Title

The Molecular Signatures Database (MSigDB) hallmark gene set collection.

Permalink

https://escholarship.org/uc/item/4mk8145t

Journal

Cell systems, 1(6)

ISSN

2405-4712

Authors

Liberzon, Arthur
Birger, Chet
Thorvaldsdóttir, Helga
et al.

Publication Date

2015-12-01

DOI

10.1016/j.cels.2015.12.004

Peer reviewed

Published in final edited form as:
Cell Syst. 2015 December 23; 1(6): 417–425. doi:10.1016/j.cels.2015.12.004.

The Molecular Signatures Database (MSigDB) hallmark gene set collection

Arthur Liberzon[1], **Chet Birger**[1], **Helga Thorvaldsdóttir**[1], **Mahmoud Ghandi**[1], **Jill P. Mesirov**[1,2,3,4,*], and **Pablo Tamayo**[1,2,3,4,*,1]

[1] Broad Institute of MIT and Harvard, 415 Main St. Cambridge, MA 02142, USA

[2] Department of Medicine, UC San Diego, La Jolla, CA 92093, USA

[3] Moores Cancer Center, UC San Diego, La Jolla, CA 92093, USA

Abstract

The Molecular Signatures Database (MSigDB) is one of the most widely used and comprehensive databases of gene sets for performing gene set enrichment analysis. Since its creation, MSigDB has grown beyond its roots in metabolic disease and cancer to include >10,000 gene sets. These better represent a wider range of biological processes and diseases, but the utility of the database is reduced by increased redundancy across, and heterogeneity within, gene sets. To address this challenge, here we use a combination of automated approaches and expert curation to develop a collection of "hallmark" gene sets as part of MSigDB. Each hallmark in this collection consists of a "refined" gene set, derived from multiple "founder" sets, that conveys a specific biological state or process and displays coherent expression. The hallmarks effectively summarize most of the relevant information of the original founder sets and, by reducing both variation and redundancy, provide more refined and concise inputs for gene set enrichment analysis.

Abstract

[4] Corresponding authors jmesirov@ucsd.edu, ptamayo@ucsd.edu.

[*] These authors contributed equally.

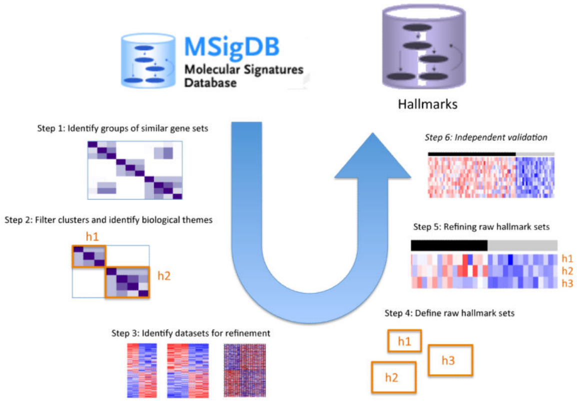
Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DISCLOSURE DECLARATION

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

TO BE ADDED AT PROOFS


MSigDB
Molecular Signatures Database
Hallmarks
Step 1: Identify groups of similar gene sets
Step 2: Filter clusters and identify biological themes
h1
h2
Step 3: Identify datasets for refinement
Step 4: Define raw hallmark sets
h1
h3
h2
Step 5: Refining raw hallmark sets
h1
h2
h3
Step 6: Independent validation

Keywords

gene expression; gene set enrichment analysis; gene sets

INTRODUCTION

High-throughput technologies, such as microarrays and next generation sequencing, generate measurements of gene activity at genomic scale. For transcription profiling, these technologies report transcript abundances for tens of thousands of genes. Analysis of this type of data usually follows one of two approaches. The first identifies genes that are differentially expressed across phenotypes of interest. This is straightforward to perform, but in practice it leads to challenges in the follow-up analysis and interpretation of results. For example, in some instances only a few genes reach statistical significance and the analysis may not produce meaningful results. Alternatively, when a large number of genes pass a significance threshold, there may be no obvious way to select the most interesting genes to follow up. Moreover, the resulting list of genes may be difficult to interpret and to identify the relevant biological process that those genes represent. An alternative approach, pioneered by Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005), focuses on coordinated differential expression of annotated groups of genes, or gene sets, and produces results that can more easily be interpreted in terms of the relevant biological processes. Since its introduction, the use of GSEA has become widespread and has motivated the development of many similar approaches (reviewed in Huang et al., 2009a) and even novel statistical methods based on groups of variables (Efron, 2010; Good, 2011). Over the last decade GSEA has proven a very successful approach in many fields of biomedical research and has become an essential part of the genomic analysis toolbox.

The Molecular Signatures Database (MSigDB) (Liberzon et al., 2011), originally developed for use with GSEA and now employed by many similar approaches, remains one of the largest and most popular repositories of gene sets. The latest version of MSigDB consists of seven collections C1-C7 which include: genes grouped by their location in the human genome (C1), canonical pathways and experimental signatures curated from publications (C2), genes sharing *cis*-regulatory motifs up- or downstream of their coding sequences (C3), clusters of genes co-expressed in microarray compendia (C4), genes grouped according to gene ontology (GO) categories (C5), signatures of oncogenic pathway activation (C6), and a large collection of immunological conditions (C7). All of the gene sets in MSigDB are reviewed, curated, and annotated manually by the MSigDB curator. They are all represented as lists of human gene symbols from the HUGO Gene Nomenclature Committee at the European Bioinformatics Institute (Gray et al., 2015).

The usefulness of GSEA and other gene-set-based analysis methods depends on the availability of independent compendia of gene sets such as MSigDB. The growth of these compendia over time can provide the benefits of better representation and coverage of biological processes but it can also pose new challenges. These challenges derive from the intrinsic redundancy and heterogeneity associated with a larger universe of gene sets.

Redundancy can take different forms, e.g. gene sets may simply share a large proportion of their comprising genes. Another more subtle form of redundancy can occur when gene sets have only a partial overlap but their annotations refer to similar or the same biological process. In the latter case, the gene sets may actually represent partial transcriptional readouts of the same processes, and in both cases the sets may attain similar GSEA. As a consequence of this redundancy, gene set enrichment analysis could produce long lists of statistically significant results with multiple occurrences of essentially the same biological process. Moreover, many high scoring, but overlapping or redundant, gene sets can dominate the top of a result set and effectively hide other potentially relevant hits further down the list. In this scenario one can easily fail to notice important and relevant findings and thus not realize the full potential of GSEA. In addition, the overrepresentation of a biological process at the top of a gene set list can skew the tail of the *observed* distribution of enrichment scores, thereby increasing the significance of top scoring gene sets that represent the same signal.

