High-Throughput Functional Annotation of Natural Products by Integrated Activity Profiling

Suzie K. Hight[1], Kenji L. Kurita[3,6], Elizabeth A. McMillan[1,5] Walter Bray[2], Trevor N. Clark[3], Anam F. Shaikh[4], F. P. Jake Haeckl[3], Fausto Carnevale-Neto[3], Scott La[2], Akshar Lohith[2], Rachel M. Vaden[1,5], Jeon Lee[1], Shuguang Wei[4], R. Scott Lokey[2], Michael A. White[1,5], Roger G. Linington[4*], John B. MacMillan[2,3*]*

[1]Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

[2]Department of Chemistry, University of California Santa Cruz, Santa Cruz, CA 95064, USA.

[3]Department of Chemistry, Simon Fraser University, Burnaby, BC, Canada.

[4]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

[5]Present Address: Oncology Research and Development, Pfizer, Inc., San Diego, CA 92121, USA

[6] Present Address: Department of Small Molecule Pharmaceutical Science, Genentech, 1, DNA Way, South San Francisco, CA 94080, USA

Abstract

Determining mechanism of action (MOA) is one of the biggest challenges in natural products discovery. Here, we report a comprehensive platform that uses Similarity Network Fusion (SNF) to improve MOA predictions by integrating data from the cytological profiling high-content imaging platform and the gene expression platform FUSION. The predictive value of the integrative approach was assessed using a library of target-annotated small molecules as benchmarks. Using KS-tests to compare in-class to out-of-class similarity, we found that SNF resulted in improved power to correctly assign MOA over either dataset alone. Furthermore, we integrated untargeted metabolomics of complex natural product fractions to map biological signatures to specific metabolites. Three examples are presented where SNF coupled with metabolomics was used to directly functionally characterize natural products and accelerate identification of bioactive metabolites. Our results support SNF integration of multiple phenotypic screening approaches along with untargeted metabolomics as powerful approach for advancing natural products drug discovery.

Introduction & Background

Assigning the mechanism of action to botanicals, natural products and synthetic chemicals remains a major challenge in chemical biology. Despite the technological advances in isolation, synthesis and screening strategies that make many bioactive substances available, in most cases their biological targets remain unknown[1,2]. This challenge is exacerbated when taking a systems-level approach to gain mechanistic information about entire collections of molecules, such as natural product libraries from microorganisms, marine invertebrates or plants. Due to the difficulties associated with isolation of natural products, the majority of academic and industrial natural product collections are complex fractions or crude extracts. Thus, in addition to the biological complexity of assigning mechanism of action, there is the added complication that typical one-compound-one-well assay formats are not applicable[3].

There has been a concerted effort in academia as well as biopharma to return to phenotypic screening approaches in drug discovery efforts[4]. This paradigm shift has come with the development of novel, information-rich phenotypic approaches that have the ability to provide a level of mechanistic understanding. These methods take advantage of gene expression profiling, high content imaging[5,6], yeast chemical genetics[7], proteomics[8,9], and others. While these platforms are valuable individually, each one is subject to the limitations of the reporter system used and thus can still fail to detect real biological associations. Here we test the idea that using computational tools to

integrate screening results from orthogonal screening platforms will enhance mechanism of action prediction and provide greater confidence in MOA predictions. In this study, we integrate gene expression-based (FUSION) and high content imaging-based (Cytological Profiling) screening platforms using Similarity Network Fusion (SNF), resulting in a novel and improved framework for the functional annotation of natural products.

Our interest in the functional characterization of microbe-derived natural products and botanicals led us to independently develop phenotypic screening strategies to evaluate natural product fraction libraries from marine bacteria. One platform, termed Functional Signatures of Ontology (FUSION) takes advantage of the power of perturbation-induced gene expression signatures coupled with pattern-matching tools to produce verifiable guilt-by-association mode-of-action (MOA) hypotheses[10]. The platform used the colon cancer cell line HCT116 and six reporter genes to generate expression signatures of "known" perturbagens such as siRNAs or mechanistically validated small molecules, for comparison to "unknown" perturbagens such as natural product fractions and pure natural products. This method has been used to characterize a series of microbially-derived molecules with unique mechanisms of action[10-14]. In this study, we have adapted the FUSION approach to a non-small cell lung cancer background using the cell line NCI-H23 and a new set of 14 reporter genes that form the basis for pattern-matching between known and unknown perturbagens (manuscript in preparation). A limitation of this approach, as well as other gene expression-based approaches such as the Connectivity Map [15,16], is that the sensitivity and specificity of the signature a bioactive molecule can

produce is dependent on the biological context of the assay. For example, interrogation of signaling pathways which are not intact in the chosen biological model may result in either a null or off-target pharmacological effect. Consideration of these factors is important when utilizing these types of data to inform MOA hypotheses.

A biologically orthogonal platform, cytological profiling (CP), takes advantage of high content image analysis to cluster perturbations based on features extracted directly from microscopy images from cells stained with a panel of fluorescent probes[17]. Consistent with existing approaches to high-content screening, our CP platform utilizes unsynchronized HeLa cells, which, after treatment with perturbagen for 24 h, are fixed and stained with probes for S-phase progression, total DNA, phospho-histone H3, tubulin and actin[3]. Automated fluorescence microscopy yields separate images for each probe, from which a total of 251 unique cytological features are extracted for each perturbagen. The distribution of these features is normalized to the distributions in DMSO-treated control wells to generate a “histogram difference” (HD) score, and the aggregate HD scores of all >200 features constitute a compound’s cytological fingerprint. Clustering compounds by their CP fingerprints reveals both well-established associations among compounds with the same target or MOA, as well as novel or unexpected associations and unique phenotypes of natural products [3,18]. Again, while CP and related high-content imaging platforms have demonstrated success for evaluating mechanism of action, they can be subject to limited resolution of bioactive compounds with broad cellular effects or limited sensitivity to bioactive compounds with that engage morphologically silent mechanisms[18].

These technologies have been used by our respective groups to evaluate the mechanism of action of complex natural product mixtures and pure natural products[3,10,11,19]. We hypothesized that a bioinformatic approach to integrate the two platforms would expand the biological space covered, decrease the number of natural products with a null signature, enhance weaker signatures from individual platforms, and provide greater confidence in mechanistic predictions. A challenge with integration of diverse data types is how to handle the different numbers and scales of features in individual datasets. To solve this problem, we have adopted similarity network fusion (SNF[20], which overcomes this challenge by constructing similarity networks individually for each available data type and then fusing these into a single network based on shared similarity across multiple datasets. SNF has been used efficiently in a range of applications, including integration of image-based profiling data to capture cell-to-cell heterogeneity[21], integration of drug sensitivity and structural data to generate a reference dataset that can be used to infer MOA of uncharacterized compounds[22] and for identification of cancer subtypes from multiple datatypes[20].

Unlike synthetic screening libraries, which are typically arrayed in one-compound-one-well format, natural products screening libraries are typically prepared as complex mixtures. To relate phenotypes to specific components in these mixtures we required both a detailed description of the chemical constitution of each fraction, and informatics tools to define the associations between constituents and phenotypes. Accurate

characterization of chemical composition for natural products libraries remains an unsolved challenge in the field. This objective is particularly difficult because natural products can be present at a large range of concentrations, with variable ionization efficiencies and a wide range of polarities and molecular weights.

Current mass spectrometry-based methods often yield peak lists with very high false discovery rates, making it difficult to identify biologically relevant features from these large results files. We developed bespoke acquisition and data processing methods designed to describe the chemical constitution of the natural products fraction library while removing interference signals caused by instrument noise and systemic contaminants from the sample processing workflow. These methods included appropriate replicates, blanks, and sample preparation workflow, employing an ion-mobility spectroscopy-enabled ESI-qTOF. The inclusion of ion mobility spectroscopy improved separation of co-eluting components and improved the binning of analytes between samples by providing a third axis of separation (collision cross-sectional area) over traditional methods. Using a modified version of our Compound Activity Mapping platform[19], we then defined activity scores for all analytes based on SNF clustering results and developed a custom data visualization platform to directly relate analytes to specific biological phenotypes.

Here we describe the integration of Cytological Profiling and FUSION screening data to create a platform for functional annotation that significantly outperforms either individual method for *de novo* mode of action prediction. By combining this with next-generation

metabolomics analysis of natural products libraries, we have created a unique framework for natural product biological characterization (Figure 1A). The value of this platform is illustrated by three examples that demonstrate how integration of these three data types can help drive natural product discovery.

