TempO-seq and RNA-seq Gene Expression Levels are Highly Correlated for Most Genes: A Comparison Using 39 Human Cell Lines

Dr. Laura Word

ASCCT Webinar

The views are the author's and do not necessarily represent the views of the U.S.EPA. The author declares no conflicts of interest.

Why care about transcriptomics data?

- The EPA high-throughput transcriptomics (HTTr) team is working on identifying patterns of effect when chemicals impact the same gene target

  - This research can help us to predict the bioactivity of chemicals (without animal exposures)

Messenger RNA Sequencing for Transcriptomics

- Quantifying levels of mRNA in cells is **helpful for understanding changes in gene expression** (such as in response to chemical exposure)

- There are **different technologies** for mRNA sequencing, including:
  - RNA-seq using Illumina
  - TempO-seq from BioSpyder

- Can sequence mRNA across the human genome (approximately 20,000 genes)

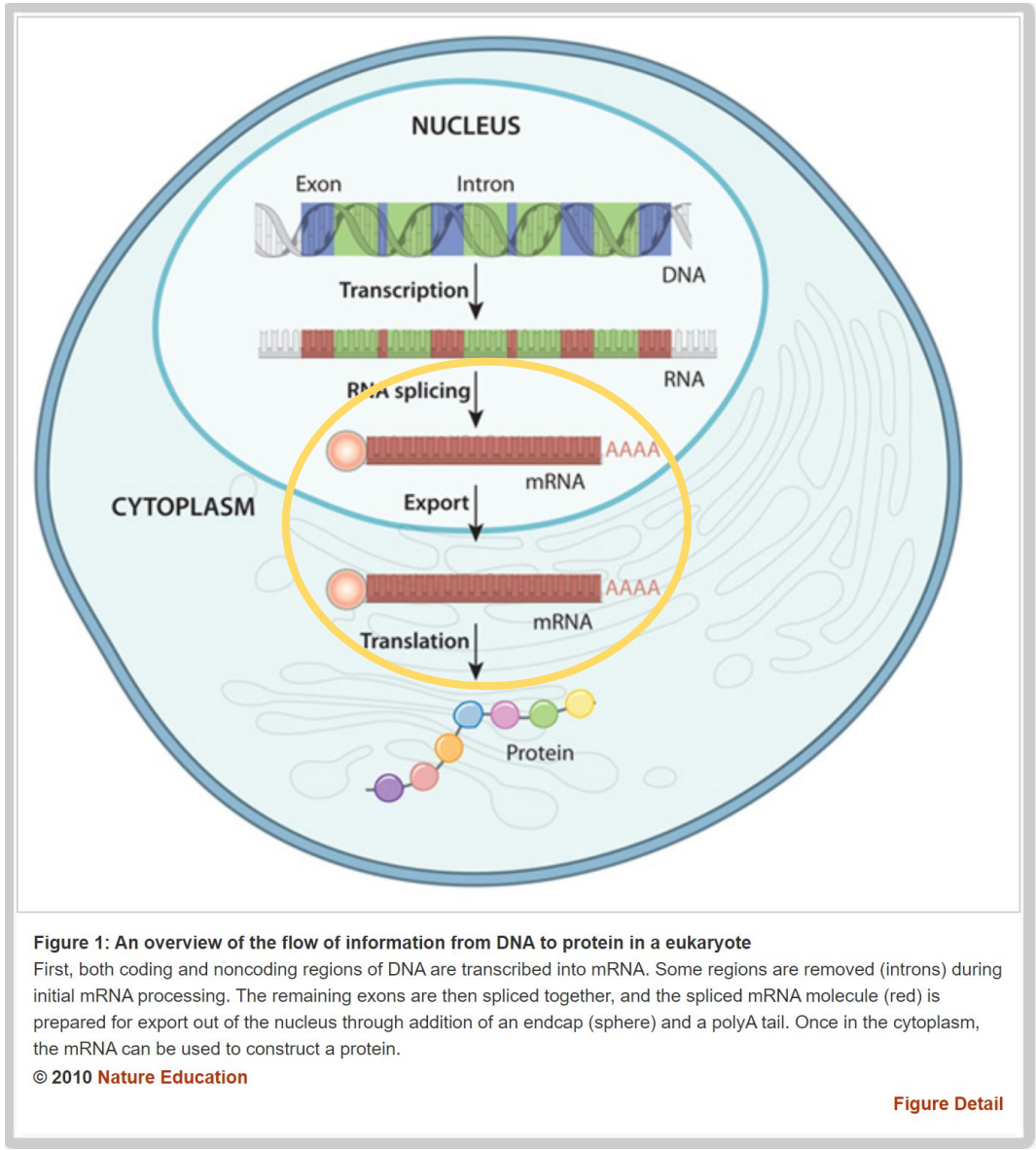


Figure 1: An overview of the flow of information from DNA to protein in a eukaryote
First, both coding and noncoding regions of DNA are transcribed into mRNA. Some regions are removed (introns) during initial mRNA processing. The remaining exons are then spliced together, and the spliced mRNA molecule (red) is prepared for export out of the nucleus through addition of an endcap (sphere) and a polyA tail. Once in the cytoplasm, the mRNA can be used to construct a protein.

Figure Detail

RNA-seq Method

(more established)

Key features:

- **Gold-standard, established method**
- **Non-targeted sequencing of RNA**, so all RNA is quantified and species type does not have to be known
- **Requires purification of RNA** before quantification
- Fragments of RNA are sequenced and later aligned for data analysis, **requiring significant computing resources**

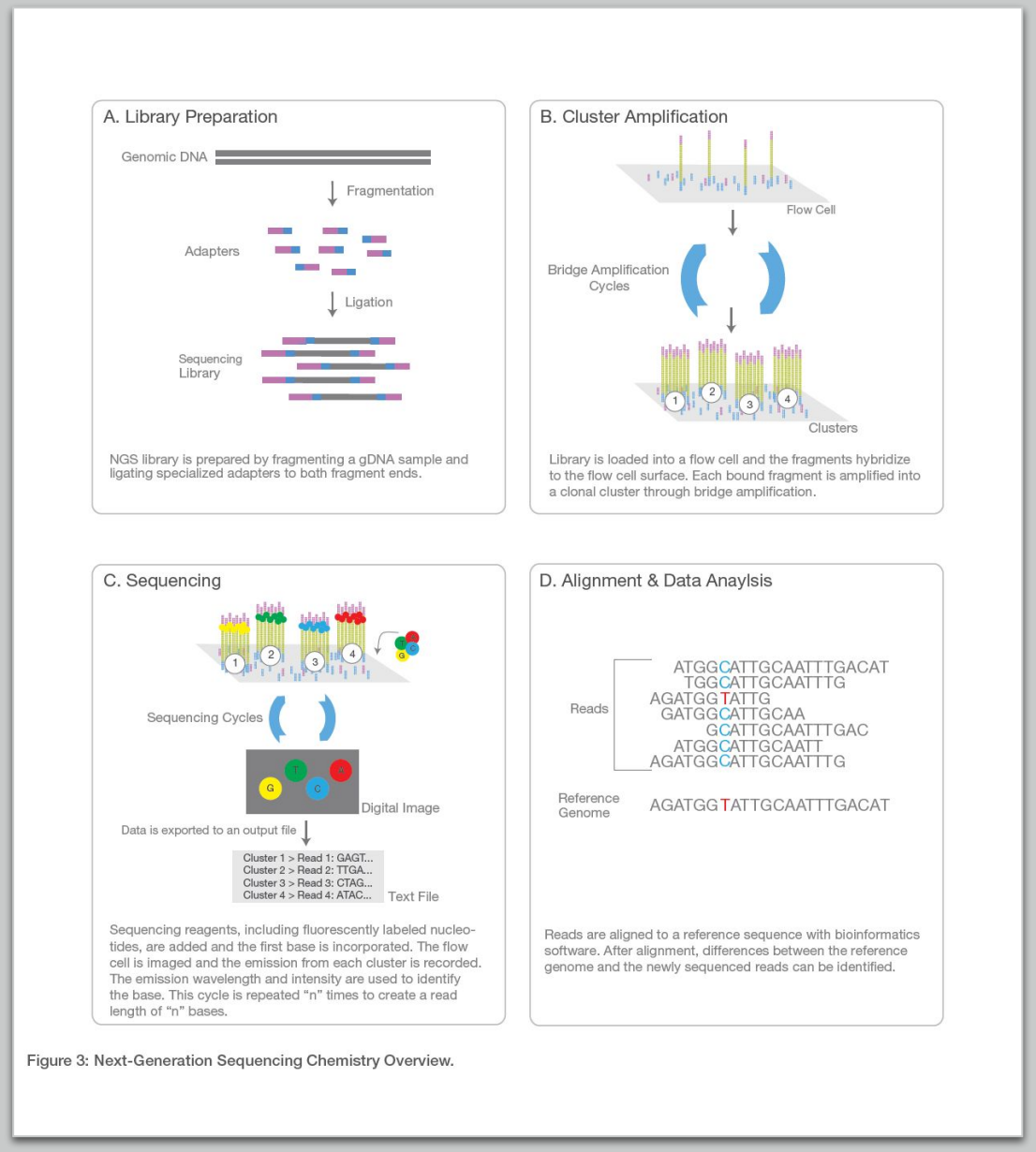


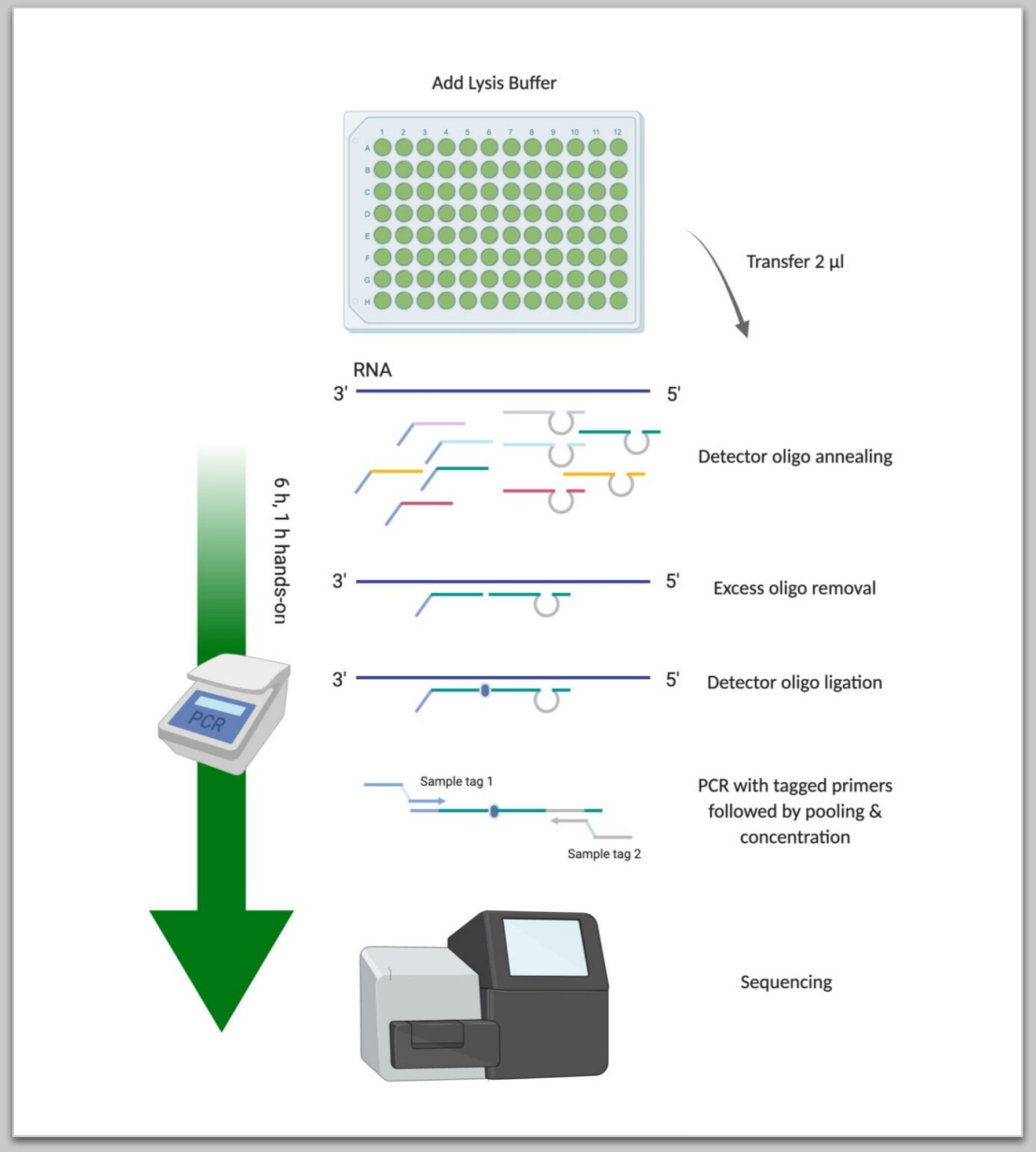
Figure 3: Next-Generation Sequencing Chemistry Overview.

TempO-seq Method

(newer technology)

Key features:

- **Easier sample prep** because lysed cells can be used
- **Less sample material is needed** (picograms instead of nanograms)
- **Possible to customize** which transcripts are quantified
- **Can be less expensive** per sample at high scale
- **Must have detector oligo (DO) probes for the species**, only quantifies RNA for which there is a tag to measure it



Previous Research: prior case studies show TempO-seq is as consistent and sensitive at detecting changes in gene expression as RNA-seq

Fresh cell and tissue samples:

- **Yeakley 2017:** found that TempO-Seq had high correlation with fold differences measured by RNA-seq (R^2 = 0.9) for more than 20,000 targets following exposure of MCF-7 cells to the histone deacetylase inhibitor Trichostatin A (TSA).
- **Bushel 2018**: compared data from the TempO-seq S1500+ surrogate transcriptome (2,284 genes) to whole transcriptome RNA-seq. Purified RNA from liver samples of rats showed some technological platform differences but the statistical analysis grouped by the 5 different mechanisms of action (MOAs) for the 15 chemicals.
    - TempO-seq data had a higher (better) signal to noise ratio, less unexplained variance, and more reproducibility between biological replicates compared to RNA-seq, which they found to be partly due to TempO-seq having less variation in detection of lowly expressed genes.

Frozen and formalin-fixed paraffin-embedded (FFPE) samples:

- **Turnbull 2020:** recommended TempO-seq as the preferable choice when analyzing human breast cancer samples with very limited quantity.
- **Cannizzo 2022:** determined that TempO-seq provided more consistent fold-change results for differentially expressed genes (DEGs) within frozen and FFPE mouse liver samples.

