Comparing RNA-seq and TempO-seq mRNA data sets: a case study

Dr. Laura Word-Taylor
U.S. Environmental Protection Agency Post-Doc
Advisor: Dr. Richard Judson
SOT March 20th, 2023

The views expressed are those of the authors and do not necessarily represent the views or the policies of the U.S. EPA

Brief overview of mRNA sequencing

Transcriptomics quantifies messenger RNA

- The human genome contains approximately **20,000 genes**
- **Measuring changes in gene expression** can help elucidate chemical mechanisms of action in cells
- There are **different technologies** for mRNA sequencing, including:
  - **RNA-seq**
  - **TempO-seq**

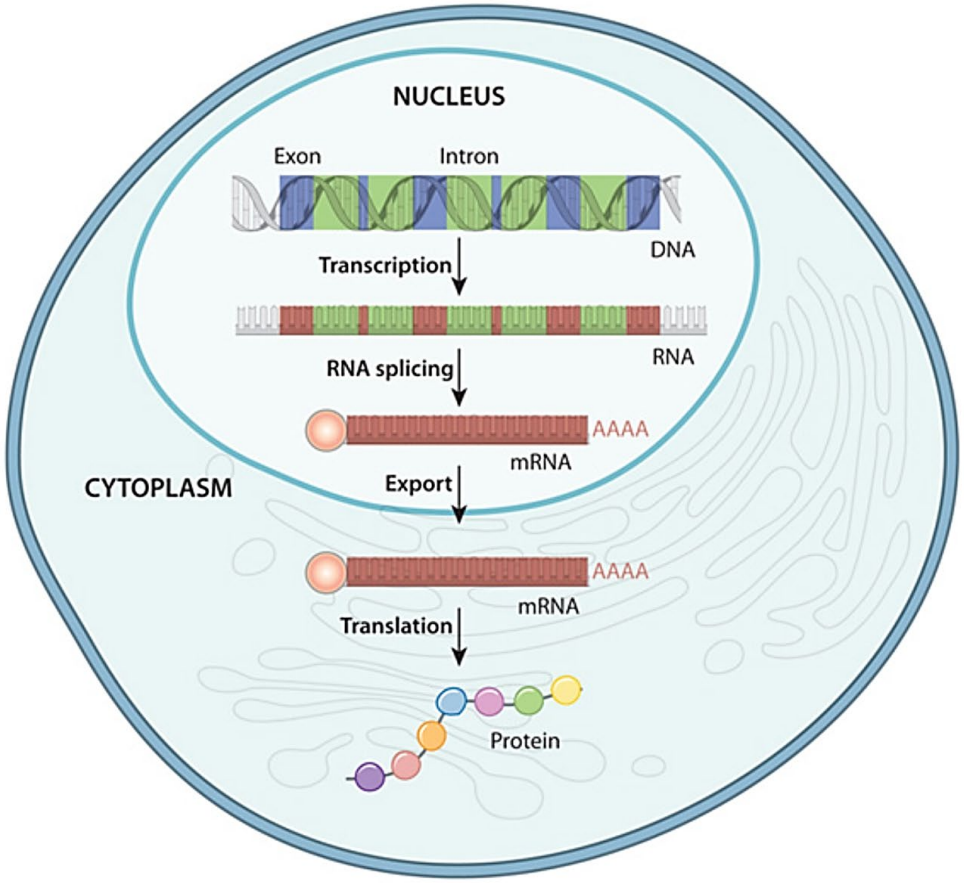


Figure 1: An overview of the flow of information from DNA to protein in a eukaryote
First, both coding and noncoding regions of DNA are transcribed into mRNA. Some regions are removed (introns) during initial mRNA processing. The remaining exons are then spliced together, and the spliced mRNA molecule (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.

© 2010 Nature Education

Kuang, Jujiao, et al. "An overview of technical considerations when using quantitative real-time PCR analysis of gene expression in human exercise research." *PloS one* 13.5 (2018): e0196438.

Figure Detail

Differences in the Technologies

RNA-seq:

- **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 mRNA** before quantification and is **more expensive and time-intensive**

TempO-seq:

- **Easier sample prep** because lysed cells can be used
- **Less sample material is needed** (picograms instead of nanograms)
- **Can be less expensive** per sample when done in bulk
- **No specialized equipment** is necessary (regular PCR)
- **Can attenuate highly expressed genes**, meaning less read-depth is needed
- **Must have probes for the species**, only quantifies RNA for which there is a tag to measure it

OBJECTIVE

Need to Compare the Two Technologies

TempO-seq is a new method that needs to be validated against the established RNA-seq method

Previous Research

Study using rat liver samples showed TempO-seq and RNA-seq showed platform differences but the mechanism of action (MOA) grouped by chemical treatment instead of by platform

Work comparing human samples and more types of cells is still needed

Front Genet. 2018 Oct 30;9:485. doi: 10.3389/fgene.2018.00485. eCollection 2018.

A Comparison of the TempO-Seq S1500+ Platform to RNA-Seq and Microarray Using Rat Liver Mode of Action Samples

Pierre R Bushel [1], Richard S Paules [2], Scott S Auerbach [2]

Affiliations + expand

PMID: 30420870 PMCID: PMC6217592 DOI: 10.3389/fgene.2018.00485

Free PMC article

Abstract

The TempO-Seq™ platform allows for targeted transcriptomic analysis and is currently used by many groups to perform high-throughput gene expression analysis. Herein we performed a comparison of gene expression characteristics measured using 45 purified RNA samples from the livers of rats exposed to chemicals that fall into one of five modes of action (MOAs). These samples have been previously evaluated using Affymetrix™ rat genome 230 2.0 microarrays and Illumina® whole transcriptome RNA-Seq. Comparison of these data with TempO-Seq analysis using the rat S1500+ beta gene set identified clear differences in the platforms related to signal to noise, root mean squared error, and/or sources of variability. Microarray and TempO-Seq captured the most variability in terms of MOA and chemical treatment whereas RNA-Seq had higher noise and larger differences between samples within a MOA. However, analysis of the data by hierarchical clustering, gene subnetwork connectivity and biological process representation of MOA-varying genes revealed that the samples clearly grouped by treatment as opposed to gene expression platform. Overall these findings demonstrate that the results from the TempO-Seq platform are consistent with findings on other more established approaches for measuring the genome-wide transcriptome.

Keywords: RNA-Seq; S1500+; TempO-Seq; chemicals; microarray; mode of action; toxicants; toxicogenomics.

Case Study for Comparing TempO-seq and RNA-seq Data

Baseline Expression Data Sets for the Case Study

EPA Internal data [TempO-seq]

Cell Atlas and Tox21 CPP2

External data [RNA-seq]

Human Protein Atlas (HPA)

TempO-seq Cell Types

Cell atlas
A549
A-704
ARPE-19
ASC52telo
BHY
BJ-5ta
CCD-18Co
CHON-001
Daudi
DMS-454
DV-90
HBEC3-KT
HepaRG_2D
HepaRG_3D
HepG2
HME-1
HPNE
HSAEC-1
HTB-9
HUVEC
Ker-CT
KP-N-RT-BM-1
MCF-7
NCI-H1092
RPE-1
RPTEC
SET-2
SK-MEL-28
TeloHAEC
TIME
U-2 OS

Tox21 cpp2
BT-483
C3A
Cal-148
Cal-78
CCD-18Co
Daudi
Detroit-551
DoHH2
EFM-19
HepG2
hNP1
HOS
Hs.839.T
Hs-5
Huh-1
Huh-7
MCF-7
MG-63
MHH-CALL-4
NCI-H1092
NCI-H1105
NCI-H1436
NCI-H2106
NCI-H2171
PLC/PRF/5
Saos-2
SU-DHL-6
T47-D
U-2 OS

RNA-seq Cell Types

Human Protein Atlas (HPA)	HPA Cont.
A-431	JURKAT
A549	K-562
AF22	Karpas-707
AN3-CA	LHCN-M2
ASC diff	MCF7
ASC TERT1	MOLT-4
BEWO	NB-4
BJ	NTERA-2
BJ hTERT+	OE19
BJ hTERT+ SV40 Large T+	PC-3
BJ hTERT+ SV40 Large T+ RasG12V	REH
CACO-2	RH-30
CAPAN-2	RPMI-8226
Daudi	RPTEC TERT1
EFO-21	RT4
fHDF/TERT166	SCLC-21H
GAMG	SH-SY5Y
HaCaT	SiHa
HAP1	SK-BR-3
HBEC3-KT	SK-MEL-30
HBF TERT88	SuSa
HDLM-2	T-47d
HEK 293	THP-1
HEL	TIME
HeLa	U-138 MG
Hep G2	U-2 OS
HHSteC	U-2197
HL-60	U-251 MG
HMC-1	U-266/70
HSkMC	U-266/84
hTCEpi	U-698
hTEC/SVTERT24-B	U-87 MG
hTERT-HME1	U-937
hTERT-RPE1	WM-115
HUVEC TERT2	

Step 1. Compare the Cell Atlas and Tox21 CPP2 Data Sets [TempO-seq]

Baseline Expression Data Sets for the Case Study

EPA Internal data [TempO-seq]

Cell Atlas and Tox21 CPP2

6 overlapping cell types

TempO-seq Cell Types

Cell atlas
A549
A-704
ARPE-19
ASC52telo
BHY
BJ-5ta
CCD-18Co
CHON-001
Daudi
DMS-454
DV-90
HBEC3-KT
HepaRG_2D
HepaRG_3D
HepG2
HME-1
HPNE
HSAEC-1
HTB-9
HUVEC
Ker-CT
KP-N-RT-BM-1
MCF-7
NCI-H1092
RPE-1
RPTEC
SET-2
SK-MEL-28
TeloHAEC
TIME
U-2 OS

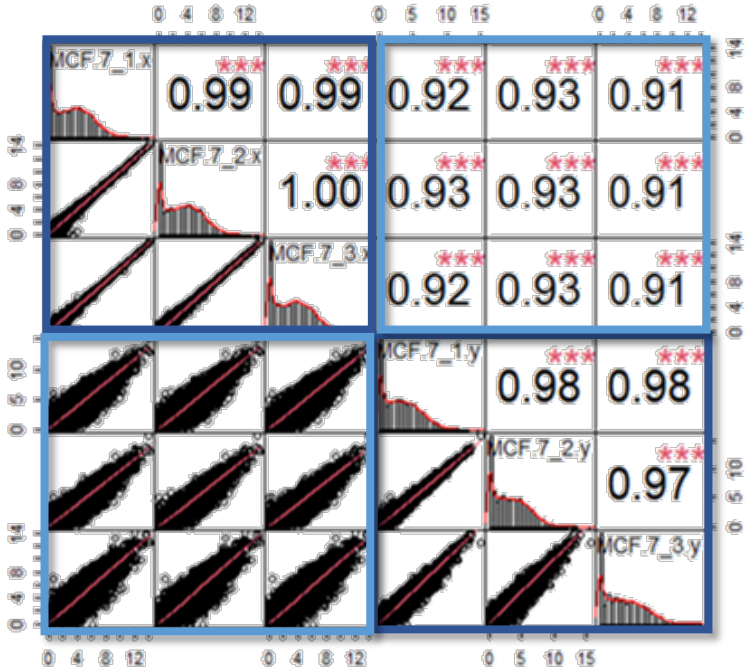
Tox21 cpp2
BT-483
C3A
Cal-148
Cal-78
CCD-18Co
Daudi
Detroit-551
DoHH2
EFM-19
HepG2
hNP1
HOS
Hs.839.T
Hs-5
Huh-1
Huh-7
MCF-7
MG-63
MHH-CALL-4
NCI-H1092
NCI-H1105
NCI-H1436
NCI-H2106
NCI-H2171
PLC/PRF/5
Saos-2
SU-DHL-6
T47-D
U-2 OS

Cell Line	Tissue Origin
MCF-7	Breast
U-2 OS	Bone
HepG2	Liver
Daudi	Peripheral Blood (B lymphoblast)
CCD-18Co	Colon
NCI-H1092	Lung

Pearson Correlations show high reproducibility for the 6 common cell types within Cell Atlas and Tox21 CPP2, log_2(CPM+1) TempO-seq Data

Notation: Cell atlas = ".x", Tox21 CPP2 = ".y", and _1,_2,_3 denotes replicate number

