DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–16
doi: 10.1093/database/baw100
Original article

Original article

The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins

Andrew D. Rouillard, Gregory W. Gundersen, Nicolas F. Fernandez, Zichen Wang, Caroline D. Monteiro, Michael G. McDermott and Avi Ma'ayan*

Department of Pharmacology and Systems Therapeutics, Department of Genetics and Genomic Sciences, BD2K-LINCS Data Coordination and Integration Center (DCIC), Mount Sinai's Knowledge Management Center for Illuminating the Druggable Genome (KMC-IDG), Icahn School of Medicine at Mount Sinai, New York, NY, USA

***Corresponding Author**: avi.maayan@mssm.edu

Citation details: Rouillard,A.D., Gundersen,G.W., Fernandez,N.F. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* (2016) Vol. 2016: article ID baw100; doi:10.1093/database/baw100

Received 13 February 2016; Revised 15 May 2016; Accepted 31 May 2016

Abstract

Genomics, epigenomics, transcriptomics, proteomics and metabolomics efforts rapidly generate a plethora of data on the activity and levels of biomolecules within mammalian cells. At the same time, curation projects that organize knowledge from the biomedical literature into online databases are expanding. Hence, there is a wealth of information about genes, proteins and their associations, with an urgent need for data integration to achieve better knowledge extraction and data reuse. For this purpose, we developed the Harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins from over 70 major online resources. We extracted, abstracted and organized data into ~72 million functional associations between genes/proteins and their attributes. Such attributes could be physical relationships with other biomolecules, expression in cell lines and tissues, genetic associations with knockout mouse or human phenotypes, or changes in expression after drug treatment. We stored these associations in a relational database along with rich metadata for the genes/proteins, their attributes and the original resources. The freely available Harmonizome web portal provides a graphical user interface, a web service and a mobile app for querying, browsing and downloading all of the collected data. To demonstrate the utility of the Harmonizome, we computed and visualized gene–gene and attribute–attribute similarity networks, and through unsupervised clustering, identified many unexpected relationships by combining pairs of datasets such as the association between kinase perturbations and disease signatures. We also applied supervised machine learning methods to predict novel

substrates for kinases, endogenous ligands for G-protein coupled receptors, mouse phenotypes for knockout genes, and classified unannotated transmembrane proteins for likelihood of being ion channels. The Harmonizome is a comprehensive resource of knowledge about genes and proteins, and as such, it enables researchers to discover novel relationships between biological entities, as well as form novel data-driven hypotheses for experimental validation.

Database URL: http://amp.pharm.mssm.edu/Harmonizome.

Introduction

Currently, biomolecular data are stored in many disjoint online databases. The data within these databases is structured, and thus suitable for data integration; however, most attempts to integrate knowledge from multiple resources have only succeeded in accomplishing this for a few resources. For example, web-based platforms such as BioGPS (1), NCBI's Entrez Gene Database (2), UniProt (3), GeneWeaver (4), MSigDB (5), GO-Elite (6) or Ingenuity Target Explorer, provide knowledge about genes from the Gene Ontology (GO) (7), protein domains, protein–protein interactions, expression in tissues, membership in pathways, and literature references but there are many other sources that these sites are missing. The knowledge that is commonly missing includes, e.g. gene-phenotype associations, putative regulation of genes by transcription factors, membership of proteins in complexes, putative regulation of genes by microRNAs, and changes in expression after drug treatment, or changes in expression in disease, or after single gene perturbations such as knockdown, knockout, mutation or overexpression. GeneCards (8) is becoming one of the most comprehensive resources for collective knowledge about genes and proteins, aggregating information from over 120 resources. However, GeneCards is a commercial product that does not provide the data through an open and free application programming interface (API). GeneCards is advertising commercial products such as antibodies, compounds, recombinant proteins, and gene sequencing services. This limits the utility of GeneCards for integrative knowledge discovery and pure data mining. Another leading resource is UniProt (9). UniProt focuses on sequence information and employs careful manually curated protein pages with less emphasis on data from omics resources. Other resources use text-mining strategies to collect information about genes and proteins. For example, resources such as WikiGenes (10), iHOP (11), Genes2Wordcloud (12) and EvidenceFinder (http://labs.europepmc.org/evf) identify and highlight genes and other semantic entities in sentences from abstracts and full-text publications to summarize gene and protein functions. These resources suffer from literature research focus biases (13); the uneven attention researchers give to well-studied genes and proteins (14).

One of the challenges related to integrating knowledge about genes and proteins is the standardization of data formats and harmonizing identifiers (15). Integration efforts made in subdomain areas such as protein–protein interactions have already developed successful solutions (16, 17). These solutions require some level of abstraction (15, 18, 19), i.e. ignoring quantitative details specific to a data resource (20, 21). Here we demonstrate that such an abstraction approach is feasible for integrating data about genes and proteins from many online resources. Using a simple schema, such data integration effort directly translates to a useful web service and a gateway to knowledge discovery with many applications (Figure S1D).

Results

Datasets and data resources

To create the Harmonizome, we collected information about human and mouse, genes and proteins, from 125 unique datasets (Tables 1–9, Supplementary Table S1) hosted by 72 open online resources (Supplementary Table S2). The collected datasets cover six broad categories of information about mammalian genes or proteins: (i) disease and phenotype associations, (ii) genomic profiles, (iii) physical interactions, (iv) proteomic profiles, (v) structural or functional annotations and (vi) transcriptomic profiles (Supplementary Figure S1A). The datasets provide evidence for associations between genes/proteins and biological entities spanning nine broad categories (Supplementary Table S3 and Supplementary Figure S1B), whereas the evidence types supporting the gene-entity associations span five broad categories (Supplementary Table S4 and Figure S1C). Half of the datasets are from high-throughput, data-driven studies, a third are from low-throughput, hypothesis-driven studies, and the remainder are from mixed sources.

To harmonize the 125 datasets we: (i) organized each incoming dataset into a matrix with genes labeling the

Table 1 Datasets. List of datasets group by attribute, with dataset citations

Dataset	Citations
Achilles Cell Line Gene Essentiality Profiles	(22–24)
BioGPS Cell Line Gene Expression Profiles	(1, 25, 26)
CCLE Cell Line Gene CNV Profiles	(27)
CCLE Cell Line Gene Expression Profiles	(27)
CCLE Cell Line Gene Mutation Profiles	(27)
COSMIC Cell Line Gene CNV Profiles	(28, 29)
COSMIC Cell Line Gene Mutation Profiles	(28, 29)
GDSC Cell Line Gene Expression Profiles	(30)
Heiser et al., PNAS, 2011 Cell Line Gene Expression Profiles	(31)
HPA Cell Line Gene Expression Profiles	(32)
Klijn et al., Nat. Biotechnol., 2015 Cell Line Gene CNV Profiles	(33)
Klijn et al., Nat. Biotechnol., 2015 Cell Line Gene Expression Profiles	(33)
Klijn et al., Nat. Biotechnol., 2015 Cell Line Gene Mutation Profiles	(33)
BioGPS Human Cell Type and Tissue Gene Expression Profiles	(1, 25, 26)
BioGPS Mouse Cell Type and Tissue Gene Expression Profiles	(1, 25, 26)
HPM Cell Type and Tissue Protein Expression Profiles	(34)
ProteomicsDB Cell Type and Tissue Protein Expression Profiles	(35)
Roadmap Epigenomics Cell and Tissue DNA Methylation Profiles	(36, 37)
Roadmap Epigenomics Cell and Tissue Gene Expression Profiles	(36, 37)
Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by Microarray	(38–40)
Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by RNA-seq	(38–40)
GTEx Tissue Sample Gene Expression Profiles	(41, 42)
HPA Tissue Sample Gene Expression Profiles	(32)
TCGA Signatures of DEGs for Tumors	(43)
Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles	(37–40, 44)
Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles	(38, 39, 45)
Allen Brain Atlas Prenatal Human Brain Tissue Gene Expression Profiles	(38–40, 46)
GTEx Tissue Gene Expression Profiles	(41, 42)
HPA Tissue Gene Expression Profiles	(32)
HPA Tissue Protein Expression Profiles	(32)
TISSUES Curated Tissue Protein Expression Evidence Scores	(47)
TISSUES Experimental Tissue Protein Expression Evidence Scores	(47)
TISSUES Text-mining Tissue Protein Expression Evidence Scores	(47)

List of datasets group by attribute, with dataset citations. Datasets providing evidence for associations between genes and 'cell lines, cell types or tissues'.

