### Title

Gel-seq: whole-genome and transcriptome sequencing by simultaneous low-input DNA and RNA library preparation using semi-permeable hydrogel barriers




### Journal

Lab on a Chip, 17(15)

### ISSN

1473-0197



### Authors

Hoople, Gordon D
Richards, Andrew
Wu, Yan
et al.



### Publication Date

2017-07-25

### DOI

10.1039/c7lc00430c

Peer reviewed

Lab on a Chip

PAPER

View Article Online
View Journal | View Issue





Cite this: *Lab Chip*, 2017, **17**, 2619

# Gel-seq: whole-genome and transcriptome sequencing by simultaneous low-input DNA and RNA library preparation using semi-permeable hydrogel barriers†

Gordon D. Hoople, ‡*[ab] Andrew Richards, ‡[c] Yan Wu,[c] Kota Kaneko,[d] Xiaolin Luo,[d] Gen-Sheng Feng,[d] Kun Zhang[c] and Albert P. Pisano[b]

Received 19th April 2017,
Accepted 21st June 2017

DOI: 10.1039/c7lc00430c

rsc.li/loc

The advent of next generation sequencing has fundamentally changed genomics research. Unfortunately, standard protocols for sequencing the genome and the transcriptome are incompatible. This forces researchers to choose between examining either the DNA or the RNA for a particular sample. Here we describe a new device and method, collectively dubbed Gel-seq, that enables researchers to simultaneously sequence both DNA and RNA from the same sample. This technology makes it possible to directly examine the ways that changes in the genome impact the transcriptome in as few as 100 cells. The heart of the Gel-seq protocol is the physical separation of DNA from RNA. This separation is achieved electrophoretically using a newly designed device that contains several different polyacrylamide membranes. Here we report on the development and validation of this device. We present both the manufacturing protocol for the device and the biological protocol for preparing genetic libraries. Using cell lines with uniform expression (PC3 and Hela), we show that the libraries generated with Gel-seq are similar to those developed using standard methods for either RNA or DNA. Furthermore, we demonstrate the power of Gel-seq by generating a matched genome and transcriptome library from a sample of 100 cells collected from a mouse liver tumor.

## Introduction

Genomicists strive to understand how the information encoded by our DNA is turned into life. Understanding the way variations in DNA impact RNA expression is critical to decoding cell behavior. Recent advances in sequencing technology have made it possible to examine either the genome or the transcriptome of increasingly small samples.[1–3] Both approaches are extremely powerful, however the protocols are generally incompatible. This presents a challenge for simultaneously investigating both DNA and RNA.

When samples are sufficiently large, they can be split in half and processed for either for DNA or RNA sequencing. Unfortunately, large samples tend to average out interesting variations between cells.[4] Researchers are increasingly interested in investigating the variations present in small populations of cells.[4] To illustrate the importance of studying small cell populations, consider that tumors are often composed of heterogeneous cell populations.[5] Evidence suggests this heterogeneity may be responsible for treatment failure.[6] In order to understand tumor genomics, it would be useful to profile small groups of cells from different locations. When collecting just a few hundred cells from such a tumor, splitting a sample in half could result in two distinctly different cell populations, making it difficult to establish a causal link between genomic and transcriptomic variations. Gel-seq is our solution to this problem. Rather than splitting the sample, researchers can instead use Gel-seq to generate DNA and RNA libraries from the same starting cells. This method allows for the direct comparison of DNA and RNA data from low input samples.

The ability to sequence either DNA or RNA from low input samples has only been achieved in the last five years.[1–3] Consequently there has been very little work regarding how to sequence both DNA and RNA from the same sample. To date we are only aware of two other publications on this topic, both from 2015, and both having taken a very different approaches from our method. Dey *et al.* have developed a

[a] *Shiley-Marcos School of Engineering, University of San Diego, 5998 Alcala Park, San Diego, CA 92110, USA. E-mail: ghoople@sandiego.edu; Tel: +619 260 2753*
[b] *Department of Mechanical and Aerospace Engineering, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA*
[c] *Department of Bioengineering, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA*
[d] *Department of Pathology, School of Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7lc00430c

‡ These authors contributed equally to this work.

protocol, DR-Seq, for simultaneously amplifying and sequencing DNA and RNA from the same single cell.[7] DR-Seq takes a computational approach to distinguish between genomic DNA and the cDNA derived from RNA. To calculate DNA coverage in DR-Seq, reads where only exons are present are computationally suppressed, as those could have originated from either DNA or RNA. The genomic profile is instead determined using data based only on sequences containing introns. A drawback of this approach is that it requires *a priori* knowledge (exons *vs.* introns) of a reference genome assembly. Furthermore, intron splicing is not always conserved in disease states such as cancer. Macaulay *et al.* have developed G&T-seq, a method for separating, amplifying, and sequencing DNA and RNA from the same single cell.[8] This approach relies on a physical separation of RNA from genomic DNA by using the 3′ polyadenylated tail as a pull-down target. Messenger RNA is captured on a magnetic bead using a biotinylated oligo-dT primer, allowing it to be separated from genomic DNA.

The novel aspect of Gel-seq is the ability to separate DNA and RNA in hundreds of cells based exclusively on size. Our method requires no *a priori* knowledge of the genome and is not limited to polyadenylated transcripts. For applications where a researcher can start with a few hundred cells, or where the transcripts of interest are not polyadenylated, Gel-seq provides an alternative approach to existing methods using cheap and widely-available materials.

Our method takes advantage of the vast size differences between DNA and RNA. At the heart of the Gel-seq protocol is the electrophoretic separation of DNA and RNA/cDNA hybrids based on this size difference. Genomic DNA from humans, for example, is tens of millions of base-pairs (bp) long for the shortest chromosomes and will remain megabase-scale if shearing is minimized. Most messenger RNA, on the other hand, are only a few hundred to a few thousand nucleotides. Understanding this size difference, we developed two membranes that could be used to separate DNA from RNA. The first membrane, a low density polyacrylamide gel, allows RNA molecules to pass through but stops larger genomic DNA. The second membrane, a high density polyacrylamide gel, traps the RNA molecules. Both membranes allow small fragments (<100 bases) of unwanted artifacts, such as primers, to pass through. The membranes also allow small buffer ions to pass through unimpeded, a necessary condition for electrophoresis. While it is well documented in the literature that ion gradients can form in microfluidic systems in response to applied electric fields,[9] we see no evidence that such gradients are negatively impacting our separation. We theorize that the large size of our buffer reservoir, the high potential difference across the membrane, and the short timespan over which we run the device the mitigates the effects of any ion buildup.

Our basic approach to separating DNA and RNA is shown in Fig. 1. Fig. 1A shows DNA and RNA free floating in solution near a synthetic membrane. When an electric field is applied, as shown Fig. 1B, DNA and RNA experience an electrophoretic force that induces migration through the membrane. By tuning the membrane properties, we created a semi-permeable membrane that separates DNA from RNA. The genomic DNA molecules are pushed against the membrane, but become trapped at the edge due to their large size. Smaller RNA molecules, on the other hand, are able to weave their way through the low density membrane much like a snake through grass, a process known as reptation.[10] These RNA molecules are then stopped by a second, high density membrane. Once they have been physically separated, the DNA and RNA can be recovered and processed into genomic and transcriptomic sequencing libraries.

