Electronic Supplemental Information:

Gel-seq: Whole-Genome and Transcriptome Sequencing by Simultaneous Low-Input DNA and RNA Library Preparation Using Semi-Permeable Hydrogel Barriers

Additional Experimental Techniques

Positive Control Library Preparation: Standard protocols were used to generate reference libraries as a comparison to our Gel-seq protocol. To generate libraries from RNA, we followed the Smart-Seq[1] and Nextera XT[2] manuals. The only modification we made to these protocols was to use half volume reactions and the addition of random priming to the reverse transcription step of Smart-Seq. To generate libraries from genomic DNA, we lysed cells using a simple lysis buffer developed by Shatzkes.[3] Once the cells were lysed, we followed the standard Nextera XT manual using half volume reactions.[2]

Cell Culture: PC3 was cultured in F-12K media (Gibco) supplemented with 10% heat-inactivated (HI) FBS (Gibco) and 1% penicillin/streptomycin (P/S) (Gibco). HeLa was cultured in Eagle's Minimum Essential Medium (ATCC) supplemented with 10% HI FBS and 1% P/S. 3T3 was cultured in high-glucose Dulbecco's Modified Eagle's Medium (4.5 g/L glucose and L-glutamine) supplemented 10% HI FBS and 1% P/S and 3T3 cell lines were cultured in DMEM with 4.5 g/ml glucose and 1 mM sodium pyruvate (ThermoFisher Scientific) supplemented with 10% heat-inactivated FBS (ThermoFisher Scientific) and 1% penicillin-streptomycin (ThermoFisher Scientific).

Mouse Primary Hepatocyte Collection: Mouse livers were perfused with a classic two-step method. Briefly, livers were perfused via the portal vein with 20 ml of pre-warmed wash buffer followed by 20 ml of digestion buffer containing 5000 U collagenase Type IV (Gibco) and 5000 U collagenase type I (Worthington). After perfusion, tissue was cut as small as possible, passed through 100-μm cell strainer, and centrifuged at 50g for 5 min to pellet hepatocytes. The animal protocols (s09108) for all procedures were approved by the UCSD Institutional Animal Care and Use Committee (IACUC). All methods were performed in accordance with the relevant guidelines and regulations.

Analysis of Sequencing Data

Sequencing and De-multiplexing: Libraries were sequenced on a MiSeq (Illumina) using v3 kits and standard sequencing primers. Libraries were loaded at 27-30 pM and at least 50 cycles were obtained for read 1 for each experiment, plus 8 cycles for Index 1 and 8 cycles for Index 2. Base calls were de-multiplexed to fastq using bcl2fastq.

Extrapolation Simulations: Library complexity and genomic coverage simulations were performed with preseq[4] using extrac_lc and extrap_gc, respectively, with 100 bootstrapping iterations each.

DNA Mapping and CNV Calling: Copy number profiling on DNA libraries was performed as described in Baslan et al. with minor modifications.[5] Briefly, reads were trimmed to 36 bases using fastx (14 bases from the start and all bases after 50) and mapped to GRCh38 or mm10 with bowtie.[6] For both mouse and human, alignments were counted across 25,000 bins whose boundaries were calculated such that mapping their respective reference genomes would generate equal counts per bin. Bin counts were then normalized to mean for each sample and GC corrected in matlab by lowess regression based on GC content. No segmentation was performed. Pearson correlations were performed in Python using scipy[7] and plotted using matplotlib[8].

RNA Mapping and PCA: RNA fastq was mapped to a pre-annotated index constructed from either GRCh38.87 (human, hg38) or GRCm38.87 (mouse, mm10) using STAR[9] and read counts for each gene were converted to TPM. A threshold of TPM $>$ 5 was applied. Pearson correlations and PCA were performed in Python using scipy[7] and scikit-learn[10], respectively, and plotted using matplotlib[8]. Base-wise percentage of reads mapping to exons was calculated from GenCode GTF files using bedtools coverage. Base-wise percentage of reads mapping to ribosomal and transfer RNA was calculated from Ensembl BioMart BED files, also using bedtools coverage. Percentage of RNA reads mapping to mitochondrial genes was calculated from alignments using grep to search for the MT in the chromosome field.

Sequencing pool	Type	Sample
A	DNA	3T3-G1
		3T3-G2
		Liver-G1
		Liver-G2
		3T3-T1
		3T3-T2
		Liver-T1
		Liver-T2
B	RNA	3T3-G1
		3T3-G2
		3T3-T1
		3T3-T2
C	RNA	Liver-G1
		Liver-G2
		Liver-T1
		Liver-T2
D	DNA	HeLa-G1
		HeLa-G2
		HeLa-T1
		HeLa-T2
		PC3-G1
		PC3-G2
		PC3-T1
		PC3-T2
E	RNA	HeLa-G1
		HeLa-G2
		HeLa-T1
		HeLa-T2
		PC3-G1
		PC3-G2
		PC3-T1
		PC3-T2

