Pathprinting: An integrative approach to understand the functional basis of disease

Citation
Altschuler, Gabriel M, Oliver Hofmann, Irina Kalatskaya, Rebecca Payne, Shannan J Ho Sui, Uma Saxena, Andrei V Krivtsov, et al. 2013. "Pathprinting: An Integrative Approach to Understand the Functional Basis of Disease." Genome Medicine 5 (7): 68.

Published Version
doi:10.1186/gm472

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:12605447

Genome Medicine

Pathprinting: An integrative approach to understand the functional basis of disease

Genome Medicine 2013, **5**:68 doi:10.1186/gm472

Gabriel M Altschuler (gabrielaltschuler@googlemail.com)
Oliver Hofmann (ohofmann@hsph.harvard.edu)
Irina Kalatskaya (Irina.Kalatskaya@oicr.on.ca)
Rebecca Payne (rebeccapayne@gmail.com)
Shannan J Ho Sui (shosui@hsph.harvard.edu)
Uma Saxena (usukhija@partners.org)
Andrei V Krivtsov (krivtsoa@mskcc.org)
Scott A Armstrong (armstros@mskcc.org)
Tianxi Cai (tcai@hsph.harvard.edu)
Lincoln Stein (lstein@oicr.on.ca)
Winston A Hide (whide@hsph.harvard.edu)

ISSN	1756-994X
Article type	Method
Submission date	8 March 2013
Acceptance date	10 July 2013
Publication date	26 July 2013
Article URL	http://genomemedicine.com/content/5/7/68

Pathprinting: An integrative approach to understand the functional basis of disease

Altschuler G M[1], Hofmann O[1, 2, 5], Kalatskaya I[3], Payne R[1], Ho Sui S J[1, 2], Saxena U[1], Krivtsov A V[4], Armstrong S A[4, 5], Cai T[1], Stein L[3], and Hide W A*[1, 2, 5]

1 Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA
2 Bioinformatics Core, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA
3 Ontario Institute for Cancer Research, Department of Informatics and Bio-computing, MaRS Centre, South Tower, 101 College Street, Toronto, ON, M5G 0A3, Canada
4 Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA
5 Harvard Stem Cell Institute, 1350 Massachusetts Ave, Cambridge, MA 02138
*Corresponding Author:

E-mail addresses:
Gabriel M. Altschuler - gabrielaltschuler@googlemail.com
Oliver Hofmann - ohofmann@hsph.harvard.edu
Irina Kalatskaya - Irina.Kalatskaya@oicr.on.ca
Rebecca Payne - rebeccapayne@gmail.com
Shannan J. Ho Sui - shosui@hsph.harvard.edu
Uma Saxena - usukhija@partners.org
Andrei V. Krivtsov - krivtsoa@mskcc.org
Scott A. Armstrong - armstros@mskcc.org
Tianxi Cai - tcai@hsph.harvard.edu
Lincoln Stein - lincoln.stein@oicr.on.ca
Winston A. Hide - whide@hsph.harvard.edu

Abstract:

New strategies to combat complex human disease require systems approaches to biology that integrate experiments from cell lines, primary tissues and model organisms. We have developed Pathprint, a functional approach that compares gene expression profiles in a set of pathways, networks and transcriptionally-regulated targets. It can be applied universally to gene expression profiles across species. Integration of large-scale profiling methods and curation of the public repository overcomes platform, species and batch effects to yield a standard measure of functional distance between experiments. We show that Pathprints combine mouse and human blood developmental lineage, and develop new prognostic indicators in Acute Myeloid Leukemia. The code and resources are available at http://compbio.sph.harvard.edu/hidelab/pathprint.

Background:

Complex human diseases arise from perturbations of the cellular system [1]. Defining these changes from a systems biology perspective provides the opportunity to relate the function of genes, pathways and processes. The ability to compare experiments across model organisms and humans directly impacts our capacity to determine the basis of disease [2-4] and the importance of cross-species data analysis has been well illustrated: human disease genes have been identified by large scale meta-analysis of conserved human-mouse co-expression [5], gene-based cross-species distance metrics have highlighted diseases that activate similar human and mouse pathways [6], and oncogenic expression signatures have been prioritized by comparing human cancer and mouse model expression profiles [7-9]. Gene expression provides the most extensive resource to profile functional changes, and the opportunity for wide-scale meta analyses has been made possible by the development of public data repositories such as the National Center for Biotechnology Information Gene Expression Omnibus (GEO) [10] and the European Bioinformatics Institute ArrayExpress [11]. Cross-study analysis and integration is an area of highly active research, however, most gene-based approaches are confounded by the challenge of comparing gene activity between different platforms and species. Consistent and scalable methods for combining these data are now required so that researchers can perform comprehensive integration of prior knowledge with new experiments, identify consistent signals, compare heterogeneous data, and validate hypotheses.

Methods for cross-study integration of gene expression data have tended to focus on differential expression in well-matched control and experimental samples [12], as approaches based on correlation or absolute profiles [13] are dominated by lab and platform variability in cross-study analyses [14]. The ability to leverage public data to address platform-effects has been demonstrated most recently by the Gene-Expression Barcode, and Gene Expression Commons, both of which define absolute gene expression scores based on a background distribution [15, 16]. By virtue of their reliance on gene level comparisons, these compelling simplifying approaches are restricted to selected platforms and so do not address global comparison of biological function across experiments and species.

We have sought to develop a new, functional-based, approach for comparing profiles that can truly scale across the diversity of available experiments, platforms and species. Expression of biological functions across batches and divergent expression platforms shows higher concordance than genes [17], and assigning genes to pathways [18-20] or ontologies [21] is effective for revealing phenotype associations [22-25], cross-platform integration [14], and specifying disease subgroups [26]. On this basis, we have developed Pathprint, a global pathway activation map spanning 6 species and 31 array technologies that represents expression profiles as a ternary score (under-expressed {-1}, intermediately-expressed {0}, over-expressed {+1}) in a set of 633 pathways, networks and transcriptionally-regulated targets. The method leverages a static background built from public data repositories, integrating pathway annotation and prediction with large-scale profiling.

Pathprint provides a quantitative definition of cellular phenotype, and a functional distance between all experiments based on their global pathway activity. It presents a significant methodological advance over single-study, relative enrichment methods such as Gene Set Enrichment Analysis (GSEA) [27] and existing gene-based methods for comparison between platforms and species. Pathprinting provides a robust framework for large-scale meta-analysis of clinical data, and allows phylogenetic reconstruction of developmental lineages from a functional perspective. We demonstrate the use of Pathprinting for retrieval of functionally matched samples from cross platform expression databases, reconstruction of the blood developmental lineage across species, and the integration of data from mouse experiments, human samples, and clinical studies to develop new prognostic indicators and drug targets in Acute Myeloid Leukemia.

Methods:

The pipeline to create a Pathprint of an array is shown in Figure 1. A score of 0 in the final Pathprint vector represents pathway expression at a similar level to the majority of arrays of the same platform in the GEO database; scores of 1 and -1 reflect significantly high and low expression respectively. Below we describe the individual steps used to construct the method.

Expression data for building pathway background distributions

A list of arrays from 31 of the most highly represented one-channel gene expression platforms in GEO that profiled *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans* was compiled (please see Additional file 1) and the normalized expression tables retrieved, all normalization methods were accepted. After discarding incomplete records, this list contained 176,971 arrays. It was necessary to restrict the platform coverage to one-channel arrays as two channel arrays provide the relative expression of genes between test and control samples, hindering direct comparison of the test sample between experiments when the control sample differs. The expression data were mapped to Entrez Gene IDs using systematically updated annotations from AILUN [28]. Multiple probes were merged to unique Entrez Gene IDs by the mean expression level. It should be noted that although the mean expression level will produce stable gene expression values it also ‘averages out’ the effects of alternative promoter usage and splice variants. Tissue specific splicing has been recognized as an important factor in defining cellular function [29], however at the present time, there is insufficient data to consistently map individual splice variants to pathways.

Pathway databases

Canonical pathway gene sets were compiled from Reactome [18], Wikipathways [20] and KEGG [19], which were chosen as they include pathways relating to metabolism, signaling, cellular processes, and disease. For the major signaling pathways experimentally derived transcriptionally up and down regulated gene sets were obtained from Netpath [30]. The pathways provide structured relationships between genes, unlike ontologies such as GO [21] that define relationships between but not within terms.

