OXFORD

Databases and ontologies

MACE: mutation-oriented profiling of chemical response and gene expression in cancers

Euna Jeong[1,†], Ningning He[2,†], Hyerin Park[1], Mee Song[1], Nayoung Kim[2], Seongjoon Lee[1] and Sukjoon Yoon[1,2,*]

[1]Center for Advanced Bioinformatics and Systems Medicine and [2]Department of Biological Sciences, Sookmyung Women's University, Seoul 140-742, Republic of Korea

*To whom correspondence should be addressed

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Abstract

Summary: The mutational status of specific cancer lineages can affect the sensitivity to or resistance against cancer drugs. The MACE database provides web-based interactive tools for interpreting large chemical screening and gene expression datasets of cancer cell lines in terms of mutation and lineage categories. GI50 data of chemicals against individual NCI60 cell lines were normalized and organized to statistically identify mutation- or lineage-specific chemical responses. Similarly, DNA microarray data on NCI60 cell lines were processed to analyze mutation- or lineage-specific gene expression signatures. A combined analysis of GI50 and gene expression data to find potential associations between chemicals and genes is also a capability of this system. This database will provide extensive, systematic information to identify lineage- or mutation-specific anticancer agents and related gene targets.

Availability and implementation: The MACE web database is available at http://mace.sookmyung.ac.kr/.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: yoonsj@sookmyung.ac.kr

1 Introduction

Various drug responses are induced by somatic mutations in cancers. Approaches to the analysis of gene functions related to drug sensitivity are becoming a popular strategy for overcoming drug resistance and for predicting drug sensitivity (Tan, *et al.*, 2010). Some efforts have focused on the lineage-dependent association between cell lines and drug response (Amatschek, *et al.*, 2004; Dawany and Tozeren, 2010) and the identification of mutation-dependent compound sensitivity (He, *et al.*, 2013a; Kim, *et al.*, 2012). Currently, many small molecule databases have been developed, including several of the following databases: chemical information with biological activity and target information (Hastings, *et al.*, 2013; Law, *et al.*, 2014; Seiler, *et al.*, 2008), chemical search engines with a variety of cheminformatics tools (Chen, *et al.*, 2007; Goede, *et al.*, 2005; Hsin,

et al., 2011; Masciocchi, *et al.*, 2009), and chemical information with extended response profiles on biological samples (Schmidt, *et al.*, 2009; Wang, *et al.*, 2014). Most of these databases are focused on the collection of chemical information, not organized for quantitative analysis. Some studies integrating drug response with genomic data have been established on a plentiful of cell lines, but for limited numbers of chemicals (Barretina, *et al.*, 2012; Garnett, *et al.*, 2012). Thus, we developed a web-based database, MACE (Mutation-oriented Analysis of Chemical response and gene Expression), to provide mutation- or lineage-oriented profiling of chemical response and gene expression on a wide variety of cancer cell lines from NCI60 datasets.

The NCI60 by the Developmental Therapeutics Program (DTP) of NCI/NIH provides anticancer compound screening data for
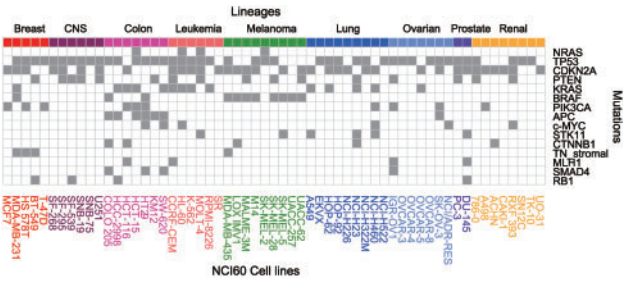
Fig. 1. Distribution of lineages and mutations in NCI60 cell lines

