Published in final edited form as:

Drug Discov Today Dis Models. 2013 ; 10(1): . doi:10.1016/j.ddmod.2011.12.001.

Making Models Work: Library Annotation through Phenoclustering

CH Williams¹ and CC Hong^{2,3,4,5}

¹Division of Cardiovascular Medicine, Veterans Affairs TVHS, Nashville, TN, United States of America

²Department of Pharmacology, Veterans Affairs TVHS, Nashville, TN, United States of America

³Department of Cell and Developmental Biology, Veterans Affairs TVHS, Nashville, TN, United States of America

⁴Vanderbilt Institute of Chemical Biology, Vanderbilt University School of Medicine, Veterans Affairs TVHS, Nashville, TN, United States of America

⁵Research Medicine, Veterans Affairs TVHS, Nashville, TN, United States of America

Abstract

For the chemical biologist, the promise of the post-genomic era has yet to be fulfilled. In the past decade, a flurry of phenotype-based chemical genetic screens in *in vivo* and cultured cell models have yielded numerous small molecules with interesting biological properties with potential to reveal plethora of novel insights. However, these screens have also led to the bottleneck of target identification. This article will focus on recent progress in phenoclustering in various model systems as an option for target identification.

Introduction

Chemical genetic screening for small molecules that affect *in vitro* and *in vivo* phenotypes is drug development paradigm that is increasingly recognized as a viable alternative to the classical target-based drug discovery paradigm. As a model organism for such screening, the zebrafish is becoming a favorite because of its rapid development, high fecundity, low cost, vertebrate orthologies and liquid aquaculture allowing for precise and scalable screening. However, because of the organism's complexity, screening can yield numerous and often complex phenotypes. The traditional method of target identification is affinity chromatography, which is both time and labor intensive with predilection for identifying the most abundant proteins that may or may not be biologically relevant. Although newer approaches such as drug affinity responsive target stability (DARTS) and yeast three hybrid systems are promising, target identification requires a separate platform distinct from the original *in vivo* screening models¹. By contrast, a key advantage of an *in vivo* phenotype-based chemical screen using zebrafish and other animal models is that the developing animal itself can provide crucial clues as to the pathway being disrupted. As such, biological responses to small molecules with known activity can be quantified and used as a reference

Corresponding author: Charles C. Hong, charles.c.hong@vanderbilt.edu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

for clustering responses of unknown compounds, with a supposition that tightly clustered compounds will have similar mechanisms of action (Fig 1).

***In silico* clustering**

High-throughput screening has become a staple of drug development; and has resulted in the formation of a large repertoire of information on PubChem. As of August, 2011, over 30 million chemically unique compounds have been deposited in the PubChem database. Furthermore, over 500 thousand bioassay records have been uploaded, representing over 130 million experimental bioactivity results. With this wealth of data Han and colleagues were able to mine the PubMed bioactivity spectra by hierarchical clustering and were able to understand the biological mechanisms of target-small molecule interactions². One such example was the compound myricetin (PubChem CID:5281672), a flavonoid that is commonly found in natural food source. By examining the bioactivity spectra they found that the molecule is an inhibitor of several proteins such as aldehyde dehydrogenase, Leishmania Mexicana pyruvate kinase, and stress-activated protein kinase. This finding was confirmed in a literature search where the mechanism of action of this molecule had been described². This was a principle proof that structurally similar compounds have similar bioactivity spectra. Therefore, the same group later used a similar method and investigated 37 small molecules in the context of their PubChem bioactivity spectra and chemical similarity. They found that compounds that were then examined in the context of the NCI-60 clustered into groups with similar mode of actions, which strongly correlated with chemical structures³. The NCI-60 is a project run by DTP (Developmental Therapeutic Program) and NCI (National Cancer Institute), designed to screen up to 3,000 compounds per year for potential anticancer activity against 60 different human tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney. The service is provided at no cost to the individual who submits their compound. Given the results of this study, the authors suggest that the NCI-60 activity spectra could be used as a standardized resource for identifying compounds that have similar mechanisms of action through a clustering approach.