Static Modules

Pathprint is built to leverage expertly curated biological knowledge found in canonical pathway databases in a systematic framework. This approach provides a consistent biological annotation of datasets in terms that are well understood by the community. However, a uniquely pathway-centric approach would introduce an inherent curation bias towards well-studied genes and processes. Therefore, we have supplemented the curated pathways with non-curated sources of interactions by including highly connected modules from a functional interaction network, termed ‘static modules’. This functional interaction network was constructed by extending curated pathways with non-curated sources of information, including protein-protein interactions, gene co-expression, protein domain interaction, Gene Ontology (GO) annotations and text-mined protein interactions. The final functional-interaction network contains 181706 interactions between 9452 genes [31], representing close to 50% of the total human proteome. A Markov Cluster Algorithm was applied to decompose the network, yielding 144 closely related functional interaction clusters, ‘static modules’, ranging from 10 to 743 nodes. Each cluster was named according to the member gene with the highest interaction degree. The modules cover 6458 genes, 1542 of which are not represented in any of the pathway databases. These static modules offer the opportunity to examine the activity of less studied or annotated biological processes, and also to compare their activity to that of the canonical pathways. To provide biological context for the static modules the top GO terms associated with all the pathways have been compiled (please see Additional file 2).

Compiling cross species gene sets

M. musculus, *R. norvegicus*, *D. rerio*, *D. melanogaster* and *C. elegans* gene sets were inferred using homology based on the HomoloGene database [32]. HomoloGene uses pairwise gene comparison combined with a guide tree and gene neighborhood conservation. HomoloGene was selected as, when compared to alternative inference methods, it provides a better functional proxy and higher specificity for the resolution of shared cellular ontogeny, albeit with lower overall coverage [33].

Summary of the Pathprint gene sets

All of the modules and pathways were converted to flat gene sets, so intra-pathway gene-level interaction data are not used. Combined, the canonical pathways, downstream targets, and static modules total 633 human gene sets. The gene membership of these sets is described in Table 1, within the R package Pathprint, and on the Pathprint website, for the number of genes overlapping between each of the data sources please see Additional file 3. Specific pathway sub-subsets may also be used in individual analyses.

Calculating pathway expression

Genes were ranked by expression level, from 1 (low expression) to N (high expression), where N is the total number of genes in the array. For a pathway, P of size k, represented in an array by genes G_1, G_2…G_n, the pathway expression score, En(P), is defined by the mean squared-rank,

$$En(P) = \frac{1}{n} \times \sum_{i=1}^{n} R_i^2$$

Where R_i is the rank of gene G_i in a pathway containing n genes. Rank normalizations provide robust summary statistics to calculate pathway expression scores [6, 13] that can be applied across all technologies, and does not depend on the dynamic range of an array. The mean squared rank was used based on a survey of statistical approaches for gene set analysis [34] and out-performed other summary statistics in a series of classification benchmarks based on tissue-specific pathway expression (see benchmarking section below).

Normalization and probability of expression (POE)

When comparing gene set expression scores between experiments it is essential to assess the expression against a suitable null hypothesis [35]. In this case, comparing the expression of a gene set in one array to its expression in all other arrays, i.e. sample permutation, is required to account for the internal gene expression correlation structure within gene sets, expected to be particularly high within pathways [36]. For each gene set, the expression score was normalized against a background built using all arrays of the same platform type. This is the first study comparing database-wide gene set expression and the expected distribution scores is not known. We adopted a similar approach to the Gene

Expression Barcode [15] that estimates which genes are expressed and which are unexpressed in data from single microarrays. The Gene Expression Barcode converts gene expression levels to binary scores based on a static background distribution built from public expression data for 3 distinct platforms. Here, we constructed static pathway expression background distributions for each pathway across 31 platforms in GEO [10]. Each of these distributions was then fit to two-component uniform-normal mixture model [37]. The normal component represents the core distribution of pathway expression scores for a particular pathway, i.e. not significantly high or low expression. The uniform component represents outlying pathway expression due to significantly high or low expression. A signed Probability of Expression (POE) can be calculated representing the probability that a pathway expression score belongs the uniform component of the fitted mixture model. We took advantage of the increase in computation speed afforded by the expectation-maximization implementation of POE in the R package metaArray [38].

Application of a ternary threshold

High/low thresholds [15], or filters with weight vectors approaching the thresholding limit [6], operate as an effective noise filter to remove uninformative signal variation. POE values were converted to a ternary score by the transformation

$$\begin{aligned} F_i = 1 & \quad T \leq POE_i \\ 0 & \quad -T < POE_i < T \\ -1 & \quad POE_i \leq -T \end{aligned}$$

Where POE$_i$ (i = 1,2...n) represents the POE for gene set i, T is the threshold and the F$_i$ are components of the Pathprint vector. Selection of the threshold, T, is of vital importance as this directly modulates the sensitivity and specificity at which gene sets are scored as significant. Large values of T - high stringency - is appropriate for gene expression [15]. Small values of T - low stringency - increase the weighting of subtle differences in expression, and may be required to discriminate arrays at the pathway level, where the coordinated effects of multiple genes are under consideration. The threshold was optimized by combining multiple benchmarks (see below). Thresholding improves sample clustering (see below), provides a read-out for sample annotation, and simplifies quantification of sample relationships.

Constructing consensus pathprints

To summarize the activity of a group of pathprints we define the consensus score for each pathway as

$$\begin{aligned} C_i = 1 & \quad \mu_i > t \\ -1 & \quad \mu_i < -t \end{aligned}$$

Where μ_i is the mean score for pathway i across the group of pathprints, and t is a consensus threshold value. The consensus pathprint is the vector constructed by calculating the consensus score for each pathway, representing the consistently significantly expressed pathways across the group. The rationale behind introducing a threshold is to associate a set of pathways with a phenotype, and so provide a discrete functional representation of a cell type based on a collection of pathprints.

Defining distance between pathprints

A functional distance between experiments is defined as the distance between two pathprint vectors. We define the distance by the Manhattan distance, providing a simple readout for the number of pathway scores that differ between two samples. We define the distance from a consensus pathprint to any other pathprint by the Manhattan distance between the subset of the pathprint vectors that contain only the pathways for which the consensus pathprint is non-zero. This ensures that only differences in the consistently expressed pathways that make up the consensus pathprint are considered.

Optimizing threshold value

The threshold value was optimized using cross-platform, cross-species gene expression data from a panel of human and mouse tissue samples [15] and an independent dataset profiling brain sub-regions in human, mouse and rat [39]. Four approaches were used to determine the optimum threshold.

i) Cross-validation

The data sets were divided into 5 subsets of equal, or approximately equal, size. One of the subsets (the test set) was omitted and mean pathprints were calculated for each tissue from the remaining samples (the training set). Next, the samples in the test set were assigned to the tissue with the closest mean tissue pathprint in the training set by Euclidean or Manhattan distance (both yielded similar results). An error rate was calculated by comparing these assignments to the known annotations. This was repeated, omitting each of the subsets in turn, to obtain a mean error rate. The cross validation procedure was performed 10 times for each threshold value to estimate the mean and standard deviation of the error rate. The standard deviation was small relative to the change in mean error rate over the thresholds and so this number of repetitions was deemed sufficient (Additional file 4). The procedure was also performed as a leave-one-out cross validation, equivalent to dividing the data into a number of subsets equal to the same number of samples, with similar results.

ii) Cluster validity (intra- vs. inter-tissue distance and principle component analysis)

Cluster validity was determined by the ratio of the intra- to inter-tissue variance, where variance was defined as sum of the squared Euclidean distance between each sample and the mean pathprint for each tissue. A lower ratio indicates tighter clustering within tissues and/or better separation of the tissue type clusters. The clusters formed by pathprints had an intra/inter cluster distance ratio of 0.63, compared to 1.26 for the Barcode and 0.92 for Spearman correlation (Additional file 4).

iii) Retrieval: Precision-Recall of cross-species tissue data

The combined human and mouse dataset was ranked by distance from each sample (*Manhattan*, *Euclidean* or *Spearman* correlation). These ordered retrieval lists were used to calculate average interpolated precision-recall curves at a range of threshold values. Decreasing the stringency of the threshold initially improved performance but at thresholds below 0.001 the difference became less significant, summarized by the plot of mean average precision (Additional file 4). Pathprinting improves the performance of tissue retrieval across species over gene expression measurements (both Barcode and Spearman correlation) and results obtained with randomly constructed gene sets (Figure 2). Pathway expression scores based on the mean squared-rank out-performed the mean rank, as assessed by precision-recall curves for the tissue-species data. In addition, an identical analysis pipeline was also constructed using the GSEA algorithm, as applied to single samples [25], as the initial step, in place of the mean squared rank. It was found that the enrichment scores were highly correlated and yielded no significant improvements in precision or recall. There was also a much greater computational burden associated with running GSEA on 180,000 arrays compared to the mean squared rank.

iv) Comparison to randomly constructed gene sets

A Pathprint based on ‘random’ gene sets was constructed to test whether the ‘expert’ knowledge contained within the pathways and modules contribute to the success of the Pathprint over and above the effect of simply reducing the dimensionality of the data. These random gene sets contained genes sampled without replacement from the genes used in the original pathways and retained the size distribution of the original pathway list. The precision-recall curves for Pathprint based on random gene sets (Figure 2, Additional file 4), demonstrate inferior performance compared to Pathprint. This is especially pronounced at stringent thresholds. At less stringent thresholds, the difference between the curves is smaller, implying that both the reduction in data dimensionality and the integration of biological knowledge contribute to the effectiveness of Pathprint.