NCI60 cancer cell lines. The quantified response data, GI50—the concentration that inhibits growth by 50%, are available for more than 50,000 chemical compounds (Scudiero, *et al.*, 1988). This resource can be used to find relationships among compound structure, mechanism of action, cell lineage, and tumor mutations (He and Yoon, 2012). The NCI60 cell lines represent nine cancer lineage types (breast, CNS—central nervous system, colon, leukemia, melanoma, lung, ovarian, prostate and renal). MACE includes top 15 most frequently mutated genes which are found in more than 4 cell lines (TP53, CDKN2A, PTEN, KRAS, BRAF, PIK3CA, APC, c-MYC-Amp, STK11, CTNNB1, SMAD4, RB1, MLH1, NRAS and TN_stromal) (Figure 1). Mutations in these genes are known to be involved in cancer-related cellular pathways. TP53 is a tumor suppressor, and somatic mutations in TP53 are the most common genetic changes found in human cancer (Olivier, *et al.*, 2010). Defects in the gene cyclin-dependent kinase inhibitor 2A (CDKN2A) is involved in the genesis of many tumor types (Sulong, *et al.*, 2009). PTEN is a tumor suppressor, and PIK3CA is an oncogene. These proteins make up the PTEN–PI3K signaling pathway, which is involved in the regulation of many cancer related pathways (Carracedo and Pandolfi, 2008). KRAS, NRAS and BRAF are oncogenes involved in the ErbB signaling pathway, which mediates cell growth and malignant transformation through kinase pathway activation (Brose, *et al.*, 2002). APC is a tumor suppressor, and CTNNB1 is an oncogene. Mutations in these genes cause defects in Wnt signaling (Polakis, 2012). High expression of the oncogene c-MYC (c-MYC-Amp) and mutational inactivation of the tumor suppressor RB1 cause dysregulation of cell cycle control (Lin, *et al.*, 2012). STK11 has been identified as a tumor suppressor involved in the development of cancer, especially in lung cancer (Sanchez-Cespedes, *et al.*, 2002). Mutation of SMAD4 results in disruption of TGFbeta signaling and occurs frequently in association with malignant progression (Miyaki and Kuroki, 2003). Genetic mutation of MLH1 is associated with colon cancer risk (Bonadona, *et al.*, 2011), and activated expression of TN_stromal is related to human breast cancer (Ishihara, *et al.*, 1995).

MACE provides tools for 1) category-based analysis of chemical response or gene expression in cancer cell lines (Table 1), 2) finding a correlation between chemical response and gene expression on NCI60 cell lines, and 3) 2D- or 3D-based chemical structural searches in real-time. The term "category" refers to mutational or lineage groups of cell lines, and thus "a specific category" indicates a specific mutation or lineage (e.g. BRAF or lung).

2 MACE overview

The overall data preprocessing and data flow of MACE are depicted in Figure 2. We collected chemical response and gene expression data for NCI60 cell lines, and chemical structure information as the

Table 1. Types of profiles analyzed in MACE

	59 Cell lines	Category	
		15 Mutations	Nine lineages
Chemicals	z-score	Fold change	
Genes	$\log_2$(gene expression)	$\log_2$(fold change)	

For the 50 839 chemicals (75 446 records) and the 54 613 gene probes, the three types of profiles are analyzed. Note that the cell line MDA-MB-468 is excluded. Data are obtained as follows: i) z-score by normalization of $-\log$GI50, ii) $\log_2$ (gene expression) by $\log_2$ transformation of RMA normalization, iii) fold change and $\log_2$ (fold change) by subtracting the median of all cell lines from the average of in-category cell lines based on z-score and $\log_2$ (gene expression), respectively.

primary dataset. NCI60 drug response data were obtained from the NCI DTP (http://dtp.nci.nih.gov). The latest version, released in 2012, contains 50,839 compounds (75,446 records) with negative log-transformed GI50 values to characterize sensitivity across NCI60 cancer cell lines. The distribution pattern of average –log GI50 values for total chemicals are displayed in Supplementary Figure 1. The figure shows that NCI60 response data were biased in leukemia cell lines. To eliminate this bias, z-score normalization was used against individual cell lines. Furthermore, microarray gene expression data for the NCI60 cancer cell lines were obtained from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) dataset series GSE32474 (Pfister, *et al.*, 2009). This dataset was derived from 174 arrays performed in triplicate for NCI60 cancer cell lines. The experiments were carried out with the Affymetrix U133 Plus 2.0 Array chip, which includes 54,613 probes. The data were normalized using a Robust Multi-array Average (RMA) algorithm (Irizarry, *et al.*, 2003). In brief, the data preprocessing includes normalization of the primary data, organization of category-oriented profiles, building 2D structure and 3D shape databases for chemicals, and calculation of cross correlation between chemicals, between genes, and between chemicals and genes.

MACE provides multiple query methods for chemical profiles using simple text (including chemical name, NSC ID—the NCI's internal ID number, and CAS number) or structure, and for gene expression profiles, both gene symbol and gene probe ID are used (Figure 3). A simple text search for a chemical retrieves a particular chemical structure and its information. The structure search has options to select 2D structure- and 3D shape-based similarity searches for a chemical structure drawn via the structure editor. In addition, groups of structures can be searched by filtering chemicals satisfying user-defined criteria in terms of response difference and statistical significance in a specific category. Likewise, groups of genes can also be retrieved by filtering genes satisfying user-defined conditions.