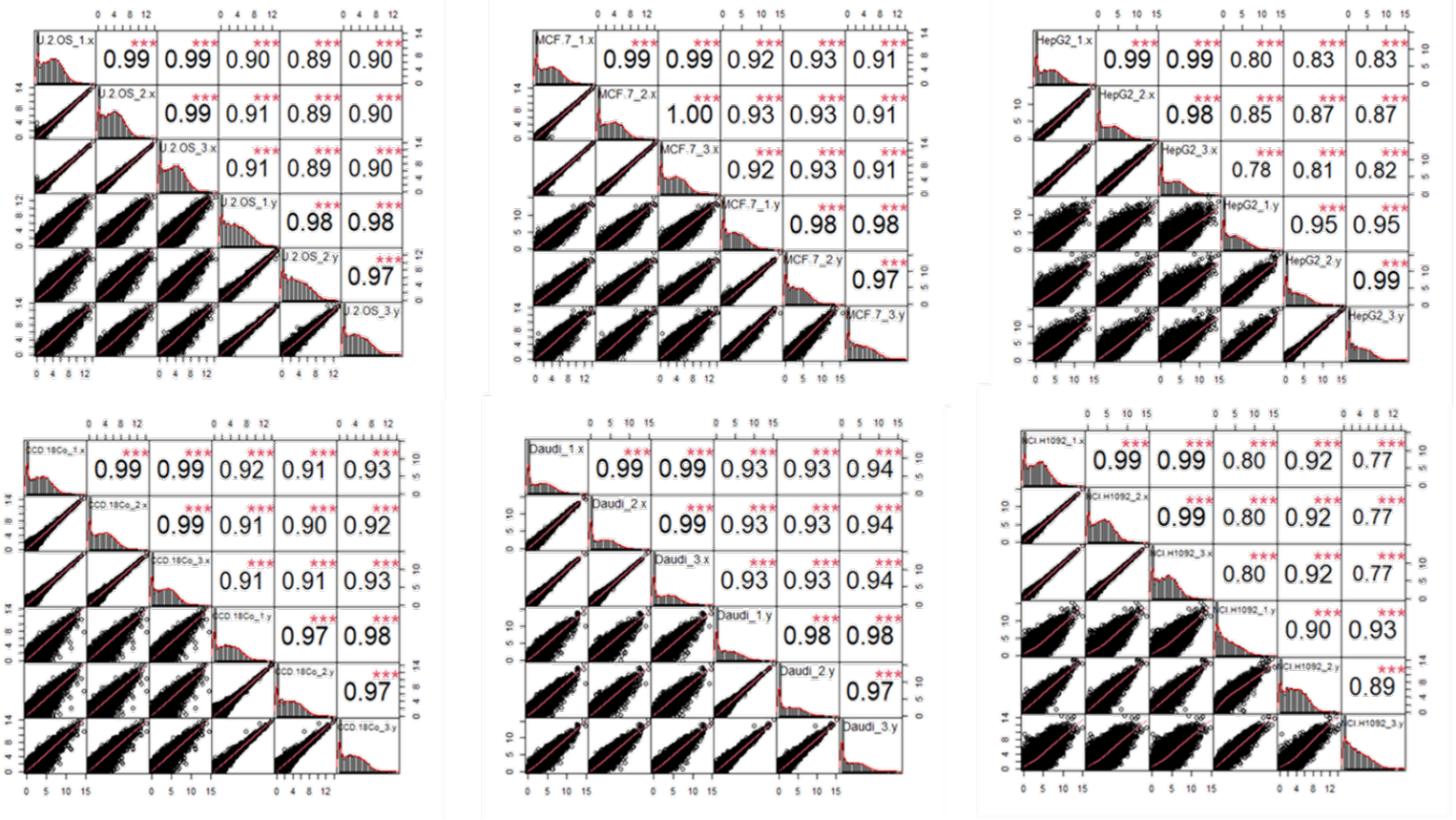
MCF7 Cells



- **Pearson correlations between technical replicates in the same data set**
- **Pearson correlations between the two TempO-seq data sets**

Pearson Correlations show high reproducibility for the 6 common cell types within Cell Atlas and Tox21 CPP2, log_2(CPM+1) TempO-seq Data

Notation: Cell atlas = ".x", Tox21 CPP2 = ".y", and _1,_2,_3 denotes replicate number



- **Very high reproducibility within technical replicates in the same data set**
  - **Pearson correlation ave 0.98 (range 0.89-1.00)**

- **Strong reproducibility between the two TempO-seq data sets**
  - **Pearson correlation ave 0.90 (range 0.77-0.94)**

Principal Component Analysis (PCA) shows that the two TempO-seq data sets were reproducible

Cell Atlas vs TOX21 CPP2 $\log_2$(CPM) TempO-seq Data

Step 2. Compare the Cell Atlas and Tox21 CPP2 Data Sets [TempO-seq] to the Human Protein Atlas (HPA) Data [RNA-seq]

Baseline Expression Data Sets for the Case Study

EPA Internal data [TempO-seq]

Cell Atlas and Tox21 CPP2

External data [RNA-seq]

Human Protein Atlas (HPA)

12 overlapping cell types

TempO-seq Cell Types

Cell atlas	Tox21 cpp2
A549	BT-483
A-704	C3A
ARPE-19	Cal-148
ASC52telo	Cal-78
BHY	CCD-18Co
BJ-5ta	Daudi
CCD-18Co	Detroit-551
CHON-001	DoHH2
Daudi	EFM-19
DMS-454	HepG2
DV-90	hNP1
HBEC3-KT	HOS
HepaRG_2D	Hs.839.T
HepaRG_3D	Hs-5
HepG2	Huh-1
HME-1	Huh-7
HPNE	MCF-7
HSAEC-1	MG-63
HTB-9	MHH-CALL-4
HUVEC	NCI-H1092
Ker-CT	NCI-H1105
KP-N-RT-BM-1	NCI-H1436
MCF-7	NCI-H2106
NCI-H1092	NCI-H2171
RPE-1	PLC/PRF/5
RPTEC	Saos-2
SET-2	SU-DHL-6
SK-MEL-28	T47-D
TeloHAEC	U-2 OS
TIME	
U-2 OS	

RNA-seq Cell Types

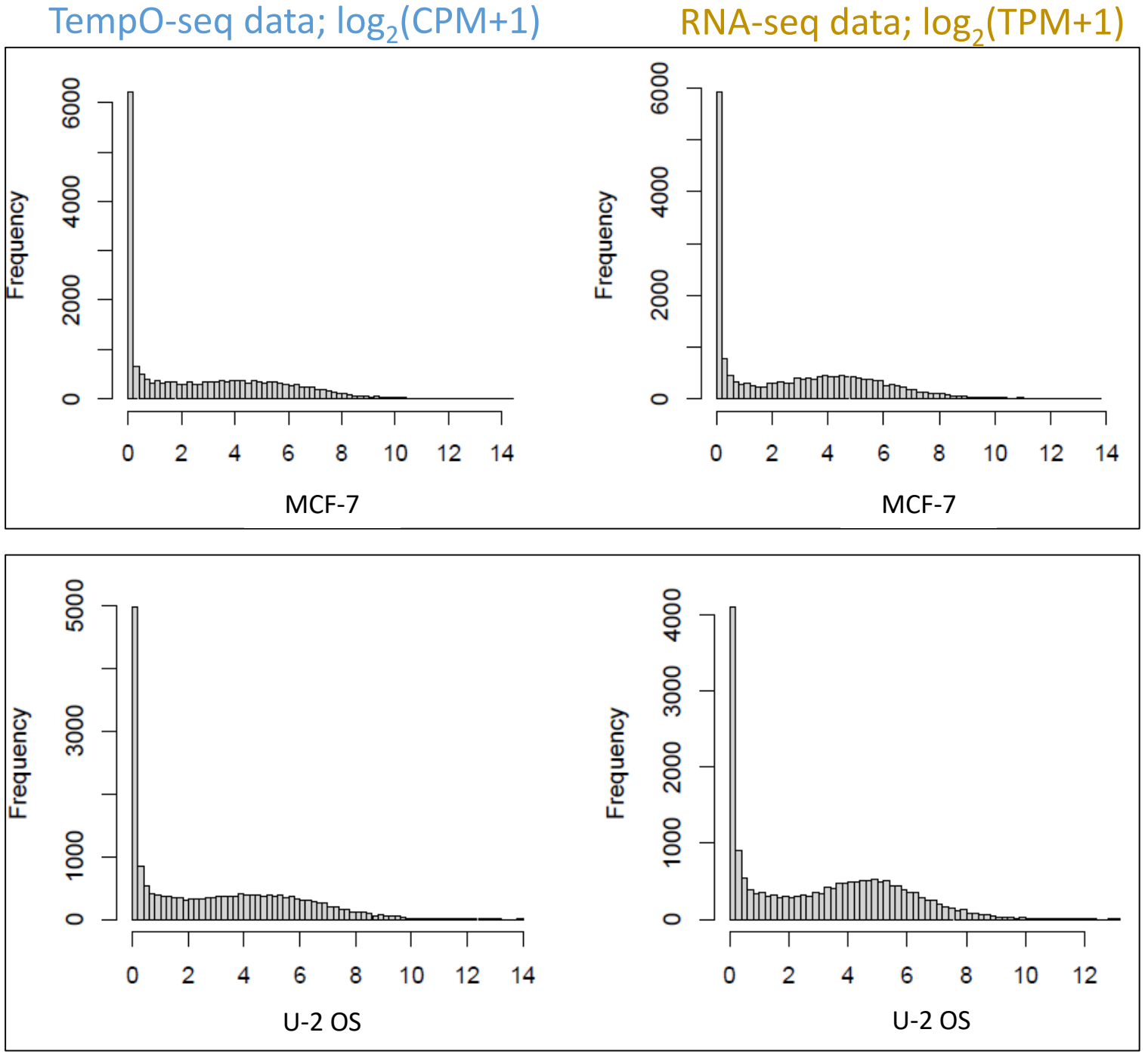
Human Protein Atlas (HPA)	HPA Cont.
A-431	JURKAT
A549	K-562
AF22	Karpas-707
AN3-CA	LHCN-M2
ASC diff	MCF7
ASC TERT1	MOLT-4
BEWO	NB-4
BJ	NTERA-2
BJ hTERT+	OE19
BJ hTERT+ SV40 Large T+	PC-3
BJ hTERT+ SV40 Large T+ RasG12V	REH
CACO-2	RH-30
CAPAN-2	RPMI-8226
Daudi	RPTEC TERT1
EFO-21	RT4
fHDF/TERT166	SCLC-21H
GAMG	SH-SY5Y
HaCaT	SiHa
HAP1	SK-BR-3
HBEC3-KT	SK-MEL-30
HBF TERT88	SuSa
HDLM-2	T-47d
HEK 293	THP-1
HEL	TIME
HeLa	U-138 MG
Hep G2	U-2 OS
HHSteC	U-2197
HL-60	U-251 MG
HMC-1	U-266/70
HSkMC	U-266/84
hTCEpi	U-698
hTEC/SVTERT24-B	U-87 MG
hTERT-HME1	U-937
hTERT-RPE1	WM-115
HUVEC TERT2	

Cell Type	Tissue
A549	Lung
Daudi	Haematopoietic and lymphoid
HepG2	Liver
MCF-7	Breast
U-2 OS	Bone
T47-D	Breast
HBEC3-KT	Lung
HME-1	Breast, mammary gland
HUVEC/TERT-2	Blood vessel wall
RPE-1	Eye
RPTEC/TERT-1	Kidney
TIME	Skin

The data distributions are very similar for each cell type

Histograms for internal TempO-seq data (left) vs Human Protein Atlas RNA-seq data (right)

Showing two cell types of interest



Strong Pearson correlation coefficients

Different cell types intra-platform: normally around 0.80 (0.69-0.97)

Matching cell types inter-platform: 0.80 (0.75-0.82)

x = TempO-seq data

y = RNA-seq data

	A549.x	Daudi.x	HBEC3.KT.x	HepG2.x	HME.1.x	HUVEC.x	MCF.7.x	RPE.1.x	RPTEC.x	TIME.x	U.2.OS.x	T47.D.x	A549.y	Daudi.y	HBEC3.KT.y	HepG2.y	HME.1.y	HUVEC.y	MCF.7.y	RPE.1.y	RPTEC.y	TIME.y	U.2.OS.y	T47.D.y
A549.x	1.00	0.79	0.80	0.81	0.88	0.87	0.88	0.91	0.89	0.88	0.88	0.83	0.80	0.65	0.73	0.70	0.73	0.72	0.71	0.72	0.73	0.71	0.73	0.71
Daudi.x	0.79	1.00	0.69	0.73	0.78	0.77	0.80	0.78	0.75	0.78	0.78	0.77	0.61	0.82	0.62	0.61	0.63	0.62	0.62	0.58	0.60	0.61	0.62	0.62
HBEC3.KT.x	0.80	0.69	1.00	0.75	0.90	0.78	0.79	0.81	0.79	0.79	0.78	0.75	0.69	0.60	0.78	0.65	0.75	0.68	0.68	0.70	0.70	0.69	0.68	0.69
HepG2.x	0.81	0.73	0.75	1.00	0.79	0.77	0.80	0.79	0.78	0.77	0.79	0.79	0.66	0.61	0.65	0.79	0.65	0.64	0.65	0.64	0.65	0.64	0.65	0.65
HME.1.x	0.88	0.78	0.90	0.79	1.00	0.88	0.84	0.91	0.85	0.89	0.87	0.79	0.74	0.67	0.80	0.69	0.82	0.75	0.71	0.75	0.74	0.76	0.75	0.71
HUVEC.x	0.87	0.77	0.78	0.77	0.88	1.00	0.83	0.91	0.86	0.97	0.86	0.78	0.70	0.64	0.72	0.65	0.73	0.82	0.65	0.73	0.70	0.81	0.71	0.66
MCF.7.x	0.88	0.80	0.79	0.80	0.84	0.83	1.00	0.86	0.87	0.83	0.86	0.91	0.70	0.64	0.69	0.66	0.69	0.66	0.76	0.66	0.68	0.66	0.68	0.74
RPE.1.x	0.91	0.78	0.81	0.79	0.91	0.91	0.86	1.00	0.89	0.92	0.90	0.82	0.73	0.65	0.75	0.67	0.76	0.75	0.68	0.80	0.73	0.76	0.75	0.69
RPTEC.x	0.89	0.75	0.79	0.78	0.85	0.86	0.87	0.89	1.00	0.86	0.85	0.82	0.70	0.60	0.71	0.65	0.71	0.69	0.67	0.70	0.80	0.69	0.68	0.68
TIME.x	0.88	0.78	0.79	0.77	0.89	0.97	0.83	0.92	0.86	1.00	0.87	0.78	0.70	0.64	0.73	0.65	0.74	0.80	0.66	0.73	0.71	0.82	0.71	0.66
U.2.OS.x	0.88	0.78	0.78	0.79	0.87	0.86	0.86	0.90	0.85	0.87	1.00	0.83	0.68	0.61	0.69	0.64	0.70	0.68	0.65	0.70	0.67	0.68	0.78	0.66
T47.D.x	0.83	0.77	0.75	0.79	0.79	0.78	0.91	0.82	0.82	0.78	0.83	1.00	0.64	0.60	0.65	0.62	0.64	0.61	0.67	0.62	0.64	0.60	0.63	0.75
A549.y	**0.80**	0.61	0.69	0.66	0.74	0.70	0.70	0.73	0.70	0.70	0.68	0.64	1.00	0.78	0.88	0.86	0.86	0.86	0.89	0.86	0.88	0.85	0.88	0.87
Daudi.y	0.65	**0.82**	0.60	0.61	0.67	0.64	0.64	0.65	0.60	0.64	0.61	0.60	0.78	1.00	0.78	0.77	0.78	0.77	0.79	0.73	0.75	0.76	0.78	0.78
HBEC3.KT.y	0.73	0.62	**0.78**	0.65	0.80	0.72	0.69	0.75	0.71	0.73	0.69	0.65	0.88	0.78	1.00	0.83	0.94	0.88	0.86	0.87	0.89	0.87	0.87	0.86
HepG2.y	0.70	0.61	0.65	**0.79**	0.69	0.65	0.66	0.67	0.65	0.65	0.64	0.62	0.86	0.77	0.83	1.00	0.81	0.81	0.85	0.79	0.82	0.79	0.82	0.83
HME.1.y	0.73	0.63	0.75	0.65	**0.82**	0.73	0.69	0.76	0.71	0.74	0.70	0.64	0.86	0.78	0.94	0.81	1.00	0.88	0.83	0.89	0.88	0.87	0.87	0.83
HUVEC.y	0.72	0.62	0.68	0.64	0.75	**0.82**	0.66	0.75	0.69	0.80	0.68	0.61	0.86	0.77	0.88	0.81	0.88	1.00	0.81	0.89	0.86	0.95	0.86	0.81
MCF.7.y	0.71	0.62	0.68	0.65	0.71	0.65	**0.76**	0.68	0.67	0.66	0.65	0.67	0.89	0.79	0.86	0.85	0.83	0.81	1.00	0.80	0.84	0.80	0.85	0.91
RPE.1.y	0.72	0.58	0.70	0.64	0.75	0.73	0.66	**0.80**	0.70	0.73	0.70	0.62	0.86	0.73	0.87	0.79	0.89	0.89	0.80	1.00	0.88	0.89	0.87	0.82
RPTEC.y	0.73	0.60	0.70	0.65	0.74	0.70	0.68	0.73	**0.80**	0.71	0.67	0.64	0.88	0.75	0.89	0.82	0.88	0.86	0.84	0.88	1.00	0.85	0.85	0.85
TIME.y	0.71	0.61	0.69	0.64	0.76	0.81	0.66	0.76	0.69	**0.82**	0.68	0.60	0.85	0.76	0.87	0.79	0.87	0.95	0.80	0.89	0.85	1.00	0.85	0.80
U.2.OS.y	0.73	0.62	0.68	0.65	0.75	0.71	0.68	0.75	0.68	0.71	**0.78**	0.63	0.88	0.78	0.87	0.82	0.87	0.86	0.85	0.87	0.85	0.85	1.00	0.85
T47.D.y	0.71	0.62	0.69	0.65	0.71	0.66	0.74	0.69	0.68	0.66	0.66	**0.75**	0.87	0.78	0.86	0.83	0.83	0.81	0.91	0.82	0.85	0.80	0.85	1.00

