

SCIENTIFIC REPORTS

OPEN

Computational Drug Repositioning for Gastric Cancer using Reversal Gene Expression Profiles

In-Wha Kim¹, Hayoung Jang¹, Jae Hyun Kim¹, Myeong Gyu Kim², Sangsoo Kim³ & Jung Mi Oh¹ 

Treatment of gastric cancer (GC) often produces poor outcomes. Moreover, predicting which GC treatments will be effective remains challenging. Computational drug repositioning using public databases is a promising and efficient tool for discovering new uses for existing drugs. Here we used a computational reversal of gene expression approach based on effects on gene expression signatures by GC disease and drugs to explore new GC drug candidates. Gene expression profiles for individual GC tumoral and normal gastric tissue samples were downloaded from the Gene Expression Omnibus (GEO) and differentially expressed genes (DEGs) in GC were determined with a meta-signature analysis. Profiles drug activity and drug-induced gene expression were downloaded from the ChEMBL and the LINCS databases, respectively. Candidate drugs to treat GC were predicted using reversal gene expression score (RGES). Drug candidates including sorafenib, olaparib, elesclomol, tanespimycin, selumetinib, and ponatinib were predicted to be active for treatment of GC. Meanwhile, GC-related genes such as *PLOD3*, *COL4A1*, *UBE2C*, *MIF*, and *PRPF5* were identified as having gene expression profiles that can be reversed by drugs. These findings support the use of a computational reversal gene expression approach to identify new drug candidates that can be used to treat GC.

Gastric cancer (GC) is the fifth most common cancer worldwide and the third leading cause of cancer death, with 1.3 million incident cases and 819,000 deaths occurring globally in 2015¹. Although GC rates have declined in most developed countries, the incidence of non-cardia GC among Caucasians aged 25–39 years has increased in the United States over the past two decades².

Increased rates of early GC detection have increased survival rates for GC patients, but treatment outcomes for GC remain low and difficult to predict³. Moreover, GC is a highly heterogeneous disease as reflected by the numerous histological and molecular classifications⁴.

The development of new drugs to treat diseases, especially cancer, is dependent on the identification of novel drug targets. In recent years, an increasing number of innovations have promised to improve our understanding of disease biology, provide novel targets, and catalyze a new era in the development of medicines. However, despite impressive advances in technologies, the situation has remained relatively static in terms of new molecular entities⁵. After some success in targeted therapies for the treatment of several human cancers^{6,7}, research has focused more on new approaches for the identification of novel targets in cancer therapy. Although large numbers of potential targets have been identified by advanced technologies, it has proven difficult to find targets that are causally involved in the disease.

The number of drugs approved by the US Food and Drug Administration has continuously declined because traditional methods of drug development do not support highly efficient drug discovery. Traditional approaches to develop of new drugs are expensive and time-consuming, with an average of 15 years and a price tag of more than \$2 billion necessary to bring a drug to market^{8,9}. Over 90% of drugs fail during the early development stage due to safety concerns or a lack of efficacy¹⁰.

The increasing availability of large public datasets such as the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI)¹¹, the Cancer Cell Line Encyclopedia (CCLE)¹², DrugBank¹³, and the Library of Integrated Network-Based Cellular Signatures (LINCS)^{14,15}, which together catalog disease-specific and

¹College of Pharmacy and Research Institute of Pharmaceutical Sciences, Seoul National University, Seoul, Republic of Korea. ²Graduate School of Clinical Pharmacy, CHA University, Pocheon, Republic of Korea. ³Department of Bioinformatics and Life Science, Soongsil University, Seoul, Republic of Korea. In-Wha Kim and Jung Mi Oh Contributed equally. Correspondence and requests for materials should be addressed to I.-W.K. (email: iwkim2@hanmail.net) or J.M.O. (email: jmoh@snu.ac.kr)