Table 2 Datasets providing evidence for associations between genes and 'chemicals'

Dataset	Citations
CTD Gene-Chemical Interactions	(48, 49)
SILAC Phosphoproteomics Signatures of Differentially Phosphorylated Proteins for Drugs	
DrugBank Drug Targets	(50, 51)
Guide to Pharmacology Chemical Ligands of Receptors	(52)
HMDB Metabolites of Enzymes	(53, 54)
CMAP Signatures of DEGs for Small Molecules	(55)
GEO Signatures of DEGs for Small Molecules	(56–58)
LINCS L1000 CMAP Signatures of DEGs for Small Molecules	(59)
KinomeScan Kinase Inhibitor Targets	

Table 3 Datasets providing evidence for associations between genes and 'diseases, phenotypes or traits'

Dataset	Citations
GEO Signatures of DEGs for Diseases	(56, 57)
CTD Gene-Disease Associations	(48, 49)
DISEASES Curated Gene-Disease Assocation Evidence Scores	(61)
DISEASES Experimental Gene-Disease Assocation Evidence Scores	(60)
DISEASES Text-mining Gene-Disease Assocation Evidence Scores	(60)
GAD Gene-Disease Associations	(61)
GAD High Level Gene-Disease Associations	(61)
GWASdb SNP-Disease Associations	(62)
PhosphoSitePlus Phosphosite-Disease Associations	(63, 64)
ClinVar SNP-Phenotype Associations	(65)
GWAS Catalog SNP-Phenotype Associations	(66)
GWASdb SNP-Phenotype Associations	(62)
HPO Gene-Disease Associations	(67)
HuGE Navigator Gene-Phenotype Associations	(68)
MPO Gene-Phenotype Associations	(69–72)
OMIM Gene-Disease Associations	(73, 74)
dbGAP Gene-Trait Associations	(75, 76)

rows and biological entities (attributes) labeling the columns; and (ii) standardized identifiers for genes and biological entities; while also (iii) calculated standardized scores for gene-biological entity associations; and (iv) computed gene–gene and entity–entity similarity matrices. These matrices were then: (v) saved to text files and (vi) loaded into a relational database. To manage gene or protein identifiers, we mapped them all to NCBI Entrez Gene Symbols. To consolidate biological entity identifiers we mapped these to existing ontologies for tissues, cell lines, chemicals, functional terms, phenotypes and diseases (Supplementary Table S5). To serve the data in useful formats, we provide

Table 4 Datasets providing evidence for associations between genes and 'functional terms, phrases or references'

Dataset	Citations
GO Biological Process Annotations	(7, 77)
GeneRIF Biological Term Annotations	(78)
Phosphosite Textmining Biological Term Annotations	
COMPARTMENTS Curated Protein Localization Evidence Scores	(79)
COMPARTMENTS Experimental Protein Localization Evidence Scores	(79)
COMPARTMENTS Text-mining Protein Localization Evidence Scores	(79)
GO Cellular Component Annotations	(7, 77)
LOCATE Curated Protein Localization Annotations	(80)
LOCATE Predicted Protein Localization Annotations	(80)
GO Molecular Function Annotations	(7, 77)
Biocarta Pathways	
HumanCyc Pathways	(81, 82)
KEGG Pathways	(83, 84)
PANTHER Pathways	(85, 86)
PID Pathways	(87)
Reactome Pathways	(88, 89)
Wikipathways Pathways	(90)
CORUM Protein Complexes	(91, 92)
NURSA Protein Complexes	(93, 94)
ESCAPE Omics Signatures of Genes and Proteins for Stem Cells	(95)
GeneSigDB Published Gene Signatures	(96, 97)

Table 5 Datasets providing evidence for associations between genes and 'other genes, proteins or microRNAs'

Dataset	Citations
MSigDB Cancer Gene Co-expression Modules	(98)
GEO Signatures of DEGs for Gene Perturbations	(56, 57)
LINCS L1000 CMAP Signatures of DEGs for Gene Knockdowns	(59)
MSigDB Signatures of DEGs for Cancer Gene Perturbations	(98)
SILAC Phosphoproteomics Signatures of Differentially Phosphorylated Proteins for Gene Perturbations	
Hub Proteins Protein–Protein Interactions	(99)
BIND Biomolecular Interactions	(100, 101)
BioGRID Protein–Protein Interactions	(102, 103)
DIP Protein–Protein Interactions	(104)
HPRD Protein–Protein Interactions	(105, 106)
IntAct Biomolecular Interactions	(107, 108)
NURSA Protein–Protein Interactions	(93, 94)
Pathway Commons Protein–Protein Interactions	(109)
GEO Signatures of DEGs for Kinase Perturbations	(56, 57)
KEA Substrates of Kinases	(110)
PhosphoSitePlus Substrates of Kinases	(63, 64)
SILAC Phosphoproteomics Signatures of Differentially Phosphorylated Proteins for Protein Ligands	
Guide to Pharmacology Protein Ligands of Receptors	(52)
MiRTarBase microRNA Targets	(111, 112)
TargetScan Predicted Conserved microRNA Targets	(113–115)
TargetScan Predicted Nonconserved microRNA Targets	(113–115)
DEPOD Substrates of Phosphatases	(116)
GEO Signatures of DEGs for Transcription Factor Perturbations	(56, 57)
CHEA Transcription Factor Targets	(117)
ENCODE Transcription Factor Targets	(118, 119)
JASPAR Predicted Transcription Factor Targets	(120, 121)
TRANSFAC Curated Transcription Factor Targets	(122, 123)
TRANSFAC Predicted Transcription Factor Targets	(122, 123)
Virus MINT Protein-Viral Protein Interactions	(124)

Table 6 Datasets providing evidence for associations between genes and 'molecular profiles'

Dataset	Citations
Kinativ Kinase Inhibitor Bioactivity Profiles	
ENCODE Histone Modification Site Profiles	(118, 119)
Roadmap Epigenomics Histone Modification Site Profiles	(36, 37)
CHEA Transcription Factor Binding Site Profiles	(117)
ENCODE Transcription Factor Binding Site Profiles	(118, 119)

all gene–entity-value triplets for download as text files in matrix, gene-set library, biological entity-set library and bipartite graph formats. In addition, gene–gene and entity–entity similarity networks for each dataset are also available.

The harmonizome web resource

To accommodate users who seek information about a single gene, as well as computational biologists who can programmatically operate on the data, the Harmonizome includes advanced search functionality, and serves the data in text file and JSON formats through an API. The Harmonizome landing page displays a search bar where users can type in any search term with autocomplete capabilities (Supplementary Figure S2A). The engine searches for matching datasets, genes and attributes. On the search results pages users can choose to view datasets, genes or attributes pages (Supplementary Figure S2B). These pages contain metadata and provide various views. The Harmonizome site also has a global summary visualization of the knowledge about each gene across all of the datasets. This interactive heat map, called the Harmonogram, displays the genes as the rows and the datasets as the columns. The intensity of each square on the Harmonogram indicates the relative number of functional associations that each gene has in each dataset (Supplementary Figure S2C). This visualization reveals gaps in knowledge about genes, and suggests where to

focus future experiments to illuminate functions of unannotated genes to increase potential for novel discoveries.

For further visual exploration of the data, the Harmonizome includes interactive heat maps of hierarchically clustered: (i) datasets (gene–biological entity relationships matrices), (ii) gene–gene similarity matrices, (iii) entity–entity similarity matrices and (iv) dataset pairs (matrices comparing biological entities from one dataset to biological entities from another dataset based on similarity of their gene associations).

Hierarchically clustered data matrices in the Harmonizome collection can uncover new knowledge. For example, we organized phenotype data from the Mammalian Phenotype Ontology (MPO) (129) into a binary matrix with genes labeling the rows, phenotypes labeling the columns, and matrix elements set equal to 1 to indicate which phenotypes were observed following knockout of a gene. Hierarchical clustering of this matrix shows patches of common phenotypes for groups of genes (Figure 1A). By exploring the clustered heat map visualization of the MPO dataset, we noticed a small group of genes (NCOR1, BAG3, SIRT7, STEAP4, CXCL14, CEBPA, PROX1, AGPAT2, BSCL2, LIPA, NR1H4 and PPARG) that are associated with abnormalities of both the immune system and metabolism, such as glucose homeostasis, lipid homeostasis and feeding behavior (Figure 1B). Interactive hierarchical clustering plots with zooming and panning capabilities are available on the Harmonizome site, enabling further exploration of this type of clustering analysis.