x = TempO-seq

y = RNA-seq

Percentage of genes with matching expression magnitude at different levels in both data sets has high agreement

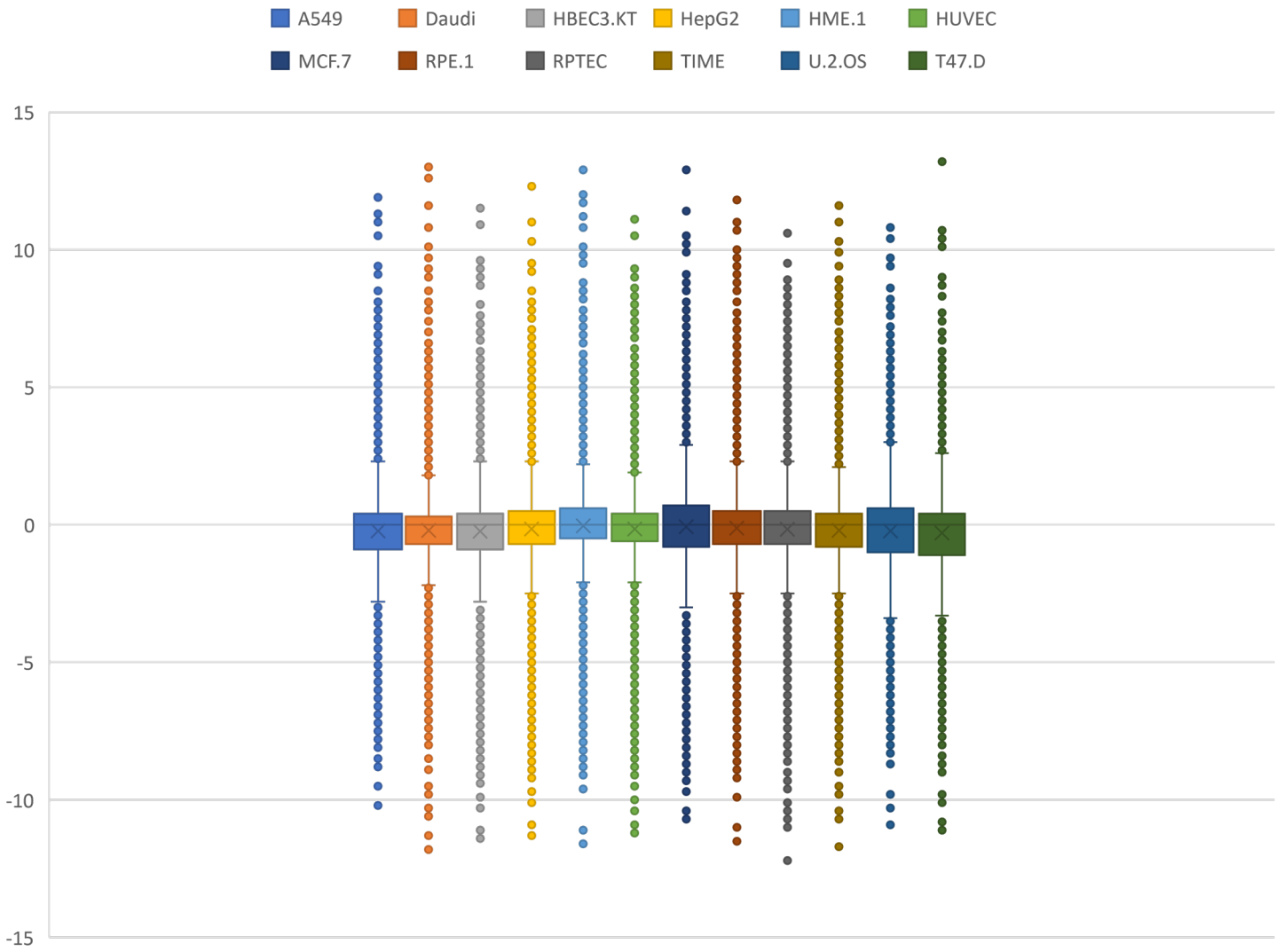
Low expression in both platforms → High expression in both platforms

	TempO-seq log_2(CPM+1) data vs RNA-seq log_2(TPM+1) data					
Cell_type	match_level_1	match_level_2	match_level_5	match_level_7	match_level_10	match_level_14
A549	90%	87%	82%	92%	99.0%	99.98%
Daudi	90%	89%	85%	92%	99.1%	99.98%
HBEC3.KT	89%	86%	82%	93%	99.2%	99.97%
HepG2	90%	87%	82%	92%	99.0%	99.98%
HME.1	91%	89%	83%	92%	99.1%	99.99%
HUVEC	92%	89%	83%	92%	99.1%	99.99%
MCF.7	88%	85%	81%	91%	98.9%	99.99%
RPE.1	91%	88%	83%	92%	99.1%	99.99%
RPTEC	91%	87%	83%	92%	99.1%	99.98%
TIME	89%	85%	79%	90%	98.9%	99.99%
U.2.OS	91%	89%	83%	92%	99.1%	99.99%
T47.D	88%	86%	80%	91%	99.1%	99.99%
average	**90%**	**87%**	**82%**	**92%**	**99.1%**	**99.99%**
min	**88%**	**85%**	**79%**	**90%**	**98.9%**	**99.97%**
max	**92%**	**89%**	**85%**	**93%**	**99.2%**	**99.99%**

match_level_1: Percentage that internal data >= 1 & HPA data >= 1 OR internal data < 1 & HPA data < 1

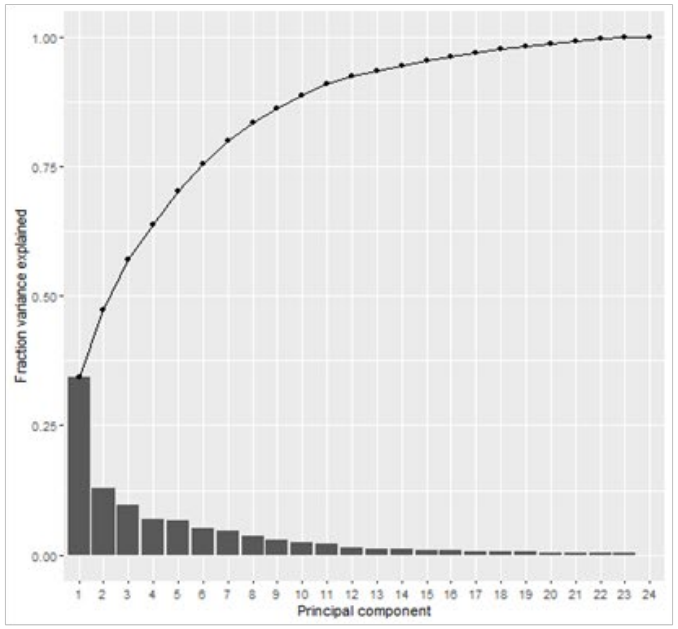
TempO-seq minus RNA-seq $\log_2$(expression level) is centered around zero across cell types, showing **strong replicability**

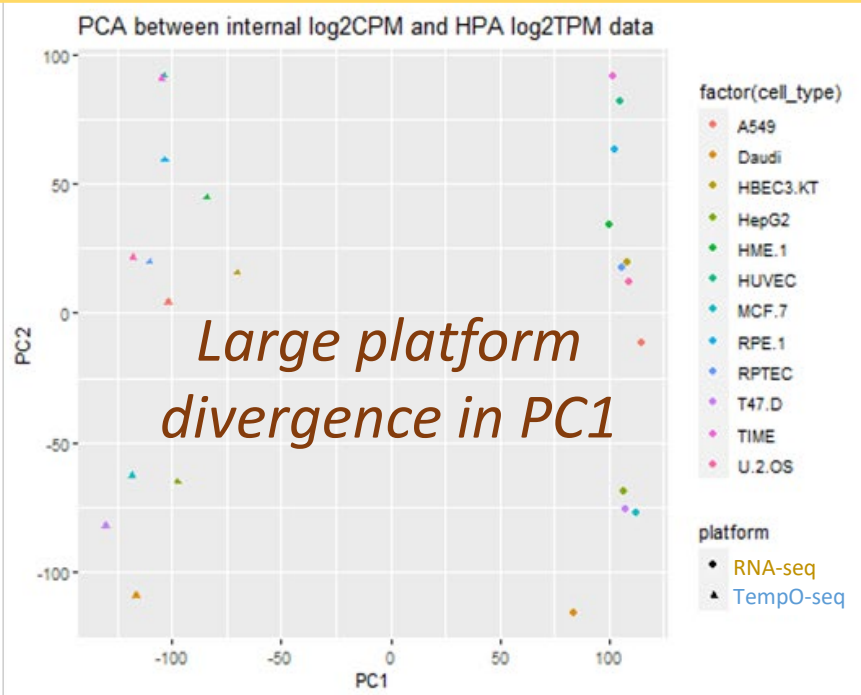
Log_2 difference in expression by cell type (TempO-seq minus RNA-seq)



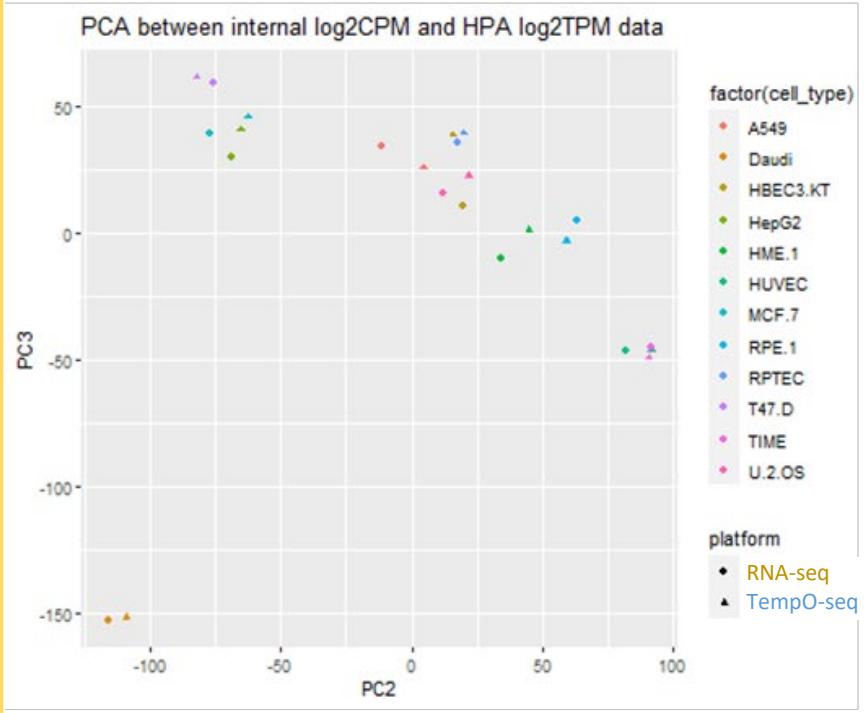
Principal component analysis (PCA)