A threshold value of 0.001 was chosen on the basis that it performed optimally across the majority of the performance measures. It is interesting to note that a highly stringent threshold, ~0.9, did not perform well in cross-validation but yielded good results for the precision-recall and cluster validity tests, and produced the greatest difference in performance compared to the equivalently thresholded random genesets. These results show that moderate pathway expression levels best characterize samples, but the most highly expressed pathway expression scores are also informative. Further work is required to determine whether combining more than one thresholding regimen would be beneficial.

Phenotype matching using the GEO database

Any set of arrays can be used as a ‘seed’ to construct a consensus pathprint profile representing the commonly expressed functions of the set, e.g. tissue-specific arrays (Figure 2). The distance of every array in the GEO pathprint collection then can be measured to produce a table of GEO samples, ordered according to their phenotypic similarity to the seed set, i.e. a ranked list of retrieved samples (please see Additional file 5).

Distribution of distances

In considering the distribution of distances from a consensus pathprint, a major problem is how to assign a measure of significance. This is particularly important if it is necessary to impose a cutoff at which to evaluate retrieved results. Calculating significance based on the distribution of pathprint scores across the full GEO database is complicated as a) each pathway has a different distribution of ternary scores and b) the pathways scores are known to be correlated. An alternative strategy is to use the distribution of the database to define a background distribution, based on the following assumption. Firstly, that there are two distinct populations, a small number of closely matched and a large number of non-matched samples, and secondly that the distances of the non-matched samples are normally distributed. The estimated distribution of the non-matched samples is derived from the inter-quartile range of the full distribution. The significance with which an array is matched with a pathprint, or with a consensus pathprint, is then calculated using the p-value based on the normal distribution function based on this estimated distribution. This approach is clearly an over-simplification and a more complete significance model will form the basis of further study. We expect a large number of the samples contained in GEO to be disease related, representative of a research focus bias inherent in the scientific literature, and so we are aware that the underlying distribution could be multimodal due to perturbed transcriptional programs and copy number variations associated with disease, and specifically cancer cell types. The correlation between this estimated p-value and the precision for each of the 6 tissue samples is shown in Additional file 5.

Phylogenetic analysis

Pathprints corresponding to hematopoietic gene expression datasets GSE24759 [40] and GSE6506 [41] were calculated using the Pathprint pipeline. A consensus pathprint was constructed for each of cell types using an arbitrarily selected threshold of 0.75. Phylogenetic analysis was performed using the R package Phangorn [42]. Optimized parsimony and (non-parametric) bootstrapped trees were found by nearest neighbor interchange with a cost matrix based on the difference between pathprint scores.

Self-renewal signature and survival analysis

Gene expression data for leukemia stem cell, normal stem cells and progenitor cells in mouse and human were obtained from the GEO (GSE24006 and GSE3722). Pathprints were calculated for each sample using the Pathprint package in R. Pathways shared by leukemic and normal stem cells that are differentially expressed in progenitor cells were identified for the human and mouse datasets. The self-renewal signature (SRAS) was defined as the set of pathways common to the human and mouse signatures. Gene expression arrays and the associated survival data were obtained for 4 clinical studies of acute myeloid leukemia from GEO (GSE10358, GSE12417, GSE1159, and GSE14468). Pathprints were calculated for each sample in these datasets. Survival plots and associated p-values were derived using the Kaplan-Meier method by stratifying patient samples into two groups by the sum of their pathprint scores across the SRAS pathways. For each dataset the approach was repeated 1000 times using random permutations of the pathprint pathways with the same number of member pathways as the SRAS set to produce a background distribution of p-values against which to compare the SRAS result.

Code and Pathprint R package

The code and data to process gene expression arrays to pathprints have been compiled into the R package Pathprint. Pathprints have also been pre-calculated for approximately ~180,000 gene expression profiles from the GEO repository and included in the R package, along with their associated metadata to create a search-able cross-platform matrix covering 31 platforms and 6 species (please see Additional file 1). Future versions of Pathprint will extend the acquisition pipeline to encompass the remaining platforms and incorporate data from other repositories. The package as well as the complete R code (as Sweave documents) required to reproduce the analysis and figures contained within this manuscript are available online[43].

Results and Discussion:

The ability of pathprints to classify cross-platform and species data was tested on a series of tissue specific datasets, and compared to the Gene Expression Barcode [15], gene-expression correlation, and a pathprint based on random gene sets (Figure 2, Additional file 4). In each test, Pathprint improves sample classification, and clusters tissues together across platform and species. The biological and technical variation across pathprints in the tissue-specific dataset was investigated by principal component analysis (Figure 2c). The first two principal components separate most tissue types, irrespective of their originating platform and species, with some convolution of the lung and spleen samples. Notably, a corresponding plot produced from Gene Expression Barcode data clusters samples first by platform and then tissue type (Figure 2d).

A high degree of overlap in gene membership is introduced when combining multiple pathway databases. Overlapping gene membership can be due to redundancy in the pathway sets, for example different views of the Wnt pathway in the Reactome, Wikipathway and KEGG databases, or due to pathways being closely biologically related and so sharing a subset of there genes, such as ‘G1 to S cell cycle control’ and ‘DNA replication’. Overlapping genes will result in correlation between the gene expression scores of these pathways. In addition to the correlation due to overlapping genes, it is well recognized that pathways do not function as discrete elements but are organized into cascades and co-regulatory networks. We have not attempted to make a quantitative definition of the second source of correlation but have tested the effect of correcting for overlapping genes by incorporating a pathway covariance matrix to adjust the contribution of each gene set using the Mahalanobis distance. The covariance matrix was calculated using pathway expression scores from 10,000 randomly permuted expression profiles to providing a measure of the covariance due to the gene member overlap, without the additional complication of gene-gene expression correlations. In the benchmark tests the Mahalanobis distance did not improve performance over the simpler Euclidean and Manhattan distances (Figure 2b). Accordingly, all pathways, irrespective of size and including overlapping gene sets, were retained in the pathprint. No additional correction was made as we wish to maximize the utility of the pathprint as a source of annotation of samples as well as for sample clustering and organization. Plans to include feature selection of gene sets that contribute the most towards performance, for example by non-negative matrix factorization are the subject of ongoing algorithmic development.

We will now outline a series of case studies demonstrating major applications of pathprinting that focus on integrating data from human and mouse.

Tissue-specific pathway profiles

The consensus pathprints derived from the tissue specific datasets described above define consistent functional identities for each tissue, for example skeletal muscle expresses myogenesis, liver and kidney express metabolic pathways, and brain expresses neuroactive ligand receptors (Figure 2e). To validate these tissue-specific pathway combinations, the full GEO matrix of pathprints, approximately 180,000 samples, were ranked based on pathprint distance from each tissue profile. Originating tissue types were assigned for each GEO sample using the metadata in the database allowing validation of the matched samples and the construction of precision-recall curves for each tissue. The results demonstrate remarkable specificity (Table 2, Additional file 5): the 50 human and mouse brain Affymetrix arrays used to build a brain profile retrieved ~8,500 brain samples at 95% precision, spanning 4 species (human, mouse, rat and zebrafish) and 25 different platforms (from Affymetrix, Illumina and ABI). For 5 of 6 tissues over 1,000 correctly matched arrays were retrieved at 95% precision. Although performance is noticeably worse for the spleen, a high proportion of spleen mass is blood, and therefore blood samples, predominantly leukocytes, ranked highly in the retrieval list, lowering the observed precision. We tested the ability of the brain and liver mouse and human consensus pathprints to retrieve samples from each of the other species covered by the pathprint; rat, zebrafish, fruit fly and nematode. The top matches for the brain consensus were all brain samples for rat and zebrafish, head samples for fruit fly and a more heterogeneous set that included neuron samples for nematode. The top samples retrieved by the liver consensus were liver in rat and zebrafish, and whole samples for nematode and fruit fly (please see Additional file 6).