Table 7 Datasets providing evidence for associations between genes and 'organisms'

Dataset	Citations
GEO Signatures of DEGs for Viral Infections	(56, 57)
Virus MINT Protein-Virus Interactions	(124)

Table 8. Datasets providing evidence for associations between genes and 'sequence features'

Dataset	Citations
GTEx eQTL	(41, 42)

Table 9. Datasets providing evidence for associations between genes and 'structural features'

Dataset	Citations
InterPro Predicted Protein Domain Annotations	(125–128)

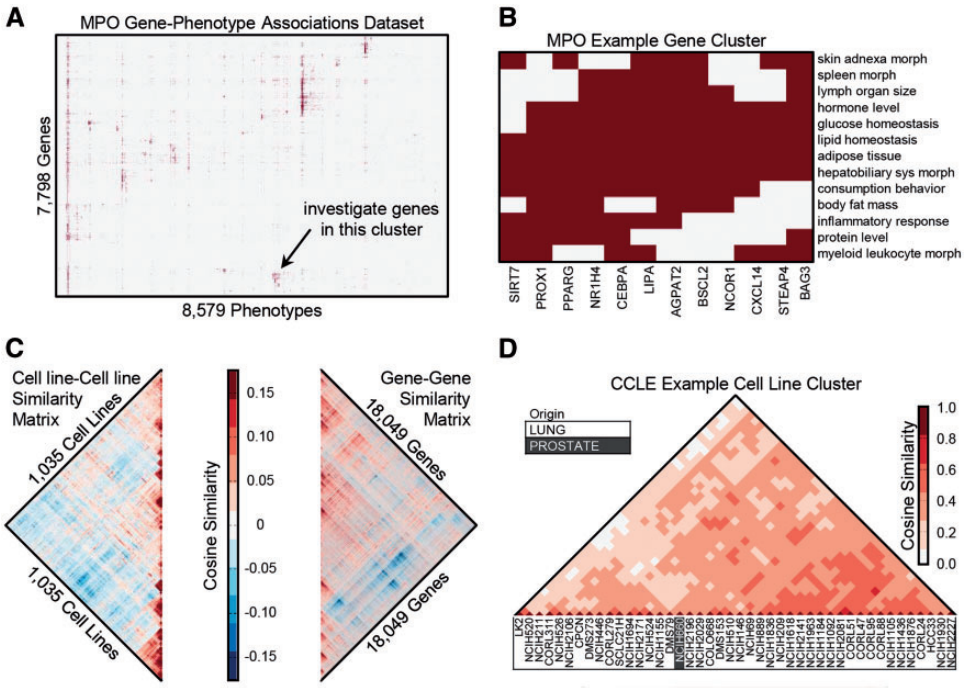

Figure 1. Hierarchical clustering of gene-term, term-term and gene-gene matrices. **(A)** Gene-phenotype associations from the MPO organized into a binary matrix and clustered using hierarchical clustering. **(B)** Zooming into a cluster of genes with similar associated phenotypes, filtered to show higher level phenotypes associated with at least half of the genes in the cluster but no > 10% of all genes. **(C)** The gene–gene and cell-line/cell-line similarity matrices are from the CCLE gene expression dataset. Along the main diagonal of both matrices, there are several distinct zones of high red intensity, indicating clusters of cell lines with similar differentially expressed genes (DEGs) and clusters of genes with similar patterns of expression across cell lines. **(D)** Zooming into the lung cancer cell-lines cluster.

Hierarchically clustered functional association networks (130) can also be explored for each dataset. We derived gene–gene and entity–entity functional association networks by computing the cosine similarity of the rows and columns of each dataset, respectively. In the cancer cell-line encyclopedia (CCLE) dataset, as an example, we can observe correlated gene expression modules and groups of cell lines (Fig. 1C). The cell lines from CCLE predominantly cluster by tissue of origin. However, in a few interesting instances, some cell lines are in clusters of a different tissue; e.g. NCI-H660 is marked as prostate tissue, but appears within a cluster of 43 lung cancer cell-lines (Figure 1D). The ATCC website states that NCI-H660 was originally a small-cell lung carcinoma cell-line, but this cell line was later reclassified to extra-pulmonary lymph node metastatic cancer originating from the prostate (131–133). The cell-line similarity heat map strongly supports a lung origin/phenotype. Interactive gene-gene and attribute–attribute functional association networks with zooming and panning capabilities are available on the Harmonizome site, potentially uncovering many other unexpected relationships.

Users of the Harmonizome can combine two or more datasets to identify relationships that are only possible to uncover once these datasets have been abstracted, normalized, organized and combined. We devised two related case studies to demonstrate this concept. For the first case study, we integrated differentially expressed gene (DEG) signatures for kinase perturbations with DEG signatures for diseases. The similarity scores for 233 disease signatures paired with 285 kinase perturbation signatures mostly did not match; however, we observed clear patches of positive and negative correlations (Figure 2A). The positive correlations (red patches) suggest that the kinase, or its pathway, is likely perturbed in the disease. The negative correlations (blue patches) suggest diseases in which down-regulating the kinase may reverse expression toward the normal tissue expression and promote a more favorable phenotype. Hence, these kinases are potential drug targets for the specific disease. To confirm this conjecture, we found that some of the similarity scores were predictive of kinase-disease associations obtained from genome-wide association studies (GWAS) and other genetic association datasets in the Harmonizome (Figure 2B). Finally, we integrated knowledge about small molecules that inhibit kinases by combining the kinase-disease similarity network with the LINCS KinomeScan dataset to create a tri-partite graph connecting small molecules to kinases to diseases as potential therapeutics (Figure 2C).

In the second case study, we performed a similar analysis, but here we replaced the disease signatures with signatures of DEGs for cancer cell lines from CCLE to derive similarity scores for 1037 cancer cell lines paired with 285 kinase perturbations (Figure 2D). These similarity scores were predictive of driver gene mutations in the cancer cell lines as reported by the COSMIC resource (28) (Figure 2E). Finally, we integrated the LINCS KinomeScan dataset to create a tri-partite graph connecting cancer cell lines to likely driver kinases to kinase-inhibitor compounds (Figure 2F). Experimental methods can assess whether some of these compounds selectively influence the phenotype of these cells. Integration with the recently published cancer cell-line sensitivity data is an alternative (27,30, 31,134). Indeed, some of the predicted small molecules have already been tested and shown to have favorable effects on the cancer cell lines and diseases suggested by our analysis. For example, sorafenib has shown promise for the treatment of colorectal cancer (135); dinaciclib for the treatment of malignant gliomas (136); and bosutinib for melanoma (137), prostate cancer (138) and pancreatic cancer (139). These confirmations suggest that some of our predictions are correct, and some can serve as a global reference point for further analyses to provide other rational and novel hypotheses for experimental validation. These case studies illustrate just two of many ways to combine the Harmonizome datasets for discovery and hypothesis generation. The Harmonizome website provides the ability to explore similar relationships between pairs of datasets by performing unsupervised hierarchical clustering of similarity matrices comparing biological entities between datasets.

The harmonizome mobile app

The Harmonizome mobile application serves the biological knowledge we collected in an easy-to-access interface where a user can enter a gene of interest to discover properties and functions for the gene (Supplementary Figure S3). Developed using the Facebook React Native platform, the Harmonizome mobile app serves knowledge about genes organized into eight categories, and provides links to external sources for further exploration of gene-function associations. The Harmonizome mobile application is free and available at the Google Play Store (http://goo.gl/JWlI8H) for Android devices, and the App Store (http://appstore.com/harmonizome) for iOS devices. A demonstration video with a case study is available on YouTube at: https://youtu.be/dkYcD51pnfY.

Machine learning case studies

On its own, the Harmonizome web resource is a valuable tool for discovery and hypothesis generation by enabling exploration of functional associations between mammalian genes and diseases/phenotypes, tissues and other biological

