White Paper

A Proof-of-concept Exploration of Transcriptome-informed Drug Discovery Using the HTG Transcriptome Panel

Introduction

In recent years there has been significant interest in using transcriptomic profiling in drug discovery and repurposing, to support or gain information about target(s), method of action (MoA), pharmacologic or toxicologic effects, or for elucidation of cellular effects via pathway analysis or gene set enrichments.[1-3] A simple example of using RNA during drug development is the assessment of cytochrome P450 (CYP) family gene expression in hepatocytes as part of early drug metabolism testing.[4] A larger effort by the Connectivity Map (CMap) and Library of Integrated Network-Based Cellular Signatures (LINCS) groups[5,6] seeks to characterize perturbagen (compounds, siRNA, CRISPR) treatments via their resulting RNA expression profiles. Large-scale generation of RNA profiles from cells treated with various perturbagens with known targets/MoA allowed a large database of differential expression signatures to be assembled. This signature database can be queried using new data to find either matching or reversed profiles, potentially indicating on- and off-target effects, MoA, or undesired effects for *e.g.* target identification, drug repurposing, or discovery. Importantly, it is designed to be queried by other gene expression profiling technologies.

HTG EdgeSeq™ is a targeted RNA sequencing technology that couples a nuclease protection assay with next-generation sequencing for rapid and accurate RNA quantification. The recently released HTG Transcriptome Panel (HTP) measures approximately 20,000 protein-coding genes (mRNAs), allowing assessment of the human transcriptome.[7] The unique advantages of the HTG EdgeSeq technology include a 96-well plate format, low sample input requirement, no RNA extraction, and rapid assay and analysis time, and make this an attractive technology for applying transcriptomic profiling to drug design and discovery. RNA profiling of compounds at early design stages ("transcriptome-informed drug design"), coupled with traditional binding assays, could potentially aid in refining candidate molecules to favor desired effects and minimize unfavorable ones.[1-3] Of interest is whether transcriptomic profiling using the HTP can differentiate between closely related compounds to allow the data to be used in such a manner.

In this white paper, we establish the utility of the HTP in transcriptome-informed drug discovery using a proof-of-concept experiment. Cultured cells were treated with a set of compounds that share the same target, and an expression profile was generated from those compound-treated cells using the HTP. The resulting data were used to establish accuracy and reproducibility metrics, generate differential expression profiles, observe the relative effects of different compounds at different doses, determine correlation between these data and the publicly available CMap/LINCS database, and to investigate the relationship between transcriptomic profiles and compound molecular structure.

Experimental Overview and Performance Metrics

This study was designed to generate initial transcriptomic profile data on compound-treated cells. Briefly, twelve compounds, all known to target mTOR (mechanistic target of rapamycin), were selected and used to treat two cell lines (PC3 and HepG2) for 24 hours at high and low doses (in triplicate). All compounds were solubilized in DMSO. Cells were treated at 10 µM and 100 nM; doses were chosen to match those used for these compounds in