Development of a pluripotent pathprint

The study and characterization of embryonic stem (ES) cells is dominated by subjective choices of selection markers. ES cells express consistent transcriptional profiles that provide benchmarks for pluripotency [44], however to date, it has not been possible to consistently assess ES signatures across all available data and platforms and it is becoming increasingly important to provide biologically interpretable functional signatures that are robust across a range of experimental origins. An ES pathprint was derived from 127 human and mouse samples (please see Additional file 7) that includes high expression of known ES-related functions such as DNA repair, one-carbon metabolism [45] and a network centered on *SUMO1*, the ubiquitin-related modifier thought to target and stabilize Oct4 [46]. The profile is a consistent indicator of pluripotency; 90% of the 1,000 closest pathprint-matched samples in GEO are ES and induced pluripotent stem (iPS) cells from 140 different human and mouse studies and 13 platforms (please see Additional files 8 and 9). The non-ES/iPS samples retrieved were cancer cell lines known to express ES, consistent with the concept that pathways required for stem cell specification play fundamental roles in tissue regeneration and cancer. Systematically profiling stem cells using pathprints to integrate data from mouse models, human primary tissue and clinical studies will resolve the contributions of these stem pathways to developing and aberrant systems and reveal pathways of clinical relevance.

Integration of the human and mouse hematopoietic lineage

Mapping cellular lineages has traditionally relied on direct observation, or by endogenous or genetically engineered markers. Defining cell types using a combination of markers is not always possible, and often the link between marker and cellular function is not understood. Hematopoietic differentiation has been analyzed in the context of the canonical view of blood lineage using gene expression profiles of surface marker purified populations in human [40] and mouse [41]. Pathprinting allows a novel pathway-based phylogenetic approach for an unsupervised definition of this lineage by maximum-parsimony reconstruction using the discrete pathprint states. The reconstruction recapitulates the known lineage ontogeny and allows integration of human and mouse data, using the common informative pathways (Figure 3, Additional file 10). The phylogeny resolves the major myeloid and lymphoid branches independent of species. Species-specific contributions overcome some cell-type groupings, but this is unsurprising as marker selection and immune presentation differ between the experiments. A comprehensive survey of mouse immune-cell gene expression is in progress [47]. As these and further data becomes available, pathprints will allow integration with the existing human and mouse ontogenies, providing functional differences, and resolving problems of data availability and incomplete lineage coverage.

Self-renewal pathways in acute myeloid leukemia

Well-characterized mouse models of acute myeloid leukemia (AML) have been used to explore the molecular basis for stem-like behavior of sub-populations of leukemia cells [48]. A self-renewal associated gene signature (SRAS) has been identified that is activated in both hematopoietic stem cells and leukemia initiating cells. An analogous study of human AML has identified a clinically relevant stem-associated signature expressed in human normal hematopoietic and leukemia stem cells [49]. There are only 4 genes common to the published human and mouse signatures, and the extent to which the mouse model functionally recapitulates the human system is unknown. A pathprint analysis systematically extracts and compares the pathways defining stem phenotypes in each of these studies, identifying 4 common human and mouse stem-associated pathways. The common pathways are *translation factors* and *class B secretin-like GPCRs* from Wikipathways, and static modules centered on *PLCG2* and *RAN* (Figure 4a). There is no overlap between these pathways at the gene level. The combinatorial clinical relevance of these pathways was tested by calculating pathprints for 4 independent clinical studies of gene expression in AML patients [50-53]. The patient samples were grouped into high and low expression groups by k-means clustering of the sum of their pathprint scores in the common self-renewal pathways. High scores are associated with poor prognosis in each of the studies and is also significant compared to a background of random pathway permutations (Figure 4b, Additional file 11). The identification of translation factors suggests that modulation of translation might be a therapeutic approach in poor-prognosis AML, consistent with studies targeting this process in early phase clinical trials [54]. The set of stem cell pathways that are conserved across human and mouse have significantly greater clinical relevance than either the human or mouse pathways on their own, demonstrating the value of a cross species analysis in this case study (Additional file 12). The GPCRs, PLCG2 and RAN modules may represent new pathways for clinical investigation; a clear relationship between the pathprint score and clinical outcome is observed for the PLCG2 module, a tightly connected set of genes involved in signaling and metabolism (Figure 4c, Additional file 13).

Conclusions:

The Pathprinting project provides the community with a consistent, functional annotation of gene expression across a fixed ‘set’ of pathways. It moves beyond traditional approaches, resolving the major bottleneck on the road towards efficient systems-biology based modeling by addressing the inherent experimental and platform biases that confound microarray analyses. Pathprinting is now being applied to group the function of datasets within the Harvard Stem Cell Institute Stem Cell Commons (stemcellcommons.org) so that samples that have similar function can be discovered within stem cell data. A cytoscape plugin is also in development as part of the NHLBI Progenitor consortium[55] and we have integrated the method into the Stem Cell Discovery Engine [56] (SCDE) to provide web-based accessibility. The SCDE is a portal for integrated access to tissue and cancer stem cell experimental information and molecular profiling analysis tools via a web-based Galaxy instance. Pathprinting is also embedded within the toolbench distribution of Galaxy. We encourage the community to employ Pathprinting to communicate functional findings more consistently. It is important to note that Pathprinting is effective for use on single samples - a sample can easily be pathprinted and compared to ‘what is there’. This has significant implications for applications in personalized medicine and single cell analyses.

The R package Pathprint is provided to calculate pathprints (or continuous pathway scores) from expression arrays and pathway enrichments from input gene lists. The package also contains a database of approximately 180,000 pathprints from GEO. The packages, along with Sweave files detailing the package usage and analysis in this paper are available online[43]. A supplementary package, pathprintTF is also provided, containing a similar framework and database to pathprint but built upon protein interaction modules centered on transcription factors rather than pathways to enable cross-platform comparison of transcriptional control elements. The transcription factor modules are based on protein-protein interaction sub-networks centered on a series of 1022 transcription factors, the package and more details are provided on the pathprint website.

The correlation of mRNA expression to protein levels, and also to phenotype, depends on a variety of factors such as translation efficiency, mRNA abundance, ribosome occupancy, protein abundance and turnover. Gene expression levels are a good surrogate for protein levels for housekeeping genes (ribosomal proteins, glycolytic enzymes and TCA cycle proteins) but mRNA levels correlate less well with protein levels for kinases, proteases, secreted proteins and transcription factors, and overall mRNA variability explains approximately 40% of the variability in protein levels. Pathprinting establishes a standardized method for large-scale quantitative comparisons of cellular function, and any analysis of this type depends on the availability of large-scale quantitative genome-wide data-sets. Gene expression data repositories are currently the only resource expansive enough to address this need. Future versions of the Pathprint will extend the value of existing array data by integrating RNA-seq, epigenetic and proteomic profiles, providing context for new experiments from the existing body of microarray data, and helping resolve the links between regulation and expression of cellular function.

Abbreviations

GEO: Gene Expression Omnibus; GSEA: Gene Set Enrichment Analysis; GO: Gene Ontology; POE: Probability of Expression; SRAS: Self-Renewal Signature; ES: Embryonic Stem; iPS: Induced Pluripotent Stem; AML: Acute Myeloid Leukemia; SCDE: Stem Cell Discovery Engine; TCA: tricarboxylic acid; mRNA: messenger ribonucleic acid.

Authors’ contributions

GA, OH and WH conceived the study. GA designed the algorithm, analyzed the data and prepared the manuscript, WH supervised the project, WH and OH edited the manuscript, IK and LS constructed the static networks, RP and TC designed the benchmarking, AK and SA designed the biological validations. GA and US compiled the R package. SH integrated the package into SCDE. All authors read and approved the final manuscript.

Acknowledgements
The authors would like to acknowledge the following for conceptual and technical advice.
Dr. R. Gazit, Dr. T. Mehan, Prof. F. Michor, Prof. C. Seoighe, Dr. A. Sinha, Dr. R. Sompallae, C Spencer, I. Sytchev.
This work was supported by the National Institute of Health [1RC2CA148222-01 to W.H.]

Competing interests
The authors declare that they have no competing interests.

References:

1. Wang X, Gulbahce N, Yu H: **Network-based methods for human disease gene prediction.** *Brief Funct Genomics* 2011, **10:**280-293.
2. Liu Y, Koyuturk M, Barnholtz-Sloan JS, Chance MR: **Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases.** *BMC Syst Biol* 2012, **6:**65.
3. Yang X: **Use of functional genomics to identify candidate genes underlying human genetic association studies of vascular diseases.** *Arterioscler Thromb Vasc Biol* 2012, **32:**216-222.
4. Yang X, Zhang B, Zhu J: **Functional genomics- and network-driven systems biology approaches for pharmacogenomics and toxicogenomics.** *Curr Drug Metab* 2012, **13:**952-967.
5. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F: **Prediction of human disease genes by human-mouse conserved coexpression analysis.** *PLoS Comput Biol* 2008, **4:**e1000043.
6. Le H-S, Oltvai ZN, Bar-Joseph Z: **Cross Species Queries of Large Gene Expression Databases.** *Bioinformatics (Oxford, England)* 2010.
7. Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, Mesirov J, Golub TR, Jacks T: **An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis.** *Nat Genet* 2005, **37:**48-55.
8. Sadanandam A, Futakuchi M, Lyssiotis CA, Gibb WJ, Singh RK: **A cross-species analysis of a mouse model of breast cancer-specific osteolysis and human bone metastases using gene expression profiling.** *BMC Cancer* 2011, **11:**304.