Previous Research: prior case studies show TempO-seq is as consistent and sensitive at detecting changes in gene expression as RNA-seq

Fresh cell and tissue samples:

- **Yeakley 2017:** found that TempO-Seq had high correlation with fold differences measured by RNA-seq (R^2 = 0.9) for more than 20,000 targets following exposure of MCF-7 cells to the histone deacetylase inhibitor Trichostatin A (TSA).
- **Bushel 2018**: compared data from the TempO-seq S1500+ surrogate transcriptome (2,284 genes) to whole transcriptome RNA-seq. Purified RNA from liver samples of rats showed some technological platform differences but the statistical analysis grouped by the 5 different mechanisms of action (MOAs) for the 15 chemicals.
    - TempO-seq data had a higher (better) signal to noise ratio, less unexplained variance, and more reproducibility between biological replicates compared to RNA-seq, which they found to be partly due to TempO-seq having less variation in detection of lowly expressed genes.

Frozen and formalin-fixed paraffin-embedded (FFPE) samples:

- **Turnbull 2020:** recommended TempO-seq as the preferable choice when analyzing human breast cancer samples with very limited quantity.
- **Cannizzo 2022:** determined that TempO-seq provided more consistent fold-change results for differentially expressed genes (DEGs) within frozen and FFPE mouse liver samples.

A need remained for comparing lysed cells for the full transcriptome baseline gene expression in human samples across more cell types

TempO-seq: EPA Phase 1 and Phase 2 Data

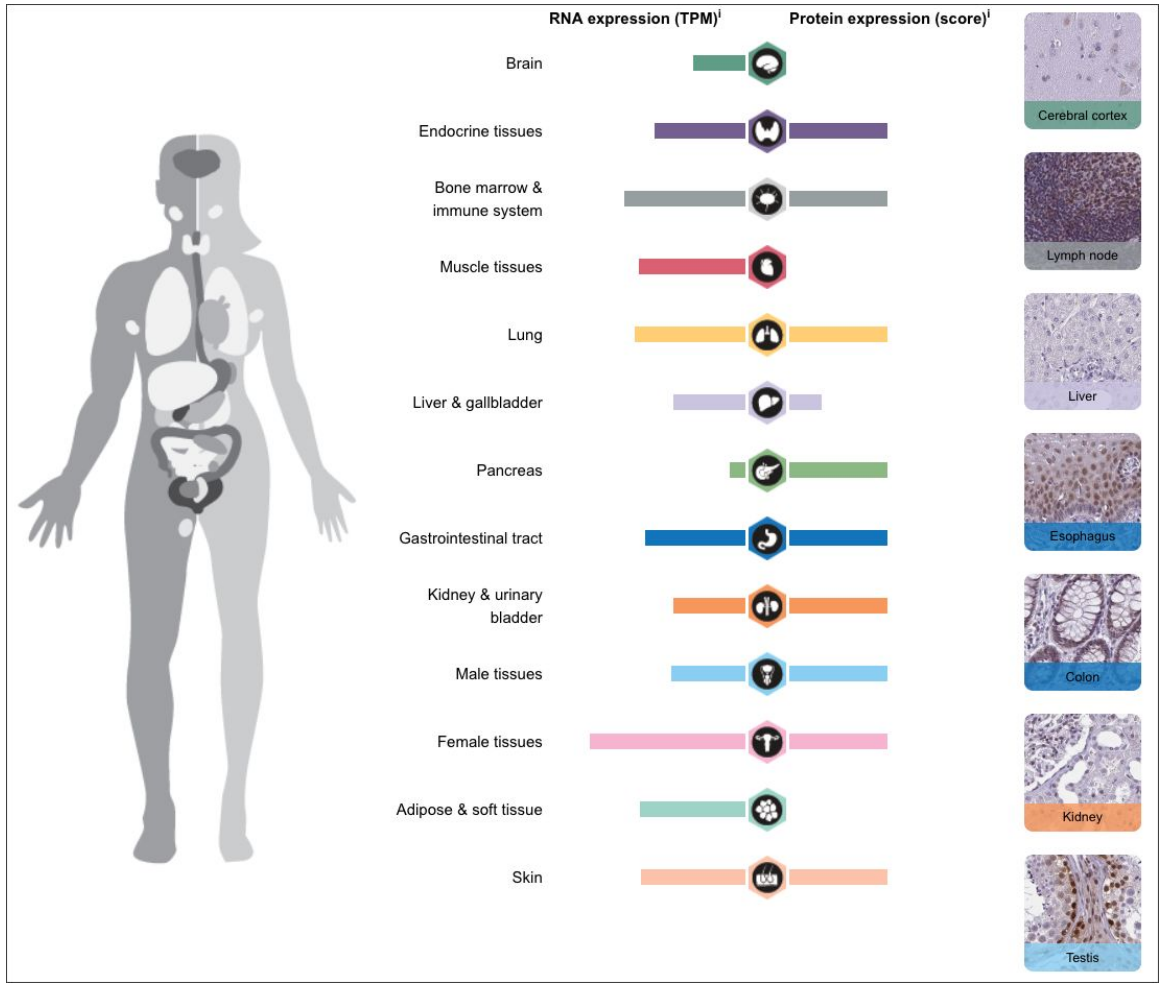
- Baseline gene expression
- Both of these TempO-seq data sets were generated at the EPA in 2018-2019
  - Phase 1 = 6 million read depth
  - Phase 2 = 4.5 million read depth
- Clinton Willis performed sample collection for both data sets
- Cells came from independent cultures but were from the same cryostocks

Unilever

RNA-seq data: Human Protein Atlas

THE HUMAN PROTEIN ATLAS

- Publicly available RNA and protein baseline expression data for many tissues of the human body
- RNA-seq data at approximately 20 million reads depth
- More details: *HPA is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology*



Step 1. Compare the TempO-seq Phase 1 and Phase 2 Data Sets

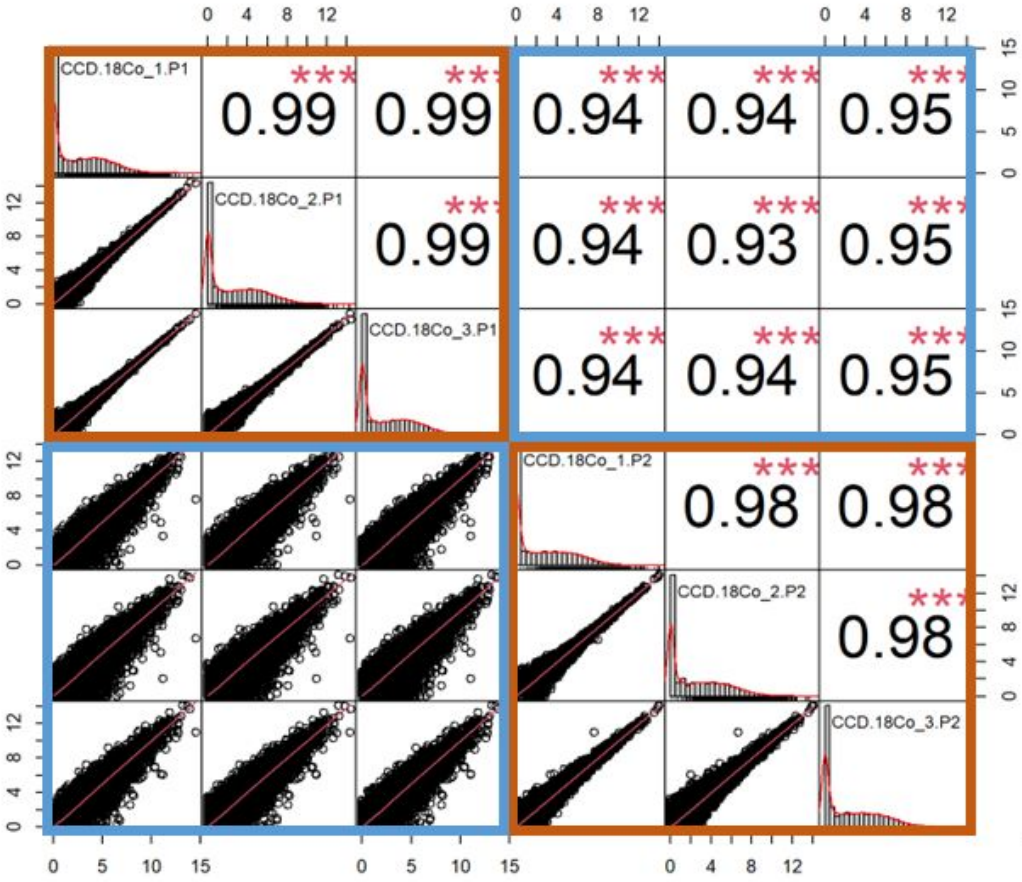
Common Cell Types: TempO-seq Phase 1 and Phase 2

Cell Line	ExPASy CelloSaurus Accession	Tissue Origin	Disease	Growth Mode	Morphology	Source
MCF-7	CVCL_0031	Breast	Adenocarcinoma	adherent	epithelial	ATCC (HTB-22™)
U-2 OS	CVCL_0042	Bone	Osteosarcoma	adherent	epithelial	ATCC (HTB-96™)
HepG2	CVCL_0027	Liver	Hepatoblastoma	adherent	epithelial	ATCC (HB-8065™)
Daudi	CVCL_0008	Peripheral Blood (B lymphoblast)	Burkitt's Lymphoma	suspension	lymphoblast	ATCC (CCL-213™)
CCD-18Co	CVCL_2379	Colon	none	adherent	fibroblast	ATCC (CRL-1459™)
NCI-H1092	CVCL_1454	Lung	Small cell lung cancer (stage E carcinoma)	suspension	n/a	ATCC (CRL-5855™)

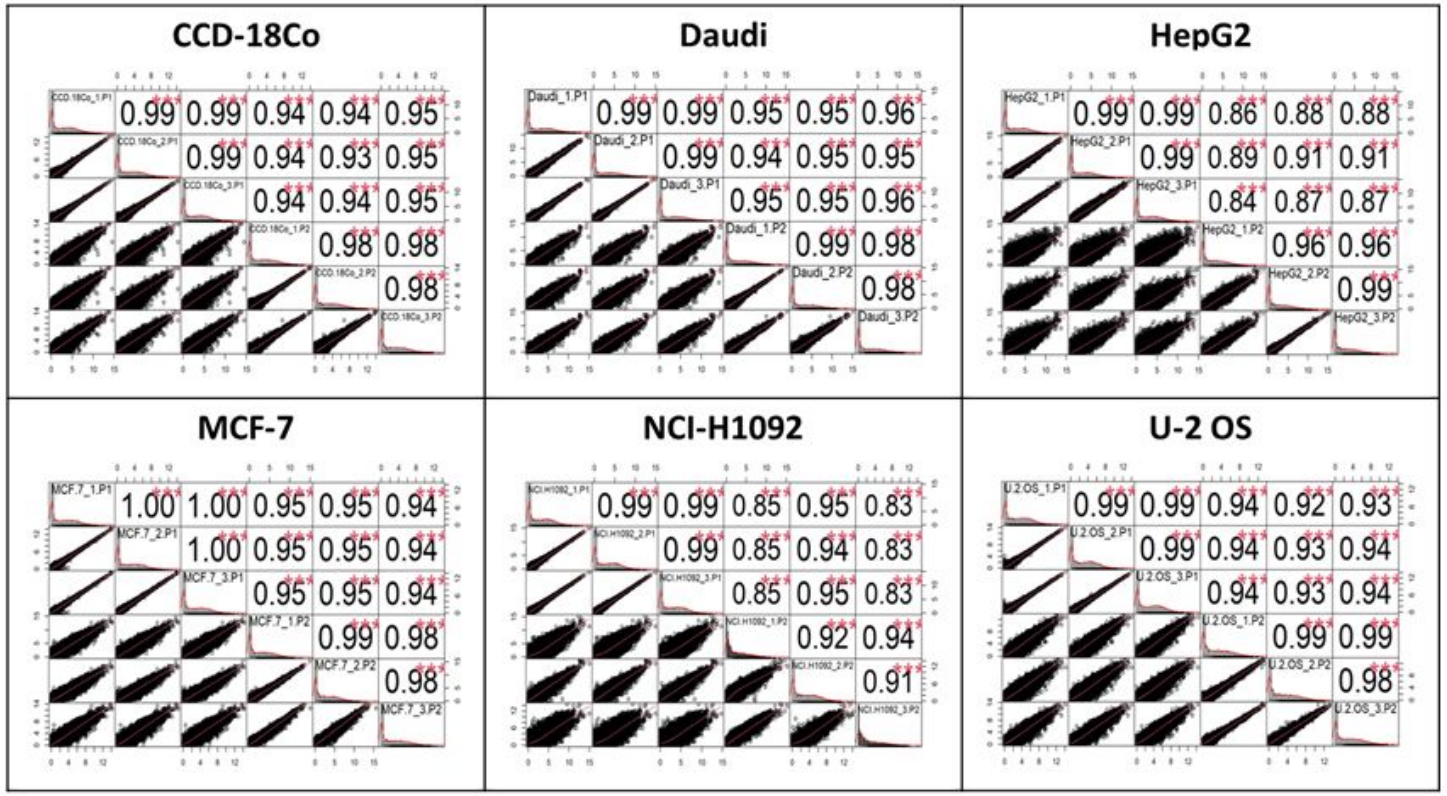
Pearson correlations for TempO-seq Phase 1 and Phase 2 show strong reproducibility

The average across technical replicates was 0.98 (95% CI: 0.97–0.99) when averaged across both Phase 1 and Phase 2.
When comparing the technical replicate data across the two TempO-seq phases, the average was 0.93 (95% CI: 0.90–0.96).