From the retrieved results, three types of profiles are available for both datasets: mutation-based, lineage-based, and cell line-based profiles (Table 1). For each profile, the data are visualized as three types of graphs including a bar graph, a heatmap, and a waterfall graph (Figure 4).

3 Category-specific analysis of chemical response and gene expression

Category-specific profiles show differential sensitivity of chemical response or gene expression in a group of cell lines within a specific
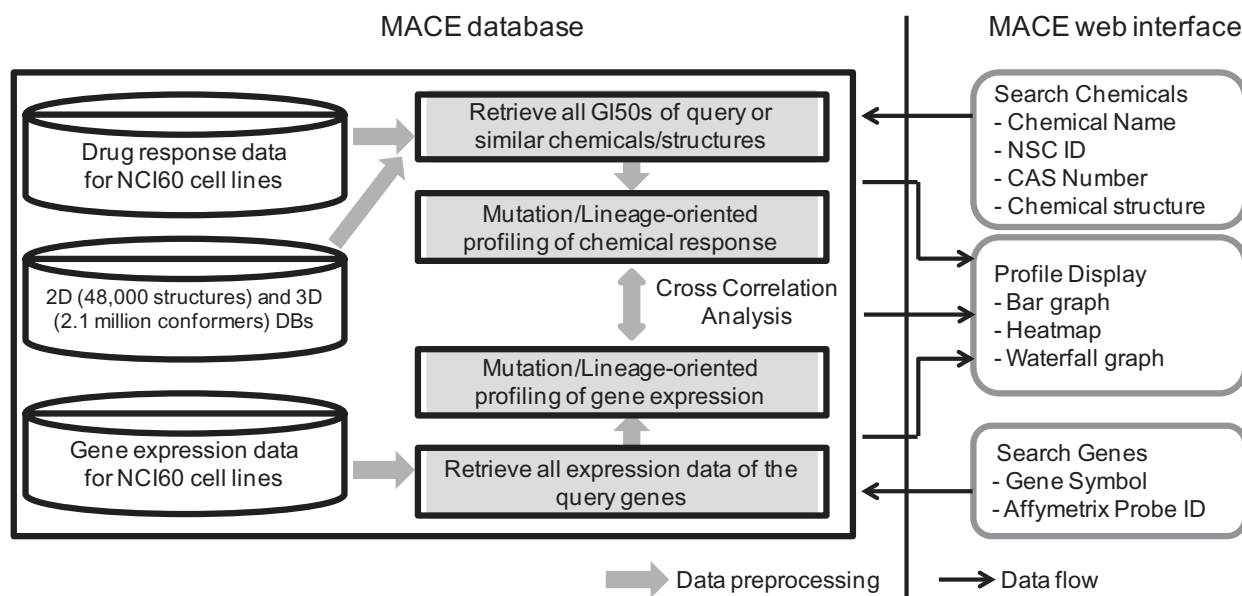
Fig. 2. Overall diagram of MACE. Drug response data (GI50) and gene expression data (DNA microarray) for NCI60 cell lines are normalized and organized to represent category-specific profiles. Through the user interface, three different types of profiles are visualized, and similar profile searches between chemicals and genes are available
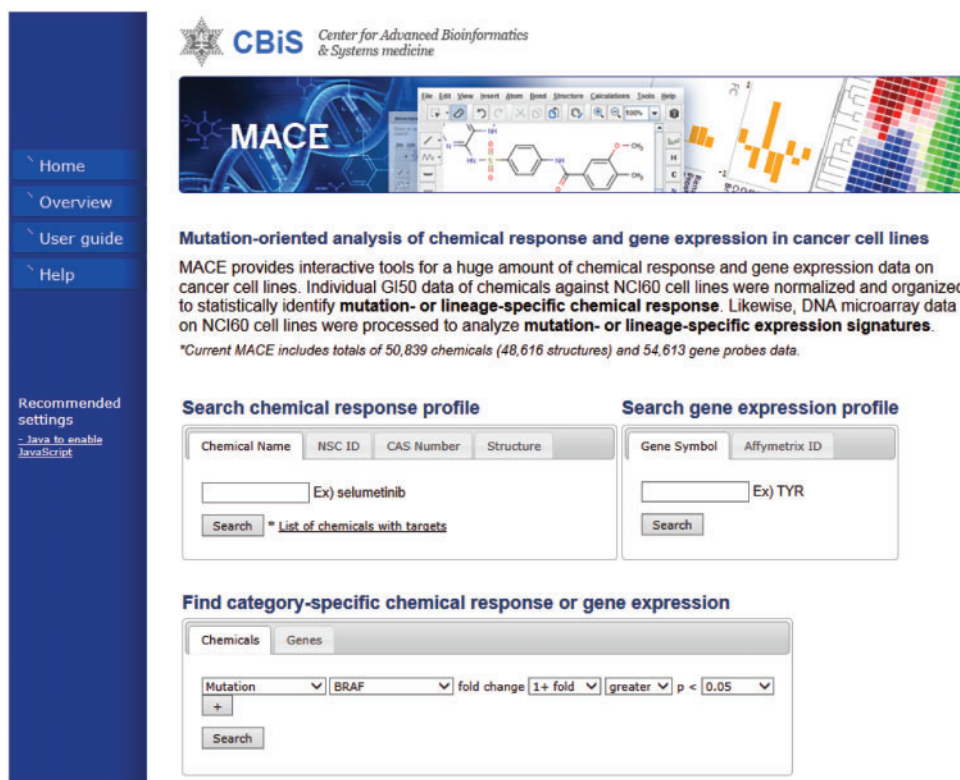
Fig. 3. A screen shot of MACE including chemical search, gene search and filtering search

