cells are used to make individual sequencing libraries, but the positional identities of the cells are lost in the context of the tissue and must be inferred from their genomic data—we propose to call this *reverse RNA-seq* (Figure 6), by loose analogy to reverse genetics. In contrast, LCM Smart-3SEQ allows for standard *forward RNA-seq* on specific single cells: the cells of interest can be identified under the microscope by markers such as morphology, location, or biomarker staining, and

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

this prior classification is already known when the cells are profiled for gene expression. Our study is an illustration of the knowledge gained: we identified two sets of macrophages, one present in the stroma and one present within the duct associated with ductal carcinoma *in situ*. Conventional single-cell sequencing would fail to distinguish these two populations: t-SNE analysis groups these two populations of cells together.

We have demonstrated the success of our method for single-cell expression profiling on a number of levels: 1) our approach is efficient: we have obtained a clear gene-expression profile from every one of our single-cell libraries; 2) the single-cell libraries recapitulate bulk gene-expression data; 3) we identified two distinct cell-type signatures, the DCIS profile and the macrophage profile, with clinically recognized biomarkers of each cell type; 4) the data are sufficiently quantitative to be used in conventional single-cell RNA-seq approaches; 5) the data are sufficiently quantitative that we can detect a DNA copy-number change; and 6) we can use the data to discover potentially new cell phenotypes, such as stromal macrophages vs. ductal macrophages, and describe gene-expression profile differences between the two cell subtypes. There are limitations to LCM Smart-3SEQ. LCM is labor-intensive and this limits the number of cells that can be collected and observed. The quality of RNA from FFPE material is poor. Recent FFPE clinical material that has been well archived can have an RNA integrity number (RIN) above 3. However, RNA quality from FFPE material decreases with time and decades-old FFPE material will yield degraded RNA that is challenging for the LCM Smart-3SEQ method. As single cells already have a limited amount of RNA, the age of the FFPE specimen may influence the cell-to-cell variation in absolute RNA expression and cell profile differences will be altered. The RNA obtained from a single cell may represent a fraction of the total possible RNA as the LCM may only capture part of the cell.

Our approach addresses a number of experimental-design problems that are difficult to overcome with conventional single-cell RNA-seq. FFPE archival clinical tissue cannot be examined by conventional single-cell RNA-seq as the cells cannot be physically dissociated. Non-archival tissue samples, both fresh and frozen, are difficult to collect for clinical studies, especially those that require large numbers of samples for assessing clinical outcomes. LCM Smart-3SEQ also enables studies on uncommon or microscopic lesions that are difficult to collect as fresh or fresh-frozen material. Thus, a much wider variety of cells and diseases can be assessed using archival material with Smart-3SEQ. There are also technical advantages of LCM Smart-3SEQ due to the ability to bank isolated single cells for future experiments. This includes the spiking in of known cells to control for batch effects with biological replicates. There is also potential bias in the representation of cell populations in fresh single-cell RNA-seq if some cell types survive the dissociation steps better than others.

The LCM Smart-3SEQ method creates a new histologically focused approach to studying small cell populations and individual cells. Cells of interest can be chosen by morphology, microenvironment location, or *in situ* biomarker status and then profiled to uncover previously unappreciated heterogeneity in gene expression, including subtle changes that might escape reverse RNA-seq analyses. This approach will allow transcriptome profiling to distinguish each cell from all others.

Materials and methods

Reference RNA preparation

ERCC standards and Human Brain Reference RNA were purchased from Thermo Fisher Scientific (catalog #4456739 and #AM6050 respectively). Universal Human Reference RNA was purchased from Agilent Technologies (#740000) and resuspended in RNA Storage Solution (Thermo Fisher #AM7001), which was also used for all RNA dilutions. HBRR and UHRR stocks were measured with a Quant-iT RNA Assay Kit (Thermo Fisher #Q10213) on a Qubit fluorometer and these concentrations were used to calculate dilutions. All RNA stocks were stored at $-80\,^{\circ}\mathrm{C}$.

Tissue preparation

Samples were collected with the approval of a HIPAA-compliant Stanford University Medical Center institutional review board. The FFPE tissue blocks were archived with the Stanford University Hospital Department of Pathology.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Slide preparation

Consecutive sections of the FFPE block were taken on a microtome at 7 µm thickness and mounted on glass slides with polyethylene naphthalate membranes (Thermo Fisher #LCM0522). Slides were stored overnight in a nitrogen chamber. The next day, slides were immersed 20 s each in xylenes (3 times), 100% ethanol (3 times), 95% ethanol (2 times), 70% ethanol (2 times), water, hematoxylin (Dako #S3309), water, bluing reagent (Thermo Fisher #7301), water, 70% ethanol (2 times), 95% ethanol (2 times), 100% ethanol (3 times), xylenes (3 times).

