Bioinformatics, bty027, https://doi.org/10.1093/bioinformatics/bty027 (2018)

ADEPTUS: A discovery tool for disease prediction, enrichment and network analysis based on profiles from many diseases

David Amar[1,2], Amir Vizel[2], Carmit Levy[3], and Ron Shamir[2]

1. Stanford Center for Inherited Cardiovascular Disease, Stanford University, Stanford, CA 94305, USA.
2. The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel.
3. Department of Human Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Abstract

ADEPTUS is a web-tool that enables various functional genomics analyses based on a high quality curated database spanning >38,000 gene expression profiles and >100 diseases. It offers four types of analysis. (1) For a gene list provided by the user it computes disease ontology (DO), pathway, and gene ontology (GO) enrichment and displays the genes as a network. (2) For a given disease, it enables exploration of drug repurposing by creating a gene network summarizing the genomic events in it. (3) For a gene of interest, it generates a report summarizing its behavior across several studies. (4) It can predict the tissue of origin and the disease of a sample based on its gene expression or its somatic mutation profile. Such analyses open novel ways to understand new datasets and to predict primary site of cancer.

Availability: data and tool: http://acgt_adeptus_home.cs.tau.ac.il/home; a description of the analyses: **Supplementary Text**.

Introduction

Case-control studies seek discriminatory signals between different phenotypes, in order to decipher the molecular basis of disease. Most analyses address a small number of studies, and typically span very few phenotypes and tissues and a modest number of samples. These limitations reduce the reliability and specificity of disease data analysis. We introduce ADEPTUS, a new web-tool that aims to overcome these issues by conducting analyses based on multiple diseases and numerous studies simultaneously. It employs a novel high quality database of >37,000 gene expression profiles and >9,500 cancer somatic mutation profiles, covering >200 different disease and tissue labels. We developed a classifier that gave high quality predictions for 68 of the disease and tissue labels. Good predictors also produced biomarkers that were scored for replicability, intensity, and specificity. ADEPTUS uses these results for a variety of functional analyses, including disease and function enrichment of gene sets, network analysis of a disease, and disease label prediction (**Figure 1A**).

ADEPTUS

The gene expression database

We assembled and manually annotated 37,337 gene expression profiles. Each profile was annotated with the tissue of origin and either with a set of DO terms or as a control. The *label* of a profile is a phenotype of a disease, a tissue, or both. For example, melanoma, skin, and melanoma of skin are all labels. The profiles cover 10,501 genes that were shared across all samples, 190 disease labels and 18 tissue labels. Using this database we performed for each label and gene a conservative leave-study-out cross validation (Amar *et al.*, 2015), obtaining three scores for label-gene association: *signal strength* and *specificity* (computed using ROC, default threshold: 0.7), and *replicability* (computed using *meta-analysis* over the datasets, default threshold-value: 0.01; >50% p-values<0.05 was defined as replicated signal). See the **Supplementary Text** for details. 68 labels passed all thresholds. Those were defined as *well-classified* and used for subsequent analyses.

Analysis of a gene list

Users can upload a gene list and perform several enrichment analyses: (i) GO enrichments, computed using TANGO (Ulitsky *et al.*, 2010), (ii) pathway enrichment, computed using Fisher's exact test on KEGG (Kanehisa and Goto, 2000) and WikiPathways (Kelder *et al.*, 2012) pathways, (iii) DO enrichment, using a GSEA-like analysis (Subramanian *et al.*, 2005). To the best of our knowledge, this is the first tool to provide disease enrichment.

Analysis of a disease

For each well-classified label the tool displays its biomarker genes in a summary network. Nodes are genes and edges are protein or genetic interactions taken from GeneMANIA (Vlasblom *et al.*, 2014). Color-coding of the gene's node gives information on it: up- or down-regulation compared to the negatives or the background, and availability of a drug targeting the gene (Law *et al.*, 2014). For cancer labels, a color indicates when the gene is associated with the label based on somatic mutation data (Amar *et al.*, 2017). The set of genes displayed can be changed by adjusting score thresholds.

Analysis of new profiles

Our classifiers (for the well-classified labels) can be applied on new profiles. Given an input matrix of patients x profiles (gene expression or mutated genes), the classifiers output a table of patients x labels, whose values are the predicted association probabilities. When the input is the mutated genes in a biopsy, we also predict its cancer subtype (e.g. the primary cancer of a metastasis sample) using our multi-label classifiers (Amar *et al.*, 2017).