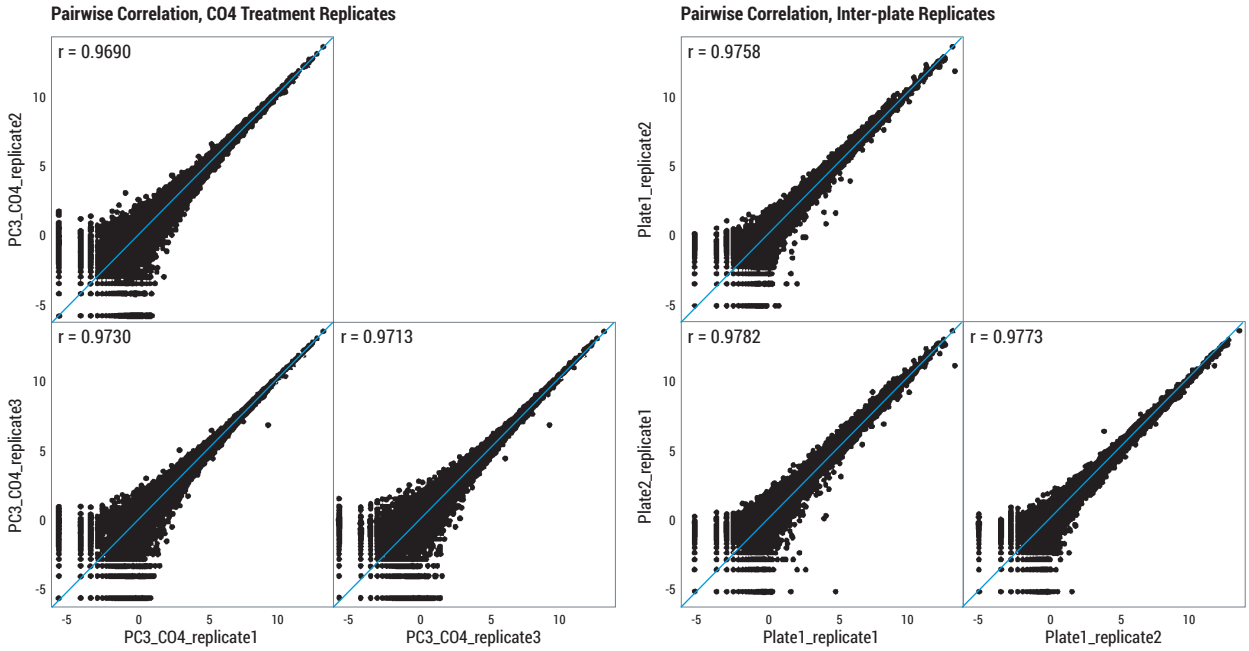


Figure 1. *Replicates (treatment or inter-plate) are highly correlated. Scatterplots show pairwise correlations for technical treatment replicates (left) and inter-plate replicates (right).*

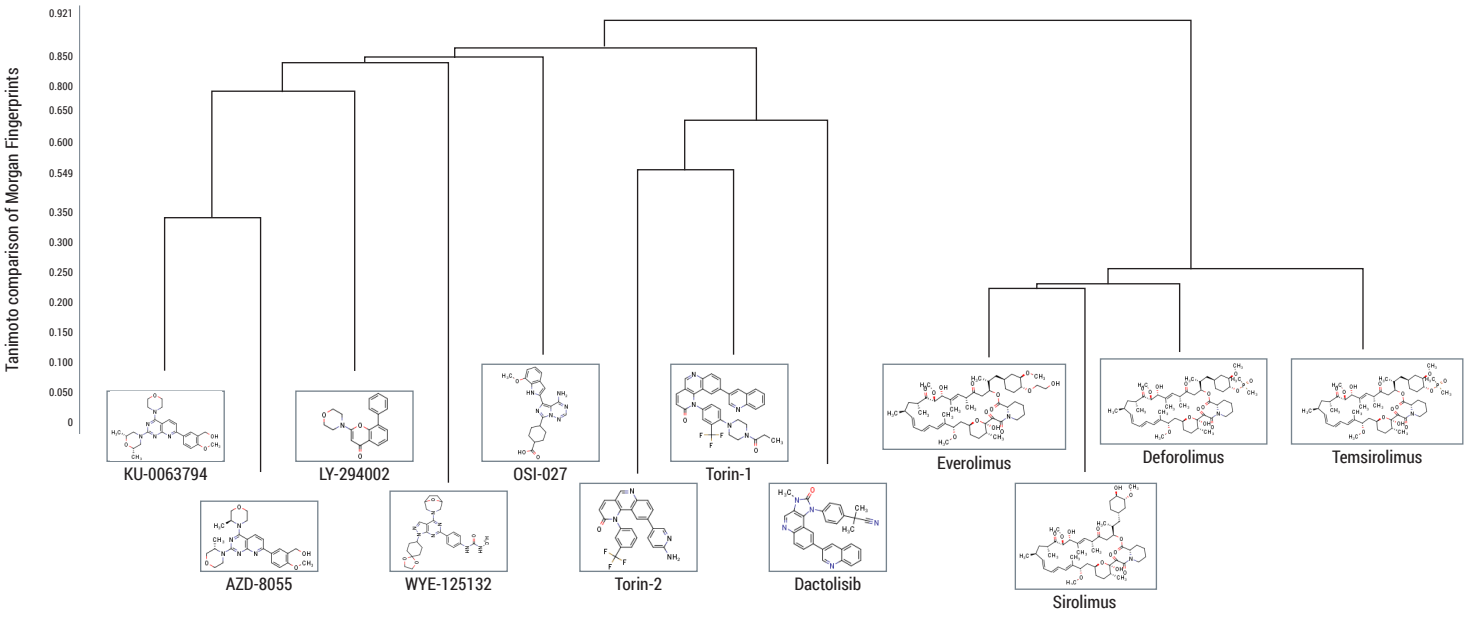


Figure 2. *Dendrogram showing the structural relationship of the 12 compounds used in this experiment. Relationships were calculated based on the Tanimoto distance.*