Results

Integration of multiple platforms leads to improved in-class target classification

As a test-of-concept, we profiled a small collection of randomly selected 628 microbial natural product fractions of unknown composition through our FUSION, CP, and metabolomics profiling platforms. For reference benchmarks, we also profiled a library of 2034 known synthetic small molecules that were selected for known bioactivity and have been annotated for mechanism of action and/or molecular target (Supplementary Table 1). Briefly, all perturbagens were screened in triplicate and at the same dose in both platforms. FUSION gene expression signatures were normalized to non-treated wells, and CP fingerprints were normalized to DMSO-treated wells (see Methods). A Z-score transformation was then applied to both normalized datasets two-way hierarchical clustering was performed (Figure 1B). Analysis of this clustering revealed that natural product fractions were interspersed throughout each dataset, with more dispersion in FUSION than in CP. Comparison of the pair-wise Pearson correlations between NPFs and all other perturbagens in FUSION vs CP showed that while some positive correlations

trend in the same direction, the overall concordance between the two datasets is relatively low (Figure 1C). This suggests that each dataset may report on the same biological space differently.

Generation of a fused similarity network across CP and Fusion signatures could aid in boosting signal in individual datasets by grouping together perturbagens based on similarity in both phenotypic and cellular readouts. However, integration of orthogonal datasets is a computational challenge due to inherent differences in experimental collection, measured features, noise, and overall scale between methods[23]. In order to test the idea that combining the information from FUSION and CP would lead to an improved platform for MOA assignment, we turned to a recently developed approach, called Similarity Network Fusion (SNF)[20]. This method addresses challenges associated with differences in scale and feature measurement by first constructing within-sample similarity networks for each data type. A single similarity matrix is then generated by iteratively propagating similarity information simultaneously across all individual networks to generate a single, fused similarity matrix where perturbagens with evidence of similarity across multiple datasets result in higher similarity measures. (see Supplementary Figure 1, and Methods). This similarity matrix was then treated as a matrix of weights to calculate a new Euclidean distance or Pearson correlation matrix, and subjected to hierarchical affinity propagation clustering [24,25] to group perturbagens based on this metric (see Methods).

In order to assess the performance of the individual and fused datasets in assigning MOA, we used our collection of commercial compounds and their target annotations as benchmark references (SelleckChem). Among the 195 pre-annotated target classes within this collection, 89 classes contained five or more chemicals. A two-sample, one-sided Kolmogorov–Smirnov (KS) test was applied to each of these 89 target classes to determine if the pairwise similarities between chemicals, as determined by FUSION or CP, with the same target annotation ("in-class") were significantly closer or more correlated than the pairwise similarities between these chemicals and those from other target classes ("out-of-class"). We used both Euclidean distance and Pearson correlation as similarity metrics. Perturbagens with high Pearson correlation will have signatures whose overall trend is in the same direction, but whose magnitudes may be very different. This can be useful when considering perturbagens which may have similar biological effects but different levels of potency but will also have the effect of dispersing noise throughout the dataset. Meanwhile perturbagens with small Euclidean distances will have signatures which are closely related in both direction and magnitude. Thus, this metric can be particularly useful to make fine distinctions between different mechanisms of action but may group noisy signatures together. Comparison of the p-values for each target class across all datasets indicated that SNF identified more target classes as significantly self-associated than either FUSION or CP alone (Figure 2A; all CDF plots for each target class are included in Supplementary Figure 2). For example, in Euclidean distance networks, pairwise distances between chemicals in the Bcl-2 target class are clearly smaller than out-of-class pairwise distances in SNF, but not in either FUSION or

CP (Figure 2B). When Euclidean distance was used as the similarity metric, SNF performed either as well or better than individual datasets for most target classes. In classes where Euclidean distance SNF performed worse, the difference between p-values from individual datasets and SNF was relatively small. By contrast, many target classes in individual datasets performed better than SNF when Pearson correlation was used (Supplementary Figure 3). This may reflect a lower resolving power by the Pearson metric, in that it would reflect trends but does not provide the finer discrimination between magnitude of values that Euclidean distance metric does. Notably, most target classes identified as significant by SNF-Euclidean were also identified by SNF-Pearson (Supplementary Figure 4A). Euclidean distance also showed greater improvement over individual datasets in terms of total classes identified as significant compared to Pearson, possibly due to the power of this metric in distinguishing between noise and signal (Supplementary Figure 4). Taken together, these analyses demonstrated that integrating the FUSION and CP datasets together using similarity network fusion resulted in improved power to distinguish unique mechanisms of action from the background.

SNF integration drives clustering of natural product fractions

We next used SNF values to construct a relational network among the reference compounds and natural product fractions using hierarchical affinity propagation clustering (APC) as described previously[25] (Figure 3). This clustering method was chosen as it is a deterministic clustering method that defines, in a data-driven fashion, both the number

and membership of clusters emerging from a given similarity matrix[24]. Coloring edges based on the contribution from each individual dataset revealed that some associations were driven by single datasets, but the majority were derived by fusing information from both FUSION and CP (see Methods; Figure 3A). Notably, most of the perturbagens that were flagged as "dead" in either platform clustered together in the SNF network, and this list included compounds for which cytotoxicity would be expected at the doses used in these assays (i.e., topoisomerase inhibitors; Supplementary Figure 5). In general agreement with our KS-test results, many clusters were significantly enriched for chemicals with the same target annotation, as assessed by a hypergeometric test (Figure 3B; Supplementary Table 2). We also observed that some clusters were significantly enriched for multiple classes, which may be a reflection of similar mechanisms of action and/or convergence of downstream signaling effects (e.g., enrichment of PI3Ki, mTORi and EGFRi in Cluster 123). It is also possible that overlap of multiple target classes in the same cluster may reflect a limitation of the gene set, cytological features, or the cell lines selected for profiling in both platforms, in that these reporters may not be sufficient to distinguish between those mechanistic classes.

Untargeted metabolomics relates chemical constitution to functional signatures via the SNF-similarity score

The chemical complexity of natural product fractions increases the difficulty in relating phenotypes to specific molecules or sets of molecules for a given sample. However, in

most cases biological activities are driven by a single compound or a small subset of compounds in each extract. By determining the distribution of secondary metabolites across the full sample set, it is possible to test the hypothesis that extracts with similar phenotypes contain the same or similar bioactive species. To accomplish this objective, it is necessary to create a clear picture of chemical constitution across the sample set.

In this study we performed untargeted metabolomics on the full set of natural products extracts using a UPLC-IMS-qTOF instrument operating in data-independent acquisition mode (DIA). Inclusion of ion mobility spectrometry affords an additional axis of separation over standard LCMS systems that improves separation of complex mixtures and provides an additional physicochemical measure (collisional cross-sectional area) for matching analytes between samples. Use of DIA increases the percentage of analytes that are subjected to MS^2 fragmentation compared to traditional data-dependent acquisition (DDA). These fragmentation patterns are useful for comparing analytes between samples, and for comparing to external reference libraries for compound identification[26].

Samples were analyzed as three independent technical replicates, and consensus feature lists generated for each sample using a suite of in-house data processing scripts. Features were required to appear in at least two of three replicates to be included in the consensus feature list. These sample by sample feature lists were then 'basketed' to produce a single list of unique mass spectrometric features across the full sample set. This feature list included information about mass spectrometric properties (e.g. retention

time, mass to charge ratio, collision cross-sectional area) as well as distribution across the sample set.

To examine the relationship between individual features and the SNF network, we employed a variation of our previously developed Compound Activity Mapping method to score predicted feature activities[19]. Using the feature list from the metabolomics profiling data, we sequentially identified the subset of samples containing each feature and calculated the average of the SNF similarity scores within this set (see Supplementary Figure 6 and Methods). This score provides a numerical evaluation of how closely the presence of a specific mass feature is correlated with the presence of a specific biological phenotype in the APC network. In cases where a given feature is responsible for an observed activity, it is expected that the phenotypes of the associated set should be similar, and that the average SNF similarity score (SNF score) should be correspondingly high. By contrast, compounds that do not impart a biological response should not correlate with a specific biological signature, and the SNF score should be correspondingly weak. Analysis of 668 natural product fractions identified a total of 8108 features, of which 3498 appear only once in the sample set. Of the 4610 features that appear at least twice, just 315 have SNF scores ≥ 0.25, indicating that most features have low or no correlation with specific biological phenotypes; this is in line with typical hit rates for natural product screening programs[27].