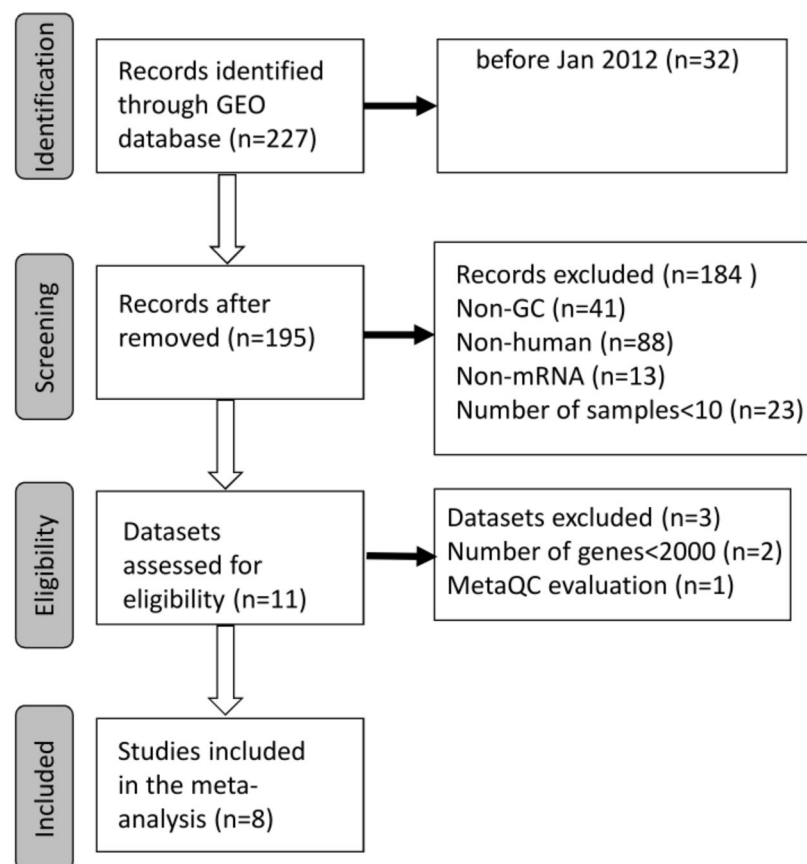


Figure 1. Flowchart of the selected process to select gene expression datasets for meta-analysis of gastric cancer. GEO, Gene Expression Omnibus; GC, gastric cancer.

drug-induced gene expression signatures, offers a time-efficient approach to reposition existing drugs for new indications^{9,16}. Several computational methods, such as bioinformatics, system biology, machine learning, and network analysis can be used for drug repositioning or repurposing as well as to identify new indications for drugs¹⁷.

Most computational drug repositioning approaches are based on a “guilt by association” strategy¹⁸, wherein agents having similar properties are predicted to have similar effects. Many drug repositioning strategies are based on different data, including similar chemical structures, genetic variations, and gene expression profiles¹⁹. Recently, interest in the use of genomics-based drug repositioning to aid and accelerate the drug discovery process has increased⁹. Drug development strategies based on gene expression signatures are advantageous in that they do not require a large amount of a priori knowledge pertaining to particular diseases or drugs^{20,21}.

The purpose of this study is to predict drug candidates that can treat GC using a computational method that integrates publicly available gene expression profiles of GC patient tumors and GC cell lines and cellular drug response activity profiles.

Results

Short Overview of Included Studies. The study selection process is outlined in Fig. 1. Following the search and selection steps, eight studies: GSE2689, GSE29272, GSE30727, GSE33335, GSE51575, GSE63089, GSE63288, and GSE65801, were included in the final analysis. An additional dataset, GSE54129, was excluded due to lower quantitative QC scores after a MetaQC analysis (Supplementary Table S1). Detailed information about the downloaded datasets is summarized in Supplementary Table S2. Tumor gene expression signatures were analyzed for 719 GC samples by comparing RNA expression data for 410 tumors and 326 adjacent normal tissues from the GEO. The samples originated from 410 patients, of whom 152 (37.1%) were Korean, 236 (57.6%) were Chinese, and 22 (5.4%) were Caucasians. The samples of patients who had no prior therapy were from GSE29272, GSE65801, and GSE63288. The sample information was not available in GSE30727 nor GSE26899, while the sample information was not mentioned in GSE33335 nor GSE51575. All patients received some type of pre-treatment in GSE63089.

Tumor Gene Expression Signatures. The workflow for the exploration of compounds using the calculated RGES values is presented in Fig. 2. All probe sets were re-annotated with the most recent NCBI Entrez Gene IDs and then mapped manually to yield 9,113 unique common genes across the different platforms. A fixed-effect model method was utilized by combining the *P* values in the MetaDE package. Among the gene expression signatures, 136 genes showed increased expression levels in tumors compared to normal tissues (adjusted *P* < 0.001,

