Title
Applications of high-throughput genome and transcriptome analysis in human disease

Author
Inkeles, Megan So

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Applications of high-throughput genome and transcriptome analysis in human disease

A dissertation submitted in satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Megan So Inkeles

2014

ABSTRACT OF THE DISSERTATION

Applications of high-throughput genome and transcriptome analysis in human disease

by

Megan So Inkeles

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2014

Professor Matteo Pellegrini, Chair

The development of gene expression profiling technology has enabled the high-throughput discovery of the genes and pathways that underlie disease pathophysiology and phenotype. This work analyzes microarray and RNA sequencing data to identify genes and functional pathways associated with human diseases. In the first part, gene expression profiles derived from pancreatic ductal adenocarcinoma tumors are correlated to patient disease free survival time in order to find genes that confer a protective advantage. Four genes found to be significantly correlated with disease free survival were validated in tissue using PCR. In the second part, publicly available gene expression profiles for 16 skin diseases were integrated to build a disease classifier as well as characterize genes, functions, and pathways associated with each condition. Since data was drawn from different laboratories and experiment batches, we used Frozen Robust MultiArray Average to normalize the data and identified disease specific gene signatures using a ranking algorithm. Finally, we integrated this skin database with public data on interferon-regulated gene programs to find a negative inverse correlation between Type I and Type II interferon. The final part of this work applies the principles of comparisons in multiple diseases to the problem of characterizing subtypes of one disease. mRNA-seq techniques were briefly explored to probe for genes which historically have been difficult to

detect on microarray. We compared microarray gene expression profiles from four subtypes of leprosy—lepromatous leprosy (L-lep), tuberculoid leprosy (T-lep), reversal reaction, and erythema nodosum leprosum—to build a proportional median-random forest classifier and perform functional analyses, such as weighted gene correlation network analysis (WGCNA), to find genes and pathways associated with each leprosy subtype. Integrating our proportional median subtype signature for T-lep with the WGCNA module associated with T-lep, we identified MMP12 as a novel differentiator of T-lep from L-lep. This gene was verified in tissue sections of leprosy using immunohistochemistry. The use of high throughput gene expression profile analysis in these three projects demonstrates the versatility and utility of transcriptome analysis when applied to human disease systems.

The dissertation of Megan So Inkeles is approved

Thomas G. Graeber

Robert L. Modlin

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2014

Contents

List of Figures and Tables

Chapter 3:

Chapter 4:

Chapter 5

Acknowledgments

Chapter 4 is a version of a manuscript in submission with the same title.

MSI, POS, WRS, JTE, RLM, and MP wrote and edited the manuscript. DL built the website. MSI, WRS, JTE, RLM, RMBT, TGG, and MP designed the components of the analysis. MG, BH, and SM provided patient samples and microarray data. WRS performed deconvolution analysis. MSI performed all other analyses.

Chapter 5 is a version a manuscript in preparation.

William Swindell performed the deconvolution analysis. Rosane Teles performed the Ingenuity Pathways Analysis and the immunohistochemistry experiments. Megan Inkeles performed all other analyses and wrote the manuscript.

This work was funded in part by the UCLA Whitcome Pre-Doctoral Fellowship, the UCLA Medical Scientist Training Program, the CHANEL-CERES award and NIH/NIAMS grant P50 5P50AR063020.

Vita

Timeline

2004-2007	Residential teaching assistant, Education Program for Gifted Youth summer program (Stanford University) Taught C++ and Java programming
2005-2007	Undergraduate Research, Stanford University
2006	Stanford University Undergraduate Major Research Grant
2007	B.S., Engineering (Biomedical Informatics) Stanford University
2012-2014	Whitcome Pre-Doctoral Training Program Fellowship

Publications and presentations

Inkeles MS, Scumpia Po, Swindell WR, Lopez D, Teles RMB, Graeber TG, Meller S, Homey B, Elder JT, Gilliet M, Modlin RL, Pellegrini M. Comparison of Molecular Signatures from Multiple Skin Diseases Identifies Mechanisms of Immunopathogenesis. *In Submission*

Wheelwright M, Kim EW, Inkeles MS, De Leon A, Pellegrini M, Krutzik SR, Liu PT. All-trans retinoic acid-triggered antimicrobial activity against Mycobacterium tuberculosis is dependent on NPC2. *J Immunol*. 2014 Mar 1;192(5):2280-90.

Marinelli LJ, Fitz-Gibbon S, Hayes C, Bowman C, Inkeles M, Loncaric A, Russell DA, Jacobs-Sera D, Cokus S, Pellegrini M, Kim J, Miller JF, Hatfull GF, Modlin RL. Propionibacterium acnes Bacteriophages Display Limited Genetic Diversity and Broad Killing Activity against Bacterial Skin Isolates. *MBio*. 2012 Sep 25;3(5).

Presentations and Posters

Poster presentation at the Society for Investigative Dermatology Annual Meeting, Albuquerque, New Mexico (May 8, 2014).

“Skin disease bioinformatics: Molecular classification and insights into skin disease pathogenesis.” (May 21, 2013). Talk given at the UCLA Dermatology Basic Sciences Research Symposium.

Poster Presentation at the International Investigative Dermatology Meeting, Edinburgh, Scotland (May 10, 2013).

Poster presentation at the Education Academy of Computational Life Science, Tokyo, Japan (Sept 5, 2012).

“Gene expression profiles in non-melanoma skin cancer” (Feb 2, 2012). Talk given at the UCLA Bioinformatics Departmental Retreat

“A microarray meta-analysis of publicly available skin disease” (Jan 23, 2012). Talk given as part of UCLA medical student Clinical Research Certificate program.