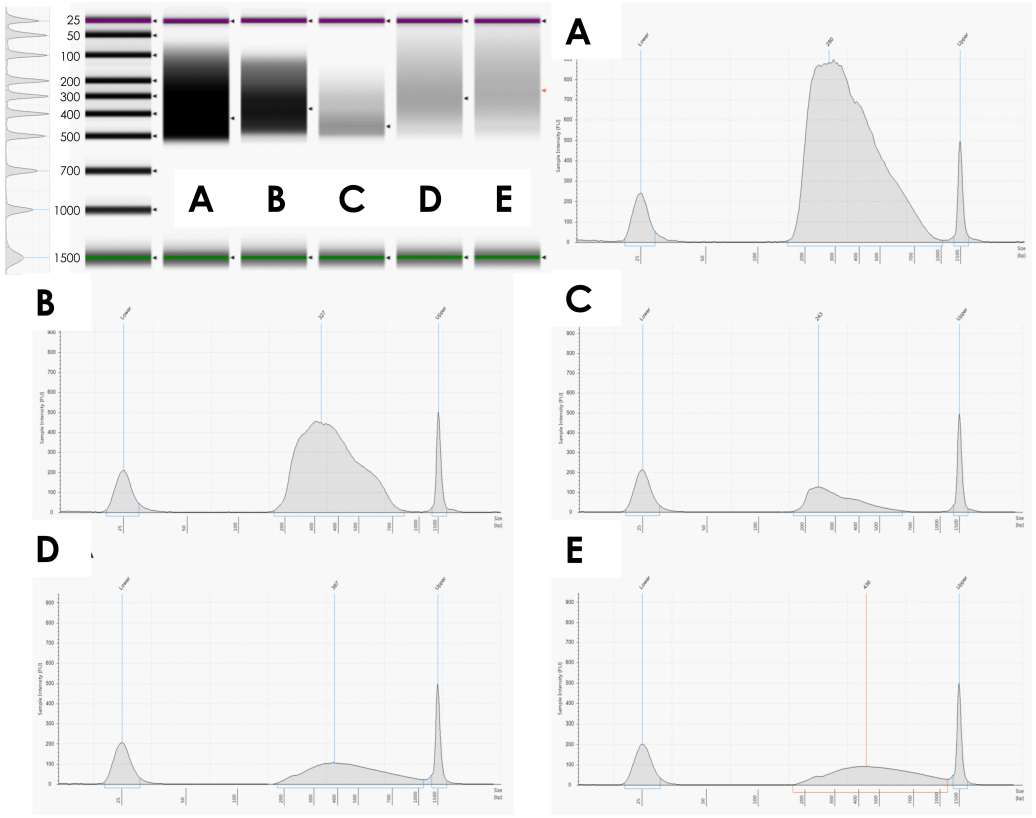

Fig. S1 TapeStation traces for all sequencing library pools.

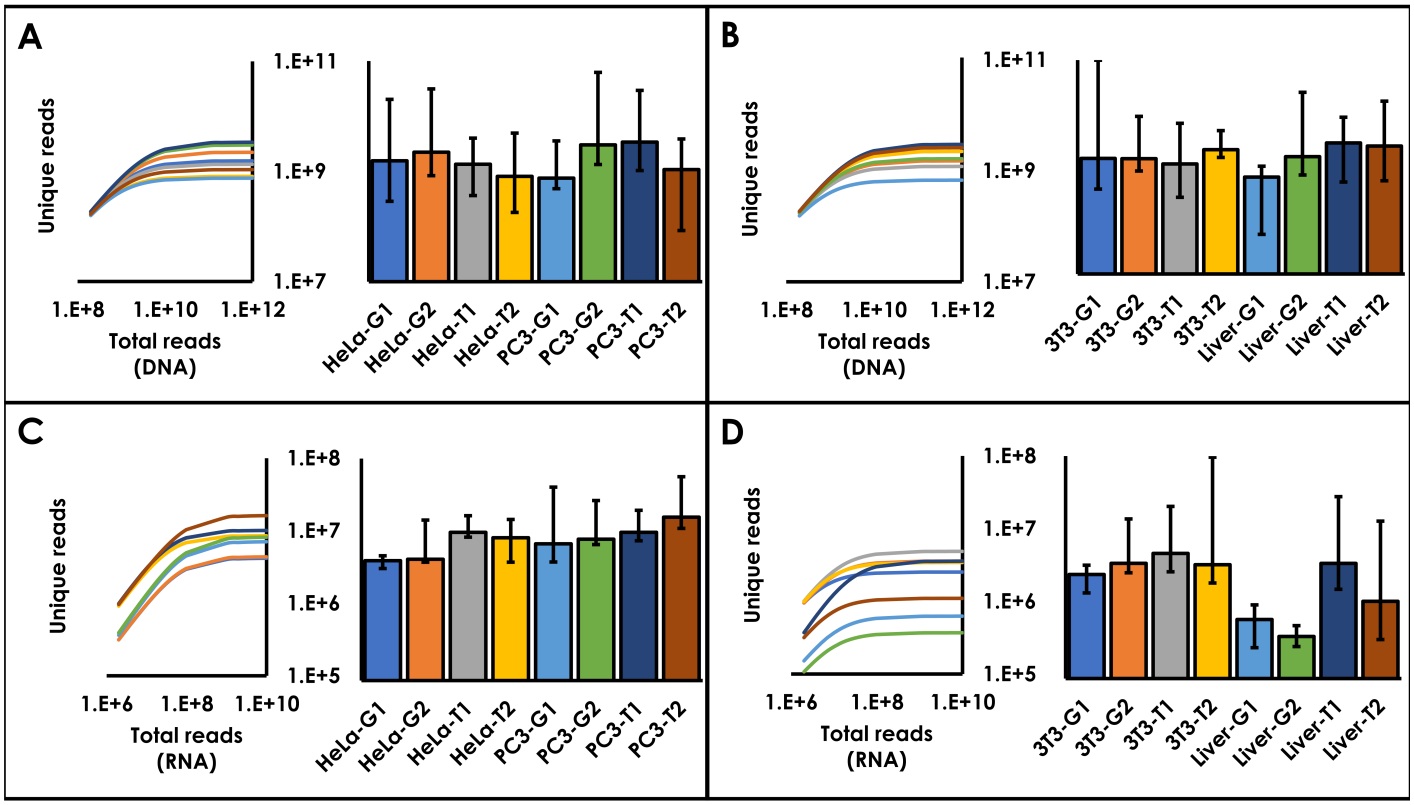

Fig. S2 Panels A and B show predicted genomic coverage as a function of depth of coverage in DNA libraries from human and mouse samples, respectively. Bar graphs represent maximum predicted coverage at saturation. Panels C and D show predicted library complexity for RNA libraries from human and mouse samples, respectively. Error bars are 95% confidence intervals created from 100 bootstrapping simulations.