Results

We analyzed melanoma, the most lethal and treatment-resistant human skin cancer, for which new treatment and prevention approaches are needed (Hodis *et al.*, 2012). The ADEPTUS database had three relevant labels: melanoma, skin melanoma, and uveal melanoma. Only melanoma was well-classified, mainly because no reliable classifier distinguished between skin and uveal melanoma.

Melanocytes, the melanoma origin cells, are specialized in releasing pigment vesicles, termed melanosomes (Raposo and Marks, 2007). Melanoma keeps this ability in order to directly affect

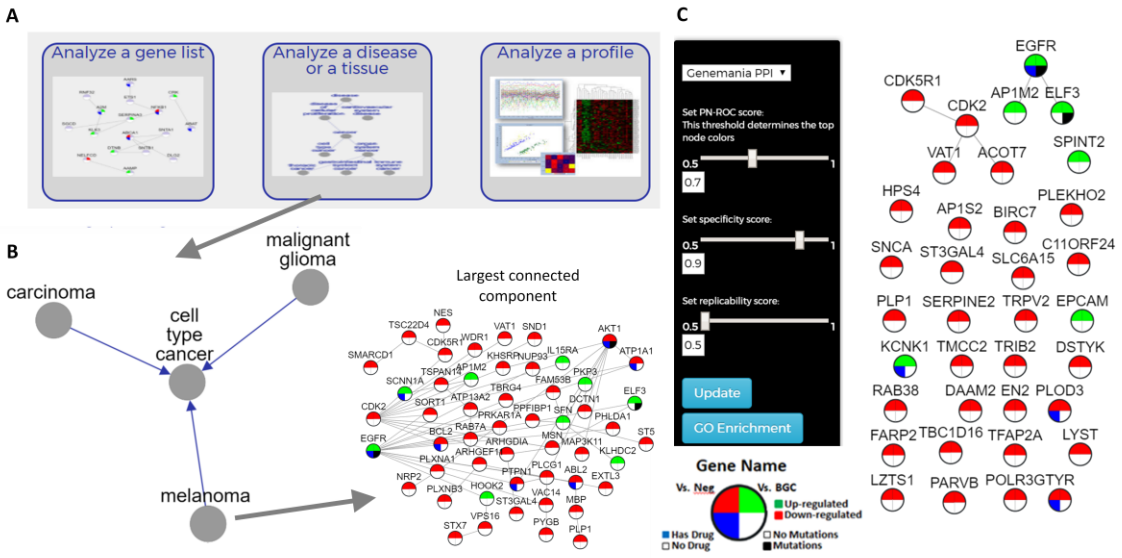
the formation of their tumor niche by microRNA trafficking via melanosomes (Dror *et al.*, 2016). Remarkably, ADEPTUS selected genes (**Figure 1B**) that were enriched for vesicles transport and pigmentation (q < 1e-03). When we compared melanoma to other diseases by increasing the specificity score to 0.9 (**Figure 1C**), melanoma vesicles were again found as the most significant GO enrichment. Dror et al. further found that inhibition of melanosome trafficking by melanoma can block melanoma formation (Dror *et al.*, 2016). ADEPTUS offered new drug targets (**Figure 1B,C**) with promising therapeutic potential to block melanosome trafficking in addition to the drug used in (3). Taken together, ADEPTUS is a strong tool for identifying disease related genes and for proposing new potential drugs.

Acknowledgments

C.L. thanks the support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 726225). DA was supported in part by a fellowship from the Safra Center for Bioinformatics at Tel Aviv University. RS was supported in part by the Israel Science Foundation (grant 317/13) and by the Bella Walter Memorial Fund of the Israel Cancer Association.

Figure Legends

Figure 1. Analysis of melanoma in ADEPTUS. A) Three analysis types. B) The resulting DO subnetwork, containing the term melanoma, and the largest connected component of its biomarker genes. C) The network analysis options and the subnetwork produced by setting specificity ROC=0.9 for melanoma.



References

Amar,D. *et al.* (2015) Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Res.*, **43**, 7779–7789.

Amar,D. *et al.* (2017) Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications. *Oncogene*.

Dror,S. *et al.* (2016) Melanoma miRNA trafficking controls tumour primary niche formation. *Nat. Cell Biol.*, **18**, 1006–1017.