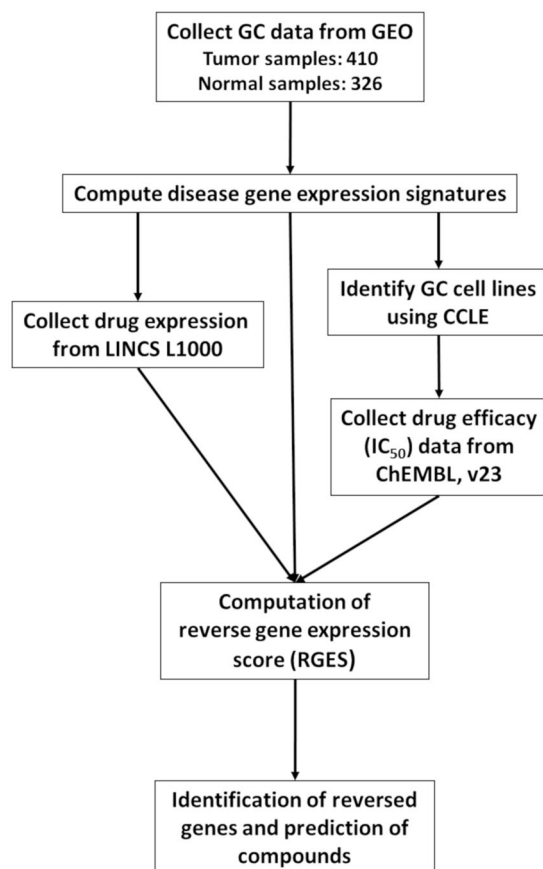


Figure 2. Workflow to determine the reverse gene expression score (RGES) using disease and drug gene expression profiles. The public database GEO was used to create cancer gene expression signatures; LINCS L1000 was used as the drug signature database; ChEMBL was used as the drug activity database; and CCLE was used to map cell lines among databases. GEO, Gene Expression Omnibus; GC, gastric cancer; LINCS, The Library of Integrated Network-Based Cellular Signatures; ChEMBL, a chemical database of bioactive molecules maintained by the European Bioinformatics Institute of the European Molecular Biology Laboratory; CCLE, Cancer Cell Line Encyclopedia.

$\log_2\text{foldchange} > 1.2$), whereas 53 genes showed decreased expression levels in tumors compared to normal tissues (adjusted $P < 0.001$, $\log_2\text{foldchange} > 1.2$; Supplementary Table S3).

Similarity in Gene Expressions between Tumor Samples and GC Cell Lines. The degree of similarity in the gene expressions between tumor samples from the GEO and GC cell lines from the CCLE was assessed by a ranked based Spearman's correlation test. Gene expression profiles for 40 GC cell lines were included in the CCLE (Supplementary Table S4). The top 5000 genes in these cell lines were ranked according to their interquartile range across all cell lines used. Among them, less than 0.3% of genes had expression levels in tumor samples from the GEO that did not correlate with those in all GC cell lines. These genes included the ECC10 (0.62%), ECC12 (0.34%), and HGC (0.32%) cell lines (adjusted P value < 0.05).

RGES Computation. LINCS data for changes in the expressions of 978 landmark genes after treatment of AGS cell lines with 25 compounds used to treat human gastric adenocarcinoma were used for the RGES computations. The median IC_{50} s values for 2025 compounds used to treat GC cancer cell lines listed in the ChEMBL were used for computation. Disease signatures including 189 DEGs after extraction from the set of LINCS landmark genes were also used for the RGES computation. Variations in the RGES outcomes were evaluated under various biological conditions. The RGES showed larger variations across different cell lines relative to those within different replicates of the same cell line when the same concentration and treatment duration for a compound were used ($P < 2.2 \times 10^{-16}$; Fig. 3A). In addition, longer treatment durations (≥ 24 h) were associated with lower RGES outcomes compared to shorter durations (< 24 h) when a compound was tested on the same cell line at the same concentration ($P < 2.2 \times 10^{-16}$; Fig. 3B). Likewise, higher compound concentrations ($\geq 10 \mu\text{M}$) had lower RGES values than lower concentrations ($< 10 \mu\text{M}$, $P < 2.2 \times 10^{-16}$; Fig. 3C). The RGES values for the compounds were evaluated by examining correlations with their activities in the same cell line. Finally, the RGES outcomes were correlated with the IC_{50} values (Spearman correlation $\rho = 0.3$, $P = 5.61 \times 10^{-3}$; Fig. 4).