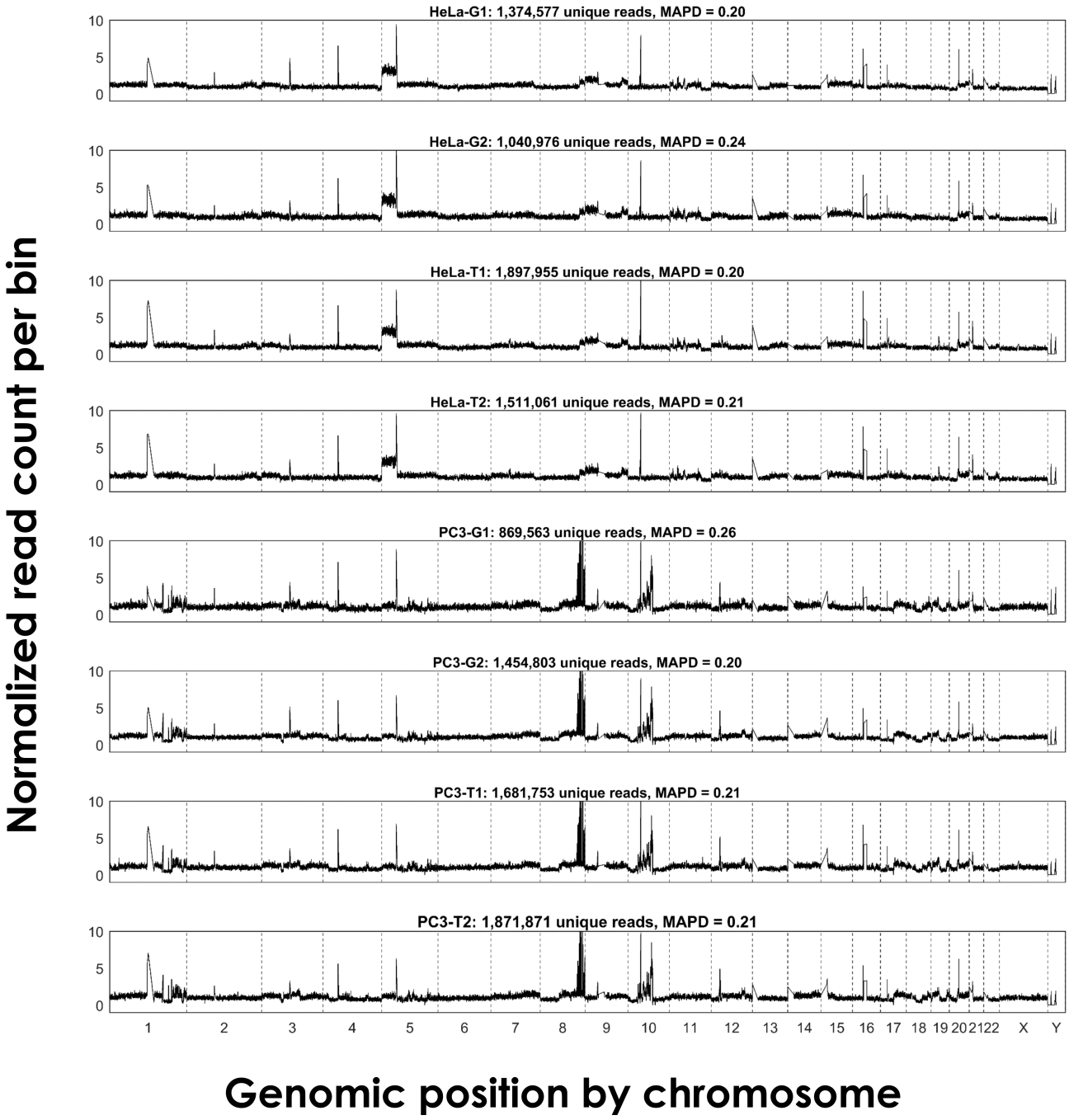

Fig. S3 Copy number profiles across 25,000 bins for human DNA libraries from HeLa and PC3 cell lines. G and T in sample names indicate Gel device and Tube controls, respectively, while numbering indicates technical replicates. Horizontal axis denotes chromosome and bin position, vertical axis denotes mean normalized bin count (corrected for GC content).