Chapter 1: Introduction

DNA and mRNA profiling methods

The publication of the first draft of the human genome in 2001 was a medical and scientific milestone that facilitated the high-throughput study of genes and their association to disease (http://www.genome.gov). Although the human genome is a finite length, the processes by which genes and their regulatory elements are encoded and expressed are incredibly complex. Over ten years later, the fields of genome analysis, genome assembly, and messenger RNA (mRNA) transcriptome analysis are still evolving with new tools and strategies for identifying genes and pathways that shed light on human disease.

This work will focus on the analysis of gene-expressing mRNA. Since strategies for mRNA transcriptome analysis are based on gene-encoding DNA, these methods will also be briefly touched upon. Both types of analysis involve high throughput amplification and sequence detection, either by array hybridization or direct sequencing. Current DNA methods primarily involve high throughput sequencing of nuclear DNA, whereas RNA methods involve either microarray characterization of gene expression or high throughput sequencing of mRNA.

De novo genome assembly

DNA sequencing describes the characterization of nucleic acids that make up a strand of genomic material. Solving the sequence of an organism's genomic DNA is an important first step in the high-throughput study of gene expression since it allows for the detection and prediction of genes and provides a reference sequence to which transcribed mRNA can be mapped. Genomic sequences are also used to find point mutations and perform comparative studies across species, although those techniques will not be discussed here.

First attempts at sequencing long strands of DNA used the Sanger method, a time-consuming and expensive process in which the entire sequence is determined by synthesizing sequentially

longer DNA fragments with differently tagged, fluorescent terminal nucleic acids (*1*). A major breakthrough in the field came with the shotgun sequencing approach, in which long DNA sequences are broken into smaller fragments, which are individually synthesized and sequenced then reassembled *in silico* into a final sequence (*2*). Current next generation sequencing technology uses this concept to concurrently sequence massive numbers of short fragments, typically 50 to 200 base pairs in length, yielding a total sequence length many times the sequence of interest. However, both Sanger and next generation sequencing still depend on so-called random amplification DNA or RNA molecules in order to generate the short reads ("so-called" because this amplification process has been shown to favor certain sequences) (*3*). Technology is currently in development to analyze individual DNA or RNA molecules as they are built ("single molecule real time sequencing"), thus eliminating much of the bias and ambiguity introduced by fragmented, PCR-amplified sequencing (*1, 4*). However, this single molecule technology has still not been widely used due to cost and accuracy issues.

As of April 2014, tens of thousands organisms have had their genomes sequenced. Of these, just under 3000 are eukaryotes, of which approximately 1000 are multicellular plants and higher level animals. Genome sequences are hosted online by the NCBI (http://www.ncbi.nlm.nih.gov/genome) for public use by the scientific and general population. For organisms whose genome sequence has not yet been determined, *de novo* sequencing and assembly must be carried out. This process involves extracting and amplifying genomic DNA from cells, using high-throughput sequencing to obtain enough reads that the projected genome length is covered approximately ten times, and using an assembly method to construct contiguous runs (or "contigs") of sequence. Finally, the reference sequence is annotated with genes and regulatory elements, either using known gene sequences or gene prediction algorithms (*2*).

Gene expression profiling

For widely studied organisms (such as mouse, human, and *Arabidopsis*), well annotated genomes exist and are used to study complex questions about gene expression. Although it is an oversimplification, levels of transcribed mRNA have been successfully used as a proxy for gene expression levels. Two common technologies used to obtain a comprehensive profile of mRNA expression are microarrays and mRNA sequencing (RNA-seq). Both methods utilize mRNA extracted and purified from samples comprised of whole cells or tissues in order to probe for the presence or absence of certain genomic sequences.

Microarray gene expression profiling

Microarrays were developed in the mid-1990s using the principle of single stranded DNA hybridization (*5*). Developers pre-defined a set of sequences (called probes or probe sets) of interest encoding the genomic sequences of genes, pseudogenes, and microRNAs, and affixed these sequences to a chip, or microarray. Sample mRNA is isolated and amplified into single stranded cDNA fragments that fluoresce when hybridized to probes spotted on known locations on the chip. High resolution images of the chips are taken in which degree of binding is indicated by the brightness, or “intensity” of each spot. Microarrays are still widely used today due to their reproducibility, relatively low cost, and rigorous level of standardization. Additionally, the fact that some of the more popular microarray products (such as Affymetrix HG U133 Plus 2.0) have been in use for over a decade has resulted in a large body of publicly available microarray data.

Microarrays do have limitations, the most obvious of which is an inability to detect expression of genes or sequences not included in the set of probes. Additionally, standard microarrays do not

provide information on splice variants or give quantitative measures of gene expression levels. Finally, there are multiple microarray platforms in common use, which makes integration of datasets difficult: although the set of genes represented on differing platforms is largely overlapping, the probes that hybridize to each gene are designed differently, and may have variable properties across platforms regarding binding affinity or cross-hybridization with other genes (*6*). A 2012 study successfully combined data from multiple microarray platforms in skin lesional biopsy expression profiles in order to compare the interferon signatures of different diseases. However, this study normalized each data set separately and compared fold changes over normal, an approach that is not applicable to all data sets (*7*).