Hodis,E. *et al.* (2012) A landscape of driver mutations in melanoma. *Cell*, **150**, 251–263.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucl. Acids Res.*, **28**, 27–30.

Kelder,T. *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301--7.

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

Raposo,G. and Marks,M.S. (2007) Melanosomes--dark organelles enlighten endosomal membrane transport. *Nat. Rev. Mol. Cell Biol.*, **8**, 786–97.

Subramanian,A. *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**, 15545–50.

Ulitsky,I. *et al.* (2010) Expander: from expression microarrays to networks and functions. *Nat. Protoc.*, **5**, 303–322.

Vlasblom,J. *et al.* (2014) Novel function discovery with GeneMANIA: a new integrated resource for gene function prediction in Escherichia coli. *Bioinformatics*, 1–5.

Supplementary Text

Table of Contents

Outline

We developed ADEPTUS: a database and a web tool for disease and tissue-oriented analysis of gene sets and profiles. This document describes the three aspects of ADEPTUS: (1) the database, (2) our multi-label analyses, and (3) the tools that are available in the webtool. The document is ordered as follows. First, we describe the new gene expression database, which covers >37,000 gene expression profiles each manually annotated with tissue and disease labels. Second, we explain our univariate analysis. This is the main statistical tool that we used to validate the database and for extracting biomarkers. Third, we explain our multivariate classification analysis, which was used for quality assurance of the database and for identifying well-classified diseases. Fourth, we explain the enrichment analyses that are provided in the web tool. Fifth, we show the capabilities and the utility of our database and tools via analysis of melanoma. Finally, we briefly explain the development of the web tool, its code, and how it can be expanded.

1. Gene expression profiles: collection and annotation

1.1 The original ADEPTUS database

In our previous work (1) we created a large compendium of expression profiles. The profiles, which were generated using different technologies, were manually annotated with disease attributes. This database, which we call ADEPTUS V.0, was used for inferring disease-specific reliable biomarkers and differential genes. Further integration with additional information sources such as gene networks and gene-drug associations was used to suggest novel drug repurposing candidates.

The ADEPTUS V.0 database contained 174 gene expression studies from GEO, each with at least 20 samples. This amounted to 13,314 samples from 17 different microarray technologies and 1,526 RNA-Seq samples from TCGA. Each sample was either assigned a set of disease ontology (DO) terms based on the its textual description, or labeled as 'control'. In our original analysis we kept only DO terms that were represented by at least five different datasets, which resulted in 48 disease terms. For each study we used the preprocessed expression matrix given in the database. Each expression profile was normalized separately based on its weighted ranks (details below). This normalization allows joint analysis of samples from different technologies and studies, at the expense of some loss of information.

1.2 The new database

We reanalyzed the data and corrected labeling errors in our original annotations in four GEO datasets (GSE21374, GDS1746, GDS4387, and GDS2113). We then added 31 new studies from GEO and all RNA-seq samples from TCGA (2) and GTeX (3). The new database contained 37,337 samples from 234 studies. The TCGA data was partitioned into studies based on the cancer subtypes. The GTeX data were considered as a single study. In total, 17,732 of the samples are of microarray platforms, and 19,605 are of RNA-Seq.

We originally had >18,500 microarray samples in our data. However, using the metadata in GEO for mapping probes to genes resulted in a very small set of genes that are common to all platforms. To improve gene coverage we used mapping of genes to GenBank and RefSeq ids on top of what is given in the platform data in GEO. That is, if a probe p is mapped to a RefSeq id r and r is mapped to a gene g (according to the RefSeq database) then we added p to the probe set of g. In addition, we excluded samples from platform GPL6102 as its genes set poorly overlapped with the other platforms. These improvements resulted in 10,081 genes that were shared across all remaining platforms, while keeping almost all samples.

A new feature of the current database is annotation of samples to tissue. We defined a set of 43 "tissue slim" terms and mapped each sample into one of them. The tissue terms are: prostate, immune system or blood, intestines, brain, epithelial cells, liver, connective tissue, adrenal glands, bronchi, genitalia, female, blood vessels, peripheral nervous system, gastrointestinal tract, heart, kidney, lung, lymph nodes, breast, mouth, pharynx, musculoskeletal system, central nervous system, spleen, genitalia, male, thyroid gland, tongue, trachea, skin, sputum, saliva, neurons, embryonic structures, germ cells, circulating endothelial cells, bile ducts, pancreas, stem cells, soft tissue, paraganglia, chromaffin, urogenital system, thymus gland, head and neck, eye.