Though we conceived of the method independently, our approach harkens back to the disc gel electrophoresis

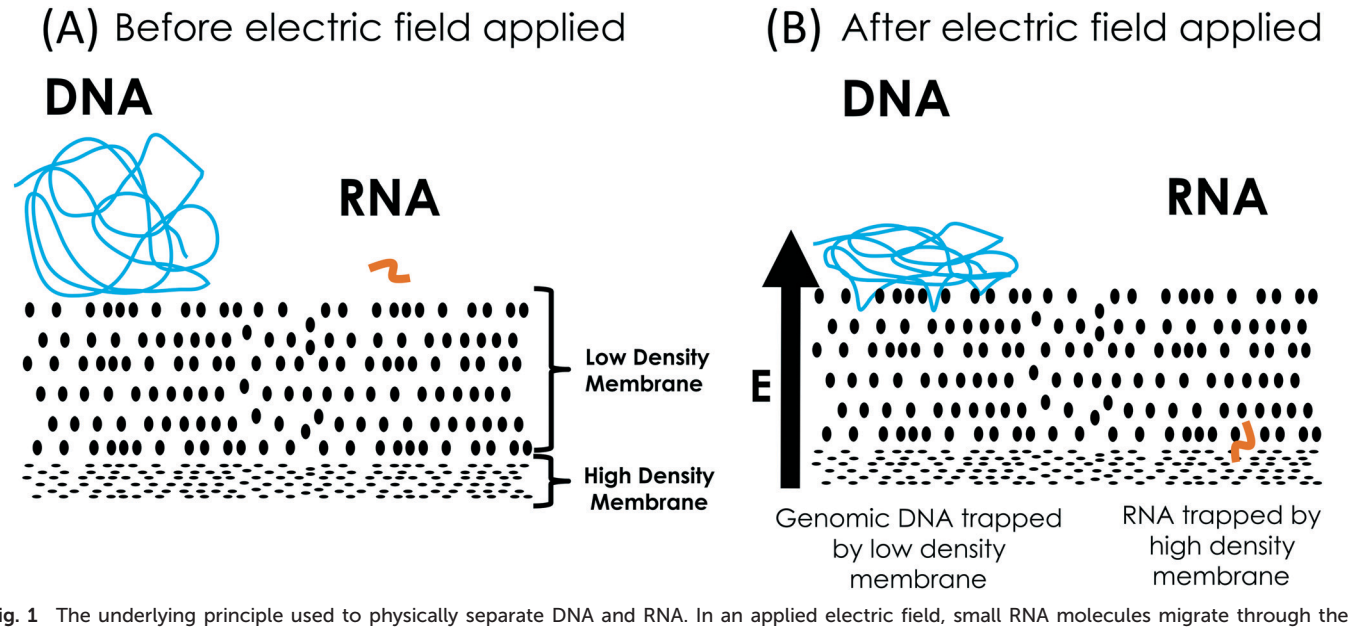


**Fig. 1** The underlying principle used to physically separate DNA and RNA. In an applied electric field, small RNA molecules migrate through the low density membrane but large DNA molecules are trapped at the surface.

 

invented by Orstein and Davis in the 1960s.[11,12] In disc gel electrophoresis, hydrogels with discontinuous pore sizes are used to increase the separation resolution for proteins. Our method differs from traditional disc gel electrophoresis in that our high density membrane is designed to stop a species of interest rather than improve the resolution between bands.

## Experimental

### Gel-seq overview

An overview of the Gel-seq protocol is shown in Fig. 2A. We used a protocol adapted with minor modifications from Nextera XT to prepare DNA libraries after separation. To prepare RNA libraries, we first converted RNA to cDNA using a modification of the Smart-Seq protocol developed by Ramskold followed by a modified version of Nextera XT.[3,13] While we can separate DNA and RNA, we have found that converting the RNA to cDNA before separation helps mitigate problems associated with RNAse contamination. We begin the protocol with between 100 and 1000 intact cells, apply a lysis buffer, and perform reverse transcription with template switching. This generates cDNA/RNA hybrids that are more stable than RNA alone. This protocol does not have a measurable impact on the quality of the genomic DNA (gDNA). The resulting cDNA/RNA hybrids are orders of magnitude smaller than the genomic DNA, enabling size-based separation as shown in Fig. 1 using a custom fabricated gel system.

The Gel-seq device shown in Fig. 2B consists of three regions of polyacrylamide gel. The top layer, highlighted with false color in pink, consists of a low density membrane of 4% total (T) acrylamide and 3% cross-linker (C) bis-acrylamide. A standard gel electrophoresis comb is used to define loading wells. This layer stops genomic DNA but allows transcripts less than 10 000 nucleotides to pass through. The second layer, highlighted in purple, is a high density membrane of 30% T acrylamide cross linked with 5% C bis-acrylamide. This layer stops RNA/cDNA but permits the passage of ions necessary for electrophoresis. The bottom layer, shown in green, fills the remainder of the gel cassette but is not used in the separation. The filler gel is also a 4% T acrylamide cross-linked with 3% C bis-acrylamide. Using a low density filler gel, rather than filling the rest of the cassette with high density gel, ensures that there is a sufficiently large potential drop across the separation region to induce RNA/cDNA migration. The resulting gel cassette is compatible with standard buffer chambers and power supplies commonly found in life science laboratories. The fabrication protocol, described in detail in the next section, is straightforward and utilizes commonly available equipment and materials.

After placing the device into a buffer chamber, we then pipette the DNA and reverse transcription products into the wells. We induce electrophoresis by applying 210 V across the cassette for 30 minutes. Once the genomic DNA and RNA/cDNA have been separated, we cut out the gel sections to recover the nucleic acids using a modified crush and soak procedure. We prepare a DNA sequencing library directly from the genomic DNA using the Nextera XT protocol. For RNA, we first PCR amplify the cDNA fraction and then prepare a sequencing library by Nextera XT.

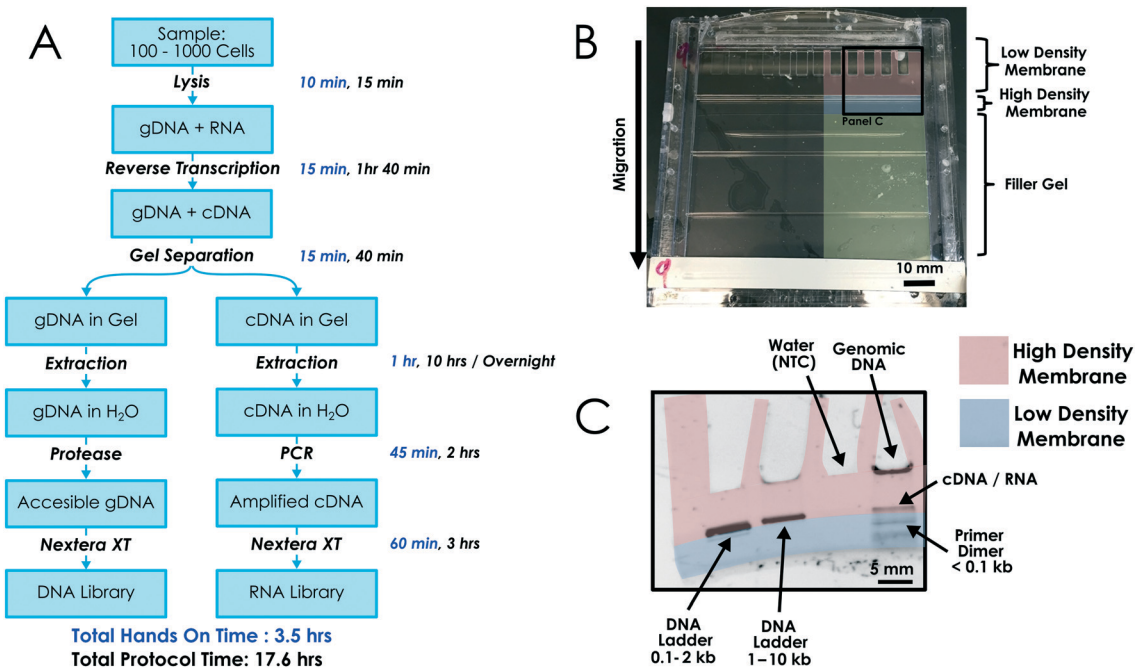


Fig. 2 An overview of the Gel-seq protocol (A) and device (B). False color has been added to half of the device to clearly demarcate the different regions of polyacrylamide gel. The third panel (C) is a fluorescent image showing the separation of genomic DNA and cDNA/RNA hybrids. Black bands indicate the presence of nucleic acids. Lanes loaded with only DNA ladder show a single band that has been trapped by the high density membrane. Lanes loaded with genomic DNA and RNA/cDNA show two bands, suggesting that genomic DNA has been separated from RNA/cDNA.