RNA-seq gene expression profiling

RNA-seq is a newer technology that addresses many of the issues seen in microarray analysis. Similar to microarrays, messenger RNA is also isolated, fragmented, and randomly amplified into single stranded cDNA molecules. These cDNA strands are affixed to a substrate while their partner strands are simultaneously and synchronously synthesized using differently colored fluorescent nucleotides. A camera takes pictures of each synthesis cycle to track the sequence of each strand. Although longer sequences will yield more accurate information, sequenced fragments ranging from 50 to 150 base pairs in length are generally a good balance between cost, accuracy, and sequencer error. The major advantage of RNA-seq is that sequences are not constrained to a predefined set of probes as in microarrays, thus enabling additional analyses such as characterization of novel transcribed regulatory factors and differential expression of splice variants. Additionally, quantification of gene transcript levels is possible. RNA-seq is gaining popularity over microarrays as it provides a more complete picture of gene transcription, but it is still a relatively expensive and specialized method and has yet to make microarray techniques obsolete. Other downsides to RNA-seq are its more complicated

analysis techniques compared to microarrays, and the often prohibitively large size of its data sets, which can require access to high power computing clusters and specialized data storage systems (*8*).

Gene expression profile analysis strategies

Gene expression profile analysis strategies – differential expression

The development of gene expression profiling technology has allowed for an unprecedented high throughput exploration of relationships between gene expression and disease phenotypes. A simple yet incredibly powerful analysis approach is the identification of differentially expressed genes. Samples are divided into two or more groups, and tests such as the Student's t-test or ANOVA are run to identify genes that are expressed at significantly higher or lower levels in one group. Fold change of gene expression relative to a common control can be used to supplement p-values to filter for genes that are most differentially expressed. Notably, these methods are optimized for tests of differential expression between two groups. While ANOVA compares values from three or more groups, it only identifies genes that are differentially expressed in at least one group, and a post-hoc test such as Tukey's test must be run to determine which group is differentially expressed. Furthermore, there are few standardized methods for calculating fold changes that capture comparisons between more than two groups (*9, 10*).

Gene expression profile strategies - classifiers

Classifier training is another common use for gene expression profiles. The most simple classification scheme is unsupervised hierarchical clustering, in which correlations between gene expression patterns are used to group samples (*11*). Clustering is often used as an initial step to both visualize data and confirm that a robust biological signal is present in the data.

Data that fails to cluster properly merits further action, such as a different normalization scheme or filtering methods. Clustering can also indicate whether data from different batches manifest significant batch effect. Additionally, visualizing clustered genes in heatmaps can highlight groups of genes that are associated with each disease.

Classifiers fall into two categories: binary or multi-class. Binary classifiers distinguish between two phenotypes and are commonly implemented using linear classifiers (such as support vector machines or logistic regression), which can be conceptualized as drawing a line that divides a series of points into two sets (*12*). Multi-class classifiers (multi-classifiers) are a more complex problem in which samples are divided into three or more categories. Although some binary classification methods can be modified for three or more classes, there are fewer methods for multi-classification. Common approaches include decision trees and random forests.

All classifiers use a set of features, or sample characteristics to distinguish between phenotypes – in the case of gene expression profiles, features are genes or microarray probes. Gene expression profiles contain on the order of 10,000 genes or probes; however, the use of tens of thousands of features to build a classifier can lead to overfitting, or spuriously good classifier performance due to artifact. Often, this results from having many more classifier features than training samples (*13*). Therefore, feature selection, or the inclusion of only those probes most informative to the classification process, is crucial to building a classifier that is not only accurate but widely applicable to other data sets. Feature selection can be as simple as a gene count or microarray intensity threshold, or can be a complex iterative process implemented before, during, and after the classifier has been built. It is crucial to separate feature selection from any samples used to validate the classifier, and ideal to not use phenotypic information to select features, as these can introduce bias in the classifier (*14*).

Molecular classifiers have enormous potential in translational medical research. Current clinical diagnostic procedures rely heavily upon subjective measures such as histological observations, which can lead to ambiguous or inaccurate diagnoses. Accurate classification based upon objective criteria such as gene expression could simplify the diagnostic process, as well as provide clinicians with treatment strategies. For example, a 2009 study used microarrays to successfully distinguish between benign nevus, a condition that calls for watchful waiting, and melanoma, a neoplasm that requires immediate excision (*15*). Gene expression classifiers also have the potential to identify biomarkers of diagnosis or prognosis, or discover genes that provide insight into disease pathogenesis. However, the relationship between genes that are useful for classification and genes that impact disease pathogenesis is not clear. Classifier based studies have yet to produce gene signatures that have directly impacted therapeutic practices at a level that other high throughput genomic technologies, such as cancer genome sequencing, have achieved (*16, 17*).

Gene expression profile studies – functional analysis

Historically, gene expression studies that have impacted clinical management have used high throughput expression profiling methods to obtain candidate gene signatures, which are then confirmed via previous studies, bench experiments, and patient studies. For example, microarray studies of breast cancer tumor samples has helped shape the clinical management of the disease by identifying at least five major subtypes of breast cancer using gene expression signatures (*18*). This is particularly useful in a heterogeneous disease like breast cancer, where tumor samples have a wide range of histological features and patients have variable responses to treatment. Single gene biomarkers have been identified for four of these five subtypes, and this subtype structure has been incorporated by the Saint Gallen Consensus Conference

guidelines for breast cancer management since 2011 (*19, 20*). In the specific subtype of hormone-responsive breast tumors, a 2004 study by Ma, et al. used microarrays to obtain gene expression profiles from hormone-responsive patients who had been treated with tamoxifen (*21*). Profiles from patients whose breast tumors had recurred were compared with patients who remained disease free, and a gene signature was identified that was correlated to patient survival and confirmed in lesions by RT-PCR. Further analysis of survival time allowed the signature to be refined to the ratio of two genes that accurately predicted patient response to tamoxifen treatment and advised the best course of therapy for each patient.