Laser-capture microdissection

Slides were dissected immediately after staining. Histology was categorized by a board-certified pathologist (RBW) according to initial sections stained with hematoxylin and eosin. Cells were dissected on an Arcturus XT LCM System using both the ultraviolet (UV) laser to cut out each sample and the infrared laser to adhere it to a CapSure HS LCM Cap (Thermo Fisher #LCM0215). For bulk samples, roughly 500 cells were captured by area, according to density estimates by cell counting on small areas. For single cells, a cell was dissected from the same area as the corresponding bulk sample, then any additional cells adhering to the cap were ablated with the UV laser. For the ablation validation experiment, two regions of both types (still roughly 500 cells each) were captured on the same cap, and then one region or the other was ablated with the UV laser, except the no-ablation controls (therefore they had roughly 1000 cells). After LCM, the cap was sealed in an 0.5 mL tube (Thermo Fisher #N8010611) and stored at −80 °C until library preparation, which was performed within 3 days of dissection.

Microscopy image preparation

LCM photographs were captured in JPEG format with the built-in camera and software of the Arcturus XT system. Photographs in Figures 3 and S18 were color-corrected to enhance contrast with the automatic "Stretch HSV" function in GNU Image Manipulation Program 2.8.20. All photographs were recompressed to reduce file size with the "jpeg-recompress" feature in JPEG Archive 2.1.1, using "veryhigh" quality, "accurate" mode, and the mean pixel error algorithm. All photographs are presented at the original resolution.

Library preparation

Sequencing libraries were prepared according to the Smart-3SEQ protocol (Supplementary File 2). Reference RNA libraries were prepared using the standard protocol for non-degraded RNA and the pre-SPRI pooling option, one batch for the ERCC experiment and one for the human reference RNAs, with the numbers of PCR conditions shown in Figures S4 and S9. The LCM experiments used the special protocol for FFPE tissue on an Arcturus LCM HS cap and the pre-SPRI pooling option, a single batch for all LCM samples, using 22 PCR cycles for the bulk and ablated samples with roughly 500 cells, 25 cycles for the single cells, and 20 cycles for the no-ablation controls with roughly 1000 cells. No-template controls were prepared identically to single-cell samples in the same batch using empty LCM caps. Libraries were characterized immediately and stored at −20 °C until sequencing.

Library characterization and sequencing

Libraries were profiled for size distribution on an Agilent 2200 TapeStation with High Sensitivity D1000 reagent kits and quantified by qPCR with a dual-labeled probe [17], and libraries were mixed to equimolarity according to the qPCR measurements. In the reference-RNA experiments, libraries were prepared according to the manufacturer's instructions with a 1% spike-in of the ϕX174 control library (Illumina #FC-110-3002) and sequenced on an Illumina NextSeq 500 instrument with a High Output v2 reagent kit (Illumina #FC-404-2005), reading 86 nt for read 1 and 6 nt for the P7 index read. The LCM experiments used 76 nt for read 1. In the experiment for testing the effect of read lengths, libraries were prepared according to the manufacturer's instructions with a 1% ϕX174 spike-in and sequenced on an Illumina MiSeq instrument with a Reagent Kit v3 (Illumina #MS-102-3001), reading 169 nt for read 1 and 6 nt for the P7 index read.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data preprocessing

In most experiments, base calls from the NextSeq were demultiplexed and converted to FASTQ format with bcl2fastq (Illumina Inc); adapter trimming was enabled and short trimmed sequences were retained for diagnostic purposes. For the MiSeq run, FASTQ generation, demultiplexing, and adapter trimming were performed by the MiSeq Reporter in "Generate FASTQ" mode. Only reads that passed the chastity filter were included in the total read count and further analysis. The first 5 nt of each read were removed from its sequence and appended to the read name as its UMI; the following 3 nt were assumed to derive from the G-overhang of the template-switch oligonucleotide and were discarded. Reads shorter than 8 nt after adapter trimming were assumed to be primer dimers and were excluded from other analysis, but still counted toward the total number of sequenced reads. Software versions and command-line arguments are listed in Table S2.