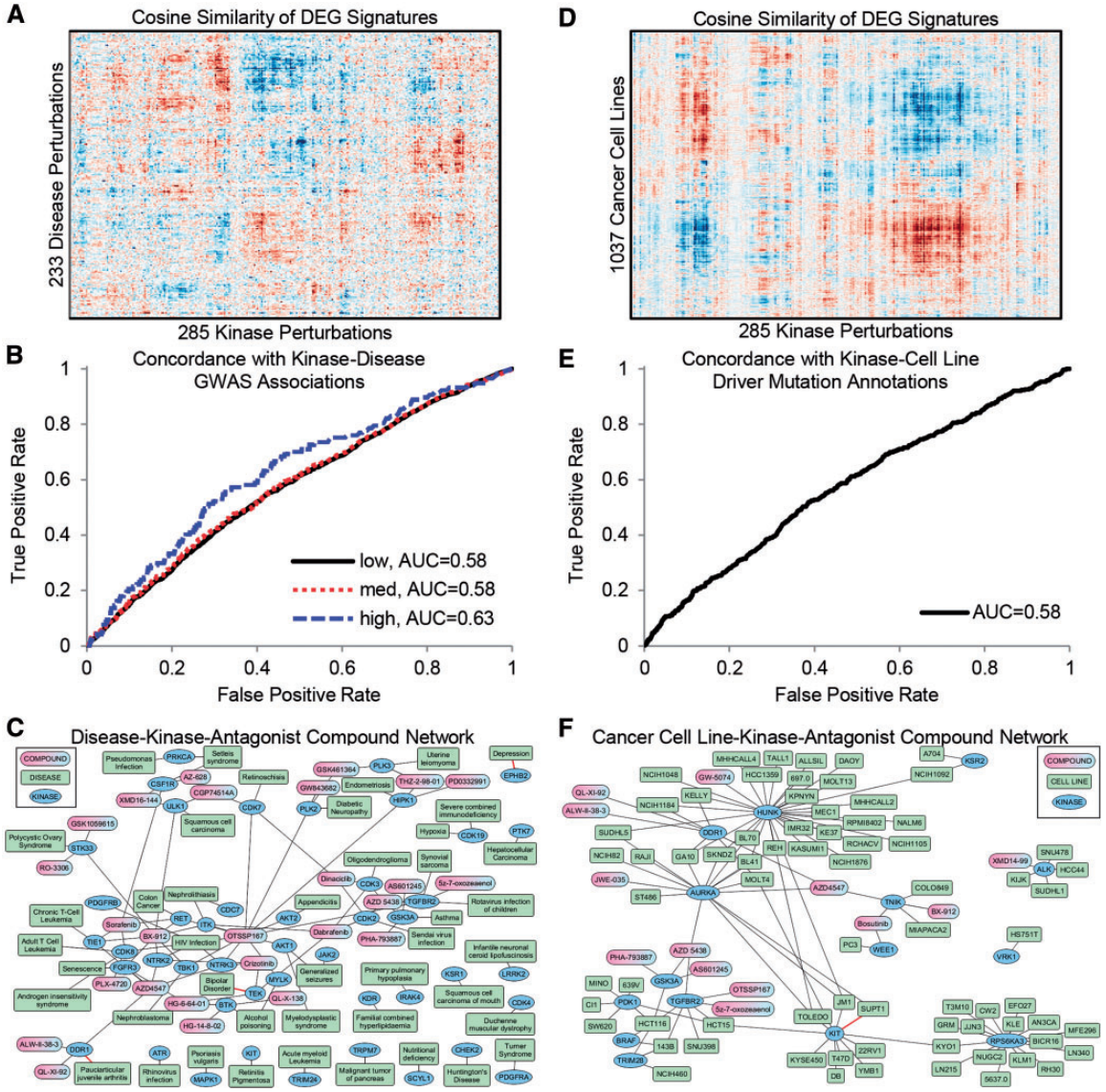

Figure 2. Example of combining datasets: matching kinases with diseases and drugs. **(A)** Hierarchical clustering of kinase perturbation signatures extracted from GEO and disease signatures extracted from GEO. **(B)** Validation of kinase-disease associations with genomics datasets. ROC curve showing concordance of kinase-disease associations derived by comparing gene expression profiles and kinase-disease associations collected from GWAS and other genetic association datasets. Low, medium and high labels correspond to confidence levels of associations from GWAS datasets. **(C)** Network showing top predictions of drug-kinase-disease associations. Red edges indicate kinase-disease associations that have supporting GWAS evidence. **(D)** Hierarchical clustering of signatures of DEGs for kinase perturbations extracted from GEO compared with signatures for cancer cell lines from CCLE. **(E)** ROC curve showing concordance of kinase-cell line associations derived by comparing gene expression profiles and driver kinase mutations for cell lines from COSMIC. **(F)** Network showing top predictions of drug-kinase-cell line associations. Red edges indicate kinase-cell line associations supported by COSMIC as having a driver mutation in the cell line.

entities collected from over a hundred diverse datasets. However, there is also the opportunity for discovering new knowledge about mammalian genes and proteins by the 'guilt-by-association' concept, i.e. genes and proteins that share some common functional properties are likely to share more of those properties. To demonstrate this concept we utilized the Harmonizome data for developing four predictive models using Machine Learning. These case studies demonstrate how to use the Harmonizome data for predicting novel properties for genes and proteins.

Predicting ion channels from uncharacterized transmembrane proteins

Discovery of novel ion channels could open new lines of research and reveal potential drug targets (140). Ion channels have diverse structures and this makes it challenging to discover ion channels based on sequence information alone. For example, ion channels vary in their number of transmembrane domains and are commonly part of macromolecular complexes (141). Searching gene or protein sequences for transmembrane domains is useful for

predicting proteins that are located in the plasma membrane, but channel activity is much more difficult to predict computationally from sequence alone. Roughly 5500 genes have been predicted to give rise to transmembrane proteins (142). We can use the omics data within the Harmonizome to construct a Machine Learning classifier to predict if any uncharacterized transmembrane proteins are likely ion channels. Our overall modeling approach can be broken down into three stages: (i) gene and dataset selection; (ii) dimensionality reduction and feature selection; and (iii) model training, cross-validation and finally making predictions.

We began with 5428 human genes predicted by Fagerberg *et al.* (142) to encode for transmembrane proteins. We next divided these genes into three classes: 341 known ion channels, 4510 non-ion channels and 577 uncharacterized genes. Next, we selected datasets from the Harmonizome to obtain attributes for the ion channel classifier. We considered only omics datasets, ranked each dataset by the predictive value of its attributes, and retained a final set of 8 datasets covering 320 ion channels (94%), 3928 non-ion channels (87%) and 396 uncharacterized transmembrane genes/proteins (69%).

From each of those eight datasets we next selected the best attributes as predictors for training the ion channel classifier. For this we performed principal component analysis on each of the selected datasets, retained the principal components needed to capture 99% of the variance of each dataset, and then concatenated the principal components from all datasets into a single matrix. This process yielded 6985 total predictors. We performed receiver operating characteristic (ROC) analysis to rank the value of each predictor for discriminating between ion channels and non-ion channels. We used the Breiman Random Forest algorithm with decision trees to train ion channel classifiers and found that 70 features and 300 trees were sufficient to achieve near minimal out-of-bag error. The final set of 70 features contained contributions from all eight datasets, with the majority of the features coming from the InterPro structural domains dataset (Supplementary Table S6).

The area under the ROC curve of the final classifier was 0.99 (Figure 3A). The F1 score and Matthew's Correlation Coefficients (MCC) had maximum values of 0.922 and 0.918 (Figure 3B, Supplementary Figure S4 and Supplementary Table S7). These performance statistics, calculated from the out-of-bag data, estimate how well the classifier generalizes to data not seen while training the model. We used the model to predict and rank ion channel probabilities for the 396 uncharacterized genes (Supplementary Table S8). To provide context for these predictions, we computed a network connecting each predicted ion channel to its most similar known ion channels (Figure 3C). In summary, we can determine with high confidence the molecular function of uncharacterized transmembrane genes/proteins that are likely ion channels and have the potential to become drug targets. The first step for experimentally validating such predictions is to express these genes in artificial systems that can test channel activity.

Predicting mouse phenotypes for single gene knockouts

The Mouse Phenotype Ontology (129) currently contains phenotype data for ~7000 single gene knockouts in mice. Knockout phenotype data are valuable for generating hypotheses about the function, tissue specificity and disease relevance of mammalian genes. The International Mouse Phenotyping Consortium is working toward systematically phenotyping single gene knockouts for the remainder of the genome (143). This is an expensive and time-consuming effort projected to complete in 2021.

In a similar way as described earlier for ion channels, we used omics datasets from the Harmonizome to build a model to predict phenotypes for single gene mouse knockouts. Instead of training a single model to predict a single gene label, i.e. an ion channel, we trained many models to predict many labels (2666 phenotypes). Observed phenotypes of mice harboring single gene knockout mutations obtained from the Mouse Genome Database (71, 129) were the positive training examples, while single gene knockouts with unobserved phenotypes were the negative training examples. The area under the ROC curve of the phenotype classifier was 0.88 (Figure 3A). The F1-score and MCC had maximum values of 0.24 and 0.23 (Figure 3B, Supplementary Figure S5 and Supplementary Table S7). We used the model to predict phenotypes for 7934 single gene mouse knockouts (Supplementary Table S9), and created a gene-phenotype network to visualize a subset of the top predictions (Figure 3D). Our computational predictions of phenotypes for single gene knockouts can assist in prioritizing genes for experimental phenotyping. Furthermore, such predictions, if combined with mouse models of disease, have the potential to identify novel drug-target candidates.