### Device fabrication

Many companies sell standard gel electrophoresis systems that come with a power supply, electrophoresis chamber, and empty cassettes. These systems dramatically simplify the process of conducting experiments with gel electrophoresis. End users simply fill the cassette with the desired density polyacrylamide based on their needs. Once the gel has polymerized, the cassette is placed in the electrophoresis buffer chamber, sample is added, and the chamber is connected to a power supply to apply an electric field. In this paper we based our fabrication protocol around the XCell SureLock® Mini-Cell system (Lonza); however any similar system could be used.

Device fabrication builds on skills that will be familiar to researchers who use standard polyacrylamide gel cassettes. Before fabrication, monomer solutions are made for each layer by combining acrylamide/bis-acrylamide solution, 10× Tris-borate-EDTA (TBE), water, and sucrose solution (50% w/v) as shown in Table 1. The addition of sucrose to the polyacrylamide precursor solution is key to the formation of a smooth interface layers between the different densities, but has minimal impact on electrophoresis. Stock acrylamide/bis-acrylamide solutions used in these recipes can be made by combining acrylamide (monomer) and bis-acrylamide (crosslinker) powders using the following formulas:

$$\%\,\mathrm{T} = \frac{\text{monomer mass (g)} + \text{crosslinker mass (g)}}{\text{solvent volume (mL)}} \quad (1)$$

$$\%\,\mathrm{C} = \frac{\text{crosslinker mass (g)}}{\text{monomer mass (g)} + \text{crosslinker mass (g)}} \quad (2)$$

The gel precursor solutions are mixed in a tube and vortexed to ensure thorough mixing, and then immersed in a sonicator under house vacuum. This helps to remove dissolved gases that could inhibit the polymerization process. Immediately before transferring the precursor solution to the cassette, a polymerization initiator containing ammonium persulfate (APS) and catalyst (TEMED) are added and the mix is briefly vortex again. Note that the high density gel does not contain any TBE. While it could be included, we find it easier to mix the precursor solution when it is not included as we are approaching the solubility limit of acrylamide and bisacrylamide. We have noticed no negative impact on device performance from the omission of TBE in this region.

An overview of the protocol is shown in Fig. 3. Layers are fabricated from bottom to the top. We first add 6 mL of filler gel precursor to the cassette. The remainder of the cassette is filled with de-ionized, degassed water. The filler gel is allowed to polymerize for at least one hour or up to overnight. The water overlay ensures the formation of a smooth interface. After polymerization, we remove the water overlay by simply inverting the cassette and shaking. Compressed air can be used to assist in the removal of any trapped water droplets. We then add 350 μL of the high density precursor to the cassette. Due to the small volume of high density gel, it is important to ensure the precursor is evenly distributed by tilting the cassette back and forth to allow the liquid to uniformly spread out over the filler gel. Once the high density precursor has been uniformly distributed, we again add a water overlay. In order to obtain the best interface, it is important to add the water slowly to the center of the cassette in order to minimize mixing with the high density precursor. We allow the high density gel to polymerize for at least 10 minutes before the water overlay is removed. Finally, we add the low density precursor to fill the remainder of the cassette, approximately 1.65 mL. In order to define the loading wells, we insert a standard gel comb into the cassette. Cassettes can be fabricated with different numbers and sizes of wells by using different combs. In this work, we fabricated gels with either 10 or 12 well combs. We allow the low density gel to polymerize overnight before using the cassettes. Cassettes can be stored immersed in TBE buffer for several weeks.

### Gel-seq protocol

In addition to device development, there was a need to adapt existing biochemical protocols to be compatible with physical separation of gDNA and RNA and to prepare libraries from both. Recognizing the susceptibility of RNA to degradation, we reverse transcribe RNA to cDNA before separating it from gDNA. Once we separate gDNA and RNA/cDNA, we then prepare a sequencing library from the gDNA using Nextera XT. In parallel, we amplify the cDNA sample by PCR and prepare a sequencing library, also using Nextera XT. In order to

**Table 1** Recipes for mixing polyacrylamide gel precursors

| Filler gel precursor (4% T, 3% C) | | High density gel precursor (30% T, 5% C) | | Low density gel precursor (4% T, 3% C) | |
|---|---|---|---|---|---|
| 40% T, 3.3% C acrylamide Bisacrylamide solution | 0.8 mL | 50% T, 5% C acrylamide Bisacrylamide solution | 1.2 mL | 40% T, 3.3% C acrylamide Bisacrylamide solution | 0.4 mL |
| Deionized water | 5.12 mL | Deionized water | 0.48 mL | Deionized water | 3.2 mL |
| Sucrose (50% w/v) | 1.28 mL | Sucrose (50% w/v) | 0.32 mL | 10× TBE | 0.4 mL |
| 10× TBE | 0.8 mL | APS (10%) | 25 μL | APS | 26 μL |
| APS (10%) | 52 μL | TEMED | 0.5 μL | TEMED | 1.5 μL |
| TEMED | 3 μL | | | | |
| Total volume | 8.1 mL | Total volume | 2 mL | Total volume | 4.0 mL |

 

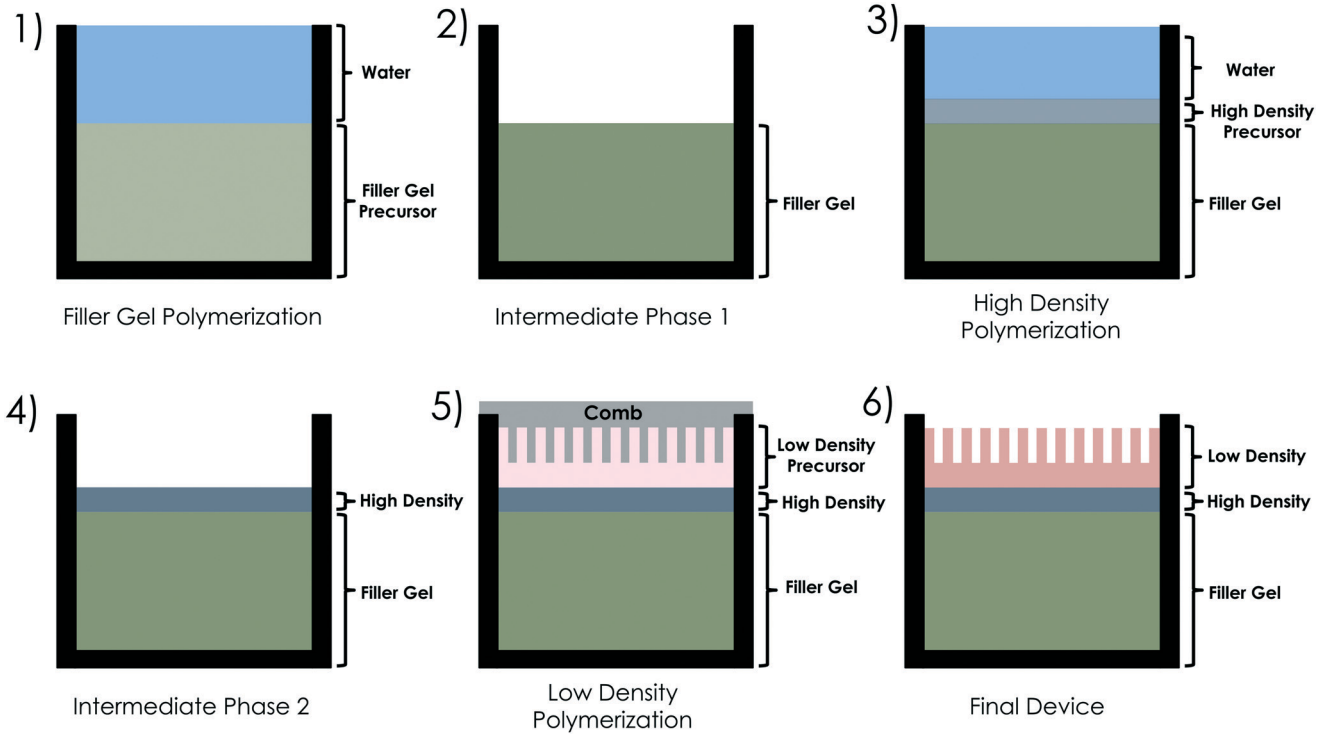


**Fig. 3** The fabrication protocol for the cassette based devices. Each layer of gel is allowed to polymerize before the next layer of gel is poured on top of it. A water overlay helps to create a smooth interface between layers.