However, individual genes that provide insight into disease phenotype are not always so easily identified. In more ambiguous cases, downstream functional analysis of gene signatures can detect groups of genes that participate in a common molecular function or biological pathway. Most functional analyses involve the comparison of expected versus observed genes that are annotated with the same biological process or pathway, calculated using the hypergeometric distribution (*22, 23*). Functional annotation resources such as the Gene Ontology, Kegg, and Biocarta store comprehensive databases of gene-function relationships, which not only facilitate such searches but also provide a standardized, central resource. Free and paid services such as DAVID Functional Annotation and Ingenuity Pathways Analysis synthesize relationships from multiple annotation databases, as well as the body of scientific literature, in order to identify the most statistically significant enriched pathways (*23-25*). Many of these services also use known gene interactions to build networks that visualize relationships between genes. These resources enable researchers with little computational background to easily perform functional enrichment analyses.

The relationships in annotation databases and network building tools are based on known gene interactions from previously published data, but gene expression profiles can also be mined for *de novo*, or previously undiscovered, gene interactions. One strategy for finding such connections is to compare gene expression patterns across many diseases, searching for genes that are consistently co-expressed. Calculating the pairwise Pearson correlation, in which each gene is treated as a vector of expression values across diseases, is one such metric that identifies groups of genes that either all have high expression or all have low expression across a range of samples. Such groups are often biologically relevant since the co-expression may indicate a shared pathway or upstream regulator. Weighted Gene Correlation Network Analysis (WGCNA) is a commonly used tool that uses correlation to group genes, but adds a weighting step that gives more weight to high correlations while still considering the information encoded by lower correlations (*26*). Importantly, by calculating gene connectivity within each module of highly correlated genes, WGCNA enables the construction of novel gene networks not based upon previously published data.

Publicly available gene expression profile data

High throughput gene expression profiling technology generates large amounts of data that can be re-used and re-mined in additional analyses. Furthermore, gene expression profiling experiments are expensive and may require biopsy specimens from rare or difficult to obtain conditions. In order to facilitate the sharing of gene expression profile data, and to encourage time and cost efficient scientific practices, public data repositories such as the NCBI Gene Expression Omnibus (GEO) were established (*27*). High impact publishing groups such as Nature and PLoS have contributed to the body of public data by requiring authors to deposit data in public repositories such as GEO prior to article publication. GEO provides standard-format data from high throughput experiments, including microarray and RNA-seq data. Raw

data files are available to download from many sets, making the information even more amenable to a variety of uses.

There are a number of common strategies in the analysis of public data. The simplest approach involves re-analyzing published data: for example, calculating differential expression using a sample partitioning that differs from the original study. Additionally, data from multiple experiments may be combined in a meta-analysis, as in a 2013 study that integrated 5 batches of microarrays into a single analysis to find gene expression differences between lesional and non-lesional skin from psoriasis patients (*28*). This combination strategy is appealing since it enables an increase in statistical power without the associated effort and cost of generating additional data. Integration of data from the same condition and platform is usually straightforward, especially if sample size is sufficient or there is a common control to aid in normalization. However, further considerations are necessary if data must be integrated across conditions, batches, or platforms.

A major issue when integrating data from public sources is batch effect, or noise that is non-random, specific to samples obtained at one time and place, and does not represent a true biological difference. Batch effect has been shown to cause up to a third of the variation present in data from multiple sources, and can result in false identification of differentially expressed genes (*29, 30*). Mathematical removal of batch effect can be achieved using software packages such as ComBat, which uses PCA-based methods to identify axes of greatest variation between sample, which are assumed to contain the majority of batch effect noise (*29*). However, these methods are limited in that all batches must contain samples from every condition in the analysis in order to distinguish between batch effect artifact and true

biological effects. For analyses containing singleton data sets, which contain only one condition per batch, batch effect removal is therefore not possible.

The effect of combining data from different batches can also be mitigated by an alternative normalization scheme, such as frozen Robust Multi-Array Average (fRMA). Instead of using within-set variation to build a normalization distribution, fRMA draws from a prebuilt, standardized external data set (*31*). This method assumes that the larger sample size of the external data set will better capture the full spectrum of variation between gene expression profiles, thus minimizing the effect of between-batch differences. Additionally, the use of the same external data set for normalization allows for data to be normalized separately and integrated at any step in the analysis, compared to traditional RMA where all samples must be normalized in one step. fRMA currently only has pre-built reference datasets for three microarray platforms, but a 2011 release of the software allows users to create customized reference databases for any platform (*32*). However, all data must come from the same platform, and the curation of such a reference database is not trivial: the original fRMA publication built reference databases with a balance of data from different tissue and experiment types, and drew from approximately 6000 public data samples (*31*).

Conclusion

Gene expression profiling methods have enabled the widespread discovery of genes and pathways associated with disease. While analyses of gene expression profiling data have contributed to major advances in understanding the causes and possible treatments of human disease, the immense complexity of genetic information, the cost of obtaining new samples, and conflicting or competing analysis techniques have left many questions unanswered.

The dissemination of publicly available gene expression profiles has provided further options for researchers in the form of existing datasets that can be mined to study additional questions. The vast body of public gene expression profile data is ever expanding, presenting researchers with instant access to low cost high-throughput data on a scale beyond what is feasible for the average research lab. However, the integration of public data from a variety platforms, sources, and batches still presents a difficulty, especially in one-off studies containing one disease state per batch. While this particular problem has not been solved yet, part of this work is devoted to the identification of robust signals that persist beyond batch effect.

Gene transcription is only one step of the process by which the information encoded in DNA affects biological processes and ultimately phenotypes. Pre or post transcriptional processes such as epigenetic methylation of DNA or degradation of transcribed mRNA by small RNAs affect gene expression levels in ways that are totally invisible to those studying gene expression profiles. However, gene expression profiles remain a backbone of translational research as an easily accessible, well developed technology that provides a comprehensive snapshot of cellular or tissue gene expression.