SNF scoring is feature independent, meaning that a high score for one mass spectrometry

feature has no impact on the scores of other features in the sample. This is important because the mass spectrometry data are not deconvoluted by either adduct (e.g. $[M+H]^+$ vs $[M+Na]^+$) or in-source fragments (e.g. $[M-H_2O+H]^+$). It is therefore common to identify a suite of mass spectrometry features with the same retention time that all possess strong SNF scores. These features can be used in concert to determine the correct accurate mass for the active component (which aids in dereplication) and to reconstitute mass spectrometry fragments (which can help with metabolite identification).

Calculating SNF scores for every mass spectrometric feature provides a metric to filter and prioritize compounds for subsequent isolation. For example, identifying features with high SNF scores present in a specific cluster in the APC plot can be used to directly target molecules with prioritized biological properties. Conversely, in situations where clusters contain several classes of bioactive compounds, SNF scoring can be used to subdivide these clusters based on biologically active chemistry, even in cases where these samples are mechanistically indistinguishable. Using the open source Bokeh server library, we created a data visualization tool that enables direct examination and filtering of the untargeted metabolomics data with a range of data display options (Supplementary Figure 6).

In order to evaluate the efficiency of this new platform for *de novo* bioactivity prediction from complex mixtures, we tested two different query approaches. These included querying the SNF network for natural products clustering predominately near other

reference compounds and filtering for metabolites with highly correlated biological activity. These strategies were selected to test the platform under different conditions, from simple situations where the annotations were unanimous, to complex situations with multiple reference compound types and multiple natural product fractions. In each approach, we highlight the contribution of SNF and metabolomics towards identification of both the natural product driving the signature and its biological mechanism of action.

Identification of trichostatin A from an HDAC-inhibitor enriched cluster validates the integrated SNF platform

We first sought to validate the SNF network by querying the dataset for natural product fractions that clustered mainly with reference compounds from a single target class. In the SNF-Euclidean APC, there were 6 clusters that were highly enriched for chemicals belonging to the same target class ($p<1E-10$): Cluster 48 (HSP), Cluster 49 (HDAC), Cluster 76 (Epigenetic Reader Domain), Cluster 100 (mTOR), and Cluster 103 (Proteasome) (Figure 3B; Supplementary Table 2, SNF-Euclidean APC Cytoscape file). Of these, Clusters 49, 100 and 103 contained natural product fractions (3 in Cluster 49, 1 in Cluster 100, and 1 in Cluster 103), thus identifying readily testable MOA hypotheses for each of these. A KS-test also confirmed that association between chemicals in the HDAC inhibitor target class is preserved in the full dataset, and that these associations were still significantly improved in SNF ($p=1.8e-61$; Figure 4A).

We chose to focus on the HDAC inhibitor cluster for validation because it contained 3 sequential natural product fractions from the Linington library (SW218953, SW218954, and SW218955) (Figure 4B). The presence of multiple NPFs from the same series suggests the presence of common metabolite profiles. Filtering of the metabolomics data by setting the minimum SNF score to 0.25 and limiting the display to only those features present in the three natural products fractions in this cluster revealed a vertical "stripe" of mass spectrometry features at 3.13 minutes with a parent mass feature of 323.1713 *m/z* (Figure 4C; red box). This pattern of signals is indicative of both a parent mass and associated in-source fragments from the LCMS analysis (extracted ion chromatograms for the NP fractions are included in Supplementary Figure 7). Subsequent chromatographic optimization, purification and NMR analysis from SW218953 (Figure 4D) identified this product as the known bacterial metabolite trichostatin A (Figure 4E). Trichostatin A has been extensively studied for its activity as an HDAC inhibitor[28,29]. Notably, Cluster 49 also contained pure trichostatin A from the Selleck library (cluster members are listed in Supplementary Table 3). An analysis of the top 50 nearest neighbors to trichostatin A in the Euclidean SNF dataset confirms that these three natural product fractions are tightly associated with HDAC inhibitors (Figure 4F, CP images can be found in Supplementary Figure 8). Therefore the *a priori* MOA prediction from the HIFAN platform aligns well with the literature data for this class of bioactive molecules, demonstrating the power of this annotation strategy.

Metabolomics-driven SNF cluster identifies desferrioxamine class natural products

Next we explored whether filtering for highly interrelated SNF scores could identify novel metabolites in our natural product libraries. Filtering the metabolomics data using the standard approach outlined above, with an SNF score of 0.15, identified two related molecules with parent mass to charge ratios of 587.3395 and 601.3567, present in SW218716, SW218717 and SW218718, SW218850, SW218755 and a few other natural product fractions (Figure 5A, Supplementary Figure 9). Chromatographic optimization, purification and structure elucidation identified these metabolites as desferrioxamines D2 and E (DFOD and DFOE) (Figure 5B). Querying the SNF-Euclidean network for the top 50 nearest neighbors to SW218850 revealed that it is closely associated with other DFOE-containing natural product fractions, with two formulations of Ciclopirox (SelleckChem catalog numbers S2528 and S3019), a known bidentate iron chelator, and another iron chelator Deferasirox (SelleckChem S1712) (Figure 5C). In addition to associating with natural product fractions and iron chelators, when the region of the APC map that contains these natural product fractions is examined, it is enriched with reference compounds that inhibit DNA synthesis and/or inhibit cell cycle progression (Figure 5D). Iron depletion is known to cause G1/S arrest in a variety of cell types [30-32]. Desferrioxamine specifically has been has shown to induce G1/S arrest due to activation of HIF-1a[33], activation of members of the MAPK signaling pathway[34], and upregulation of the GADD family of stress response genes[35]. Thus, the prediction is entirely consistent with literature evidence for the desferrioxamine family of metabolites.

These data confirm that the use of the SNF score with untargeted metabolomics can be used to quickly identify the active metabolite in fractions that share a phenotype. This is also an example of a commonly encountered scenario in natural products research, where several natural product fractions show a distinctive phenotype that is driven by the presence of a commonly encountered compound class. In this case, we were able to rapidly assess this natural product fractions in this grouping to be dominated by the DFO family of iron chelators with the clustered biology being associated with inhibition of cell cycle progression. With this assignment in hand, future profiling efforts will directly annotate and exclude these fractions from further consideration.

Discussion

The natural products literature contains thousands of examples of novel compounds with biological activities reported from simple end-point assays, such as MTT cytotoxicity and antimicrobial growth inhibition assays. While this provides a handle for further investigation, the lack of detailed mechanistic information means that the majority of these molecules are never followed up on for biological characterization. This is due to the aforementioned challenges associated with characterizing the mode of action of pharmacological agents. Previous biological screening platforms developed by our laboratories (cytological profiling and FUSION) have been successful at characterizing new natural products with detailed mechanistic assignments. While powerful, both platforms encountered scenarios where no prediction for a natural product fraction was

possible due to weak or non-standard signatures. Differences in both resolution and sensitivity between platforms can limit their utility, either because a given mechanism is not reported on by the assay system, or because the resolving power of the platform is insufficient to differentiate between mechanistic classes. To address this issue, we applied the Similarity Network Fusion (SNF) data integration method to integrate data from both cytological profiling and FUSION. In principle, application of SNF to these two datasets should identify unique biological associations that would not have been identified from either platform independently.

The SNF network demonstrated improved assignment of pure chemicals to their annotated target class compared to either FUSION or CP, delivering an enhanced capacity for untargeted mode of action prediction. We validated the utility of this network in assigning MOA by demonstrating that natural product fractions containing trichostatin A were clustered with pure trichostatin A and other HDAC inhibitors. We then further developed a robust pipeline to assign the mechanistic annotation that the SNF network provides to specific natural product structures. Using untargeted metabolomic profiling of the full natural product fraction library and creating a scoring method (SNF score) to relate these mass spectrometry features to defined phenotypes, it is possible to directly predict the contributions of all mass spectrometry features to the biological landscape of the sample set. SNF scores provide a rich perspective on chemical interpretation from the natural products library. For example, in situations where two different compound classes cause the same biological phenotype (i.e. one cluster in the APC network), SNF scores

will correctly identify these two compounds as high priority candidates, even though neither compound is present in all members of the biological cluster, provided that each molecule is predominantly found within that cluster. Similarly, in situations where extracts contain many features, most features will be quickly deprioritized because their distributions throughout the sample set do not correlate to specific biological phenotypes. This mechanism for compound prioritization is therefore a robust and powerful strategy for directly targeting biologically relevant compounds from large, complex natural product libraries. Development of the Bokeh server visualization suite (Supplemental Figure 5) provides a facile platform for data filtering and visualization that enables the rapid exploration of these data using a range of different viewpoints.