a) Correlations: Layout Example



b) Correlations: All cell types

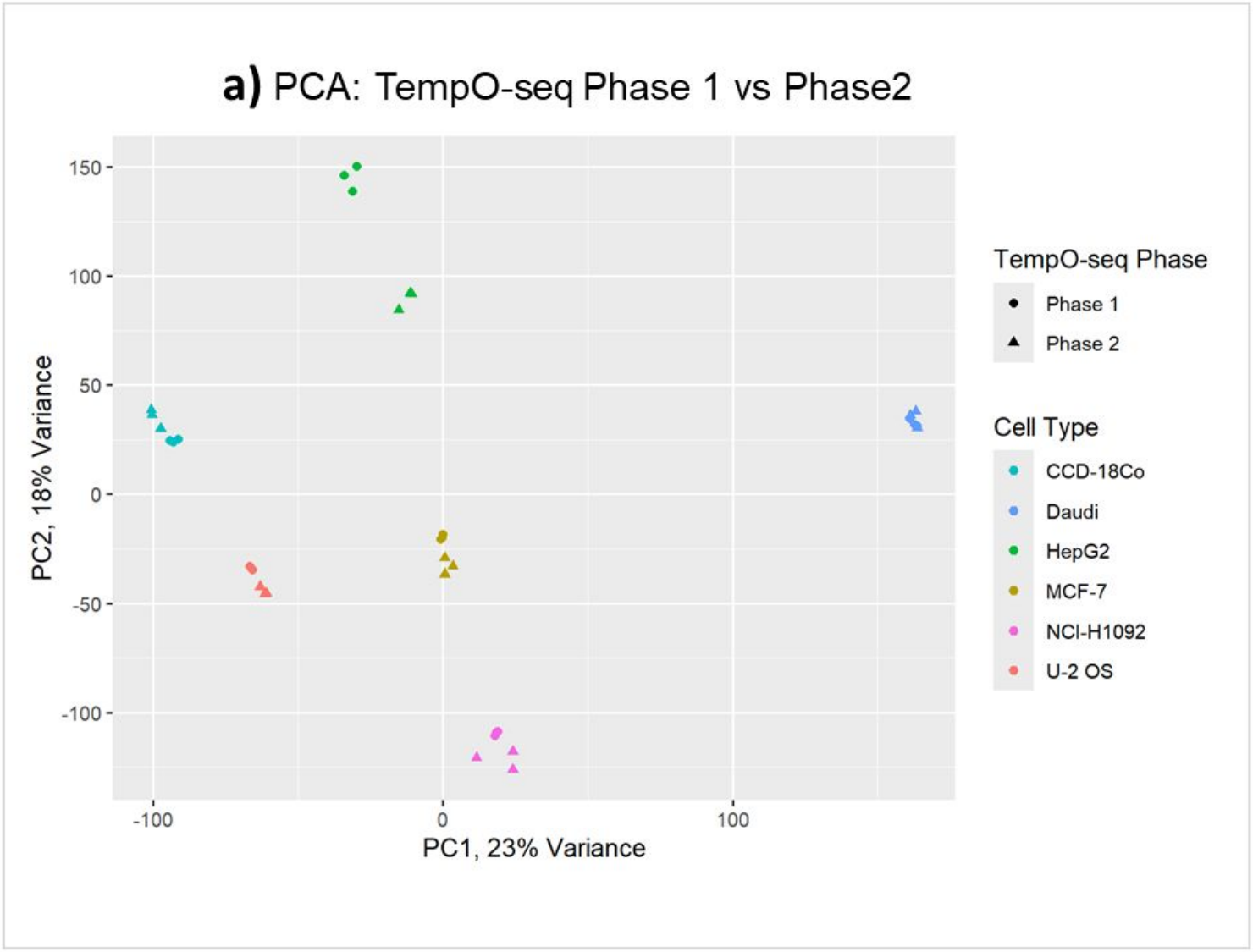


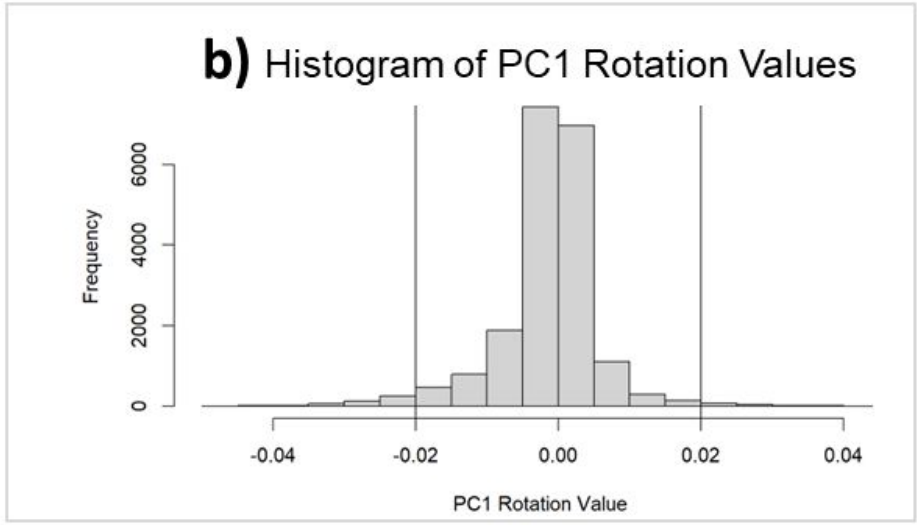
Principal Component Analysis (PCA)

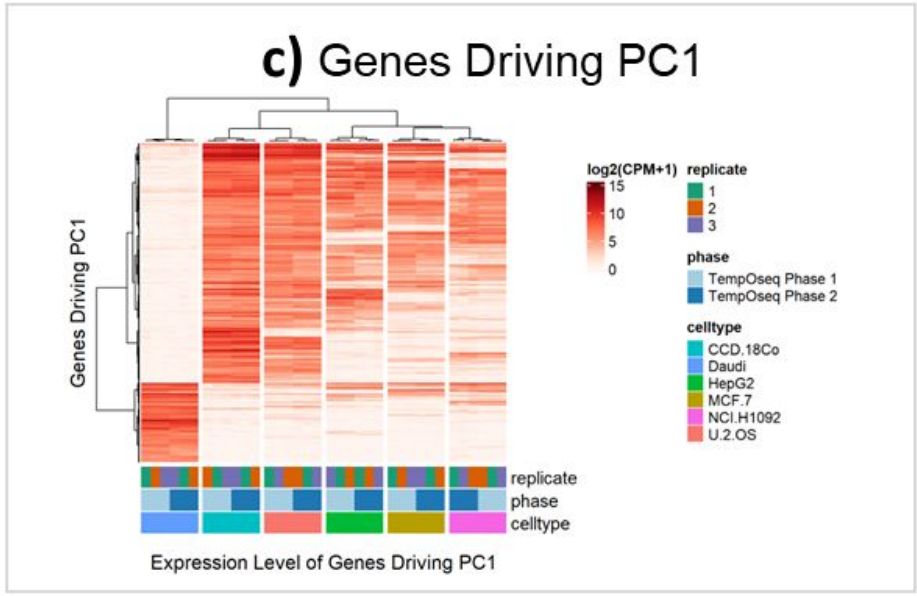
PCA is an unsupervised dimensionality reduction method for visualizing patterns in data

Principal Component Analysis (PCA)

PCA shows that the replicate data from the two TempO-seq data sets group well by cell line

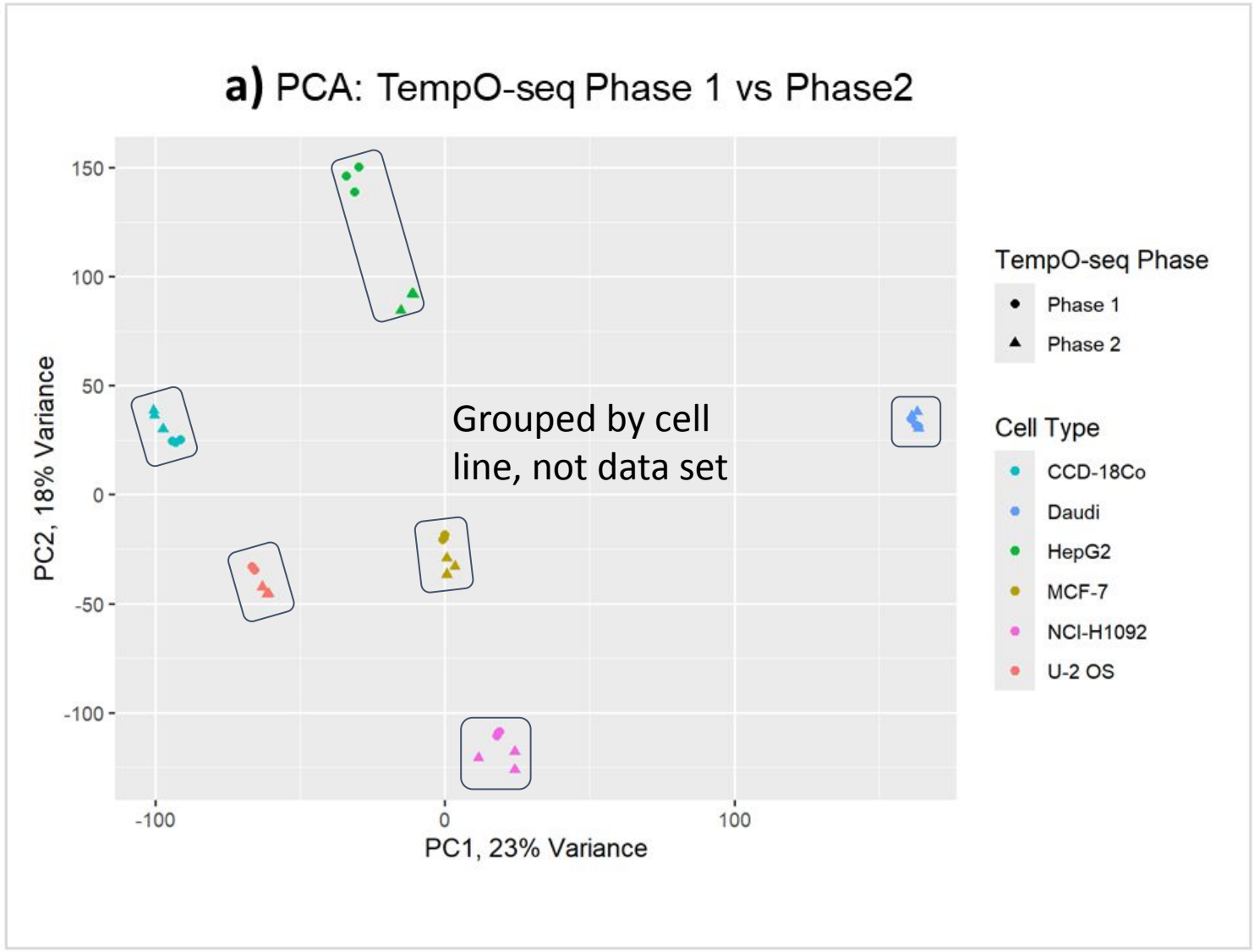


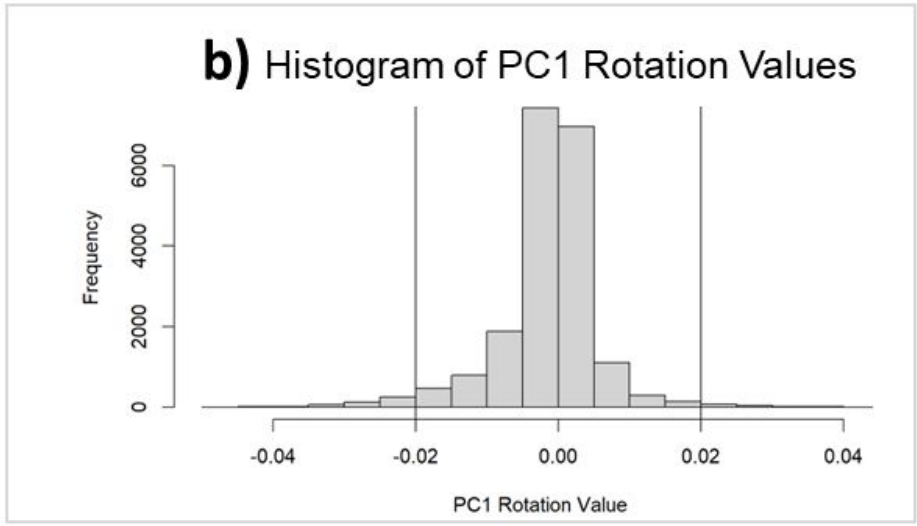


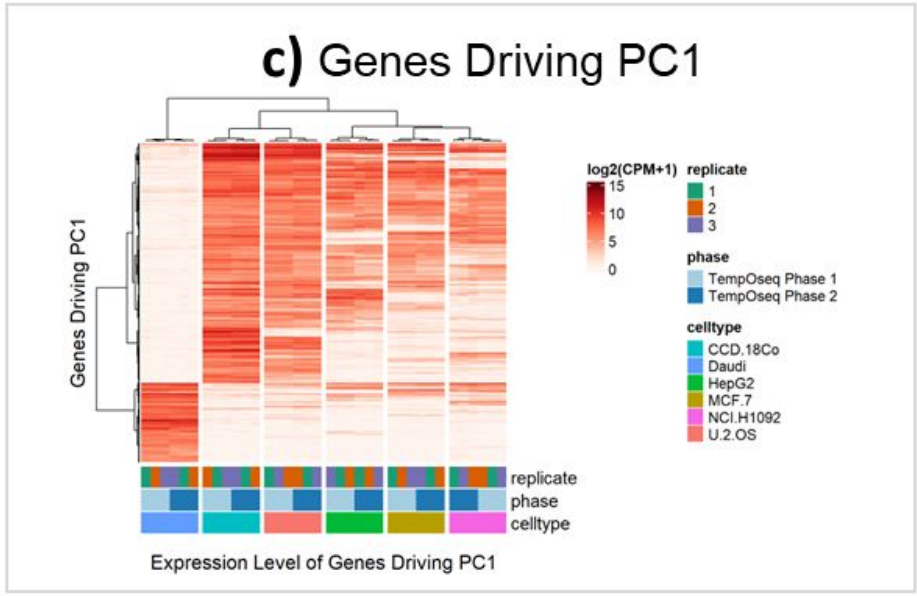


Principal Component Analysis (PCA)

PCA shows that the replicate data from the two TempO-seq data sets group well by cell type