1.3 Label definition

In our analysis, a label corresponds to a phenotype shared by group of samples. The labeling is summarized by a binary matrix Y of samples (rows) over labels (columns). That is, each label l has a column in Y that induces a binary partition of the samples into those with the phenotype that l represents and those without it.

To formally define the labels, let us start with some notation. Let D be the set of all DO terms in the database. That is, a DO term d belongs to D if we have at least one sample in the database has the disease d. Let D_G be the subgraph of the complete DO directed acyclic graph (DAG) induced by D. D_G is also a DAG and its arcs represent is-a relations among the node terms. For each node d that has children in D_G we add a new dummy term d^* to D, and in D_G we added an arc from d^* to its parent d. This term is added to represent the samples that are assigned to d but not to any of its children (see the example in **Figure SF1B**). Let D^* be the set of all d^* terms. Let T be the set of all tissue slim terms.

A label l is a pair $< d, t >$, where d is either a disease term or 'control' (i.e., $d \in \{D \cup D^* \cup \{'control'\}\}$), and t is either a tissue term from T or 'Any' (i.e., $t \in \{T \cup \{'Any'\}\}$). 'Any' means that the label is defined solely by the information in d. **Figure SF1A** gives an example related to melanoma. The melanoma parent label has $d = melanoma$ and, $t = Any$. Its children also have $d = melanoma$, but they differ in their tissue information (skin vs. uveal).

Using the notation above, we define four types of labels. (1) *control-tissue*: labels with $d = control$. These are samples that are not annotated with any disease and the tissue can take any of the values defined above (including 'Any'). (2) *DO-tissue*: labels with $d \in D \backslash D^*$, and $t \in T$. (3) *DO-only*: labels in which $t = Any$. (4) DO*: labels with $d \in D^*$. **Figure SF1B** shows some label examples. All labels have $t = immune\ system\ or\ blood$.

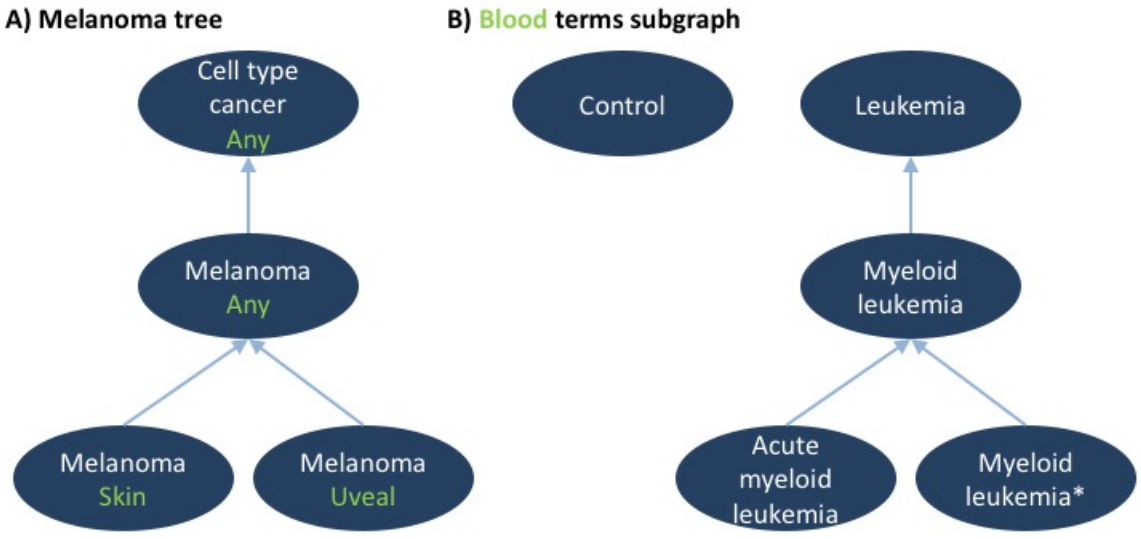


Figure SF1. Example of labels in the ADEPTUS database. Nodes are labels. Arcs denote is-a relations. Tissue information is written in green. A) The disease ontology subgraph of melanoma related terms. The top two terms have no tissue annotation. The melanoma parent term has two children: the disease is the same but the tissue differs. B) A subgraph of the terms of blood tissues. The control term marks healthy blood samples. A subgraph of leukemia related terms is shown. The children of myeloid leukemia are acute myeloid leukemia (AML) and myeloid leukemia*, which marks all myeloid leukemia patients that are not AML (e.g., subacute myeloid leukemia and chronic myeloid leukemia).