CMap/LINCS. DMSO alone was used as a control. Two 96-well plates were used for each cell line, and DMSO controls were included on both plates. Following treatment, each well of cells was lysed and ~4000 cells/sample were processed using the HTP. Samples were tagged and sequenced on an Illumina NextSeq 2000 sequencer.[7]

All samples passed quality-control and were used in downstream analyses. Repeatability was evaluated for three types of replicates:

1: Treatment replicates (*i.e.,* individual wells of cells each treated with the same compound and run as independent samples),

2: HTP technical replicates (*i.e.,* the same sample run multiple times), and

3: Inter-plate replicates for the assessment of potential plate effects (DMSO controls).

Pairwise comparisons were performed for each replicate type and the Pearson correlation coefficient calculated. *Figure 1* summarizes the high overall replicate correlation; the median Pearson correlation for all replicates was 0.97.

The strong repeatability and 100% sample pass rate in this experiment suggest that for screening purposes, a single treatment replicate could be used without compromising data quality. Additionally, no significant plate effects were observed, suggesting that plate controls could be kept to a minimum.

Transcriptome Profiles Differentiate Between Compounds and Doses

A goal of this experiment was to determine if closely related compounds (by target and/or chemical structure) can be differentiated by their transcriptomic profiles. To illustrate structural "relatedness", the Morgan fingerprint of each of the twelve compounds utilized was used to generate similarity scores using the Tanimoto equation.[8] *Figure 2* displays the resulting dendrogram. All twelve compounds target mTOR, a protein kinase involved in the regulation of important cellular processes such as proliferation, autophagy, and cell survival.[9] Among the compounds, the most significant separation is between rapamycin and three rapamycin homologs or "rapalogs" on the right and the other mTOR-targeting compounds on the left. The former are allosteric inhibitors of mTOR that bind at the same location; the latter are kinase inhibitors that compete for ATP to inhibit all kinase-dependent functions of mTOR and may also target other phosphoinositide 3-kinases (PIK3 kinases).

A holistic view of the relative differences between the transcriptomic profiles of the treated cells is provided by a principal components analysis (PCA) plot. *Figure 3* shows the PCA results for profiles of PC3 cells treated at the low dose *(left)* and the high dose *(right)* of the twelve compounds. Both plots include the DMSO control. Three conclusions can be derived from these plots. One, there is variation in the profiles from the different compounds, including between very closely-related compounds such as the rapalogs, as shown by the distance between the data