Predicting endogenous ligands for G protein-coupled receptors

G-protein-coupled-receptors (GPCRs) are important biologically and pharmacologically due to their roles as sensors and signal transducers (144). GPCRs are the most successful protein family currently serving as targets for drugs; yet most research efforts have focused on relatively few GPCRs (145). At present, there are over 140 orphan GPCRs, which are GPCRs with no known ligand (146). So far,

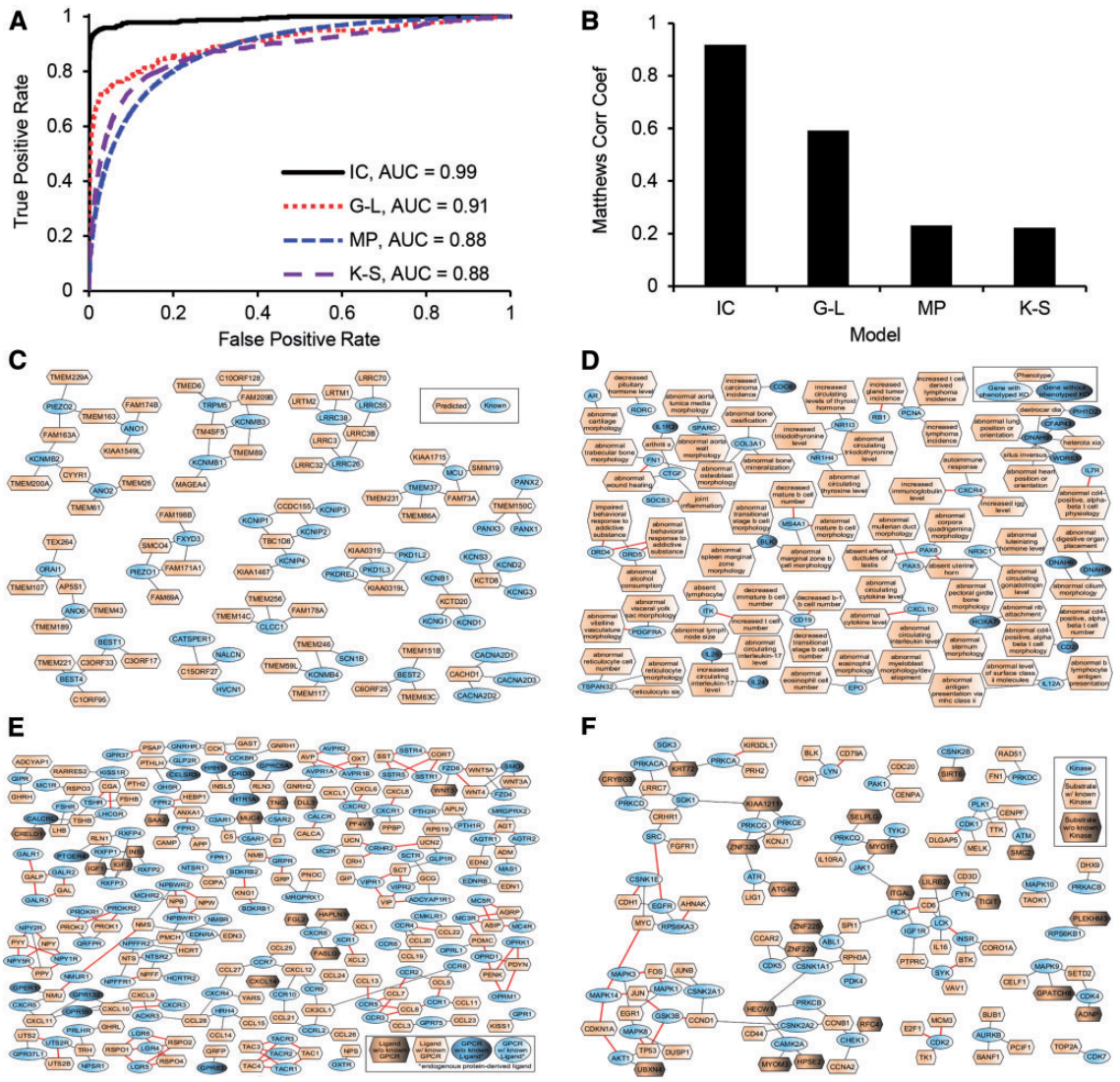

Figure 3. Example of supervised machine learning: classifiers to predict ion channels (IC), phenotypes of single gene knockouts in mice (MP), ligands of GPCRs (G-L), and substrates of kinases (K-S). **(A)** ROC curve of the classifiers. **(B)** MCC as a function of the fraction of correct predictions. **(C)** Network showing candidate ion channels, predicted at a false discovery rate (FDR) of 0.67, connected to their most similar known ion channels, and limited to no more than three edges per node. **(D)** Network showing candidate gene-phenotype associations, predicted at a FDR of 0.33, limited to no more than three edges per node, and trimmed to remove clusters with all edges supported by prior knowledge. Red edges indicate known associations. **(E)** Network showing candidate GPCR-ligand interactions; predicted at a FDR of 0.67 and limited to no more than three edges per node. Red edges indicate known interactions. **(F)** Network showing candidate kinase-substrate interactions predicted at a FDR of 0.67 and limited to no more than three edges per node. Red edges indicate known interactions.

most computational approaches have attempted to predict ligands for GPCRs using structure-based methods. As a complementary method, we used datasets from the Harmonizome to build a classifier to predict protein ligands for GPCRs, although we are aware that GPCRs can bind non-protein ligands. First, we extracted known GPCR-ligand interactions from the Guide to Pharmacology (52). This allowed us to assign GPCR-candidate ligand pairs to positive, negative, or unknown classes for model training and predictions. Using the same procedure as described above for ion channels, the area under the ROC curve of the GPCR-ligand interaction classifier was 0.91 (Figure 3A). The F1-score and MCC had maximum values of 0.59 (Figure 3B, Supplementary Figure S6 and Supplementary Table S7). We used the model to classify 368 953 GPCR-ligand pairs involving either a GPCR with no known endogenous protein ligand, or a candidate ligand with no known GPCR interaction (Supplementary Table S10). Finally, we created a GPCR-ligand network to visualize a subset of the top predictions (Figure 3E). The discovery of endogenous ligands for these GPCRs could open new lines of biological and pharmacological research. Methods that screen ligands for GPCRs rapidly emerge (147, 148) and these predictions can inform such efforts.

Predicting substrates of kinases

Protein kinases are well-studied enzymes that regulate almost all cellular processes by reversible phosphorylation of their substrates (149, 150). Kinases are also a promising family of drug targets. While phosphoproteomics studies have revealed many phosphorylation sites, and the human kinome is highly annotated, our knowledge of kinase-substrate interactions remains vastly incomplete. For developing the PhosphoSitePlus database, investigators from Cell Signaling Inc. (151) aggregated information about phosphorylation sites on proteins from low-throughput published studies in the literature, and high-throughput mass spectrometry studies, finding ~108 000 phosphorylation sites on 12 500 human proteins. We also aggregated information about kinase-substrate phosphorylation reactions from few databases and found about 3500 human proteins with at least one known kinase that phosphorylates them (110). This leaves thousands of proteins with at least one phosphorylation site but with no known upstream regulatory kinase.

To attempt filling this knowledge gap, we used the Harmonizome data collection to build a classifier to predict substrates for kinases. We began with 8293 human proteins with reported phosphorylation sites. We used the kinase enrichment analysis (KEA) dataset (110) to divide these proteins into two classes: 3552 substrates with a known upstream kinase, and 4741 substrates with unknown upstream kinase. Next, we selected datasets from the Harmonizome to build the classifier. We initially considered 34 datasets that cover at least 95% of the substrates. After an initial dataset selection process, we ultimately left with a final set of 12 datasets covering 3,363 substrates with at least one known kinase (95%), and 4270 substrates with unknown kinase (90%). We then performed feature selection using principal component analysis, retaining features that capture 99% of the variance of each dataset. This analysis reduced the number of features to 75. Using a similar scheme as described earlier, we predicted novel kinase–substrate interactions between kinases and substrates with no known kinase. Known kinase–substrate interactions from KEA (110) were used to define positive and negative classes for training the model. The area under the ROC curve of the kinase-substrate interaction classifier was 0.88 (Figure 3A). The F1-score and MCC had maximum values of 0.23 and 0.22 (Figure 3B, Supplementary Figure S7 and Supplementary Table S7). We used the model to classify 2 993 096 potential kinase-substrate pairs involving either a kinase with no known substrate, or a candidate substrate with no known regulatory kinase (Supplementary Table S11). Finally, we created a kinase-substrate network to visualize a subset of the top predictions (Figure 3F). The prediction of kinase-substrate associations is still missing the site of the phosphorylation, the functional effect of the phosphorylation, and the context of the phosphorylation. However, it provides a reliable mapping at a more abstract level, and a resource that can direct experimental testing towards detailed direction of discovery. It can also assist in the reconstruction of the human kinome network, i.e. how kinases regulate each other.