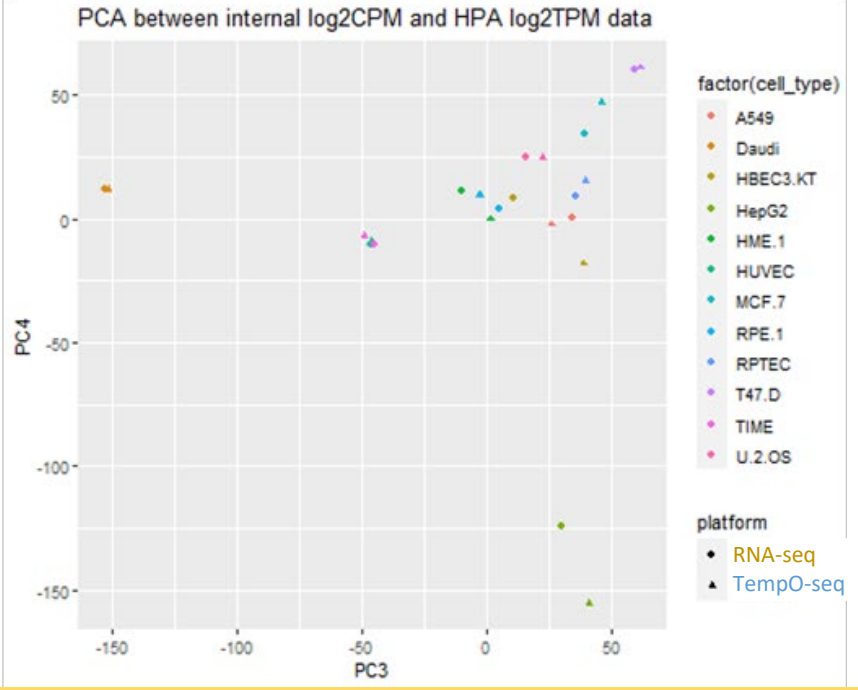
TempO-seq vs RNA-seq



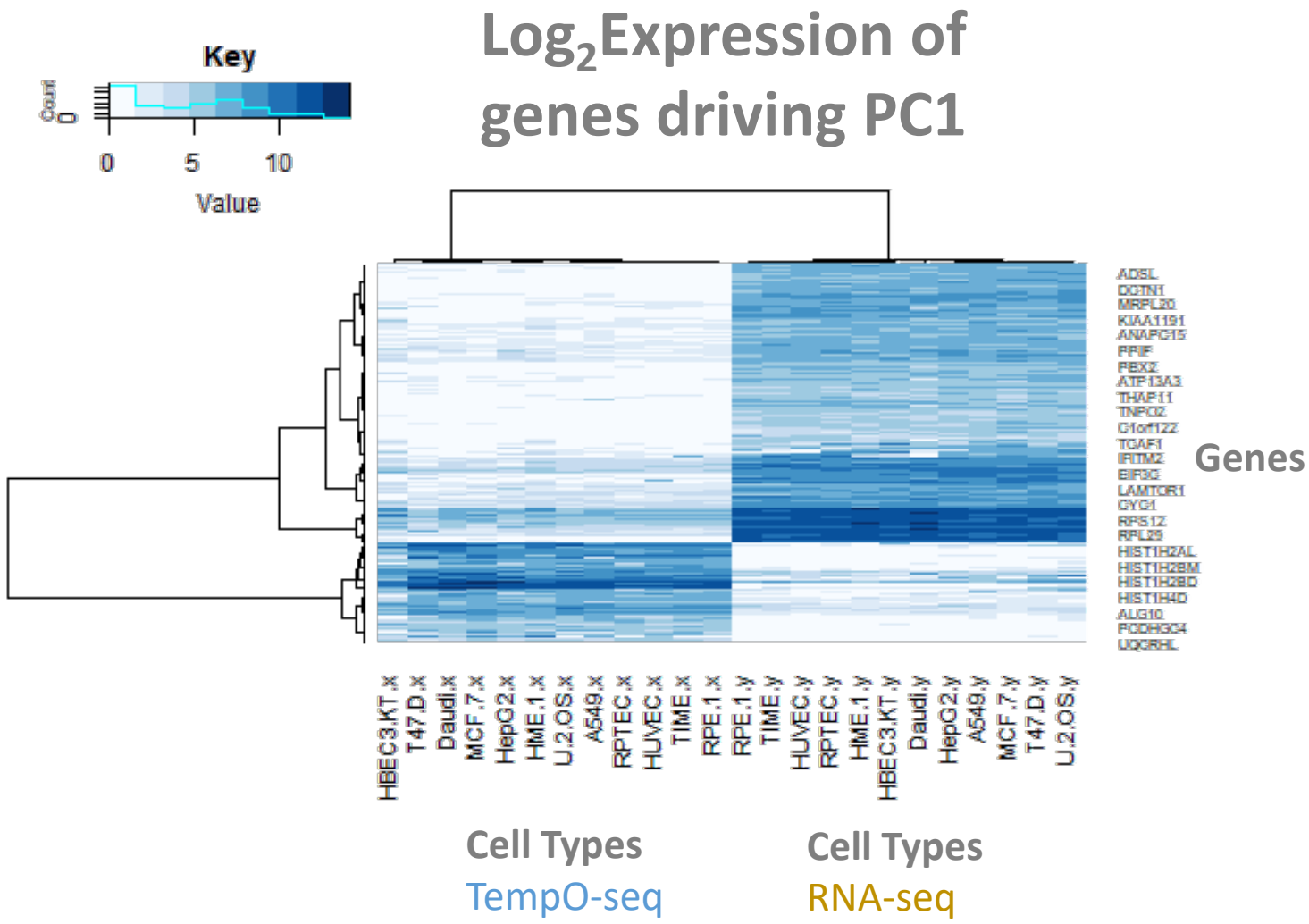


Large platform divergence in PC1





The expression level in genes driving the divergence in PC1 consistently have low expression in one platform and high expression in the other platform, across cell types

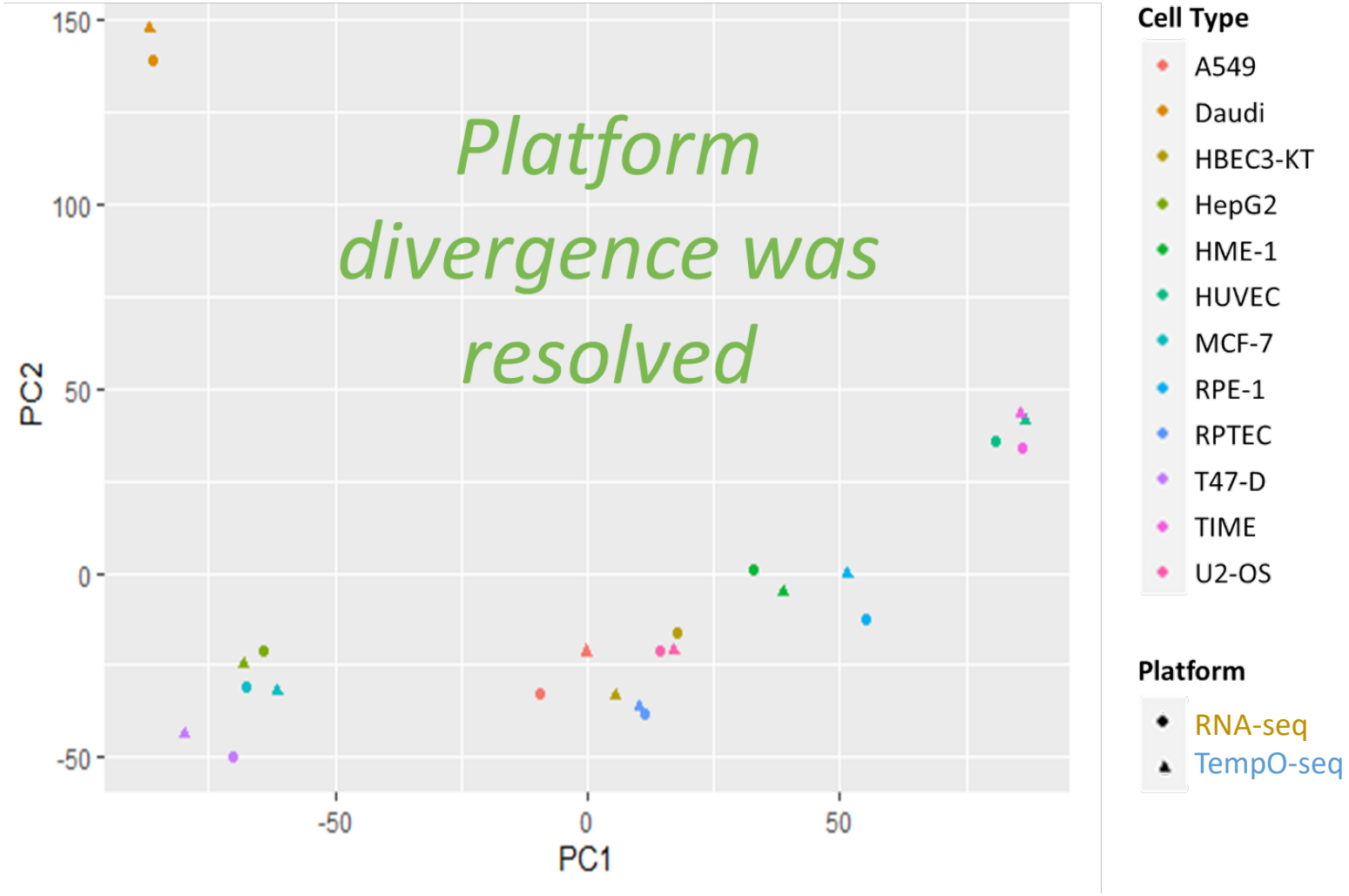


Method a) Removing genes with the highest difference in gene expression resolved the PC1 platform divergence

4,398 of the 19,119 genes were removed that had abs(ave $\log_2$CPM diff) > 1.5

TempO-seq vs RNA-seq PCA

Subset of genes with best replicability



Relative Log Expression (RLE) Method

- Calculated average across cell types for the TempO-seq data and separately for the RNA-seq data
- Divided the $\log_2$(CPM) for each cell line by the average expression for that data set

ensembl_gene	A549.x	Daudi.x	HBEC3.KT.x	HepG2.x	HME.1.x	HUVEC.x	MCF.7.x	RPE.1.x	RPTEC.x	TIME.x	U.2.OS.x	T47.D.x	Average
ENSG00000167658	10.1	11.5	12.5	11.6	11.6	11.3	11.3	11.0	11.0	10.7	10.9	11.4	11.2
ENSG00000213145	10.0	9.8	4.1	8.2	0.8	1.5	3.2	0.9	1.7	4.4	1.4	8.1	4.5
ENSG00000115648	10.0	0.4	6.7	0.5	9.9	0.0	11.1	8.4	2.7	6.8	5.2	9.6	5.9
ENSG00000057019	10.0	0.3	7.4	6.9	10.2	7.5	5.7	10.1	8.7	6.0	9.2	4.7	7.2
ENSG00000104687	10.0	7.5	7.2	8.4	6.9	7.1	7.8	6.8	8.0	7.3	6.5	7.2	7.6
ENSG00000182220	10.0	8.5	10.1	8.6	9.5	10.0	9.7	9.7	10.3	10.1	10.4	10.4	9.8
ENSG00000108298	10.0	10.0	6.3	7.0	7.7	9.8	8.5	9.7	10.1	9.2	9.0	7.8	8.8
ENSG00000126261	10.0	9.5	6.3	8.2	8.2	8.6	9.0	8.7	8.3	8.7	9.3	8.3	8.6
ENSG00000082898	10.0	9.8	7.8	8.6	9.4	9.2	10.0	9.5	9.0	9.4	10.3	10.1	9.4
ENSG00000112245	10.0	7.4	5.9	9.7	6.5	8.3	9.0	8.0	8.3	7.9	9.3	8.8	8.3
ENSG00000111142	10.0	10.5	8.8	8.2	9.5	9.8	9.6	9.6	9.5	9.6	9.1	9.6	9.5
ENSG00000108953	10.0	9.7	9.5	7.3	10.4	10.6	9.5	10.6	10.1	10.6	9.4	7.8	9.6
ENSG00000135046	12.5	1.1	8.9	1.9	10.8	11.5	6.2	11.2	11.9	11.9	11.1	5.0	8.7
ENSG00000109971	12.2	12.2	9.4	8.8	12.3	12.6	11.9	12.1	12.8	12.3	11.8	12.0	11.7
ENSG00000111716	12.1	10.9	11.0	1.5	11.2	11.0	2.3	11.0	13.2	10.0	10.6	1.4	8.9

ensembl_gene	A549.x	Daudi.x	HBEC3.KT.x	HepG2.x	HME.1.x	HUVEC.x	MCF.7.x	RPE.1.x	RPTEC.x	TIME.x	U.2.OS.x	T47.D.x
ENSG00000269028	0.9	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ENSG00000167996	2.2	2.2	0.9	1.8	0.2	0.3	0.7	0.2	0.4	1.0	0.3	1.8
ENSG00000161970	1.7	0.1	1.1	0.1	1.7	0.0	1.9	1.4	0.5	1.1	0.9	1.6
ENSG00000108518	1.4	0.0	1.0	1.0	1.4	1.0	0.8	1.4	1.2	0.8	1.3	0.7
ENSG00000177954	1.3	1.0	1.0	1.1	0.9	0.9	1.0	0.9	1.1	1.0	0.9	1.0
ENSG00000111640	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.1	1.0	1.1	1.1
ENSG00000067225	1.1	1.1	0.7	0.8	0.9	1.1	1.0	1.1	1.2	1.1	1.0	0.9
ENSG00000182718	1.2	1.1	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.0
ENSG00000205426	1.1	1.0	0.8	0.9	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.1
ENSG00000184009	1.2	0.9	0.7	1.2	0.8	1.0	1.1	1.0	1.0	1.0	1.1	1.1
ENSG00000213741	1.1	1.1	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ENSG00000181019	1.0	1.0	1.0	0.8	1.1	1.1	1.0	1.1	1.0	1.1	1.0	0.8
ENSG00000135046	1.4	0.1	1.0	0.2	1.2	1.3	0.7	1.3	1.4	1.4	1.3	0.6
ENSG00000109971	1.0	1.0	0.8	0.8	1.1	1.1	1.0	1.0	1.1	1.1	1.0	1.0
ENSG00000111716	1.4	1.2	1.2	0.2	1.3	1.2	0.3	1.2	1.5	1.1	1.2	0.2