minimize the shearing of genomic DNA, which could cause it to enter the separation gel, we avoided vortexing samples. Instead all samples were mixed by gently pipetting up and down approximately 10 times. While this will shear the chromosomes somewhat, the fragments are still orders of magnitude larger than the RNA/cDNA hybrids.

We begin the protocol by preparing cells in PBS at a concentration of 100 to 1000 cells per μL. Using the reagents provided in the Smart-Seq v4 kit (Clontech Laboratories), we mix 19 μL of lysis buffer and 1 μL of RNase inhibitor to prepare a 10× stock solution of reaction buffer. We then combine 1 μL of the cell suspension, 0.5 μL of 10× reaction buffer and 2.75 μL of nuclease-free water and mix by pipetting up and down 5 times. We then add 1 μL of 3′ SMART-Seq CDS Primer II and 1 μL of 20 μM random hexamer with SMART-Seq adapter (Integrated DNA Technologies (IDT); 5′ AAGCAGTGGTATCAACGCAGAGTACNNNNNN 3′). Each sample is incubated at 72 °C in a preheated thermal cycler with heated lid for 3 minutes to lyse the cells. Note that the addition of random hexamer seemed to have minimal impact and the mapping rates to rRNA remain below 1% (see Fig. S9†).

After lysis, we add a master mix containing 2 μL of 5× Ultra Low First-Strand Buffer, 0.5 μL of SMART-Seq v4 Oligos, 0.25 μL RNase Inhibitor, and 1 μL SMARTScribe Reverse Transcriptase. We mix the sample by pipetting up and down 5 times and then immediately place it in a preheated thermal cycler at 42 °C with a heated lid for 90 min, followed by a heat inactivation step at 70 °C for 10 min.

Following the completion of reverse transcription, we mix the samples with 2 μL of 6× DNA Gel Loading Dye (ThermoFisher). We load the entire reaction volume into the Gel-seq device (one sample per well) and apply an electric field of 210 V across the device for 30 minutes to separate RNA from DNA. After separation, we stain the gel in 30 mL of 0.5× TBE with 3 μl SYBR Gold (ThermoFisher) for 5 minutes. We image the gel using a 30 second exposure on a Bio-Rad Gel Doc. We then cut out the regions containing gDNA and cDNA/RNA using a scalpel. Visualizing the cDNA from the 100 cell input samples sometimes presented a challenge due to the small amount of nucleic acids present. Fortunately, the ability to visualize the location of the gDNA or cDNA is not a requirement for recovering it from the gel. We designed the device so the gDNA stops at the start of the well and the cDNA stops at the start of the high density gel. As these locations are both visible to the naked eye, the gel can be cut without the use of a UV backlight. In practice we found using the UV backlight convenient as most samples could be visualized, but this is not a strict requirement.

Once cut from the gel, each gel section is placed into a separate tube and ground up using the end of a pipette. We add 40 μL of nuclease-free water to the gel containing gDNA and 80 μL of nuclease-free water to the gel containing cDNA/RNA. We then tape the tubes containing the gel and water to a vortex mixer inside 37 °C incubator and shake them for 8 to 12 hours. This allows the nucleic acids to diffuse out from the gel into the water.

After incubating the samples, we pipette the samples into an 8 μm mesh filter plate (Corning HTS Transwell 96-well permeable support) and spin the plate at 2600 RCF for 5 minutes to strain out the gel fragments. We then pipette the gel-free water into a new 200 μL tube.

For the gDNA sample, we add 1 μL of protease (Qiagen, diluted to 0.9 AU $mL^{-1}$) and incubate at 50 °C for 15 min followed by 70 °C for 15 min. This step is critical for depleting nucleosomes, making the DNA accessible for Nextera XT library preparation. Next we use an 18 gauge needle to create holes in the caps of all samples tube before spinning them in a vacufuge to reduce sample volume. The cDNA/RNA samples are reduced to 10 μL and the gDNA samples reduced to 5 μL. This step takes 30–60 minutes, depending on the number of samples in the vacufuge. If samples were found to be below the target volume, 1–2 μL of clean nuclease free water was added to bring them to the correct target volume.

We generate libraries from the gDNA samples by following the standard Nextera XT protocol.[13] To conserve reagents, we have found that using half volume reactions does not significantly impact our library quality. The protocol is otherwise identical from this point on.

To generate libraries from the cDNA/RNA samples, we first amplify the sample using PCR. We combine a 10 μL sample with 12.5 μL 2× KAPA SYBR Fast qPCR MasterMix (KAPA Biosystems), 0.5 μL PCR Primer II A (12 μM, from the Smart-Seq kit), and 2 μL nuclease-free water. We perform qPCR in a Bio-Rad thermocycler using the following protocol: hot-start at 95 °C for 3 min, followed by 20–30 cycles of 98 °C for 10 seconds, 65 °C for 30 s, and 72 °C for 3 min. We adjust the number of cycles depending on the amount of starting sample and the shape of the qPCR curves to avoid over-amplification. After amplification, we clean the product using AMPure XP beads following the protocol described in the Smart-Seq Manual.[14] Finally, once the amplified cDNA has been purified, we prepare libraries using the Nextera XT protocol with half volume reactions.

The entire protocol requires 3.5 hours of hands on time and can be completed in 17.6 hours. We recommend starting the protocol in the afternoon so that the crush and soak step can take place overnight.

## Results and discussion

### Validation of DNA and RNA/cDNA separation

To validate our separation approach, we tested the device using four samples: a low mass DNA ladder (0.1–2 kilobases (kb)), a high mass DNA ladder (1–10 kb), water as negative control, and genomic DNA and RNA/cDNA hybrids. Commercially purchased DNA is not generally appropriate as a control for genomic DNA in this case, as it tends to be sheared somewhat during production. The best solution is to use freshly lysed cultured cells in each experiment. After electrophoresis, the device was stained with SYBR Gold and imaged. The resulting fluorescent image is shown in Fig. 2C; false color has been added to distinguish between the different regions of the gel.

The negative control (lane 3) showed no signal, demonstrating that the device is not auto-fluorescent. The first two lanes, loaded with DNA ladder, show the presence of black bands indicating that nucleic acid has been trapped in a specific location. The first lane, which was loaded with the low mass DNA ladder, contains only one band at the interface between the low and high density gels. This band contains fragments ranging from 100–2000 basepairs. Rather than spreading throughout the gel, as is typical in standard gel electrophoresis, the bands stack on top of each other at the interface. This is exactly the desired behavior; small fragments of cDNA and RNA should move through the low density gel and collect at the interface of the high density region. Importantly, this ladder also demonstrates that fragments as small as 100 bp are stopped by the high density membrane.