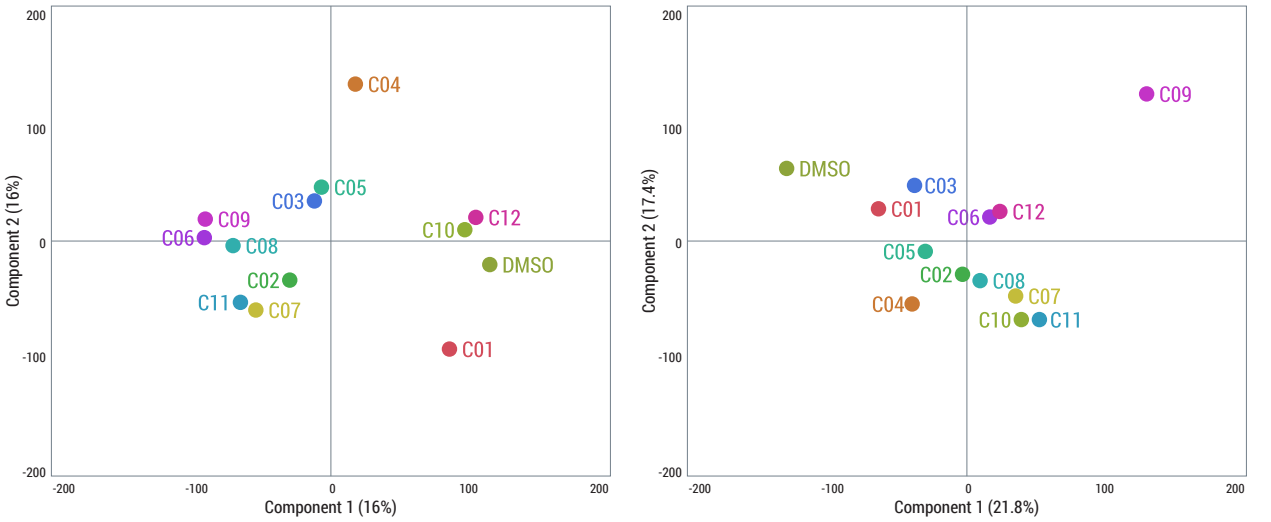


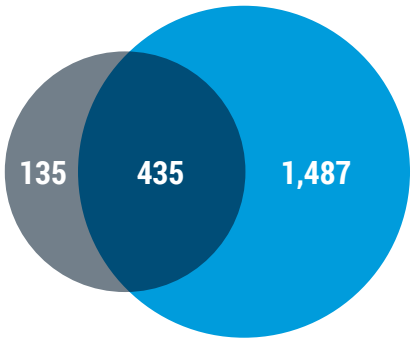
Figure 3. *Principal component plot of expression profiles for the twelve compounds (see Figure 2) and DMSO control. Treatment of PC3 cells with low dose (left) and high dose (right) are shown in the plots. Each sample is an average of triplicates, from log_2CPM standardized data. Note the movement away from the DMSO control as the dose increases. In some cases (e.g. C01) the lower dose and the control almost overlap.*

points on the plot. Two, dose effects also are differentiated by the results, and these dose effects vary by compound. This is shown by the movement of points away from the DMSO control on the PCA plots. For compounds such as C01, C10, and C12, the lower dose plots close to the DMSO control, meaning that they vary less than compounds such as C09 and C04, which plot further from the control. At the higher dose, all compounds have moved away from the control, indicating differentiation by dose as well as by compound. Third, the compounds form clusters on the PCA plot that may represent closely-related effects. Overall, this is a clear demonstration that closely related compounds that share the same target are indeed differentiated using transcriptomic profiles from the HTP.

While there was no *a priori* reason to assume that differences in the transcriptomic profiles would follow the differences in compound structure, it is interesting to note that the clustering of structures *(Figure 2)* and the PCA plots *(Figure 3)* do not reflect one another exactly.

An important question when considering the effect of dose is to consider whether the difference in doses causes qualitatively different changes, or only a different magnitude of the same changes. A comparison of the differentially expressed genes between cells treated with high and low doses of each compound showed that the latter is the case. *Figure 4* shows the overlap of differentially expressed genes between doses of C09. There is also a clear directional correlation for the majority of genes measured (84.9% when using a p-value cutoff of 0.05). Overall, the directionality of the data and the limited changes at the lower doses indicate that a single dose is sufficient for screening purposes. For compounds of high interest, additional doses and/or treatment lengths could be added.

Figure 4. *The effect of dose on transcriptomic profiles. C09 was used as an example, as it had the largest number of differentially expressed genes. The diagram indicates the number of significantly differentially regulated genes unique to low or high doses, and the overlap.*



Using HTP Data to Query the CMap/LINCS Database

As discussed above, CMap/LINCS is a large and well-known repository of transcriptomic profiles generated from treatment of various cell lines. LINCS data are generated by directly measuring the expression level of 978 landmark genes (the "L1000") using a luminescent assay on isolated RNA, and using those signals to infer the expression of other genes, for a total of 12,328 reported genes. To query the LINCS database, a differential expression profile is compared to the database of known profiles using an algorithm described by Subramanian *et al.*[5] A "connectivity score" or Tau value is one of the results returned. This connectivity score is ranked on a scale from 100% to -100%. On the positive scale, the percentage represents the rank in the entire database of the returned compound, such that a score of 99% means that only 1% of other compounds are more similar to the query profile. This approach was used to query the LINCS database, using differential expression profiles from HTP data (data were restricted to the 12,328 LINCS gene space). As the identity of each compound was known, the Tau of the self-compound (*i.e.,* the expected match) was recorded, as was the rank of the self-compound.

HTP data were processed for comparison as follows. First, differential expression values were calculated for PC3 cells treated with the higher compound dose vs the DMSO control, using the DESeq2 algorithm. The HTP data were then subsetted to only include the genes that overlap those reported in LINCS/CMap (n=12,096). Fold-change values from up- and down-regulated genes were used to query the LINCS database, using an R package (signatureSearch[10]) with a built-in LINCS algorithm that uses the calculation described in the 2017 Subramanian *et al.* paper.[5] The results are summarized in *Table 1*.