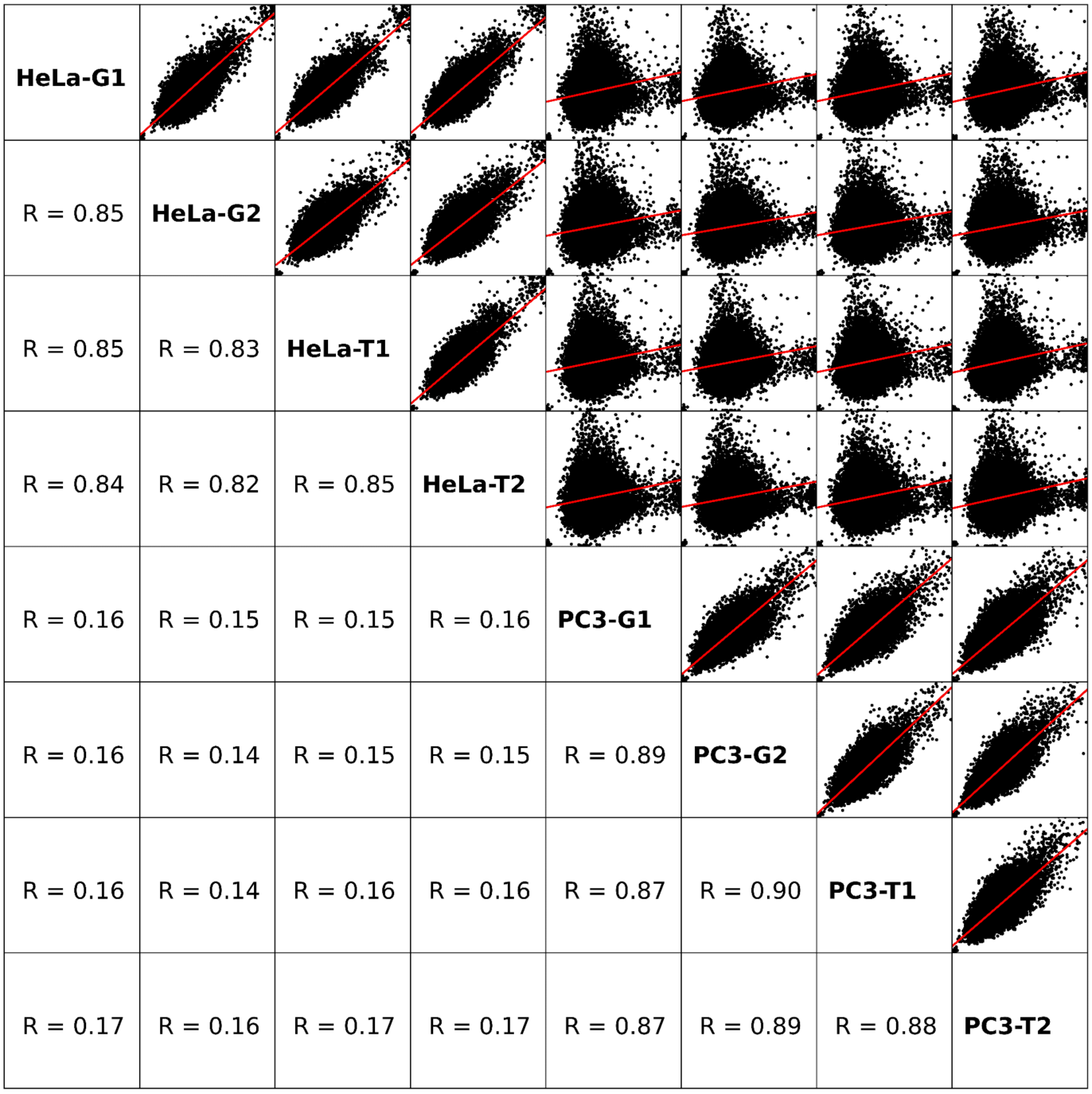

Fig. S4 Pairwise correlations between bin counts for human DNA libraries from HeLa and PC3 cell lines. Main diagonal entries are sample names, lower diagonal entries are Pearson correlation coefficients.

