Bioinformatics, 34, 2018, i555–i564
doi: 10.1093/bioinformatics/bty271
ISMB 2018

GSEA-InContext: identifying novel and common patterns in expression experiments

Rani K. Powers[1,2], Andrew Goodspeed[2], Harrison Pielke-Lombardo[1,2], Aik-Choon Tan[3] and James C. Costello[1,2,*]

[1]Computational Bioscience Program, [2]Department of Pharmacology and [3]Department of Medical Oncology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Gene Set Enrichment Analysis (GSEA) is routinely used to analyze and interpret coordinate pathway-level changes in transcriptomics experiments. For an experiment where less than seven samples per condition are compared, GSEA employs a competitive null hypothesis to test significance. A gene set enrichment score is tested against a null distribution of enrichment scores generated from permuted gene sets, where genes are randomly selected from the input experiment. Looking across a variety of biological conditions, however, genes are not randomly distributed with many showing consistent patterns of up- or down-regulation. As a result, common patterns of positively and negatively enriched gene sets are observed across experiments. Placing a single experiment into the context of a relevant set of background experiments allows us to identify both the common and experiment-specific patterns of gene set enrichment.

Results: We compiled a compendium of 442 small molecule transcriptomic experiments and used GSEA to characterize common patterns of positively and negatively enriched gene sets. To identify experiment-specific gene set enrichment, we developed the GSEA-InContext method that accounts for gene expression patterns within a background set of experiments to identify statistically significantly enriched gene sets. We evaluated GSEA-InContext on experiments using small molecules with known targets to show that it successfully prioritizes gene sets that are specific to each experiment, thus providing valuable insights that complement standard GSEA analysis.

Availability and implementation: GSEA-InContext implemented in Python, Supplementary results and the background expression compendium are available at: https://github.com/CostelloLab/GSEA-InContext.

Contact: james.costello@ucdenver.edu

1 Introduction

Gene Set Enrichment Analysis (GSEA) (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) was developed to help with the analysis and interpretation of the long lists of genes produced from high-throughput transcriptomic experiments. By summarizing genome-wide gene expression changes into gene sets—groups of functionally related genes—a user can gain insight into how biological pathways and processes are affected under the tested experimental conditions. Since its initial application to microarray experiments, GSEA has demonstrated utility across many applications, including RNA-seq gene expression experiments, genome-wide associations studies (de Leeuw *et al.*, 2016; Zhang *et al.*, 2010), proteomics (Lavallée-Adam *et al.*, 2014) and metabolomics studies (Xia and Wishart, 2010).

The power of GSEA lies in its use of gene sets, which provide a more stable and interpretable measure of biological functions compared to individual genes that can show greater experimental and technical variation (Eklund and Szallasi, 2008). Custom gene sets can be defined, but more commonly, researchers rely on pre-compiled sets, such as the widely-used Molecular Signatures Database (MSigDB) (Subramanian *et al.*, 2005). Additional online resources have become available to provide pre-compiled gene sets specific to drug response (Yoo *et al.*, 2015), human disease and pharmacology (Araki, 2012), molecular phenotypes (Huang *et al.*, 2012) and patient prognosis (Culhane *et al.*, 2012), to name a few.

Similar to other Functional Class Scoring (FCS) methods (Khatri *et al.*, 2012), the underlying hypothesis of GSEA is that genes involved in a similar biological process or pathway (grouped into gene sets) are coordinately regulated. Thus, if an experimental perturbation activates a pathway, the genes in the associated gene set will be coordinately up-regulated and this pattern can be identified using statistical tests. The enrichment score, which reflects the degree to which genes in a gene set are over-represented at either end of a ranked gene list, is a fundamental aspect of FCS methods. Accordingly, a great deal of effort has been devoted to the development and evaluation of statistical models, from simple mean/median gene level statistics (Jiang and Gentleman, 2007) or maxmean statistics (Efron and Tibshirani, 2007) to the Wilcoxon rank sum tests (Barry *et al.*, 2005) and a modified version of the Kolmogorov-Smirnov test that is used in GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005). Finally, the significance of the enrichment score is estimated against the null hypothesis. Two categories of null hypotheses are used across FCS methods: i) self-contained or ii) competitive null hypothesis. When running GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005), these options can be found under the 'Permutation type' field with options, `phenotype` (self-contained) or `gene_set` (competitive).

The self-contained null hypothesis states that *no genes in a given gene set are differentially expressed*. To test this hypothesis for any given gene set, the phenotype labels defining the experimental condition of individual samples are permuted. This approach focuses on the genes in a given gene set and ignores genes outside the set, providing strong statistical power and rejecting more null hypotheses (Goeman and Bühlmann, 2007; Khatri *et al.*, 2012; Tian *et al.*, 2005). However, this approach has several drawbacks. For experiments with a high number of differentially expressed genes, this approach will produce many significantly enriched gene sets. Conversely, if few genes are differentially expressed, correspondingly few gene sets will be significantly enriched. Because phenotype labels are permuted under this null hypothesis, the statistical power of the test is determined by the number of samples in the experiment. As a result, the GSEA documentation recommends providing at least seven samples per phenotype label when running GSEA with the `phenotype` option selected in the 'Permutation type' field (GSEA User Guide, 2018). Experiments with fewer than three samples per phenotype cannot be run, and tens to hundreds of samples per experimental condition are needed to achieve robust statistics.

For the large number of experiments generating less than seven samples per condition, the alternative to the self-contained null hypothesis is the competitive null hypothesis. The null hypothesis for this approach states that *genes in a given gene set are at most as often differentially expressed as the genes not in the set*. To test this, random sets of genes of equal size to a given gene set are scored. Thus, this approach compares genes within a set to genes outside the set. When sample sizes are numerous and the data follow the assumptions of the underlying statistical models, then the self-contained null hypothesis is preferred as it offers greater statistical power than the competitive null hypothesis to reject the null hypothesis (Goeman and Bühlmann, 2007; Khatri *et al.*, 2012; Tian *et al.*, 2005). However, when these assumptions are not met or the focus of an analysis is on an individual sample, the competitive hypothesis is needed. When running GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005), the competitive hypothesis can be selected using the `gene_set` option under the 'Permutation type' field (GSEA User Guide, 2018). It is also the only option when running the 'GSEAPreranked' mode, where the user supplies a pre-ranked list of genes based on whatever method they choose, most often this is a list of differentially expressed genes.

There are many experiments that require the use of the competitive null hypothesis for proper comparison. Accordingly, this requirement motivated a series of methods to address the statistical challenges in single-sample analysis of ranked gene lists (Barbie *et al.*, 2009; Hänzelmann *et al.*, 2013; Lee *et al.*, 2008; Tomfohr *et al.*, 2005). By selecting random sets of genes outside the set being tested, the competitive null hypothesis approach breaks the inherent correlation structure of genes in the tested set. Methods like GSVA (Hänzelmann *et al.*, 2013) nicely address this challenge by incorporating gene-specific variation directly in the calculation of a sample-wise gene set enrichment score within a given input dataset.