***In vitro* Clustering**

Phenoclustering is being applied *in vitro* with high-throughput annotation of cell morphology after exposure to known and previously undescribed bioactive molecules. For example, Tanaka and colleagues screened 107 compounds that were structurally similar in 5 separate cell lines. By looking at features such as area form factor and staining with Hoechst and alpha-tubulin they were able to identify a more potent a structural analog of PP, hydroxy-PP, which binds not only the src-family kinase Fyn but rather a completely different biomolecule, the oxidoreductase CBR1⁴. This method could be expanded and utilized with a more diverse chemical library, and possibly more read outs to allow for *in vitro* phenotypic target identification.

Another *in vitro* clustering approach utilizes the gene responses (transcriptomic fingerprint) elicited by exposure to small molecule. Connectivity Map⁵ and Mantra⁶ are two platforms developed for mouse and human genomes. In short, the gene response for an unknown compound is ranked and then clustered against a database of compounds with known mechanisms of actions and the mechanism is inferred. With this new technique, it was possible to infer a new mechanism of action for the FDA approved drug Fausudil, suggesting that it could be used as a therapeutic agent to induce autophagy⁶.

***In vivo* clustering**

Conceptually, phenoclustering has been widely used by developmental biologists for many years. Arguably, the pioneers of the “*in vivo* phenomics” are Eric Wieschaus and Christiane Nusslein-Volhard, whose seminal contribution was, not just in identification of mutations that cause patterning defects in *Drosophila* embryo, but in classification the mutant loci into groups based on distinct patterning defects⁷. From *Drosophila* to other model organisms, clustering of mutant loci resulting from large-scale mutagenesis screens based on morphologic phenotypes has become a central theme in developmental genetics. One such example of a large-scale forward genetic screen in vertebrates was the Tübingen Screen carried out at the Max-Planck Institute. Mutants were first clustered according to defects in areas such as jaw/ craniofacial development, retinal development, early arrest, pigmentation, and neural development⁸. They were then further clustered according to similarity of the defects - for example, the early arrest mutants were clustered by the timing of developmental arrest. The core logic behind such clustering is that mutations that have similar phenotypes are related by developmental pathway or mechanism. In other words, mutants that exhibited similar phenotypes were found to be caused by mutations in the same gene or in the genes in the same developmental pathway. By analogy, compounds found to cause specific phenotypes in a high content chemical genetic screen can be clustered based on distinct phenotypes, and compounds that elicit similar phenotypes presumed to target a common gene or distinct genes in a common developmental pathway.

To facilitate target identification of bioactive compounds discovered in phenotype based chemical screens, phenomic clustering will involve comparison of the small molecule-induced phenotypes to an annotation of phenotypes generated when each gene in the organism's genome is disrupted. In recent years, RNAi-based phenoclustering has been successfully used in to elucidate individual gene functions in *C. elegans*^{9,10} and *Drosophila*¹¹. Moreover, Sugimoto and colleagues have developed a database of RNAi knockdown phenotypes in *C. elegans* which can be mined for genes that exhibit the phenotype a small molecule elicits¹². The wealth of phenome-genome data makes *C. elegans* a choice model for phenotypic chemical screens¹.

Few studies have undertaken the use of phenoclustering in vertebrate models. Kokel and colleagues recently used the zebrafish model in a high-throughput screen to identify neuroactive small molecules¹³. The major hurdles to using a hierarchical clustering analysis in the context of a chemical screening are deciding the appropriate measurements that need to be made; and generating enough data points to be able to harness the full power of clustering analysis. By utilizing photomotor response (PMR), a startle response to high-intensity light, Kokel, et al. screened 14,000 small molecules¹³. Simply put, this assay measured whether the zebrafish embryos moved more or less in response to light stimuli. Since a single qualitative reading, such as increased or decreased movement, is insufficient, they developed multiple quantifiable read outs. Moreover, rather than simply quantifying motion in four broad phases of PMT (Background, Latency, Excitation, and Refractory), Kokel further divided the Excitation period into three segments and the Refractory period into two, obtaining a 14 data-point barcode for each small molecule¹³.

A pilot study with several known small molecules with known neuroactivity across different mechanisms (adrenergic, dopaminergic, and serotonergic) demonstrated that, not only were the PMR profiles reproducible across days, embryos, and replicates, but the pharmacological effects mirrored that of what happens in mammalian systems. For example, isoproterenol, a psycho-stimulant, increased activity throughout PMR while apomorphine, a dopamine agonist, lengthened PMR latency. With these profiles and others in place, a large-scale screen of 14,000 compounds yielded 1,627 hit compounds that were then clustered¹³.