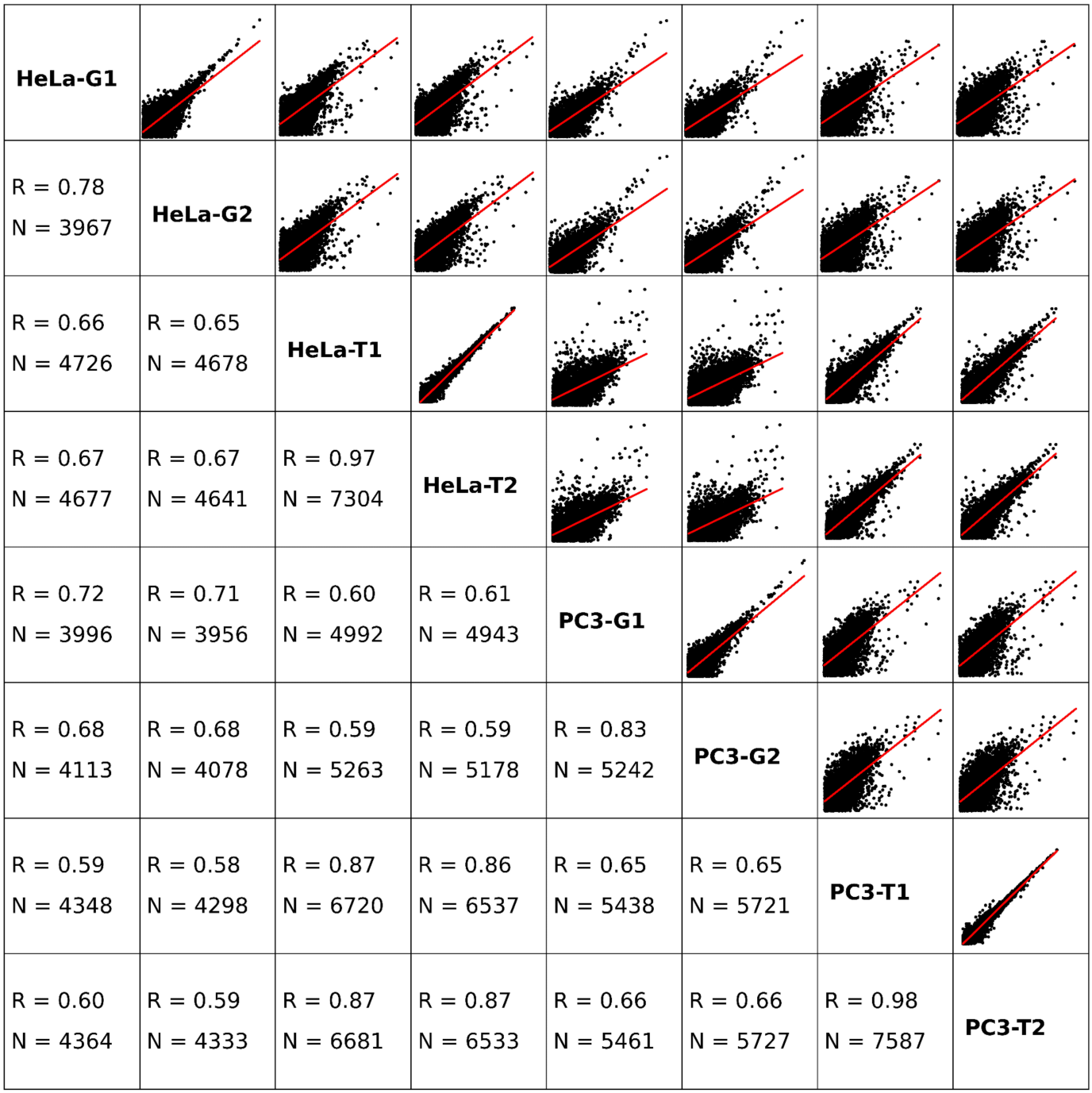

Fig. S5 Pairwise correlations between detected gene counts for all human RNA libraries from HeLa and PC3. Lower diagonal entries are Pearson correlation coefficients (R value) and number of detected genes (N values, genes with non-zero counts) in common for each comparison.

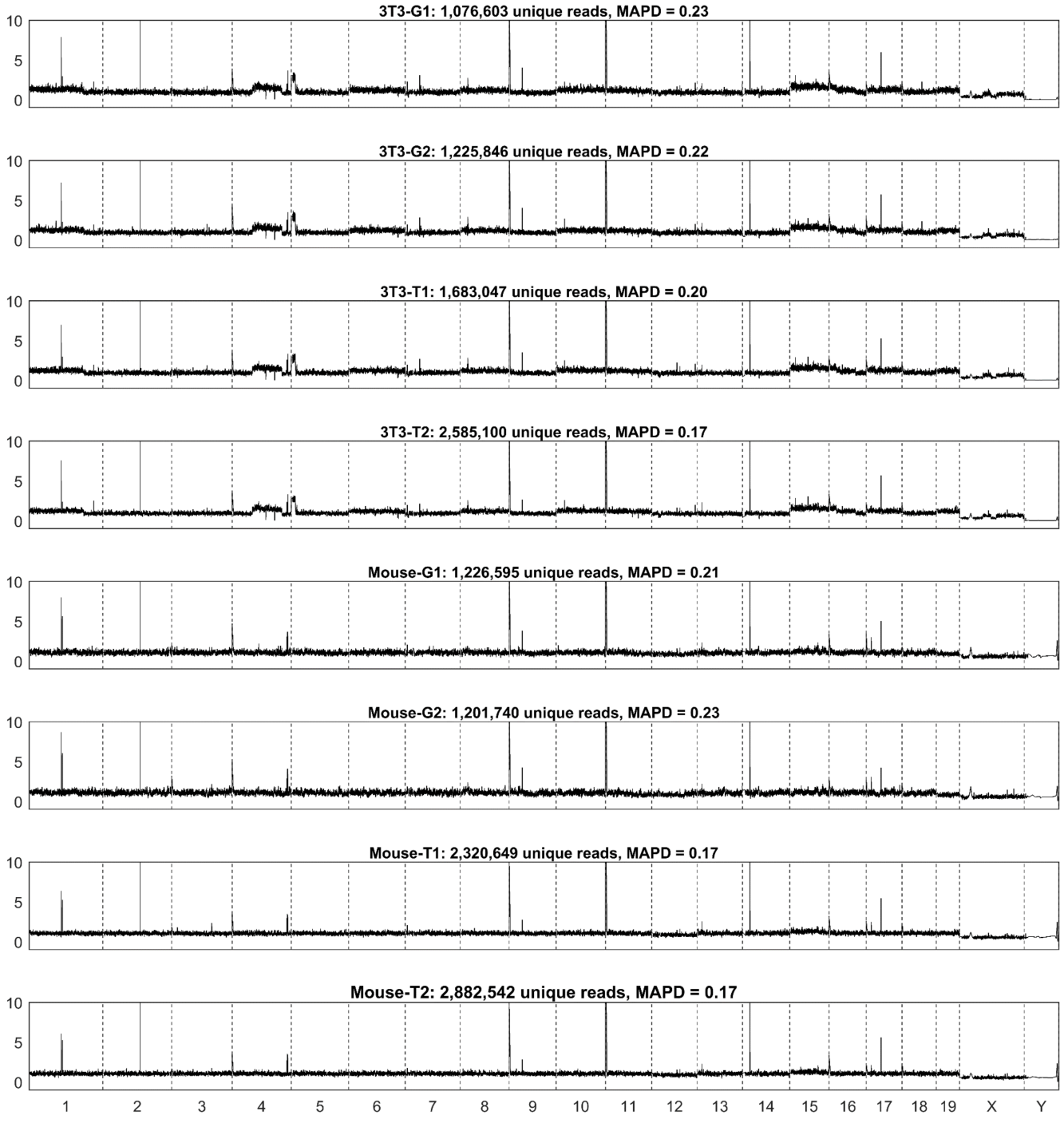

Fig. S6 Copy number profiles across 25,000 bins for mouse DNA libraries from a 3T3 cell line and mouse primary tissue. G and T in sample names indicate Gel device and Tube controls, respectively, while numbering indicates technical replicates. Horizontal axis denotes chromosome and bin position, vertical axis denotes mean normalized bin count (corrected for GC content). Extreme peaks in mouse primary samples are due to "bad bins" in the reference genome, in which repetitive sequences present in the true genome are not included in the reference genome, leading to a false pile up of reads from experimental data.