Here we take a different approach to analyze and adjust for patterns in differentially enriched gene sets produced using GSEA with the competitive null hypothesis. Specifically, we account for gene-specific variation estimated from a set of background experiments. Our approach is motivated by the fact that there are no methods available for a user to easily compare their GSEA results to GSEA results obtained from other experiments to discern similar and/or distinct patterns affected across experiments. Overall, the goal of this research is to address two questions: i) which gene sets are commonly enriched across a compendium of experiments, and ii) which gene sets are uniquely enriched in a single experiment compared to many other, independent experiments? By addressing these questions, we intend to complement and enrich the results generated by GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005).

To accomplish these goals, we first curated a compendium of gene expression experiments encompassing a variety of experimental conditions and identified patterns of positive and negative enrichment by applying GSEA. We then leverage these patterns to help place single experiments into context. Accordingly, we developed an extension for GSEA that uses these context-specific patterns to inform the statistical testing procedure. Specifically, while GSEA tests for the significance of an enrichment score against a null distribution of enrichment scores calculated for *random* permuted gene sets, our algorithm generates permuted gene lists based on a set of background experiments. Because we allow the user to define the context of the background set of experiments, we have termed our method, GSEA-InContext, which stands for GSEA—Identifying novel and Common patterns in expression experiments.

We applied GSEA-InContext to a compendium of gene expression experiments testing small molecule treatments in human cell lines. Small molecules remain the gold standard of treatment for numerous diseases, and in the context of cancer, human cell lines have been widely used to study mechanisms of drug action and present a robust pharmacogenomic platform (Barretina *et al.*, 2012; Garnett *et al.*, 2012; Goodspeed *et al.*, 2016; Shoemaker, 2006). Gene expression experiments are regularly performed to study the direct effect of a small molecule, but expression profiles will capture both on- and off-target effects of the small molecule and disentangling these effects remains a challenge. At the same time, patterns of positive or negative enrichment can provide insights into common (i.e. not tissue- or drug-related) responses to small molecule treatment. In this article, we demonstrate how GSEA-InContext can be used to gain insights into both aspects of small molecule treatment. We proceed by first describing our curated background compendium of small molecule gene expression experiments. We present an analysis of this compendium and identify commonalities in differentially expressed genes and significantly enriched gene sets, motivating the development of the GSEA-InContext method. Finally, we demonstrate GSEA-InContext on two example applications: Notch inhibition in T-cell acute lymphoblastic leukemia and investigating gene

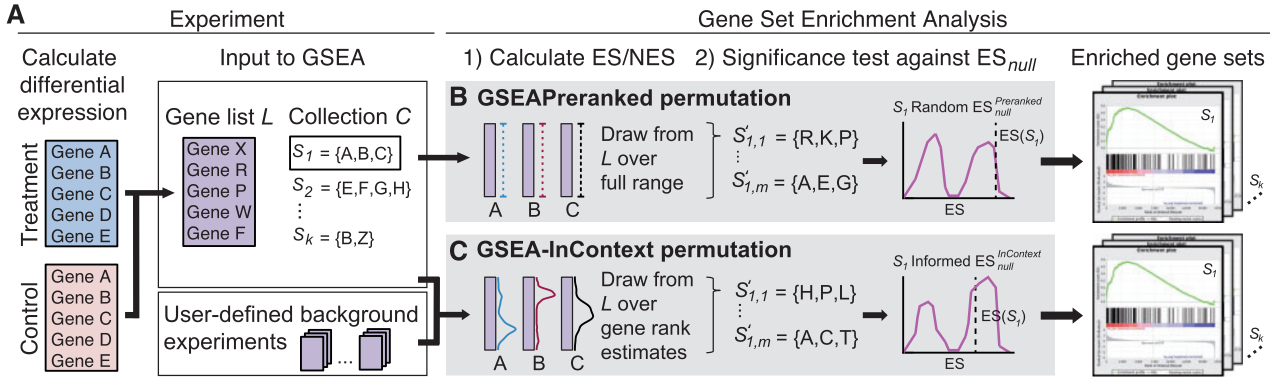

Fig. 1. Overview of the statistical test for GSEAPreranked and GSEA-InContext. (**A**) A workflow for using GSEA to identify significantly enriched gene sets in a vehicle control vs drug-treated experiment. Calculating the expression fold change between the two conditions produces a ranked gene list *L*. This list is input into GSEA along with a collection of gene sets *C*. (**B**) To test whether a gene set S_1 is significantly enriched in *L*, the enrichment score, $ES(S_1)$, is tested against a null distribution ES_{null}. GSEAPreranked creates $ES^{Preranked}_{null}$ by calculating the *ES* for *m* gene sets the same size as S_1, which are created by randomly selecting from teh full range of *L*. (**C**) The GSEA-InContext approach takes as input *L*, *C*, and a user-defined set of background experiments. Instead of randomly generating gene sets, $ES^{InContext}_{null}$ is created by selecting genes based on how they are distributed in the background set of experiments. The *ES* for each of *m* informed gene sets is calculated and used to evaluate the significance of $ES(S_1)$

expression changes in response to dexamethasone and estradiol treatment in breast cancer cell lines.

2 Materials and methods

2.1 Data collection and normalization

We queried the Gene Expression Omnibus (GEO) database (Edgar *et al.*, 2002) for human gene expression studies performed on the Affymetrix Human Genome U133 Plus 2.0 Array that tested small molecules. We excluded studies that had fewer than two replicates per condition, or that did not have an appropriate vehicle control condition, which was needed to calculate differentially expressed gene lists across all experiments. We proceeded with a total of 128 studies comprised of 2812 individual microarrays that met the search criteria. Meta-data for each study was parsed from GEO in order to annotate tissue type, cell line and small molecule. The CEL files for each study were downloaded with the *GEOQuery* R package (Davis and Meltzer, 2007). Within each study, the expression data was background corrected, quantile normalized and probe sets were summarized using RMA (Bolstad *et al.*, 2003) with the *affy* R package (Gautier *et al.*, 2004). For each study, control and treatment conditions were identified and differential expression between all control/treatment pairs was calculated with the *limma* package (Ritchie *et al.*, 2015). Probe sets were annotated to genes using the *hgu133plus2.db* R package (Carlson, 2016), keeping one probe set per gene with the highest average expression across all samples. For each experimental comparison, genes were ranked according to their $\log_2$ fold change and saved as a ranked list *L* (Fig. 1A) for input into GSEAPreranked and GSEA-InContext. In total, we generated a compendium of 442 ranked lists.

All gene set collections were downloaded from MSigDB, v6.1 (Liberzon *et al.*, 2015; Mootha *et al.*, 2003; Subramanian *et al.*, 2005). The Hallmarks collection (Liberzon *et al.*, 2015) was selected to be used for all analyses because it is comprised of 50 gene sets, thus full results can be reported and displayed through this manuscript. Analyses performed with additional gene sets are supplied as described in Section 2.4.

To annotate mechanisms of action for the small molecules, we grouped them based on their targets using the Drug Repurposing Hub (Corsello *et al.*, 2017) and DSigDB (Yoo *et al.*, 2015).