Pearson correlations comparing Relative Log Expression (RLE) for TempO-seq data vs RLE RNA-seq data made cell line differences more pronounced

x = TempO-seq data

y = RNA-seq data

	A549.x	Daudi.x	HBEC3.KT.x	HepG2.x	HME.1.x	HUVEC.x	MCF.7.x	RPE.1.x	RPTEC.x	TIME.x	U.2.OS.x	T47.D.x	A549.y	Daudi.y	HBEC3.KT.y	HepG2.y	HME.1.y	HUVEC.y	MCF.7.y	RPE.1.y	RPTEC.y	TIME.y	U.2.OS.y	T47.D.y
A549.x	1.00	-0.12	-0.09	-0.08	-0.07	-0.09	-0.04	-0.02	-0.03	-0.07	-0.10	-0.09	0.49	-0.09	-0.03	-0.03	-0.02	-0.05	0.00	-0.03	-0.02	-0.06	-0.05	-0.05
Daudi.x	-0.12	1.00	-0.12	-0.11	-0.13	-0.11	-0.12	-0.13	-0.13	-0.09	-0.14	-0.12	-0.10	0.66	-0.10	-0.09	-0.08	-0.10	-0.06	-0.10	-0.10	-0.06	-0.05	-0.07
HBEC3.KT.x	-0.09	-0.12	1.00	-0.10	0.11	-0.11	-0.07	-0.08	-0.06	-0.09	-0.14	-0.10	-0.01	-0.06	0.37	-0.05	0.15	-0.06	0.01	0.00	-0.01	-0.03	-0.09	0.01
HepG2.x	-0.08	-0.11	-0.10	1.00	-0.14	-0.14	-0.12	-0.15	-0.11	-0.15	-0.15	-0.10	-0.03	-0.07	-0.11	0.72	-0.11	-0.11	-0.04	-0.10	-0.09	-0.09	-0.10	-0.09
HME.1.x	-0.07	-0.13	0.11	-0.14	1.00	-0.03	-0.10	0.07	-0.07	0.03	-0.13	-0.15	-0.03	-0.07	0.22	-0.10	0.40	0.03	-0.05	0.07	-0.04	0.03	-0.08	-0.09
HUVEC.x	-0.09	-0.11	-0.11	-0.14	-0.03	1.00	-0.14	0.01	-0.09	0.30	-0.15	-0.17	-0.07	-0.07	-0.02	-0.11	0.01	0.58	-0.08	0.03	-0.06	0.30	-0.11	-0.10
MCF.7.x	-0.04	-0.12	-0.07	-0.12	-0.10	-0.14	1.00	-0.11	-0.06	-0.14	-0.12	0.10	0.03	-0.07	0.00	-0.07	-0.06	-0.11	0.42	-0.07	-0.04	-0.09	-0.07	0.15
RPE.1.x	-0.02	-0.13	-0.08	-0.15	0.07	0.01	-0.11	1.00	-0.03	0.05	-0.09	-0.14	-0.03	-0.10	0.04	-0.11	0.12	0.06	-0.06	0.46	0.00	0.05	-0.07	-0.11
RPTEC.x	-0.03	-0.13	-0.06	-0.11	-0.07	-0.09	-0.06	-0.03	1.00	-0.07	-0.14	-0.08	-0.01	-0.11	0.00	-0.08	-0.02	-0.05	-0.06	0.01	0.69	-0.07	-0.11	-0.06
TIME.x	-0.07	-0.09	-0.09	-0.15	0.03	0.30	-0.14	0.05	-0.07	1.00	-0.17	-0.19	-0.05	-0.05	0.04	-0.12	0.08	0.38	-0.10	0.05	-0.05	0.47	-0.15	-0.15
U.2.OS.x	-0.10	-0.14	-0.14	-0.15	-0.13	-0.15	-0.12	-0.09	-0.14	-0.17	1.00	-0.11	-0.06	-0.11	-0.13	-0.13	-0.12	-0.15	-0.08	-0.04	-0.12	-0.12	0.64	-0.08
T47.D.x	-0.09	-0.12	-0.10	-0.10	-0.15	-0.17	0.10	-0.14	-0.08	-0.19	-0.11	1.00	-0.03	-0.06	-0.08	-0.08	-0.11	-0.13	0.12	-0.06	-0.06	-0.09	-0.07	0.56
A549.y	0.49	-0.10	-0.01	-0.03	-0.03	-0.07	0.03	-0.03	-0.01	-0.05	-0.06	-0.03	1.00	-0.14	-0.06	-0.06	-0.07	-0.10	-0.03	-0.06	-0.04	-0.10	-0.13	-0.08
Daudi.y	-0.09	0.66	-0.06	-0.07	-0.07	-0.07	-0.07	-0.10	-0.11	-0.05	-0.11	-0.06	-0.14	1.00	-0.12	-0.10	-0.11	-0.12	-0.10	-0.13	-0.13	-0.08	-0.14	-0.12
HBEC3.KT.y	-0.03	-0.10	0.37	-0.11	0.22	-0.02	0.00	0.04	0.00	0.04	-0.13	-0.08	-0.06	-0.12	1.00	-0.12	0.27	0.02	-0.07	-0.02	0.01	-0.03	-0.21	-0.09
HepG2.y	-0.03	-0.09	-0.05	0.72	-0.10	-0.11	-0.07	-0.11	-0.08	-0.12	-0.13	-0.08	-0.06	-0.10	-0.12	1.00	-0.14	-0.13	-0.09	-0.14	-0.11	-0.11	-0.16	-0.11
HME.1.y	-0.02	-0.08	0.15	-0.11	0.40	0.01	-0.06	0.12	-0.02	0.08	-0.12	-0.11	-0.07	-0.11	0.27	-0.14	1.00	0.04	-0.10	0.05	-0.02	-0.01	-0.18	-0.14
HUVEC.y	-0.05	-0.10	-0.06	-0.11	0.03	0.58	-0.11	0.06	-0.05	0.38	-0.15	-0.13	-0.10	-0.12	0.02	-0.13	0.04	1.00	-0.14	0.03	-0.04	0.28	-0.20	-0.17
MCF.7.y	0.00	-0.06	0.01	-0.04	-0.05	-0.08	0.42	-0.06	-0.06	-0.10	-0.08	0.12	-0.03	-0.10	-0.07	-0.09	-0.10	-0.14	1.00	-0.13	-0.11	-0.14	-0.14	0.09
RPE.1.y	-0.03	-0.10	0.00	-0.10	0.07	0.03	-0.07	0.46	0.01	0.05	-0.04	-0.06	-0.06	-0.13	-0.02	-0.14	0.05	0.03	-0.13	1.00	0.00	0.01	-0.13	-0.14
RPTEC.y	-0.02	-0.10	-0.01	-0.09	-0.04	-0.06	-0.04	0.00	0.69	-0.05	-0.12	-0.06	-0.04	-0.13	0.01	-0.11	-0.02	-0.04	-0.11	0.00	1.00	-0.09	-0.17	-0.10
TIME.y	-0.06	-0.06	-0.03	-0.09	0.03	0.30	-0.09	0.05	-0.07	0.47	-0.12	-0.09	-0.10	-0.08	-0.03	-0.11	-0.01	0.28	-0.14	0.01	-0.09	1.00	-0.19	-0.16
U.2.OS.y	-0.05	-0.05	-0.09	-0.10	-0.08	-0.11	-0.07	-0.07	-0.11	-0.15	0.64	-0.07	-0.13	-0.14	-0.21	-0.16	-0.18	-0.20	-0.14	-0.13	-0.17	-0.19	1.00	-0.13
T47.D.y	-0.05	-0.07	0.01	-0.09	-0.09	-0.10	0.15	-0.11	-0.06	-0.15	-0.08	0.56	-0.08	-0.12	-0.09	-0.11	-0.14	-0.17	0.09	-0.14	-0.10	-0.16	-0.13	1.00

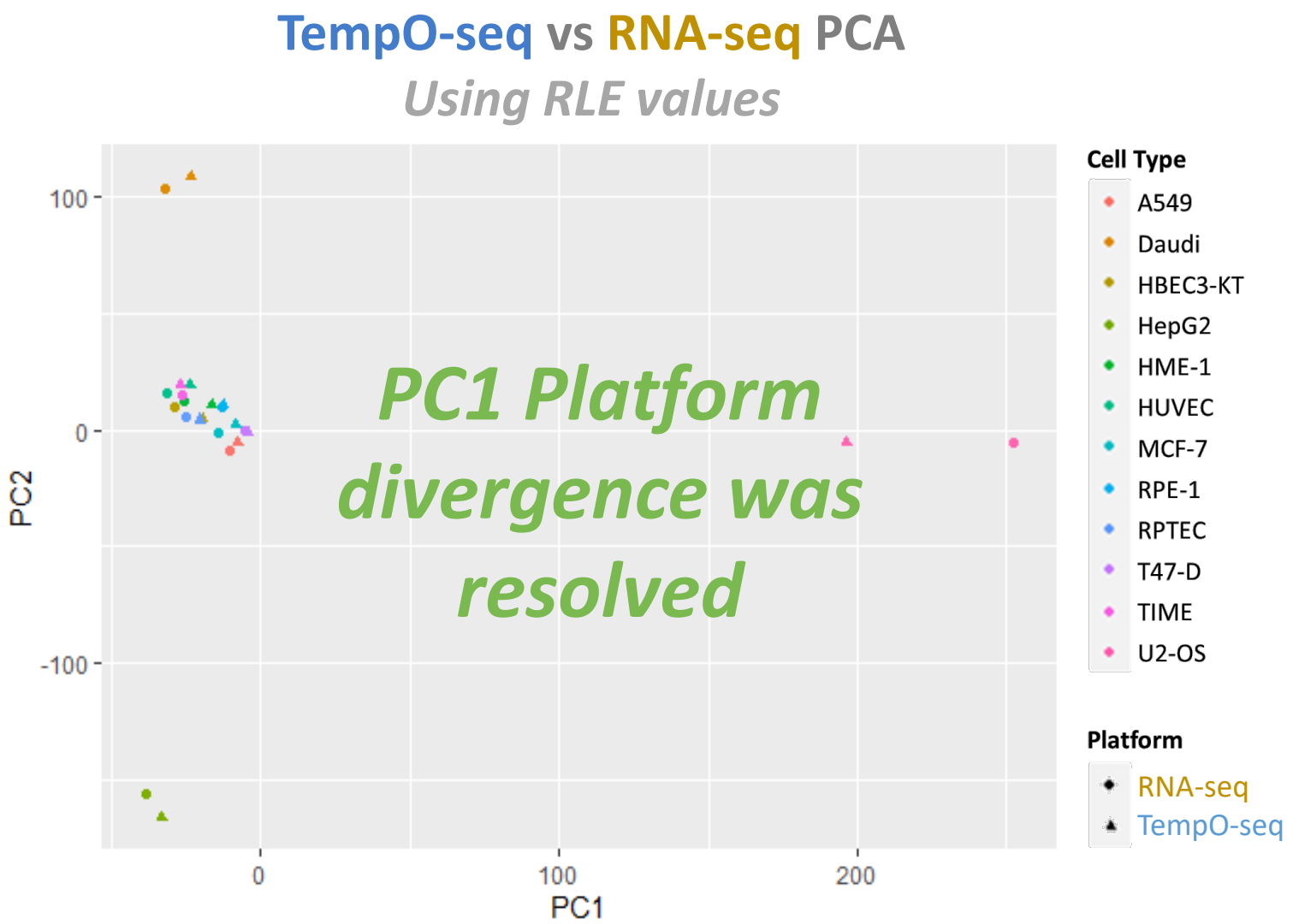
x = TempO-seq

y = RNA-seq

Pearson correlations for different cell types intra-platform: generally around 0.07 (0.00-0.19)

Pearson correlation for matching cell types inter-platform: 0.54 (0.37-0.72)

Method b) Normalizing the data by calculating relative log expression also resolved the platform divergence along PC1



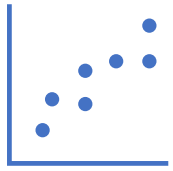
Conclusions and Future Work

RNA-seq and TempO-seq showed consistent gene expression findings

Similar data distributions

High pearson correlations

And after normalization, the data grouped by cell type and not by technology platform



This work can help increase confidence in using TempO-seq transcriptomic data

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

Future work: Determine whether RNA-seq and TempO-seq data can be combined in chemical studies

Need data for comparing chemical perturbation data with both platforms to see if the data sets can be combined. This work using baseline expression data is a good foundation for such work.

Many thanks to the high-throughput transcriptomics team for their input on these methods!