category compared to all cell lines. For a given chemical, the fold change was calculated by subtracting the median z-score of all cell lines to the average z-score of in-category cell lines (see Table 1). Likewise, for a given gene, the $\log_2$(fold change) was calculated by subtracting the median $\log_2$(gene expression) of all cell lines to the average $\log_2$(gene expression) of in-category cell lines. A t-test

P-value is used to evaluate the statistical significance of the fold change. In addition, we also calculated enrichment scores using an odds ratio between the observed odds and the expected odds to describe specificity. The observed odds is the ratio of the number of cell lines with a significant difference (fold change>1 and $P < 0.01$) in a specific category to the number of all cell lines in that category.
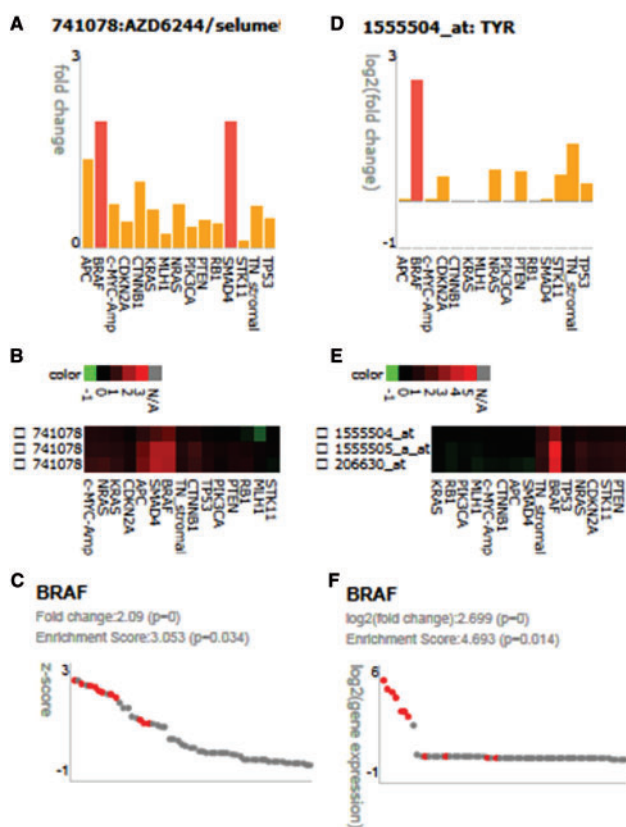
Fig. 4. Mutation-oriented analysis of chemicals and gene expression. Examples of mutation-based profiles (**A–C**) for a chemical, selumetinib (NSC ID: 741078) and (**D–F**) for a gene, TYR (probe ID: 1555504_at), are displayed. (A, D) Bar graphs for each mutation ($P < 0.01$ as indicated by red colored bars). (B, E) Heatmap display for mutation categories. (C, F) Waterfall graphs for BRAF-mutant cell lines compared with other cell lines. The horizontal axis represents all cell lines sorted by the *y*-axis value and the red dots indicate the cell lines with the BRAF mutation

The expected odds is the ratio of cell lines with significant differences (fold change>1 and $P < 0.01$) among the NCI60 cell lines.