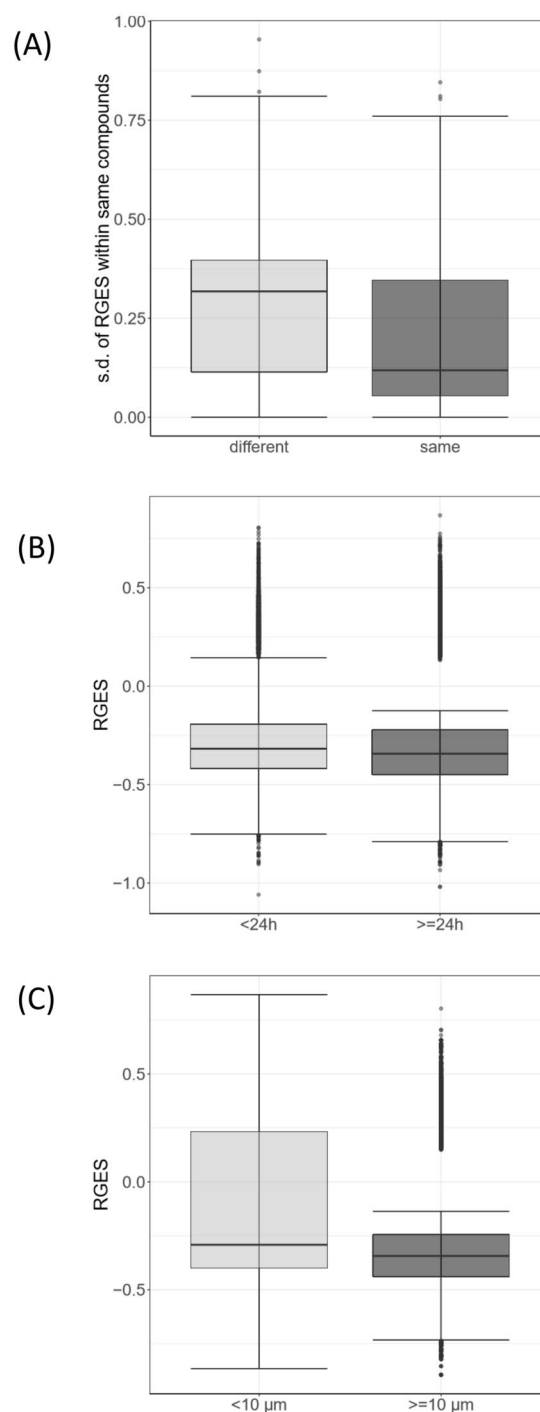


Figure 3. Differences in RGES under the various biological conditions. **(A)** Standard deviation (s.d.) of RGES of individual compounds across different cell lines (grey) vs. across replicates within the same cell line (black grey). **(B)** RGES distribution between treatment durations <24 hr (grey) and ≥24 hr (black grey). **(C)** RGES distribution between drug concentrations <10 μM (grey) and ≥10 μM (black grey). Treatment duration and compound concentration were categorized based on compound profiles in LINCS. *P* values were calculated using a Wilcoxon signed-rank test.

RGES Summarization and Evaluation. Summarized RGES (sRGES) values were computed by weighting various cell lines, compound concentrations, and treatment durations. A number of known methods were used to summarize the RGES and obtain sRGES values (Supplementary Table S5). The calculated sRGES scores for each compound were significantly correlated with drug activity (Spearman correlation $\rho = 0.27$ and $P = 1.04 \times 10^{-2}$; Fig. 5). Additionally, CTRP was used as an external dataset to confirm the correlation between reversal potency and compound activity. Activity data expressed as AUC values for 546 compounds tested in GC cell lines were collected from CTRP. After the sRGES computation, the median AUC values across multiple cell lines were

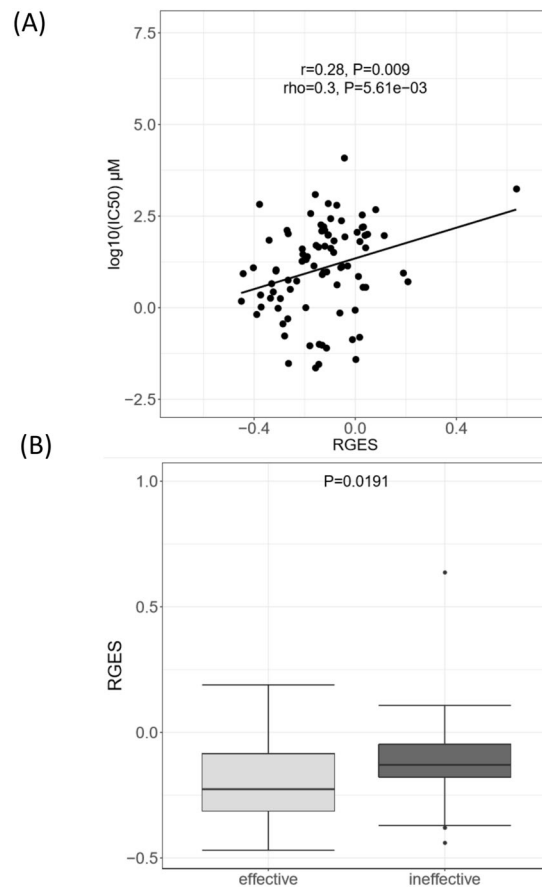


Figure 4. Correlation between drug activity and reverse gene expression score (RGES) in AGS cancer cell lines. (A) Correlation between RGES and drug activity (IC_{50}) by linear regression and Spearman's correlation tests.