Notwithstanding the value of this approach, there are several situations which remain difficult to resolve. Currently, the SNF score is not weighted by relative intensity of each MS feature. This is because determining relative concentrations of unknown analytes in complex samples remains an unsolved challenge in mass spectrometry-based metabolomics. In situations where a bioactive metabolite is present both above and below its EC_{50} the SNF score will be reduced, as there will be no measurable phenotype in extracts where the concentration is low. Secondly, the system cannot differentiate between active and inactive metabolites if they are always co-expressed. Review of the metabolomics data suggests that this circumstance is rare, however in these cases both metabolites would be scored as active candidates, requiring downstream deconvolution. Finally, in situations where bioactive compounds are frequently encountered with other

unrelated bioactives the resulting phenotypes could bear little relationship to one another. Review of the dataset suggests that this situation is also unusual, however in these cases SNF scores will also deteriorate because of the reduced similarity scores between samples with different phenotypic signatures.

Natural products research brings with it a number of challenges, such as the chemical complexity of extracts, re-isolation of known compounds and characterization of biological activity. These challenges limit the pace of natural product research and leave knowledge gaps around the value of a given natural product structure or class. Recent initiatives to develop resources to better understand the genomics of natural product biosynthetic gene clusters[36] and the development of the Global Natural Products Social (GNPS) molecular networking platform[26] have fundamentally changed how natural product research is conducted, but the field as a whole is far behind in leveraging 'Big Data' to address outstanding challenges. The approach we have detailed here provides an unbiased, data-driven platform that can be used to integrate biological assay and metabolomics results to provide a comprehensive viewpoint on chemical/biological relationships in the natural product sphere.

Acknowledgements

This research was supported by NIH grants U41 AT008718 to R.G.L, M.A.W and J.B.M (NCCIH and ODS) and R01 CA149833 to J.B.M (NCI). E.A.M was supported by NIH training grant 5T32GM8203-27. R.M.V was supported by CPRIT training grant RP140110

and NIH training grant 5T32 CA124334-09. A.F.S was supported by NSF GRF#2017247469

Author Contributions

Conceptualization: J.B.M., R.G.L., M.A.W., R.S.L., E.A.M., S.K.H., K.L.K. Methodology: J.B.M., R.G.L., M.A.W., R.S.L., E.A.M., S.K.H., K.L.K., W.B., A.F.S. Software: E.A.M., S.K.H., K.L.K., A.L. Formal Analysis: S.K.H., K.L.K., E.A.M., W.B., T.N.C., F.P.J.H., A.F.S., A.L. Investigation: S.K.H., K.L.K., E.A.M., W.B., T.N.C., F.P.J.H., A.F.S., S.L., F.C.-N., R.M.V., A.L. Writing: S.K.H., M.A.W., R.G.L., R.S.L., J.B.M. Supervision: J.B.M., R.G.L., R.S.L., M.A.W. Funding: J.B.M., R.G.L., M.A.W., R.S.L.

Declarations of Interest

The authors have no conflicts of interest to report. Michael White, Rachel Vaden, and Elizabeth McMillan are current employees of Pfizer, Inc. Kenji Kurita is an employee of Genentech, Inc.

Lead Contact and Materials Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contacts, John B. MacMillan (jomacmil@ucsc.edu) and Roger G. Linington (rliningt@sfu.ca).

Figure legends

Figure 1. Overview of screening platforms and initial data collection. A) Experimental outline. Natural product fractions are isolated from marine bacteria, screened through two biological screening platforms (FUSION and CP), and subjected to high-resolution mass spectrometry-based metabolomics profiling. FUSION and CP data are integrated using Similarity Network Fusion (SNF) which is then provides biological annotation on individual metabolites identified. B) Two-way hierarchical clustering of Z-scores from FUSION and CP using Euclidean distance and complete linkage. NPFs are indicated with pink flags. C) Heat-scatter plot comparing Pearson correlations between NPFs and all other perturbagens in FUSION vs CP. Color scale ranging from low data density to high data density = gray – blue – red – orange – yellow.

Figure 2. Comparison of in-class target annotation vs out-of-class annotation KS-test p-values in FUSION, CP, and SNF. A) Dot plot of -log10 KS-test p-values for each dataset, using Euclidean distance as the similarity metric, by target annotation class. Significance threshold is represented by the horizontal line (p=0.01). B) CDF plots comparing pairwise Euclidean distances within one example target class, Bcl-2, against out-of-class associations (red, in-class; blue, out-of-class; p-value calculated by KS-test). Here, in-class vs. out-of-class associations were not significant in either FUSION or CP but are significant in SNF.

Figure 3. Affinity propagation clustering map of the SNF network preserves in-class target associations. A) Hierarchical affinity propagation clustering map of the SNF network using Euclidean distance as the similarity metric. Edges are colored based on contribution from individual datasets: Pink = driven by FUSION, blue = driven by both datasets, and green = driven by CP. Perturbagen type is indicated by node color: black = NPF, gray = pure chemical. B) Lollipop plot showing results of hypergeometric tests for each target annotation class, per cluster. Significance threshold is indicated by a horizontal line (Bonferroni-corrected alpha = 0.0016). For clusters with multiple target classes scoring as significant, points above the significance threshold are colored in order of increasing significance: Pink = 1st significant class, orange = 2nd significant class, green = 3rd significant class, and purple = 4th significant class. All data is available in Supplementary Table 2.

Figure 4. SNF correctly assigns MOA to the major metabolite in a series of natural product fractions. A) CDF plots comparing pairwise Euclidean distances between chemicals annotated as HDAC inhibitors in the full dataset, versus out-of-class associations. Gray, in-class; Black, out-of-class. Colored circles correspond to cluster membership in the associated APC map. The p-value was calculated by KS-test. B) Cluster 49 from the SNF-Euclidean map, drawn using a spring-embedded layout such that edge length is proportional to Euclidean distance. C) Retention plot showing

common mass spec features present in the NPFs highlighted panel B. D) Chemical structure of trichostatin A. E) NMR spectra confirming trichostatin A. F) The top 50 nearest neighbors to trichostatin A in the SNF-Euclidean network, with the three natural product fractions found in Cluster 49 labeled.

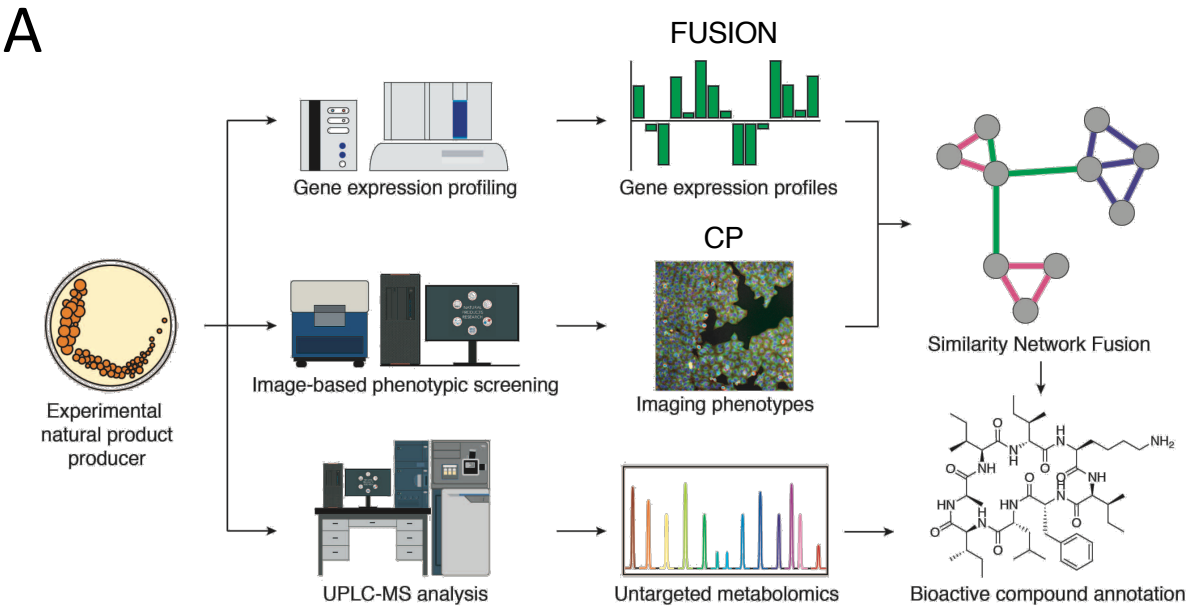
Figure 5. Filtering for highly correlated SNF scores can identify single metabolites with biological activity. A) Retention plot showing common mass spec features associated with SNF scores above 0.15. B) Chemical structures for desferrioxamines E and D2. C) The top 50 nearest neighbors to one of the natural product fractions that contained DFO-E, SW218850, in the SNF-Euclidean network. Other DFO-E containing natural product fractions and iron chelators present in the reference library are labeled. D) Region of the SNF-Euclidean APC map showing that DFOE-containing SWIDs are near each other and a cluster enriched for topoisomerase inhibitors (green).