Read alignment and counting

Reads from the experiment that used only ERCC RNAs were aligned to NIST's empirical ERCC sequences [18] by NovoAlign (Novocraft Technologies Sdn Bhd), using A_{30} as the "adapter" sequence, and alignments with MAPQ < 10 (posterior probability < 0.9) were discarded. Reads from the experiment that mixed human reference RNAs with ERCC spike-ins were aligned by STAR [19] to a combination of the ERCC sequences and the hg38 reference genome sequence, masked for SNPs according to dbSNP version 150 (obtained from the UCSC Genome Browser) [20], and the GENCODE version 27 gene annotations [21] were provided as the reference transcriptome. Reads from the experiment with human tissue were aligned in the same way except their reference index contained no ERCC sequences. Alignments per human transcript were counted by featureCounts from the Subread suite [22], including only non-duplicate alignments in the correct (sense) orientation.

Quantification of transcript abundance

Transcripts per million were computed differently between Smart-3SEQ and whole-transcript RNA-seq because the latter must be normalized by transcript length:

$$\mathrm{TPM}_{\text{RNA-seq}}(i) = \frac{\frac{c_i}{l_i}}{\sum_i \frac{c_i}{l_i}} \tag{1}$$

$$\mathrm{TPM}_{\text{3SEQ}}(i) = \frac{c_i}{\sum_i c_i} \tag{2}$$

where c_i is the number of reads aligned to transcript i and l_i is the length of the annotated transcript.

Because transcript abundances and Smart-3SEQ read counts by transcript were log-distributed, but some counts were zero, we defined the linear correlation between read counts and other abundance measurements as

$$r' = \mathrm{corr}(\log_{10} A, \log_{10}(C+1)) \tag{3}$$

where A and C are the random variables of which the abundance measure a and read count c, respectively, are particular realizations, and $\mathrm{corr}(X, Y)$ is Pearson's product-moment correlation.

Analysis of gene expression

Differential gene expression was analyzed with DESeq2 1.6.1 [23]. All default options were used except dispersions were estimated with local fitting. The regularized log transformation was used to normalize read counts. PCA was performed on the normalized data with the prcomp function and t-SNE was performed with the Rtsne package on the Euclidean distance matrix of the normalized data.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Code availability

Scripts used to perform the analyses in this study are collected in Supplementary File 5.

Data availability

All sequencing data generated in this study are deposited in the NCBI Sequence Read Archive under accession number SRP119788.

Author contributions

JWF conceived the Smart-3SEQ method. JWF, PJ, CZ and SXZ developed the protocol. JWF and CZ performed the experiments shown here. JWF developed the data-processing pipeline. JWF and PL analyzed the data. JWF and RBW prepared the manuscript. MJM supervised JWF and PJ in the initial Smart-3SEQ phase of the project. RBW supervised JWF, CZ, SXZ, and PL in the later LCM phase of the project.

Acknowledgments

We are grateful to the laboratory of Carlos D. Bustamante and especially to Shirley Sutton and Alexandra Sockell for use of the TapeStation and NextSeq instruments. We are also grateful to the McGill Group for Suicide Studies and especially to Gary G. Chen for use of the MiSeq instrument and guidance with sequencing. Some of the computations were performed on the Sherlock cluster of the Stanford Research Computing Center and the Guillimin cluster of McGill University's High Performance Computing Centre. JWF was supported by a Foundation MIRI grant from Brain Canada. JWF and PJ were supported by the Ludmer Centre for Neuroinformatics and Mental Health. JWF, CZ, SXZ, PL, and RBW were supported by the National Cancer Institute of the National Institutes of Health under award number R01CA183904.

Competing interests

The authors declare no competing financial interests.

Supplementary files

1. Supplementary text, figures, and tables.
2. Smart-3SEQ protocol.
3. Tables of gene-expression measurements and differential-expression analysis from the FFPE LCM experiment.
   - `bulk comparison`: bulk DCIS vs. bulk macrophage
   - `single comparison`: single DCIS vs. single macrophage, excluding DCIS cells that lack the *ERBB2* amplification
   - `amp comparison`: single DCIS with the amplification vs. single DCIS without the amplification
   - `site comparison`: single DCIS without the amplification (inferred ductal macrophage) vs. single macrophage (stromal)
   - `single vs. bulk`: single cells vs. bulk tissue, controlling for tissue type, excluding DCIS cells without the amplification
   - `DCIS vs. macrophage`: DCIS vs. macrophage, single cells and bulk combined and controlled for, excluding DCIS cells without the amplification
4. Complete set of LCM photos.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Analysis scripts.