The second lane, loaded with the high mass DNA ladder, shows similar behavior. The major difference here is that the ladder fragments range in size up to 10 kb. Again, the ladder has stacked at the interface with the high density gel, except for a small fraction at the top of the low density gel. This suggests a size cut-off somewhere between 2 and 10 kb, and perhaps a range of partial migration efficiency above 2 kb, however the great majority of cDNA/RNA species of interest are below this size.[15]

Finally the fourth lane demonstrates the separation of genomic DNA and cDNA/RNA hybrids. A clear dark band present at the top of the start of the low density membrane represents megabase scale genomic DNA, which is unable to enter the gel, while cDNA/RNA hybrids are stacked at the interface of the low and high density regions. Unlike the lanes loaded with ladder only, however, there are several bands present within the high density region of the gel. These fragments, smaller than 100 bp, are off-target products generated from primer oligonucleotides during reverse transcription. By allowing these bands to pass through the high density membrane, we can easily remove them from the experiment by only cutting out the cDNA/RNA hybrids stacked at the membrane interface.

As mentioned previously, there is no commercially purchased genomic DNA control shown in this example, as purified DNA tends to be sheared somewhat during production, and does not accurately represent the full native size of mammalian chromosomes. Furthermore, DNA library preparations in early iterations of Gel-seq failed until the addition of a protease digestion step to the protocol after gel separation, indicating that genomic DNA as loaded into our device is still complexed with nucleosomes. We hypothesize that these protein components of DNA in fact assist in trapping virtually all genomic DNA at the gel surface, aiding recovery by preventing nucleic acids from embedding in the gel during electrophoresis.

In order to validate the conclusions inferred from this image, as well as assess the data quality of sequencing libraries, we cut out sections of the gel with the genomic DNA and cDNA/RNA hybrids and generated sequencing libraries.

 

### Validation of DNA and RNA libraries

We compared Gel-seq against standard methods common in the genomics field using commercially available kits, which we refer to as "tube controls", to prepare a total of 32 sequencing libraries (see Table 2) from two human cell lines (PC3 prostate cancer and HeLa cervical cancer), a mouse cell line (3T3 fibroblasts), and primary derived hepatocytes from mouse liver. PC3 and HeLa were chosen because they are representative of cancers with extensive copy number variations (CNVs). CNVs are either duplications or deletions of large regions of the genome, and can be detected by coverage density with shallow sequencing. CNVs are known to play a role in many cancers and are a widely studied area in cancer genomics.[16–19] In addition, CNVs provide a useful signal for genomic data that lends itself to easy comparison between different approaches for whole genome sequencing library preparation. Primary derived hepatocytes from mouse were chosen in order to validate Gel-seq using cells from a complete organ, which presents additional challenges in terms of sample prep and reaction efficiency due to the presence of extracellular matrix and other inhibitory factors. 3T3 fibroblasts were included as a positive control against liver tissue samples.

Gel-seq and tube control experiments were performed in parallel for all samples to assess the level of agreement between methods. DNA and RNA libraries were prepared for both human and mouse samples. For Gel-seq samples, RNA data was generated from the exact same cells as the DNA data, because DNA and RNA are separated after lysis, while cells used in the tube controls were split 50/50 before lysis. Technical replicates were generated for all samples in order to assess reproducibility of both genomic and transcriptomic profiles from Gel-seq data. Finally, we compared transcriptomic profiles between the different samples types within each species to assess whether Gel-seq can distinguish cell type on the basis of RNA expression.

Fig. 4A shows a comparison of genome-wide CNV profiles generated from PC3 using either Gel-seq or a standard tube reaction. Each point is a mean normalized bin count; bins are defined from reference genome data such that each bin has equal expected count in a healthy diploid cell, *i.e.*, a flat line, representing equal copies for each region of all autosomal (excluding X and Y) chromosomes. In PC3, many CNVs can be seen as spikes above a background copy number of two, and Gel-seq yields a qualitatively similar CNV profile as standard tube reaction. Agreement between the two plots can be assessed quantitatively by linear regression in Fig. 4B. A Pearson correlation of $R = 0.90$ indicates that genomic data gathered from either method is functionally equivalent. Fig. 4C shows maximum predicted library coverage at saturation sequencing depth, indicating that Gel-seq yields high-coverage libraries similar to standard methods. Full coverage extrapolations as a function of depth are shown in Fig. S2.†

Similarly, we compared the transcriptome data from our Gel-seq protocol to the standard in-tube Smart-Seq protocol. Fig. 5A shows the correlation between both Gel-seq technical replicates and between Gel-seq and the standard method. Each point is a count in transcripts per kilobase per million (TPM) for each gene detected at TPM $> 5$ in both dataset. The linear regressions are shown as red lines, and the Pearson correlation coefficient is shown in the upper left corner. Technical replicates from Gel-seq agree with each other ($R \sim 0.8$), but correlate less well with the standard method ($R < 0.7$). This suggests that Gel-seq introduces a bias in gene counts, but that the bias is systematic and meaningful conclusions are still possible between different biological samples. We performed linear regression for all pair-wise combinations of the 8 human RNA datasets: PC3 and HeLa, Gel-seq and tube, two technical replicates each (Fig. S5†). Pearson correlation coefficients for all 28 pairs are condensed in Fig. 5B by comparison type. The green bars represent correlations between pairs of datasets from the same cell type generated from either standard tube reactions, Gel-seq, or Gel-seq *versus* standard. The red bars represent correlations between pairs of datasets from different cell types, which are expected to have lower $R$ values due to biologically different transcriptomic profiles. Although Gel-seq does not agree well with the standard method ($R = 0.66$ for matched samples), it shows similar difference in correlation between matched and mis-matched samples ($R = 0.81$ *versus* $R = 0.70$, respectively) to the standard method ($R = 0.97$ *versus* $R = 0.86$), suggesting that Gel-seq still provides powerful insight into transcriptional variation between different cell types. Indeed, Fig. 5C demonstrates that RNA-seq data generated from Gel-seq (left plot) discriminates well between HeLa and PC3 cell types based on principal component analysis (PCA), as does the standard in-tube method (right plot). Fig. 5C shows that samples separate by method on the first principal component with 96.3% variance explained, confirming that Gel-seq introduces a systematic bias, but that different cell types (HeLa and PC3, red and blue clusters) still separate well on principal component 2.

As reported by the SEQC/MAQC-III Consortium, all RNA-seq methods show some gene specific bias.[20] The key for any new approach is to demonstrate reproducibility so that differences observed between samples can be attributed to a biologically

**Table 2** All 16 samples for both human and mouse. For each sample, both DNA and RNA libraries were generated (32 in total). Tube samples (standard method performed in tube as control) were split before lysis for subsequent DNA and RNA library prep protocols in parallel. Gel-seq samples were lysed first, and DNA and RNA were separated in device before library prep

| Human | | | Mouse | | |
|---|---|---|---|---|---|
| Cell type | Method | Sample name | Cell type | Method | Sample name |
| HeLa | Gel-seq | HeLa-G1 | 3T3 | Gel-seq | 3T3-G1 |
| | | HeLa-G2 | | | 3T3-G2 |
| | Tube | HeLa-T1 | | | 3T3-T1 |
| | | HeLa-T2 | | Tube | 3T3-T2 |
| PC3 | Gel-seq | PC3-G1 | Hepatocytes | Gel-seq | Liver-G1 |
| | | PC3-G2 | | | Liver-G2 |
| | Tube | PC3-T1 | | | Liver-T1 |
| | | PC3-T2 | | Tube | Liver-T2 |