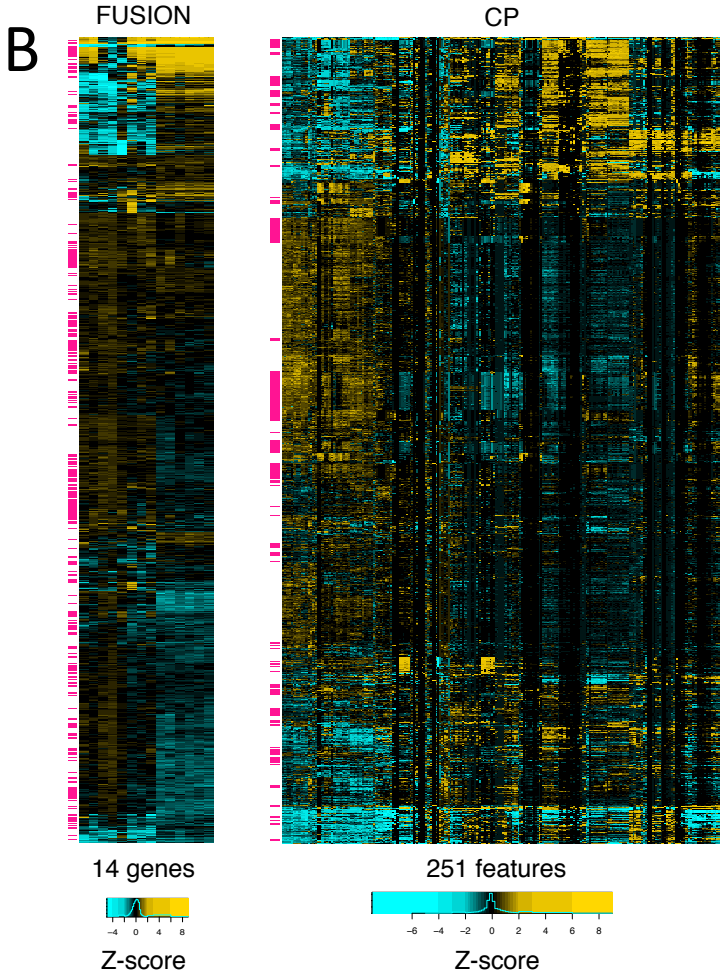
References

1. Carter, G. T. Natural products and Pharma 2011: Strategic changes spur new opportunities. *Nat. Prod. Rep.* **28,** 1783–7 (2011).
2. Wagner, B. K. & Schreiber, S. L. The Power of Sophisticated Phenotypic Screening and Modern Mechanism-of-Action Methods. *Cell Chemical Biology* **23,** 3–9 (2016).
3. Schulze, C. J. *et al.* Function-First Lead Discovery: Mode of Action Profiling of Natural Product Libraries Using Image-Based Screening. *Chemistry & Biology* **20,** 285–295 (2013).
4. Zheng, W., Thorne, N. & McKew, J. C. Phenotypic screens as a renewed approach for drug discovery. *Drug Discovery Today* **18,** 1067–1073 (2013).
5. Bray, M.-A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protoc* **11,** 1757–1774 (2016).
6. Bray, M.-A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6,** 1–5 (2017).
7. Parsons, A. B. *et al.* Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast. *Cell* **126,** 611–625 (2006).
8. Schenone, M., k, V. D. C. I., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* **9,** 232–240 (2013).
9. Rix, U. & Superti-Furga, G. Target profiling of small molecules by chemical proteomics. *Nat Rev Drug Discov* **5,** 616–624 (2009).
10. Potts, M. B. *et al.* Using functional signature ontology (FUSION) to identify mechanisms of action for natural products. *Science Signaling* **6,** ra90–ra90 (2013).
11. Potts, M. B. *et al.* Mode of action and pharmacogenomic biomarkers for exceptional responders to didemnin B. *Nat Chem Biol* **11,** 401–408 (2015).
12. Hu, Y. *et al.* Discoipyrroles A–D: Isolation, Structure Determination, and Synthesis of Potent Migration Inhibitors from Bacillus hunanensis. *J. Am. Chem. Soc.* **135,** 13387–13392 (2013).
13. Das, B. *et al.* A Functional Signature Ontology (FUSION) screen detects an AMPK inhibitor with selective toxicity toward human colon tumor cells. *Scientific Reports* 1–10 (2018). doi:10.1038/s41598-018-22090-6
14. Vaden, R., Oswald, N., Potts, M., MacMillan, J. & White, M. FUSION-Guided Hypothesis Development Leads to the Identification of N6,N6-Dimethyladenosine, a Marine-Derived AKT Pathway Inhibitor. *Marine Drugs* **15,** 75–10 (2017).
15. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
16. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171,** 1437–1452.e17 (2017).

17. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306,** 1194–1198 (2004).
18. Woehrmann, M. H. *et al.* Large-scale cytological profiling for functional analysis of bioactive compounds. *Mol. BioSyst.* **9,** 2604–14 (2013).
19. Kurita, K. L., Glassey, E. & Linington, R. G. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc Natl Acad Sci USA* **112,** 11999–12004 (2015).
20. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Meth* **11,** 333–337 (2014).
21. Rohban, M. H., Abbasi, H. S., Singh, S. & Carpenter, A. E. Capturing single-cell heterogeneity via data fusion improves image-based profiling. *Nature Communications* 1–6 (2019). doi:10.1038/s41467-019-10154-8
22. El-Hachem, N. *et al.* Integrative cancer pharmacogenomics to infer large-scale drug taxonomy. *Cancer Research* canres.0096.2017–41 (2017). doi:10.1158/0008-5472.CAN-17-0096
23. Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. & Wild, D. L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28,** 3290–3297 (2012).
24. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972–976 (2007).
25. Kim, J. *et al.* XPO1-dependent nuclear export is a druggable vulnerability in KRAS-mutant lung cancer. *Nature* **538,** 114–117 (2016).
26. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Rev Drug Discov* **34,** 828–837 (2016).
27. Amirkia, V. & Heinrich, M. Natural products and drug discovery: a survey of stakeholders in industry and academia. *Front. Pharmacol.* **6,** xlviii–8 (2015).
28. Hoshikawa, Y., Kwon, H. J., Yoshida, M., Horinouchi, S. & Beppu, T. Trichostatin A induces morphological changes and gelsolin expression by inhibiting histone deacetylase in human carcinoma cell lines. *Exp. Cell Res.* **214,** 189–197 (1994).
29. Wood, M., Rymarchyk, S., Zheng, S. & Cen, Y. Trichostatin A inhibits deacetylation of histone H3 and p53 by SIRT6. *Archives of Biochemistry and Biophysics* **638,** 8–17 (2018).
30. Lucas, J. J. *et al.* Effects of iron-depletion on cell cycle progression in normal human T lymphocytes: selective inhibition of the appearance of the cyclin A-associated component of the p33cdk2 kinase. *Blood* **86,** 2268–2280 (1995).
31. Brodie, C. *et al.* Neuroblastoma sensitivity to growth inhibition by deferrioxamine: evidence for a block in G1 phase of the cell cycle. *Cancer Research* **53,** 3968–3975 (1993).
32. Yu, Y., Kovacevic, Z. & Richardson, D. R. Tuning Cell Cycle Regulation with an Iron Key. *Cell Cycle* **6,** 1982–1994 (2014).
33. Woo, K. J., Lee, T.-J., Park, J.-W. & Kwon, T. K. Desferrioxamine, an iron chelator, enhances HIF-1α accumulation via cyclooxygenase-2 signaling

pathway. *Biochemical and Biophysical Research Communications* **343,** 8–14 (2006).
34. Yu, Y. & Richardson, D. R. Cellular iron depletion stimulates the JNK and p38 MAPK signaling transduction pathways, dissociation of ASK1-thioredoxin, and activation of ASK1. *J. Biol. Chem.* **286,** 15413–15427 (2011).
35. Saletta, F., Rahmanto, Y. S., Siafakas, A. R. & Richardson, D. R. Cellular Iron Depletion and the Mechanisms Involved in the Iron-dependent Regulation of the Growth Arrest and DNA Damage Family of Genes. *Journal of Biological Chemistry* **286,** 35396–35406 (2011).
36. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11,** 625–631 (2015).