A second challenge stems from heterogeneity within a gene set. For example, genes in a given gene set do not always behave consistently or coherently. This could be due to several causes: variation because of context dependencies, the existence of multiple modalities of biological response, intrinsic variation in the original dataset from which the gene set is experimentally or computationally derived, limitations of manual curation, or poor biological resolution with respect to the relevant biological process.

Here we present a new MSigDB collection of “hallmark” gene sets and show how it can help to overcome these challenges. These hallmark gene sets are generated by a hybrid approach that combines an automated computational procedure with manual expert curation. The computational methodology identifies gene set overlaps and generates coherent representatives of them. The manual curation makes critical use of domain expert knowledge in order to: i) assign biological themes to groups of the original overlapping gene sets, ii) identify expression data for refinement and validation of the hallmark signatures, and iii) properly annotate the refined hallmarks. The hallmarks summarize information across multiple gene sets by emphasizing genes that display coordinate expression and represent well-defined biological processes, thereby reducing variation and redundancy, and providing a better delineated biological space for GSEA analysis.

RESULTS

Generating the Hallmark Collection

Here we give an overview of the hallmarks generation procedure (see Methods for details). We first identified groups of similar gene sets according to their individual gene membership overlaps using consensus clustering. Starting with 8,380 gene sets from MSigDB v4.0 collections C1-C6, the consensus clustering grouped them into 600 clusters. We manually reviewed the clusters and were able to annotate 43 of them with 50 clear biological themes. While 36 clusters had only one theme assigned to them, seven clusters were assigned to two themes due to the heterogeneity of their founder gene sets (see Supplemental Experimental Procedures, Note 1 for details). These themes, and their associated clusters, served as candidates for an initial collection of hallmark signatures. We defined “raw” sets, one for each candidate hallmark, as the union of a cluster’s gene sets. We refined each of these raw sets according to its gene expression profile in a number of datasets relevant to the corresponding biological theme. The refinement excluded genes that did not well discriminate the relevant phenotype. In this way, only coordinately expressed and biologically relevant genes remained in the final hallmark to be added to the collection. An additional validation procedure determined whether the final hallmark generalized, i.e., performed as expected in an independent dataset that was not used for the refinement. Founders for the final set of 50 hallmarks (Table 1) comprise 4,022 of the original 8,380 MSigDB gene sets.

Examples using the Hallmark Collection

In this section, we give three examples that illustrate different aspects of the use of the hallmark collection for GSEA. The first example shows how using only the hallmarks can form the basis for a concise and sensitive comparison between subtypes of medulloblastoma. It also illustrates the type of summarization that hallmarks provide for their founder sets. The next example necrosis in glioblastoma, demonstrates the redundancy reduction gained from using the hallmark collection instead of the original MSigDB collections C1 – C6. Finally, in the last example we show that hallmarks can effectively associate with their corresponding protein activation phenotypes thus confirming their biological relevance.

Analysis of Hedgehog signaling in medulloblastoma

Medulloblastomas comprise a diverse group of malignant tumors of the cerebellum and are the most common pediatric brain cancers (Northcott et al., 2012). In 2011, Cho et al. analyzed transcriptomes of a collection of 189 primary medulloblastoma tumors. Unsupervised clustering of this dataset identified six distinct molecular subgroups of medulloblastoma, including one that is driven by the tumorigenic activation of the Hedgehog pathway (Cho et al., 2011). Here we considered only the samples of the Hedgehog subtype. As a “control”, we used samples of another subtype from this dataset where the relevant oncogenic process is not Hedgehog signaling but rather photoreceptor activation and *GABRA5* up-regulation. We projected the samples from these two subtypes into the space of the 50 hallmarks by means of single sample GSEA (Barbie et al., 2009). Single sample GSEA (ssGSEA, see Methods) estimates the degree of enrichment of gene sets in individual samples. Then we estimated the degree of association between each hallmark's ssGSEA profile and the Hedgehog vs. photoreceptor activation phenotypic distinction using the Information Coefficient (IC, see Methods). In fact, any standard method for differential analysis could be used for this purpose. An empirical sample-label permutation test (500,000 random permutations) was used to estimate the statistical significance of the scores (*i.e.*, the p-value and False Discovery Rates (FDR) in Figure 1A). Notably, if we rank the hallmarks according to their association with the Hedgehog versus photoreceptor phenotypes then the top-scoring hallmark, *i.e.*, the one with the highest IC value (IC: 0.7346, p-value: 2×10^{5}, FDR:1×10^{5}), is indeed the Hedgehog hallmark, consistent with prior findings in the literature (Cho et al., 2011). A heat map of the results along with the hallmarks ranked by IC appears in Figure 1A. Thus we see the sensitivity of the Hedgehog hallmark in detecting its relevant biological process *i.e.*, oncogenic activation of the sonic hedgehog pathway in the context of a disease subtype.

Repeating the analysis with the original Hedgehog hallmark's founder sets, augmented by the hallmark itself, reveals that the hallmark attains the 4th highest IC measure of association with the oncogenic phenotype (Figure S1). The first 3 highest scoring gene sets refer to embryonic development of the nervous system (Figure 1B and Figure S1).

Analysis of necrosis in glioblastoma

Here we show how hallmarks address the problem of gene set redundancy, by illustrating the difference between GSEA performed with the hallmarks and one that uses thousands of gene sets from the MSigDB collections C1-C6. The dataset contains expression data for 200 glioblastoma multiforme (GBM) and two normal brain samples from the Cancer Genome Atlas Research Network (TCGA) (Verhaak et al., 2010). GBM is the most common, most aggressive malignant primary brain tumor in adults (Ostrom et al., 2013). Necrosis, resulting from a limited supply of oxygen and nutrients, is a critical diagnostic feature of GBM (Karsy et al., 2012). We performed standard GSEA on this dataset with the MSigDB v4.0 collections C1-C6 using the samples' clinical annotation of percentage of necrosis as a continuous phenotype, the Pearson correlation as the ranking metric, and 1,000 permutations of sample labels to estimate significance.