Table 1. Querying CMap/LINCS with HTP data. *Differential expression profiles for PC3 cells treated with the indicated compounds were used to query the LINCS database. The compound self-rank (restricted to PC3 cells) and Tau score (for the entire database) are shown. Note that C06 was not present in the LINCS database.*

Compound ID	Name	Rank	Tau
C01	LY-294002	2	99.88
C02	temsirolimus	100	99.14
C03	sirolimus	16	99.78
C04	deforolimus	26	99.69
C05	everolimus	410	92.94
C06	dactolisib	not present	N/A
C07	AZD-8055	1	99.99
C08	torin-1	2	99.99
C09	torin-2	1	100.00
C10	OSI-027	55	99.87
C11	WYE-125132	32	99.87
C12	KU-0063794	5	99.94

The outputs fell into two categories: Compounds whose self-retrieval had a high rank and high Tau, and those who had a lower rank but a high Tau. Of note, ten of eleven compounds had a Tau score better than 99% (the exception was everolimus, with a Tau score of 92%). Overall, these results show that HTP and LINCS data are directionally aligned, especially for target identification. With the caveat that this is a relatively small dataset, it is feasible to consider using LINCS data in conjunction with HTP data.

These results also point to a limitation of using the LINCS database. LINCS results aid in identifying signatures or profiles that are similar to (or opposite of) the query signature; they are not intended to specifically differentiate between closely related compounds. Indeed, in all our query results, many highly similar profiles (Tau of 99%+) were returned; these highly similar profiles principally hailed from other compounds targeting mTOR or PIK3 kinases. This expected behavior reinforces that HTP data can be used effectively to query the LINCS database, but also suggests that subtle effects could be missed.

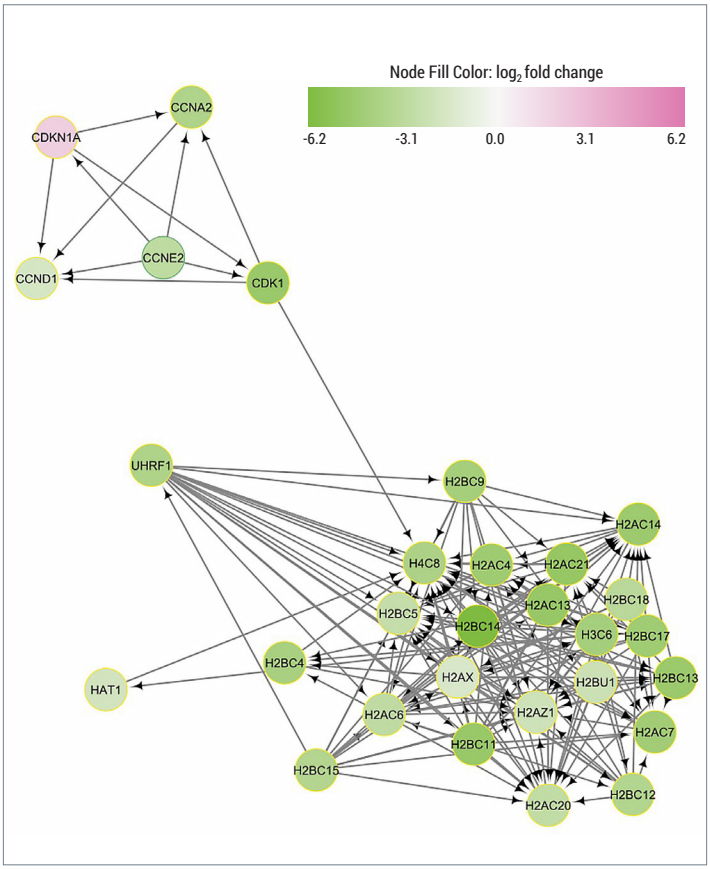
HTP Data Provide Insights that Extend Beyond the Capabilities of LINCS

Given the results from the LINCS queries, an important question is whether there is value to using an assay such as the HTP, which covers the entire human transcriptome, when the data from LINCS are readily available. A recent paper[11] reports that the number of significantly differentially expressed genes within a signature was a major factor in the quality of the signature and its subsequent self-retrieval from LINCS (*i.e.*, the more changes measured, the easier it is to differentiate between different compounds). Since the differential expression of LINCS is determined by the 978 genes in the L1000 itself (all other values are inferred), an obvious question is: of the number of differentially expressed genes measured in the HTP data, how many of these are also in the L1000 set? In the high dose/PC3 HTP data (which had the highest number of differentially expressed genes), only about 8% of the differentially expressed genes are part of the L1000 subset. Therefore, it seems likely that HTP data could provide additional information using its higher resolution of differential expression.