Discussion

To create the Harmonizome resource, we had to make many decisions in regards to cutoffs for significance of differential expression analysis, data normalization methods, similarity measures between genes and terms, merging IDs for genes and proteins, and combining IDs across mammalian organisms. In addition, in many cases we had to ignore details such as the location of a single nucleotide polymorphism (SNP), the location of a binding site in proximity of a coding region, location of phosphorylation sites on a protein, physical interactions between proteins in a complex and more. This form of data abstraction was necessary for data integration (20, 21). To impute knowledge from observed functional associations between genes and their attributes, we constructed Random Forest classifiers for four supervised Machine Learning tasks. We chose the Random Forest classifier because it is nonlinear, nonparametric, regularized and simple to train (152, 153). To achieve better performance, we could have trained an ensemble of different high-performing classifiers. Furthermore, the performance of the classifiers can be improved in many ways, e.g. by using a multivariate feature selection method. Another limitation of our initial approach may be that the negative class for the training examples was not always purely negative. For example, to predict substrates of kinases, ideally, we would benefit from negative class examples. In practice, the negative class consisted of proteins where it is unknown experimentally whether the kinase phosphorylates the substrate. Regardless of these potential limitations, we believe that our predictions represent a set of credible data-driven hypotheses suitable for experimental validation.

So far, we have noticed that the Harmonizome web service has been highly accessed. Form October 2015 to May 2016, over 33 000 unique users accessed the site. In the near future, we plan to add complex querying capabilities, on the fly Machine Learning, and communities of users centered on a gene or a dataset of interest. In addition, we can organize and serve knowledge about drugs and small molecules in a similar way. Another feature that we plan to implement is providing suggestions for similar genes or drugs to those currently displayed. We plan to continually maintain and expand the Harmonizome while keeping it a free and open resource.

Methods

Data processing

We extracted gene- entity-value triplets from each dataset and stored these data in matrices with genes labeling the rows and biological entities labeling the columns. The values in these matrices are discrete or continuous, depending on the data source. We standardized continuous-valued datasets to create more harmonized datasets. Our strategy was to standardize each continuous-valued dataset to have values ranging from 0 to 1, or −1 to 1, where 1 indicates strong positive gene-entity association, –1 indicates strong negative gene-entity association and 0 indicates no observed gene-entity association. Negative values applied to datasets where it was appropriate to convey signed information, e.g. up-regulation and down-regulation for gene expression datasets. To implement this strategy, for each continuous-valued dataset, we converted the values to empirical cumulative probabilities, which transformed the values to range from 0 to 1. If the median values for the genes were different, we computed the probabilities gene-by-gene, otherwise we computed the probabilities on all of the data at once. When appropriate to convey sign information, we doubled the probabilities and subtracted unity, which transformed the values to range from −1 to 1. After creating the standardized datasets, we created binary or tertiary datasets by applying a threshold to retain only 10% of the strongest gene-biological entity associations.

The processing steps to convert each data matrix to a binary or tertiary matrix depends on the data type and processing steps already taken by the original data provider. Any of the following operations may have been part of a data processing pipeline: filtering rows or columns, averaging rows or columns, imputation, transformation/scaling and quantile normalization. Each dataset page on the Harmonizome website provides a script documenting the processing steps used for each dataset. These scripts are also available on GitHub.

Identifier mapping

For each dataset, we mapped gene or protein identifiers to NCBI Entrez Gene Symbols and Gene IDs for human genes. Overall, we encountered six types of identifiers: NCBI Entrez Gene IDs, gene symbols, Ensembl Gene IDs, UniProt Accessions, genomic coordinates given as nucleotide position(s) on a chromosome and microarray Probeset IDs. We utilized ID mapping tables maintained by NCBI Entrez Gene, Ensembl, UniProt, Hugo Gene Nomenclature Committee (HGNC), Mouse Genome Informatics (MGI) and the Gene Expression Omnibus (GEO) to convert identifiers to NCBI Entrez Gene Symbols and Gene IDs. We used the mapping table maintained by NCBI Homologene to convert mouse NCBI Entrez Gene IDs to human Entrez Gene IDs.

Specifically, for gene symbols, we obtained lists of retired or synonymous gene symbols for NCBI Entrez Gene IDs from NCBI Entrez Gene, HGNC and MGI. From these lists, we created a table mapping gene symbols to NCBI Entrez Gene IDs and official Gene Symbols. We then filtered the table, removing symbols that mapped to more than one NCBI Entrez Gene ID and removing symbols that were identical to official Gene Symbols. For Ensembl Gene IDs, we downloaded tables from Ensembl mapping Ensembl Gene IDs to NCBI Entrez Gene IDs for human and mouse genes. For UniProt Accessions, we downloaded tables from UniProt mapping UniProt Accessions to NCBI Entrez Gene IDs for human and mouse proteins. For genomic coordinates, we downloaded tables from Ensembl listing the chromosome, gene start position, gene end position and transcription start site of each Ensembl Gene ID for human and mouse genes. We joined these tables with the previously described tables mapping Ensembl Gene IDs to NCBI Entrez Gene IDs to derive a table mapping genomic coordinates to NCBI Entrez Gene IDs. For microarray Probeset IDs, we downloaded the platform annotation tables from GEO, mapping Probeset IDs to gene symbols, NCBI Entrez Gene IDs, or Ensembl Gene IDs. We joined these tables with the mapping tables described above to derive tables mapping Probeset IDs to NCBI Entrez Gene IDs. We discarded data for unconverted gene or protein identifiers. We documented the original identifier, number of identifiers and fraction of unmapped identifiers (Supplementary Table S12). The median fraction of unmapped identifiers was 3%. Many of the unmapped identifiers correspond to predicted genes and other forms of untranslated to protein non-coding genes.

We mapped labels for tissues, cell lines, chemicals, functional terms, phenotypes and diseases to terms in relevant ontologies and dictionaries, which we refer to as naming authorities. If we matched a label to a term or synonym from one naming authority, we linked the original label to that term and its metadata including name, description, identifier and persistent URL. Otherwise, we did not change the original label.

Harmonizome web resource implementation

The Harmonizome web server is a Java servlet built with Java 8 and running in an Apache Tomcat 8 container. The application and all its dependencies are running within a Docker virtual machine and deployed to a 16-node cluster. The cluster distributes resources using Apache Mesos. With Mesos, the Harmonizome can run on any of the 16

nodes and switch to a new machine if its current node goes down. The Harmonizome database runs on an internal MariaDB server. MariaDB is a drop-in replacement for MySQL. The application communicates with the database through Hibernate object-relational mapping (ORM). An ORM is a framework that maps a tabular schema onto an object paradigm. For example, a single row in the Harmonizome Gene Table is an instance of a Gene class in Java. The search engine uses exact and full-text MariaDB queries to search the database for relevant matches. MariaDB's natural language search functionality prioritizes the results. We implemented JavaServer Pages (JSP) for most of the user interface. Styling is specified with Less, a Cascading Style Sheets (CSS) pre-processor. We used JavaScript for front-end scripting.

Funding

This work was supported from the national institutes of health (NIH) (R01GM098316, U54HL127624 and U54CA189201 to A.M.).

Acknowledgements

We would like to thank Dr Kathleen Jagodnik for copyediting this article.

Conflict of interest. None declared.