Chapter 2: Methods

Microarray data normalization, filtering, and preparation

Data is obtained by isolating mRNA from tissue or cell samples, as described in Bleharski et al., 2003, and hybridizing to microarrays (*33*). The microarray platform used in these analyses is Affymetrix HG U133 Plus 2.0, a comprehensive human genome expression array that contains 54675 probe sets representing over 20,000 unique genes, pseudogenes, and transcripts. This array is one of the more commonly used microarray platforms.

Before analysis can be carried out, data must be normalized to make the distribution of probe set intensities from each sample approximately equal. Normalization can be accomplished by comparing intensity values in each microarray to itself (such as in MAS5), or to the other microarrays in the data set (RMA, frozen RMA). The MAS5 algorithm is exclusive to Affymetrix arrays and uses the difference between perfect match (PM) and mismatch (MM) probes (i.e. signal and negative control, respectively) to act as an indicator of background noise and cross hybridization (*34*). Robust Multi-array Average (RMA) is a newer technique that uses information from all data sets to build a background distribution that is used to correct probe set intensities, after which data is quantile normalized and summarized (*35*). RMA does not use Affymetrix MM probes to perform background correction; however, expression of MM probes has been shown to be noisy and unreliable (*34*). Frozen RMA (fRMA) uses the same principles, with the exception that the background distribution is generated using an external, fixed set of microarrays curated by the authors of the software (*31*).

Before analysis, normalized data must be filtered to remove low intensity probes. Historically, using a flat threshold of mean intensity greater than 100 has yielded satisfactory results, based on the assumption that measurements below this threshold are unreliable and may be noisy. A more inclusive scheme for data sets with multiple conditions retains out probe sets with intensity greater than 100 in any one condition. More rigorous methods can be used to filter, including

filtering by probe variance or using detection call rates from the microarray. Microarray experiments are performed under the assumption that only a minority of probes will be differentially expressed, therefore a large minority of probes may be discarded in the filtering process (*36*). Final data processing steps may be taken after filtering, such as normalizing each probe by its mean intensity across all samples, or log transforming the data.

Proportional Median

When comparing three or more conditions, conventional methods such as fold change are not sufficient to capture relationships. We used the proportional median (PM) metric to identify probe sets that had relatively higher expression in one condition compared to two or more other conditions. Like fold change, PM is calculated for each probe set in each disease. The PM of probe X in disease Y is equal to the median intensity of X across the samples in Y, divided by the median intensity of X across all samples in all conditions. This yields a list of values for each disease, which can then be sorted in descending order. It is important to note that PM values should be considered relative to each other within a disease, but not compared across diseases. Rather, PM can be used to rank probe sets within a disease, after which probe sets can be referred to by rank in subsequent analyses.

Random forest classification

Random forest (RF) is a classification method that can be used high-dimension data for which there is a large feature space, include those for which there are more features than samples. For classification using a training set consisting of microarrays, the RF algorithm builds a user-defined number of decision trees, each using a randomly selected subset of the training samples and a randomly selected subset of probe sets. A random subset of samples is used to train each tree, with the remaining samples (termed “out of bag”) used to assess tree

performance. For each tree, a random subset of probe sets is selected to build that decision tree. Out of bag or test samples are evaluated by each decision tree independently and majority voting determines the aggregate classification (*37, 38*).

For this work, we performed feature selection using PM. We calculated PM values for each disease, ranked probes by PM, and took the top 25 probes. We used the unique set of these genes as input for our classifier.

Weighted Gene Co-expression Network Analysis

Weighted Gene Co-expression Network Analysis (WGCNA) is an analysis technique developed by the Horvath lab at UCLA and available as an R package (*26*). WGCNA takes a gene expression profile as input and detects modules of genes that share similar expression patterns. WGCNA is similar to traditional clustering analyses in that nodes (genes or probe sets) are treated as vectors of expression values for all samples in the analysis. Pairwise correlation between nodes is computed and stored in a similarity matrix. WGCNA then computes an adjacency matrix from the similarity matrix based on a weighted soft thresholding scheme, where each value in the similarity matrix is raised to a power supplied by the user. The power can be selected empirically, and the WGCNA package provides functions to estimate suitable values.

Once modules are constructed, module eigengenes are calculated for each module by taking its first principle component. These eigengenes are used as representatives that capture gene expression in the entire module. Additionally, intramodular connectivity is calculated for all genes within each module by correlating a particular gene's expression to its module eigengene. The genes with highest intramodular connectivity will typically have functional

annotations that reflect the biological pathways enriched in that module (*39*). Module eigengenes can also be correlated to a binary matrix representation of sample phenotype (where a sample has a '1' value for its phenotype and '0' values for all other phenotypes) to identify modules that are significantly correlated with specific phenotypes. Finally, gene networks are constructed for the most highly connected genes in each module by displaying their topological overlap, which is a measure that identifies the number of shared connections between pairs of genes (*26, 40*)

Chapter 3. Characterization of gene expression profiles from pancreatic cancer

Introduction

Human pancreatic ductal adenocarcinoma (PDAC) is an extremely aggressive form of cancer with a high mortality rate and limited options for treatment. As the tenth most common cancer diagnosis, PDAC contributes the fourth highest number of annual cancer deaths, not only indicating a heavy disease burden, but also representing a disproportionate mortality rate compared to other malignant cancers (*41*). The current five year survival for PDAC is less than 4% (*42*).

Multiple factors contribute to the high mortality rate in PDAC. On an anatomical level, the pancreas sits in close proximity with vital organs and blood vessels that provide an easy path of metastatic spread. Additionally, the placement of the pancreas in the dorsal portion of the abdomen prevents early stage tumors from being physically palpated or visualized unless specific imaging tests are run. PDAC tumors tend to be asymptomatic until the disease has reached a very late stage, and even these late stage symptoms can be non-specific. Finally, for reasons unknown to researchers, PDAC spreads aggressively such that even early stage tumors may have already spread to vital organs such as the liver or lung. These factors make the detection of PDAC in its early stages especially difficult (*41*).

