MOSAIC: a chemical-genetic interaction data repository and web resource for exploring chemical modes of action

Justin Nelson[1], Scott W. Simpkins[1], Hamid Safizadeh[2,3], Sheena C. Li[4], Jeff S. Piotrowski[5], Hiroyuki Hirano[4], Yoko Yashiroda[4], Hiroyuki Osada[4,*], Minoru Yoshida[4,*], Charles Boone[4,6,*] and Chad L. Myers[1,2,*]

[1]Bioinformatics and Computational Biology Program, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, USA, [2]Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, USA, [3]Department of Electrical and Computer Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, USA, [4]RIKEN Center for Sustainable Resource Science, Wako, Saitama, Japan 3510198, [5]Yumanity Therapeutics, Cambridge, MA, USA 02139 and [6]University of Toronto, Donnelly Centre, Toronto, Ontario, Canada

*To whom correspondence should be addressed.

Abstract

Summary: Chemical-genomic approaches that map interactions between small molecules and genetic perturbations offer a promising strategy for functional annotation of uncharacterized bioactive compounds. We recently developed a new high-throughput platform for mapping chemical-genetic (CG) interactions in yeast that can be scaled to screen large compound collections, and we applied this system to generate CG interaction profiles for more than 13,000 compounds. When integrated with the existing global yeast genetic interaction network, CG interaction profiles can enable mode-of-action prediction for previously uncharacterized compounds as well as discover unexpected secondary effects for known drugs. To facilitate future analysis of these valuable data, we developed a public database and web interface named MOSAIC. The website provides a convenient interface for querying compounds, bioprocesses (GO terms), and genes for CG information including direct CG interactions, bioprocesses, and gene-level target predictions. MOSAIC also provides access to chemical structure information of screened molecules, chemical-genomic profiles, and the ability to search for compounds sharing structural and functional similarity. This resource will be of interest to chemical biologists for discovering new small molecule probes with specific modes-of-action as well as computational biologists interested in analyzing CG interaction networks.

Availability: MOSAIC is available at http://mosaic.cs.umn.edu.

Contact: chadm@umn.edu, charlie.boone@utoronto.ca, yoshidam@riken.jp, or hisyo@riken.jp

1 Introduction

Discovering the mode-of-action of compounds is an important step in developing effective treatments for diseases or controlling the spread of pathogens. With the advent of efficient whole-genome sequencing methods that enable comprehensive genotyping of individuals, there are an increasing number of candidate drug targets implicated in specific diseases, which presents a need for new drug discovery paradigms that maximize these opportunities. The standard methods in high-throughput screening are relatively inefficient at linking small molecules to cellular targets and thus need to be improved in order to develop a rich collection of tool compounds for probing biological pathways and generating drug leads. In particular, a typical strategy for screening chemical libraries involves developing a specific biochemical assay to detect activity against a desired target protein and then applying it to a large library of candidate small molecules (Macarron *et al*., 2011; Bibette, 2012). Such an approach requires a significant commitment to a specific target protein or pathway before one develops a specific assay. Thus, we have yet to build a collection of small molecules that are able to target a substantial fraction of the human proteome.

Alternatively, "top-down" chemical-genomic approaches reveal compound-dependent phenotypes in a systematic, unbiased manner to predict the cellular role of an uncharacterized bioactive compound. For example, studies have tested compounds against the complete gene deletion collection that is available for the genetically tractable yeast *Saccharomyces cerevisiae* (Giaever *et al*., 1999; Parsons *et al*., 2004; Parsons *et al*., 2006; Giaever and Nislow, 2014; Lee *et al*., 2014). In these approaches, a collection of genetic mutants is assayed for sensitivity or resistance to a compound to generate a compound-specific functional profile. Methods are available for haploid, homozygous, and heterozygous deletion mutant collections and have incorporated global genetic interaction networks to make mode-of-action predictions (Giaever *et al*., 1999; Parsons *et al*., 2004; Parsons *et al*., 2006; Hillenmeyer *et al*., 2008; Costanzo *et al*., 2010; Hoepfner *et al*., 2014; Lee *et al*., 2014; Wildenhain *et al*., 2016). We recently developed a new high-throughput platform for mapping CG interactions in yeast for large-scale compound screening and, using this, we generated a library of CG profiles for various compound collections (Piotrowski *et al*., 2017). This represents one of the most extensive chemical-genomic datasets to date, including informative profiles for more than 1,000 natural products and derivatives, hundreds of clinically relevant compounds, and other small molecules. Here, we present MOSAIC, a web interface that provides access to this rich chemical-genomic dataset.