Special thanks to:

- Richard Judson
- Joshua Harrill
- Logan Everett
- Woody Setzer
- Imran Shah
- Joseph Bundy
- Bryant Chambers
- Sarah Davidson

Contact information

Email: taylor.laura@epa.gov

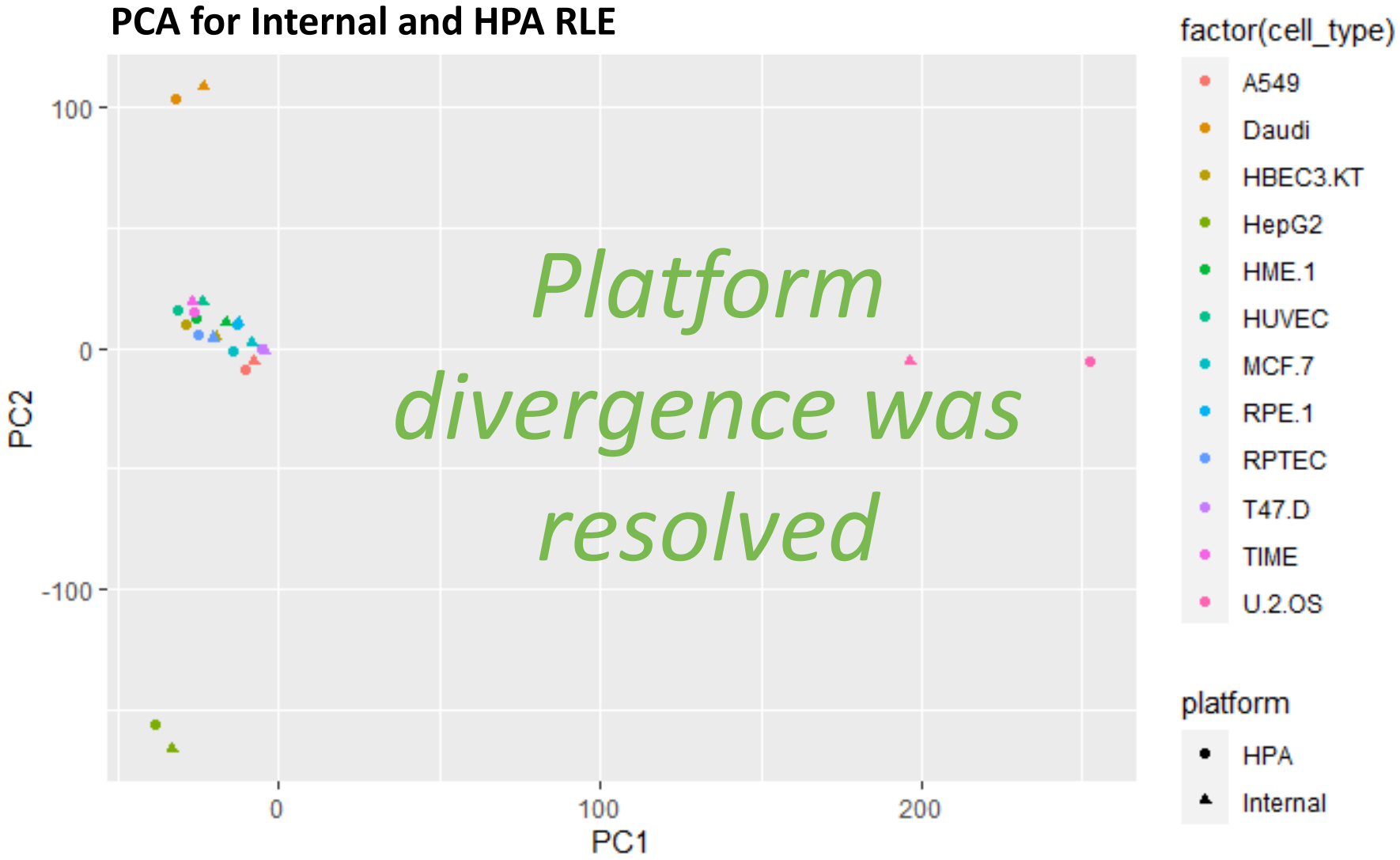
Work phone: 919-541-1060

Extra Slides

Note for application of these methods

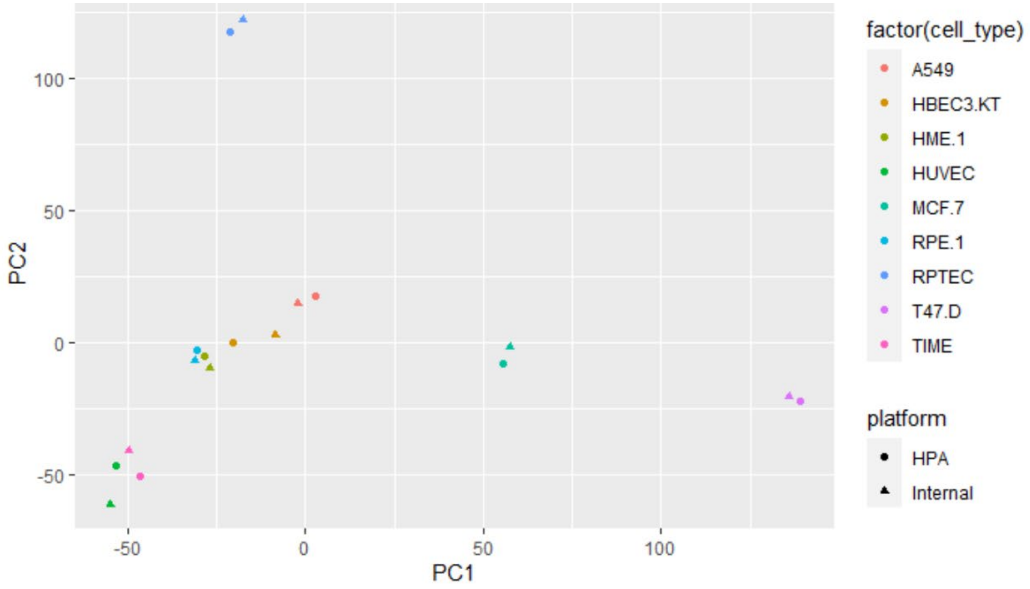
- When performing RLE normalization, gene expression levels can then only be compared across different cell types (not within a single cell type)
- Keeping only the genes that were the most replicable across the different cell types enables comparison to still be done both within a single cell type and/or across cell types, but this list of genes needs to be replicated or a new list can be generated for each study's analysis based on their own data

PCA on relative log expression (RLE) compared to average across cell types

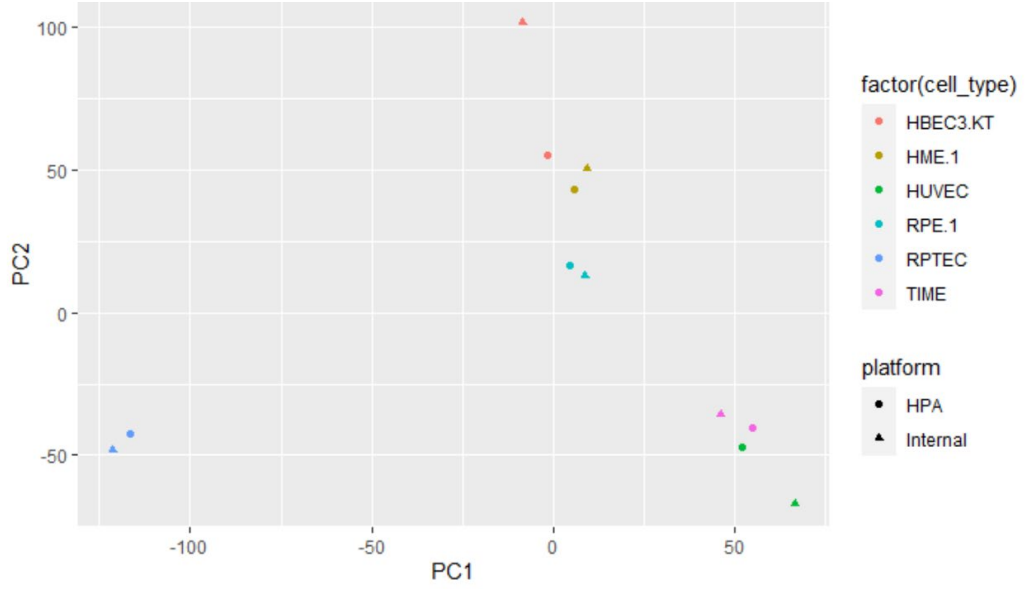


Repeating RLE PCA with a subset of cell types

PCA for Internal and HPA RLE, no Daudi/HepG2/U2-OS

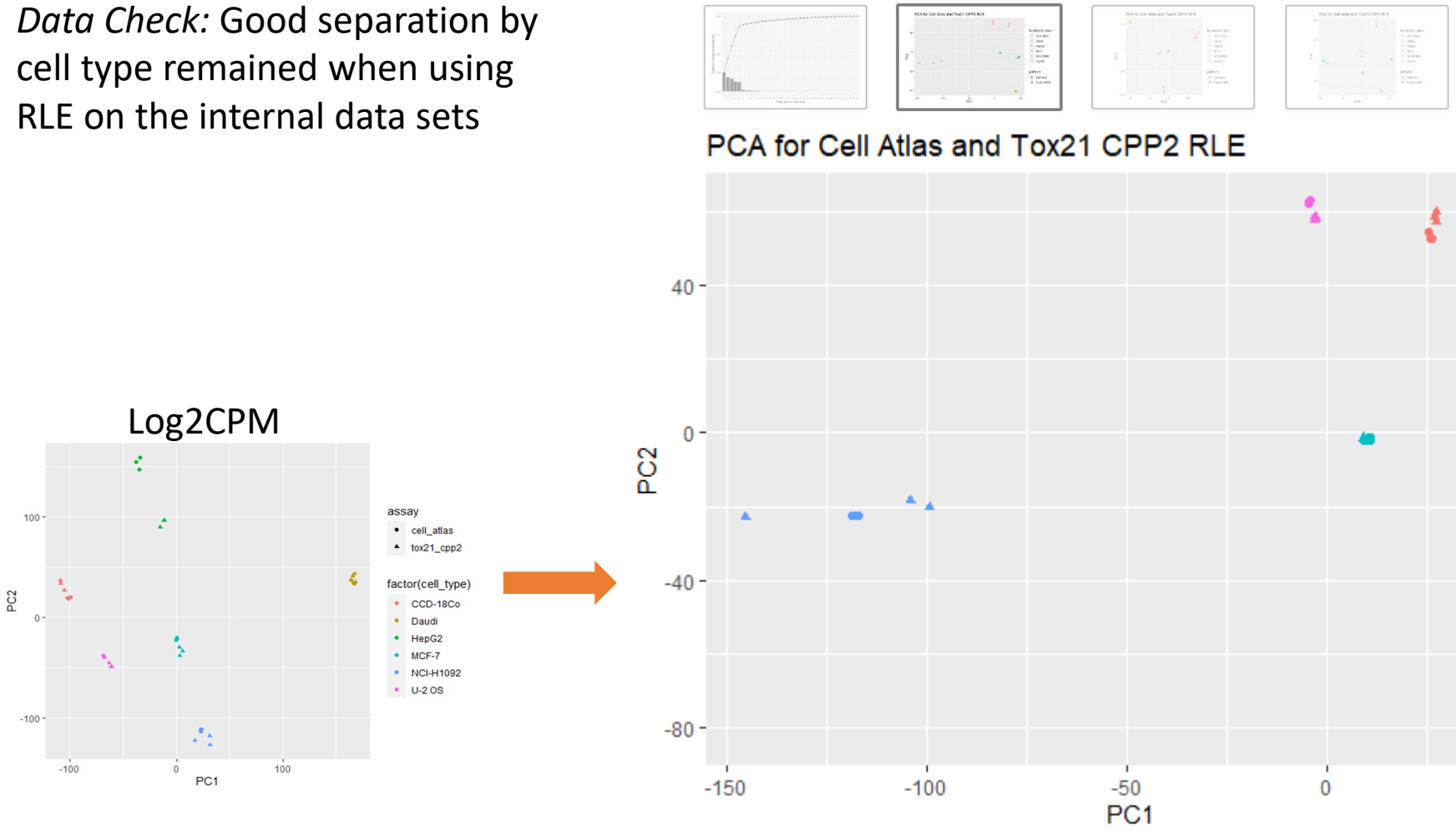


PCA for Internal and HPA RLE, no cancer cell lines



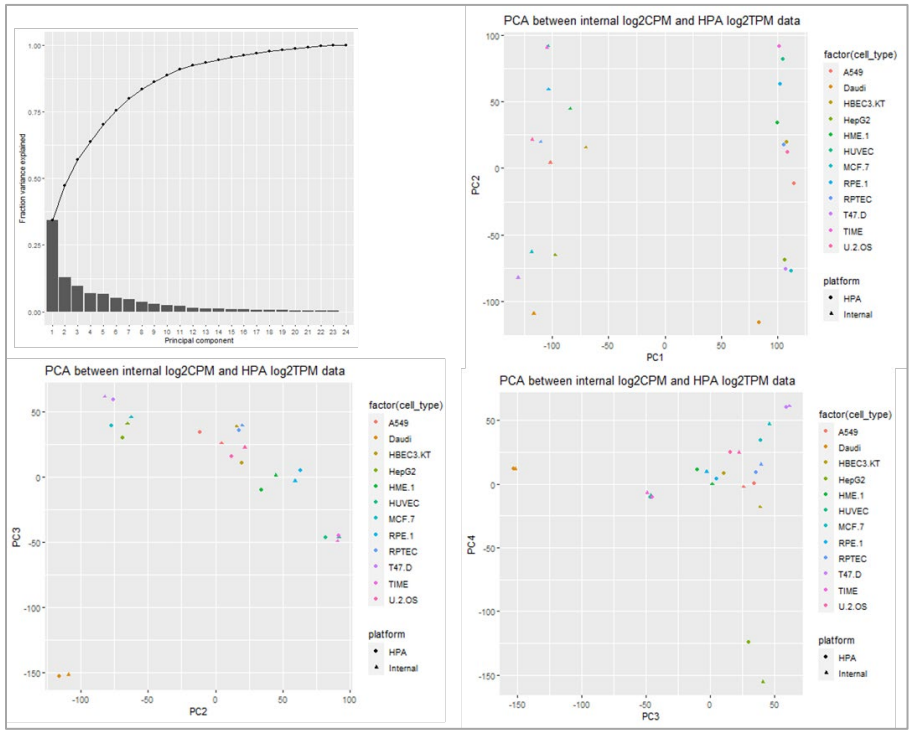
Applying RLE normalization to the Internal Data

Data Check: Good separation by cell type remained when using RLE on the internal data sets