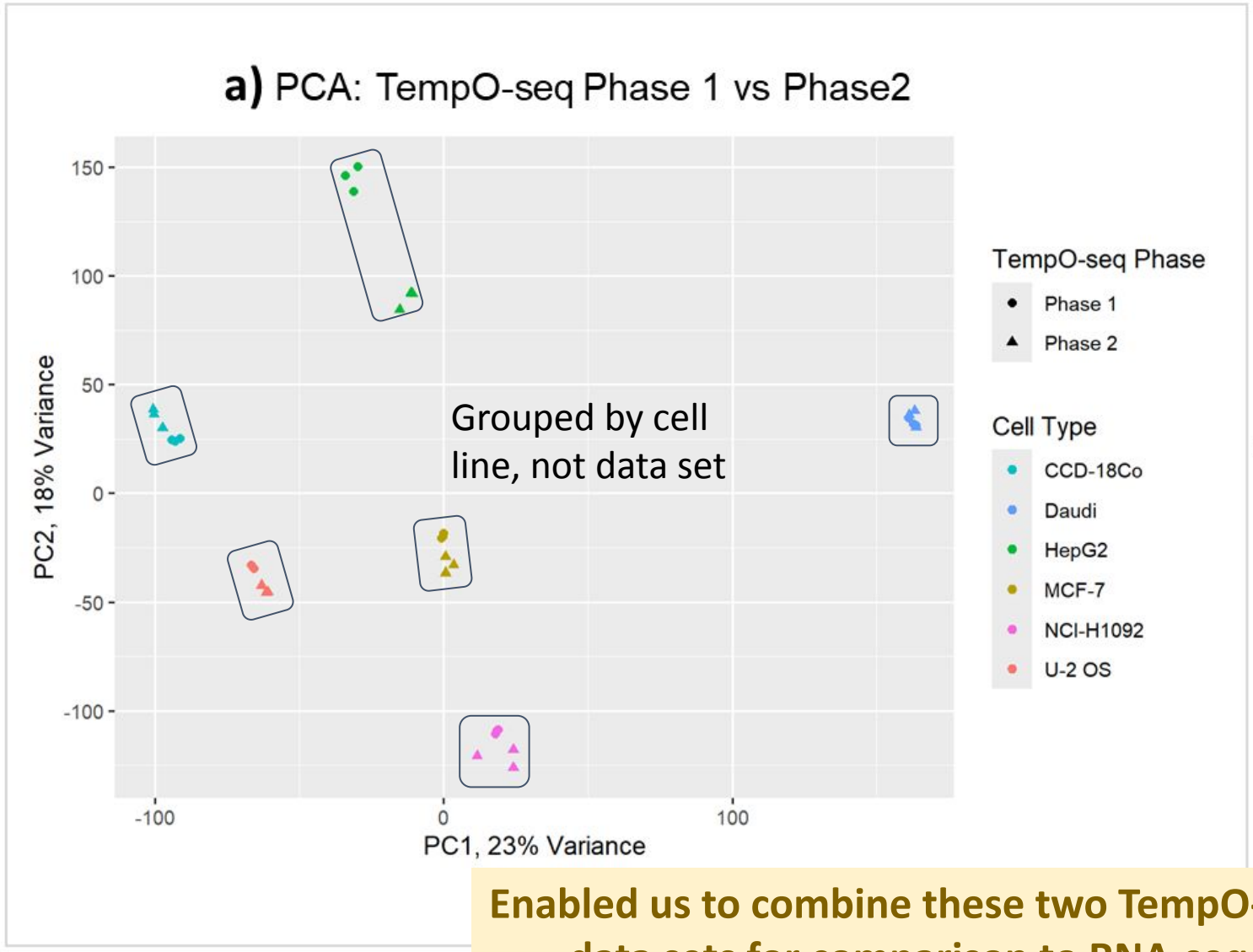


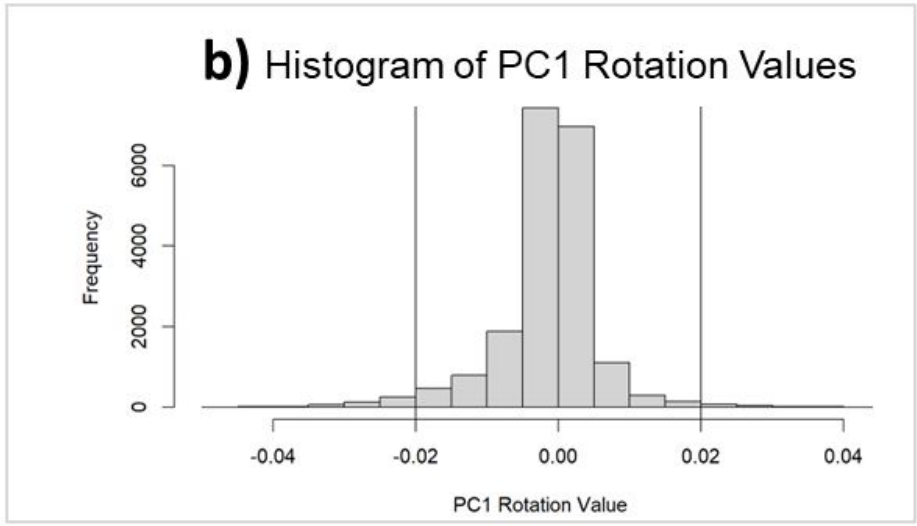


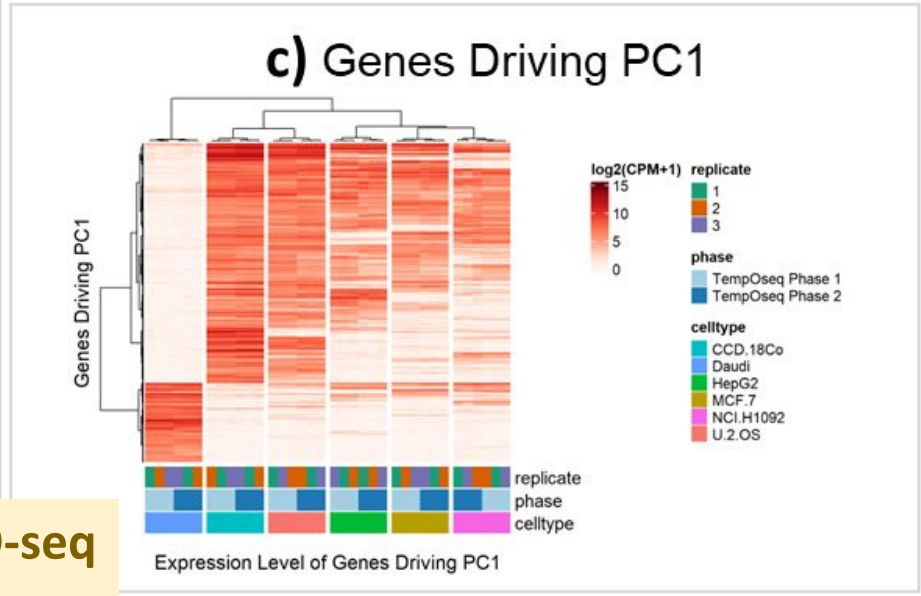


Principal Component Analysis (PCA)

PCA shows that the replicate data from the two TempO-seq data sets group well by cell type







Enabled us to combine these two TempO-seq data sets for comparison to RNA-seq

Step 2. Compare the Combined TempO-seq Data to RNA-seq

39 Cell lines were compared for TempO-seq vs RNA-seq

11 Tissue types:
Lung, blood, liver, kidney, breast, bone, eye, vascular endothelium, skin, brain, adipose

- ***This represents a significant expansion upon previous studies, covering 4 tissue types****: blood, breast, liver, and prostate cancer*

Data Comparison	TempO-seq Phase	Cell Line	Tissue Origin	Disease or Cell Line	Growth Mode
TvR	1	A549	Lung	Carcinoma	Adherent
TvR	1	A704	Kidney	Renal Cell Carcinoma	Adherent
TvR	1	ASC52Telo	Adipose Tissue	Mesenchymal Stem Cell	Adherent
TvR	1	BHY	Upper Aerodigestive Tract	Oral Squamous Cell carcinoma	Adherent
TvR	2	BT-483	Breast	Ductal Carcinoma	Adherent
TvR	2	CAL-148	Breast	Ductal Adenocarcinoma	Mixed
TvR	2	CAL-78	Muscle	Chondrosarcoma	Adherent
TvT, TVR	1, 2	CCD-18Co	Colon	None (Fibroblast)	Adherent
TvT, TvR	1, 2	Daudi	Lymphoid	Burkitt's Lymphoma	Suspension
TvR	1	DMS 454	Lung	Small Cell Lung Carcinoma	Adherent
TvR	2	DoHH2	Lymphoid	B Cell Lymphoma	Suspension
TvR	1	DV-90	Lung	Adenocarcinoma	Adherent
TvR	2	EFM-19	Breast	Ductal Carcinoma	Adherent
TvR	1	HBEC3-KT	Lung	Bronchial Epithelia	Adherent
TvT, TvR	1, 2	HepG2	Liver	Hepatoblastoma	Adherent
TvR	2	HOS	Bone	Osteosarcoma	Adherent
TvR	2	Hs.839.T	Skin	Melanoma	Adherent
TvR	1	hTERT-HME1	Breast	Breast Epithelium	Adherent
TvR	1	hTERT-RPE1	Eye	Pigmented Epithelium	Adherent
TvR	2	Huh-1	Liver	Hepatoma	Adherent
TvR	2	Huh-7	Liver	Hepatoblastoma	Adherent
TvR	1	HUVEC/TERT2	Umbilical Cord	Vascular Endothelium	Adherent
TvR	1	KP-N-RT-BM-1	Central Nervous System	Neuroblastoma	Adherent
TvT, TvR	1, 2	MCF7	Breast	Adenocarcinoma	Adherent
TvR	2	MG-63	Bone	Osteosarcoma	Adherent
TvR	2	MHH-CALL-4	Lymphoid	B Cell Lymphoma	Suspension
TvT, TvR	1, 2	NCI-H1092	Lung	Small cell lung cancer (stage E carcinoma)	Suspension
TvR	2	NCI-H1105	Lung	Small Cell Lung Cancer	Suspension
TvR	2	NCI-H1436	Lung	Small Cell Lung Cancer	Suspension
TvR	2	NCI-H2106	Lung	Non-small Cell Lung Cancer	Suspension
TvR	2	NCI-H2171	Lung	Small Cell Lung Cancer	Suspension
TvR	2	PLC/PRF/5	Liver	Hepatoma	Adherent
TvR	1	RPTEC/TERT1	Kidney	Proximal Tubule Epithelium	Adherent
TvR	2	SaOS-2	Bone	Osteosarcoma	Adherent
TvR	1	SET-2	Lymphoid	Acute Megakaryoblastic Leukemia	Suspension
TvR	1	SK-MEL-28	Skin	Melanoma	Adherent
TvR	2	SU-DHL-6	Lymphoid	Large / B Cell Lymphoma	Suspension
TvR	2	T-47D	Breast	Ductal Carcinoma	Adherent
TvR	1	TIME	Skin	Dermal Microvascular Endothelium	Adherent
TvT, TvR	1, 2	U-2 OS	Bone	Osteosarcoma	Adherent

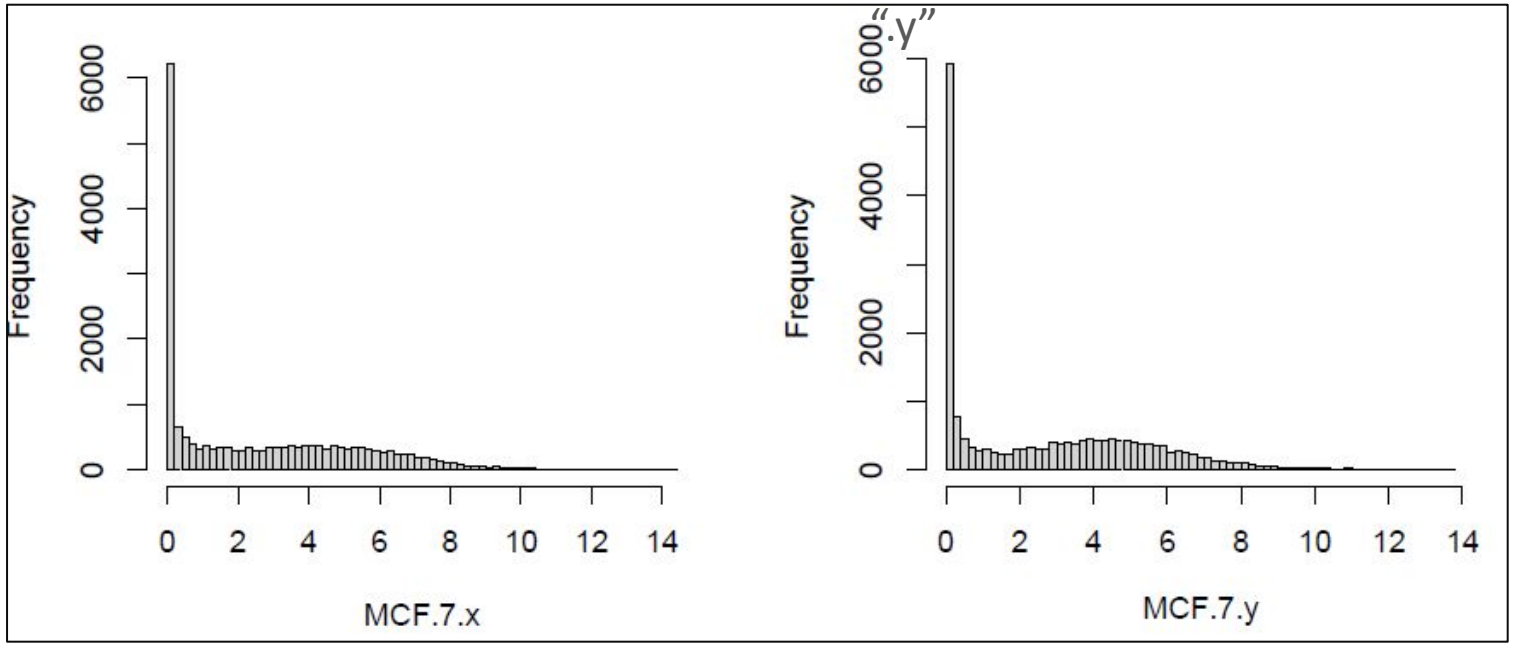
Understanding the data distributions

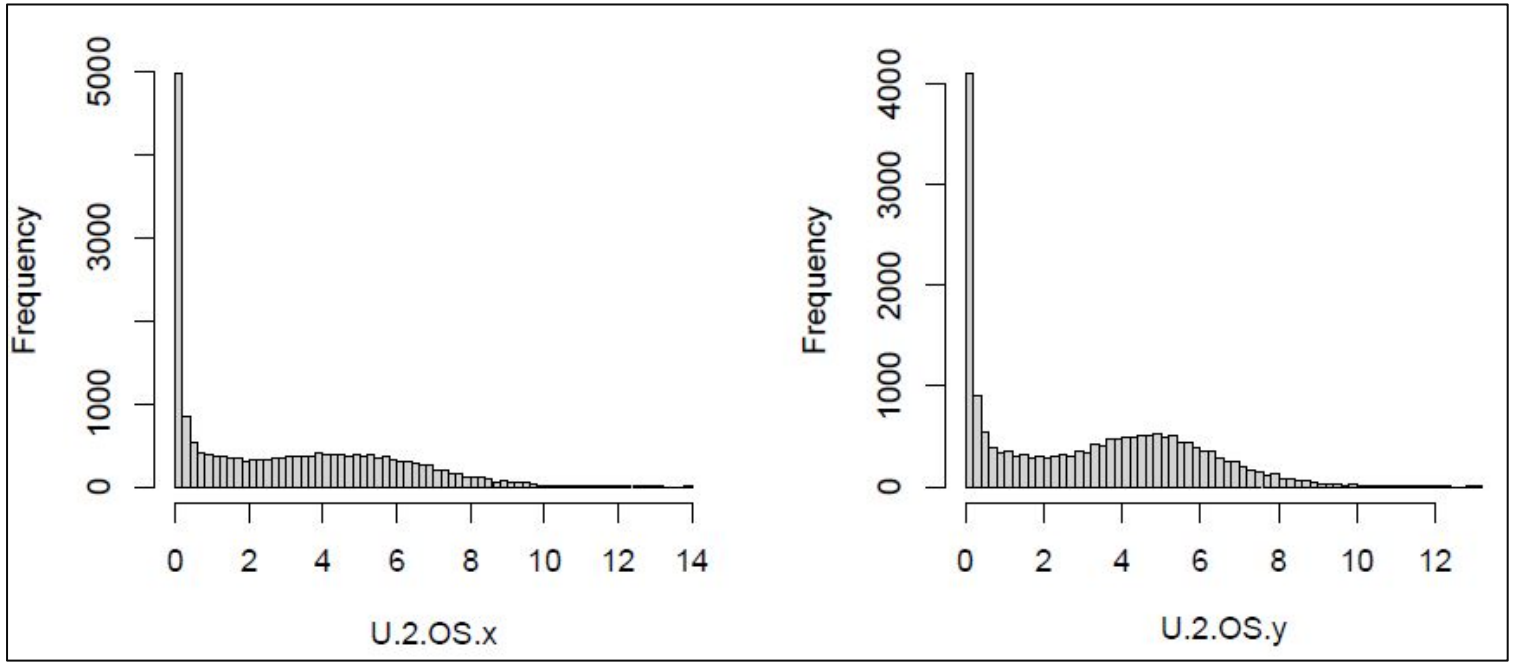
Histograms for TempO-seq data (left) vs RNA-seq data (right)