We kept only labels whose sample set contained at least 100 samples that originated from at least three datasets. However, after this filter, there may still be distinct labels that correspond to the same sample set. This can only occur (by definition) when the two labels are not controls. We therefore applied the following additional filter to removed redundancies. For each pair of labels $l_1 = < d_1, t_1 >, l_2 =< d_2, t_2 >$ that had the same sample sets we first compared their disease terms and if one is the ancestor of the other, then we removed the ancestor. In addition, if $t_1 = 'Any'$ and $t_2 \neq 'Any'$ (or vice versa) then we kept the non-'Any' label only. The filters above left us with 204 labels that covered 96 different DO terms, and included 18 different control terms (each of a different tissue).

1.4 Gene annotation and profile standardization

As in our previous work (1), given a gene expression profile of a single sample S in which k genes were measured we transformed the profile into scores based on the gene ranks as follows. We first ranked the genes by their expression levels g_1, g_2, g_3,…, g_k (with g_1 having the highest level), and assigned a score to each gene based on its rank: $W_S(g_i) = ke^{-1} - ie^{-i/k}$. This transformation produces very small differences among low rank genes (i.e., it keeps a small difference between genes with low expression intensities). **Figure SF2A** illustrates the effect of the transformation on a vector of 1000 values sampled independently from a standard uniform distribution. Note that in (1) we used the score $W^*_S(g_i) = ie^{-i/k}$, so that the highly expressed genes get the lowest scores. Here we defined $W_S(g_i)$ so that the ordering by scores matches the original gene ranking.

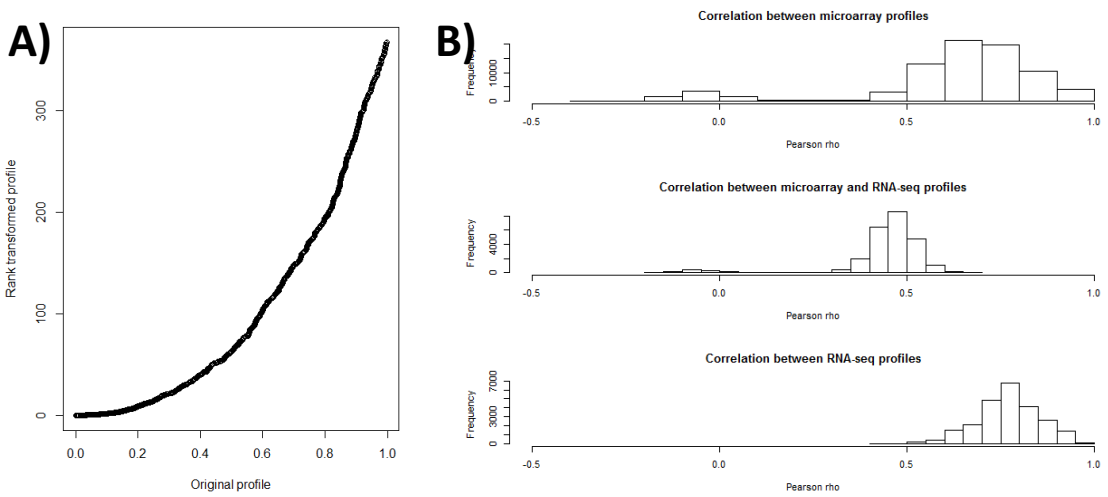


Figure SF2. Effects of the rank-based normalization. A) A profile of 1000 values sampled from a standard uniform distribution plotted against the normalized profile. B) Distribution of the correlation among rank-normalized profiles for different technologies.