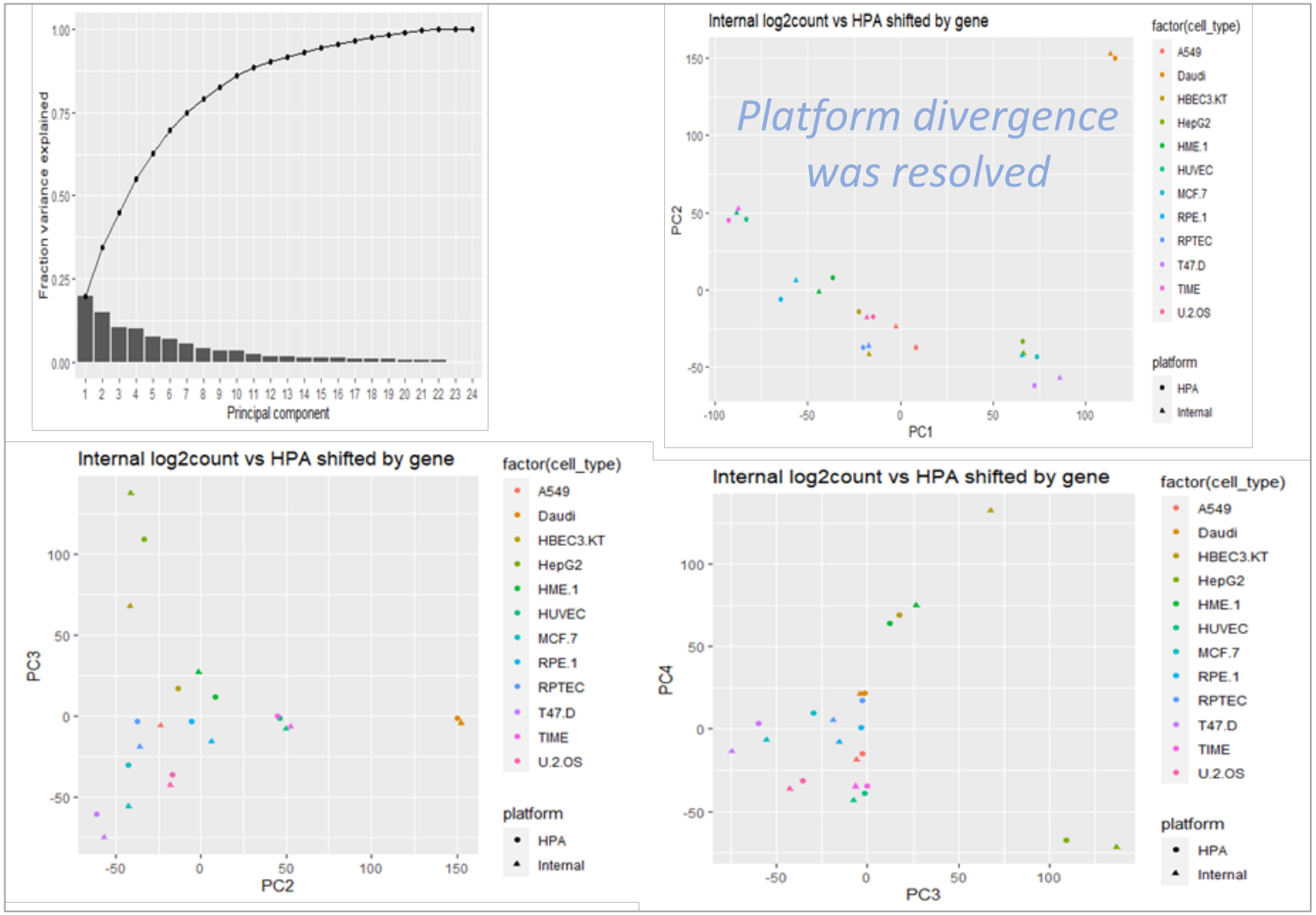


PCA: Normalization

Before



After normalizing by gene



Normalization method: calculated internal log2count minus HPA log2count for all cell types, took average difference across all cell types, then added that average difference to the HPA values

RNA-seq Method

(more established)

Key features [provided by Illumina]

Prep

2.5 days/6 hours hands-on time

TruSeq DNA Exome

A cost-effective library preparation and exome enrichment solution with industry-leading accuracy.

Illumina Experiment Manager
BaseSpace Clarity LIMS

Sequence

5 days/30 minutes hands-on

HiSeq SBS Kit V4

Ready-to-load reagents for sequencing by synthesis on enabled HiSeq sequencing systems.

HiSeq PE Cluster Kit v4 cBot

Reagents for paired-end cluster generation on the cBot cluster amplification system.

Analyze

2–3 hours

Enrichment BaseSpace App

Rapid alignment and variant detection for small, structural, and copy number variant calling, variant annotation, and enrichment metrics calculation.

Browse Sample Datasets in BaseSpace Sequence Hub

HiSeq 2500 Exome Run Data
HiSeq 2500 Exome Project Data

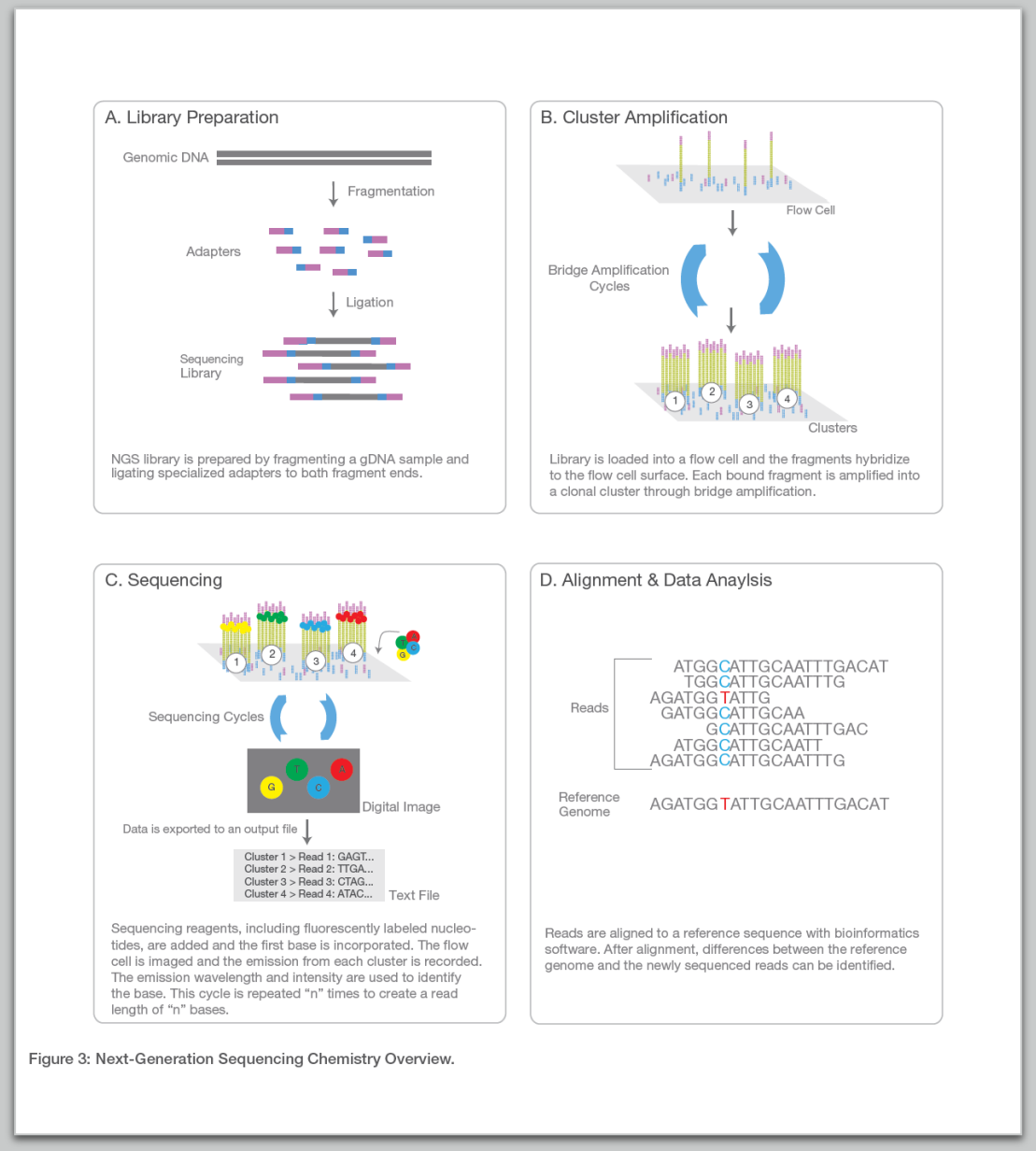
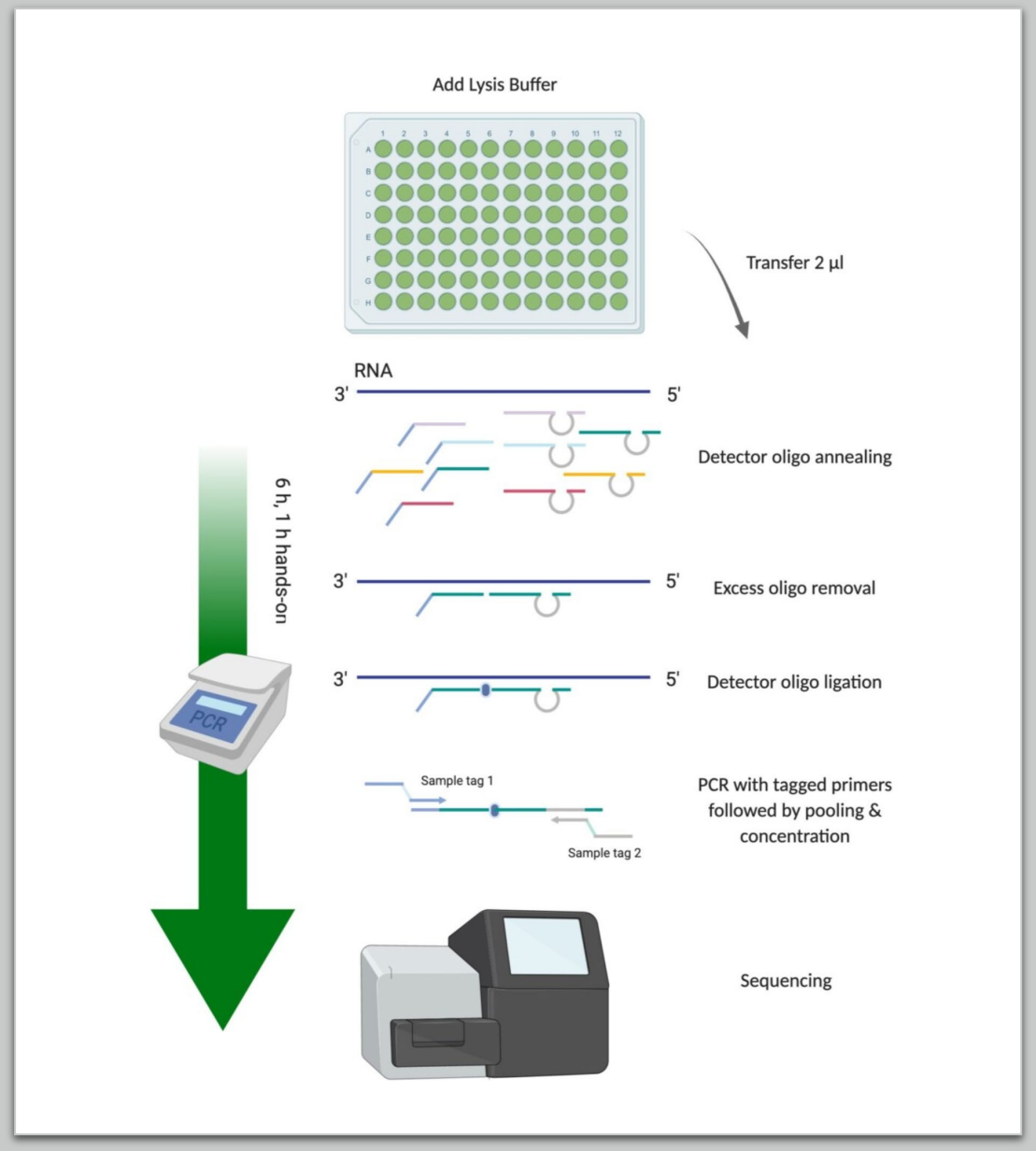


Figure 3: Next-Generation Sequencing Chemistry Overview.