2.2 Application of GSEAPreranked

To ensure consistency between implementations of GSEA, we ran each of the 442 ranked lists through the GSEAPreranked algorithm using both the javaGSEA Desktop program (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) and the GSEApy Python package (https://github.com/BioNinja/gseapy); both implementations produced equivalent results. For all analyses shown here, we applied GSEApy (pypi package version 0.9.3, Python3.6) using a `weighted` enrichment scoring statistic and 100 permutations. GSEAPreranked requires the use of the competitive null hypothesis, the `gene_set` permutation type. Default settings were used for all other parameters.

2.3 Implementation of GSEA-InContext

According to the GSEA documentation (GSEA User Guide, 2018), the GSEAPreranked algorithm takes as input a user-supplied ranked gene list L and a collection of gene sets $C = \{S_1 \ldots S_k\}$, where S_k is an *a priori* defined gene set (Fig. 1A). An enrichment score (*ES*) is calculated for each gene set $ES(S_k)$ using a weighted Kolmogorov-Smirnov-like statistic (Mootha *et al.*, 2003; Subramanian *et al.*, 2005). The *ES* reflects the degree to which genes in S_k are positively or negatively enriched at either end of the ranked gene list L.

To estimate the significance level of $ES(S_k)$, GSEAPreranked tests $ES(S_k)$ against an empirically defined null distribution, $ES^{Preranked}_{null}$. To illustrate how this distribution is created, we use the example of S_1 in Figure 1. GSEAPreranked generates m permuted gene sets of the same length as S_1 by randomly selecting genes from L (Fig. 1B). We use the notation $S'_{1,j}$ to represent the jth permutation of the randomized gene set S'_1. The nominal P-value for S_1 is calculated by comparing $ES(S_1)$ to the $ES^{Preranked}_{null}$ distribution. Note that the modified Kolmogorov-Smirnov test applied by GSEA creates a bimodal $ES^{Preranked}_{null}$.

Our method applies the same approach as GSEA to calculate the nominal P-value (GSEA User Guide, 2018; Mootha *et al.*, 2003; Subramanian *et al.*, 2005). However, in contrast to GSEAPreranked, GSEA-InContext employs an alternative procedure to generate the null distribution, in which the m permuted gene sets are generated using the gene ranks estimated from a set of pre-compiled background experiments (Fig. 1C). The background experiments are defined by the user, either by compiling their own experiments or leveraging a subset or the full set of the 442 precompiled ranked list supplied here. For a gene present in gene list L, let random discrete variable $X = \{x_1 \ldots x_n\}$ represent the set of gene ranks across all n background experiments where x_r is the gene's rank in the rth background experiment. Using the background experiments, we calculate the mean rank, μ_i, and variance, σ_i^2, where we set the default σ_i^2 to be the median of the distances between all pairwise ranks of gene i over the n background experiments. The distance is simply the difference in ranks between x_r and $x_r + 1$. Using this information we estimate $P(X = r)$, or the probability of any gene having a given rank using the beta-binomial distribution as follows:

$$P(X = r|\alpha, \beta) = \binom{n}{r} \frac{B(r + \alpha, n - r + \beta)}{B(\alpha, \beta)}$$

$B(\alpha, \beta)$ is the beta distribution with shape parameters α and β, which we calculate as follows:

$$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right)\mu^2, \quad \beta = \alpha\left(\frac{1}{\mu} - 1\right)$$

To calculate α and β, μ and σ^2 are bounded, $\mu \in (0, 1)$ and $\sigma^2 \in (0, 0.5^2)$. Thus, we scale the gene ranks, and subsequently μ and σ^2, to be in the range $(0, 1)$. After $P(X = r)$ is estimated over all the ranks for a given gene i, the values are scaled back to the original rankings. As shown in Figure 1C, this procedure is applied independently for all genes in a given gene set, such as S_1.

To create the permuted gene set $S'_{1,1}$, GSEA-InContext draws rank indices according to $P(X = r|\alpha, \beta)$ for each gene in S_1. This index is then used to select the gene at that rank in L. This procedure is repeated to create m permutations of S_1 to generate the informed $ES_{null}^{InContext}$. Should the same index be drawn randomly, we resample according to the same procedure to create a non-overlapping $ES_{null}^{InContext}$. As in GSEAPreranked, the nominal P-value for S_1 is calculated by comparing $ES(S_1)$ to the $ES_{null}^{InContext}$ distribution (Fig. 1C). Also following GSEAPreranked, the false discovery rate (FDR) is calculated as the ratio of the actual ES compared to the ESs for C against all permutations over the distribution of the actual ES compared to the ESs for C in the dataset being tested (GSEA User Guide, 2018).

Outside of the changes to the way GSEA-InContext generates the null distribution of enrichment scores, all other components of GSEA are the same for GSEAPreranked and GSEA-InContext.

When run using our Python implementation, GSEA-InContext returns a table of nominal P-values, FDR-adjusted P-values, enrichment scores and normalized enrichment scores for every gene set tested. Enrichment plots can also be generated. We also return the table of GSEAPreranked results for all of the experiments used in the background set, allowing researchers to explore common patterns in the background experiments. Finally, GSEAPreranked results are output for comparison to the GSEA-InContext results.

2.4 Code availability

To leverage the multi-threading capabilities of GSEApy, we implemented GSEA-InContext as a new class within the existing Python package. The code, documentation and Supplementary results for all gene set collections are available at: https://github.com/CostelloLab/GSEA-InContext.

The background gene expression compendium of 442 ranked lists is available at: https://www.synapse.org/GSEA_InContext.

3 Results

3.1 Overview of gene expression datasets

We curated a gene expression compendium of 442 gene lists ranked by $\log_2$ fold change between treatment versus control conditions. We required that all comparisons have at least three replicates per condition, where the conditions were either small molecule treatments or the appropriate vehicle control treatment. Raw data were processed according to the procedures outlined in Section 2.1. The tissues and small molecules included in the compendium are summarized in Figure 2. A total of 21 tissues are represented in the dataset (202 unique immortalized or primary cell lines), with the most common tissue type being breast. We captured a range of 129 small molecules that we grouped into 69 drug classes based on mechanism of action. The most commonly used small molecules were eribulin and paclitaxel, which inhibit microtubule dynamics.

3.2 Common patterns of genes and pathways across small molecule treatments

To evaluate general gene- and pathway-level patterns, we first created a distribution of the mean ranks for each gene across the 442 experiments. We compared these results to a null distribution generated by randomizing the genes in each of the 442 experiments. We found that roughly 25% of genes fell at least 3 standard deviations outside the mean rank of the null distribution, compared to the expected frequency of 0.3% (Fig. 3A). Of the 25%, 12.6% of genes ranked higher and 13.9% ranked lower than the mean rank. These results demonstrate that roughly a quarter of the genes being studied across 442 experiments are more consistently differentially regulated than expected at random.