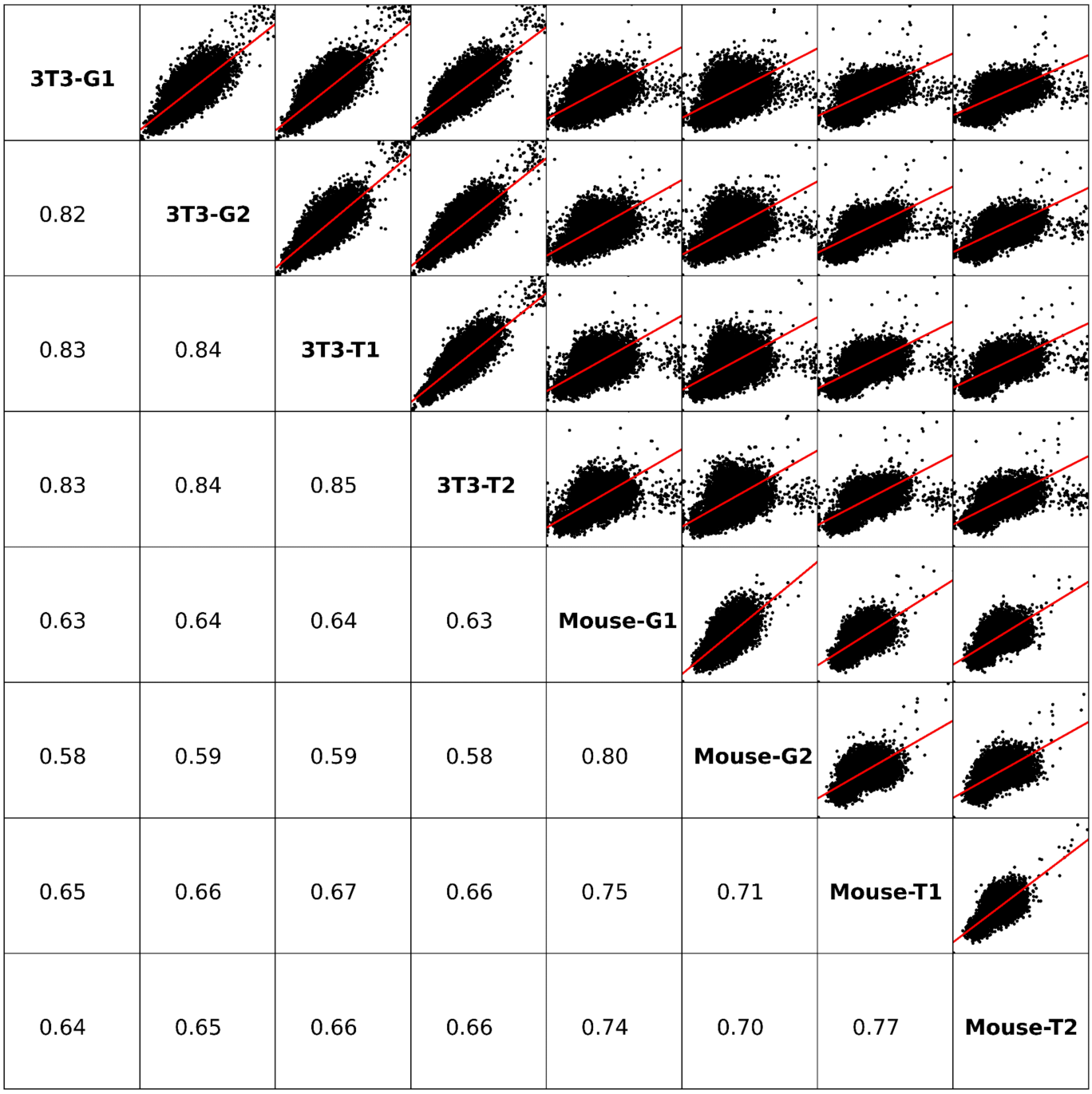

Fig. S7 Pairwise correlations between bin counts for mouse DNA libraries from 3T3 and mouse tissue. Main diagonal entries are sample names, lower diagonal entries are Pearson correlation coefficients. Correlations are weaker than for PC3 versus HeLa due to less extreme copy number variation in 3T3 and almost none in the mouse primary sample, leading to little dynamic range in bin counts.