9. Johnson RA, Wright KD, Poppleton H, Mohankumar KM, Finkelstein D, Pounds SB, Rand V, Leary SE, White E, Eden C, Hogg T, Northcott P, Mack S, Neale G, Wang YD, Coyle B, Atkinson J, DeWire M, Kranenburg TA, Gillespie Y, Allen JC, Merchant T, Boop FA, Sanford RA, Gajjar A, Ellison DW, Taylor MD, Grundy RG, Gilbertson RJ: **Cross-species genomics matches driver mutations and cell compartments to model ependymoma.** *Nature* 2010, **466:**632-636.

10. 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:**D760-765.

11. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pedro Pereira R, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U: **ArrayExpress update--trends in database growth and links to data analysis tools.** *Nucleic Acids Res* 2013, **41:**D987-990.

12. Engreitz JM, Morgan AA, Dudley JT, Chen R, Thathoo R, Altman RB, Butte AJ: **Content-based microarray search using differential expression profiles.** *BMC bioinformatics* 2010, **11:**603.

13. Fujibuchi W, Kiseleva L, Taniguchi T, Harada H, Horton P: **CellMontage: similar expression profile search server.** *Bioinformatics* 2007, **23:**3103-3104.

14. Keum C, Woo JH, Oh WS, Park S-N, No KT: **Improving gene expression similarity measurement using pathway-based analytic dimension.** *BMC Genomics* 2009, **10 Suppl 3:**S15.

15. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA: **The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes.** *Nucleic acids research* 2011, **39:**D1011-1015.

16. Seita J, Sahoo D, Rossi DJ, Bhattacharya D, Serwold T, Inlay MA, Ehrlich LI, Fathman JW, Dill DL, Weissman IL: **Gene expression commons: an open platform for absolute gene expression profiling.** *PLoS ONE* 2012, **7:**e40321.

17. Li Z, Su Z, Wen Z, Shi L, Chen T: **Microarray platform consistency is revealed by biologically functional analysis of gene expression profiles.** *BMC bioinformatics* 2009, **10 Suppl 11:**S12.

18. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L: **Reactome: a database of reactions, pathways and biological processes.** *Nucleic acids research* 2011, **39:**D691-697.

19. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38:**D355-360.

20. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: **WikiPathways: pathway editing for the people.** *PLoS Biol* 2008, **6:**e184.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
22. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, D. Crawford M, B. Datto M, Kelley M, Mathey-Prevot B, Potti A, Nevins JR: **A pathway-based classification of human breast cancer.** *Proc Natl Acad Sci USA* 2010, **107:**6994-6999.
23. Greenblum SI, Efroni S, Schaefer CF, Buetow KH: **The PathOlogist: An Automated Tool for Pathway-Centric Analysis.** *BMC bioinformatics* 2011, **12:**133.
24. Gundem G, Lopez-Bigas N: **Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types.** *Genome Med* 2012, **4:**28.
25. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, 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.
26. Lee E, Chuang HY, Kim JW, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification.** *PLoS Comput Biol* 2008, **4:**e1000217.
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102:**15545-15550.
28. Chen R, Li L, Butte AJ: **AILUN: reannotating gene expression data automatically.** *Nat Methods* 2007, **4:**879.
29. Burgess DJ: **Alternative splicing: proteomic rewiring through transcriptomic diversity.** *Nat Rev Genet* 2012, **13:**518-519.
30. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GSS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C, Gollapudi SK, Tattikota SG, Mohan S, Padhukasahasram H, Subbannayya Y, Goel R, Jacob HKC, Zhong J, Sekhar R, Nanjappa V, Balakrishnan L, Subbaiah R, Ramachandra YL, Rahiman BA, Prasad TSK, Lin J-X, Houtman JCD, Desiderio S, Renauld J-C, Constantinescu SN, et al: **NetPath: a public resource of curated signal transduction pathways.** *Genome biology* 2010, **11:**R3.
31. Wu G, Feng X, Stein L: **A human functional protein interaction network and its application to cancer data analysis.** *Genome biology* 2010, **11:**R53.
32. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY,

Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2012, **40:**D13-25.

33. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS Comput Biol* 2009, **5:**e1000262.
34. Ackermann M, Strimmer K: **A general modular framework for gene set enrichment analysis.** *BMC bioinformatics* 2009, **10:**47.
35. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102:**13544-13549.
36. Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA: **Heading down the wrong pathway: on the influence of correlation within gene sets.** *BMC Genomics* 2010, **11:**574.
37. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E: **A statistical framework for expression- based molecular classification in cancer.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002, **64:**717-736.
38. Choi H, Shen R, Chinnaiyan AM, Ghosh D: **A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments.** *BMC bioinformatics* 2007, **8:**364.
39. **http://molecularbrain.org**
40. Novershtern N, Subramanian A, Lawton L, Mak… R: **Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis.** *Cell* 2011.
41. Chambers SM, Boles NC, Lin K-YK, Tierney MP, Bowman TV, Bradfute SB, Chen AJ, Merchant AA, Sirin O, Weksberg DC, Merchant MG, Fisk CJ, Shaw CA, Goodell MA: **Hematopoietic fingerprints: an expression database of stem cells and their progeny.** *Cell Stem Cell* 2007, **1:**578-591.
42. Schliep K: **Phylogenetics in R package phangorn.** 2010:1-46.
43. **http://compbio.sph.harvard.edu/hidelab/pathprint**
44. Koeva M, Forsberg EC, Stuart JM: **Computational Integration of Homolog and Pathway Gene Module Expression Reveals General Stemness Signatures.** *PLoS ONE* 2011, **6:**e18968.
45. Wang J, Alexander P, Wu L, Hammer R, Cleaver O, McKnight SL: **Dependence of mouse embryonic stem cells on threonine catabolism.** *Science* 2009, **325:**435-439.
46. Wei F, Scholer HR, Atchison ML: **Sumoylation of Oct4 enhances its stability, DNA binding, and transactivation.** *J Biol Chem* 2007, **282:**21551-21560.
47. Painter MW, Davis S, Hardy RR, Mathis D, Benoist C: **Transcriptomes of the B and T lineages compared by multiplatform microarray profiling.** *J Immunol* 2011, **186:**3047-3057.

48. Krivtsov AV, Twomey D, Feng Z, Stubbs MC, Wang Y, Faber J, Levine JE, Wang J, Hahn WC, Gilliland DG, Golub TR, Armstrong SA: **Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9.** *Nature* 2006, **442:**818-822.
49. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA: **Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia.** *JAMA* 2010, **304:**2706-2715.
50. Metzeler KH, Hummel M, Bloomfield CD, Spiekermann K, Braess J, Sauerland MC, Heinecke A, Radmacher M, Marcucci G, Whitman SP, Maharry K, Paschka P, Larson RA, Berdel WE, Buchner T, Wormann B, Mansmann U, Hiddemann W, Bohlander SK, Buske C: **An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia.** *Blood* 2008, **112:**4193-4201.
51. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary AR, Hockenbery D, Wood B, Heimfeld S, Radich JP: **Identification of genes with abnormal expression changes in acute myeloid leukemia.** *Genes Chromosomes Cancer* 2008, **47:**8-20.
52. Tomasson MH, Xiang Z, Walgren R, Zhao Y, Kasai Y, Miner T, Ries RE, Lubman O, Fremont DH, McLellan MD, Payton JE, Westervelt P, DiPersio JF, Link DC, Walter MJ, Graubert TA, Watson M, Baty J, Heath S, Shannon WD, Nagarajan R, Bloomfield CD, Mardis ER, Wilson RK, Ley TJ: **Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia.** *Blood* 2008, **111:**4797-4808.
53. Wouters BJ, Lowenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R: **Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome.** *Blood* 2009, **113:**3088-3091.
54. Assouline S, Culjkovic B, Cocolakis E, Rousseau C, Beslu N, Amri A, Caplan S, Leber B, Roy DC, Miller WH, Jr., Borden KL: **Molecular targeting of the oncogene eIF4E in acute myeloid leukemia (AML): a proof-of-principle clinical trial with ribavirin.** *Blood* 2009, **114:**257-260.
55. **http://www.progenitorcells.org/content/bioinformatics-and-genomics-tools**
56. Ho Sui SJ, Begley K, Reilly D, Chapman B, McGovern R, Rocca-Sera P, Maguire E, Altschuler GM, Hansen TA, Sompallae R, Krivtsov A, Shivdasani RA, Armstrong SA, Culhane AC, Correll M, Sansone SA, Hofmann O, Hide W: **The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons.** *Nucleic Acids Res* 2012, **40:**D984-991.

