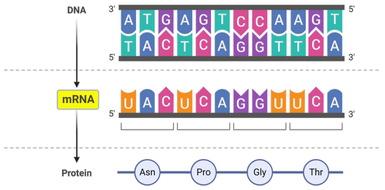
Comparing TempO-seq and RNA-seq mRNA data sets: a case study

Laura Word-Taylor, Logan Everett, Joshua Harrill, Imran Shah, Richard Judson
U.S. Environmental Protection Agency, Center for Computational Toxicology and Exposure



Increasing scientific confidence when using TempO-seq data for conducting transcriptomics research

Introduction:
There are multiple technological platforms available for quantifying mRNA levels to use for transcriptomics studies. With the increase in mRNA expression data generated using TempO-seq, it is important to determine whether TempO-seq and RNA-seq data are comparable. A previous study using rat samples by Bushel et al in 2018 demonstrated that TempO-seq and RNA-seq showed platform differences but mechanism of action for exposures grouped by treatment instead of by platform. To further that research, work comparing human samples is still needed. Here, we describe a workflow process for evaluating whether and how mRNA data sets can be combined when comparing and/or aggregating data generated by TempO-seq versus RNA-seq.

Methods:
Two different EPA-generated TempO-seq data sets for baseline expression were compared to each other using principal component analysis (PCA) for six overlapping cell types (CCD-18Co, Daudi, HepG2, MCF-7, NCI-H1092, and U-2 OS). For the four overlapping cell types within both EPA-generated TempO-seq data sets, the average of all replicates per cell type were used for this comparison to RNA-seq. The combined TempO-seq data was then compared to baseline RNA-seq data from the Human Atlas Project (HPA) for 12 cell types (A549, Daudi, HBEC3-KT, HepG2, HME-1, HUVEC, MCF-7, RPE-1, RPTEC, TIME, U-2 OS, and T47-D) using PCA.

Results:
The PCA showed that the TempO-seq data was reproducible and that the two data sets could be combined. Statistical testing using PCA on TempO-seq versus RNA-seq mRNA expression data for the 19,119 overlapping genes showed that there was a clear platform divergence pattern within the first principal component (PC1) for all cell types evaluated. This meant that these TempO-seq and RNA-seq data should not be combined without further steps. Removing genes with large average differences in expression levels between the technological platforms was found to be effective in resolving platform divergence. Normalizing the data by calculating the relative log expression (RLE) compared to the average expression level across cell types in each platform also removed the platform divergence observed in PC1.

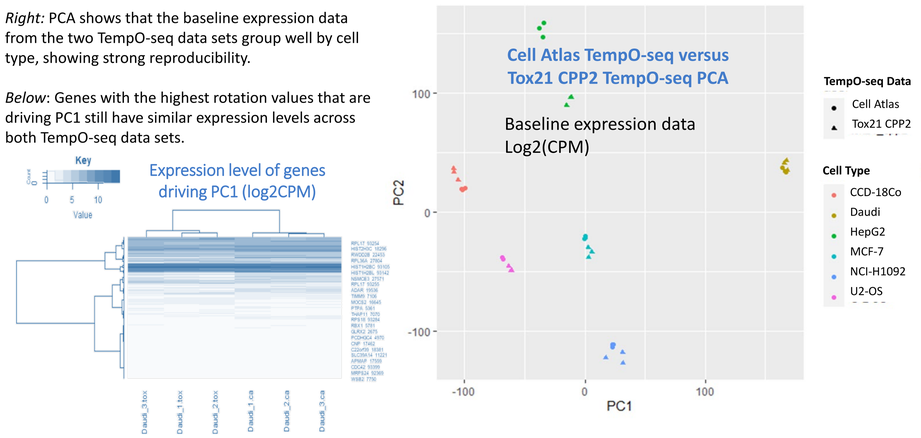
The views expressed are those of the authors and do not necessarily represent the views or the policies of the U.S. EPA.

Two EPA data sets on baseline expression showed that TempO-seq was reproducible and combinable

TempO-seq data was reproducible, as shown by principal component analysis (PCA)

Right: PCA shows that the baseline expression data from the two TempO-seq data sets group well by cell type, showing strong reproducibility.

Below: Genes with the highest rotation values that are driving PC1 still have similar expression levels across both TempO-seq data sets.



Conclusions and Future Work

TempO-seq and RNA-seq show consistent gene expression findings
TempO-seq data was reproducible, and after normalization for TempO-seq vs RNA-seq, the data grouped by cell type and not by technology platform.