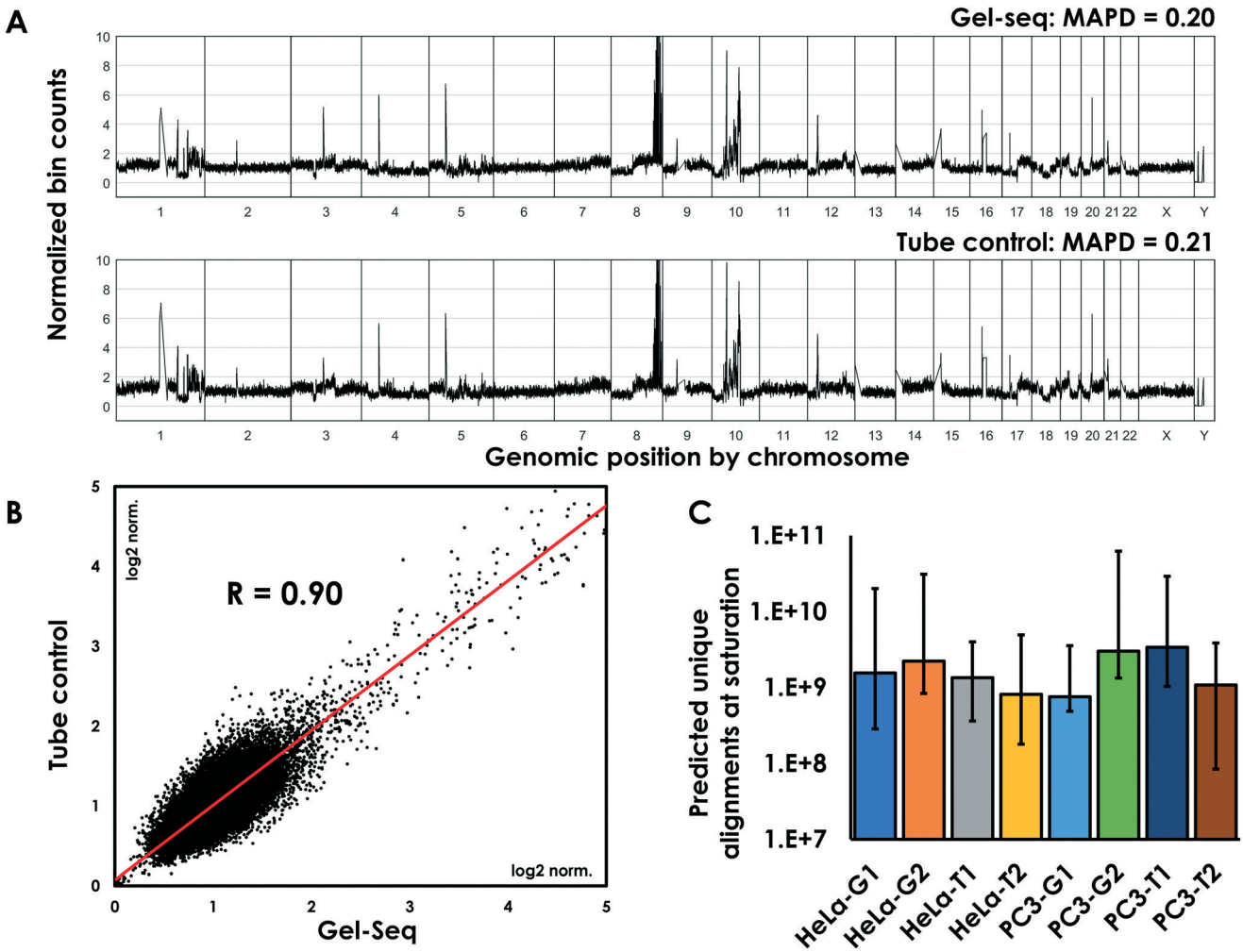


**Fig. 4** Comparing genomic data generated using the Gel-seq protocol to tube control. (A) Mean normalized bin counts for Gel-seq (top) and a tube control (bottom). Random noise is quantified by median absolute pairwise difference (MAPD, upper right). A MAPD of ~0.2 indicates very low noise. (B) Pearson correlation between two representative libraries. Full pairwise correlations are shown in Fig. S4.† (C) Maximum predicted genomic coverage for all human DNA libraries extrapolated to saturation sequencing depth. Error bars are 95% confidence intervals. Suffixes indicate Gel-seq data (–G) or tube controls (–T), numbers indicate technical replicates (1 or 2).

relevant phenomenon. While Gel-seq does not perfectly replicate the results from Smart-Seq, it gives reproducible results and can be used to identify differences between samples.

### Generating paired libraries from tissue

Gel-seq allows researchers to generate both genome and transcriptome data from the same limited sample using commonly available materials. This is useful in scarce samples, such as those collected from living tissue in a biopsy. As mentioned previously, preparing next-gen sequencing libraries from tissue rather than cell lines presents substantial additional challenges. Cell lines divide rapidly, typically doubling in number in 24 to 48 hours, and tend to be highly transcriptionally active, expressing a broader set of genes at high levels compared to an adult tissue under homeostasis. Tissue samples are also subject to the presence of additional extracellular matrix, which can severely inhibit enzymatic reactions. Several iterations of both our device and accompanying biochemical methods were tested before establishing Gel-seq as a robust protocol that works in tissue as well as cultured cell lines. We also lowered the input to 100 cells (0.61 ng DNA). Gel-seq libraries from mouse tissue displayed high quality statistics in terms of unique DNA alignments and genes detected by RNA (Table S1†). Genomic coverage for DNA data and library complexity for RNA data were extrapolated to high sequencing depth (saturation) in Fig. S2† based on bootstrapping simulations, indicating that Gel-seq yields high-quality libraries with coverage similar to standard methods for cells from both cultured lines and complex tissue.

## Conclusion

One of our goals in developing Gel-seq was to create a protocol that could be easily implemented by other researchers.

 