The analysis yielded 527 significantly enriched gene sets that were positively correlated with necrosis (FDR < 0.25, Table S1). Upon inspecting these 527 significant gene sets we were able to assign 11 biological themes to 245 of them. Figure 2 shows the ranks of those annotated gene sets according to their enrichment scores (NES values) grouped by their corresponding biological theme on the left side.

The 100 top scoring gene sets show numerous instances of three biological processes: NFκB signaling, EMT (epithelial-mesenchymal transition), and hypoxia/glycolysis. Indeed, 62 out of the 100 top scoring sets account for 27 NFkB, 25 EMT and 10 hypoxia/glycolysis annotations (Figure 2). The strong presence of EMT and NFκB signatures in association with necrosis agrees with the observations made in the original GBM study. However, an additional group of 8 biological themes (shown inside a red box in Figure 2) only appear below rank 97 in the ranked results list and are thus eclipsed. For example, the original study noted deregulation of the p53 pathway as an underlying relevant biological theme of GBM. Our analysis finds significant enrichment of 15 gene sets corresponding to the p53 pathway.

However, the first such gene set appears only at rank 114 and thus, despite being statistically significant, would more likely be overlooked in a routine interpretation of the GSEA results.

Repeating GSEA with the 50 hallmarks finds 12 that are significantly enriched (FDR < 0.25, Table S2). The three top hallmarks correspond to the biological themes in the top 100 full MSigDB analyses, but the hallmark collection is more sensitive and also highlights the other eight biological processes. Notably, the hallmarks not only produce more parsimonious but equivalent results representing the main biological themes found in the prior analysis, but they also avoid the problem of gene set redundancy and over-representation altogether.

Matching hallmark enrichment scores to protein level phenotypes

We sought biological validation that the hallmark gene sets are able to detect their annotated processes by measuring their performance against experimental data from established protein reporters of pathway activation. For this we used a subset of the Cancer Cell Line Encyclopedia (CCLE) gene expression dataset (Barretina et al., 2012). Besides gene expression, the CCLE repository (www.broadinstitute.org/ccle) maintains detailed genomic, proteomic, and pharmacologic records for about 1,000 cancer cell lines. We projected the CCLE gene expression dataset onto hallmark gene sets using ssGSEA. To define protein abundance phenotypes we utilized the CCLE reverse phase protein array (RPPA) data. RPPA is an antibody-based assay that quantifies expression of proteins and allows concordant interrogation of multiple proteins in many samples (Spurrier et al., 2008). For a large panel of CCLE cell lines, we obtained the RPPA abundances of 8 proteins: AR, BCL2, CDH2 (N-cadherin), ESR1, KDR (VEGFR2), MYC, SMAD3, STAT5A, and a variant of STAT3 phosphorylated at Tyr^{705} (STAT3_pY705). Next, we matched 9 relevant ssGSEA hallmark profiles to those phenotypes using the IC, and assessed the significance of their matching scores using an empirical permutation test as above (2,083,333 permutations). Figure 3 (A through I) shows the results. We observe that all the relevant hallmark enrichment profiles display a high degree of association against the corresponding protein profiles (IC scores and p-values):

MYC protein (Figure 3A)—Cell lines with high levels of MYC protein are associated with high ssGSEA scores for both MYC hallmarks (MYC targets V1: IC = 0.552, $p < 4.8\times10^{-7}$; MYC targets v2: IC = 0.464 $p < 4.8\times10^{-7}$).

Estrogen Response (Figure 3B)—There is a strong association of ESR1 protein expression with the ssGSEA scores of both hallmarks for estrogen response (IC = 0.518, $p < 4.8\times10^{-7}$).

Androgen Receptor (Figure 3C)—Androgens are a group of steroid hormones that regulate the development and maintenance of male characteristics (Matsumoto et al., 2013). Accordingly, the highest ssGSEA score of the hallmark set denoting androgen response corresponds to high AR protein levels (IC = 0.429, p = 0.0002).

BCL2 and Apoptosis (Figure 3D)—BCL2 blocks apoptosis and consequently its protein levels display strong association (IC = 0.476, $p < 4.8\times10^{-7}$) with a proliferation signature represented by the E2F hallmark (Topham and Taylor, 2013). The BCL2 profile has also a strong negative association with the apoptosis hallmark (IC = −0.588, $p < 4.8\times10^{-7}$).

N-cadherin and the Epithelial-Mesenchymal Transition (Figure 3E)—N-cadherin is a marker of mesenchymal cells (Zeisberg and Neilson, 2009) and the epithelial-mesenchymal transition (EMT) (Kalluri and Weinberg, 2009). This figure shows a strong correlation between N-cadherin protein levels and ssGSEA scores of the EMT hallmark (IC = 0.560, $p < 4.8\times10^{-7}$).

SMAD3 and TGF-β (Figure 3F)—TGF-β interacts with TGF-β receptors and leads to phosphorylation of SMAD2 and SMAD3 proteins (Akhurst and Hata, 2012). The profile of SMAD3 protein expression matches activity of the TGF-β hallmark (IC = 0.396, $p < 4.8\times10^{-7}$).

STAT3 and Interleukin-6 (Figure 3G)—Interleukin-6 (IL6) binds a cytokine receptor and triggers a signal transduction cascade through Janus kinases (JAK) that culminates in phosphorylation of STAT3 on the Tyr^{705} (Kaptein et al., 1996; Stark and Darnell, 2012). There is a strong correlation between high levels of the phosphorylated STAT3 (STAT3_pY705) and enrichment of the IL6 JAK STAT3 hallmark (IC = 0.422, $p = 2.38\times10^{4}$).

STAT5 and heme metabolism (Figure 3H)—STAT5 is a member of JAK/STAT signaling network (Stark and Darnell, 2012). During erythroid differentiation, erythropoietin activates STAT5, which in turn activates transcription of genes defining erythroid lineage (Ferbeyre and Moriggl, 2011). Consistent with this relationship, the heme metabolism hallmark (IC = 0.494, $p < 4.76\times10^{-7}$) attains high scores against the profile of STAT5A. Activation of STAT5 by IL2 in T lymphocytes, on the other hand, turns on expression of a different group of genes that play a role in a variety of immune responses. Accordingly, the hallmark for IL2 STAT5 signaling pathway is also associated with the STAT5A protein (IC = 0.368, $p = 2.44\times10^{-4}$).