To explore this, pathway analysis was performed using iPathwayGuide (Adviata Biosciences) and differential expression data from HTP profiles of treated PC3 cells. The initial focus was on the rapalogs, as these compounds have been extensively studied and the expectation was that findings could be rapidly confirmed by existing literature.

A meta-analysis of all four rapalog profiles showed that two of the most significantly impacted pathways were *cell cycle* and *neutrophil extracellular trap formation.* Rapalogs, and indeed all the mTOR-targeting compounds, are known to cause cell-cycle arrest at the G1/S boundary.[12] This boundary is tightly controlled by Cyclin E and Cyclin A. All four rapalogs downregulated *CCNA2* and *CDK2*, which encode Cyclin A, and most also downregulated *CCNE2* (Cyclin E), as well as other genes involved in control of the G1/S transition. Neutrophil extracellular trap (NET) formation is, as the name suggests, generally associated with neutrophils. PC3 cells are prostate cancer derived and not expected to have neutrophilic properties. A closer look at the data showed that the entire pathway was not impacted; but many of the affected genes in that pathway encode histone genes (*e.g. H2BC3, H3C4*), and these genes are strongly downregulated in rapalog-treated cells. Surprisingly, there are no reports in the literature of histone downregulation in response to mTOR inhibition (whether by rapalog treatment or otherwise). Nor is the relationship "hidden" in the LINCS database but not explicitly described in the literature, as histone genes are not included in the LINCS 12,328-gene space. There is evidence to support the link, as DNA damage (which results in G1/S arrest) results in downregulation of histone genes,[13] and G1/S arrest is already known to be associated with mTOR inhibition, but the direct link has not been previously described. Pathway mapping of histone genes and differentially expressed upstream regulators *(Figure 5)* shows the relationship of histone genes to regulators of the

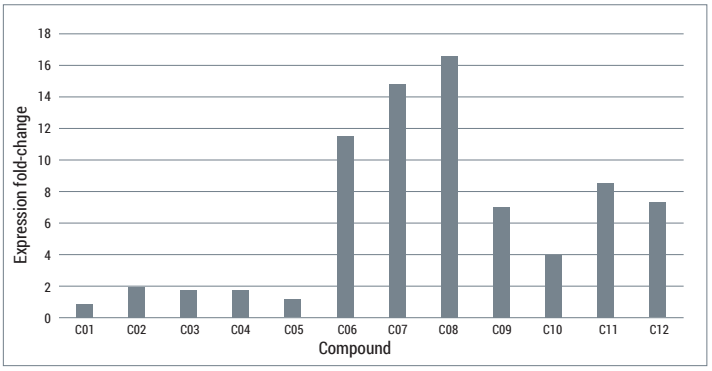
Figure 5. *Pathway analysis of compound-treated cells showing downregulated histone genes and possible upstream regulatory factors, notably two genes involved in G1/S cell cycle transition (CKD1, UHRF1). Color scale represents the log_2 fold change of expression compared to the DMSO control.*



G1/S transition (*UHRF1* and *CDK1*).[14] These findings suggest that the HTP-generated transcriptomic profiles have the potential to reveal mechanisms that have not been previously described in the scientific literature, thus increasing the body of knowledge available for drug candidates.

An additional use for gene expression data during transcriptome-informed drug discovery is in measuring Cytochrome P450 gene expression, a standard assessment of compound metabolism.[4] This is easily done using the HTP data. Given that the compounds in this study were not expected to induce the expression of canonically-measured *CYP* genes, it was not surprising to find that neither *CYP1A2, CYP2B6, CYP2C9,* nor *CYP3A4* were upregulated. However, there was upregulation of *CYP3A5,* another Phase I enzyme, particularly in the non-rapalog compounds *(Figure 6)*. While LINCS does include 44 *CYP* genes, none are directly measured via the L1000 and thus would not be useful for this type of fold-change analysis.