Tables

Table 1: **Summary of gene sets used in pathprint**

	Pathways	Mean size	Median size	Min size	Max size	Total genes
Reactome	53	153.6	108	11	932	4874
Wikipathways	173	50.14	33	6	260	3918
Netpath	36	170.08	83	8	816	3811
KEGG	227	75.51	55	6	1138	5990
Static Modules	144	44.9	21	9	733	6458
All	633	73.53	41	6	1138	10903

Table 2: **Pathprint-based retrieval of data from GEO:** Arrays retrieved from GEO from consensus tissue pathprints at 95% precision.

	Seed Arrays	Correct retrievals	Platforms	Species
Brain	50	8691	25	4
Kidney	81	1156	14	3
Liver	196	4797	22	4
Lung	142	1735	13	3
Skeletal muscle	29	2919	18	3
Spleen	33	179	5	2

Figure Legends

Figure 1: **The Pathprint pipeline:** Rank-normalized gene expression is mapped to pathway expression. A distribution of expression scores across GEO is used to produce a probability of expression (POE) for each pathway. A pathprint vector is derived by transformation of the signed POE distribution into a ternary score, representing pathway activity as significantly under-expressed [-1], intermediately-expressed [0], over-expressed [+1].

Figure 2: **Cross-species integration** a): Precision-recall within tissue training dataset for the pathprint (red, mean average precision 0.90), unthresholded POE (dashed, 0.88), random gene sets (black, 0.83), gene-expression barcode (blue, 0.73), Spearman gene-expression correlation (green, 0.71). (b) Comparison of distance metrics: Precision-recall curves for aggregated mouse to human tissue data based on a thresholded pathprint build using Euclidean (blue), Manhattan (green) and Mahalanobis distances (red). Tissue- vs platform/species-dominated clustering: Plots of the two most significant principal components (PC) for (c) the pathprint and (d) the gene expression barcode, brain = red, kidney = yellow, liver = green, lung = light blue, muscle = dark blue, spleen = pink, Mouse 430A2 = circles, Human 133plus2 = diamonds, Human 133A = crosses (e) Functional classification of tissues and blood cell types: Hierarchical clustering of consensus pathprints for human and mouse tissues on three platforms based on the Wikipathway and Reactome pathways that significantly contributing the clustering. Colors indicate 1 (red), 0 (white) and -1 (blue)).

Figure 3: **Functional classification of blood cell types:** a) Maximum parsimony phylogenetic reconstruction of the hematopoietic lineage using pathprints calculated from a) human[40] and b) mouse [41] gene expression experiments. c) Combined human-mouse tree based on shared informative pathways that resolve trees a) and b) and pathway heatmap. The myeloid (yellow) and lymphoid (purple) branches are indicated, dark branches represent agreement with the canonical lineage. See Additional file 10 for pathway annotations.

Figure 4: **Clinically important Self Renewal Associated Signature in AML:** a) Pathways differentially expressed in stem vs non-stem cell profiles in leukemic and normal samples were found in human and mouse experiments. 4 common SRAS pathways were identified. b) The SRAS pathprint scores of AML patients is significantly associated with survival. c) A single pathway of interest is highlighted, the overall PGCL2 module is upregulated in normal and cancer stem cells but individual genes differ between species. This pathway is strongly associated with survival (please see Additional file 13).

Additional data files

Additional file 1: **Table listing platforms covered by Pathprint.**
Format: XLS

Additional file 2: **Table listing pathway sources, retrieval dates, URLs, and the top GO term that is enriched in each pathway (hypergeometric distribution p-value).**
Format: XLS

Additional data 3: **Table of the overlap in the genes covered by each gene set resource across in Pathprint (human pathways).**
Format: XLS

Additional file 4, Supplementary Figure 1: **Benchmarking and threshold optimization:** Benchmarking based on the tissue dataset (above) and brain-subtypes (below).
a), d) Mean error rate based on 10 repeats of a 5-fold cross-validation over a range of POE thresholds. Error bars indicate -/+ 1 std.dev. The black line indicates the ratio for the unthresholded POE matrix, and the red for the barcode, dashed lines indicate -/+ 1 std.dev. b,d) Intra-cluster vs. inter-cluster variance ratio, over a range of POE thresholds dashed line indicates the ratio for the unthresholded POE matrix. c,f) Mean average precision over a range of POE thresholds for the pathprint (black circles) and a pathprint build on random gene sets of equivalent size distribution (blue circles). Solid lines indicate the mean average precision for Barcode (blue), Spearman correlation (green) and the unthresholded pathprint (red) N.B. barcode or gene expression correlation data were not calculated for the brain-subtype dataset
Format: PDF

Additional file 5, Supplementary Figure 2: **Precision-recall curves across the full set of GEO samples and distribution of distances of GEO samples from each tissue pathprint**: (a) Precision-recall curves for each of the tissues across the pathprint-mapped GEO database; brain (red), kidney (yellow), liver (green), lung (cyan), skeletal muscle (blue), spleen (magenta). (b) Precision curves for each of the tissues across the pathprint-mapped GEO database (red, right axis) and histogram of distance of samples in the pathprint-mapped GEO database from each tissue consensus pathprint (black, left axis). Distance scales between 0 (all pathway scores matched) to 1 (all pathway scores mis-matched, i.e. 1 vs -1) (c) Estimated p-values: A p-value was assigned to every sample in the GEO pathprint matrix to assess the likelihood of association with the consensus pathprint for each tissue. The plots the relationship between this p-value and the precision (i.e. the proportion correctly matched to each tissue), as determined from the GEO metadata, when samples are ranked according to p-value.
Format: PDF

Additional file 6: **Table of the top *R. norvegicus*, *D. rerio*, *D. melanogaster* and *C. elegans* D.rerio matching arrays to human/mouse brain and liver samples.**
Format: XLS

Additional file 7: **Table listing the pathways in the pluripotent consensus pathprint.**
Format: XLS

Additional file 8, Supplementary Figure 3: **ES differentiation timecourse** a) Distance from the embryonic stem cell pathprint signature of two mouse embryonic stem cell lines, J1 and R1, differentiating to embryoid bodies. The data were obtained from GEO accessions GSE2972 (J1) and GSE3749 (R1). b) Heatmap of pathways in the ES pathprint signature that vary over both differentiation timecourses (blue = -1, white = 0, red = +1). The column labeled ES denotes the ES pathprint signature.
Format: PDF

Additional file 9: **Table listing the pluripotent seed arrays and top arrays matching the pluripotent consensus pathprint.**
Format: XLS

Additional file 10, Supplementary Figure 4: **Combined human and mouse blood lineage tree:** Pathway heatmap based on shared informative pathways that resolve trees b) and c) in Figure 2.
Format: PDF

Additional file 11, Supplementary Figure 5: **Pathway-based survival analysis** (a) Kaplan Meier curves of patients in 4 independent AML clinical datasets

stratified by expression of common mouse and human SRAS pathways; Translation Factors (Wikipathways), GPCRs, Class B Secretin-like (Wikipathways), PLCG2 (Static Module), and RAN (Static Module). The red and blue lines indicates high and low pathprint scores respectively (b) P-value of Kaplan Meier estimator of patients stratified by expression of common mouse and human SRAS pathways in 4 independent clinical datasets, relative to a background of randomly selected pathways from the full pathprint set, (c) common genes relative to a background of randomly selected genes from expression chip (only single dataset shown), and (d) common SRAS pathways relative to a background of randomly selected human SRAS pathways. A red dot indicates the p-value, the grey cone is a bean plot representing the distribution of p-values from 1000 randomly selected sets of pathways or genes. The blue line indicates a p-value of 0.05
Format: PDF

Additional file 12, Supplementary Figure 6: **Pathway-based survival analysis by species** Kaplan Meier curves of AML patients stratified by expression of common human and mouse (a), human (b), and mouse (c) SRAS pathways in 4 independent clinical datasets. The red and blue lines indicate high and low pathprint scores respectively.
Format: PDF

Additional file 13, Supplementary Figure 7: **The PGLC2 module** a) The protein-protein interaction network of a single human/mouse common SRAS pathway - the PGLC2 module. Node color represents fold change in the combined leukemic/normal blood dataset (expression in normal and leukemia stem cells / expression in progenitor cells). b) The pathprint score of this single pathway in AML patients is associated with survival in 4 independent clinical datasets (red = +1, yellow = 0, blue = -1)
Format: PDF