6. Spreadsheet for calculating reagent costs.

References

[1] R. Virchow, *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre*. Hirschwald, 1859.

[2] J. Abbasi, "An international human cell atlas consortium takes shape," *JAMA*, vol. 318, no. 8, pp. 685–686, 2017.

[3] A. Spira, M. B. Yurgelun, L. Alexandrov, A. Rao, R. Bejar, K. Polyak, M. Giannakis, A. Shilatifard, O. J. Finn, M. Dhodapkar, N. E. Kay, E. Braggio, E. Vilar, S. A. Mazzilli, T. R. Rebbeck, J. E. Garber, V. E. Velculescu, M. L. Disis, D. C. Wallace, and S. M. Lippman, "Precancer atlas to drive precision prevention trials," *Cancer Research*, vol. 77, no. 7, pp. 1510–1541, 2017.

[4] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, pp. 1202–1214, May 2015.

[5] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, pp. 14049+, Jan. 2017.

[6] Y. Y. Zhu, E. M. Machleder, A. Chenchik, R. Li, and P. D. Siebert, "Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction," *BioTechniques*, vol. 30, pp. 892–897, Apr 2001.

[7] D. Ramskold, S. Luo, Y. C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg, "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells," *Nature Biotechnology*, vol. 30, pp. 777–782, Aug 2012.

[8] S. Picelli, A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg, "Smart-seq2 for sensitive full-length transcriptome profiling in single cells," *Nature Methods*, vol. 10, pp. 1096–1098, Nov 2013.

[9] A. H. Beck, Z. Weng, D. M. Witten, S. Zhu, J. W. Foley, P. Lacroute, C. L. Smith, R. Tibshirani, M. van de Rijn, A. Sidow, and R. B. West, "3′-end sequencing for expression quantification (3SEQ) from archival tumor samples," *PLoS ONE*, vol. 5, p. e8768, Jan 2010.

[10] T. Kivioja, A. Vaharautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, "Counting absolute numbers of molecules using unique molecular identifiers," *Nature Methods*, vol. 9, pp. 72–74, Nov 2011.

[11] A. Oshlack and M. J. Wakefield, "Transcript length bias in RNA-seq data confounds systems biology," *Biology Direct*, vol. 4, p. 14, Apr 2009.

[12] SEQC/MAQC-III Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium," *Nature Biotechnology*, vol. 32, pp. 903–914, Sep 2014.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[13] S. A. Munro, S. P. Lund, P. S. Pine, H. Binder, D. A. Clevert, A. Conesa, J. Dopazo, M. Fasold, S. Hochreiter, H. Hong, N. Jafari, D. P. Kreil, P. P. Łabaj, S. Li, Y. Liao, S. M. Lin, J. Meehan, C. E. Mason, J. Santoyo-Lopez, R. A. Setterquist, L. Shi, W. Shi, G. K. Smyth, N. Stralis-Pavese, Z. Su, W. Tong, C. Wang, J. Wang, J. Xu, Z. Ye, Y. Yang, Y. Yu, and M. Salit, "Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures," *Nature Communications*, vol. 5, p. 5125, Sep 2014.

[14] MAQC Consortium, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, pp. 1151–1161, Sept. 2006.

[15] G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold, "From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing," *Genome Research*, vol. 24, no. 3, pp. 496–510, 2014.

[16] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[17] M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner, "A large genome center's improvements to the Illumina sequencing system," *Nature Methods*, vol. 5, pp. 1005–1010, Dec 2008.

[18] H. Lee, P. S. Pine, J. McDaniel, M. Salit, and B. Oliver, "External RNA Controls Consortium Beta Update," *Journal of Genomics*, vol. 4, pp. 19–22, 2016.

[19] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, Jan 2013.

[20] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation.," *Nucleic Acids Research*, vol. 29, pp. 308–311, Jan. 2001.

[21] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Research*, vol. 22, pp. 1760–1774, Sep 2012.

[22] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, pp. 923–930, Apr 2014.