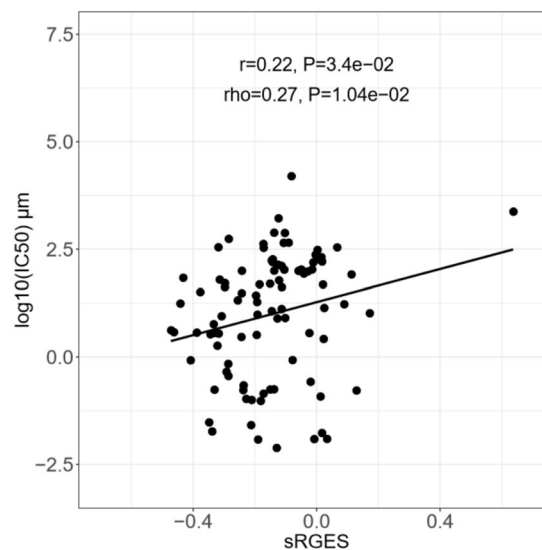
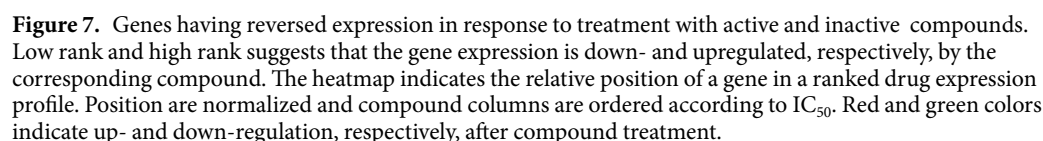
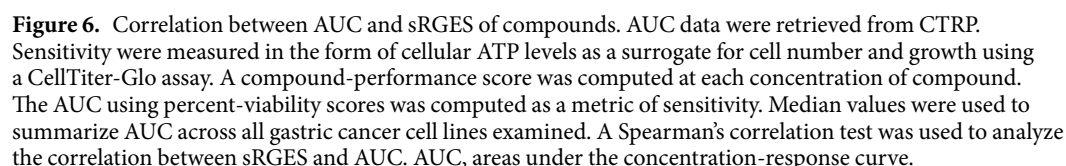


Figure 5. Correlation between drug activity (IC_{50}) and summarized reverse gene expression score (sRGES) for all cancer cell lines using a linear regression and a Spearman's correlation tests.

used to evaluate the sRGES. The sRGES values were significantly correlated with the AUC values ($\rho = 0.368$, $P = 3.8 \times 10^{-8}$; Fig. 6).



Discussion

Our results showed that the ability of drugs to reverse DEGs was correlated with drug activity in GC, although this correlation was highly dependent on the cell line as well as the drug concentration and treatment duration. The positive correlation between sRGES and IC_{50} values indicated that combining disease gene expression data

derived from clinical samples with drug gene expression profiles obtained from results with *in vitro* cell lines could be used to predict drug activity.

In our study, five GC genes, *COL4A1*, *PLOD3*, *UBE2C*, *MIF*, and *PRPF4*, showed reversed expressions in response to 15 active compounds. To the best of our knowledge, this is the first study of drug repositioning using a computational reversal gene expression approach in GC. Among these genes, *PLOD3*²⁶ and *COL4A1*²⁷ were recently shown to be overexpressed in GC. Meanwhile, the overexpression of *UBE2C* was related to poor prognosis in GC²⁸ and was a potential biomarker of intestinal type GC²⁹. *MIF* could also be a potential prognostic factor for GC³⁰. These genes showed reversed expression levels and thus may be feasible as therapeutic targets for GC. Additionally, *PRPF4* as a pre-mRNA splicing factor has been suggested as a potential therapeutic target for cancer therapy³¹.

Among the active drugs identified by our analysis, the multiple tyrosine kinase inhibitor sorafenib³² and a poly (ADP-ribose) polymerase (PARP) inhibitor, olaparib³³, have completed phase II and phase III clinical trials, respectively, for GC patients. Meanwhile, the heat shock protein 70 (Hsp70) inducer elesclomol, the novel tyrosine kinase inhibitor ponatinib, the heat shock protein 70 (Hsp90) inhibitor tanespimycin, and the mitogen-activated protein kinase inhibitor selumetinib have not been previously studied clinically for their effectiveness against GC.

GC is a heterogeneous disease that involves multiple factors associated with various molecular pathways that can function differently during the cancer development process. A limitation of this study is that the GC disease gene expression datasets from the GEO are not uniformly associated with clinical outcomes or GC etiologies. The drug activity of predicted compounds may also vary because the GC disease states varied for individual patients. Sampling time information is important, as samples obtained after the initial neo-adjuvant chemotherapy can affect the results of this meta-analysis. Nonetheless, such information was not available from some datasets.

Many recent projects focus on precision medicine to provide insights between diseases and genes. A repurposing strategy based on alterations of driver genes in each tumor can be used to identify therapeutic targets. The collection of therapeutic agents targeting driver genes and determining the connection between each patient and the targeted therapies can enhance promising drug repositioning opportunities and eventually benefit patients. Therefore, RGES may improve predictions of drug candidates because it is based on the molecular characteristics of actual tumors.