To illustrate this effect on a per gene basis, Figure 3B displays the two genes with the highest and lowest mean rank across all 442 experiments. The gene with the highest mean rank was MAF bZIP transcription factor F (MAFF), which encodes a transcription factor of the MAF family and has been shown to be essential for activation of genes involved in detoxification and the response to oxidative stress (Katsuoka *et al.*, 2005). This gene is also up-regulated in response to hypoxia (Chen *et al.*, 2006). The most lowly ranked gene was cyclin E2 (CCNE2), an activating regulatory subunit of CDK2, most highly expressed during the G1/S cell cycle transition (Gudas *et al.*, 1999). Intuitively, the rankings of these genes are consistent with small molecule treatment, given that CCNE2 is frequently down-regulated in response to drugs that arrest cell growth, and MAFF is up-regulated in response to cellular stress. However, the non-random ranking of genes does suggest there would be commonalities across enriched gene sets identified by GSEA. To investigate this, we ran GSEAPreranked on each of the 442 experiments and evaluated global gene set patterns using the Hallmarks collection (Liberzon *et al.*, 2015). We performed our analyses using all gene sets available in MSigDB (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) and found similar patterns as those reported for the Hallmarks collection; these results are available as described in Section 2.4.

In Figure 4A, we report the fraction of experiments that showed an FDR < 0.05 for each of the gene sets in the Hallmarks collection, where we found clear patterns of positive and negative enrichment. For example, proliferation and cell cycle related processes were consistently down-regulated, including `E2_TARGETS`, which was significantly down-regulated in over 45% of the experiments. Other gene sets were consistently up-regulated, such as `TNFA_SIGNALING_VIA_NFKB`, which was significantly positively enriched in approximately 53% of the experiments. These results

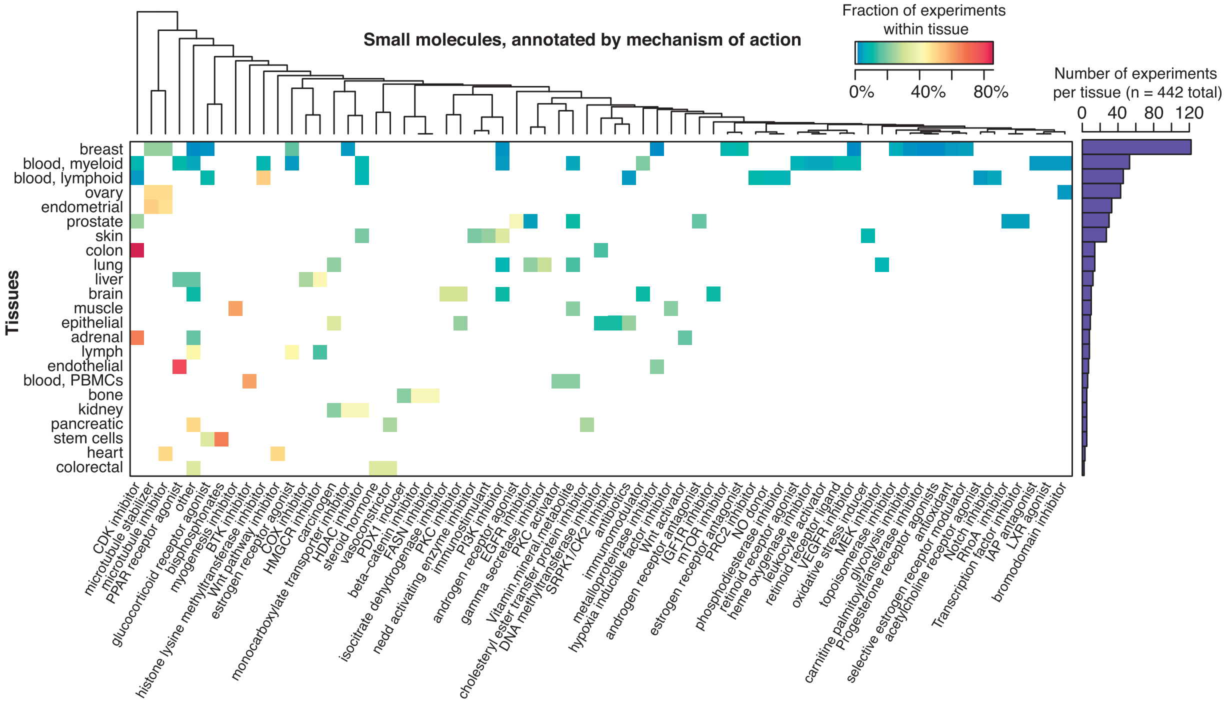

Fig. 2. Overview of gene expression datasets by tissues and small molecules. Heatmap shows the fraction of small molecules used across 442 experiments (treatment vs. control comparisons). All experiments were performed in human immortalized or primary cell lines

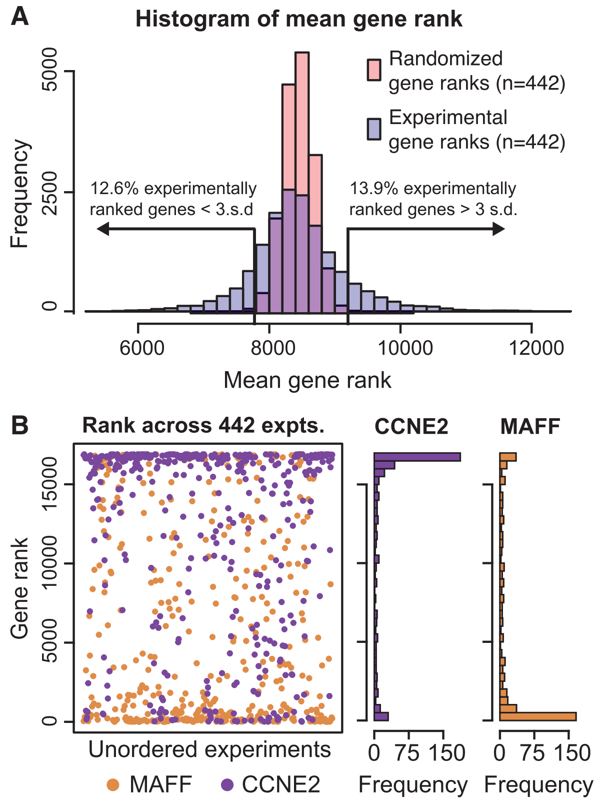

Fig. 3. Ranking of genes across 442 small molecule gene expression profiles. (**A**) Distribution of the mean rank for all genes measured across 442 small molecule experiments (blue) compared to the mean rank of genes from 442 randomized gene lists (pink). Roughly 25% of genes in the experiments fall outside 3 standard deviations from the randomly ranked genes. (**B**) The ranks of MAFF and CCNE2 across all 442 experiments. These two genes are the highest and lowest ranked genes in (A) by mean rank across all 442 experiments

are consistent with the trends we identified for CCNE2 and MAFF (Fig. 3B). CCNE2 is a member of many of the down-regulated cell cycle-related gene sets and MAFF is a member of the most up-regulated gene set, `TNFA_SIGNALING_VIA_NFKB`. In comparison, analyzing 442 randomly permuted gene lists with GSEAPreranked produced significant results in few experiments, less than 3% for any gene set in the Hallmarks collection.