A
FUSION
Gene expression profiling
Gene expression profiles
CP
Image-based phenotypic screening
Imaging phenotypes
Similarity Network Fusion
Experimental natural product producer
UPLC-MS analysis
Untargeted metabolomics
Bioactive compound annotation


B
FUSION
CP
14 genes
251 features
Z-score
Z-score

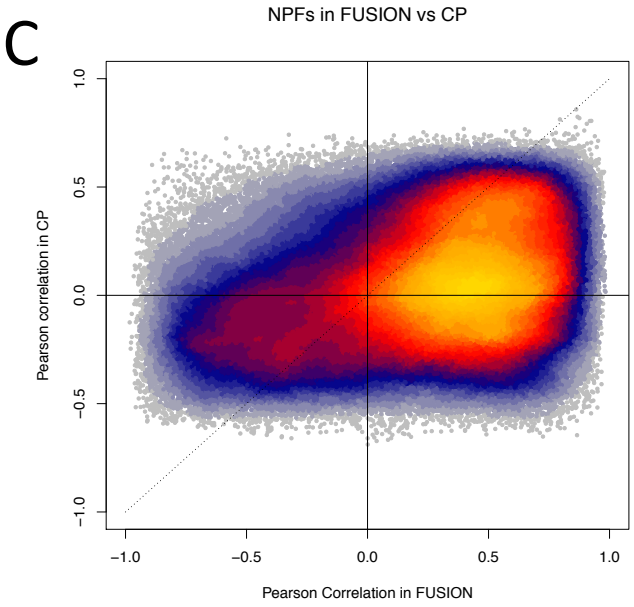
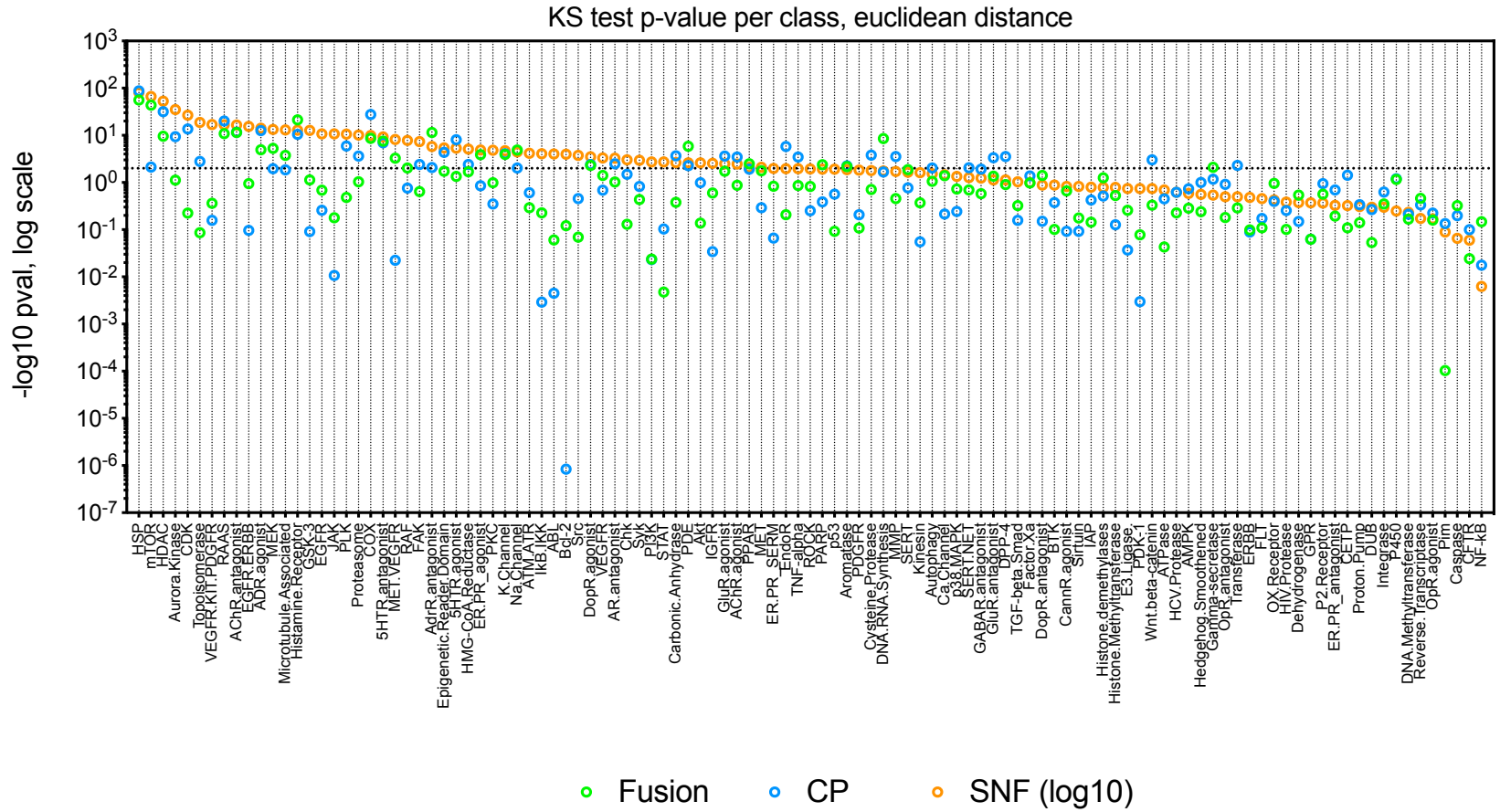

C
NPFs in FUSION vs CP
Pearson correlation in CP
Pearson Correlation in FUSION

Figure 2

A


KS test p-value per class, euclidean distance
-log10 pval, log scale
Fusion
CP
SNF (log10)