References

1. Wu,C., MacLeod,I. and Su, A.I. (2012) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res*., gks1114.
2. Brown,G.R., Hem,V., Katz,K.S. *et al*. (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*., 43, D36–D42.
3. Consortium,U. (2010) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res*., 38, D142–D148.
4. Baker,E.J., Jay,J.J., Bubier,J.A. *et al*. (2012) GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Res*., 40, D1067–D1076.
5. Liberzon,A., Subramanian,A., Pinchback,R. *et al*. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27, 1739–1740.
6. Zambon,A.C., Gaj,S., Ho,I. *et al*. (2012) GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*, 28, 2209–2210.
7. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., 25, 25–29.
8. Safran,M., Dalah,I., Alexander,J. *et al*. (2010) GeneCards Version 3: the human gene integrator. *Database*, 2010, baq020.
9. Consortium,U. (2008) The universal protein resource (UniProt). *Nucleic Acids Res*., 36, D190–D195.
10. Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet*., 40, 1047–1051.
11. Hoffmann, R. and Valencia, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21, ii252–ii258.
12. Baroukh,C., Jenkins,S.L., Dannenfelser,R. *et al*. (2011) Genes2WordCloud: a quick way to identify biological themes from gene lists and free text. *Source Code Biol. Med*., 6, 1–5.
13. Wang, Z., Clark, N.R. and Ma'ayan, A. (2015) Dynamics of the discovery process of protein-protein interactions from low content studies. *BMC Systems Biology*, 9, 26.
14. Edwards, A.M., Isserlin, R., Bader, G.D. *et al*. (2011) Too many roads not taken. *Nature*, 470, 163–165.
15. Juty,N., Le Novère,N. and Laibe,C. (2012) Identifiers. org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res*., 40, D580–D586.
16. Aranda,B., Blankenburg,H., Kerrien,S. *et al*. (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, 8, 528–529.
17. Franceschini, A., Szklarczyk, D., Frankild, S. *et al*. (2013) STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., 41, D808–D815.
18. Gaudet,P., Bairoch,A., Field,D. *et al*. (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Database*, 2011, baq027.
19. Field,D., Sansone,S.A., Collis,A. *et al*. (2009) 'Omics data sharing. *Science (New York, NY)*, 326, 234.
20. Ma'ayan,A., Rouillard,A.D., Clark,N.R. *et al*. (2014) Lean Big Data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci*., 35, 450–460.
21. Rouillard,A.D., Wang,Z. and Ma'ayan,A. (2015) Reprint of "Abstraction for data integration: Fusing mammalian molecular, cellular and phenotype big datasets for better knowledge extraction". *Comput. Biol. Chem*.
22. Cowley,G.S., Weir,B.A., Vazquez,F. *et al*. (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data*, 1, 140035.
23. Network,TCTDaD. (2010) Towards patient-based cancer therapeutics. *Nat. Biotechnol*., 28, 904–906.
24. Cheung,H.W., Cowley,G.S., Weir,B.A. *et al*. (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. USA*, 108, 12372–12377.
25. Su,A.I., Cooke,M.P., Ching,K.A. *et al*. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA*, 99, 4465–4470.
26. Su,A.I., Wiltshire,T., Batalov,S. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, 101, 6062–6067.
27. Barretina,J., Caponigro,G., Stransky,N. *et al*. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483, 603–607.
28. Forbes,S.A., Bindal,N., Bamford,S. *et al*. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*., 39, D945–D950.
29. Forbes,S.A., Tang,G., Bindal,N. *et al*. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to

investigate acquired mutations in human cancer. *Nucleic Acids Res*., 38, D652–D657.
30. Garnett, M.J., Edelman, E.J., Heidorn, S.J. *et al*. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483, 570–575.
31. Heiser,L.M., Sadanandam,A., Kuo,W.L. *et al*. (2012) Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. USA*, 109, 2724–2729.
32. Uhlen,M., Fagerberg,L., Hallstrom,B.M. *et al*. (2015) Tissue-based map of the human proteome. *Sci. Proteomics*, 347, 1260419.
33. Klijn,C., Durinck,S., Stawiski,E.W. *et al*. (2015) A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol*., 33, 306–312.
34. Kim,M.S., Pinto,S.M., Getnet,D. *et al*. (2014) A draft map of the human proteome. *Nature*, 509, 575–581.
35. Wilhelm,M., Schlegl,J., Hahne,H. *et al*. (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, 509, 582–587.
36. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F. *et al*. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol*., 28, 1045–1048.
37. Roadmap Epigenomics,C., Kundaje,A., Meuleman,W. *et al*. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 317–330.
38. Sunkin,S.M., Ng,L., Lau,C. *et al*. (2013) Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res*., 41, D996–D1008.
39. Jones,A.R., Overly,C.C. and Sunkin,S.M. (2009) The Allen Brain Atlas: 5 years and beyond. *Nat. Rev. Neurosci*., 10, 821–828.
40. Shen,E.H., Overly,C.C. and Jones,A.R. (2012) The Allen Human Brain Atlas Comprehensive gene expression mapping of the human brain. *Trends Neurosci*., 35, 711–714.
41. Consortium,G. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet*., 45, 580–585.
42. Consortium,G. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348, 648–660.
43. Cancer Genome Atlas Research,N., Weinstein,J.N., Collisson,E.A. *et al*. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet*., 45, 1113–1120.
44. Hawrylycz,M.J., Lein,E.S., Guillozet-Bongaarts,A.L. *et al*. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489, 391–399.
45. Lein,E.S., Hawrylycz,M.J., Ao,N. *et al*. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445, 168–176.
46. Miller,J.A., Ding,S.L., Sunkin,S.M. *et al*. (2014) Transcriptional landscape of the prenatal human brain. *Nature*, 508, 199–206.
47. Santos,A., Tsafou,K., Stolte,C. *et al*. (2015) Comprehensive comparison of large-scale tissue expression datasets. *PeerJ*., 3, e1054.
48. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A. *et al*. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res*., 37, D786–D792.
49. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K. *et al*. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res*., 43, D914–D920.
50. Law,V., Knox,C., Djoumbou,Y. *et al*. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*., 42, D1091–D1097.
51. Wishart,D.S., Knox,C., Guo,A.C. *et al*. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*., 34, D668–D672.
52. Pawson,A.J., Sharman,J.L., Benson,H.E. *et al*. (2014) The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res*., 42, D1098–D1106.
53. Wishart,D.S., Jewison,T., Guo,A.C. *et al*. (2013) HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res*., 41, D801–D807.
54. Wishart,D.S., Tzur,D., Knox,C. *et al*. (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res*., 35, D521–D526.
55. Lamb,J., Crawford,E.D., Peck,D. *et al*. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313, 1929–1935.
56. Barrett,T., Wilhite,S.E., Ledoux,P. *et al*. (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*., 41, D991–D995.
57. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., 30, 207–210.
58. Dudley,J.T., Tibshirani,R., Deshpande,T. *et al*. (2009) Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol*., 5, 307.
59. Duan,Q., Flynn,C., Niepel,M. *et al*. (2014) LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res*.,
60. Pletscher-Frankild,S., Palleja,A., Tsafou,K. *et al*. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, 74, 83–89.
61. Becker,K.G., Barnes,K.C., Bright,T.J. *et al*. (2004) The genetic association database. *Nat. Genet*., 36, 431–432.
62. Li,M.J., Wang,P., Liu,X. *et al*. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res*., 40, D1047–D1054.
63. Hornbeck,P.V., Chabra,I., Kornhauser,J.M. *et al*. (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4, 1551–1561.
64. Hornbeck,P.V., Zhang,B., Murray,B. *et al*. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res*., 43, D512–D520.
65. Landrum,M.J., Lee,J.M., Riley,G.R. *et al*. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*., 42, D980–D985.
66. Welter,D., MacArthur,J., Morales,J. *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., 42, D1001–D1006.
67. Kohler,S., Doelken,S.C., Mungall,C.J. *et al*. (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*., 42, D966–D974.

68. Yu,W., Gwinn,M., Clyne,M. *et al.* (2008) A navigator for human genome epidemiology. *Nat. Genet.*, 40, 124–125.
69. Blake,J.A. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, 31, 193–195.
70. Eppig,J.T., Blake,J.A., Bult,C.J. *et al.* Mouse Genome Database, G (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.*, 43, D726–D736.
71. Smith,C.L. and Eppig,J.T. (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm. Genome*, 23, 653–668.
72. Smith,C.L. and Eppig,J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *WIREs Syst. Biol. Med.*, 1, 390–399.
73. Hamosh,A., Scott,A.F., Amberger,J. *et al.* (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 30, 52–55.
74. Amberger,J.S., Bocchini,C.A., Schiettecatte,F. *et al.* (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, 43, D789–D798. 7
75. Tryka,K.A., Hao,L., Sturcke,A. *et al.* (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.*, 42, D975–D979.
76. Mailman,M.D., Feolo,M., Jin,Y. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, 39, 1181–1186.
77. Gene Ontology,C. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 43, D1049–D1056.
78. Mitchell,J.A., Aronson,A.R., Mork,J.G. *et al.* (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu. Symp. Proc.*, 2003, 460–464.
79. Binder,J.X., Pletscher-Frankild,S., Tsafou,K. *et al.* (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)*, 2014, bau012.
80. Sprenger,J., Lynn Fink,J., Karunaratne,S. *et al.* (2008) LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res.*, 36, D230–D233.
81. Caspi, R., Foerster, H., Fulcher, C.A. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 36, D623–D631.
82. Caspi,R., Altman,T., Billington,R. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 42, D459–D471. 4
83. Ogata,H., Goto,S., Sato,K. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 27, 29–34.
84. Kanehisa,M., Goto,S., Sato,Y. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42, D199–D205.
85. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, 41, D377–D386.
86. Thomas,P.D. (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.*, 31, 334–341.
87. Schaefer,C.F., Anthony,K., Krupa,S. *et al.* (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, 37, D674–D679.
88. Croft,D., Mundo,A.F., Haw,R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, 42, D472–D477.
89. Joshi-Tope,G., Gillespie, M., Vastrik, I. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33, D428–D432.
90. Kelder,T., van Iersel,M.P., Hanspers, K. *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, 40, D1301–D1307.
91. Ruepp,A., Brauner,B., Dunger-Kaltenbach,I. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, 36, D646–D650.
92. Ruepp,A., Waegele,B., Lechner,M. *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes–. *Nucleic Acids Res.*, 38, D497–D501. (2009)
93. Malovannaya,A., Lanz,R.B., Jung,S.Y. *et al.* (2011) Analysis of the human endogenous coregulator complexome. *Cell*, 145, 787–799.
94. Clark,N.R., Dannenfelser,R., Tan,C.M. *et al.* (2012) Sets2Networks: network inference from repeated observations of sets. *BMC Syst. Biol.*, 6, 89.
95. Xu,H., Baroukh,C., Dannenfelser,R. *et al.* (2013) ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database (Oxford)*, 2013, bat045.
96. Culhane,A.C., Schroder,M.S., Sultana,R. *et al.* (2012) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.*, 40, D1060–D1066.
97. Culhane,A.C., Schwarzl,T., Sultana,R. *et al.* (2010) GeneSigDB–a curated database of gene expression signatures. *Nucleic Acids Res.*, 38, D716–D725.
98. Subramanian,A., Tamayo,P., Mootha,V.K. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A*, 102, 15545–15550.
99. Chen,E.Y., Tan,C.M., Kou,Y. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14,
100. Bader,G.D. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 31, 248–250.
101. Isserlin,R., El-Badrawi,R.A. and Bader,G.D. (2011) The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)*, 2011, baq037.
102. Stark,C., Breitkreutz,B.J., Reguly,T. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34, D535–D539.
103. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, 43, D470–D478.