To investigate the potential effects of the large number of experiments coming from a limited number of studies in our dataset that use breast cells or treat with tubulin polymerization inhibitors, we repeated our GSEAPreranked analysis including and excluding these experiments (Fig. 4B). Comparing the GSEA results for 126 experiments using tubulin polymerization inhibitors to the remaining 317 experiments, we observed instances where certain gene sets increased in frequency of significance in experiments with the inhibitors and other gene sets increased under all other drugs. However, many of the general patterns shown in Figure 4A remain, demonstrating that the over-representation of tubulin polymerization inhibitors is not soley responsible for the results in Figure 4A. Similarly, we compared 107 experiments using breast cell lines to 336 experiments using cells from other tissues and, again, found gene sets such as `TNFA_SIGNALING_VIA_NFKB` were commonly significantly enriched regardless of experimental tissue type (Fig. 4C).

3.3 Global adjustment of common patterns of gene set enrichment

Our meta-analysis of GSEA results across a compendium of 442 small molecule gene expression experiments highlighted common patterns of gene set enrichment. To complement this analysis, we next asked, which gene sets are uniquely enriched in a given experiment? We addressed this question by assuming the competitive null hypothesis as in GSEAPreranked, but adjusting the empirical null distribution used in the statistical test (Fig. 1). The method we

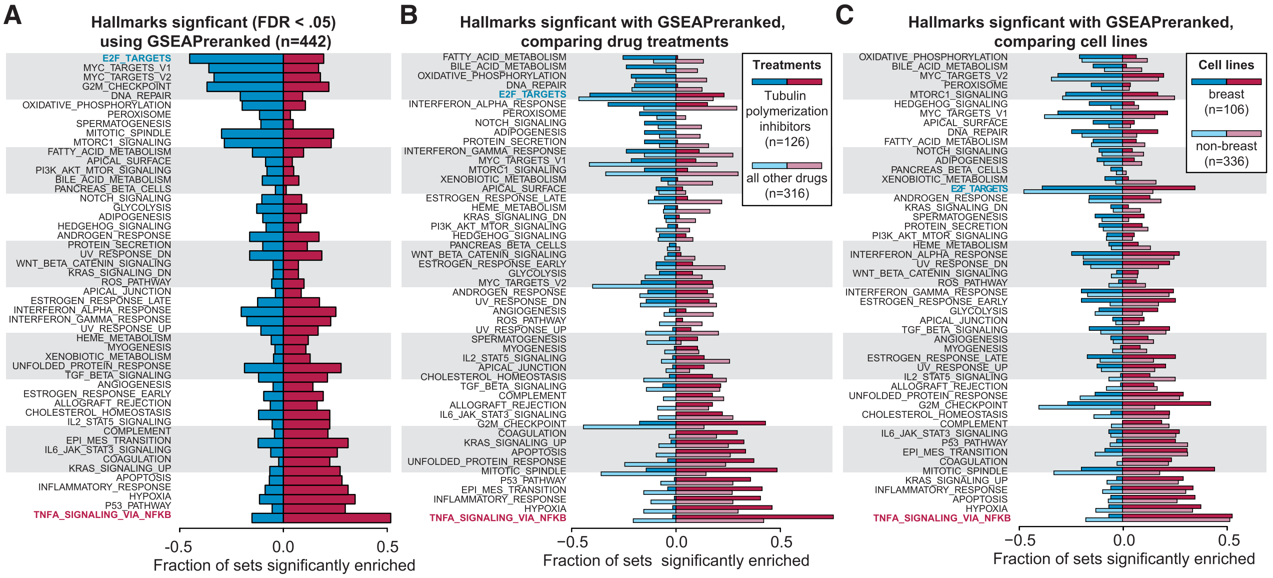

Fig. 4. Commonly enriched gene sets across 442 small molecule gene expression experiments. (**A**) The gene sets in the Hallmarks collection (Liberzon *et al.*, 2015) were tested against all 442 experiments using GSEAPreranked (competitive null hypothesis). Significant gene sets are defined as an FDR < 0.05. Gene sets are ranked by the difference in the fraction of experiments with significant positive and negative enrichment. The most frequently down-regulated pathway is E2F_TARGETS (blue text) and most commonly up-regulated pathway is TNFA_SIGNALING_VIA_NFKB (red text). (**B**) The fraction of positively and negatively enriched gene sets are shown for 126 experiments that tested response to eribulin or paclitaxel (dark bars), compared to 317 experiments that tested another compound (light bars). (**C**) The fraction of positively and negatively enriched gene sets within 107 experiments using breast cancer cell lines (dark bars), compared to 336 experiments that used non-breast cells (light bars)

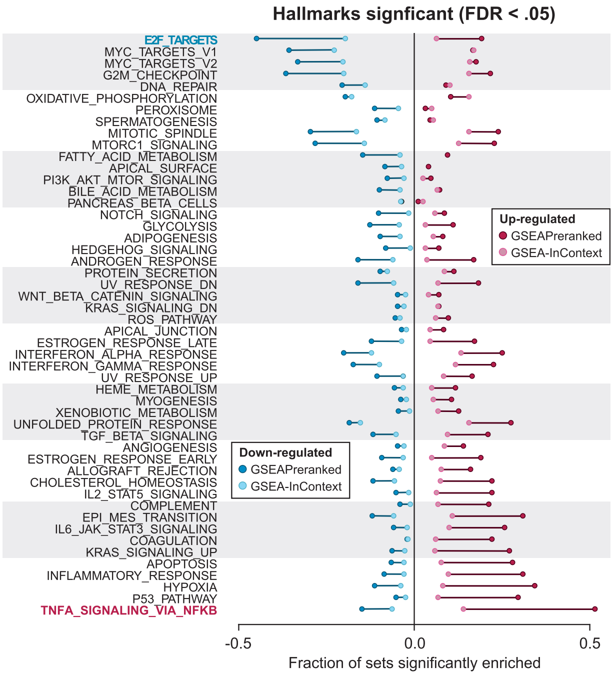

Fig. 5. Adjusting for positively and negatively enriched pathways. The points represent the fraction of gene sets that are significantly up- or down-regulated (FDR < 0.05) across all 442 experiments in GSEAPreranked (dark red, dark blue) or GSEA-InContext (light red, light blue). The bars show the difference between the fraction of significantly enriched gene sets between the analyses

propose leverages a background set of experiments to define an informed null distribution, rather than creating one with completely random permutations. As the goal of this approach is to place a single experiment in the context of a background set of user-defined experiments, we call the method GSEA-InContext. Full details of the method are described in Section 2.3.

First, we compared the results produced by GSEAPreranked on the 442 experiments to the corresponding results from GSEA-InContext. We ran GSEA-InContext on each individual experiment using the background set of the 441 other experiments and the Hallmarks collection (Liberzon *et al.*, 2015) (Fig. 5). As expected, GSEA-InContext broadly reduced the number of significantly reported gene sets per experiment. More specifically, the commonly enriched pathway `TNFA_SIGNALING_VIA_NFKB` was reduced from 53% up-regulated in GSEAPreranked to 14% in GSEA-InContext. Similarly, the most down-regulated gene set `E2F_TARGETS` was enriched in only 19% of experiments using GSEA-InContext compared to 42% in GSEAPreranked. Two gene sets, `OXIDATIVE_PHOSPHORYLATION` and `PEROXISOME`, that are uncommon in the 442 experiments become enriched at a slightly higher frequency in GSEA-InContext compared to GSEAPreranked.