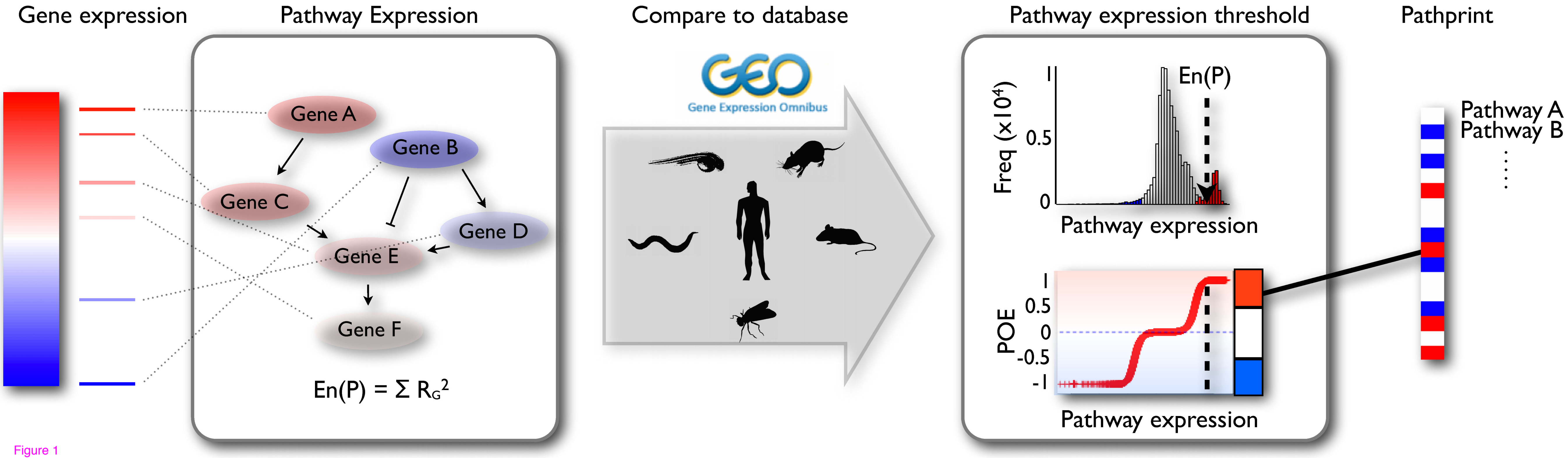
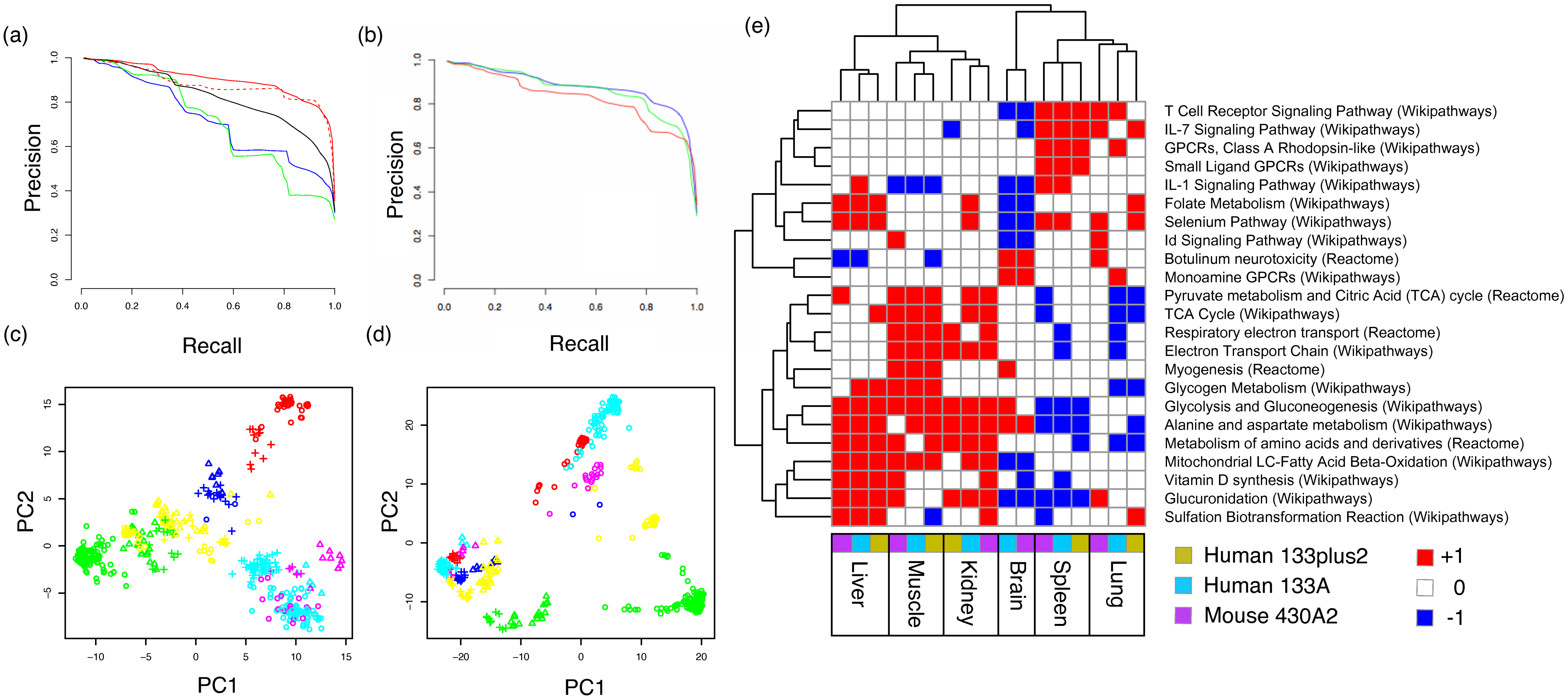

Gene expression
Pathway Expression
Compare to database
Pathway expression threshold
Pathprint
Gene A
Gene B
Gene C
Gene D
Gene E
Gene F
$En(P) = \Sigma R_G^2$
Gene Expression Omnibus
1
0.5
0
Freq ($\times 10^4$)
En(P)
Pathway expression
1
0.5
0
-0.5
-1
POE
Pathway expression
Pathway A
Pathway B

Figure 1



(a)
Precision
1.0
0.8
0.6
0.4
0.2
0.0
Recall
0.0
0.2
0.4
0.6
0.8
1.0
(b)
Precision
1.0
0.8
0.6
0.4
0.2
0.0
Recall
0.0
0.2
0.4
0.6
0.8
1.0
(c)
PC2
15
10
5
0
−5
PC1
−10
−5
0
5
10
15
(d)
PC2
20
10
0
−10
PC1
−20
−10
0
10
20
(e)
T Cell Receptor Signaling Pathway (Wikipathways)
IL-7 Signaling Pathway (Wikipathways)
GPCRs, Class A Rhodopsin-like (Wikipathways)
Small Ligand GPCRs (Wikipathways)
IL-1 Signaling Pathway (Wikipathways)
Folate Metabolism (Wikipathways)
Selenium Pathway (Wikipathways)
Id Signaling Pathway (Wikipathways)
Botulinum neurotoxicity (Reactome)
Monoamine GPCRs (Wikipathways)
Pyruvate metabolism and Citric Acid (TCA) cycle (Reactome)
TCA Cycle (Wikipathways)
Respiratory electron transport (Reactome)
Electron Transport Chain (Wikipathways)
Myogenesis (Reactome)
Glycogen Metabolism (Wikipathways)
Glycolysis and Gluconeogenesis (Wikipathways)
Alanine and aspartate metabolism (Wikipathways)
Metabolism of amino acids and derivatives (Reactome)
Mitochondrial LC-Fatty Acid Beta-Oxidation (Wikipathways)
Vitamin D synthesis (Wikipathways)
Glucuronidation (Wikipathways)
Sulfation Biotransformation Reaction (Wikipathways)
Liver
Muscle
Kidney
Brain
Spleen
Lung
Human 133plus2
Human 133A
Mouse 430A2
+1
0
-1