VEGF and angiogenesis (Figure 3I)—KDR is a receptor of VEGF growth factor. As such, it is a key regulator of blood vessel formation (Folkman and D'Amore, 1996). We observe strong correlation between KDR protein levels and activity of the angiogenesis hallmark (IC = 0.580, $p < 4.76\times10^{-7}$).

DISCUSSION

Here we introduce a collection of hallmarks, along with a methodology to generate them, and demonstrate their utility in several examples. The hallmark generation method of gene overlap yielded groups of gene sets with coherent annotation and thus eventually produced hallmarks that represented the relevant signal in related and potentially redundant gene sets. Because gene sets often convey approximate and incomplete versions of the pertinent biological conditions, we developed a hybrid approach, which combined computational and manual steps. The automated steps included clustering, microarray data processing and meta-analysis. Expert human biological review was essential to leverage prior domain knowledge for labeling clusters with biological themes because the automated clustering methods do not provide a sense of the degree of biological resolution represented by the clusters. Additional manual tasks, also requiring an experienced curator, included locating microarray datasets and annotating their phenotype classes. The refinement methodology allows the hallmark to contain the most transcriptionally coherent set of genes, which serve as more effective and accurate transcriptional signatures for detecting specific biological processes. By summarizing relevant information from thousands of founder gene sets across diverse collections in MSigDB, hallmarks greatly reduce redundancy and produce more robust and concise GSEA results that facilitate interpretation and follow-up analysis, as well as substantially reduce the chances of missing potentially important findings. The fifty hallmarks described here represent 48% (4,022 out of 8,380) of MSigDB gene sets. The hallmarks also capture 52.7% (452 out of 858) gene sets from the C4 collection of co-expression modules, which cover the global landscape of the transcriptome.

The hallmarks are freely available as the H Collection in the MSigDB v5.0 and can be used by any enrichment analysis method that relies on gene sets. The name of every gene set in this collection starts with HALLMARK to distinguish them from other MSigDB gene sets. Table 1 lists all 50 hallmarks by their names and provides their biological process category, a brief description and additional statistics such as their size and their number of founder gene sets. Table S3 contains detailed statistics for the gene sets used to make the hallmarks and their cluster assignments. Each hallmark set and all of its annotations appear on a separate web page of the MSigDB web site (see example in Figure S2). This page also provides a brief description of each hallmark and follows the standard conventions for MSigDB gene sets. In addition every hallmark page contains links to its founder gene sets and includes details about specific datasets and phenotype class comparisons that were used to refine and validate the hallmark. This information is particularly useful when interpreting enrichment analysis results and in follow up studies. In order to take full advantage of hallmarks and the exploratory nature of GSEA, we recommend proceeding through a series of GSEA analysis stages (see Box 1).

We view this group of 50 hallmarks as an initial set deriving from the gene set clusters where the relationship to a biological theme was clear during manual review. Notably, this first set already corresponds to a broad coverage of cellular processes representing about half of the gene sets in the MSigDB. We plan to move forward with a program to enhance and expand the collection, encouraged by the current results that demonstrate an increase in signal strength and the good summarization capability of the hallmarks. We believe this collection will prove to be a valuable user resource for the community and provide even more precise results when used with enrichment analysis methods.

Experimental Procedures

Hallmark generation methodology

Step 1: Identify groups of similar gene sets using consensus clustering—We first clustered all the gene sets according to their member genes' overlaps and regardless of their annotations. We used consensus clustering (Monti et al., 2003) with bootstrap resampling to allow a more robust determination of cluster stability for multiple values of *k*, the ultimate number of clusters. In order to find the optimal number of clusters, we inspected the cophenetic coefficient as a function of k and searched for a peak value indicating the most stable partition (Brunet et al., 2004). We avoid choosing solutions with high values of k that produce higher values of the cophenetic coefficient but potentially overfit and represent small numbers of gene sets in each cluster. The extreme of this behavior is, for example, when the number of clusters equals the number of items, and the fit becomes perfect.

The input dataset to this procedure consisted of 8,380 gene sets from MSigDB v4.0 (collections C1 through C6) each containing between 5 and 1,994 genes (features). We decided to include the C1 collection containing genes in cytogenetic bands because these often indicate regions of similar chromatin structure, or regions affected by oncogenic copy number alterations, which could result in co-regulation and may be important in development and cancer related datasets. We used agglomerative hierarchical clustering with average linkage as implemented in the *fastclust* R package (Müllner, 2013). For the clustering distance metric we used the Jaccard's distance (Jaccard, 1902; Levandowsky and Winter, 1971). For two gene sets S_1, S_2 the Jaccard distance is:

$$\boldsymbol{D}_{12}=1\frac{|\boldsymbol{S}_1 \cap \boldsymbol{S}_2|}{|\boldsymbol{S}_1 \cup \boldsymbol{S}_2|} \quad (1)$$

where $|S_1 \cap S_2|$is the number of elements in the intersection of $S_1 \quad \text{and} \quad S_2, \quad \text{and} \quad |S_1 \cup S_2|$ is the number of elements in the union of the sets.

The bootstrapping resampling procedure for consensus clustering involved sampling with replacement from a pool of 31,847 genes comprising the union of all the 8,380 original gene sets. We performed 100 resampling iterations and carried out consensus clustering for $50 \leq k \leq 8{,}000$ in increments of 50. We used cophenetic coefficients (ρ) of the consensus clustering results to estimate the optimal number of clusters. The cophenetic analysis showed two peaks: one at k = 450 (ρ = 0.9668) and another at k = 600 (ρ = 0.9670, Figure S3). After inspecting results for both values of k, we found the partition with k = 450 to be too coarse and heterogeneous for our purposes. On the other hand, clusters made with k = 600 seemed to be at the level of granularity that was more appropriate for making hallmark sets. We therefore chose the partition at k = 600 to produce clusters of gene sets for the subsequent steps in the hallmark methodology.

Step 2: Filter clusters and identify biological themes—After initial manual assessment, we excluded some of the clusters from further consideration based on their small size in terms of number of genes or gene sets. We left out clusters that had fewer than 150 genes total when we merged the genes in all their member gene sets to allow for sufficient number of genes for subsequent refinement by meta-analysis. We also removed clusters with fewer than six gene sets as the smaller clusters usually lacked sufficient information (in terms of descriptions of and overlaps among their constituent gene sets) to deduce meaningful biological theme. The filtering left 168 clusters for the next stage.

