Customizing the Connectivity Map Approach for Functional Evaluation in Toxicogenomics Studies

Karmaus AL[1], Kothiya P[2], Watford S[2], Thomas RS[2], Martin MT[2]

[1]US Environmental Protection Agency, National Center for Computational Toxicology, Office of Research and Development, Research Triangle Park, NC; [2]Integrated Laboratory, Systems, Inc. Research Triangle Park, NC

Agnes Karmaus
akarmaus@ils-inc.com
919-291-1110 x254

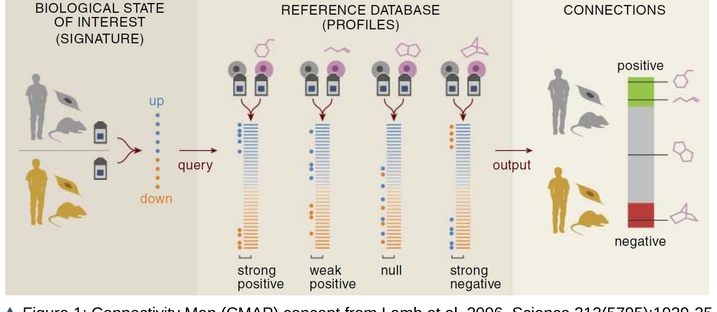
EPA

www.epa.gov

Abstract

Evaluating effects on the transcriptome can provide insight on putative chemical-specific mechanisms of action (MOAs). With whole genome transcriptomics technologies becoming more amenable to high-throughput screening, libraries of chemicals can be evaluated in vitro to produce large toxicogenomics datasets. However, developing a systematic approach for linking transcriptional changes to MOA has been challenging. This study presents a connectivity map (CMAP) inspired methodology to conduct gene set enrichment analysis using toxicogenomics datasets to identify putative MOAs for chemical-mediated effects. Our CMAP approach utilizes a reference database of differential expression profiles and a rank-based permutation test for identifying enriched molecular targets. This approach requires that profiles in the reference database represent a diversity of perturbations that are mapped or annotated to molecular targets. To satisfy these requirements, we established a reference database of ~900 whole-genome expression profiles from the original CMAP effort (Lamb et al., 2006) and annotated the chemical perturbagens to 86 unique targets. To evaluate the new custom CMAP approach, 34 chemicals were selected that encompass multiple MOAs including nuclear receptor agonists/antagonists, enzyme inhibitors, and chemicals interfering with cell integrity (tubulin disruption). MCF7 and HepaRG cells were treated with three concentrations of each chemical for six hours, and changes in whole genome expression were quantified using Affymetrix microarrays (De Abrew et al., 2016). Z-score distributions were used to identify differential gene expression (using a cutoff of z-score > 2) and profiles were matched using JG scoring (Jiang and Gentleman, 2007). Finally, a rank based permutation was applied to identify targets enriched among the significantly associated reference profiles. Of the 34 chemicals evaluated, 17 had MOAs that were not sufficiently represented in the reference database, 11 were correctly and significantly matched to their putative target, and six targets/mechanisms of action were not correctly identified. By integrating molecular target annotation and rank-based permutations for targets, this adapted CMAP approach can help identify putative MOAs for chemical-mediated effects using toxicogenomic data.

CMAP Concept



▲ Figure 1: Connectivity Map (CMAP) concept from Lamb et al. 2006. Science 313(5795):1929-35. Experimental genome-wide expression profiles are compared to reference chemical expression profiles to identify positive and negatively correlated profiles. The output of this CMAP approach is a ranking of similar and dissimilar chemicals, based solely on gene expression profiling.

CMAP Method Development

1 Process Experimental Gene Expression Data

Input data required for the CMAP analysis is a list of differentially expressed genes from any genome-wide expression profiling platform

- Raw data: Microarray CEL files, Sequencing FASTQ files, etc.
- Normalization: RMA, FPKM, etc.
- Differential Gene Expression: fold change, z-score, p-value, etc.

2 Compare to Reference Profile Database

Like the original CMAP concept, this step identifies & ranks the most like and dislike chemicals from the reference profile database

- Gene set enrichment analysis (GSEA)
- Statistical output for enriched profiles: JG score, gsealm, KS statistic, etc.