Interesting, many of the clustered hits that shared a similar activity profile also showed similar chemical scaffolding, demonstrating the strength of this approach to identify chemical motifs having similar bioactivities. The true strength of clustering utilized by this method is the potential for target identification. Of 15 compounds that exhibited a "slow-to-relax" phenotype, two, STR-1 and STR-2, were novel compounds that clustered closely with eserine, a known inhibitor acetylcholinesterase (AChE). Indeed, Kokel et al. demonstrated that STR-1 and STR-2 are novel AChE inhibitors¹³. In a similar behavioral study using zebrafish, Rihel et al. used restwake behavior as a platform for small molecule screening. Implementing a similar behavioral fingerprint and then utilizing hierarchical clustering, they uncovered novel mechanisms involved in the regulation of rest-wake behaviors, including the role of ERG potassium channels and immunomodulators like NSAIDs¹⁴. Importantly, this method could facilitate target identification. For example, MRS-1220, an adenosine A3 receptor antagonist, which clustered with monoamine oxidase (MAO) inhibiting antidepressants, was found to inhibit *in vitro* MAO activity with an IC₅₀ of ~1 μ M¹⁴.

Advances for *in vivo* phenoclustering

As an *in vivo* model for phenoclustering, the zebrafish has incredible potential. Among small animal models amenable to large scale chemical screens, zebrafish is the lone vertebrate, sharing the highest genetic homology to humans. Moreover, the structural, physiological and behavioral similarities permit elegant behavioral studies as described above as well as other studies that examine specific organ systems¹.

As for morphology-based phenoclustering, analyzing the shape of an organ in the zebrafish via standard microscopy might not yield enough quantitative data points for hierarchical clustering. Nonetheless, there are exciting emerging technologies that could overcome the limitation of morphology-based chemical screens¹⁵. For example, Canada and colleagues developed a system for high-throughput histology and image capture for whole mount zebrafish, as well as a program SHIRAZ, an automated histology image annotation system for zebrafish phenomics¹⁶. The authors focused on the retina, which consists of seven easily discernable layers, and quantified for each layer various phenotypic attributes, such as absence, necrosis, disorganization, and hypotrophy¹⁶. By utilizing this technology, it may be possible to conduct large-scale morphology-driven phenotypic small molecule screens that generate more than enough data to allow for hierarchical clustering, and, with a trained set using known bioactive compounds, target identification as well.

Model comparison

All models described above have their drawbacks (Summarized in Table 1). With *in silico* mining of phenomic data, investigators are limited to the data that is available. However, public resources with standardized annotation such as the NCI60 bioactivity spectra are promising places to start. *In vitro* phenoclustering of cells and cellular responses to chemical stimuli has potential for direct inter-actors and target identification but lacks the proper *in vivo* context for evaluation of the effects of compounds at the whole organism level. Nonetheless, this approach allows for discovery of compounds eliciting complex cellular responses and clustering of such compounds in a high-throughput manner. The *c. elegans* model, with its searchable database of systematic gene knockout and phenotypic analysis, is very promising as well. A drawback of *c. elegans* is an invertebrate model, and as such, lack certain fundamental biology found in mammals. The use of zebrafish in behavioral chemical screens and phenotype clustering provides a new paradigm for small molecule discovery and target identification. The zebrafish provides an ideal model because it is a vertebrate with behaviors and drug targets that are highly conserved to those of mammals. Finally, although phenotype clustering is yet to be used systematically in a morphology-based analysis in the

zebrafish, it is posed to make impact with the advent of systematic high-throughput methods to process and analyze morphological features.

Translations to humans

While the chemical biological approaches discussed here are exciting, there is still much to be done in the field of with regards to pharmacological targeting of the human proteome/phenome. Human tumor cell lines, tissue cell lines and pluripotent stem cells like induced pluripotent stem cells (iPSCs) can be used for probing the chemical-phenotype interactions. However, *in vitro* methods do not provide the organism-level context necessary for accurate prediction of *in vivo* effects in live animals. In this regard, while model organisms do not share complete sequence homology to humans and there is no guarantee that the targets of small molecules are conserved between species, *in vivo* phenotype-based chemical screens and characterization using model organisms like the zebrafish may provide the best springboard for translation to human biology and medicine¹.