Therapeutic efficacy is more complex than a simple correlation of gene expression profiles with drugs and diseases. Therefore, our findings with regard to drug candidates will require further preclinical testing and demonstrations in clinical trials, although our results did validate that the method of the computational analysis of public gene expression databases is a potentially useful means of drug discovery. In summary, our computational approach combined disease gene expression with drug-induced expression profiles in GC to identify new drugs and target genes for GC therapy. This approach can also be used to predict the efficacy of new drug candidates with which to treat GC. This computational approach could be broadly applied to other diseases for which reliable gene expression data are available.

Materials and Methods

Collection of Gastric Adenocarcinoma Gene Expression Profiles. Publicly available gene expression profiles for GC patients were downloaded from the GEO database of the NCBI (<https://www.ncbi.nlm.nih.gov/geo/>). A search of the GEO database was conducted in July of 2018 using 'gastric cancer' as a key search phrase. The results for deposits made since January of 2015 were filtered using the search terms *Homo sapiens*, expression profiling by array, and expression profiling by high-throughput sequencing. Only original experimental datasets that compared the expression levels of mRNAs between GC tumors and normal tissue controls were selected. Datasets containing more than ten sets of normal and tumor samples were retained. Additionally, gene expression profiles of human gastric adenocarcinoma cell lines were downloaded from the CCLE (version 2.7, updated 2015 <https://portals.broadinstitute.org/ccle>)¹².

Gene Expression Data Preprocessing. The GEO accession number, platform, sample type, numbers of cases and controls, references and expression data were extracted from each of the identified datasets, which were then individually preprocessed using a log₂ transformation and normalization approach. If there were multiple probes for the same gene, the probe values were averaged for that gene expression level. All probe sets on different platforms were re-annotated to use the most recent NCBI Entrez Gene Identifiers (Gene IDs), and the Gene IDs were used to cross-map genes among the different platforms. Only genes present in all selected platforms were considered. To combine the results from individual studies and to obtain a list of more robust DEGs between GC and normal control tissues, guidelines outlined by Ramasamy *et al.*³⁴ for meta-analyses of gene expression microarray datasets were followed. The R packages MetaQC³⁵ was used for quality control (QC). MetaQC uses six quantitative QC parameters: (i) measures of internal QC; (ii) measures of external QC; (iii) accuracy QC of featured genes; (iv) accuracy QC of the pathway; (v) consistency of QC in the ranking of featured genes; and (vi) consistency QC in the ranking of the pathway. The mean rank of all QC measures in each dataset was also determined as a quantitative summary score by calculating the ranks of each QC measure among all included datasets.

Disease Gene Expression Signatures. MetaDE^{36–38} was used to identify DEGs in GC. A moderated *t*-statistic was used to calculate the *P* values for each dataset, and a meta-analysis was conducted with a fixed-effect model³⁹ using the MetaDE package to identify DEGs. Additionally, similarities among gene expressions profiles between tumor samples from the GEO and GC cell lines from the CCLE were assessed.

Compound Gene Expression Profiles. Level 4 gene expression profiles consisting of 978 landmark genes (L1000 genes) from LINCS as of May of 2018 were downloaded from LINCS cloud storage (<https://lincscloud.org/>), hosted by the Broad Institute⁴⁰. Cell lines described in LINCS, CCLE, and ChEMBL (version 23 1st May 2017,

<https://www.ebi.ac.uk/chembl/>)⁴¹ were mapped using GC cell line names followed by manual inspections. Meta-information for compound-induced gene expressions, including the cell line types as well as the treatment durations and drug concentrations was retrieved. Only small-molecule perturbagens having high-quality gene expression profiles (is_gold = 1, annotated in the meta-information) were used for further analysis.

Compound activity profiles. Compound response activity data, described as the half-maximal inhibitory concentrations (IC₅₀) in GC cell lines, were retrieved from ChEMBL. As the IC₅₀ values for a given compound could vary for the same cell line across different studies, the median IC₅₀ value was used. Compounds included in the ChEMBL and LINCS were mapped using The International Union of Pure and Applied Chemistry International Chemical Identifier keys. Additionally, the area under the curve (AUC) values for compound activity data in GC cell lines were retrieved from the Cancer Therapeutic Response Portal (CTRP ver 2, <https://portals.broadinstitute.org/ctrp.v2.1/>)⁴². Sensitivity levels were measured in the form of cellular ATP levels as a surrogate for cell number and growth using CellTiter-Glo assays⁴³. A compound-performance score was computed at each concentration of compound. The AUC using percent-viability scores was computed as a metric of sensitivity given that AUC reflects both relative potency and the total level of inhibition observed for a compound across CCLs. Median AUC values across multiple cell lines were used. Compounds were categorized into active (IC₅₀ < 10 μM) and inactive groups (IC₅₀ ≥ 10 μM) based on their activities in cell lines. An IC₅₀ value of 10 μM was chosen as an activity threshold because compounds with IC₅₀ ≥ 10 μM in primary screenings are often not pursued⁴⁴.