Currently, the only treatment for PDAC that has a significant improvement in survival is surgical excision of all lesions. However, this is only a viable therapeutic option in early stage tumors. Due to the factor stated above, 85% of patients present with tumors in the metastatic stage, for which surgical resection is not a possibility. Furthermore, only 20% of patients survive five years past surgical treatment, perhaps due to the presence of undetectable micro-metastases (*43*). Despite decades of research, the causes and mechanisms of PDAC are still unclear to clinicians and scientists, and few advances have been made in its treatment. In this chapter, we investigated genes that are associated with survival in patients diagnosed with PDAC. Using

microarrays performed on surgically removed tumors from patients with varying survival times, we identified genes that were significantly correlated to survival time. We used gene expression profiles derived from two sets of microarray data: flash-frozen PDAC samples, and formalin-fixed, paraffin-embedded (FFPE) samples.

Results

We worked with two sets of gene expression profiles obtained from de-identified surgical biopsy samples of human PDAC tumors for the preliminary portion of this project: one set from 42 flash-frozen tumors, and one set from 30 formalin fixed, paraffin embedded (FFPE) tumors. Both batches of data were derived using the Affymetrix HG-U133 Plus 2.0 microarray platform. The first set of experiments we analyzed was derived from flash-frozen PDAC tumors, which were obtained from collaborators in the Hong Wu lab (Department of Molecular and Medical Pharmacology, UCLA). The goal of this analysis was to identify genes that are correlated with disease free survival times (DFS). The microarrays were normalized using RMA (Robust Multi-array Average) and filtered at a mean raw intensity of 100, since intensity measurements below this level tend to be unreliable (*44*). We used a Cox proportional hazards regression test to determine which genes were correlated with DFS, and conducted random permutations to find genes that were significant according to a false discovery rate (FDR) of 5%. Only about 20 genes were found to be significantly associated with survival time, perhaps due to the small range of survival times (Figure 1).

To address the issues with the narrow range of survival times, we used an FFPE set of PDAC microarray data, which were generated by a second collaborator, David Dawson (Department of Pathology, UCLA). Although mRNA in FFPE tissue can suffer from degradation, it is still an appealing alternative over snap-frozen as researchers typically have more ready access to FFPE specimens from a wide range of clinical histories. The 30 tumor samples Dr. Dawson selected represent a larger spread of DFS (Figure 1).

The data was normalized using RMA, genes were filtered with a mean intensity less than 100, and a Cox regression analysis with random permutations was performed to find genes that were significantly associated with survival above a 5% FDR. From this analysis, 37 significantly

associated genes were found, four of which our collaborators validated via quantitative polymerase chain reaction (qPCR) in a separate group of 31 FFPE PDAC samples (Table 1). These validated genes were KRT10 (keratin 10), SLC20A1 (solute carrier family 20, member 1, a phosphate carrier), SIRT5 (sirtuin 5), and PHLDA2 (pleckstrin homology-like domain, family A, member 2), and while none of these have an established link to pancreatic cancer, they function in processes such as cell structure, metabolism, and epigenetic regulation of gene expression (*45-48*). Additionally, PHLDA2 is located on a tumor suppressor locus, 11p15.5 (*48, 49*).

Patients were separated into two groups, “low survival” and “high survival,” based on the clinical knowledge of our collaborators. Hierarchical clustering of patients and heatmap visualization of gene expression, both for all 37 differentially expressed genes and for the four experimentally validated genes, effectively separated patients into high and low survival groups (Figure 2). For each of the four validated genes, Kaplan-Meier curves were drawn, which split patients between those with gene expression higher (blue) or lower (red) than the median (Figure 3). Although the number of patients is not large enough to calculate robust statistics, there is a clear separation between survival times according to gene expression of these four genes.

Discussion

Despite major advances in the field of cancer treatment, PDAC remains a devastating diagnosis for the overwhelming majority of patients. The combination of patient presentation in the later stages of disease and the particularly aggressive nature of the tumor contribute to its extraordinarily low five year survival rate, even in patients eligible for surgical treatment. In this brief analysis, we were able to identify 37 genes significantly associated with PDAC survival time, including four that were validated in independent lesions via qPCR.

These pancreatic tumor samples were obtained from patients who present with surgically resectable lesions who represent only a small portion of the total patients diagnosed with PDAC. However, the applications of this study are not limited to early stage patients. The lack of fully representative PDAC mouse models and the fact that patients with late stage disease are rarely operated upon make these early stage tumors the most viable option to study lesional gene expression profiles (*50, 51*). Identifying genes associated with survival time can give insight into the pathophysiology of PDAC, potentially providing insight into such big-picture questions as how lesions arise and why the tumors are particularly aggressive. Furthermore, genes that contribute to longer DFS times could provide biomarkers of protection, or novel targets for therapy.

Methods

Lesional samples of pancreatic ductal adenocarcinoma were obtained from surgically resected specimens. All samples came from patients eligible for surgical treatment of PDAC (i.e., no distant metastases). 42 samples were flash frozen immediately after excision, and 30 samples were obtained from archival FFPE samples. For each sample, corresponding patient data such as sex, age, tumor stage, and disease free survival (DFS) were collected. Patients were de-identified before RNA was extracted and gene expression profiles were derived via HG U133 Plus 2.0 Affymetrix microarrays. qPCR validation was carried out using the 30 FFPE samples by our collaborators.