Upon manual inspection of these clusters, we assigned biological themes to 73 clusters where the theme was clearly identifiable. To do this, we relied on the MSigDB annotations of both the gene sets, in the form of their names and descriptions, and their individual genes. In a number of cases, we also used complementary annotation tools such as *DAVID* (Huang et al., 2009b) to obtain additional clues about the most relevant pathways represented by them.

Step 3: Identify gene expression datasets for refinement—We queried the GEO (Barrett et al., 2007) and ArrayExpress (Rustici et al., 2013) to find relevant human, mouse or rat expression datasets for each of the 73 clusters of the previous step. Each dataset was required to contain at least 3 samples in each phenotypic class. We also verified that these datasets were not used to define any of the hallmarks. For subsequent steps we chose 43 of these clusters, for which we identified at least 3 datasets for refinement and a fourth independent one for validation. These 43 clusters were annotated with 50 biological themes. Seven clusters gave rise to two due to heterogeneity within founding gene sets (Supplemental Experimental Procedures, Note 1 and Table S4). We plan to continue assigning themes and processing the remaining clusters to develop additional hallmarks in the future.

Step 4: Define raw hallmark sets—We defined raw hallmarks for each of the 50 hallmarks produced by the previous step. A raw hallmark is the union of a cluster of founder gene sets' genes after excluding all the "unknown genes". We considered a gene as "unknown" if it has been identified exclusively by automatic computational predictions or represented a poorly documented sequence such as an EST. Specifically, we defined a gene as "unknown" if its official gene symbol (according to NCBI Entrez and HUGO) matched naming conventions of an EST (e.g., "KIAA", "LOC", "MGC", "FLJ", or "DKFZp" followed by digits).

Step 5: Refining raw hallmark sets—We assessed how well each gene in each raw hallmark discriminated the relevant phenotypes in each of the datasets identified in step 3. We again used the IC between the phenotype or class vector and the gene expression profiles as the discrimination metric. We assessed the statistical significance of each gene's IC score and produced nominal p-values using a sample permutation test to create an empirical null distribution. This was done independently for each gene expression test dataset. A meta-analysis produced summary p-values across these datasets using Fisher's method (Fisher, 1948) as implemented in R package *MetaDE* (Wang et al., 2012). We used summary p-values to compute False Discovery Rates (FDR) following the approach of (Benjamini and Hochberg, 1995). The genes in the raw hallmark were then sorted by their FDR values and the top scoring genes with summary FDR values less than 0.01 comprised the final hallmark set. When the number of genes obtained by this method was less than 15 (or more than 200) the top scoring 15 (or 200) genes were chosen regardless of their FDR values. Thus the refined hallmarks consist of at least 15 and at most 200 genes, which is the recommended size for use with GSEA. In the refinement procedure we focused on up-regulated genes and used one-tailed tests. The rationale for this stems from our empirical observation that expression patterns of down-regulated genes are often context dependent and tend to generalize poorly across datasets, while up-regulated genes are more consistent.

Step 6: Independent validation and final hallmark set—The final hallmark set consists of genes that, at the same time, represent multiple gene sets and also display coherent, discriminating behavior across a number of test datasets. Every hallmark was validated on at least one additional independent dataset. The validation consisted of computing the ssGSEA scores for the hallmark in the independent datasets and confirming that the nominal p-value of the IC score of the hallmark vs. target phenotype was less than 0.05.

Procedures for computing Jaccard's coefficients, microarray data processing, the information coefficient metric, gene set enrichment analysis, and data visualization are in **Supplemental Experimental Procedures**.

Box 1

Practical Guidelines for using the MSigDB Hallmark Gene Set Collection

I. Establish a "bird's-eye view" of the data by carrying out GSEA using the hallmark collection. Because the hallmarks have been carefully generated and tested, a significant result in this analysis should be considered as having high confidence and worthy of additional follow up.

II. Repeat GSEA using the founder gene sets for each of the top scoring hallmarks to explore more specific or detailed findings.

III. Complement the analysis using other sub-collections of MSigDB gene sets for specific purposes. For example the C3 (master regulators/transcription factors) or C6 (oncogenic pathways) collections can provide additional insight via signals not yet represented in the hallmarks.

In some cases, the hallmark gene sets can be useful in generating target profiles to match against genomic variables (e.g., mutation status, copy number alterations, drug sensitivity, etc.) associated with the biological themes that the hallmarks represent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank Rahi Punjabi for assistance in hallmark refinement and Tenley Archer, William Kim, Rajiv Narayan, Scott Pomeroy, and Aravind Subramanian for useful discussions. This project was supported by NIH grants: R01CA154480, R01CA121941, R01GM074024, and U54CA112962.

REFERENCES

Akhurst RJ, Hata A. Targeting the TGFβ signalling pathway in disease. Nat Rev Drug Discov. 2012; 11:790–811. [PubMed: 23000686]

Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009; 462:108–112. [PubMed: 19847166]

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–607. [PubMed: 22460905]

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Research. 2007; 35:D760–D765. [PubMed: 17099226]

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995; 57:289–300.

Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences. 2004; 101:4164–4169.

Cho Y-J, Tsherniak A, Tamayo P, Santagata S, Ligon A, Greulich H, Berhoukim R, Amani V, Goumnerova L, Eberhart CG, et al. Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J. Clin. Oncol. 2011; 29:1424–1430. [PubMed: 21098324]

Efron B. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction (Cambridge University Press). 2010

Ferbeyre G, Moriggl R. The role of Stat5 transcription factors as tumor suppressors or oncogenes. Biochimica Et Biophysica Acta (BBA) - Reviews on Cancer. 2011; 1815:104–114. [PubMed: 20969928]

Fisher RA. Combining independent tests of significance. The American Statistician. 1948; 2:30.

Folkman J, D'Amore PA. Blood vessel formation: what is its molecular basis? Cell. 1996; 87:1153–1155. [PubMed: 8980221]

Good P. Analyzing the large number of variables in biomedical imagery: a brief review. J Biopharm Stat. 2011; 21:1094–1099. [PubMed: 22023678]

Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames. rg: the HGNC resources in 2015. Nucleic Acids Research. 2015; 43:D1079–D1085.

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009a; 37:1–13. [PubMed: 19033363]

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009b; 4:44–57. [PubMed: 19131956]

Jaccard P. Lois de distribution florale dans la zone alpine. Bulletin De La Société Vaudoise Des Sciences Naturelles. 1902; 38:27–31.

Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J. Clin. Invest. 2009; 119:1420–1428. [PubMed: 19487818]

Kaptein A, Paillard V, Saunders M. Dominant negative stat3 mutant inhibits interleukin-6-induced Jak-STAT signal transduction. J. Biol. Chem. 1996; 271:5961–5964. [PubMed: 8626374]

Karsy M, Gelbman M, Shah P, Balumbu O, Moy F, Arslan E. Established and emerging variants of glioblastoma multiforme: review of morphological and molecular features. Folia Neuropathol. 2012; 50:301–321. [PubMed: 23319187]

Levandowsky M, Winter D. Distance between sets. Nature. 1971; 234:34–35.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27:1739–1740. [PubMed: 21546393]

Matsumoto T, Sakari M, Okada M, Yokoyama A, Takahashi S, Kouzmenko A, Kato S. The androgen receptor in health and disease. Annu. Rev. Physiol. 2013; 75:201–224. [PubMed: 23157556]

Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning. 2003; 52:91–118. 118.

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34:267–273. [PubMed: 12808457]

Müllner D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. 2013; 53:1–18.

Northcott PA, Jones DTW, Kool M, Robinson GW, Gilbertson RJ, Cho Y-J, Pomeroy SL, Korshunov A, Lichter P, Taylor MD, et al. Medulloblastomics: the end of the beginning. Nat Rev Cancer. 2012; 12:818–834. [PubMed: 23175120]

Ostrom QT, Gittleman H, Farah P, Ondracek A, Chen Y, Wolinsky Y, Stroup NE, Kruchko C, Barnholtz-Sloan JS. CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006-2010. Neuro-Oncology. 2013; 15(Suppl 2):ii1–ii56. [PubMed: 24137015]

Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, et al. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Research. 2013; 41:D987–D990. [PubMed: 23193272]

Spurrier B, Ramalingam S, Nishizuka S. Reverse-phase protein lysate microarrays for cell signaling analysis. Nat Protoc. 2008; 3:1796–1808. [PubMed: 18974738]

Stark GR, Darnell JE. The JAK-STAT pathway at twenty. Immunity. 2012; 36:503–514. [PubMed: 22520844]

Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics. 2007; 23:3251–3253. [PubMed: 17644558]

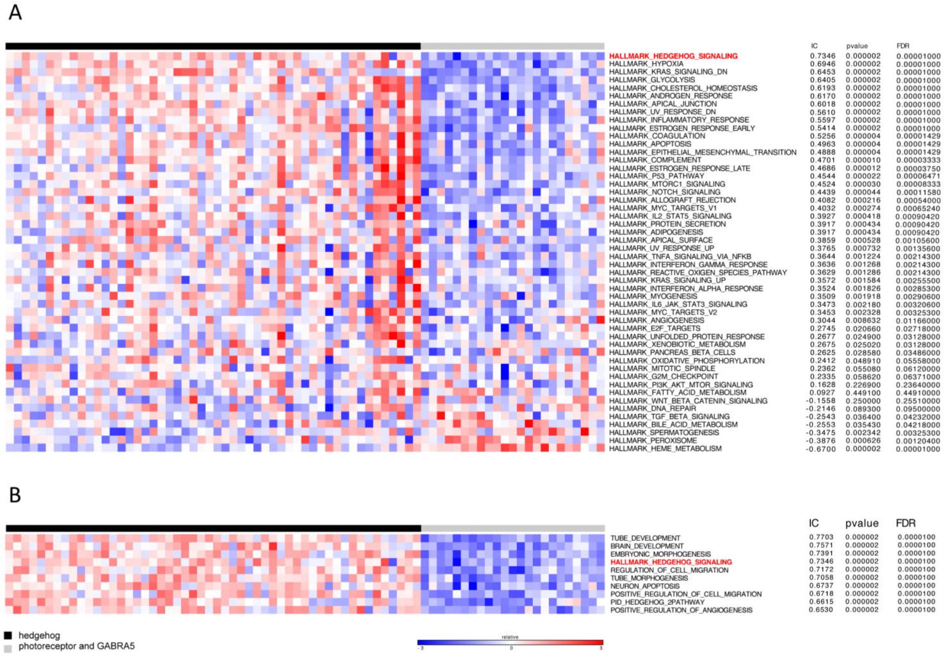
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102:15545–15550.

Topham CH, Taylor SS. Mitosis and apoptosis: how is the balance set? Curr. Opin. Cell Biol. 2013; 25:780–785.

Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]

Wang X, Kang DD, Shen K, Song C, Lu S, Chang L-C, Liao SG, Huo Z, Tang S, Ding Y, et al. An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics. 2012; 28:2534–2536. [PubMed: 22863766]

Zeisberg M, Neilson EG. Biomarkers for epithelial-mesenchymal transitions. J. Clin. Invest. 2009; 119:1429–1437. [PubMed: 19487819]