For example, the chemical selumetinib can be retrieved by typing its name or NSC ID, 741078 (Figure 3). Then, from the chemical structure information, one of three profiles can be selected. Examples of mutation-based profiles are shown in Figure 4. At first, the profile is displayed in a bar graph (Figures 4A). The bar graphs show the response difference for each of 15 mutations. In the graph, the names of mutant genes are listed on the *x*-axis and the chemical response is represented on the *y*-axis in terms of fold change (i.e., difference of normalized –log GI50 between a mutant subgroup and all cell lines). The red bar in the graph indicates a significance level (*P* value) less than 0.01 in the corresponding mutation group. Multiple bar graphs per chemical may be shown if there are redundant experiments for the given chemical. In the next step, all profiles are clustered and displayed in the form of a heatmap (Figures 4B). For selumetinib, there were three repeated experiments, which consistently show BRAF and SMAD4 sensitivity among 15 mutations. Clicking the bar graph or any row on the heatmap will direct to waterfall graphs. For mutation-based profiles, 15 waterfall graphs for 15 mutant genes are displayed. In this example, the waterfall graph for BRAF is shown in Figure 4C. In each waterfall graph, the response difference and the enrichment score are shown with each corresponding significance level. For another example, the TYR

gene, a similar analysis can be performed. The gene symbol TYR is mapped to three gene probes in the microarray dataset. The expression difference of TYR is represented in terms of log2 (fold change) for 15 mutations. TYR overexpression is highly associated with BRAF mutation (Figure 4D). Heatmap shows that all of three TYR probes consistently exhibit high expression in BRAF mutant cell lines (Figure 4E). In the waterfall graph of Figure 4F, cell lines with BRAF mutation (red dots) are highly enriched for the expression of the TYR gene.

Similarly, for lineage-based profiles, the same format of the data for each of 9 lineages will be shown for chemical response and gene expression. MACE also displays category-free bar graphs and heatmaps for all cell lines. In the case of cell line-based profiles, no waterfall graphs are provided.

4 2D and 3D similarity searches

MACE provides two types of similarity search for chemicals: 2D similarity calculated from structure fingerprints and 3D shape similarity based on volume matches. MACE supports similarity, substructure and superstructure searches for 2D structures against over 48,000 chemical structures. For 2D structure-based similarity searches, the query and the target structures are encoded into molecular fingerprints in the FP2 format using OpenBabel (O'Boyle, *et al.*, 2011). A 2D structure fingerprint is comprised of bit string (a sequence of 0's and 1's) to denote the presence of a particular substructure in a molecule. The 2D similarity is measured using the Tanimoto score which quantifies topological similarity of chemical bonding and atomic constitution in the range of 0 to 1. Chemicals with a degree of similarity 0.5 or more will be retrieved in MACE. The substructure and superstructure searches locate chemical structures that contain particular substructures and superstructures of a query chemical, respectively. For 3D shape-based similarity searches, the 3D conformer database has been built using OMEGA (Hawkins and Nicholls, 2012), which consists of up to 2.1 million conformers for ~48,000 chemicals. The maximum number of conformers per chemical was 200 for the 3D conformer database and 50 for a query, and the RMSD cutoff between conformers was 0.5 Å. The 3D shape search runs on high-end commodity GPU hardware to enable real-time shape similarity calculations. The 3D similarity measure, TanimotoCombo, is the sum of shape and color Tanimotos ranging from 0 to 2 implemented in FastROCS (Hawkins, *et al.*, 2007). The most similar 100 conformers for a query chemical are returned as hits.

The 2D and 3D similarity searches are available via the structure editor (Figure 5A). MarvinSketch is used for the structure editor, which is developed by ChemAxon (http://www.chemaxon.com). Users can draw or load structures and execute one of two searches. Figure 5 shows some of the selected chemicals with the highest similarity range from the results of 2D and 3D searches for a single query chemical. For the retrieved chemicals of the search, all three types of profiles and three formats of graphs are available as well. In the present example, while 2D and 3D searches retrieve different types of chemicals, they consistently exhibit a significant sensitivity for PIK3CA mutant cell lines. The results also show similar patterns in the heatmaps of mutation-based profiles.