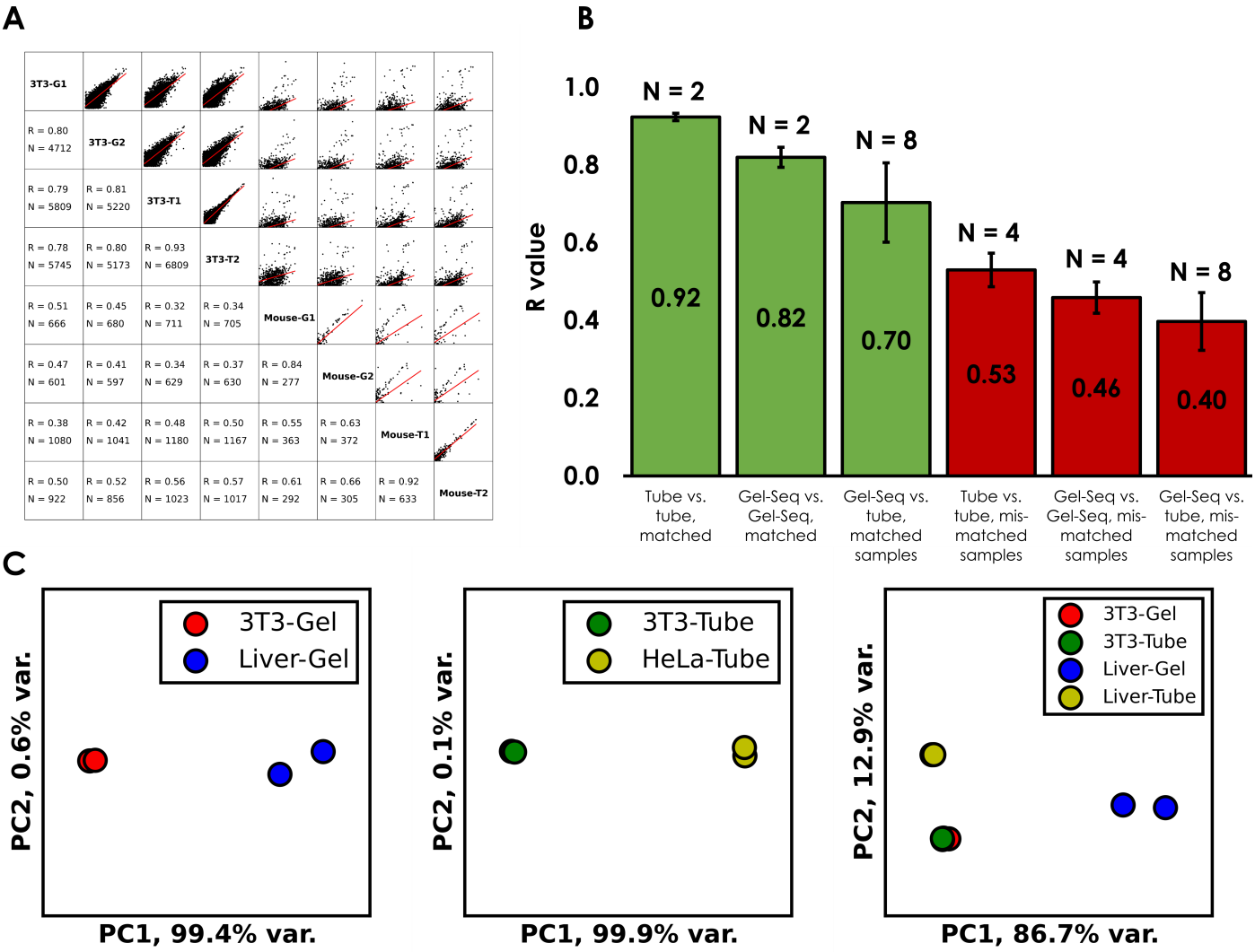

Fig. S8 Panel A shows all 28 pair-wise correlations between gene counts (TPM > 5) for RNA libraries from mouse 3T3 fibroblast cell line and mouse liver. Lower diagonal entries are Pearson correlation coefficients (R value) and number of detected genes (N values, genes with non-zero counts) in common for each comparison. Panel B shows average R values for different comparison types (Error bars are standard deviation, N is number of comparisons in each type.) Panel C shows PCA separation for Gel-seq, standard tube method controls, and all 8 together.