A

	IC	pvalue	FDR
HALLMARK_HEDGEHOG_SIGNALING	0.7346	0.000002	0.00001000
HALLMARK_HYPOXIA	0.6946	0.000002	0.00001000
HALLMARK_KRAS_SIGNALING_DN	0.6453	0.000002	0.00001000
HALLMARK_GLYCOLYSIS	0.6405	0.000002	0.00001000
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.6193	0.000002	0.00001000
HALLMARK_ANDROGEN_RESPONSE	0.6170	0.000002	0.00001000
HALLMARK_APICAL_JUNCTION	0.6018	0.000002	0.00001000
HALLMARK_UV_RESPONSE_DN	0.5610	0.000002	0.00001000
HALLMARK_INFLAMMATORY_RESPONSE	0.5597	0.000002	0.00001000
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.5414	0.000002	0.00001000
HALLMARK_COAGULATION	0.5256	0.000004	0.00001429
HALLMARK_APOPTOSIS	0.4963	0.000004	0.00001429
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.4888	0.000004	0.00001429
HALLMARK_COMPLEMENT	0.4701	0.000010	0.00003333
HALLMARK_ESTROGEN_RESPONSE_LATE	0.4686	0.000012	0.00003750
HALLMARK_P53_PATHWAY	0.4544	0.000022	0.00006471
HALLMARK_MTORC1_SIGNALING	0.4524	0.000030	0.00008333
HALLMARK_NOTCH_SIGNALING	0.4439	0.000044	0.00011580
HALLMARK_ALLOGRAFT_REJECTION	0.4082	0.000216	0.00054000
HALLMARK_MYC_TARGETS_V1	0.4032	0.000274	0.00065240
HALLMARK_IL2_STAT5_SIGNALING	0.3927	0.000418	0.00090420
HALLMARK_PROTEIN_SECRETION	0.3917	0.000434	0.00090420
HALLMARK_ADIPOGENESIS	0.3917	0.000434	0.00090420
HALLMARK_APICAL_SURFACE	0.3859	0.000528	0.00105600
HALLMARK_UV_RESPONSE_UP	0.3765	0.000732	0.00135600
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.3644	0.001224	0.00214300
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.3636	0.001268	0.00214300
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	0.3629	0.001286	0.00214300
HALLMARK_KRAS_SIGNALING_UP	0.3572	0.001584	0.00255500
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.3524	0.001826	0.00285300
HALLMARK_MYOGENESIS	0.3509	0.001918	0.00290600
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.3473	0.002180	0.00320600
HALLMARK_MYC_TARGETS_V2	0.3453	0.002328	0.00325300
HALLMARK_ANGIOGENESIS	0.3044	0.008632	0.01166000
HALLMARK_E2F_TARGETS	0.2745	0.020660	0.02718000
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.2677	0.024900	0.03128000
HALLMARK_XENOBIOTIC_METABOLISM	0.2675	0.025020	0.03128000
HALLMARK_PANCREAS_BETA_CELLS	0.2625	0.028580	0.03486000
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.2412	0.048910	0.05558000
HALLMARK_MITOTIC_SPINDLE	0.2362	0.055080	0.06120000
HALLMARK_G2M_CHECKPOINT	0.2335	0.058620	0.06371000
HALLMARK_PI3K_AKT_MTOR_SIGNALING	0.1628	0.226900	0.23640000
HALLMARK_FATTY_ACID_METABOLISM	0.0927	0.449100	0.44910000
HALLMARK_WNT_BETA_CATENIN_SIGNALING	-0.1558	0.250000	0.25510000
HALLMARK_DNA_REPAIR	-0.2146	0.089300	0.09500000
HALLMARK_TGF_BETA_SIGNALING	-0.2543	0.036400	0.04232000
HALLMARK_BILE_ACID_METABOLISM	-0.2553	0.035430	0.04218000
HALLMARK_SPERMATOGENESIS	-0.3475	0.002342	0.00325300
HALLMARK_PEROXISOME	-0.3876	0.000626	0.00120400
HALLMARK_HEME_METABOLISM	-0.6700	0.000002	0.00001000

B

	IC	pvalue	FDR
TUBE_DEVELOPMENT	0.7703	0.000002	0.0000100
BRAIN_DEVELOPMENT	0.7571	0.000002	0.0000100
EMBRYONIC_MORPHOGENESIS	0.7391	0.000002	0.0000100
HALLMARK_HEDGEHOG_SIGNALING	0.7346	0.000002	0.0000100
REGULATION_OF_CELL_MIGRATION	0.7172	0.000002	0.0000100
TUBE_MORPHOGENESIS	0.7058	0.000002	0.0000100
NEURON_APOPTOSIS	0.6737	0.000002	0.0000100
POSITIVE_REGULATION_OF_CELL_MIGRATION	0.6718	0.000002	0.0000100
PID_HEDGEHOG_2PATHWAY	0.6615	0.000002	0.0000100
POSITIVE_REGULATION_OF_ANGIOGENESIS	0.6530	0.000002	0.0000100

Figure 1.

Analysis of Hedgehog signaling in medulloblastoma. The figure shows ssGSEA scores ranked by their degree of association (IC) between the Hedgehog and photoreceptor phenotype for: A) the 50 hallmarks and, B) the Hedgehog hallmark and 9 of its top scoring founder gene sets. The IC scores, p-values and FDR's appear on the right side of the heat maps. Black and grey colors denote medulloblastoma subtypes (Hedgehog and photoreceptor subtypes respectively).

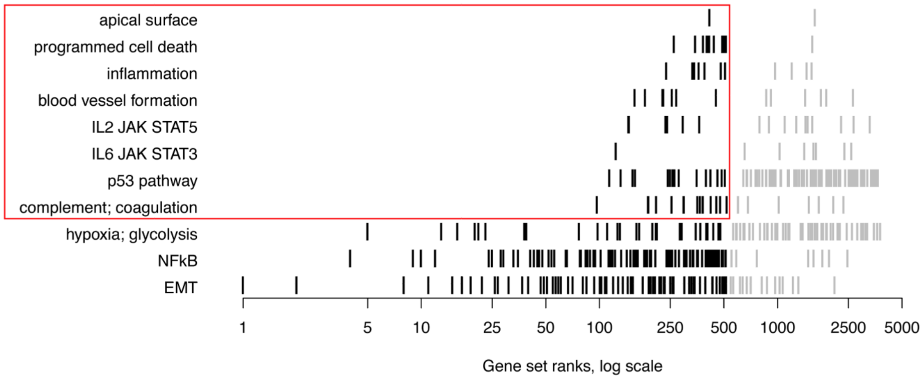


Figure 2.

Ranks of gene sets grouped by biological themes. The horizontal axis denotes rankings of gene sets enriched in the GBM data with respect to necrosis. The biological themes are on the right side of the graph. The vertical bars indicate ranks of gene sets. Black bars denote the 245 significantly enriched sets. Gray bars stand for the gene sets that were not enriched significantly. The uncategorized gene sets are not shown. The rows indicate 11 biological themes. The red box shows gene sets that are pushed down the list by high scoring gene sets representing hypoxia/glycolysis, EMT, and NFkB signaling.