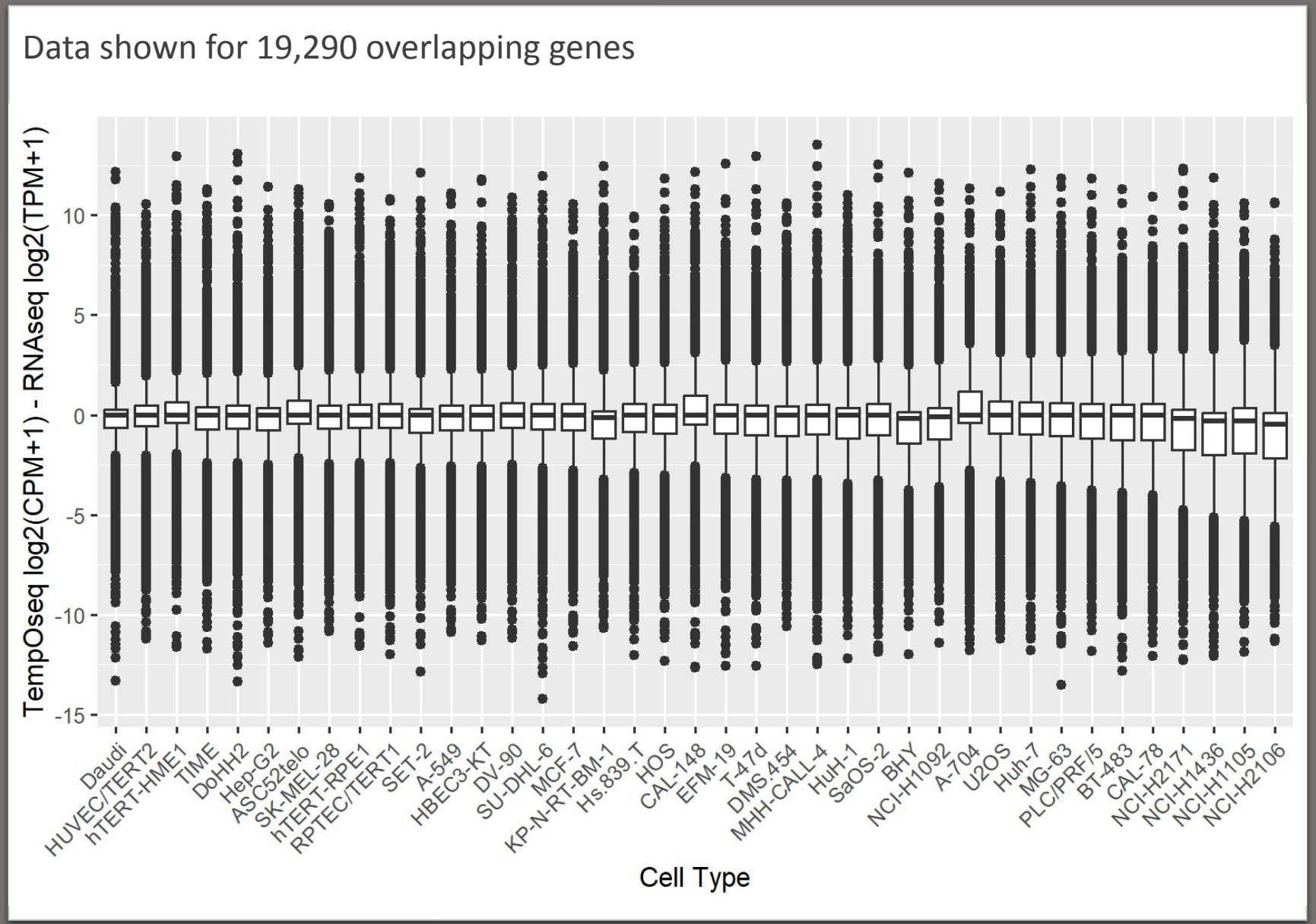
Showing two cell types of interest

TempO-seq data; log2(CPM+1); ".x"

RNA-seq data; log2(TPM+1); ".y"

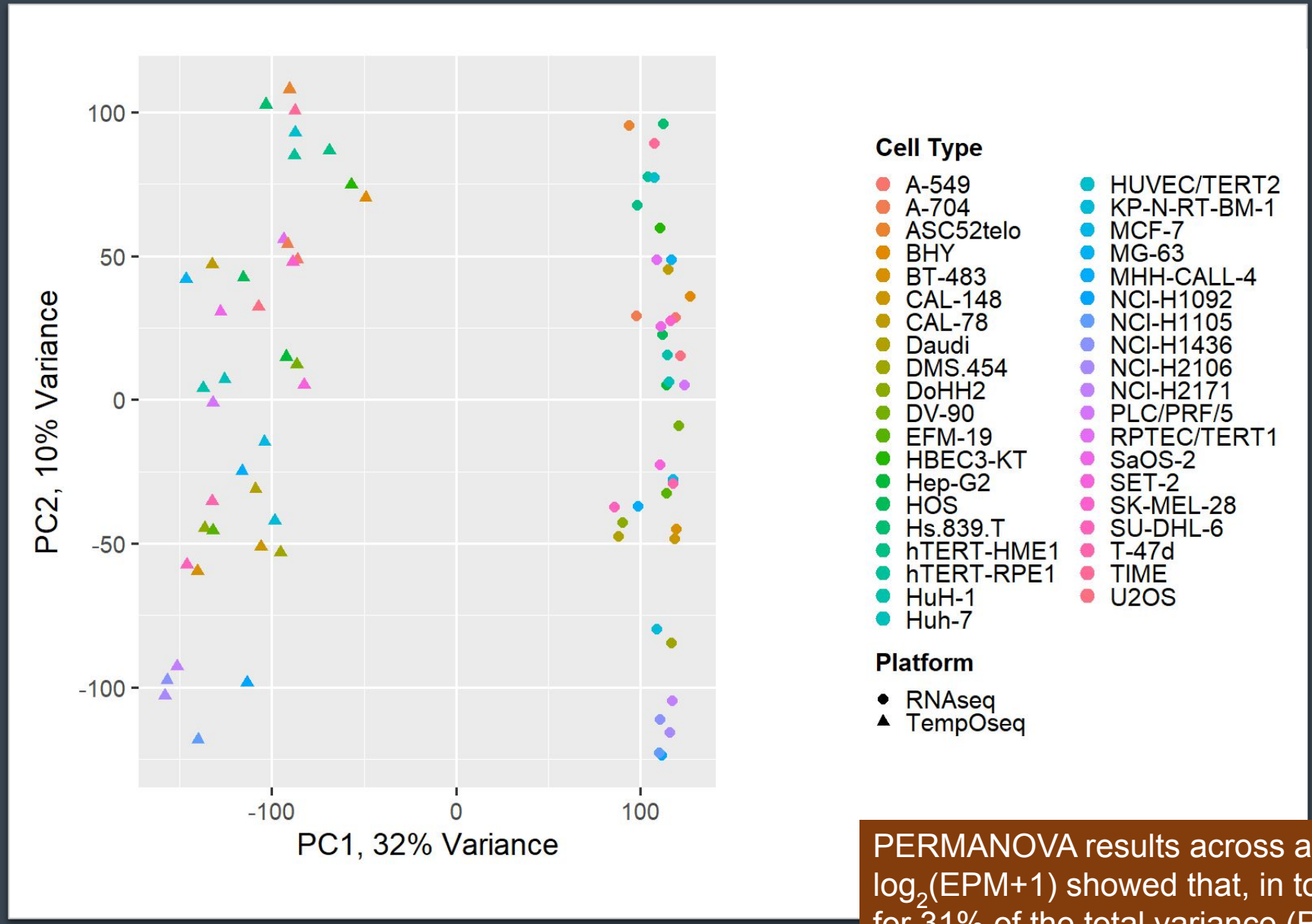






TempO-seq minus RNA-seq log2 data is centered around zero across all 39 cell types

Note: Counts Per Million (CPM) and Transcripts Per Million (TPM) were deemed comparable and will be referred to collectively as Expression Per Million (EPM)



PCA for TempO-seq vs RNA-seq shows a clear platform divergence

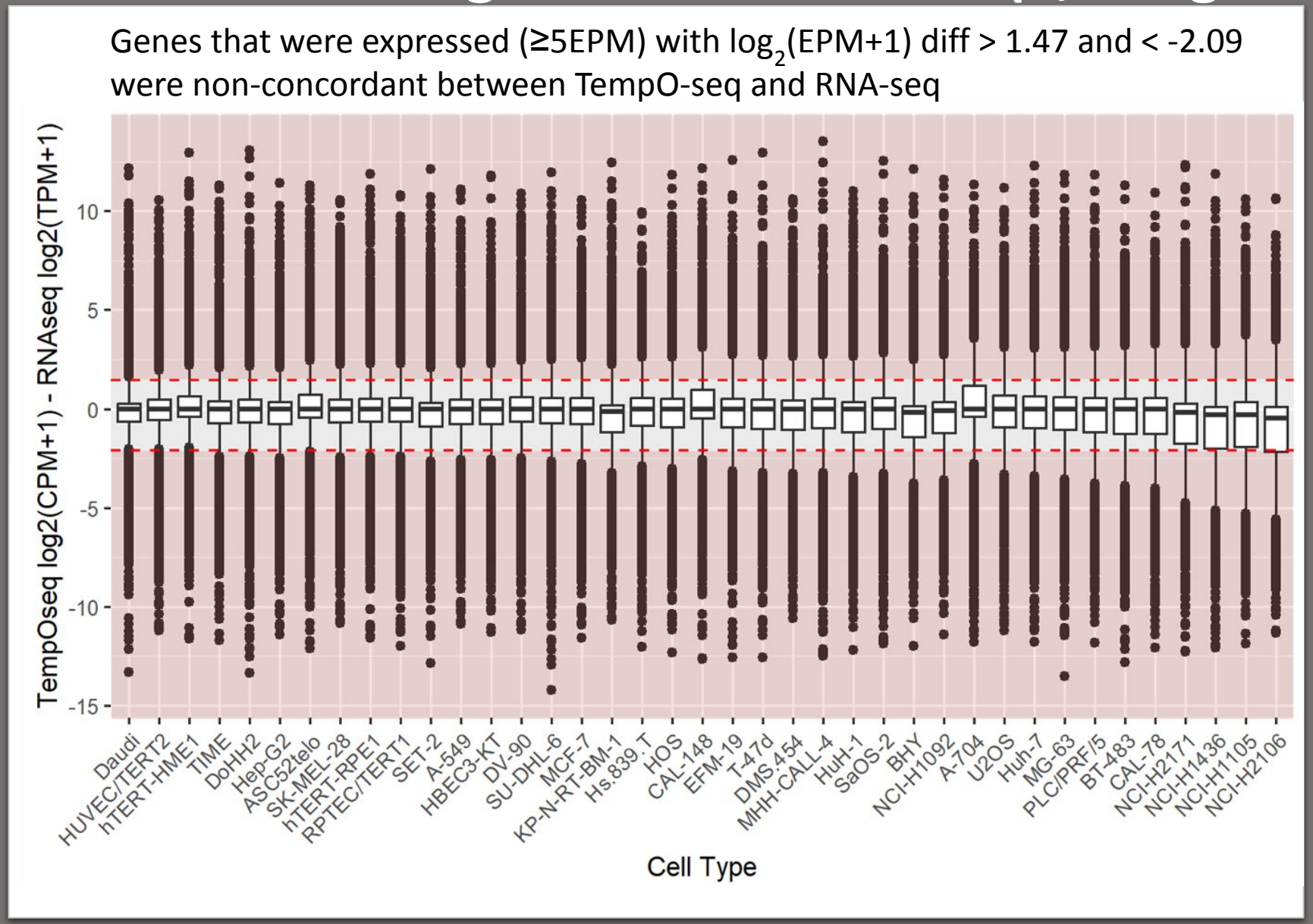
PERMANOVA results across all PCs for TempO-seq vs RNA-seq $\log_2$(EPM+1) showed that, in total, the platform effect accounted for 31% of the total variance (R^2 = 0.31, p = 0.001).

Which genes are non-concordant and are driving the platform divergence?

Which genes are non-concordant and are driving the platform divergence?