Conclusion

Phenotype-based chemical genetic screening is transforming how new drugs are discovered. However, a major hurdle to realizing the full potential of this approach to impact biology and medicine is target identification. Because traditional target identification approaches are so labor intensive, targets for only a small number of “hit” compounds have been successfully identified thus far. By contrast, hierarchical clustering has the potential to dramatically increase the throughput of small molecule target identification efforts, and mechanism of action studies. In summary, hierarchical clustering of compounds discovered in phenotype-based screen is a relative new concept posed to make a major impact in the drug discovery field.

Acknowledgments

C.C.H. was supported by Vanderbilt Institute for Clinical and Translational Research, Department of Veterans Affairs CDTA, VA Merit, NIH/NHLBI R01HL104040 and U01HL100398.

Reference

1. Williams CH, Hong CC. Multi-step usage of *in vivo* models during rational drug design and discovery. *Int J Mol Sci*. 2011; 12(4):2262–2274. [PubMed: 21731440]
2. Han L, Wang Y, Bryant SH. A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics*. 2009; 25(17):2251–2255. [PubMed: 19549631]
3. Cheng T, Li Q, Wang Y, Bryant SH. Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. *J Chem Inf Model*. 2011; 51(9): 2440–2448. [PubMed: 21834535]
4. Tanaka M, Bateman R, Rauh D, et al. An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol*. 2005; 3(5):e128. [PubMed: 15799708]
5. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*. 2006; 313(5795):1929–1935. [PubMed: 17008526]
6. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U.S.A.* 2010; 107(33):14621–14626. [PubMed: 20679242]
7. Nusslein-Volhard C, Wieschaus E. Mutations affecting segment number and polarity in *Drosophila*. *Nature*. 1980; 287(5785):795–801. [PubMed: 6776413]

8. Haffter P, Granato M, Brand M, et al. The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development*. 1996; 123(1):1–36. [PubMed: 9007226]
9. Boulton SJ, Gartner A, Reboul J, et al. Combined Functional Genomic Maps of the *C. elegans* DNA Damage Response. *Science*. 2002; 295(5552):127–131. [PubMed: 11778048]
10. Piano F, Schetter AJ, Morton DG, et al. Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. *Curr. Biol*. 2002; 12(22):1959–1964. [PubMed: 12445391]
11. Fuchs F, Boutros M. Cellular phenotyping by RNAi. *Brief Funct Genomic Proteomic*. 2006; 5(1): 52–56. [PubMed: 16769679]
12. Sugimoto A. High-throughput RNAi in *Caenorhabditis elegans*: genome-wide screens and functional genomics. *Differentiation*. 2004; 72(2–3):81–91. [PubMed: 15066188]
13. Kokel D, Bryan J, Laggner C, et al. Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nat Chem Biol*. 2010; 6(3):231–237. [PubMed: 20081854]
14. Rihel J, Prober DA, Arvanites A, et al. Zebrafish Behavioral Profiling Links Drugs to Biological Targets and Rest/Wake Regulation. *Science*. 2010; 327(5963):348–351. [PubMed: 20075256]
15. Sabaliauskas NA, Foutz CA, Mest JR, et al. High-throughput zebrafish histology. *Methods*. 2006; 39(3):246–254. [PubMed: 16870470]
16. Canada BA, Thomas GK, Cheng KC, Wang JZ. SHIRAZ: an automated histology image annotation system for zebrafish phenomics. *Multimedia Tools and Applications*. 2011; 51(2):401–440. [PubMed: 21461317]
17. Sanoh S, Horiguchi A, Sugihara K, et al. Prediction of *In vivo* Hepatic Clearance and Half-life of Drug Candidates in Human using Chimeric Mice with Humanized Liver. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*. 2011 Available at: <http://www.ncbi.nlm.nih.gov/PubMed/22048522>.

Figure 1. Annotation of small molecule libraries through hierarchical clustering can be done in a number of models; *in silico*, *in vitro* and *in vivo*

The data generated for the query compound in these models is then compared to a reference data set through clustering. The nearest neighbor should have a similar or the same target as the query compound (yellow).

Table 1

	In Vivo	In Vitro	In Silico
Subtypes	Behavioral Phenotypic RNAi	Transcriptomic Cell morphology	Bioactivity spectra
Pros	High correlation in humans	Fast High-throughput	Fast
Cons	Slightly slower Complex readouts	Highly context dependent	Only builds on pre-existing data