B

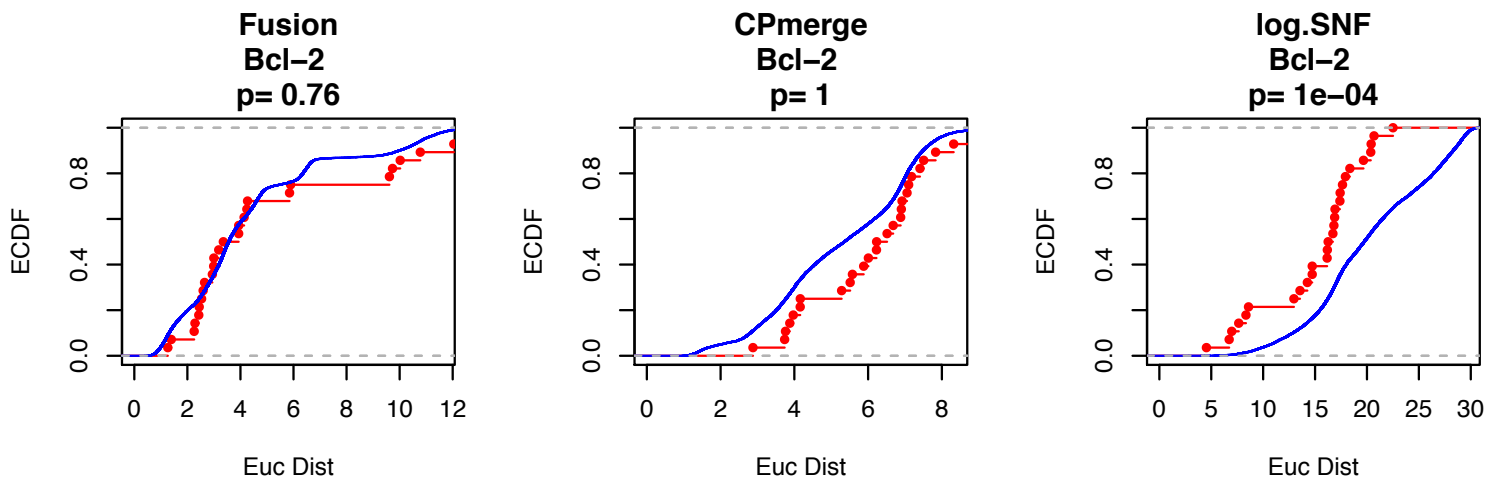
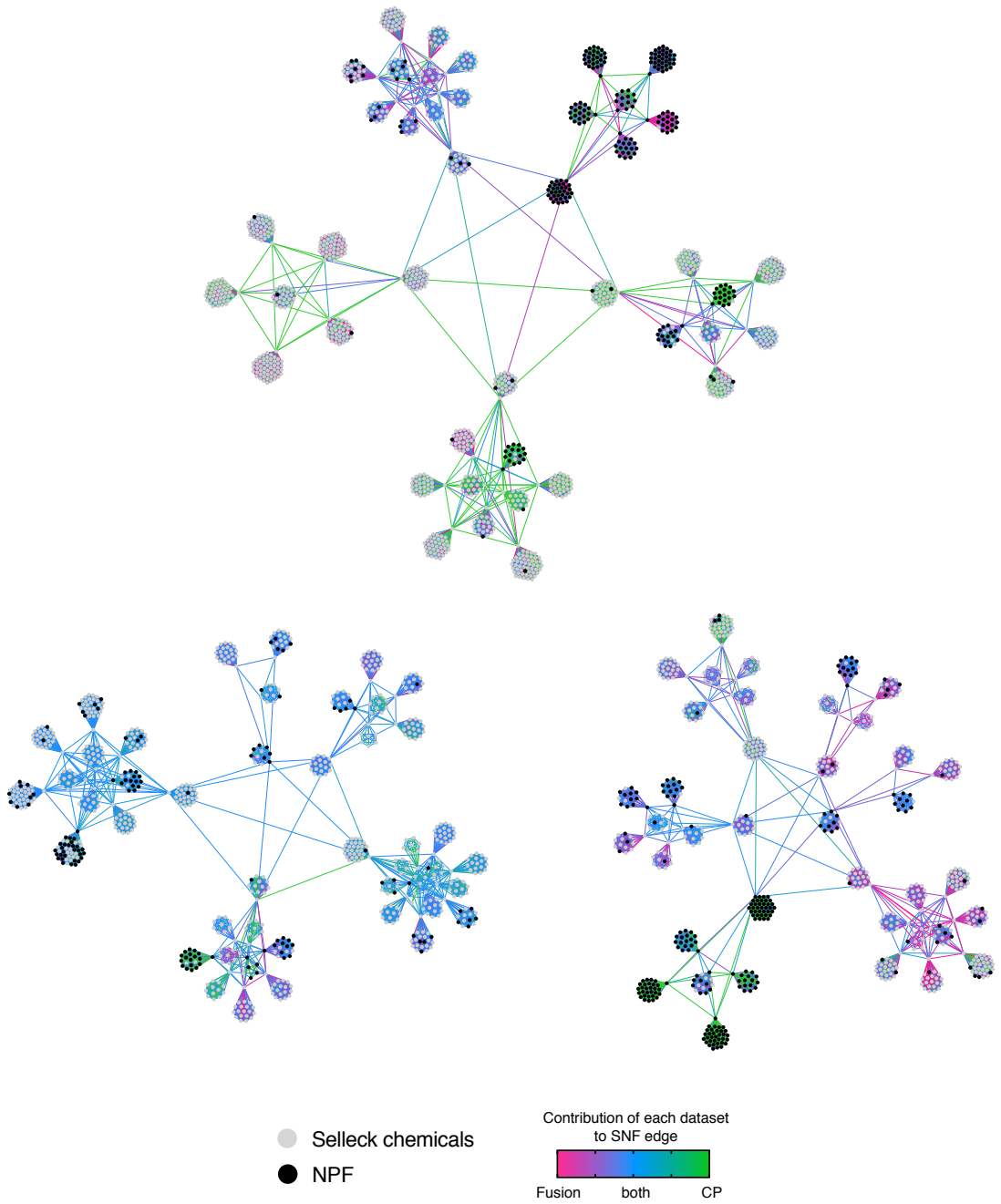

Fusion
Bcl−2
p= 0.76
CPmerge
Bcl−2
p= 1
log.SNF
Bcl−2
p= 1e−04
ECDF
Euc Dist

Figure 3

A


Selleck chemicals
NPF
Contribution of each dataset
to SNF edge
Fusion
both
CP

B

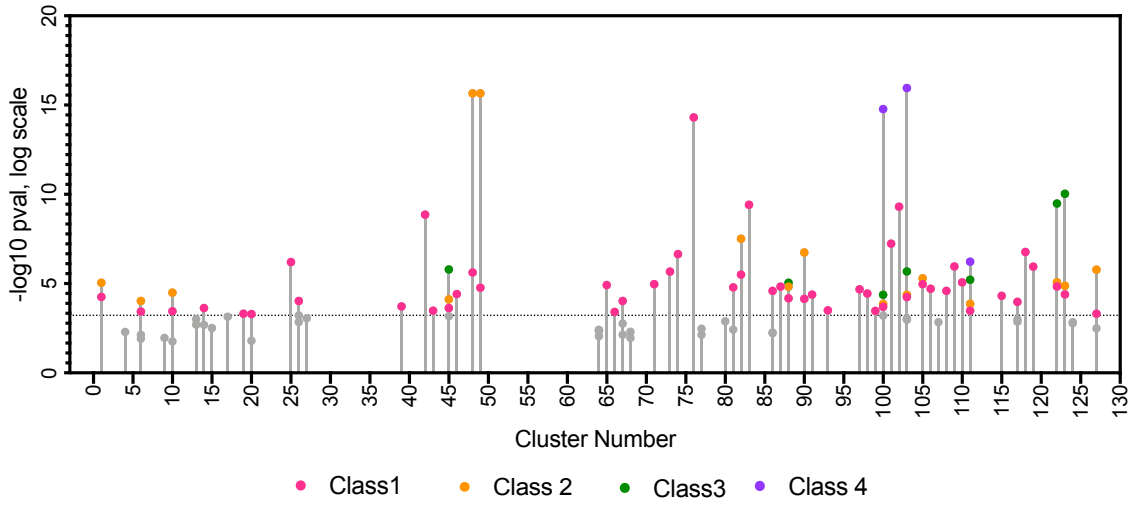

-log10 pval, log scale
0
5
10
15
20
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
Cluster Number
Class1
Class 2
Class3
Class 4