5 Profile-based similarity search

Another functionality of MACE is combinational analysis of drug response and gene expression data. Cross correlation analysis is
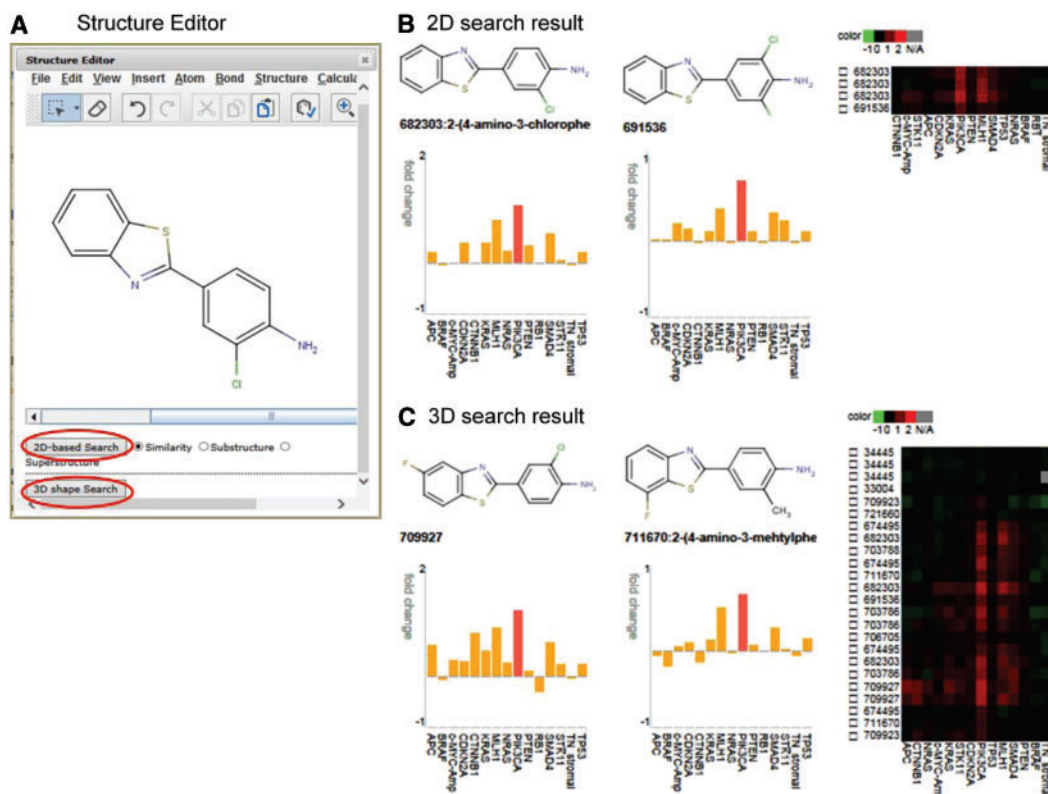
Fig. 5. 2D and 3D chemical search powered by GPU machine. (**A**) After drawing or loading a chemical in the given structure editor, clicking the 2D-based or the 3D-based shape search button will return the summary of hit chemicals and their cellular response data in three profile types. Examples of 2D and 3D search results for a query chemical are shown in (**B**, **C**), respectively, in the order of chemical structure and mutation-based response profiles in bar graphs and heatmaps

available via profile-based similarity search. MACE searches chemicals/genes with similar profiles by comparison of cell line-based profiles between all pairs of chemicals and genes. Though CellMiner provides similarity analysis among drug response, gene and microRNAs expression values, the result shows gene/drug list and Pearson correlation coefficient (PCC) values (Reinhold, *et al.*, 2012). MACE presents a heatmap of all results and makes the results more clear. The profile-based similarity measure is based on absolute PCC on all cell lines. Chemicals or genes with a correlation value ≥ 0.6 and a *P*-value < 0.01 will be returned. Users can start the search from the heatmap of the cell line-based profiles. For the query chemical, the heatmap of cell line-based profiles and two search buttons are shown in the top in Figure 6. Selecting one profile on the heatmap first and clicking one type of search will retrieve cell line-based profiles of hits with profile-based similarity measures (absolute PCC) in bar graph format. An example of hits from similar chemical and gene searches are shown in the middle and in the bottom of the Figure 6, respectively. The example results exhibited similar profiles and also show a significant sensitivity for BRAF mutations.

6 Summary

MACE is a useful tool that integrates drug response and gene expression data for the purpose of mutation- or lineage-based analysis. MACE provides fold changes for chemical responses and gene expression together with enrichment scores as statistical methods to

find unknown sensitivity or selectivity of chemical response (or gene expression) on given mutation or lineage categories.