Genes with the greatest difference in log2 expression levels between TempO-seq and RNA-seq were progressively removed until the PERMANOVA variance explained (R^2) for platform effect across all PCs was < 10%

Non-concordant genes shown in red (3,810 genes of 19,290 genes)

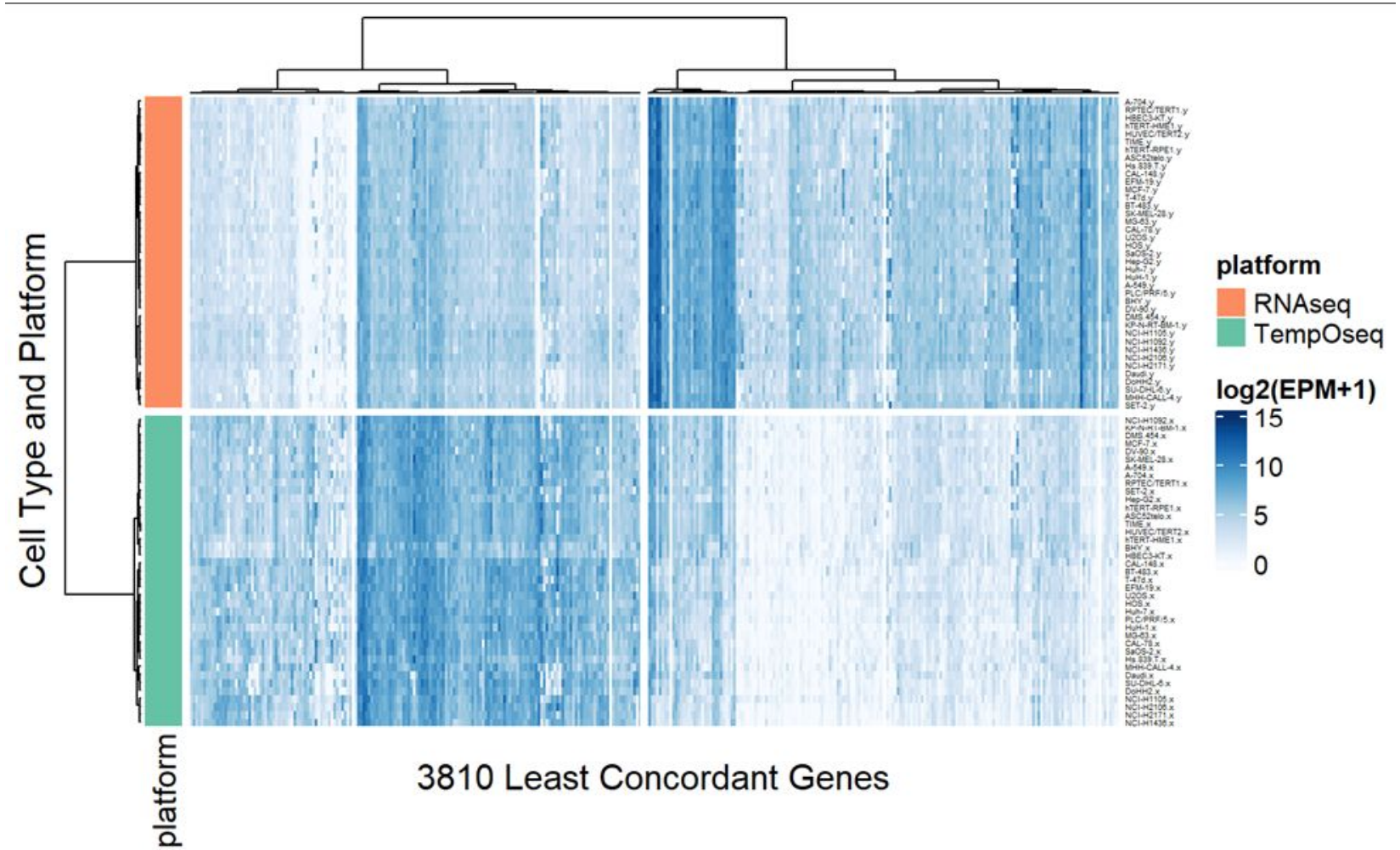


87th percentile (1.47)

13th percentile (-2.09)

After removal of the 3,810 most non-concordant genes, PERMANOVA on the PCs for the remaining 15,480 concordant genes had < 10% variance explained by platform divergence (R^2 = 0.099, p = 0.001)

The 3,810 Non-concordant genes had clear differences in expression level that were consistent across cell types



*Scientific analyses involving these genes should be interpreted with caution

Gene Ontology (GO) Analysis:
Evaluating patterns among non-concordant genes using MSigDB signatures

- **Assessed expression genes:** Required a minimum expression of ≥5 CPM in TempO-seq or ≥5 TPM in RNA-seq (10,487 genes). Of those, there were 3,810 genes that were non-concordant and 6,677 genes were concordant.
- **GO signature requirements:** We required at least 10 genes from the GO signature to be within the list of 10,487 expressed genes. We also required at least 50% of the genes within the GO signature to be in the list of 10,487 expressed genes that were retained for analysis.
    - This resulted in 3,935 GO signatures being retained in our analysis out of the full list of 10,461 GOs from Molecular Signatures Database Human Collections (MSigDB).
- **Odds ratios:** Odds of a GO signature being enriched with more non-concordant genes were calculated.

Example of GO filtering step

signature (sig)	signature_genes	sig_genecount_all	sig_genes_inlists	sig_genes_notinlists	percent_sig.genes_withinlists
GOBP_10_FORMYLTETRAHYDROFOLATE_METABOLIC_PROCESS	AASDHPPT, ALDH1L1, ALDH1L2, MTHFD1, MTHFD1L, MTHFD2L	6	5	1	83%
GOBP_3_PHOSPHOADENOSINE_5_PHOSPHOSULFATE_METABOLIC_PROCESS	ABHD14B, BPNT1, ENPP1, PAPSS1, PAPSS2, SULT1A1, SULT1A2, SULT1A3, SULT1A4, SULT1B1, SULT1C3, SULT1C4, SULT1E1, SULT2A1, SULT2B1, TPST1, TPST2	17	8	9	47%
GOBP_ACETATE_ESTER_METABOLIC_PROCESS	ACHE, BCHE, CHAT, COLQ, SLC44A4, SLC5A7	6	0	6	0%
GOBP_2FE_2S_CLUSTER_ASSEMBLY	BOLA2, BOLA2B, FDX2, FXN, GLRX3, GLRX5, HSCB, ISCU, LYRM4, NDUFAB1, NFS1	11	11	0	100%
GOBP_2_OXOGLUTARATE_METABOLIC_PROCESS	AADAT, ADHFE1, D2HGDH, DLST, GOT1, GOT2, GPT2, IDH1, IDH2, KYAT3, L2HGDH, MRPS36, OGDH, OGDHL, PHYH, TAT	16	13	3	81%

Gene ontology (GO) odds ratio (OR) calculations

GO signatures with odds ratios (ORs) > 1 had greater odds of non-concordant levels of expression between TempO-seq and RNA-seq for the genes within the signature.

$$OR = \frac{a/b}{c/d}$$

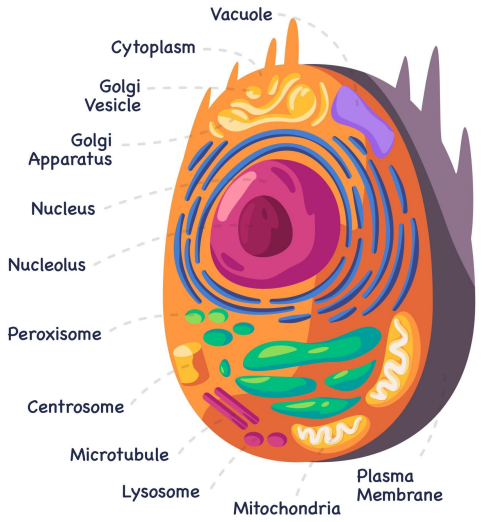
	Within GO signature	Not within GO signature	Totals
Non-concordant Genes with ≥ 5 EPM	a	b	3,810 genes
Concordant Genes with ≥ 5 EPM	c	d	6,677 genes
	(a+c)	(b+d)	(a+c)+(b+d) = 10,487 genes

Gene ontologies (GOs) relating to chromatin and ribosomes were the least concordant (OR > 1)

Gene Ontology Term from MSigDB (molecular signatures database)	Genes (n)	Genes in analysis	% Genes in analysis	OR	1/OR	FDR p-value
GOBP_PROTEIN_LOCALIZATION_TO_CENP_A_CONTAINING_CHROMATIN	18	17	94	28.15	-	5.6E-04
GOCC_CHROMOSOME_CENTROMERIC_CORE_DOMAIN	19	18	95	14.07	-	3.1E-03
GOMF_STRUCTURAL_CONSTITUENT_OF_CHROMATIN	97	67	69	10.13	-	2.2E-12
GOBP_NEGATIVE_REGULATION_OF_MEGAKARYOCYTE_DIFFERENTIATION	20	17	85	8.20	-	4.7E-02
GOCC_CYTOSOLIC_LARGE_RIBOSOMAL_SUBUNIT	60	55	92	6.34	-	4.7E-07
GOCC_CYTOSOLIC_SMALL_RIBOSOMAL_SUBUNIT	41	36	88	4.00	-	3.1E-02
GOCC_NUCLEOSOME	134	97	72	3.78	-	4.8E-07
GOCC_CYTOSOLIC_RIBOSOME	118	107	91	3.50	-	4.8E-07
GOBP_NUCLEOSOME_ORGANIZATION	138	105	76	3.40	-	1.2E-06
GOMF_STRUCTURAL_CONSTITUENT_OF_RIBOSOME	169	153	91	3.00	-	1.4E-07
GOCC_LARGE_RIBOSOMAL_SUBUNIT	117	111	95	2.80	-	1.2E-04
GOCC_RIBOSOMAL_SUBUNIT	188	177	94	2.66	-	4.5E-07
GOBP_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS	76	73	96	2.53	-	4.4E-02
GOCC_CATALYTIC_STEP_2_SPLICEOSOME	91	88	97	2.43	-	2.0E-02
GOBP_CYTOPLASMIC_TRANSLATION	156	146	94	2.41	-	1.7E-04
GOCC_PRERIBOSOME	109	105	96	2.18	-	3.6E-02
GOCC_RIBOSOME	239	215	90	2.17	-	2.6E-05
GOBP_PROTEIN_DNA_COMPLEX_ASSEMBLY	240	189	79	2.08	-	5.1E-04
GOMF_STRUCTURAL_MOLECULE_ACTIVITY	809	446	55	1.87	-	4.5E-07
GOCC_RIBONUCLEOPROTEIN_COMPLEX	1169	661	57	1.70	-	2.0E-07
GOBP_RIBOSOME_BIOGENESIS	325	308	95	1.67	-	6.2E-03
GOBP_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	502	447	89	1.62	-	5.6E-04

Gene ontologies relating to the cell structure were the most concordant (OR < 1)

Gene Ontology Term from MSigDB (molecular signatures database)	Genes (n)	Genes in analysis	% Genes in analysis	OR	1/OR	FDR
GOCC_GOLGI_APPARATUS	1634	1068	65	0.77	1.30	4.6E-02
GOBP_LYMPHOCYTE_ACTIVATION	796	405	51	0.65	1.53	4.4E-02
GOMF_PROTEIN_KINASE_ACTIVITY	577	382	66	0.60	1.66	6.2E-03
GOBP_REGULATION_OF_ANATOMICAL_STRUCTURE_MORPHOGENESIS	937	488	52	0.60	1.67	5.1E-04
GOCC_BASEMENT_MEMBRANE	90	49	54	0.15	6.46	4.4E-03



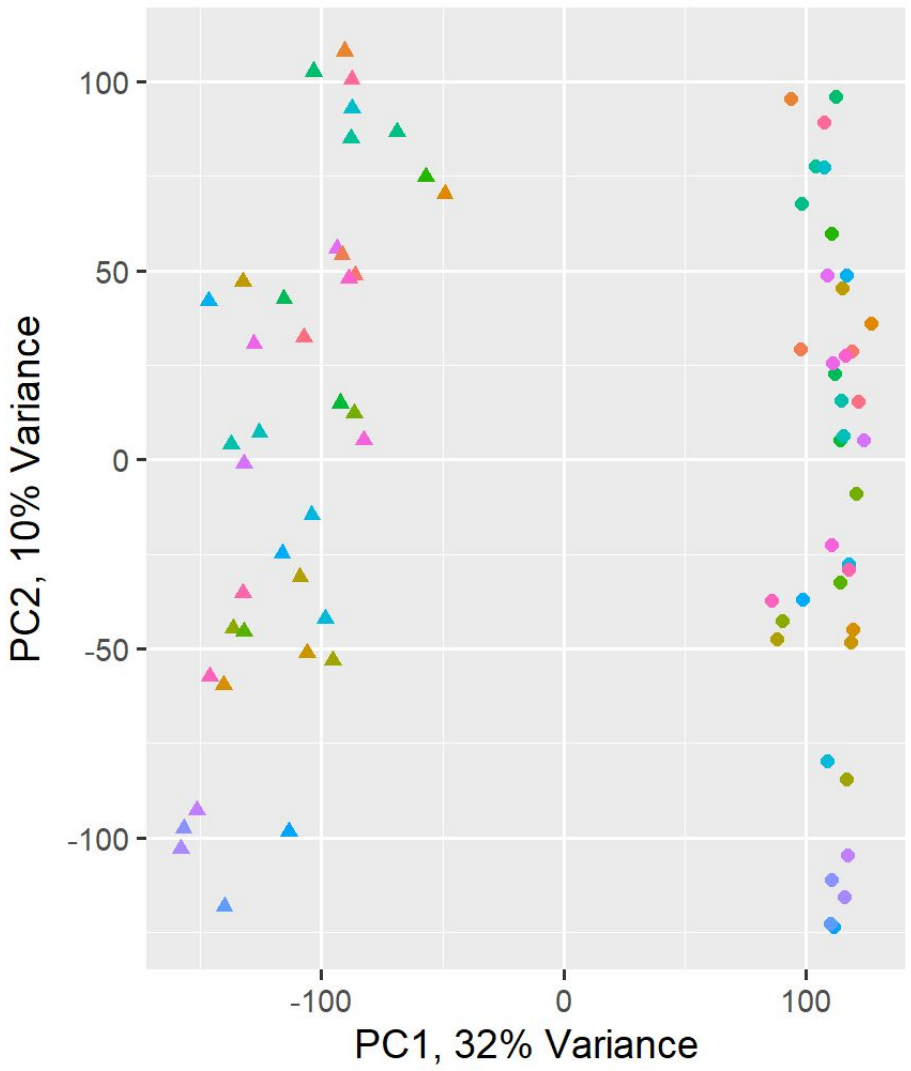
Non-concordant genes heavily featured histone and ribosomal gene families

Histone genes: 73% of all of the genes in the histone family were non-concordant

- Histone genes do not have poly-A tails
- RNA-seq preparation procedure included a poly-A tail pull-down step = had low TPM
- TempO-seq does not require poly-A tail pull-down = had high CPM
- This means that **TempO-seq may be preferable** to RNA-seq library preparations employing poly-A enrichment when interpreting expression levels for histone genes.

Ribosomal genes: more than half of the genes for ribosomal proteins were non-concordant

- TempO-seq probes were frequently not as efficient at detecting mRNA for ribosomal proteins for unclear reasons
  - One possible explanation is that the TempO-seq probe design for a subset of the ribosomal protein mRNA did not reliably capture expression for those specific genes.
- **RNA-seq may be the preferable** option when studying ribosomal protein genes.