104. Salwinski,L., Miller,C.S., Smith,A.J. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, 32, D449–D451.
105. Keshava Prasad,T.S., Goel,R., Kandasamy,K. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37, D767–D772.
106. Peri,S., Navarro,J.D., Kristiansen,T.Z. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, 32, D497–D501.
107. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, 32, D452–D455.
108. Kerrien, S., Aranda, B., Breuza, L. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40, D841–D846.
109. Cerami,E.G., Gross,B.E., Demir,E. *et al.* (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39, D685–D690.
110. Lachmann,A. and Ma'ayan,A. (2009) KEA: kinase enrichment analysis. *Bioinformatics*, 25, 684–686.
111. Hsu,S.D., Lin,F.M., Wu,W.Y. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, 39, D163–D169.
112. Hsu,S.D., Tseng,Y.T., Shrestha,S. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, 42, D78–D85.
113. Garcia,D.M., Baek,D., Shin,C. *et al.* (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, 18, 1139–1146.
114. Grimson,A., Farh,K.K., Johnston,W.K. *et al.* (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell.*, 27, 91–105.
115. Friedman,R.C., Farh,K.K., Burge,C.B. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19, 92–105.
116. Duan,G., Li,X. and Kohn,M. (2015) The human DEPhOsphorylation database DEPOD: a 2015 update. *Nucleic Acids Res.*, 43, D531–D535.
117. Lachmann,A., Xu,H., Krishnan,J. *et al.* (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26, 2438–2444.
118. Consortium,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306, 636–640.
119. Consortium,E.P. (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.*, 9, e1001046.
120. Mathelier,A., Zhao,X., Zhang,A.W. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 42, D142–D147.
121. Sandelin,A., Alkema,W., Engstrom,P. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32, D91–D94.
122. Matys,V. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31, 374–378.
123. Matys,V., Kel-Margoulis,O.V., Fricke,E. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34, D108–D110.
124. Chatr-Aryamontri,A., Ceol,A., Peluso,D. *et al.* (2009) VirusMINT: a viral protein interaction database. *Nucleic Acids Res.*, 37, D669–D673.
125. Mitchell,A., Chang,H.Y., Daugherty,L. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, 43, D213–D221.
126. Apweiler,R., Attwood,T.K., Bairoch,A. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domain and functional sites. *Nucleic Acids Res.*, 29, 37–40.
127. Apweiler,R., Bairoch,A., Wu,C.H. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32, D115–D119.
128. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204–D212.
129. Blake,J.A., Bult,C.J., Eppig,J.T. *et al.* (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, 42, D810–D817.
130. Dannenfelser,R., Clark,N. and Ma'ayan, A. (2012) Genes2FANs: connecting genes through functional association networks. *BMC Bioinformatics*, 13, 156.
131. Lai,S.L., Brauch,H., Knutsen,T. *et al.* (1995) Molecular genetic characterization of neuroendocrine lung cancer cell lines. *Anticancer Res.*, 15, 225–232.
132. Carney,D.N., Gazdar,A.F., Bepler,G. *et al.* (1985) Establishment and identification of small cell lung cancer cell lines having classic and variant features. *Cancer Res.*, 45, 2913–2923.
133. Johnson,B.E., Whang-Peng,J., Naylor,S.L. *et al.* (1989) Retention of chromosome 3 in extrapulmonary small cell cancer shown by molecular and cytogenetic studies. *J. Natl. Cancer Inst.*, 81, 1223–1228.
134. Haibe-Kains,B., El-Hachem,N., Birkbak,N.J. *et al.* (2013) Inconsistency in large pharmacogenomic studies. *Nature*, 504, 389–393.
135. Samalin,E., Bouche,O., Thezenas,S. *et al.* (2014) Sorafenib and irinotecan (NEXIRI) as second- or later-line treatment for patients with metastatic colorectal cancer and KRAS-mutated tumours: a multicentre Phase I/II trial. *Br. J. Cancer*, 110, 1148–1154.
136. Jane,E.P., Premkumar,D.R., Cavaleri,J.M. *et al.* (2015) Dinaciclib, a CDK inhibitor promotes proteasomal degradation of Mcl-1 and enhances ABT-737 mediated cell death in malignant human glioma cell lines. *J. Pharmacol. Exp. Ther.*, 356: 354–365.
137. Homsi,J., Cubitt,C.L., Zhang,S. *et al.* (2009) Src activation in melanoma and Src inhibitors as therapeutic agents in melanoma. *Melanoma Res.*, 19, 167–175.
138. Rabbani,S.A., Valentino,M.L., Arakelian,A. *et al.* (2010) SKI-606 (Bosutinib) blocks prostate cancer invasion, growth, and metastasis in vitro and in vivo through regulation of genes involved in cancer growth and skeletal metastasis. *Mol. Cancer Ther.*, 9, 1147–1157.
139. Daud,A.I., Krishnamurthi,S.S., Saleh,M.N. *et al.* (2012) Phase I study of bosutinib, a src/abl tyrosine kinase inhibitor,

administered to patients with advanced solid tumors. *Clin. Cancer Res*., 18, 1092–1100.
140. Bagal,S.K., Brown,A.D., Cox,P.J. *et al*. (2013) Ion channels as therapeutic targets: a drug discovery perspective. *J. Med. Chem*., 56, 593–624.
141. Clare,J.J. (2010) Targeting ion channels for drug discovery. *Discov. Med*., 9, 253–260.
142. Fagerberg,L., Jonasson,K., von Heijne,G. *et al.* (2010) Prediction of the human membrane proteome. *Proteomics*, 10, 1141–1149.
143. Brown,S.D. and Moore,M.W. (2012) The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. *Mamm. Genome*, 23, 632–640.
144. Lagerstrom,M.C. and Schioth,H.B. (2008) Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat. Rev. Drug Discov*., 7, 339–357.
145. Dohlman,H.G. (2015) Thematic minireview series: new directions in G protein-coupled receptor pharmacology. *J. Biol. Chem*., 290, 19469–19470.
146. Stockert,J.A. and Devi,L.A. (2015) Advancements in therapeutically targeting orphan GPCRs. *Front. Pharmacol*., 6, 100.
147. Huang,X.P., Karpiak,J., Kroeze,W.K. *et al*. (2015) Allosteric ligands for the pharmacologically dark receptors GPR68 and GPR65. *Nature*, 527: 477–483.
148. Gomes,I., Bobeck,E.N., Margolis,E.B. *et al*. (2016) Identification of GPR83 as the receptor for the neuroendocrine peptide PEN. *Sci. Signal*., 9, ra43.
149. Wu,P., Nielsen,T.E. and Clausen, M.H. (2015) FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci*., 36, 422–439.
150. Fabbro,D., Cowan-Jacob,S.W. and Moebitz, H. (2015) Ten things you should know about protein kinases: IUPHAR Review 14. *Br. J. Pharmacol*., 172, 2675–2700.
151. Hornbeck,P.V., Kornhauser,J.M., Tkachev,S. *et al*. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res*., 40, D261–D270.
152. Touw,W.G., Bayjanov,J.R., Overmars,L. *et al*. (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief. Bioinform*., 14, 315–326.
153. Breiman, L. (2001) Random forests. *Mach. Learn*., 45(1), 5–32.