3 Evaluate Significant Connections

To identify significantly enriched biological targets, the ranks from step 2 are randomized and identified chemicals/targets are evaluated for enrichment

- Reference profiles annotated by chemical, gene target, chemical class, etc.
- Resulting ranks per annotation group are randomly assigned (permuted) hundreds of times and an empirical p-value is calculated

Reference Database Landscape

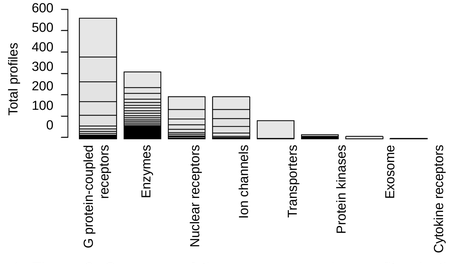


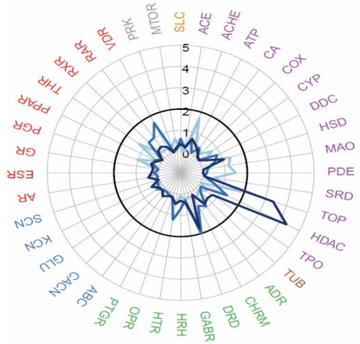
Table 1: Summary of CMAP Targets and Associated Profiles

Target Family	Target Genes	Chemicals	Cell Lines	Total Profiles
Cytokine receptors	1	1	3	3
Enzymes	40	112	5	336
Exosome	1	4	4	14
G protein-coupled receptors	16	192	4	585
Ion channels	8	65	3	194
Nuclear receptors	10	71	5	227
Protein kinases	8	6	4	19
Transporters	2	35	3	102

▲ Figure 2: Summary of the gene expression profiles in the CMAP reference database. Nearly 2,700 Affymetrix whole-genome expression profiles (as .CEL files) were obtained from the CMAP project (https://portals.broadinstitute.org/cmap/). Data were processed and chemicals were mapped to biochemical targets using DrugBank and manual curation. In total, ~45% of the chemicals in the database were mapped to a target molecule and included for our modified CMAP approach. The breakdown of data within our resulting CMAP database are summarized to highlight the number of profiles (chemicals and cell lines) per annotated target family.

Experimental Application

MCF7 and HepaRG cells were treated with three concentrations of each chemical for six hours, and changes in whole genome expression were quantified using Affymetrix microarrays. Differential gene expression was identified using a z-score cutoff of > 2, and profiles were evaluated against the reference profile database using JG scoring (Jiang and Gentleman, 2007). Finally, a rank based permutation was applied to identify targets enriched among the significantly associated reference profiles.



▲ Figure 3: Radial plot to visually depict the results of the rank permutation analysis identifying significantly enriched targets from MCF7 cells treated with valproic acid (1, 10, 100 µM; light to dark blue, respectively). The target biomolecules are color-coded by family (red: nuclear receptors; purple: enzymes, green: G-protein coupled receptors; blue: ion channels).

Table 2: Summary of CMAP Target Identification

Chemical	Target	HepaRG	MCF7
Clobetasol	GR	0	3
Clofibrate	PPAR	0	0
DHEA	AR	0	ER
DEHP	anti-AR	0	ER
dihydroxyvitamin D3	VDR	0	0
Ethynyl Estradiol	ER	0	3
Flutamide	anti-AR	0	ER
Genistein	ER	0	2
Ketoconazole	CYP	0	0
Mifepristone	anti-PR	0	0
Phenobarbital	CAR/PXR	2	0
Progesterone	PR	0	0
Retinoic Acid	RAR	3	2
Tamoxifen	anti-ER	0	1
Thyroxine	TR	0	0
Trenbolone	AR	0	ER
Troglitazone	PPAR	0	0
Valproic Acid	HDAC	2	1
Vinblastine	TUB	2	3
Vorinostat	HDAC	3	3

Only chemicals with MOAs represented in the reference database are presented. Correctly identified putative targets are highlighted.

Summary & Future Directions

- The Connectivity Map (CMAP) approach was customized to enable optimization/facilitation of modular analyses wherein different statistical approaches can be applied for the identification of positive and negatively correlated gene expression profiles from a reference database
- A rank permutation was introduced to help identify significant biochemical targets putatively mediating the effects of chemicals based on the output from the CMAP ranks
- Results reveal that the limits of this approach for toxicogenomics include:
  - Sufficient target molecule coverage must be included in the reference database for significant association to be detected
  - For chemicals with weak effects or multiple targets, the CMAP approach does not always robustly identify targets
- Future goals include the integration of effect direction (ie. agonism or antagonism of targets) as well as increasing target coverage in the reference profile database