To confirm that the GSEA-InContext method did not introduce any systematic biases, we ran GSEA-InContext on randomized rank lists for all of the 442 experiments and found that gene sets were significantly positively or negatively enriched in a small fraction ($<4\%$) of the random experiments, which was equivalent to results produced by running the same randomized experiment with GSEAPreranked.

Finally, we performed experiments to analyze the impact of the size of the background set. We randomly sampled 100 background ranked lists of size 300, 200 and 100 from our 442 ranked lists. We compared the result using 10 randomly selected experiments from the 442 and found that even at a sample size of 300, significance tended to be less stringent, resulting in greater number of lower FDRs being reported, though these findings varied across gene sets. This trend continued with background sets of 200 and 100, though the results were similar to that of the background set of 300. We report the full Supplementary results as described in Section 2.4.

3.4 Applications of GSEA-InContext

We demonstrate the application of GSEA-InContext using two biologically relevant examples. The first example illustrates that GSEA-InContext successfully removed non-specific gene set enrichment patterns in order to identify the on-target effects of a small molecule compound. The second example demonstrates how GSEA-InContext can be used to disentangle the effects of a single drug in cells treated with a drug combination.

3.4.1 Re-scoring Notch pathway inhibition in a T-ALL cell line to down-weight non-specific gene sets

Any small molecule drug will have direct (e.g. signaling) and indirect (e.g. stress) effects, whether it is due to drug promiscuity or the inherent interconnectedness of biological systems. Thus, a perennial challenge in pharmacology is to functionally characterize the on- and off-target effects of a drug treatment. Accordingly, we demonstrate how GSEA-InContext can be used to identify gene sets that are specific to a small molecule treatment by selecting an appropriate background set of experiments. One well-represented tissue type in our compendium of 442 experiments is blood, in particular leukemia cell lines, which we stratified into the lymphoblastoid ($n = 44$) and myeloid ($n = 48$) lineages. We selected a single experiment in which HBP-ALL cells treated with SAHM1, a Notch signaling inhibitor, were compared to cells treated with a vehicle control (GSE18198; Moellering *et al.*, 2009). Activation of Notch signaling has been associated with the development of T-cell acute lymphoblastic leukemia (T-ALL), with direct inhibition of Notch pathway members in tissue culture and mouse models decreasing proliferation of T-ALL cells. We found 17 gene sets significantly enriched at an FDR < 0.05 (Fig. 6A) using GSEAPreranked with the Hallmarks collection. Interestingly, while NOTCH_SIGNALING was down-regulated, it remained above the significance threshold (FDR $= 0.097$).

We next ran the same experiment through GSEA-InContext, using a set of 44 lymphoblastoid experiments as the background set. Using the Hallmarks collection, GSEA-InContext identified a total of 10 significantly enriched gene sets (FDR < 0.05). Notably, GSEA-InContext reported NOTCH_SIGNALING to be significantly down-regulated (FDR $= 0.037$) (Fig. 6A), supporting the direct inhibition of the Notch signaling pathway by SAHM1. We confirmed that the direction of enrichment (positive/negative) for all gene sets was the same in both analyses. We additionally performed a down-sampling experiment to evaluate the impact of this targeted background set and found that NOTCH_SIGNALING remained significant even when the background set was reduced to 10 ranked lists, though, as we saw in Section 3.3 the decreased background size correlated with overall lower FDRs across all tested gene sets (Supplementary analysis as reported in Section 2.4).

We compared the results of GSEAPreranked to GSEA-InContext and used these patterns to help interpret the results. A gene set that was significant in GSEAPreranked but was raised above an FDR of 0.05 in GSEA-InContext was likely commonly enriched across the background experiments. Conversely, a gene set being significant in both GSEAPreranked and GSEA-InContext suggests that the set is uniquely enriched in the experiment being tested compared to the background experiments. We found gene sets that meet both criteria. Cell cycle related gene sets (G2M_CHECKPOINT and E2F_TARGETS) were significant in GSEAPreranked, but not in GSEA-InContext (Fig. 6A), supporting the finding of Moellering, et al. that the SAHM1 inhibits cell proliferation (Moellering *et al.*, 2009). Although down-regulation of cell cycle processes is a biologically relevant result that supports the authors experimental results, GSEA-InContext indicates that this result is a common response in lymphoblastoid cells treated with an array of drugs. This is supported by the fact that approximately 70% of the 44 background experiments showed enrichment of cell cycle related processes.

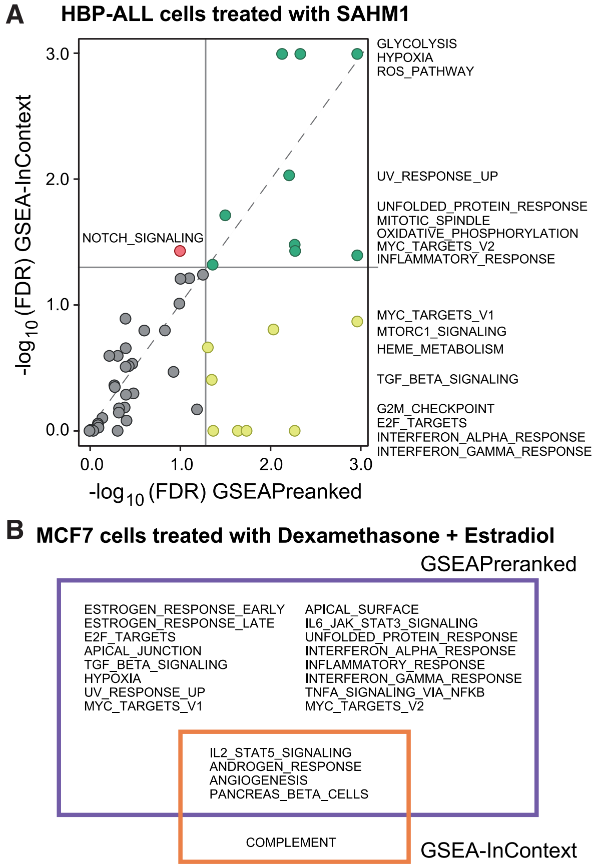

Fig. 6. Two illustrative examples using GSEA-InContext. (**A**) GSEAPreranked was run on a ranked list of differentially expressed genes from T-cell acute lymphoblastic leukaemia cells (HBP-ALL) treated with SAHM1, a Notch pathway inhibitor. GSEA-InContext was run on the same experiment using a background set of 44 lymphoblastoid cell line experiments. The plot shows the -$\log_{10}$ FDR of each analysis, with the grey lines signifying FDR < 0.05. Yellow points represent gene sets significant only in GSEAPreranked; green points were significant in both analyses; the red point (NOTCH_SIGNALING) is only significant in GSEA-InContext; grey points fell below significance in both analyses. Gene set names are listed to the right of the plot. (**B**) GSEAPreranked significant results (purple) on an MCF7 breast cancer cell line treated with a combination of dexamethasone and estradiol. GSEA-InContext results (orange) on the same experiment as in (A) using a background set of 22 experiments in which MCF7 breast cancer cell lines were treated with estradiol only. The intersection and set differences between the two analyses are shown. All analyses used the Hallmarks gene set collection (Liberzon *et al.*, 2015)

The most significantly down-regulated genes sets in GSEA-InContext are REACTIVE_OXYGEN_SPECIES_PATHWAY, GLYCOLYSIS and HYPOXIA. All three gene sets are also highly significant in GSEAPreranked, suggesting that these processes are uniquely significant when HBP-ALL cells are treated with SAHM1. The link between hypoxia and Notch signaling has been shown to play key roles in cell differentiation (Gustafsson *et al.*, 2005) and key cancer related processes of migration and invasion (Sahlgren *et al.*, 2008). Hypoxia has long been know to play a key role in controlling glycolytic metabolism, particularly in cancer cells (Lu *et al.*, 2002), and hypoxic conditions stimulate the production of reactive oxygen species (Chandel *et al.*, 2000). The tight link between these process and their regulatory link with Notch suggests that Notch inhibition could be directly down-regulating key cancer progression processes, another potential positive effect of SAHM1 treatment.