TempO-seq Method

Key features [provided by BioSpyder]

- input down to 10 pg with single-base specificity
- no RNA extraction, cDNA conversion or preamplification
- no proprietary instrumentation required: standard PCR cycler or microplate incubator is all you need
- maximize sequence capacity due to short barcode reads: up to 6,144 samples in one run
- “off the shelf” whole transcriptome or surrogate panels – or fully customizable panels assessing any number of isoforms, gene fusions, mutations, or SNPs in any species

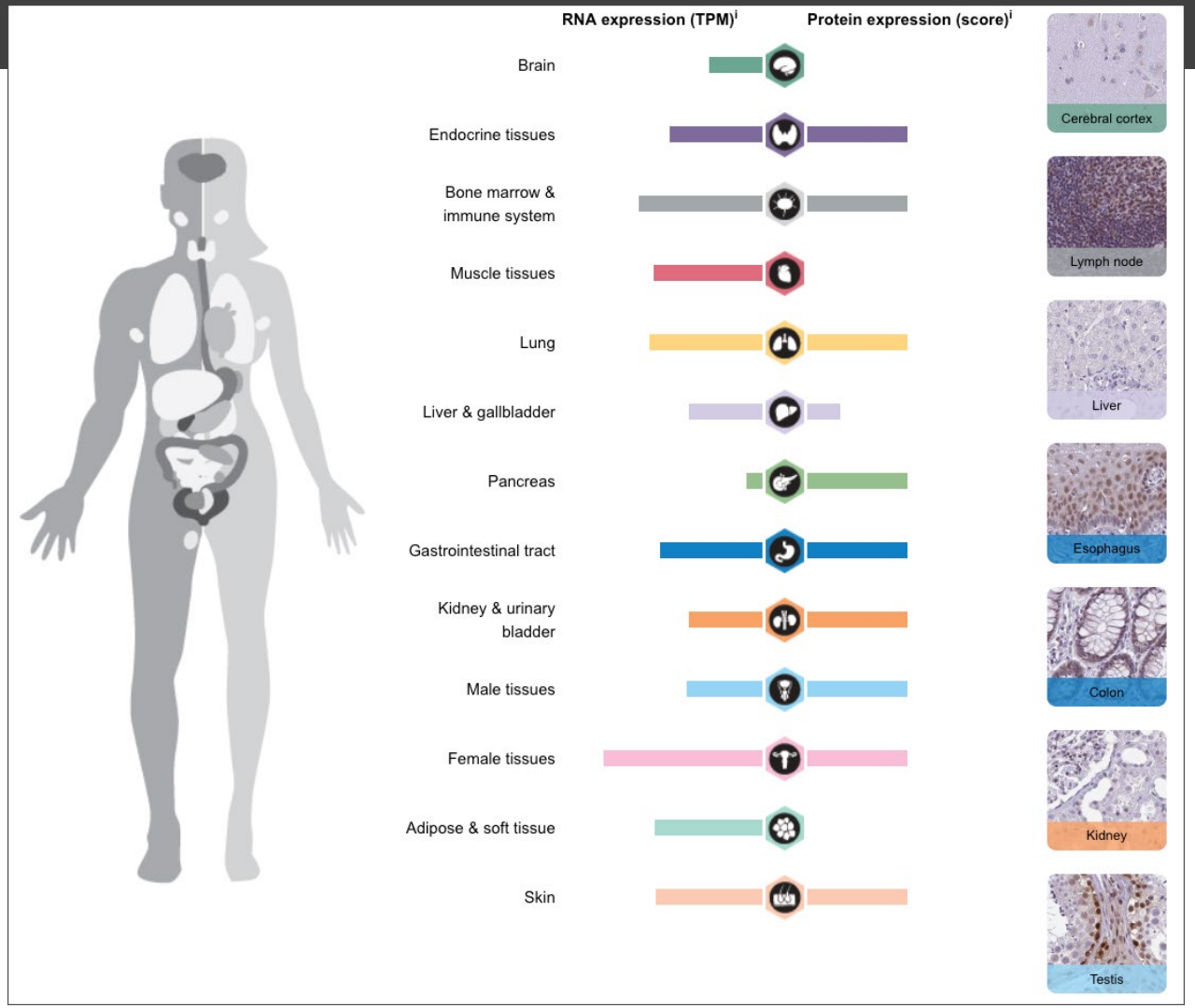


US EPA Internal Data: Cell Atlas and Tox21 CPP2

- Both of these **TempO-seq data** sets were generated at the EPA in 2018-2019
- Clinton Willis performed sample collection for both data sets
- Cells came from the same cryostocks

THE HUMAN PROTEIN ATLAS

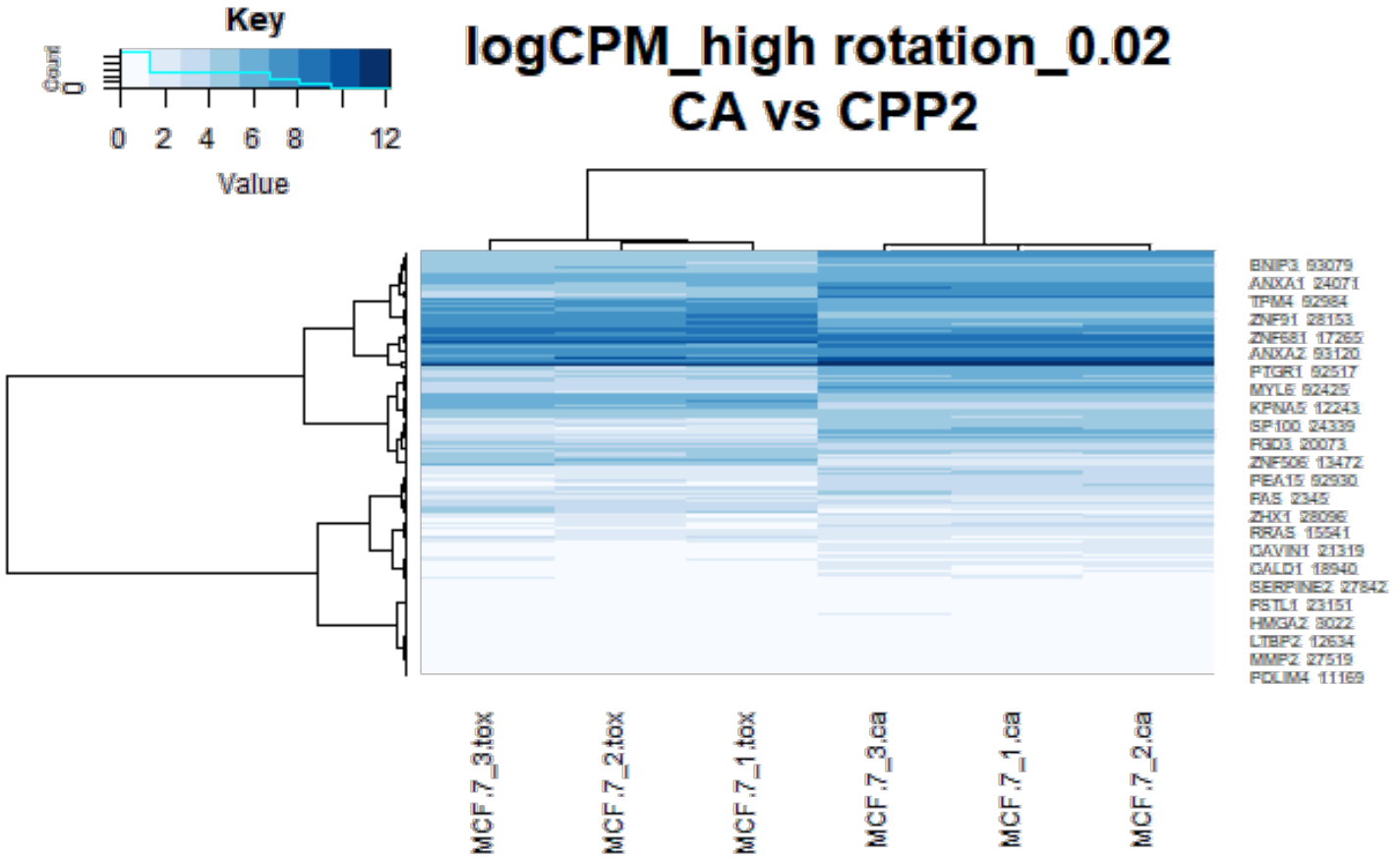
- Publicly available **RNA-seq** and protein expression data for many tissues of the human body
- Used the Illumina HiSeq2500 platform for sequencing at approximately 20 million reads depth
- More details: *HPA is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology*



Common Cell Types: Cell Atlas and Tox21 CPP2

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™)

For genes driving PC1 across the different cell types, there was consistent gene expression within the replicates for each individual cell lines. This further shows that the PCA grouping was driven by cell type differences and not by differences in the two TempO-seq data sets.



Cell line information

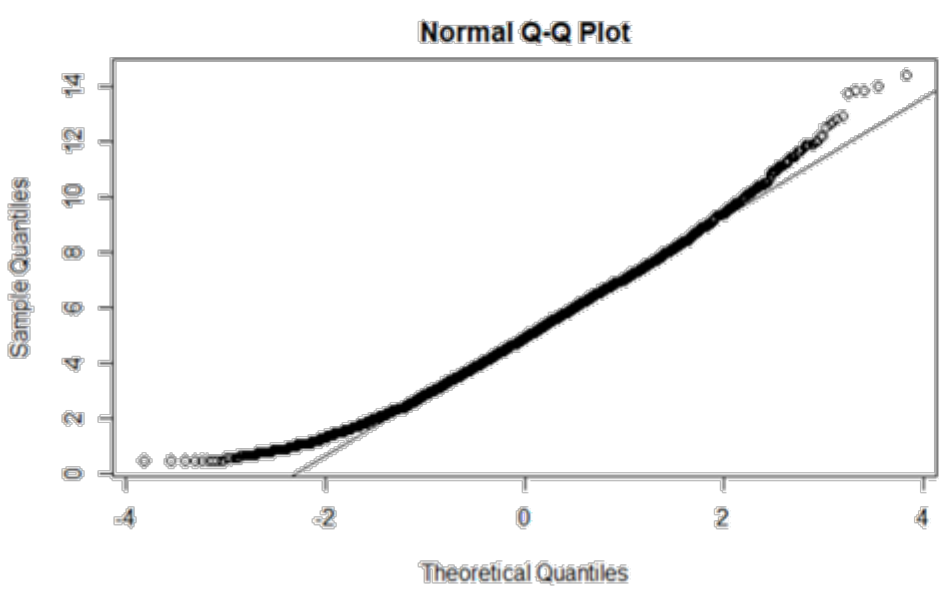
Overlap between internal TempO-seq and HPA RNA-seq cell lines

Internal Data_set	Cell Type	Vendor	Vendor Part Number	Tissue	Disease State	Growth Mode	Media Formulation (**All use 1% P.S.G. for antibiotic**)
CellAtlas, Tox21CPP2	A549	ATCC	CCL-185	Lung	Lung carcinoma	Adherent	DMEM + 10% HI-FBS
CellAtlas	Daudi	ATCC	CCL-213	Haematopoietic and lymphoid	Peripheral blood B lymphoblast cells	Suspension	RPMI-1640 Medium + 10% HI-FBS
CellAtlas, Tox21CPP2	HepG2	ATCC	HB-8065	Liver	Hepatocellular carcinoma	Adherent	DMEM + 10% HI-FBS
CellAtlas, Tox21CPP2	MCF-7	ATCC	HTB-22	Breast	Adenocarcinoma	Adherent	DMEM + 10% HI-FBS
CellAtlas, Tox21CPP2	U-2 OS	ATCC	HTB-96	Bone	Osteosarcoma	Adherent	DMEM + 10% HI-FBS
Tox21CPP2	T47-D	ATCC	HTB-133	Breast	Ductal carcinoma	Adherent	RPMI-1640 + 10% HI-FBS + 0.2 U/mL Insulin (1.03 mL of stock)
CellAtlas	HBEC3-KT	ATCC	CRL-4051	Lung, bronchial	hTERT Immortalized	Adherent	Airway Epithelial Cell Basal Medium + Bronchial Epithelial Cell Growth Kit
CellAtlas	HME-1	ATCC	CRL-4010	Breast, mammary gland	hTERT Immortalized	Adherent	SABM Basal Medium + SAGM SingleQuots BulletKit
CellAtlas	HUVEC/TERT-2	ATCC	CRL-4053	Umbilical vascular endothelium	hTERT Immortalized	Adherent	Vascular Cell Basal Medium + Vascular Endothelial Cell Growth Kit (VEGF)
CellAtlas	RPE-1	ATCC	CRL-4000	Retina, eye	hTERT Immortalized	Adherent	DMEM:F12 + 10% HI-FBS + 0.01 mg/mL Hygromycin B
CellAtlas	RPTEC/TERT-1	ATCC	CRL-4031	Renal cortex, proximal tubes	hTERT Immortalized	Adherent	DMEM:F12 + hTERT RPTEC Growth Kit
CellAtlas	TIME	ATCC	CRL-4025	Foreskin, dermal microvascular endothelium	hTERT Immortalized	Adherent	Vascular Cell Basal Medium + Microvascular Endothelial Cell Growth Kit (VEGF) + 12.5 µg/mL Blasticidine

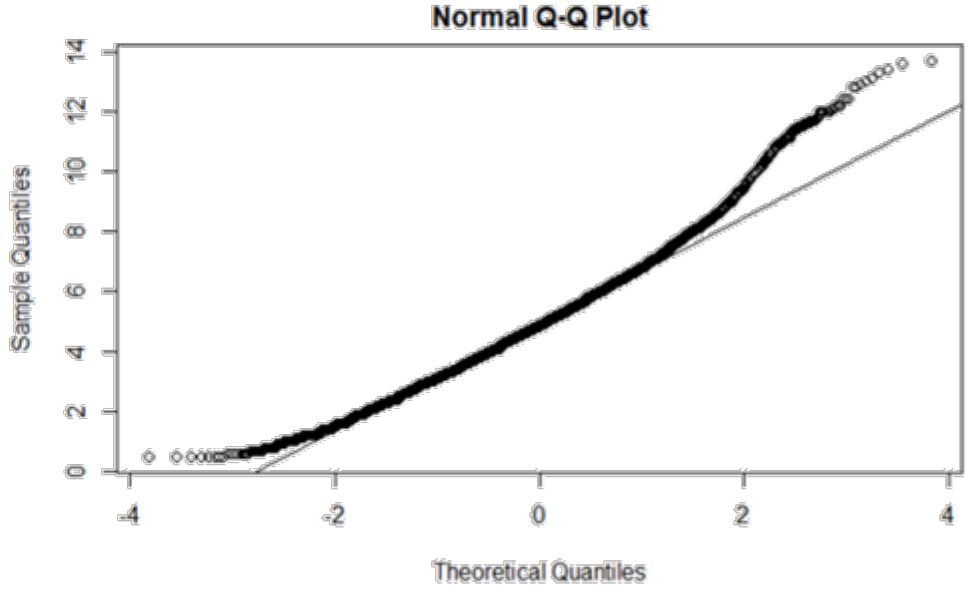
Made QQ plots

The QQ plots are better at the tails for the internal TempO-seq data than the HPA RNA-seq data for log2count >0.5. These two plots are representative of all twelve cell types:

Internal data (MCF7)



HPA data (MCF7)



Because of this, I tried removing points at the tails of the distribution and repeated PCA, but there was still a clear PC1 divergence.

Method b

Relative Log Expression (RLE)

Calculates the log expression level relative to a reference value

Example:		
Reference = 4 logCPM	Sample = 8 logCPM	RLE = 8/4 = 2