This work can help increase confidence in using TempO-seq data
This work helps to validate TempO-seq against the RNA-seq, which is the current gold-standard technique. TempO-seq is less expensive and is easier, using cell lysates instead of RNA purification.

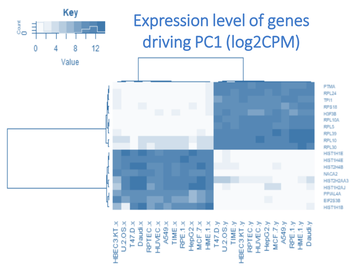
Future work: Determine whether RNA-seq and TempO-seq exposure data can be combined
Need data comparing chemical perturbation data with both platforms to see if the data sets can be combined. This work using baseline expression data is a good foundation.

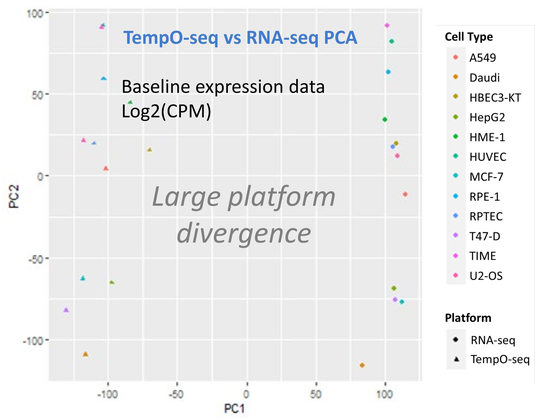
EPA's TempO-seq data compared to Human Protein Atlas' RNA-seq data required normalization steps for both data sets in order to resolve the observed platform divergence before being combinable

EPA's TempO-seq vs HPA's RNA-seq showed a platform divergence

Right: PCA shows that TempO-seq and RNA-seq data show a large platform divergence across principal component 1 (PC1). However, they do group well by cell type across PC2, which is good.

Below: Genes driving PC1 have very different levels of expression in the TempO-seq vs RNA-seq platforms. Many of these genes are histone and ribosomal genes.

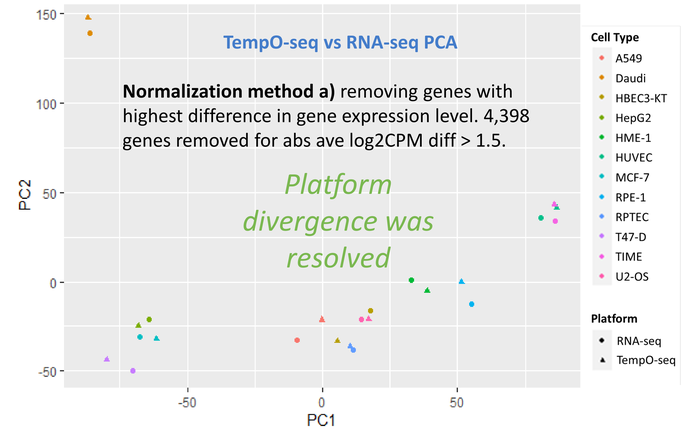




Normalization method a

Normalization method b

After normalizing the data, the platform divergence was resolved



Normalization method b: calculating relative log expression

Relative log expression (RLE) was used as another normalization method. RLE was calculated for each cell type compared to the average expression level across all of the cell types in each data set separately (TempO-seq and RNA-seq) as the reference.

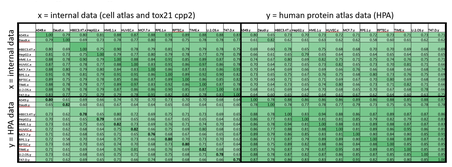
Relative Log Expression (RLE)

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

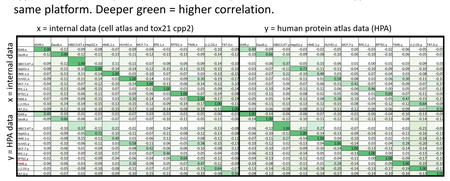
Example:

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

Before normalization: Between the same cell type for TempO-seq vs RNA-seq, the average Pearson correlation was an average of 0.80, which is similar to the correlations between matching cell types across the two different platforms. Deeper green = higher correlation.



After RLE normalization: Pearson correlations were stronger for matching cell types in TempO-seq compared to RNA-seq than for different cell types within the same platform. Deeper green = higher correlation.



The additional normalization method using RLE also removed the platform divergence

PCA on the left after RLE normalization showed a cluster of cell types that were close together. Thus, PCA was repeated without the three most divergent cell types (top right), which were all cancer lines, as well as without any of the cancer lines (bottom right). This improved cell type partitioning. The platform divergence was resolved by RLE in all three scenarios.