Samples were normalized using RMA implemented in the Matlab statistics toolbox. Data was filtered at a mean normalized intensity of at least 100 in each gene. Cox tests were performed on the 30 FFPE samples using gene expression values and right censored DFS, and a permutation analysis with 1000 permutations was carried out to adjust for a false discovery rate of 5%. Clustering and heatmap visualization was performed using the Clustergram functionality in Matlab. Kaplan-Meier curves were drawn by calculating the empirical cumulative distribution function (ecdf() in Matlab), using gene expression and right censored DFS from the 30 FFPE samples.

Figure Legends

Figure 1. **Boxplots of disease free survival (DFS) for flash frozen and formalin-fixed, paraffin-embedded (FFPE) pancreas tumor samples**. Boxplots show disease free survival time for all patients from which tumors were taken, separated by batch of data. The bottom and top edges of the boxes denote the 25th and 75th percentile, respectively; the middle bar denotes the median

Figure 2. Heatmaps of genes in PDAC tumors significantly associated with disease free survival time. A cox test with random permutations was performed to identify genes significantly associated with disease free survival time in 30 PDAC FFPE tumors. A clustered heatmap of gene expression segregates patients into high survival time (samples labeled "H") and low survival time (samples labeled "L") all 37 genes identified by cox regression (**A**) and in four genes validated by qPCR (**B**). Samples are shown on the X axis, and genes on the Y axis.

Figure 3. Kaplan-Meier curves for genes associated with PDAC survival time. Kaplan-Meier curves were drawn to show differences in disease free survival for patients with high expression of each gene (blue) versus low expression (red) in four genes: SLC20A1 (**A**), KRT10 (**B**), SIRT5 (**C**), and PHLDA2 (**D**). Disease free survival is shown in months on the X axis, and percentage of patients alive at that time point is shown on the Y axis. Plots are right censored to account for patients who were still alive at the time of the study.

Figure 1

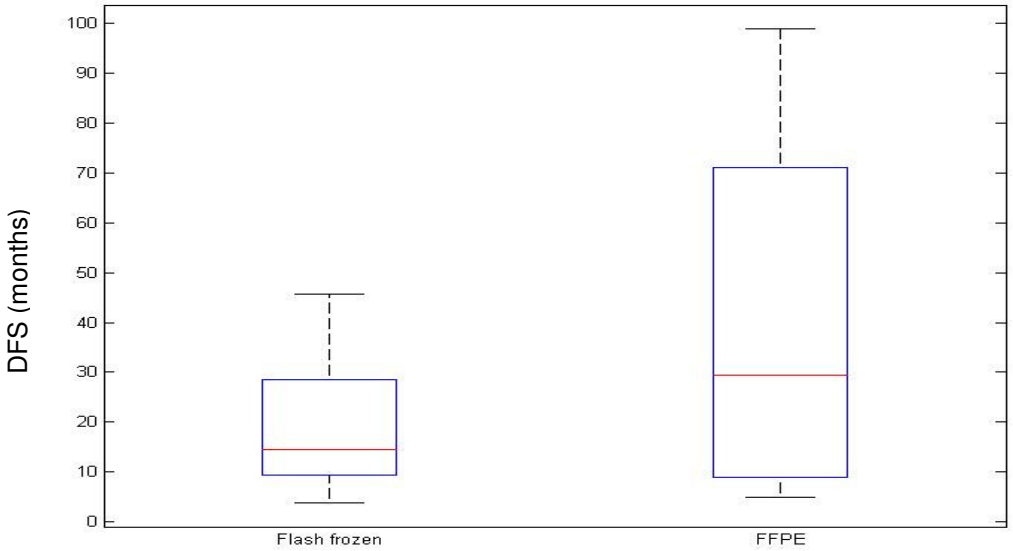
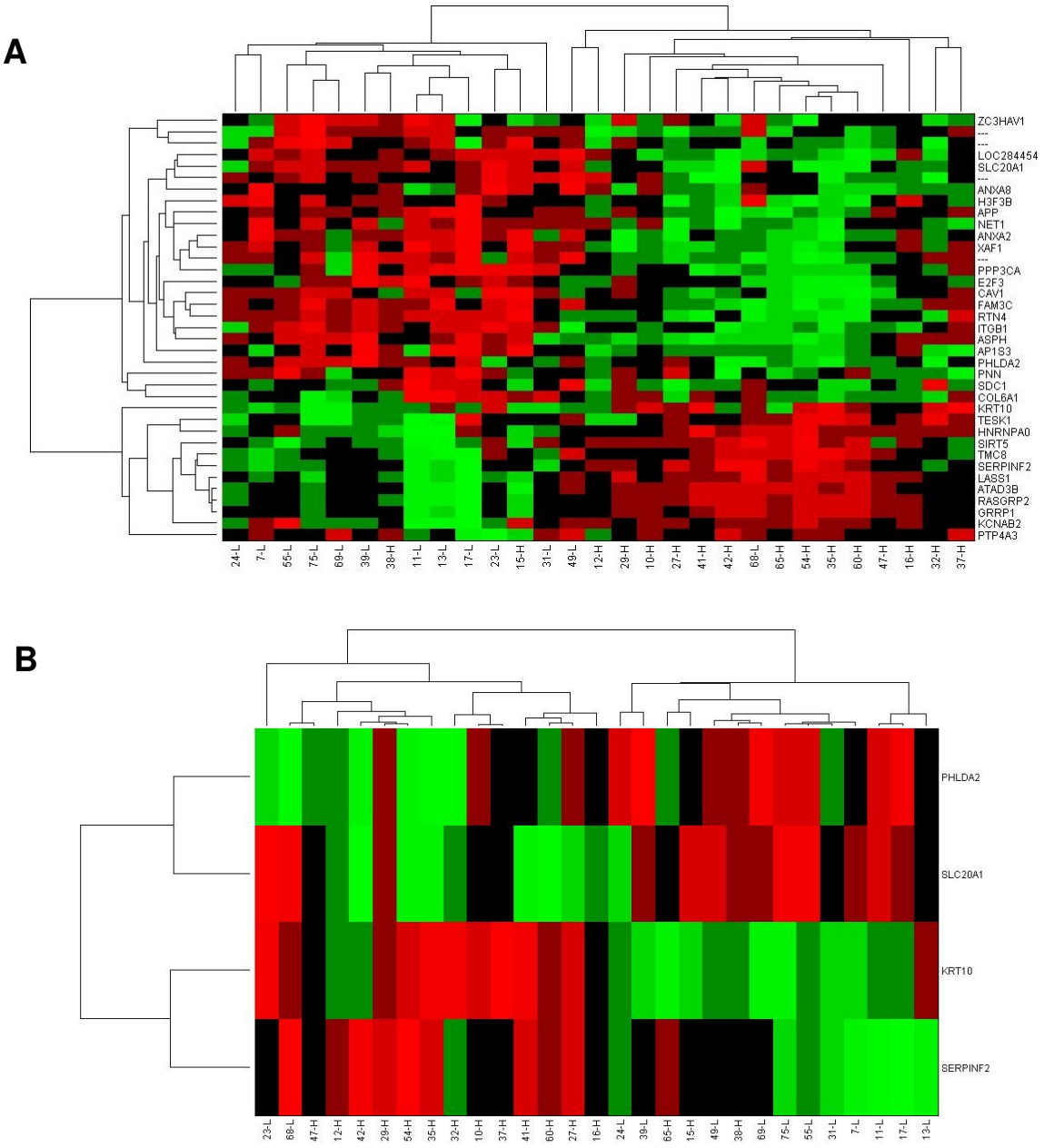