Figure 2

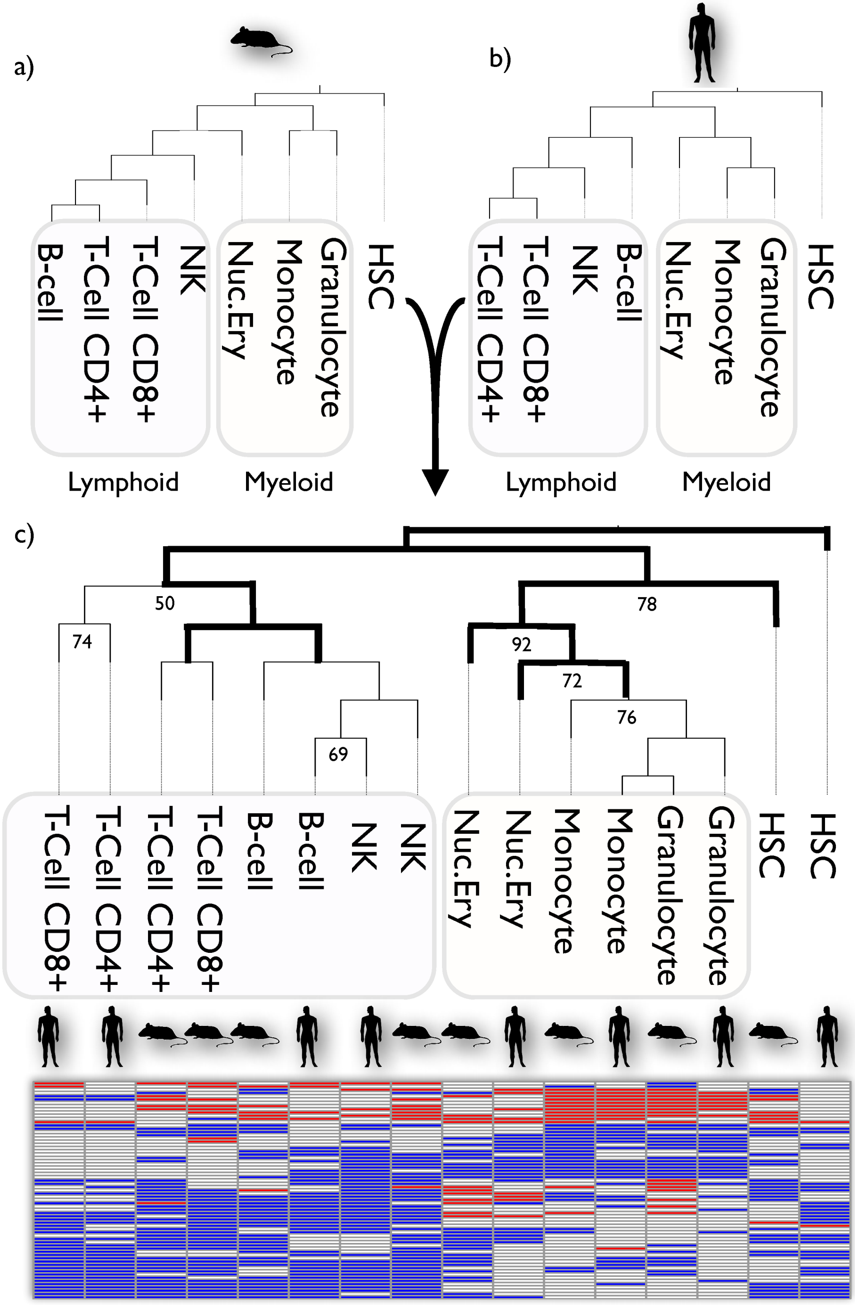

a)
B-cell
T-Cell CD4+
T-Cell CD8+
NK
Nuc.Ery
Monocyte
Granulocyte
HSC
Lymphoid
Myeloid
b)
T-Cell CD4+
T-Cell CD8+
NK
B-cell
Nuc.Ery
Monocyte
Granulocyte
HSC
Lymphoid
Myeloid
c)
50
74
78
92
72
76
69
T-Cell CD8+
T-Cell CD4+
T-Cell CD4+
T-Cell CD8+
B-cell
B-cell
NK
NK
Nuc.Ery
Nuc.Ery
Monocyte
Monocyte
Granulocyte
Granulocyte
HSC
HSC

Figure 3

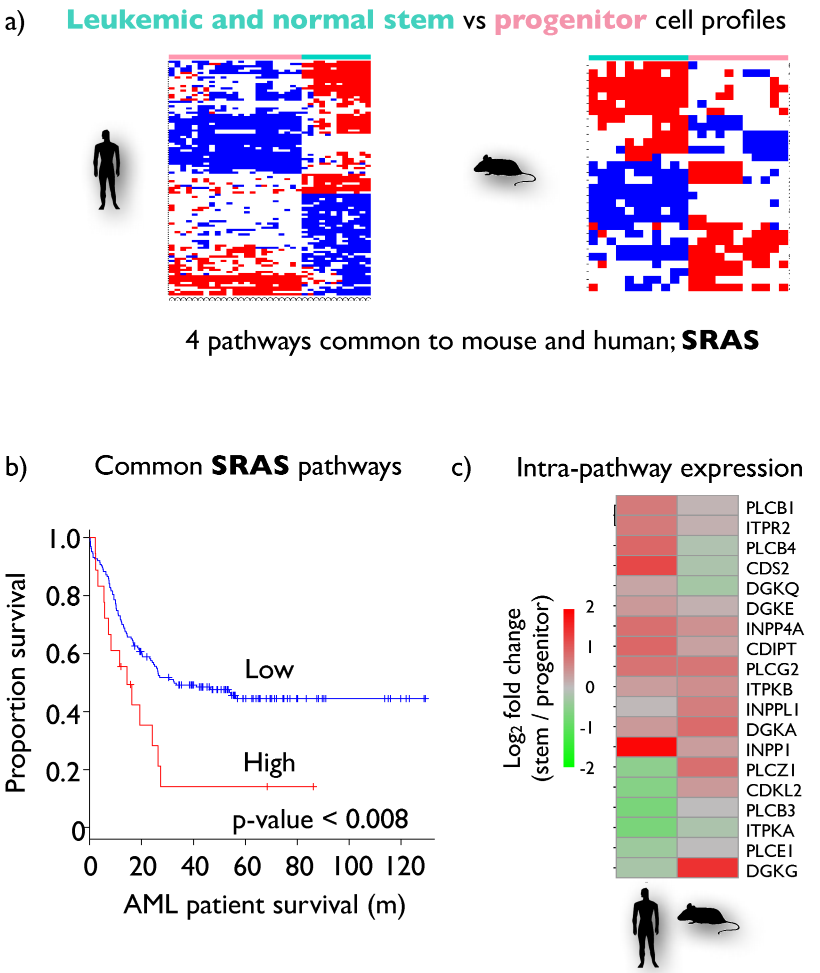

a)
Leukemic and normal stem vs progenitor cell profiles
4 pathways common to mouse and human; SRAS
b)
Common SRAS pathways
1.0
0.8
0.6
0.4
0.2
0
Proportion survival
0
20
40
60
80
100
120
AML patient survival (m)
Low
High
p-value < 0.008
c)
Intra-pathway expression
Log2 fold change (stem / progenitor)
2
1
0
-1
-2
PLCB1
ITPR2
PLCB4
CDS2
DGKQ
DGKE
INPP4A
CDIPT
PLCG2
ITPKB
INPPL1
DGKA
INPP1
PLCZ1
CDKL2
PLCB3
ITPKA
PLCE1
DGKG

Figure 4

Additional files provided with this submission:

Additional file 1: AdditionalFile1_1388493011932758.xls, 28K
http://genomemedicine.com/imedia/3611510281039323/supp1.xls
Additional file 2: AdditionalFile2_1388493011932758.xls, 224K
http://genomemedicine.com/imedia/1554373094103932/supp2.xls
Additional file 3: AdditionalFile3_1388493011932758.xls, 21K
http://genomemedicine.com/imedia/1988013179103932/supp3.xls
Additional file 4: AdditionalFile4_1388493011932758.pdf, 254K
http://genomemedicine.com/imedia/1760408058103932/supp4.pdf
Additional file 5: AdditionalFile5_1388493011932758.pdf, 164K
http://genomemedicine.com/imedia/1051944354103932/supp5.pdf
Additional file 6: AdditionalFile6_1388493011932758.xls, 80K
http://genomemedicine.com/imedia/1446663331103932/supp6.xlsx
Additional file 7: AdditionalFile7_1388493011932758.xls, 20K
http://genomemedicine.com/imedia/5214791411039324/supp7.xls
Additional file 8: AdditionalFile8_1388493011932758.pdf, 211K
http://genomemedicine.com/imedia/1051620574103932/supp8.pdf
Additional file 9: AdditionalFile9_1388493011932758.xls, 165K
http://genomemedicine.com/imedia/1481800763103932/supp9.xls
Additional file 10: AdditionalFile10_1388493011932758.pdf, 286K
http://genomemedicine.com/imedia/1626055524103932/supp10.pdf
Additional file 11: AdditionalFile11_1388493011932758.pdf, 351K
http://genomemedicine.com/imedia/1506185041103932/supp11.pdf
Additional file 12: AdditionalFile12_1388493011932758.pdf, 406K
http://genomemedicine.com/imedia/3950501721039325/supp12.pdf
Additional file 13: AdditionalFile13_1388493011932758.pdf, 781K
http://genomemedicine.com/imedia/3978054541039324/supp13.pdf