MACE provides a powerful 3D shape-based algorithm for chemical searches as well as for typical 2D searches. 3D shape comparison is a promising new approach for studying structure-activity relationships in general drug discovery, overcoming the limitations of chemical diversity in conventional 2D searches (He, *et al.*, 2013a). In a previous study, the 3D shape-based metric could provide better resolution than the 2D structure-based method for identifying compound pairs with similar cellular response profiles (He, *et al.*, 2013b). Adapting both 2D and 3D search methods, MACE provides a flexible chemical search for the identification of various chemical entities with similar profiles of cellular response in cancers.

Furthermore, MACE attempts to integrate chemical response and gene expression data based on profile similarities over NCI60 cell lines. MACE enables identification of many unexpected relationships of gene-gene, chemical-chemical or gene-chemical pairs that potentially exhibit common underlying mechanisms for chemical response and/or gene expression. MACE also provides browsing options for identification of highly sensitive or resistant chemicals or signature genes to a given category, and as such, allows users to exploit groups of chemicals with similar mutation- or lineage-specific responses and genes with similar specific differences of expression.

We believe that MACE will serve as an advanced tool in cancer research, not only for the discovery of novel anticancer candidates but also for understanding specific mechanisms of drug action and gene expression.
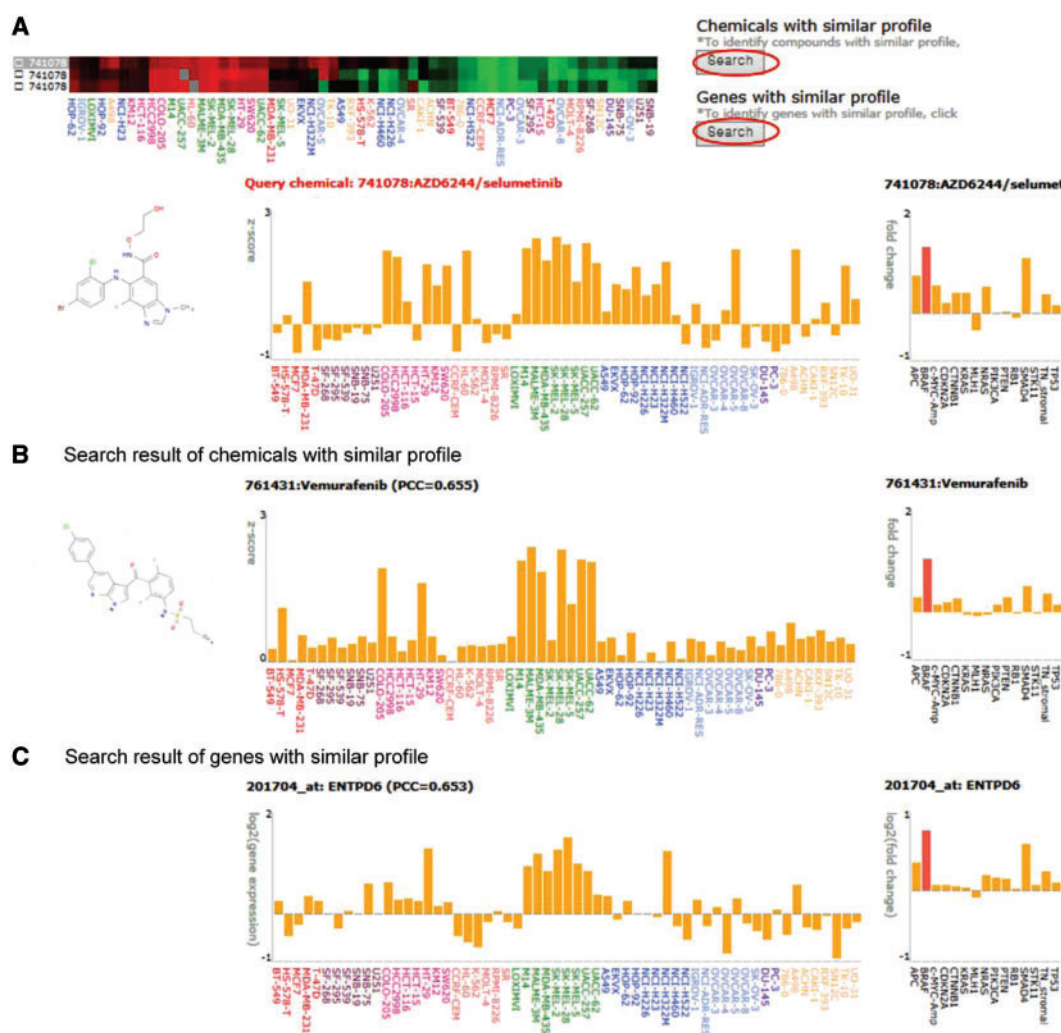
Fig. 6. Combined analysis of chemical response and gene expression data. For a query chemical in (**A**), genes or chemicals with similar NCI60 profiles are search-able. Clicking a 'search' button will retrieve cell line-based profiles as results for chemicals in (**B**) and genes in (**C**). For comparison, chemical structures for chemicals, cell line-based profiles and mutation-based profiles of a query and the query results are listed. In the bar graphs, the cell line name is colored depending on the lineage. Hovering over a cell line name will display the lineage information on the same window

Funding

Conflict of Interest: none declared.