Any integration of very heterogeneous profiles as in ADEPTUS may raise concerns regarding the ability to compare results of different platforms and preprocessing. **Figure SF2B** shows the distribution of correlation values when taking 400 samples at random from each technology type: microarray and RNA-seq. The correlation was calculated between the weighted ranks profiles of the samples. The results show that in general the correlations between platforms are high and are very similar (mean correlation > 0.45). There is a clear advantage to correlations of samples from the same technology type. The lower correlation of microarray samples can be attributed in part to the fact that in our compendium the microarray studies happened to have higher biological diversity, as well as to the much higher number of different microarray platforms (>15).

2. Univariate analysis

Our goal in this section is to quantify the association between a vector of scores v and a label l. As discussed above, each label l induces a binary partition Y_l of the samples based on whether they belong to l or not. The vector v, which is defined on the complete sample set, contains sample scores taken or calculated from the database. For example, v can be the expression level of a gene or a vector of probabilities that result from performing cross-validation.

Standard parametric or non parametric measures for association between Y_l and v such as a standard ROC score are oblivious to the grouping of the samples into different datasets. Thus,

such measures cannot check directly if the association is replicable across the datasets. In addition, such measures do not make use of the DO hierarchy to check if the association is specific. Our univariate analysis aims to quantify three different aspects of the association between v and l: (1) Study-based *Replicability and meta-analysis,* (2) *Overall separation* of the cases from their direct and indirect controls, and (3) *Specificity* to the analyzed disease.

All scores below take as input the vectors v and l and integrate them with other information on the samples. For each label l we partition the samples into three groups: (i) positives: patients with the phenotype that l represents; (ii) negatives: control non-positive samples originating from the same studies as the positive samples and (iii) background controls (BGCs): all other samples. See (1) for a discussion of the rationale of these definitions.

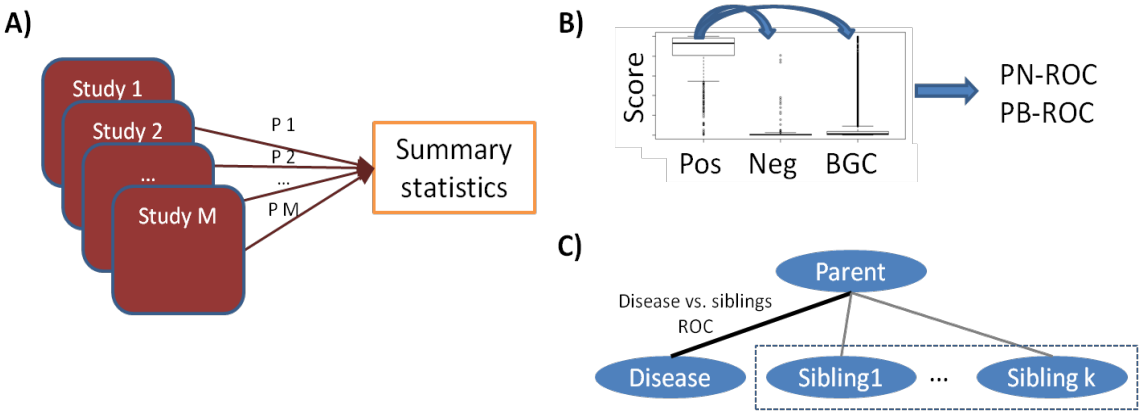


Figure SF3. Three scores of univariate analysis. A) Study-based. A p-value for the association is computed in each study separately. The output is a summary statistic of these p-values. B) Overall intensity score, which measures the separation between the label cases (Positives) from their direct (negatives) controls and from the background controls (BGCs). The output is a ROC score for each comparison. C) Specificity score. Here the separation of the cases from patients with sibling labels is computed.

Replicability and meta-analysis (Figure SF3A): Here we first check which of the datasets contain positive samples. Of those we take only the ones that contain at least ten positive samples and at least ten negative samples. Assume that M such studies remain: $DS_1, \ldots, DS_M$. In each study we use the non-parametric Wilcoxon rank-sum test to calculate a p-value for the separation between the positives and the negatives based on their scores in v. This calculation produces a vector of p-values: $P=P_1,\ldots,P_M$. We calculate two summary statistics of P. First, we perform standard meta-

analysis. Under the assumption that the datasets are independent, we use Fisher's meta-analysis to test the null hypothesis that the positives and negatives have the same distribution within each study. We call this score study-based meta analysis q-value (SMQ). Second, we compute the proportion of p-values that are lower than 0.05. We denote this value as the *replicability score*. The default threshold for selection was 0.01 for the q-value of the meta-analysis and 0.5 for replicability.