[23] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, 2014.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

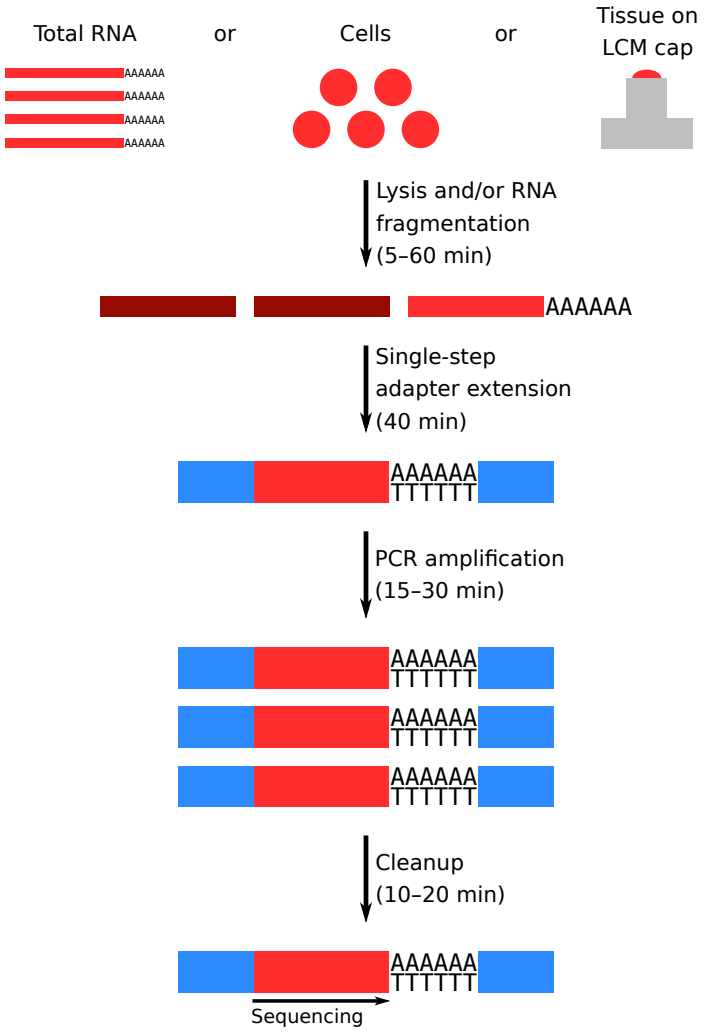

Figure 1: Schematic of the Smart-3SEQ workflow.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

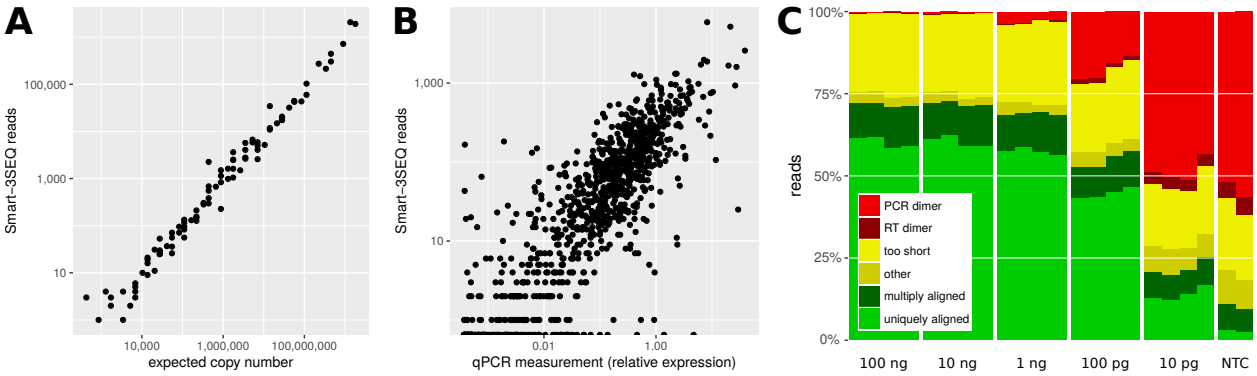

Figure 2: Technical validation of Smart-3SEQ with reference RNAs. A: Standard curve from ERCC transcripts, $r' = 0.990$. Each point represents one transcript sequence; "expected copy number" is the estimated number of copies of that transcript in the RNA sample (1 µL at 1/10 dilution, 10 fmol total), and "Smart-3SEQ reads" is the number of post-filter reads aligned uniquely to any part of that transcript's sequence in the expected orientation (6.4M). Data shown are from the first replicate of the condition with 10 fmol ERCC mix 2 and 7 PCR cycles, which had the highest read count for that condition; data from all samples are shown in Figure S6. B: Comparison with TaqMan qPCR quantification of human transcripts, $r' = 0.854$. qPCR measurements are normalized to the expression of RNA polymerase II. Data shown are from the first replicate of 100 ng UHRR, which had the highest read count for that amount of input RNA (2.7M sense-aligned to annotated genes); data from all samples are shown in Figure S10. C: Alignability of Smart-3SEQ reads from human reference RNA dilutions and no-template controls.