References

Amatschek,S. *et al.* (2004) Tissue-wide expression profiling using cDNA sub-traction and microarrays to identify tumor-specific genes. *Cancer Res.*, **64**, 844–856.

Barretina,J. *et al.* (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607.

Bonadona,V. *et al.* (2011) Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA* **305**, 2304–2310.

Brose,M.S. *et al.* (2002) BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Res.*, **62**, 6997–7000.

Carracedo,A. and Pandolfi,P.P. (2008) The PTEN-PI3K pathway: of feedbacks and cross-talks. *Oncogene* **27**, 5527–5541.

Chen,J.H. *et al.* (2007) ChemDB update—full-text search and virtual chemical space. *Bioinformatics* **23**, 2348–2351.

Dawany,N.B. and Tozeren,A. (2010) Asymmetric microarray data produces gene lists highly predictive of research literature on multiple cancer types. *BMC Bioinformatics* **11**, 483.

Garnett,M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575.

Goede,A. *et al.* (2005) SuperDrug: a conformational drug database. *Bioinformatics* **21**, 1751–1753.

Hastings,J. *et al.* (2013) The ChEBI reference database and ontology for bio-logically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.

Hawkins,P.C. and Nicholls,A. (2012) Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J. Chem. Inf. Model.*, **52**, 2919–2936.

Hawkins,P.C. *et al.* (2007) Comparison of shape-matching and docking as vir-tual screening tools. *J. Med. Chem.*, **50**, 74–82.

He,N. and Yoon,S. (2012) Gene expression signatures for compound response in cancer. *Genom. Inform.*, **9**, 8.

He,N. *et al.* (2013a) Somatic mutation patterns and compound response in cancers. *BMB Rep.*, **46**, 97–102.

He,N. *et al.* (2013b) 3D shape-based analysis of cell line-specific compound response in cancers. *J. Mol. Graph. Model.*, **43**, 41–46.

Hsin,K.Y. *et al.* (2011) EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities. *Nucleic Acids Res.*, **39**, D1042–D1048.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Ishihara,A. *et al.* (1995) Tenascin expression in cancer cells and stroma of human breast cancer and its prognostic significance. *Clin. Cancer Res.*, **1**, 1035–1041.

Kim,N. *et al.* (2012) Systematic analysis of genotype-specific drug responses in cancer. *Int. J. Cancer* **131**, 2456–2464.

Law,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

Lin,C.Y. *et al.* (2012) Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67.

Masciocchi,J. *et al.* (2009) MMsINC: a large-scale chemoinformatics database. *Nucleic Acids Res.*, **37**, D284–D290.

Miyaki,M. and Kuroki,T. (2003) Role of Smad4 (DPC4) inactivation in human cancer. *Biochem. Biophys. Res. Commun.*, **306**, 799–804.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

Olivier,M. *et al.* (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.*, **2**, a001008.

Pfister,T.D. *et al.* (2009) Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. *Mol. Cancer Ther.*, **8**, 1878–1884.

Polakis,P. (2012) Wnt signaling in cancer. *Cold Spring Harb. Perspect. Biol.*, **4**, a008052.

Reinhold,W.C. *et al.* (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.*, **72**, 3499–3511.

Sanchez-Cespedes,M. *et al.* (2002) Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. *Cancer Res.*, **62**, 3659–3662.

Schmidt,U. *et al.* (2009) SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.*, **37**, D295–D299.

Scudiero,D.A. *et al.* (1988) Evaluation of a soluble tetrazolium/formazan assay for cell growth and drug sensitivity in culture using human and other tumor cell lines. *Cancer Res.*, **48**, 4827–4833.

Seiler,K.P. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.

Sulong,S. *et al.* (2009) A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* **113**, 100–107.

Tan,D.S. *et al.* (2010) Anti-cancer drug resistance: understanding the mechanisms through the use of integrative genomics and functional RNA interference. *Eur. J. Cancer* **46**, 2166–2177.

Wang,Y. *et al.* (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**, D1075–D1082.