2 Methods

MOSAIC, the website for accessing the database, is located at http://mosaic.cs.umn.edu. This website displays data obtained by screening the RIKEN Natural Product Depository (NPDepo) as well as diversity libraries from the NCI/NIH, and the Published Kinase Inhibitor Set from Glaxo-Smith-Kline (Piotrowski *et al*., 2017). The data housed in the database were derived from the experiments described in (Piotrowski *et al*., 2017) and were processed and analyzed using the BEAN-counter (Simpkins, 2017) and CG-TARGET (Simpkins *et al*., 2017) tools.

MOSAIC was designed using a combination of hypertext markup language (HTML), javascript with jquery, asynchronous javascript and XML (AJAX), javascript object notation (JSON), PHP, and MySQL. The website uses PHP to deliver an HTML page to the client, with javascript and jquery for client side scripting. To retrieve data from the MySQL database, the website uses AJAX to run a PHP script on the server that queries the MySQL database and returns a JSON object back to the client to be displayed. This provides a responsive environment for querying the CG interaction data.

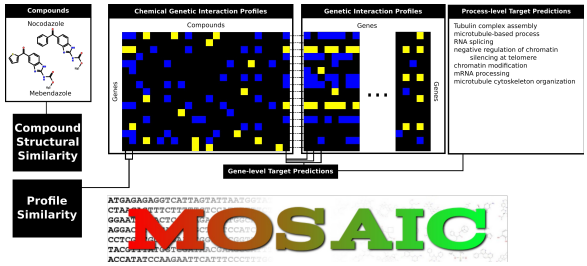

Fig. 1. Overview of data sources and mode-of-action predictions for the MOSAIC website. These include CG interaction profiles, gene and bioprocess-level target predictions, compound structural similarity and CG interaction profile similarity. Chemical-genetic interaction profiles are generated using BEAN-counter. Gene-level target predictions are derived from measuring similarities between the CG interaction profiles and genetic interaction profiles. Bioprocess-level target predictions are generated by mapping gene-level target predictions onto Gene Ontology biological process terms. Compound structural similarity and profile similarity compare compound structures or CG interaction profiles with other compounds structures and profiles respectively.

3 Usage

MOSAIC provides an interface for querying CG profiles, gene- and bioprocess-level target predictions. The MOSAIC interface has a search bar available in the middle or top of the browser that will accept a compound name, a gene common name or standard name, or a Gene Ontology biological process term. Multiple terms can be input by separating each term with a semicolon. The results page will return different data and predictions depending on whether a compound, gene, or process has been queried. If a query includes more than one type of input, the terms can be selected by using the tabs (Conditions, Genes, or GO Terms) and tables.

3.1 Chemical-genetic Interaction Profiles

The sensitivity or resistance of deletion mutants in the presence of a compound is provided as part of the CG interaction data. MOSAIC provides a table of CG interaction scores for compound queries (~13,000 compounds) or gene queries (~300 gene mutants in the yeast diagnostic pool). A negative CG interaction indicates that a deletion mutant is sensitive to a compound; conversely, a positive CG interaction indicates that a deletion mutant is resistant to a compound. These data are provided as part of MOSAIC, either in the form of a CG interaction profile (for a compound query) or as a list of conditions that a deletion mutant is sensitive or resistant to (for a gene query).

3.2 Gene-level Target Prediction

A gene-level target prediction is made based on the similarity between the CG interaction profile and a genetic interaction profile (Parsons *et al*., 2004; Parsons *et al*., 2006; Costanzo *et al*., 2010). A genetic interaction is when growth of a double mutant differs from expectation based on the single gene deletion effects. Compound and gene queries both return the table of relevant gene-level target prediction scores, which are computed using the inner product between a CG profile

and all ℓ2-normalized genetic interaction profiles. A high score between a CG profile and a genetic interaction profile indicates that the compound's effect phenocopies those of the gene deletion, suggesting that the compound perturbs the function and/or the proteins encoded by the gene. Using the MOSAIC website, researchers can quickly browse and form hypotheses based on these gene-level target predictions.

3.3 Bioprocess-level Target Prediction

A Gene Ontology biological process prediction is computed by mapping the gene-level target prediction scores onto GO biological process terms as described in (Simpkins *et al*., 2017). The table of Gene Ontology biological process predictions is displayed for compound queries, and the table of compounds predicted to a Gene Ontology biological process term is displayed for Gene Ontology queries. For example, Nocodazole binds tubulin and blocks microtubule formation. A query of "Nocodazole" will return a process prediction to "tubulin complex assembly" (GO:0007021). These Gene Ontology predictions, called bioprocess-level target predictions, are calculated by comparing the observed gene-level target prediction scores within a GO biological process to empirical null distributions of the same scores generated from experimental negative controls, resampled CG profiles, and shuffled gene-level target predictions (Simpkins *et al*., 2017). MOSAIC allows researchers to quickly generate a list of compounds predicted to target a specific cellular process of interest, enabling efficient identification of candidate compounds with a desired mode-of-action.