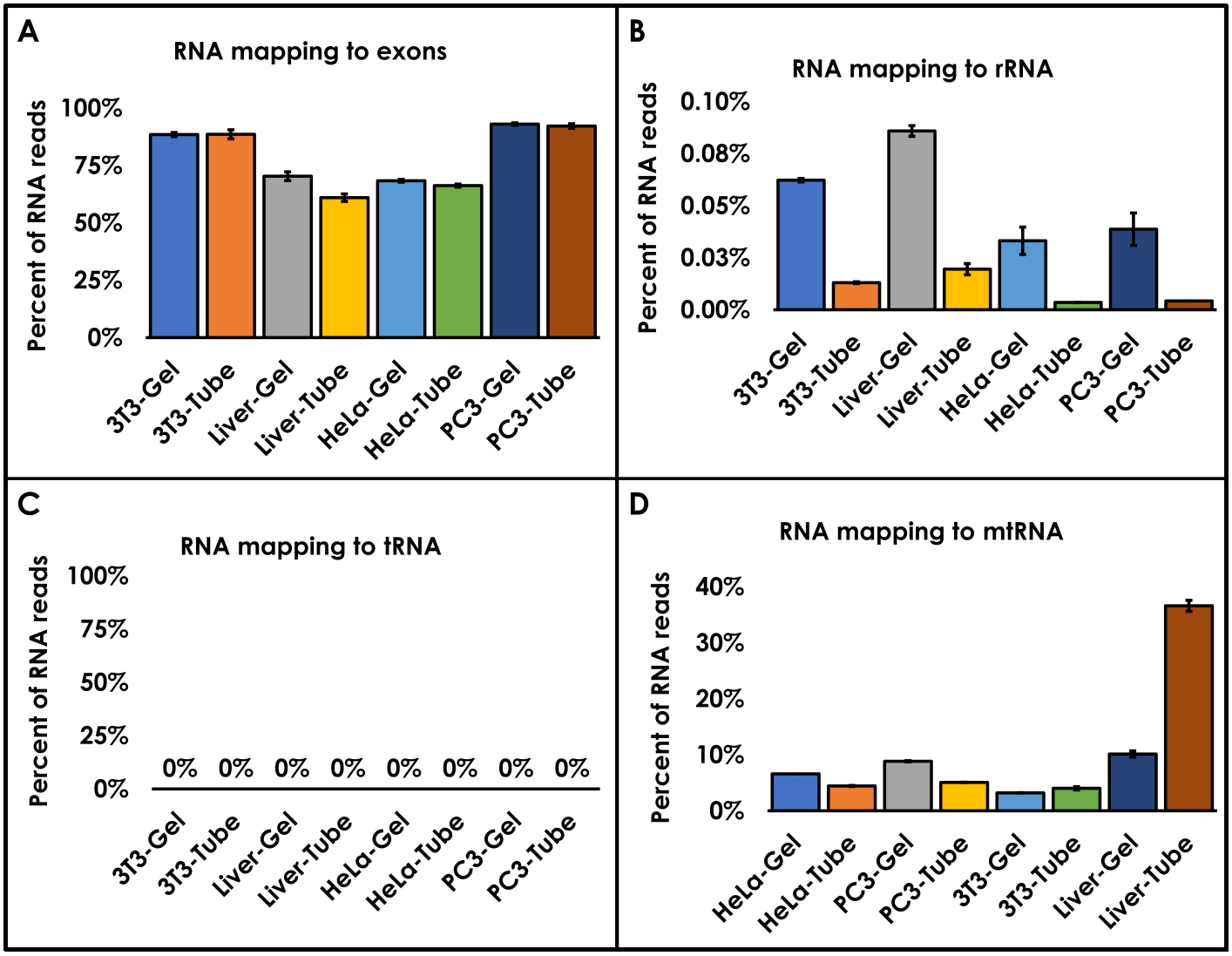

Fig. S9 Panel A shows base-wise percentage of RNA alignments in annotated exons in either GRCh38 or mm10 reference genomes. Only mouse liver samples show any significant difference, with Gel-seq data mapping to exonic regions at a slightly higher rate ($p = 0.03$, two-tailed t-test, unequal variance). Panel B shows base-wise percentage of RNA alignments in ribosomal RNA genes (rRNA). There is evidence that polyadenylation of rRNA acts as a degradation signal,[11] and short rRNA degradation fragments (less than 1000 bases) would be expected to migrate faster than the majority of mRNA, which might explain why Gel-seq detects a higher proportion of rRNA than tube controls. Panel C shows base-wise percentage of RNA alignments in transfer RNA genes (tRNA). tRNA is short and not polyadenylated, which likely explains why we see no mapping when using the Smart-Seq poly-T primers. It seems that the random priming also failed to detect tRNA in either Gel-seq or tube controls, possibly due to the short length of tRNA. Panel D shows the percentage of RNA reads mapping to the mitochondrial chromosome. Hepatocytes contain very large number of mitochondria, so a high RNA mapping rate to mitochondrial genes is not necessarily surprising in liver, although we cannot fully explain why Gel-seq detected less than the tube controls in this comparison.

Table S1 Mapping statistics for all 32 libraries. Gel samples are paired, i.e., DNA and RNA libraries are from the same cells after physical separation in Gel-seq device. Tube data were generated using standard protocols by splitting cells prior to library prep. Mouse liver Gel-seq samples have few aligned reads due to a buffer contamination issue, yet still generated enough data to discriminate from 3T3 cell line by PCA.

			DNA		RNA	
Cell type	Method	#	Aligned	Unique	Aligned	Genes detected
HeLa	Gel-seq	1	1,392,072	1,374,577 (99%)	720,874	5,795
		2	1,050,948	1,040,976 (99%)	856,478	5,791
	Tube	3	1,929,343	1,897,955 (98%)	1,046,566	8,395
		4	1,533,770	1,511,061 (99%)	834,367	8,242
PC3	Gel-seq	5	884,454	869,563 (98%)	651,622	6,965
		6	1,480,236	1,454,803 (98%)	651,622	6,965
	Tube	7	1,749,120	1,681,753 (96%)	1,021,107	8,656
		8	1,953,548	1,871, 871 (96%)	1,408,115	8,477
3T3	Gel-seq	9	1,100,593	1,076,603 (98)%	113,976	6,800
		10	1,256,040	1,225,846 (98%)	191,796	6,117
	Tube	11	1,731,667	1,683,047 (97%)	154,194	8,199
		12	2,666,911	2,585,100 (97%)	138,993	8,089
Hepatocytes	Gel-seq	13	1,255, 489	1,226,595 (98%)	70,488	873
		14	1,229,987	1,201,740 (98%)	57,882	897
	Tube	15	2,385,582	2,320,649 (97%)	64,382	2,045
		16	2,967,453	2,882,542 (97%)	26,220	1,638

References

1 Clontech, *SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing User Manual*, 2016.

2 Illumina, *Nextera XT DNA Library Preparation Guide*, Part 15031942 Rev. E. edn, 2015.

3 K. Shatzkes, B. Teferedegne and H. Murata, *Scientific Reports*, 2014, **4**, 1–7.

4 T. Daley and A. D. Smith, *Nature Methods*, 2013, **10**, 325–327.

5 T. Baslan, J. Kendall, L. Rodgers, H. Cox, M. Riggs, A. Stepansky, J. Troge, K. Ravi, D. Esposito, B. Lakshmi, M. Wigler, N. Navin and J. Hicks, *Nature Protocols*, 2012, **7**, 1024–1041.

6 B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, *Genome Biology*, 2009, **10**, R25.1–R25.10.

7 E. Jones, T. Oliphant, P. Peterson *et al.*, *SciPy: Open source scientific tools for Python*, www.scipy.org, 2001, Online; accessed April 10, 2017.

8 J. D. Hunter, *Computing In Science & Engineering*, 2007, **9**, 90–95.

9 A. Dobin, C. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. Gingeras, *Bioinformatics*, 2013, **29**, 15–21.

10 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

11 S. Slomovic, D. Laufer, D. Geiger and G. Schuster, *Nucleic Acids Research*, 2006, **34**, 2966–2975.