Taken together, the GSEAPreranked and GSEA-InContext results provide a more complete picture of the processes that are differentially regulated in HBP-ALL cell treated with SAHM1. By placing enriched gene sets in context of a lymphoblastoid experimental background, we identified both common and experiment-specific gene sets. Notably, GSEA-InContext identified NOTCH_SIGNALING as being significantly down-regulated, whereas GSEAPreranked did not (Fig. 6A).

3.4.2 Disentangling the effects of dexamethasone from estradiol response in breast cancer cell lines

As a second example, we sought to demonstrate how GSEA-InContext can be used to prioritize gene sets that are specific to a small molecule treatment by down-weighting gene sets that are enriched in the background set of experiments. In this case, we performed GSEAPreranked on an experiment in which MCF7 breast cancer cells were treated with estradiol, an estrogen receptor agonist and dexamethasone, a corticosteroid (GSE79761) (West *et al.*, 2016). We then applied GSEA-InContext to this same experiment using a background set of 22 estradiol-only treated MCF7 experiments. By defining the background this way, we aimed to down-weight gene sets related to breast cancer cells or estradiol treatment while identifying gene sets that are more specifically related to dexamethasone treatment.

We compared the results for the Hallmarks collection (Liberzon *et al.*, 2015) between each enrichment method (Fig. 6B). Gene sets shown in the purple box in Figure 6B were significantly enriched using GSEAPreranked. The gene sets in the purple box only represent pathways and processes that were commonly altered across the background experiments. In this group, we found gene sets that were expected to be enriched in MCF7 cells treated with estradiol, such as ESTROGEN_RESPONSE_EARLY and LATE. Several gene sets that we previously identified as being significantly enriched across a wide variety of cell lines and drug treatments in our compendium (Fig. 4), such as E2F_TARGETS and TNFA_SIGNALING_VIA_NFKB, were also identified as significant by GSEAPreranked. In contrast, these sets were not significantly enriched in GSEA-InContext (orange box), demonstrating that these sets were down-weighted to prioritize gene sets related to dexamethasone treatment while adjusting for the effects of estradiol.

The gene sets in the overlapping section between the purple and orange boxes were identified as significantly enriched in both GSEAPreranked and GSEA-InContext. We confirmed that the direction of enrichment (positive/negative) for these gene sets was the same in both methods. The four gene sets identified in both analyses were ANGIOGENESIS, IL2_STAT5_SIGNALING, ANDROGEN_RESPONSE and PANCREAS_BETA_CELLS. Because these gene sets are also significant in the GSEA-InContext analysis, we expect the enrichment of these gene sets to be the result of the added dexamethasone treatment in these cells.

The link between dexamethasone and the androgen signaling pathway has been investigated in several studies. Dexamethasone is a glucocorticoid receptor (GR) agonist and GR shares several transcriptional targets with the androgen receptor (AR), including SGK1, MKP1 and DUSP1 (Arora *et al.*, 2013). Indeed, SGK1 is in the ANDROGEN_RESPONSE gene set. Dexamethasone has also been linked to IL2 signaling, which we see in the IL2_STAT5_SIGNALING gene set. The ANGIOGENESIS gene set is also negatively enriched in this experiment, supporting previous results showing that dexamethasone inhibits angiogenesis (Yano *et al.*, 2006). Finally, we note that COMPLEMENT is uniquely enriched in GSEA-InContext. Interestingly, dexamethosone has been shown to be a transcriptional regulator of components in the complement pathway (Lappin and Whaley, 1991). While those results are in immune cells, this presents the potential research topic of dexamethosone regulation of complement in breast cells stimulated by estradiol.

Once again, we demonstrated that the GSEAPreranked and GSEA-InContext results taken together provide complementary perspectives into altered pathways and processes in this experiment to identify both common and experiment-specific gene sets.

4 Discussion

Extracting biological insights from the long lists of genes produced by differential expression experiments still remains a challenge. FCS methods, such as GSEA, are designed to aide in the interpretation of gene lists by identifying differentially up- and down-regulated pathways and processes. Although GSEA succeeds at summarizing the original list of genes into gene sets and identifying enrichment, the results are provided only in the context of the tested experiment. This is by design, but placing a single experiment in the context of a biologically relevant background can provide insight into the common and experiment-specific gene set patterns. In fact, common patterns of positively and negatively enriched gene sets can be observed across a variety of experimental conditions. We applied GSEAPreranked to 442 different experiments in which human cells were treated with small molecules and we identified gene sets that were commonly up- and down-regulated across a number of contexts (e.g. drugs and tissues).

The majority of drugs that we evaluated were inhibitors (most being cancer drugs). These small molecules are designed to inhibit the growth of cells. Consistent with what we expected, the gene sets representing cell cycle processes were the most down-regulated pathways, while gene sets associated with cellular damage and stress were commonly up-regulated. Interestingly, TNFA_SIGNALING_VIA_NFKB was significantly up-regulated in over 50% of the 442 experiments and NF-κB signaling downstream of TNFα has been shown to be pro-survival (Rath and Aggarwal, 1999). This suggests that inhibiting NF-κB signaling with the other small molecule could potentially be an effective drug combination treatment representing a common mechanism of drug synergy. This is one example of a testable hypothesis that can be generated from exploring commonly enriched gene sets.

Conversely, these common patterns motivate a new type of analysis: specifically, that researchers can place their own experimental results into a relevant context in order to identify uniquely enriched gene sets for their experiment compared to others. Accordingly, we introduced GSEA-InContext to perform such an analysis. By running GSEA-InContext on our compiled set of 442 expression experiments, we showed that the algorithm successfully down-weighted the gene sets such as TNFA_SIGNALING_VIA_NFKB that are commonly enriched in many experiments. Additionally, we applied GSEA-InContext to two example experiments, showing that in each case our method highlighted biological pathways relevant to the small molecule compound in each experiment.