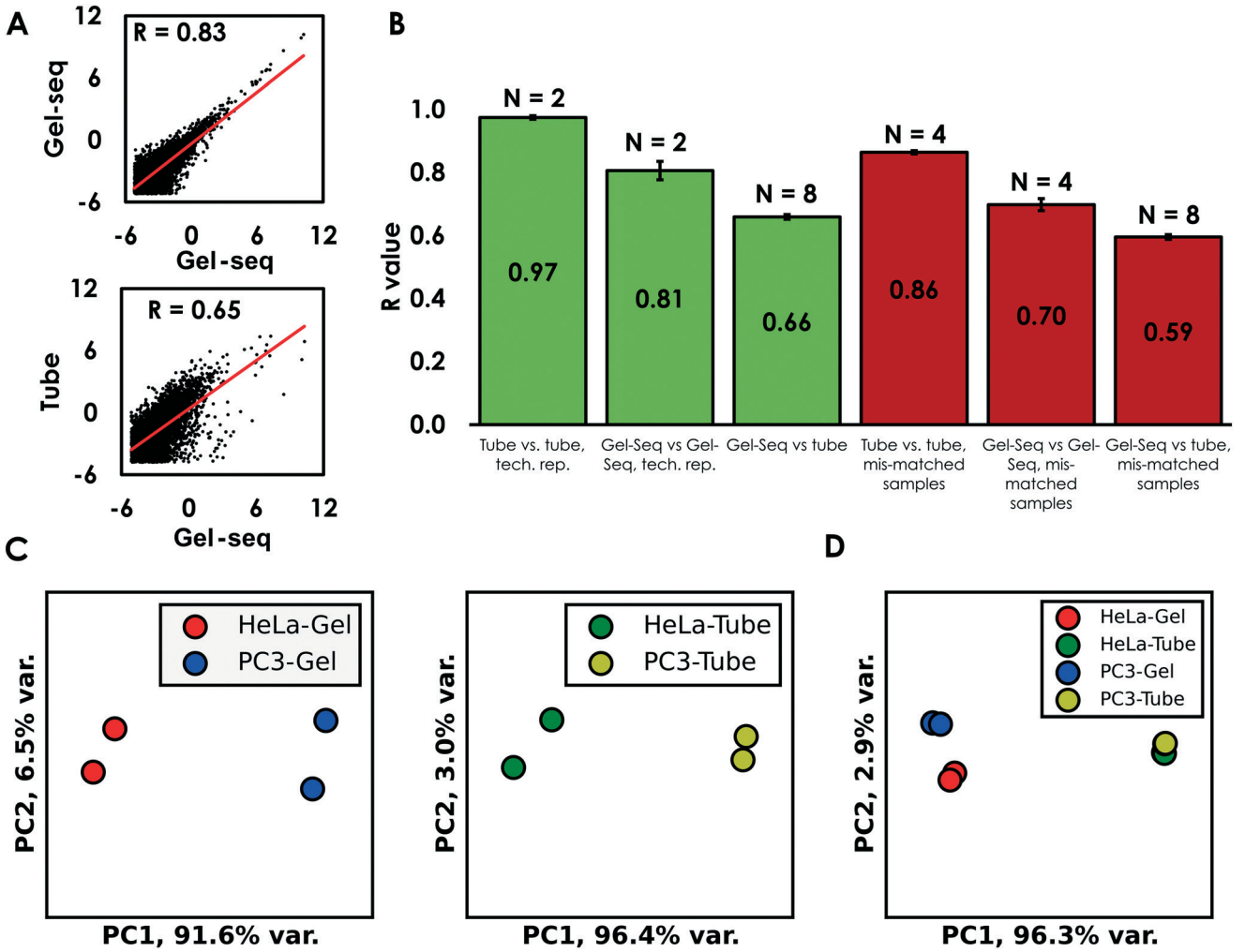


**Fig. 5** Comparing transcriptomic data generated using the Gel-seq protocol to tube controls. (A) Two representative scatter plots of TPM per gene (above threshold of TPM $>$ 5) with an overlaid linear regression and Pearson correlation coefficient. The plot on the top compares two technical replicates using Gel-seq, while the plot on the bottom shows a comparison between a Gel-seq sample and a standard method performed in a tube as control. (B) Pearson coefficients from all 28 pair-wise linear regressions for all 8 HeLa and PC3 RNA datasets generate from with Gel-seq and tube controls. Full table of scatter-plots and regressions are shown in Fig. S5.† (C) PCA for Gel-seq datasets on the left and tube controls on the right. First two principal components are plotted for each, with a total of 98.1% and 99.4% of variance explained for Gel-seq and tube controls, respectively. (D) PCA for all 8 human samples, with total of 99.2% variance explained by the first two principal components.

We therefore decided to fabricate devices within the standard form factor of a polyacrylamide gel cassette. While the technique we used to define our different membranes is novel, most genetics labs already have all of the necessary equipment to fabricate the Gel-seq device. Furthermore the cost of the device is trivial – just $5.25 for a device that can process 12 samples. We believe researchers will find it straightforward to implement Gel-seq in their own labs and hope this will facilitate the rapid adoption of the technology.

As with any library preparation protocol using commercial reagents, the overall cost for generating libraries with Gel-seq remains high. Our reagent cost per sample was $28 for Nextera XT and $50 for Smart-Seq. As cheaper alternatives for library preparation are developed, however, our protocol can be adapted to work with these new techniques. We focused on creating a device that could be adapted for different applications. While in this paper we demonstrated the Gel-seq protocol using Nextera XT and a modified Smart-Seq, the device itself can be used with a wide range of library preparation approaches. For example, during development we successfully tested the device using an older RNA library amplification protocol CellAmp.[21] The core innovation in this technology, separating DNA and RNA based on size using polyacrylamide membranes, is agnostic to the library preparation approach. We anticipate that future biological innovations in library preparation could be integrated into our work flow.

We were successful in generating RNA libraries from cell lines regardless of whether we generated the cDNA either before or, as in earlier iterations using Cell Amp, after separation from the genomic DNA. An unforeseen aspect in the development of the Gel-seq protocol, however, was the

challenge of starting from whole tissue. We found that it was important to adhere strictly to the Smart-seq protocol to generate cDNA from tissue samples as soon as possible. We also experimented with freezing tissue or cell suspensions from tissue in liquid nitrogen, but we found that the best results were obtained when processing fresh samples. We suspect that the extracellular matrix in our tissue samples may have contained RNases, proteases, or other inhibitory factors. Fortunately, Gel-seq is a flexible protocol and proved to be adaptable to liver samples. Although Gel-seq showed generally higher random noise in technical replicates compared with our tube controls, the ability to include genomic data from the same cells in the downstream analysis may justify the trade-off in many applications. Newly developed RNA library preparation methods or optimization of separation and recovery may improve the precision of the RNA data in the future.

An interesting phenomenon observed in the RNA data was that in all 4 samples types (HeLa, PC3, 3T3, and primary hepatocytes) Gel-seq technical replicates agreed with each other, but did not have high correlations with the standard in-tube method. This suggests an underlying systematic difference between methods, which some day might be corrected with either additional optimization of separation and recovery, or accounted for computationally based on known parameters. Our first suspicion was exonic transcript length, with the assumption that very long or very short genes could be lost or trapped in the device. While we did observe a weak relationship between RNA gene counts and gene length in Gel-seq data, with medium length genes showing the highest gene counts, we observed an identical effect in tube control data. Attempting to normalize by a lowess fitted correction function did not improve the correlation between Gel-seq and tube (not shown). This could suggest that additional factors beyond gene length are affecting the data. For many applications the addition of synthetic RNA spike-ins at a range of known concentrations (*e.g.*, ERCC control[22]) could be used to quantify systematic biases in sample data. This is already a common approach in the field for correcting systematic biases introduced by different kits. Future work will focus on addressing these challenges and improving the Gel-seq method. For the time being, however, Gel-seq is already a powerful and sensitive tool for finding differences in expression between samples.

Unfortunately, Gel-seq cannot be used in this embodiment to generate data at the single cell level. The geometry and low throughput of the device presented here makes it infeasible to process meaningful numbers of single cell datasets, although it is possible to fabricate qualitatively similar devices on the micron scale that could achieve this goal.[23] While the sample loss in Gel-seq is variable and hard to accurately quantify, we have observed that anywhere from 10% to 50% of the nucleic acids cannot be recovered from the gel after separation. This number agrees with the literature for similar crush and soak extraction protocols from polyacrylamide gel.[24] When working with 100 to 1000 cells, these losses do not appear to substantially change the resulting libraries. To analyze samples below this limit, however, we will need to modify our protocol.

One approach to improve the protocol could be the use of dissolvable gels to increase sample recovery. We made several attempts at using dissolvable gels during development of the device, but none were successful. Agarose is too porous to be used for the high density gel region and a hybrid device with a separation layer made from agarose and a high density layer made from polyacrylamide was too fragile to handle. We tried using BAC crosslinked polyacrylamide following protocols developed by Hansen,[25] but found low density BAC gels for the separation layer were more fragile than their standard BIS counterparts. For the high density region we found that the gels could not be dissolved, a result Hansen also reported in his work. That said, there are many other dissolvable polymer chemistries, such as DHEBA, that might improve device performance.

We explored the use of a Phi-29 MDA whole-genome amplification, but found it was not necessary, as we were able to recover sufficient starting material from our target input of approximately 100 cells for the Nextera XT protocol. A pre-amplification step before library prep could be added either before or after separation. This could potentially reduce the required cell input, but scaling down cell inputs in our experiments introduced substantial inconsistencies in performance, most likely due to a large coefficient of variation in input when attempting to load small numbers of cells. Even with pre-amplification, we suspect that this issue would hamper meaningful comparisons between samples. Alternatively, recent work has shown that with optimization of lysis conditions, high-quality sequencing libraries can be prepared directly from single cells using Tn5 without pre-amplification.[26]

Although the protocol we adapted from Smart-Seq relies on a poly-T primer, we also added primers with random binding sequences early in our experiments in an attempt to improve performance based on previous work on RNA sequencing from nuclei. We saw no effect, but kept the protocol unchanged for consistency.