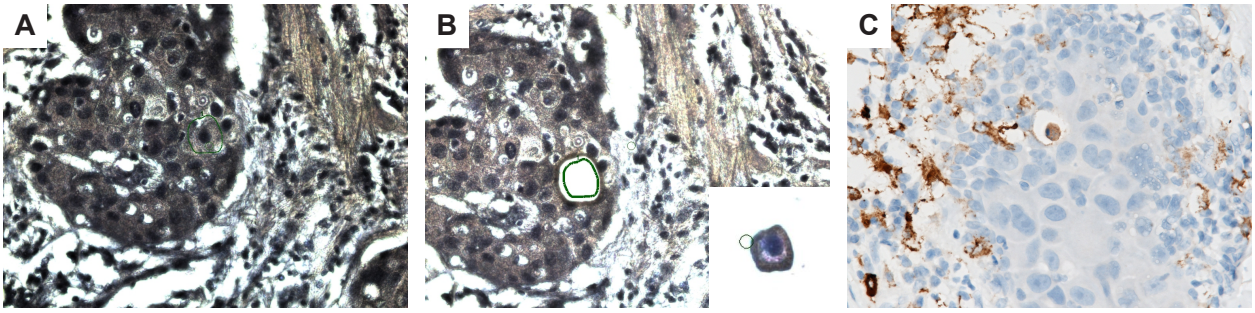

Figure 3: Laser-capture microdissection of ductal carcinoma *in situ*. A: Single cell within a duct involved by DCIS, targeted for dissection (green outline). B: Duct post-dissection and (inset) the captured cell on the LCM cap. C: Immunohistochemistry for macrophage marker CD163.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