Figure 4

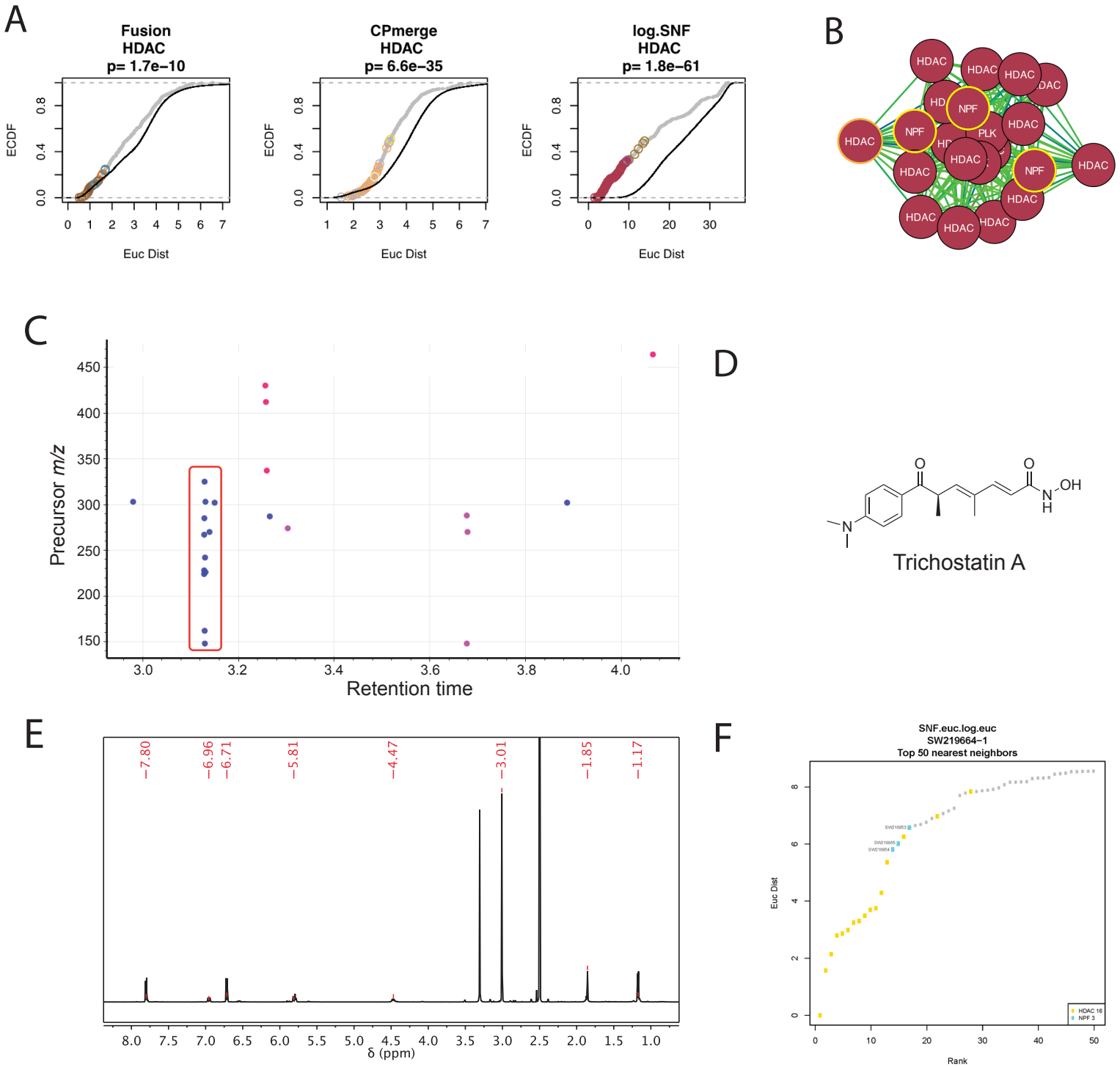
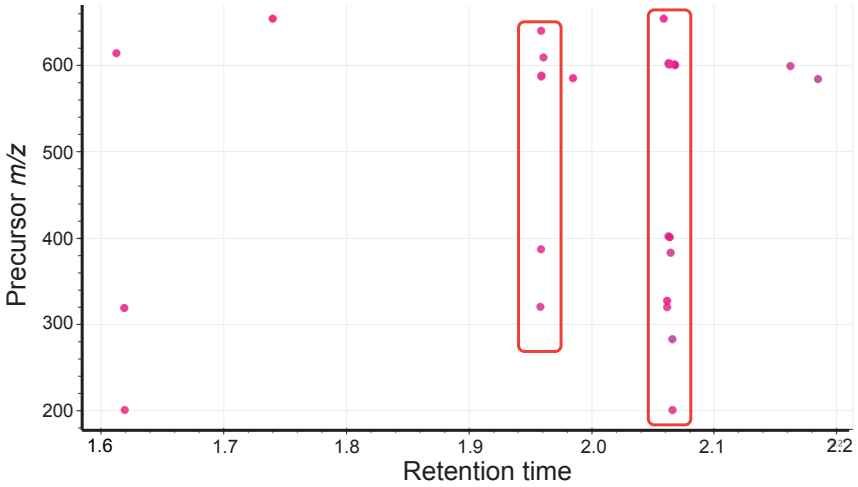
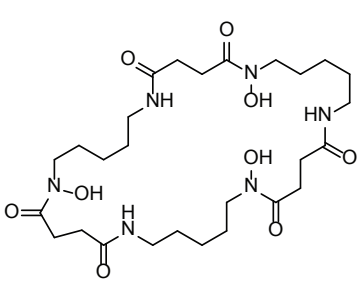

A
Fusion
HDAC
p= 1.7e−10
CPmerge
HDAC
p= 6.6e−35
log.SNF
HDAC
p= 1.8e−61
ECDF
Euc Dist
B
HDAC
NPF
PLK
C
Precursor m/z
Retention time
D
Trichostatin A
E
δ (ppm)
F
SNF.euc.log.euc
SW219664−1
Top 50 nearest neighbors
Euc Dist
Rank
HDAC 16
NPF 3

Figure 5

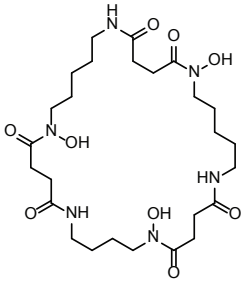
A



B

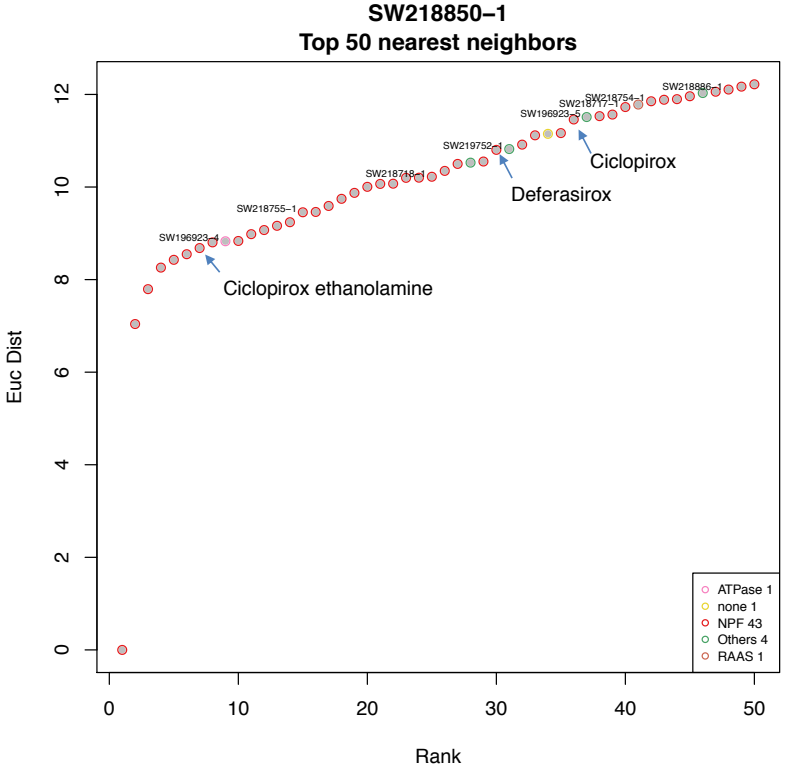
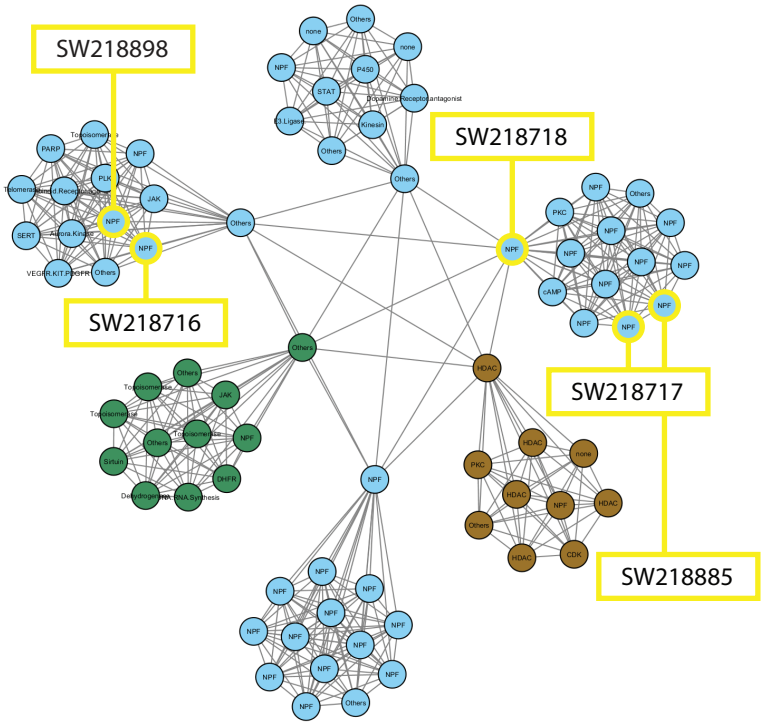


Desferrioxamine E



Desferrioxamine D2

C



D