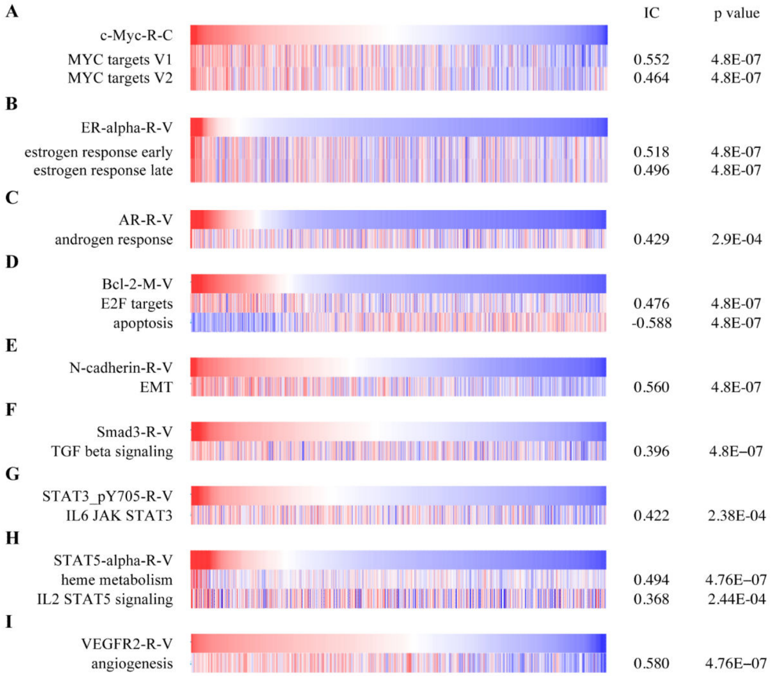


Figure 3.

Matching hallmark enrichment scores to phenotypes defined by protein levels. The top row of the heat maps shows Reverse Phase Protein Array (RPPA) profiles of selected proteins sorted in descending order from left to right. The chosen protein expression profiles are from top to bottom: A) MYC (c-Myc-R-C), B) ESR1 (ER-alpha-R-V), C) AR (AR-R-V), D) BCL2 (Bcl-2-M-V), E) CDH2 (N-cadherin-R-V), F) SMAD3 (Smad3-R-V), G) STAT3 pY705 (STAT3_pY705-R-V), H) STAT5A (STAT5-alpha-R-V) and I) KDR scores.

Table 1

Summary of the hallmark gene sets: name, process category, description, number of founder sets and number of genes it contains.

	Hallmark name	Process category	Description	Number of founder sets	Number of genes
1	APICAL_JUNCTION	cellular component	Apical junction complex consisting of adherens and tight junctions	37	200
2	APICAL_SURFACE	cellular component	Membrane proteins in the apical domain	12	44
3	PEROXISOME	cellular component	Peroxisomes	28	107
4	ADIPOGENESIS	development	Adipocyte development	36	200
5	ANGIOGENESIS	development	Blood vessel formation	14	36
6	EPITHELIAL_MESENCHYMAL_TRANSITION	development	Epithelial mesenchymal transition	107	200
7	MYOGENESIS	development	Muscle differentiation	64	200
8	SPERMATOGENESIS	development	Sperm development and male fertility	24	135
9	PANCREAS_BETA_CELL	development	Genes specific to pancreatic beta cells	24	40
10	DNA_REPAIR	DNA damage	DNA repair	44	150
11	UV_RESPONSE_DOWN	DNA damage	UV response: down-regulated genes	17	144
12	UV_RESPONSE_UP	DNA damage	UV response: up-regulated genes	16	158
13	ALLOGRAFT_REJECTION	immune	Allograft rejection	190	200
14	COAGULATION	immune	Coagulation cascade	71	138
15	COMPLEMENT	immune	Complement cascade	71	200
16	INTERFERON_ALPHA_RESPONSE	immune	Interferon alpha response	82	97
17	INTERFERON_GAMMA_RESPONSE	immune	Interferon gamma response	82	200
18	IL6_JAK_STAT3_SIGNALING	immune	IL6 STAT3 signaling during acute phase response	24	87
19	INFLAMMATORY_RESPONSE	immune	Inflammation	120	200
20	BILE_ACID_METABOLISM	metabolic	Biosynthesis of bile acids	28	112
21	CHOLESTEROL_HOMEOSTASIS	metabolic	Cholesterol homeostasis	28	74
22	FATTY_ACID_METABOLISM	metabolic	Fatty acid metabolism	53	158
23	GLYCOLYSIS	metabolic	Glycolysis and gluconeogenesis	87	200
24	HEME_METABOLISM	metabolic	Heme metabolism	36	200
25	OXIDATIVE_PHOSPHORYLATION	metabolic	Oxidative phosphorylation and citric acid cycle	93	200
26	XENOBIOTIC_METABOLISM	metabolic	Metabolism of xenobiotics	124	200
27	APOPTOSIS	pathway	Programmed cell death; caspase pathway	80	161
28	HYPOXIA	pathway	Response to hypoxia; HIF1A targets	87	200
29	PROTEIN_SECRETION	pathway	Protein secretion	74	96

	Hallmark name	Process category	Description	Number of founder sets	Number of genes
30	UNFOLDED_PROTEIN_RESPONSE	pathway	Unfolded protein response; ER stress	22	113
31	REACTIVE_OXYGEN_SPECIES_PATHWAY	pathway	Reactive oxygen species (ROS) pathway	13	49
32	E2F_TARGETS	proliferation	E2F targets	420	200
33	G2M_CHECKPOINT	proliferation	Cell cycle G2/M checkpoint	420	200
34	MYC_TARGETS_V1	proliferation	MYC targets variant 1	404	200
35	MYC_TARGETS_V2	proliferation	MYC targets variant 2	6	58
36	P53_PATHWAY	proliferation	p53 pathway	85	200
37	MITOTIC_SPINDLE	proliferation	Mitotic spindle assembly	108	200
38	ANDROGEN_RESPONSE	signaling	Androgen response	8	117
39	ESTROGEN_RESPONSE_EARLY	signaling	Early estrogen response	61	200
40	ESTROGEN_RESPONSE_LATE	signaling	Late estrogen response	61	200
41	IL2_JAK_STAT5_SIGNALING	signaling	IL2 STAT5 signaling	13	200
42	KRAS_SIGNALING_UP	signaling	KRAS signaling, up-regulated genes	14	200
43	KRAS_SIGNALING_DOWN	signaling	KRAS signaling, down-regulated genes	16	200
44	MTORC1_SIGNALING	signaling	mTORC1 signaling	487	200
45	NOTCH_SIGNALING	signaling	Notch signaling	49	32
46	PI3K_AKT_MTOR_SIGNALING	signaling	PI3K signaling via AKT to mTORC1	591	105
47	HEDGEHOG_SIGNALING	signaling	Hedgehog signaling	79	36
48	TGF_BETA_SIGNALING	signaling	TGF beta signaling	29	54
49	TNFA_SIGNALING_VIA_NFKB	signaling	TNFA signaling via NFkB	132	200
50	WNT_BETA_CATENIN_SIGNALING	signaling	Cannonical beta catenin pathway	49	42