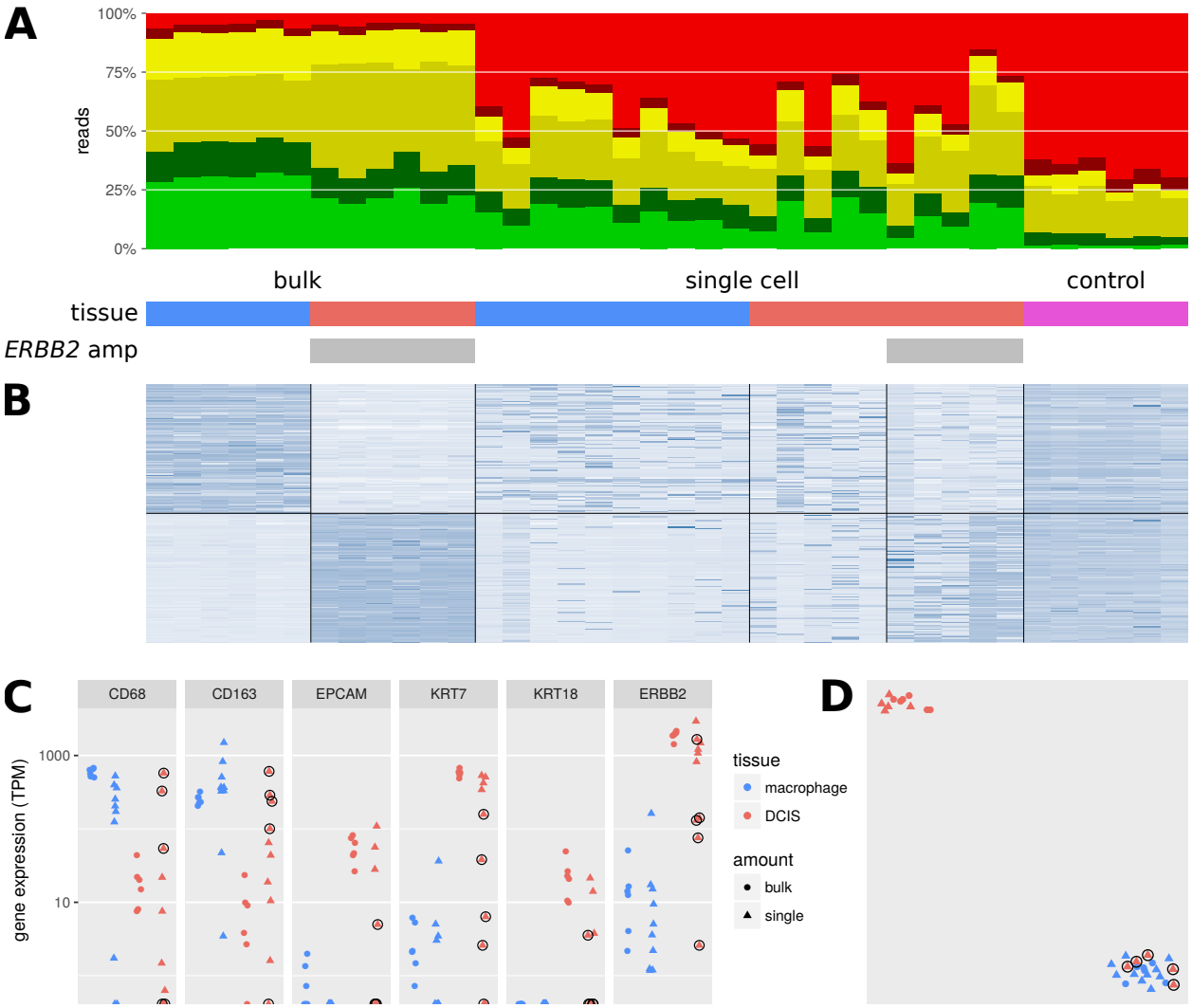

Figure 4: Gene-expression profiling on bulk and single-cell samples from FFPE tissue dissected by LCM. A: Alignability of Smart-3SEQ reads from FFPE bulk tissues, single cells, and negative controls. B: Expression (regularized log read count, normalized by row) of the 100 genes with the greatest enrichment in bulk macrophage relative to bulk DCIS and the 100 genes with the opposite enrichment, all significant at $p_{\mathrm{adj}} < 0.05$. C: Expression (transcripts per million) of known marker genes for macrophage (*CD68*, *CD163*) and DCIS (*EPCAM*, *KRT7*, *KRT18*, *ERBB2* (*HER2*)). Single DCIS cells lacking the *ERBB2* amplification are circled. D: t-SNE analysis of all genes; same plotting scheme as C.

bioRxiv preprint doi: https://doi.org/10.1101/207340; this version posted January 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

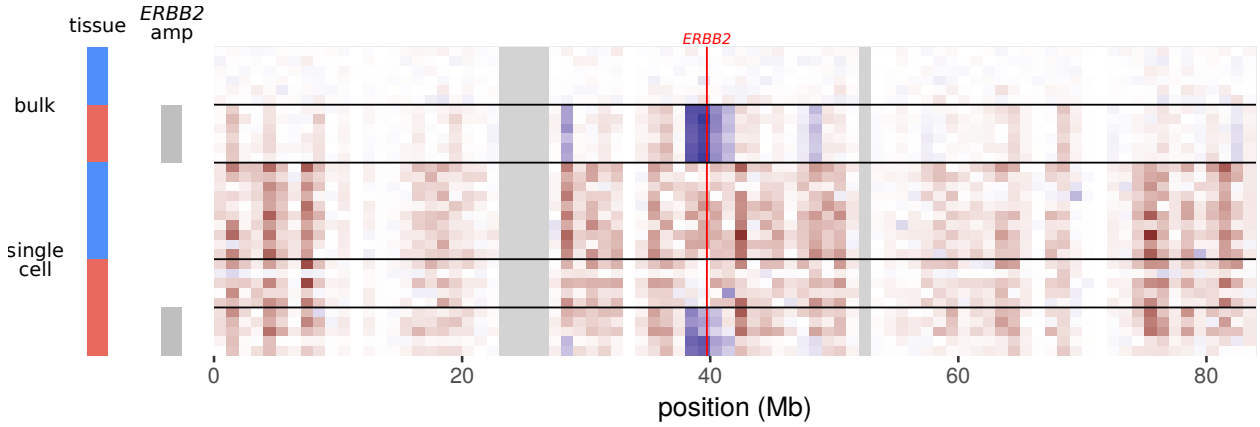

Figure 5: Inferred DNA copy number for chromosome 17 in bulk and single cell samples. Genes are aggregated in blocks of 1 Mb by transcription termination site. Heatmap cells show expression normalized to the mean of the bulk macrophages. Red: lower expression than bulk macrophages; blue: higher expression than bulk macrophages; gray: no data. The bright red line shows the position of the *ERBB2* locus. Legend colors for cell types (left) match Figure 4.