Figure 6. *Cytochrome P450 gene expression measurement in HepG2 cells. Expression data for each of the twelve compounds was compared to the DMSO control and a fold-change calculated. CYP1A2, CYP2B6, CYP2C9, and CYP3A4 were not significantly expressed in any sample and are not shown.*



Conclusions

The data described above illustrate opportunities to leverage transcriptomic profiling in drug discovery. First, the HTP can be used to generate high-quality and highly repeatable transcriptomic profiles from compound-treated cells using small sample input amounts. The quality of the data may allow for single-dose/single-replicate testing, which would minimize screening costs. These profiles may be used to differentiate between compounds, even closely-related compounds with similar targets and mechanism of action. Second, the effects measured may be used to investigate specific cellular pathways and enrich understanding of individual molecules. Third, HTP data may be used to successfully query the CMap/LINCS database. The size and richness of this database is attractive when considering the addition of machine learning approaches to drug design. However, using HTP data is expected to add more nuance and biological information, as well as enable screening of compounds not found within the current database. By directly measuring an additional ~18,000 genes over the CMap/LINCS L1000 assay, the HTP affords significant additional transcriptomic granularity.

The results of this study provide additional support for the utility of transcriptomic profiling to early drug discovery as part of a transcriptome-informed paradigm. In particular, this study shows the clear advantages of the HTP as a profiling tool for these purposes. Compared to the L1000 assay, the much larger number of directly measured genes provides improved resolution between profiles, and thus more refined differential expression analyses. Conversely, RNA-seq also directly measures more genes than the L1000 assay, but has known drawbacks to at-scale work. HTG EdgeSeq and the HTP provide comprehensive and focused measurement of the transcriptome, with the advantages of rapid assay and analysis time and the ability to

scale to higher-throughput profiling of compound libraries at early stages of design. Those profiles are expected to provide the resolution needed to further uncover and understand biological mechanisms, including more thorough assessments of on- and off-target effects of compounds intended as therapeutic entities.

References

1: Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. (2013). Transcriptional data: a new gateway to drug repositioning? *Drug Discovery Today* 18, 350-357.

2: Alexander-Dann B, *et al.* (2018). Developments in toxico-genomics: understanding and predicting compound-induced toxicity from gene expression data. *Mol. Omics* 14, 218-236.

3: Baillif B, Wichard J, Méndez-Lucio O, Rouquié D. (2020). Exploring the Use of Compound-Induced Transcriptomic Data Generated From Cell Lines to Predict Compound Activity Toward Molecular Targets. *Front. Chem.* 8, 296.

4: Zanger UM, Schwab M. (2013). Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics* 138, 103-141.

5: Subramanian A, *et al.* (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171, 1437-1452.e17.

6: Koleti A, *et al.* (2018). Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Research* 46, D558-D566.

7: HTG white papers *(https://www.htgmolecular.com/assays/htp).*

8: Bajusz D, Rácz A, Héberger K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 7, 20.

9: Saxton RA, Sabatini DM. (2017). mTOR Signaling in Growth, Metabolism, and Disease. *Cell* 168, 960-976.

10: Duan Y,*et al.* (2020). signatureSearch: environment for gene expression signature searching and functional interpretation. *Nucleic Acids Research* 48, e124-e124.

11: Lim N, Pavlidis P. (2021). Evaluation of connectivity map shows limited reproducibility in drug repositioning. *Sci Rep* 11, 17624.

12: Hashemolhosseini S, *et al.* (1998). Rapamycin Inhibition of the G1 to S Transition Is Mediated by Effects on Cyclin D1 mRNA and Protein Stability. *Journal of Biological Chemistry* 273, 14424-14429.

13: Su C, *et al.* (2004). DNA damage induces downregulation of histone gene expression through the G1 checkpoint pathway. *EMBO J* 23, 1133-1143.

14: Arima Y, *et al.* (2004). Down-regulation of nuclear protein ICBP90 by p53/p21Cip1/WAF1-dependent DNA-damage checkpoint signals contributes to cell cycle arrest at G1/S transition: Down-regulation of ICBP90 in G1 checkpoint. *Genes to Cells* 9, 131-142.

HTG Molecular Diagnostics, Inc.
Call 1-877-289-2615
www.htgmolecular.com

For Research Use Only. Not for use in diagnostic procedures.
HTG EdgeSeq is a trademark of HTG Molecular Diagnostics, Inc.
© 2022 All Rights Reserved.

C-122-03.28.R0