GSEA-InContext uses the competitive null hypothesis for statistically evaluating gene set enrichment. While the self-contained null hypothesis is preferred because it offers greater statistical power than the competitive null hypothesis (Goeman and Bühlmann, 2007; Khatri *et al.*, 2012; Tian *et al.*, 2005), there are many instances when the self-contained null hypothesis cannot be used, particularly when the number of samples per condition are low. The majority of experiments that aim to test two conditions generate far less than seven samples per condition, which requires the competitive null hypothesis to be used. Thus, while GSEA-InContext is not applicable using the self-contained hypothesis, it is is readily usable for the majority of gene expression experiments that require the use of the competitive null hypothesis.

For the purposes of this analysis, we focused our efforts on small molecule treatments of human cell lines. With over a million expression datasets currently in the GEO database (Baker, 2012), compiling a properly defined background set can be a daunting task, as each dataset requires manual curation of the control and treatment groups. However, the 442 treatment-control comparisons that we compiled and made available present a robust set of data to begin exploring common and experiment-specific gene set patterns. The results from GSEA-InContext are fully dependent on the compendium of background experiments, and as such, the approach will become more robust as the background compendium is expanded to include other drugs and cell line experiments, such as ref. (Subramanian *et al.*, 2017). Future work will also include compiling background sets to study other biological contexts and other organisms. Leveraging efforts such as CREEDS (CRowd Extracted Expression of Differential Signatures) will also rapidly expand the potential user-defined background sets (Wang *et al.*, 2016). Additionally, we reported that the background set influences statistical tests, in particular the FDR estimates; thus, future work is needed to fully characterize the effect of the background set on the results. Finally, comparing results across platforms (e.g. microarray, RNAseq) will help identify which commonly enriched gene sets can be attributed to technical differences between platforms and which patterns are robust across platforms reflecting true biological results.

We would like to close by stating that the goal of GSEA-InContext is not to replace GSEA, but to complement the original implementation of GSEA. Comparing the results obtained from GSEA and the contextualized results from GSEA-InContext, we were able to gain insights into not only the pathway-level changes in an experiment, but also the common and experiment-specific patterns.

Acknowledgements

We would like to thank Nicolle Witte for early contributions to the project.

Funding

This work was supported by the Boettcher Foundation (J.C.C.), the Front Range Cancer Challenge (A.G.) and NIH grants T32GM007635 (A.G.), T15LM009451 (H.P-L.).

Conflict of Interest: none declared.

References

Araki,H. (2012) GeneSetDB: a comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, **2**, 76–82.

Arora,V.K. *et al.* (2013) Glucocorticoid receptor confers resistance to antiandrogens by bypassing androgen receptor blockade. *Cell*, **155**, 1309–1322.

Baker,M. (2012) Gene data to hit milestone. *Nature*, **487**, 282–283.

Barbie,D.A. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Barry,W.T. *et al.* (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Carlson,M. (2016) hgu133plus2.db: Affymetrix human genome u133 plus 2.0 array annotation data (chip hgu133plus2). *R Package version 3.2.3*.

Chandel,N.S. *et al.* (2000) Reactive oxygen species generated at mitochondrial complex III stabilize hypoxia-inducible factor-1alpha during hypoxia: a mechanism of O_2 sensing. *J. Biol. Chem.*, **275**, 25130–25138.

Chen,L. *et al.* (2006) Temporal transcriptome of mouse ATDC5 chondroprogenitors differentiating under hypoxic conditions. *Exp. Cell Res.*, **312**, 1727–1744.

Corsello,S.M. *et al.* (2017) The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.*, **23**, 405–408.

Culhane,A.C. *et al.* (2012) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.*, **40**, D1060–D1066.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, **23**, 1846–1847.

de Leeuw,C.A. *et al.* (2016) The statistical properties of gene-set analysis. *Nat. Rev. Genet.*, **17**, 353–364.

Edgar,R. *et al.* (2002) Gene expression omnibus: ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Eklund,A.C. and Szallasi,Z. (2008) Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol.*, **9**, R26.

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

Gautier,L. *et al.* (2004) affy – analysis of affymetrix genechip data at the probe level. *Bioinformatics*, **20**, 307–315.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Goodspeed,A. *et al.* (2016) Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Mol. Cancer Res.*, **14**, 3–13.

GSEA User Guide (2018) http://software.broadinstitute.org/gsea/ doc/gseauser guideframe.html.

Gudas,J.M. *et al.* (1999) Cyclin E2, a novel G1 cyclin that binds Cdk2 and is aberrantly expressed in human cancers. *Mol. Cell. Biol.*, **19**, 612–622.

Gustafsson,M.V. *et al.* (2005) Hypoxia requires notch signaling to maintain the undifferentiated cell state. *Dev. Cell*, **9**, 617–628.

Hänzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, **14**, 7.

Huang,H. *et al.* (2012) PAGED: a pathway and gene-set enrichment database to enable molecular phenotype discoveries. *BMC Bioinformatics*, **13**, S2.

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

Katsuoka,F. *et al.* (2005) Genetic evidence that small maf proteins are essential for the activation of antioxidant response element-dependent genes. *Mol. Cell. Biol.*, **25**, 8044–8051.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Comput. Biol.*, **8**, e1002375.

Lappin,D.F. and Whaley,K. (1991) Modulation of complement gene expression by glucocorticoids. *Biochem. J.*, **280**, 117–123.

Lavallée-Adam,M. *et al.* (2014) PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. *J. Proteome Res.*, **13**, 5496–5509.

Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLOS Comput. Biol.*, **4**, e1000217–e1000219.

Liberzon,A. *et al.* (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.

Lu,H. *et al*. (2002) Hypoxia-inducible factor 1 activation by aerobic glycolysis implicates the Warburg effect in carcinogenesis. *J. Biol. Chem*., **277**, 23111–23115.

Moellering,R.E. *et al*. (2009) Direct inhibition of the NOTCH transcription factor complex. *Nature*, **462**, 182–188.

Mootha,V.K. *et al*. (2003) Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*., **34**, 267–167–73.

Rath,P.C. and Aggarwal,B.B. (1999) TNF-induced signaling in apoptosis. *J. Clin. Immunol*., **19**, 350–364.

Ritchie,M.E. *et al*. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*., **43**, e47.

Sahlgren,C. *et al*. (2008) Notch signaling mediates hypoxia-induced tumor cell migration and invasion. *PNAS*, **105**, 6392–6397.

Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550.

Subramanian,A. *et al*. (2017) A Next Generation Connectivity Map: l 1000 Platform and the First 1,000,000 Profiles. *Cell*, **171**, 1437–1452.e17.

Tian,L. *et al*. (2005) Discovering statistically significant pathways in expression profiling studies. *PNAS*, **102**, 13544–13549.

Tomfohr,J. *et al*. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Wang,Z. *et al*. (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun*., **7**, 12846.

West,D. *et al*. (2016) GR and ER coactivation alters the expression of differentiation genes and associates with improved ER+ breast cancer outcome. *Mol. Cancer Res*., **14**, 707–719.

Xia,J. and Wishart,D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res*., **38**, W71–W77.

Yano,A. *et al*. (2006) Glucocorticoids suppress tumor angiogenesis and in vivo growth of prostate cancer cells. *Clin. Cancer Res*., **12**, 3003–3009.

Yoo,M. *et al*. (2015) DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, **31**, 3069–3071.

Zhang,K. *et al*. (2010) i-GSEA4gwas: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res*., **38**, W90–W95.