As Gel-seq relies on hydrogel immobilization of sample material, it offers interesting possibilities when applied to new methods, such as the potential to change buffer between incompatible protocols without loss of sample material, or to amplify material inside the gel before attempting to extract. Future work in both device fabrication and protocol development could decrease input into the single cell range. A very recent publication from Adam Abate's group shows that single bacterial cells can be encapsulated in agarose hydrogels and uniquely barcoded, allowing 50 000 single-cell whole-genome libraries to be generated in a few hours.[27] The fundamental concepts of separation and library preparation demonstrated in Gel-seq *via* bulk-scale 100 to 1000 cell experiments are also relevant at the single-cell level, and many of the challenges that we faced in developing Gel-seq likely also apply at smaller scales. We believe that the solutions we present in this manuscript are a valuable resource for future work in single-cell genomics using hydrogels.

 

Since Gel-seq does not require a poly-A tail to achieve separation, it is also uniquely positioned for microbial studies, as prokaryotes typically do not polyadenylate their coding transcripts. A modification to the library prep would be required, as we relied primarily on a poly-T Smart-Seq primer, but Gel-seq benefits from an inherent flexibility in terms of different biochemical approaches. Gel-immobilized material can be washed or transferred, for example, into buffers suitable for either a poly-A tailing step or some other total RNA prep method, as long as RNAses are inhibited.

As for input, with microbial studies it might not be necessary to start with the same total mass of DNA as with mammalian genomes. While typical bacteria have only about 0.1% the nucleic acid content of mammalian cells, this also means that far less sequencing effort is needed to reconstruct either the genome or transcriptome. Previous work in the Zhang lab has shown 90% complete *de novo* assembly from a single *E. coli* bacterium after MDA pre-amplification in 12 picoliter PDMS microwells.[1] Even one million paired end 100-base reads yields 200 million bases, which, for a single *E. coli* with 6 million bases total, gives 33× coverage. Assuming sufficiently uniform coverage, this is a enough reads to perform *de novo* assembly. Even the smallest visible colony of *E. coli* that a researcher might pick from a plate using a toothpick may contain more than enough material for Gel-seq. The question that remains to be answered is what amount of material is irrecoverable from the gel barrier. We suspect that the amount of irrecoverable material is likely a function of surface area. Reducing the device geometry to suit a toothpick sized sample might achieve the same goal as pre-amplification when working with microbes.

We have shown in this paper that Gel-seq can be used to generate high quality libraries from vanishingly small populations of cells. It is a flexible protocol that can be used to quickly process samples with an inexpensive and easy-to-fabricate device. The development of a gel based method for preparing next-generation DNA and RNA sequencing libraries from the same cells opens news doors for genomics, allowing researchers to ask if DNA mutations in small numbers of cells affect RNA expression in those same cells. It is also our hope that the physical principals described here might some day be translated to a single-cell technique to allow simultaneous profiling of tens of thousands of single-cell genomes and transcriptomes. Such a device would provide a more general approach for linking DNA variation to RNA expression in complex samples such as tumors or microbial populations.

## Author contributions

GH, AR, YW, and KZ conceived the experiments, GH and AR conducted the experiments, GH, AR, YW, and KZ analyzed the results, KK and XL collected samples from the mice, GF advised KK and XL. GH, AR, and YW wrote the manuscript. AP advised on all aspects of the project. All authors reviewed the manuscript.

## Acknowledgements

The authors would like to acknowledge all the members of the Pisano and Zhang labs. In particular, Blue Lake provided valuable assistance in the development of this protocol. We also thank Erik Rodriguez, Ana Morena, and Dongxin Zhao for their generous donations of PC3, 3T3, and HeLa cell lines. Funding for this work was provided by the National Science Foundation Graduate Research Fellowship Program, NIH grant R01-HG007836, and by the Korean Ministry of Science, ICT and Future Planning.

## References

1 J. Gole, A. Gore, A. Richards, Y.-J. Chiu, H.-L. Fung, D. Bushman, H.-I. Chiang, J. Chun, Y.-H. Lo and K. Zhang, *Nat. Biotechnol.*, 2013, **31**, 1126–1132.
2 Y. Sasagawa, I. Nikaido, T. Hayashi, H. Danno, K. D. Uno, T. Imai and H. R. Ueda, *Genome Biol.*, 2013, **14**, R31.
3 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, *Nat. Biotechnol.*, 2012, **30**, 777–782.
4 E. Shapiro, T. Biezuner and S. Linnarsson, *Nat. Rev. Genet.*, 2013, 1–13.
5 D. E. Spratt, Z. S. Zumsteg, F. Y. Feng and S. A. Tomlins, *Nat. Rev. Clin. Oncol.*, 2016, **13**, 597–610.
6 A. Sottoriva, I. Spiteri, S. G. Piccirillo, A. Touloumis, V. P. Collins, J. C. Marioni, C. Curtis, C. Watts and S. Tavaré, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 4009–4014.
7 S. S. Dey, L. Kester, B. Spanjaard, M. Bienko and A. van Oudenaarden, *Nat. Biotechnol.*, 2015, 1–7.
8 I. C. Macaulay, W. Haerty, P. Kumar, Y. I. Li, T. X. Hu, M. J. Teng, M. Goolam, N. Saurat, P. Coupland, L. M. Shirley, M. Smith, N. Van der Aa, R. Banerjee, P. D. Ellis, M. A. Quail, H. P. Swerdlow, M. Zernicka-Goetz, F. J. Livesey, C. P. Ponting and T. Voet, *Nat. Methods*, 2015, **12**, 519–522.
9 T. A. Zangle, A. Mani and J. G. Santiago, *Chem. Soc. Rev.*, 2010, **39**, 1014–1035.
10 J.-L. Viovy, *Rev. Mod. Phys.*, 2000, **72**, 813.
11 L. Ornstein, *Ann. N. Y. Acad. Sci.*, 1964, **121**, 321–349.
12 B. J. Davis, *Ann. N. Y. Acad. Sci.*, 1964, **121**, 404–427.
13 Illumina, *Nextera XT DNA Library Preparation Guide, Part 15031942 Rev. E. edn*, 2015.
14 Clontech, *SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing User Manual*, 2016.
15 Y. Suzuki, D. Ishihara, M. Sasaki, H. Nakagawa, H. Hata, T. Tsunoda, M. Watanabe, T. Komatsu, T. Ota and T. Isogai, *et al.*, *Genomics*, 2000, **64**, 286–297.
16 R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat and J. Troge, *et al.*, *Genome Res.*, 2003, **13**, 2291–2305.
17 J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker and M. Chi, *et al.*, *Science*, 2004, **305**, 525–528.
18 G. Guffanti, F. Torri, J. Rasmussen, A. P. Clark, A. Lakatos, J. A. Turner, J. H. Fallon, A. J. Saykin, M. Weiner and M. P. Vawter, *et al.*, *Genomics*, 2013, **102**, 112–122.

19 J. T. Glessner, K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune and J. P. Bradfield, *et al.*, *Nature*, 2009, **459**, 569–573.
20 S.-I. Consortium, *et al.*, *Nat. Biotechnol.*, 2014, **32**, 903–914.
21 K. Kurimoto, Y. Yabuta, Y. Ohinata and M. Saitou, *Nat. Protoc.*, 2007, **2**, 739–752.
22 A. Lemire, K. Lea, D. Batten, S. J. Gu, P. Whitley, K. Bramlett and L. Qu, *J. Biomol. Tech.*, 2011, **22**, S46.
23 H. S. Lee, W. K. Chu, K. Zhang and X. Huang, *Lab Chip*, 2013, **13**, 3389–3397.
24 J. Sambrook and D. W. Russell, *Cold Spring Harb. Protoc.*, 2006.
25 J. N. Hansen, *Anal. Biochem.*, 1981, **116**, 146–151.
26 H. Zhan, A. Steif, E. Laks, P. Eirew, M. VanInsberghe, S. P. Shah, S. Aparicio and C. L. Hansen, *Nat. Methods*, 2017, **14**, 167–173.
27 F. Lan, B. Demaree, N. Ahmed and A. R. Abate, *Nat. Biotechnol.*, 2017, DOI: 10.1038/nbt.3880.

 