Overall separation analysis (Figure SF3B): Here we compute two ROC scores: PN-ROC: the ROC score obtained by comparing the positives to the negatives, and PB-ROC: the ROC scores obtained by comparing the positives to the BGCs. We computed the PB-ROC scores with and without the GTEx samples – the results were very similar (e.g., on average the correlation between the genes was 0.9).

Specificity analysis: (Figure SF3) This analysis is a generalization of the test performed in (1). Let (P,S) be a pair of labels such that their disease terms are connected by an edge in the disease ontology structure, where P is the parent and S is the child, and both P and S are labels defined on the same tissue term. To test the specificity of the scores in v with respect to the child S we compare the scores of the samples of S to the scores of the samples of P that are not of S (i.e., samples from the siblings of S including P*) and quantify the separation using a ROC score. This analysis asks whether the v separates S from similar and related disease terms. Thus, it tests for whether v contains S-specific information.

3. Classification analysis and selection of well-classified labels

Each label was estimated using leave-study-out cross validation. For each label, in each fold a set of studies is kept as a test set. An SVM-based classifier is learned on the remaining samples and the predictions on the test set are kept. After all folds are completed, each label has a vector of predictions over all subjects. We then use the univariate scores explained in section 2 to measure the classification performance as follows.

Our tested classifier was based on linear SVM using LiblineaR (4). Our previous work showed that SVM-based linear classifiers are both simpler, faster, and more accurate than other alternatives (1). Briefly, for each learning task (i.e., learning a classifier for a specific label) we learn a set of 50 SVM classifiers on randomly subsampled datasets. Prediction on a new sample is done by

averaging the predictions of the 50 classifiers. This analysis provided extremely fast and robust results.

3.1 Well-classified diseases

We designated a disease *well-classified* if it had PB-ROC > 0.7, PN-ROC > 0.7 specificity > 0.7m SMQ < 0.1 and replicability > 0.5. 68 labels were selected using these criteria: 13 control labels and 55 disease-related labels.

4. Scoring a gene set for DO enrichment

In this enrichment analysis we take a user-given gene list L and a well-classified disease label and test for enrichment of L in our predefined gene ranking based on our PN-ROC scores for the label. We use Wilcoxon rank-sum test to test for enrichment. The analysis outputs an enrichment p-value and a direction for each well-classified label. Thus, users can take a set of genes and link it (if it turns as significant) to out phenotypes using a simple online graphical user interface. This analysis can be seen as a revered GSEA: we take our gene rankings as static object and use the provided gene list as the input for the enrichment analysis.

5. Cancer mutated genes

In our network visualization we mark genes that are associated with cancer subtypes based on somatic mutation data. The gene-cancer subtype associations were taken from three sources: (1) a meta-analysis done by Lawrence et al. (5), (2) the pan-cancer analysis of Tokheim et al. (6), and (3) our previous work (7). For each cancer subtype in ADEPTUS we added all genes that are associated with it or with an ancestor of it (in the DO DAG) that appear in either of the resources above. When presenting the melanoma network we noticed an omission in the DO hierarchy in (7), and added a link from melanoma to the parent term “integumentary system cancer” in this analysis .

6. Architecture and technologies

The server side (i.e. the backend) of the web-tool was implemented using the python flask package (http://flask.pocoo.org/). Flask manages the web tool logic using standard python commands. Whenever needed, we run R scripts from our python code.

The client side is based on basic html and JavaScript. The main JavaScript package is Cytoscape.js (8), which provides an easy to use API for interactive graphs visualization.

References

1. Amar D, Hait T, Izraeli S, Shamir R (2015) Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Res* 43(16):7779–7789.

2. Weinstein JN, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45:1113–20.

3. GTEx Consortium TGte (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45(6):580–5.

4. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: A Library for Large Linear Classification. *J Mach Learn Res* 9:1871–1874.

5. Lawrence MS, et al. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484):495–501.

6. Tokheim C, Papadopoulis N, Kinzler KW, Vogelstein B, Karchin R (2016) Evaluating the Evaluation of Cancer Driver Genes. *bioRxiv* 113(50):60426.

7. Amar D, Izraeli S, Shamir R (2017) Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications. *Oncogene*:36, 3375–3383.

8. Franz M, et al. (2015) Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics* 32(2):309–311.