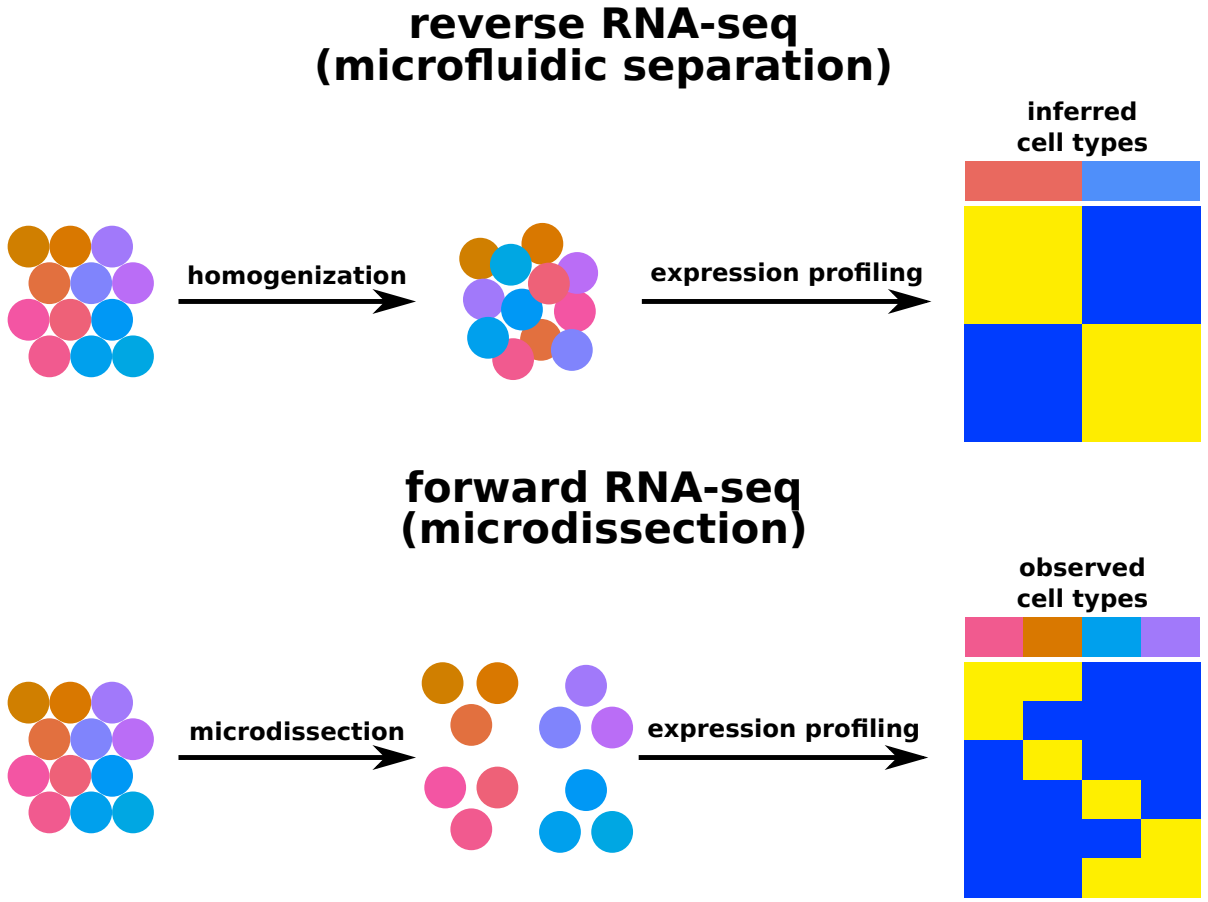

Figure 6: Conceptual diagram of different RNA-seq approaches. Reverse RNA-seq: Cells are disaggregated from tissue, destroying information about histological context and organization. After single-cell RNA-seq, expression profiles are used to retroactively infer categories of cells. Forward RNA-seq: Cells are dissected *in situ* according to their histology, so these *a priori* classes can inform differential gene-expression analysis.