Cell Type

- A-549
- A-704
- ASC52telo
- BHY
- BT-483
- CAL-148
- CAL-78
- Daudi
- DMS.454
- DoHH2
- DV-90
- EFM-19
- HBEC3-KT
- Hep-G2
- HOS
- Hs.839.T
- hTERT-HME1
- hTERT-RPE1
- HuH-1
- Huh-7
- HUVEC/TERT2
- KP-N-RT-BM-1
- MCF-7
- MG-63
- MHH-CALL-4
- NCI-H1092
- NCI-H1105
- NCI-H1436
- NCI-H2106
- NCI-H2171
- PLC/PRF/5
- RPTEC/TERT1
- SaOS-2
- SET-2
- SK-MEL-28
- SU-DHL-6
- T-47d
- TIME
- U2OS

Platform

- ● RNAseq
- ▲ TempOseq

Is there a good way to resolve the platform divergence?

Relative Log Expression (RLE)

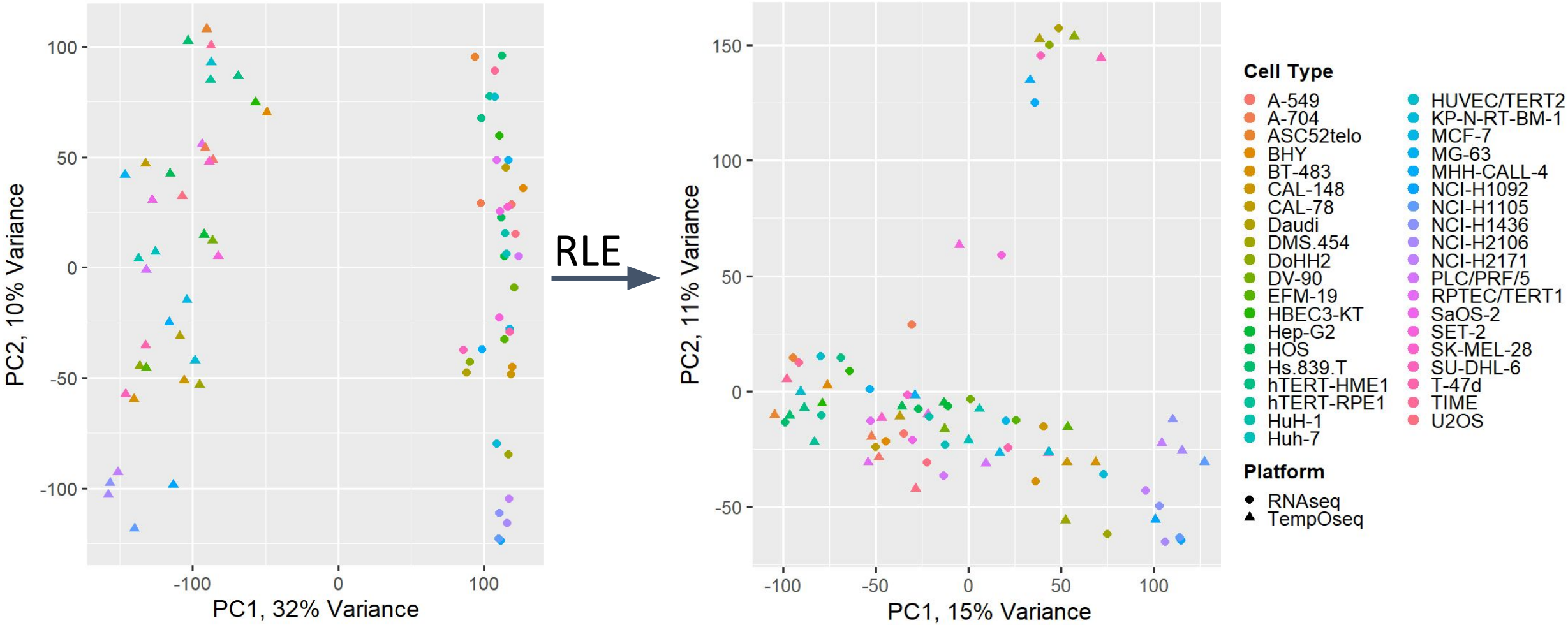
Method to calculate the log expression level relative to a reference value

Example:		
Sample = 8 $\log_2$CPM	Reference = 5 $\log_2$CPM	RLE = (8 - 5) = 3

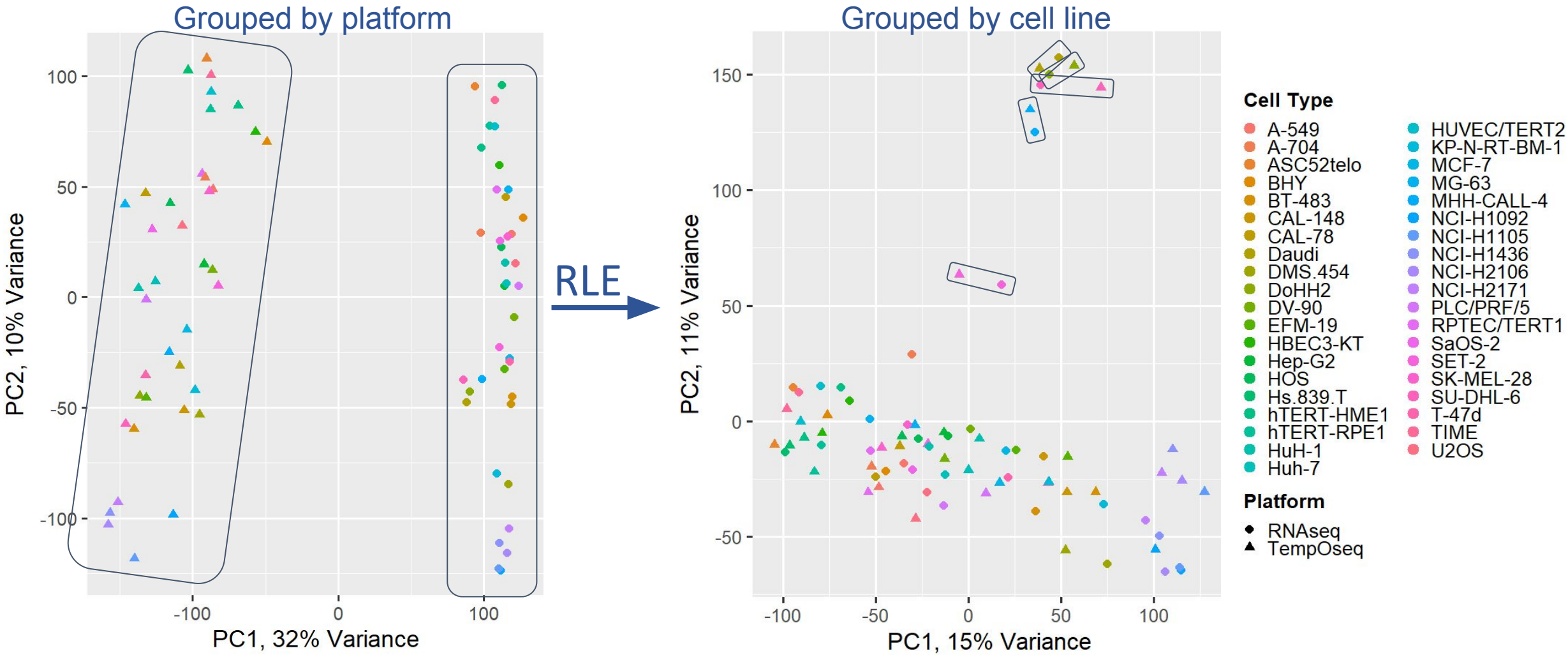
$$RLE\ for\ Gene\ X = \log_2\left(\frac{(EPM+1)\ for\ gene\ X\ within\ a\ single\ cell\ line}{Average\ (EPM+1)\ for\ gene\ X\ across\ all\ 39\ cell\ lines}\right)$$

= [$\log_2$ (EPM + 1) for gene X within a single cell line] − [Average ($\log_2$(EPM + 1) for gene X across all 39 cell lines)]

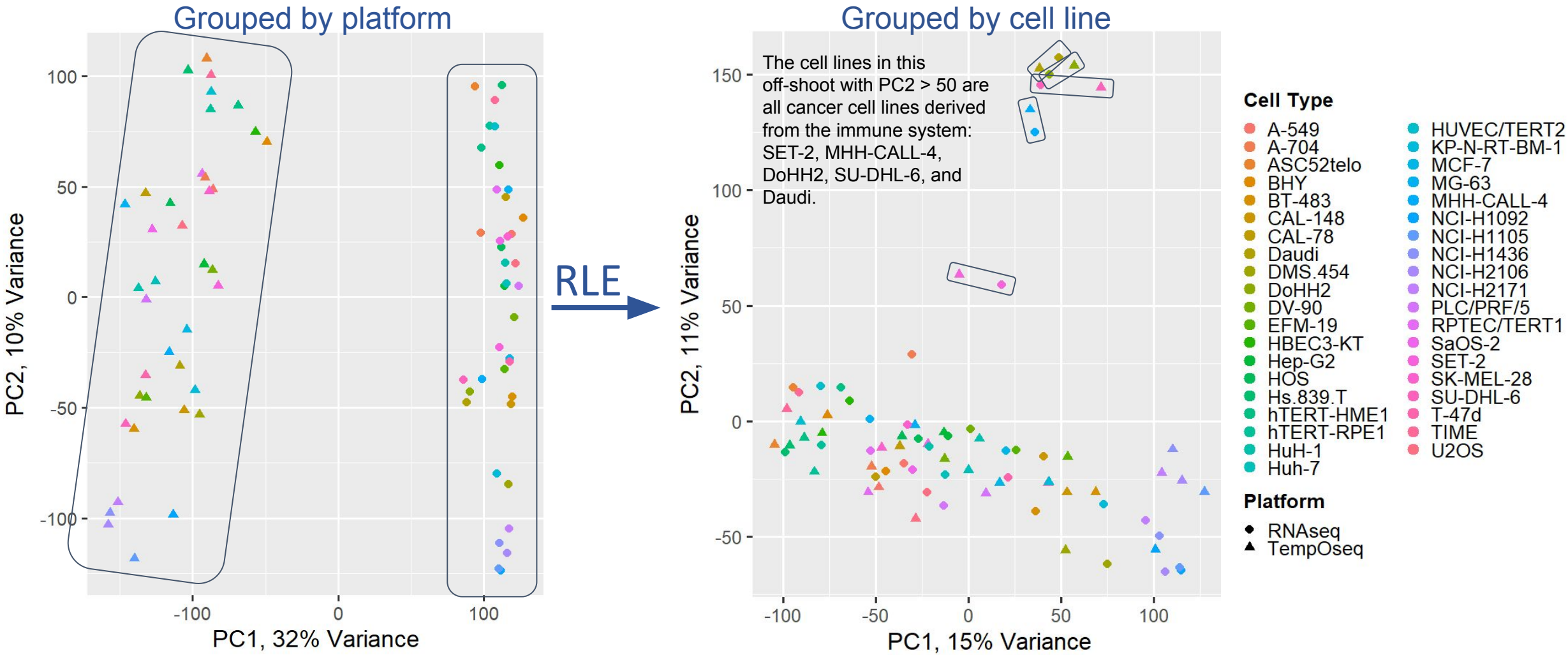
Calculated Relative Log Expression (RLE) for each cell line compared to the average across cell lines within each platform. This resolved the platform divergence without removing any genes.



Calculated Relative Log Expression (RLE) for each cell line compared to the average across cell lines within each platform. This resolved the platform divergence without removing any genes.

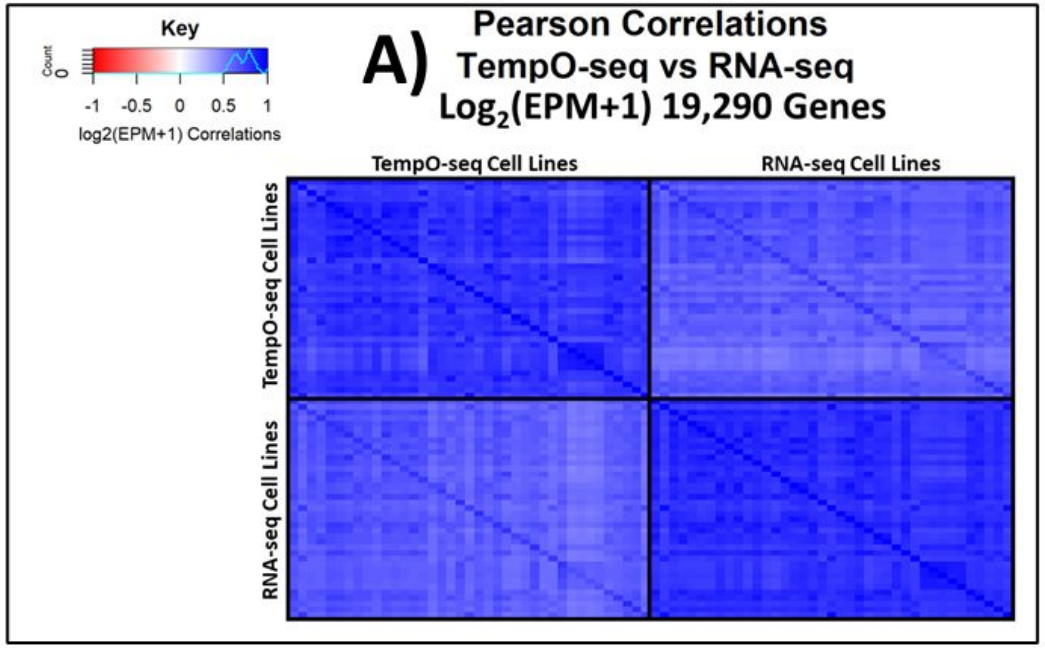


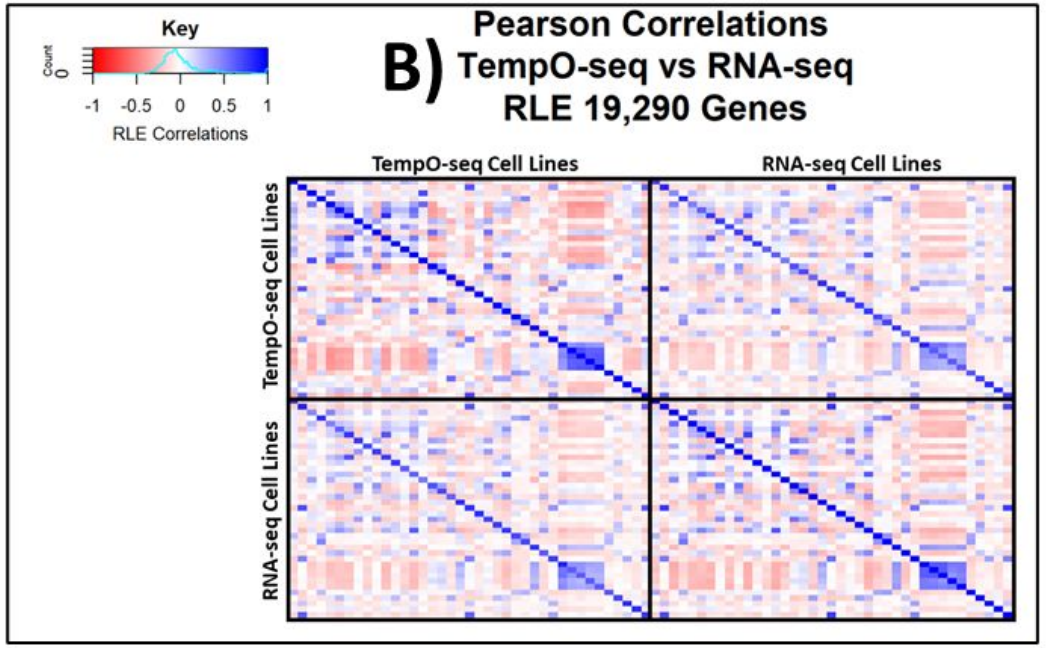
Calculated Relative Log Expression (RLE) for each cell line compared to the average across cell lines within each platform. This resolved the platform divergence without removing any genes.