Reverse Gene Expression Score (RGES) Computation and Summarization. The method used to calculate RGES outcomes was adapted from the previously described Connectivity Map method⁴⁵. Briefly, genes were initially ranked by their expression values for each compound. An enrichment score for each set of upregulated and downregulated disease genes was computed based on the positions of the genes in the ranked list. RGES values emphasize the reversal correlation by capturing the reversal relationship between the DEGs and compound-induced changes in gene expressions. Therefore, a lower negative RGES indicates a greater likelihood of reversing changes in disease gene expressions, and *vice versa*. In addition, Spearman's correlation coefficient, Pearson correlation coefficient, and cosine similarity were computed between the DEGs and compound activities as an alternate means of computing the reversal relationship between DEGs and active compounds⁴⁶. The databases can list multiple gene expression profiles associated with one compound due to testing in various cell lines, compound treatment concentrations, and compound treatment durations, which resulted in multiple RGES outcomes for one compound that could reverse disease gene expression. Given these variations, sRGES were weighted and calculated. Results obtained for 10 μM drug concentrations and 24 h treatments were used to define the reference conditions. The analysis code and an example are provided at <https://github.com/Bin-Chen-Lab/RGES>.

Identification of Reversed Genes. In cases for which multiple compound activity IC₅₀ data were available for one compound, median IC₅₀ values were calculated. In cases for which multiple gene expression profiles yielded multiple RGES values for one compound, a median RGES value was calculated from the GC cell lines. Each gene expression profile was sorted according to its expression value. Upregulated genes were ranked high (i.e., on the top), whereas downregulated genes were ranked low (i.e. on the bottom). Among the upregulated genes, reversal genes were defined as those that were ranked lower in the inactive group (IC₅₀ < 10 μM) than in the inactive group (IC₅₀ ≥ 10 μM). In contrast, among the downregulated genes, the reversal genes were defined as those that were ranked higher in the active group than in the inactive group. A leave-one-compound-out cross-validation approach was used to find genes having reversed expressions⁴⁷. For each trial, one compound was removed and the reversed genes were then identified using the approach described above. Only those genes that were significantly reversed in all trials were retained. Genes having *P* < 0.1 in all trials were considered as reversal genes.

Statistical Analysis. The degrees of similarity in the gene expressions between tumor samples from the GEO and GC cell lines from the CCLE were assessed by Spearman's rank correlation testing, as was the similarity of RGES and IC₅₀ from ChEMBL or AUC from CCLE. A Wilcoxon signed-rank test was used to assess differences between RGES across the same and different cell lines, longer (different cell lines, <24 h) treatment durations, and higher (≥10 μM) and lower (<10 μM) drug concentrations. *P* values were adjusted with a Benjamini and Hochberg's false discovery rate method to correct for multiple testing.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Global Burden of Disease Cancer Collaboration *et al.* Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol* 3, 524–548 (2017).
2. Anderson, W. F. *et al.* Age-specific trends in incidence of noncardia gastric cancer in US adults. *JAMA* 303, 1723–1728 (2010).
3. Park, J. M. *et al.* Prognostic factors for advanced gastric cancer: stage-stratified analysis of patients who underwent curative resection. *Cancer Res. Treat.* 38, 13–18 (2006).
4. Wong, S. S. *et al.* Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing. *Nat Commun* 5, 5477 (2014).
5. Kauselmann, G., Dopazo, A. & Link, W. Identification of disease-relevant genes for molecularly-targeted drug discovery. *Curr. Cancer Drug Targets* 12, 1–13 (2012).
6. Gschwind, A., Fischer, O. M. & Ullrich, A. The discovery of receptor tyrosine kinases: targets for cancer therapy. *Nat. Rev. Cancer* 4, 361–370 (2004).
7. Ren, R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat. Rev. Cancer* 5, 172–183 (2005).

8. DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **22**, 151–185 (2003).
9. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3**, 96ra77 (2011).
10. Chong, C. R. & Sullivan, D. J. Jr. New uses for old drugs. *Nature* **448**, 645–646 (2007).
11. Barrett, T. & Edgar, R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol. Biol.* **338**, 175–190 (2006).
12. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
13. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–1097 (2014).
14. Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Syst* **6**, 13–24 (2018).
15. Koleti, A. *et al.* 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 Res.* **46**, D558–D566 (2018).
16. Xue, H., Li, J., Xie, H. & Wang, Y. Review of Drug Repositioning Approaches and Resources. *Int. J. Biol. Sci.* **14**, 1232–1244 (2018).
17. Huang, G., Li, J., Wang, P. & Li, W. A Review of Computational Drug Repositioning Approaches. *Comb. Chem. High Throughput Screen.* (2017).
18. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
19. Dovrolis, N., Kolios, G., Spyrou, G. & Maroulakou, I. Laying in silico pipelines for drug repositioning: a paradigm in ensemble analysis for neurodegenerative diseases. *Drug Discov Today* **22**, 805–813 (2017).
20. Iorio, F., Rittman, T., Ge, H., Menden, M. & Saez-Rodriguez, J. Transcriptional data: a new gateway to drug repositioning? *Drug Discov Today* **18**, 350–357 (2013).
21. Li, J. *et al.* A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **17**, 2–12 (2016).
22. Chen, M. H. *et al.* Gene expression-based chemical genomics identifies potential therapeutic drugs in hepatocellular carcinoma. *PLoS One* **6**, e27186 (2011).
23. Jahchan, N. S. *et al.* A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov.* **3**, 1364–1377 (2013).
24. Kidd, B. A. *et al.* Mapping the effects of drugs on the immune system. *Nat. Biotechnol.* **34**, 47–54 (2016).
25. van Noort, V. *et al.* Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res.* **74**, 5690–5699 (2014).
26. Cheng, L. *et al.* Identification of genes with a correlation between copy number and expression in gastric cancer. *BMC Med. Genomics* **5**, 14 (2012).
27. Huang, R., Gu, W., Sun, B. & Gao, L. Identification of COL4A1 as a potential gene conferring trastuzumab resistance in gastric cancer based on bioinformatics analysis. *Mol Med Rep* **17**, 6387–6396 (2018).
28. Zhang, H. Q. *et al.* Overexpression of UBE2C correlates with poor prognosis in gastric cancer patients. *Eur. Rev. Med. Pharmacol. Sci.* **22**, 1665–1671 (2018).
29. Zhang, J. *et al.* UBE2C Is a Potential Biomarker of Intestinal-Type Gastric Cancer With Chromosomal Instability. *Front. Pharmacol.* **9**, 847 (2018).
30. He, L. J. *et al.* Macrophage migration inhibitory factor as a potential prognostic factor in gastric cancer. *World J. Gastroenterol.* **21**, 9916–9926 (2015).
31. Gao, Q. *et al.* Evaluation of cancer dependence and druggability of PRP4 kinase using cellular, biochemical, and structural approaches. *J. Biol. Chem.* **288**, 30125–30138 (2013).
32. Janjigian, Y. Y. *et al.* Phase II Trial of Sorafenib in Patients with Chemotherapy Refractory Metastatic Esophageal and Gastroesophageal (GE) Junction Cancer. *PLoS One* **10**, e0134731 (2015).
33. Bang, Y. J. *et al.* Olaparib in combination with paclitaxel in patients with advanced gastric cancer who have progressed following first-line therapy (GOLD): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.* **18**, 1637–1651 (2017).
34. Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **5**, e184 (2008).
35. Kang, D. D., Sibille, E., Kaminski, N. & Tseng, G. C. MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.* **40**, e15 (2012).
36. Lu, S., Li, J., Song, C., Shen, K. & Tseng, G. C. Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics* **26**, 333–340 (2010).
37. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40**, 3785–3799 (2012).
38. Wang, X., Lin, Y., Song, C., Sibille, E. & Tseng, G. C. Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder. *BMC Bioinformatics* **13**, 52 (2012).
39. Choi, J. K., Yu, U., Kim, S. & Yoo, O. J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19**(Suppl 1), i84–90 (2003).
40. Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449–460 (2014).
41. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–1107 (2012).
42. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
43. Basu, A. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
44. van Westen, G. J., Gaulton, A. & Overington, J. P. Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput. Biol.* **10**, e1003559 (2014).
45. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
46. Cheng, J. *et al.* Evaluation of analytical methods for connectivity map data. *Pac. Symp. Biocomput.* 5–16 (2013).
47. Zhang, M., Chen, M. & Tong, W. Is toxicogenomics a more reliable and sensitive biomarker than conventional indicators from rats to predict drug-induced liver injury in humans? *Chem. Res. Toxicol.* **25**, 122–129 (2012).

Acknowledgements

This study was supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT) (No. NRF-2018R1A2B6001859).

Author Contributions

I.W.K. and J.M.O. created and designed the study. I.W.K., J.H.K. and H.J. contributed to data collection and analyzed the data. I.W.K. contributed to write manuscript. J.M.O. and M.G.K. reviewed the manuscript. S.K. reviewed the method and the manuscript. All authors read and approved the final manuscript. J.M.O. and I.W.K. contributed equally to this work.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-39228-9>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019