3.4 Chemical Structure and CG Profile Similarity

Compound similarity searching can be useful for expanding lists of candidate compounds. MOSAIC provides access to both structural similarity and functional similarity (CG interaction profile similarity) for compound queries. The compound structural similarity is calculated using the All-Shortest Path (ASP) structure descriptor (depth 8) combined with the Braun-Blanquet similarity coefficient, whose performance was superior to several other alternatives in connecting chemical structures to compound functions (Hinselmann *et al*., 2011; Safizadeh *et al*., 2017). Functional similarity measures are calculated using Pearson correlation applied to the CG profile. These two similarity approaches can be compared to understand how structure modulates the effect of a compound on a cell.

4 Conclusion

The MOSAIC database provides convenient access to CG interaction data, mode-of-action predictions, and structural information for a large collection of both characterized and uncharacterized compounds. This resource can serve as a starting point for discovering new probes with specific modes-of-action or further characterization of novel compounds.

Funding

This work was partially supported by the National Institutes of Health (R01HG005084, R01GM104975) and the National Science Foundation (DBI 0953881) to CM, the Canadian Institutes of Health Research, FDN-143264, to CB, the JSPS KAKENHI Grant Numbers 15H04483, to CB. MY is supported by JSPS KAKENHI Grant Number 26221204. HO is a research fellow of the Japan Society for the Promotion of Science. CM and CB are fellows in the Canadian Institute for Advanced Research (CIFAR) Genetic Networks Program. Computing resources and data storage services were partially provided by the Minnesota Supercomputing Institute and the UMN Office of Information Technology, respectively. SWS was supported by an NSF Graduate Research Fellowship (00039202), an NIH Biotechnology Training Grant (T32GM008347), and a one year BICB fellowship from the University of Minnesota.

Conflict of Interest: none declared.

References

Balkwill, F. (2004). Cancer and the chemokine network. Nature Reviews Cancer, **4(7)**, 540–550.

Bibette, J. (2012). Gaining confidence in high-throughput screening. Proceedings of the National Academy of Sciences, **109**, 649-650.

Costanzo, M. et al. (2010). The genetic landscape of a cell. Science, **327(5964)**: 425-431

Giaever, G. and Nislow, C. (2014). The Yeast Deletion Collection: A Decade of Functional Genomics. Genetics, **197(2)**: 451-465

Giaever, G. *et al*. (1999). Genomic profiling of drug sensitivities via induced haploinsufficiency. Nature Genetics, **21(3)**, 278–83.

Hillenmeyer, M. E. et al. (2008). The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. Science, **320(5874)**, 362–365.

Hinselmann G, et al. (2011) jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. J. Cheminformatics. **3**:1-14

Hoepfner, D. *et al*. (2014). High-resolution chemical dissection of a model eukaryote reveals targets, pathways and gene functions. Microbiology Research, **169(2-3)**:107-120

Kleinke, J. D. (2001). The Price Of Progress: Prescription Drugs In The Health Care Market. Health Affairs, **20(5)**, 43–60.

Lee, A. Y. et al. (2014). Mapping the cellular response to small molecules using chemogenomic fitness signatures. Science, **344(6180)**, 208–11.

Macarron, R. et al. (2011). Impact of high-throughput screening in biomedical research. Nature Reviews Drug Discovery, **10**, 188-195.

Parsons, A. B. et al. (2004). Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nature Biotechnology, **22(1)**, 62–9.

Parsons, A. B. et al. (2006). Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast. Cell, **126(3)**, 611–625.

Piotrowski, J. S. et al. (2017). Functional Annotation of Chemical Libraries across Diverse Biological Processes. Nature Chemical Biology. Advanced online copy.

Safizadeh, H. et al. (2017). Improving prediction of compound function from chemical structure using chemical-genetic networks. bioRxiv. doi:10.1101/112698.

Simpkins, S. W. (2017), BEAN-counter, GitHub repository, https://github.com/csbio/BEAN-counter

Simpkins, S. W. et al. (2017). Large-scale interpretation of chemical-genetic interaction profiles using a genetic interaction network. bioRxiv. doi:10.1101/111252.

van 't Veer, L. J. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, **415(6871)**, 530–536.

Viatour, P. et al. (2005). Phosphorylation of NF-κB and IκB proteins: implications in cancer and inflammation. Trends in Biochemical Sciences, **30(1)**, 43–52.

Wildenhain, J. et al. (2016). Systematic chemical-genetic and chemical-chemical interaction datasets for prediction of compound synergism. *Scientific Data*, 3(160095)