Pearson correlations for TempO-seq vs RNA-seq show:

- **The correlation structure is preserved between TempO-seq and RNA-seq,** providing more weight of evidence suggesting the technologies give the same/similar response
- **RLE highlighted differences between cell lines,** maintaining good correlations between matching cell lines and bringing non-matching cell line correlations to nearly zero





Initial Pearson correlations:
Matching cell types: 0.77 (95% CI: 0.76 – 0.78)
Non-matching cell lines: 0.64 (95% CI: 0.64 – 0.65)

After RLE normalization:
Matching cell types: 0.71 (95% CI: 0.67 – 0.74)
Non-matching cell lines: -0.02 (95% CI: -0.03 – -0.01)

Summary of Baseline Gene Expression Comparison Findings

TempO-seq vs TempO-seq:

- TempO-seq was highly reproducible at different read depths (Pearson Correlations, PCA)

TempO-seq vs RNA-seq:

- 80% of genes for TempO-seq vs RNA-seq log2EPM data are comparable (PERMANOVA)
- The 20% of genes that were non-concordant related primarily to histone and ribosomal gene families (Gene Ontology)
- TempO-seq vs RNA-seq has a PC1 platform divergence that was able to be resolved using Relative Log Expression (RLE) normalization (PCA)
- RLE accentuates inter- and intra-platform differences in cell line gene expression patterns (Pearson correlations)

Study strengths and weaknesses

Strengths

- This comparison includes 39 cell lines for the full transcriptome and **showed consistently high correlations across all cell lines for TempO-seq vs RNA-seq**
- This is the **first study to compare cell lysates to purified RNA samples**
- **The 39 cell lines were from 11 different types of tissue**: lung, lymphatic (lymphoma), liver, kidney, breast, bone, eye, blood (leukemia), endothelium (microvascular), skin, adipose, and brain
  - ***This represents a significant expansion upon previous studies, covering 4 tissue types****: blood, breast, liver, and prostate cancer*

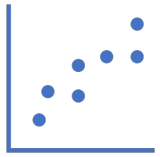
Weaknesses/Complications

- **Part of the differences could be due to the data being from different cell stocks and from being generated by different groups**
  - *However, the variation is within normally observed levels for transcriptomics data from different laboratories and provides proof of real-world replicability across labs*
- **Approximately 8,800 genes were not expressed at baseline** in any cell type in either platform, making it important for this analysis to be repeated with a chemical exposure dataset to try to induce the expression of those genes for comparison

Conclusions and Future Work

TempO-seq is highly reproducible at different read depths, and shows consistent gene expression findings as RNA-seq

After normalization, the data grouped by cell type and not by technology platform in PCA



This work can help increase confidence in using TempO-seq data and/or for using RLE normalization to combine with RNA-seq data

This helps to validate TempO-seq against the RNA-seq gold-standard technique

Future work: TempO-seq from lysed cells and RNA-seq data need to be compared from the same cell stocks and after inducing more gene expression

This work using baseline expression data is a good foundation for such work

Many thanks to the co-authors and HTTr team for their input on these methods!

Special thanks to:

- Joshua Harrill
- Logan Everett
- Clinton Willis
- Sarah Davidson-Fritz
- Richard Judson
- Woody Setzer
- Imran Shah
- Joseph Bundy
- Jesse Rogers
- Bryant Chambers
- Nisha Sipes

Manuscript Summary Hyperlink:

Gene expression technologies TempO-seq and RNA-seq are largely concordant

Citation:

Word LJ, Willis CM, Judson RS, Everett LJ, Davidson-Fritz SE, Haggard DE, Chambers BA, Rogers JD, Bundy JL, Shah I, Sipes NS, Harrill JA. TempO-seq and RNA-seq gene expression levels are highly correlated for most genes: A comparison using 39 human cell lines. PLoS One. 2025 May 9;20(5):e0320862. doi: 10.1371/journal.pone.0320862. PMID: 40344165; PMCID: PMC12064016.

TempO-seq and RNA-seq Gene Expression Levels are Highly Correlated for Most Genes: A Comparison Using 39 Human Cell Lines

Laura Word*, Clinton Willis, Richard Judson, Logan Everett, Sarah Davidson-Fritz, Derik Haggard, Bryant Chambers, Jesse Rogers, Joseph Bundy, Imran Shah, Nisha Sipes, Joshua Harrill

United States Environmental Protection Agency, Office Of Research And Development, Center For Computational Toxicology And Exposure, Research Triangle Park, NC 27709

Background

As transcriptomics data from new targeted sequencing platforms accumulates in the literature, it is important to evaluate their similarity to traditional whole transcriptome RNA-seq. The present study evaluated the comparability of one such targeted RNA-seq platform, TempO-seq, to traditional RNA-Seq using baseline gene expression profiles from human cell lines. In this study, TempO-Seq data was generated from cell lysates with no RNA purification while RNA-Seq data that was from purified RNA was downloaded from the Human Protein Atlas project. The current analysis used baseline expression and future work should repeat this comparison with chemical exposure data.

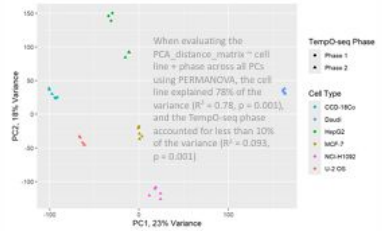
Conclusions: We found that ***normalized baseline gene expression TempO-seq data from lysed cells is reproducible and comparable to RNA-seq from purified samples for most genes,*** even when data were generated by different laboratories using different cell stocks

Methods and Results

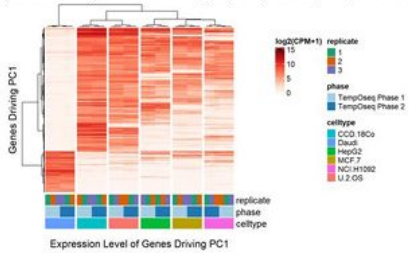
PART 1: TempO-seq data with different read depths on samples prepared months apart from the same cryostocks are highly reproducible

First, two TempO-seq data sets from the same set of six human cell lines that were generated several months apart and at different read depths were compared using principal component analysis (PCA). Phase 1 and Phase 2 data were sequenced to depths of 6 and 4.5 million reads, respectively. Average Pearson correlation was 0.93 (95% CI: 0.90 – 0.96).

F1) PCA: TempO-seq Phase 1 vs Phase 2. These two TempO-seq data sets were highly reproducible and suitable to combine.



F2) Genes driving the main variance grouped by cell line, not by data set.

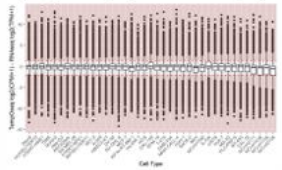


PART 2: TempO-seq and RNA-seq data is highly correlated: a platform divergence was observed, but it was readily resolved by calculating Relative Log Expression (RLE)

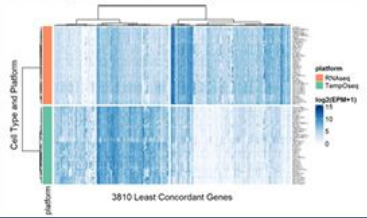
The $\log_2$ expression per million (EPM) data for 19,290 overlapping genes were well correlated between the two platforms across the 39 cell lines (0.77, 95% CI: 0.76 – 0.78). Non-concordance was determined by removing genes with the greatest $\log_2$ differences in expression between TempO-seq and RNA-seq until the percent variance explained by platform effects was resolved to less than 10% (PERMANOVA platform $R^2 < 0.10$). This determined that the majority of genes (15,480 genes, 80%) had concordant baseline gene expression levels. Additionally, relative log expression (RLE) normalization calculated for each platform resolved the observed platform divergence. RLE calculation:

$$RLE\ for\ Gene\ X = log_2\left(\frac{(EPM+1)\ for\ gene\ X\ within\ a\ single\ cell\ line}{Average\ (EPM+1)\ for\ gene\ X\ across\ all\ 39\ cell\ lines}\right)$$

F3) TempO-seq and RNA-seq relative $\log_2$ difference was centered around zero for the 19,290 overlapping genes. Non-concordant genes (shaded in red) had a $\log_2$(EPM + 1) difference of less than -2.09 (13th percentile) or greater than 1.47 (87th percentile).



F4) Expression of 3,810 least concordant genes. The genes are split about 50/50 between whether they had higher expression within TempO-seq or within RNA-seq.



F5) Odds ratios (ORs) evaluated which gene ontologies (GO) contained a higher proportion of non-concordant genes. The ontologies with more non-concordant genes (OR > 1, orange color) contained many ribosomal and histone family genes, and ontologies enriched for concordant genes (OR < 1, green color) pertained to cell structure and kinase, immune, and golgi functions.

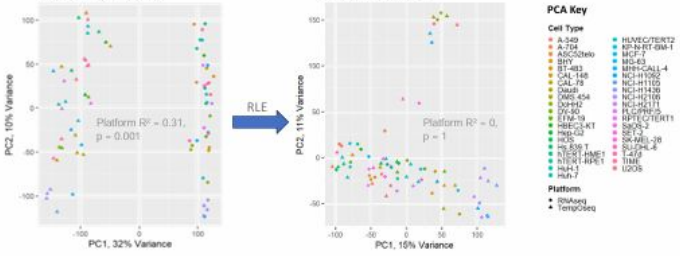
Gene Ontology OR = $\frac{a/b}{c/d}$	Within GO signature	Not within GO signature	Total
Non-concordant genes with EPM ≥ 5	a	b	3,810 Genes
Concordant genes with EPM ≥ 5	c	d	6,677 Genes

Gene Ontology Signature	GO Genes, Total	GO Genes in Analysis (a + c)	% GO Genes in Analysis	OR	1/OR	FDR p-Value
GOBP_PROTEIN_LOCALIZATION_TO_CENP_A_CONTAINING_CHROMATIN	18	17	94%	28	-	5.6E-04
GOCC_CHROMOSOME_CENTROMERIC_CORE_DOMAIN	19	18	95%	14	-	3.1E-03
GOMF_STRUCTURAL_CONSTITUENT_OF_CHROMATIN	97	67	69%	10	-	2.2E-12
GOBP_NEGATIVE_REGULATION_OF_MEGAKARYOCYTE_DIFFERENTIATION	20	17	85%	8.2	-	4.7E-02
GOCC_CYTOSOLIC_LARGE_RIBOSOMAL_SUBUNIT	60	55	92%	6.3	-	4.7E-07
GOCC_CYTOSOLIC_SMALL_RIBOSOMAL_SUBUNIT	41	36	88%	4.0	-	3.1E-02
GOCC_NUCLEOSOME	134	97	72%	3.8	-	4.8E-07
GOCC_CYTOSOLIC_RIBOSOME	118	107	91%	3.5	-	4.8E-07
GOBP_NUCLEOSOME_ORGANIZATION	138	105	76%	3.4	-	1.2E-06
GOMF_STRUCTURAL_CONSTITUENT_OF_RIBOSOME	169	153	91%	3.0	-	1.4E-07
GOCC_LARGE_RIBOSOMAL_SUBUNIT	117	111	95%	2.8	-	1.2E-04
GOCC_RIBOSOMAL_SUBUNIT	188	177	94%	2.7	-	4.1E-07
GOBP_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS	76	73	96%	2.5	-	4.4E-02
GOCC_CATALYTIC_STEP_2_SPLICEOSOME	91	88	97%	2.4	-	2.0E-02
GOBP_CYTOPLASMIC_TRANSLATION	156	146	94%	2.4	-	1.7E-04
GOCC_PRERIBOSOME	109	105	96%	2.2	-	3.6E-02
GOCC_RIBOSOME	239	215	90%	2.2	-	2.6E-05
GOBP_PROTEIN_DNA_COMPLEX_ASSEMBLY	240	189	79%	2.1	-	5.1E-04
GOMF_STRUCTURAL_MOLECULE_ACTIVITY	809	446	55%	1.9	-	4.5E-07
GOCC_RIBONUCLEOPROTEIN_COMPLEX	1,169	661	57%	1.7	-	2.0E-07
GOBP_RIBOSOME_BIOGENESIS	325	308	95%	1.7	-	6.2E-03
GOBP_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	502	447	89%	1.6	-	5.6E-04
GOCC_GOLGI_APPARATUS	1,634	1,068	65%	0.72	1.3	4.6E-02
GOBP_LYMPHOCYTE_ACTIVATION	796	405	51%	0.65	1.5	4.4E-02
GOMF_PROTEIN_KINASE_ACTIVITY	577	382	66%	0.60	1.7	6.2E-03
GOBP_REGULATION_OF_ANATOMICAL_STRUCTURE_MORPHOGENESIS	937	488	52%	0.60	1.7	5.1E-04
GOCC_BASEMENT_MEMBRANE	90	49	54%	0.15	6.5	4.4E-03

F6) PCA and Pearson correlations before vs after relative log expression (RLE) normalization, which resolved the platform divergence to <10%. It also improved cell line clustering, likely driven by each cell lines' unique gene expression patterns.

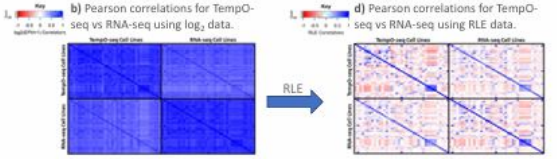
a) Initial PCA on $\log_2$ data showed a platform divergence.

c) RLE normalization resolved the platform divergence.



b) Pearson correlations for TempO-seq vs RNA-seq using $\log_2$ data.

d) Pearson correlations for TempO-seq vs RNA-seq using RLE data.



*Presenter: word.laura@epa.gov
Advisor: harrill.joshua@epa.gov