Figure 2

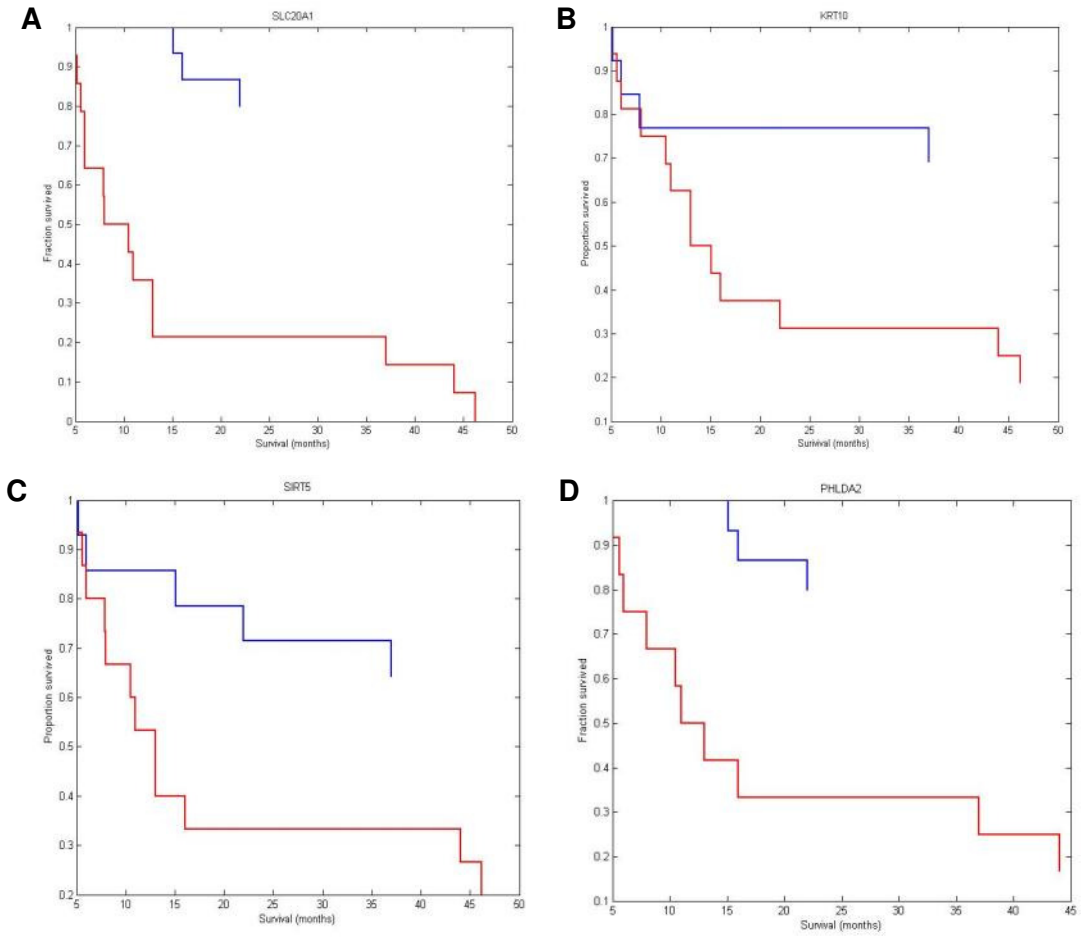

A
ZC3HAV1

LOC284454
SLC20A1

ANXA8
H3F3B
APP
NET1
ANXA2
XAF1

PPP3CA
E2F3
CAV1
FAM3C
RTN4
ITGB1
ASPH
AP1S3
PHLDA2
PNN
SDC1
COL6A1
KRT10
TESK1
HNRNPA0
SIRT5
TMC8
SERPINF2
LASS1
ATAD3B
RASGRP2
GRRP1
KCNAB2
PTP4A3
24-L 7-L 55-L 75-L 69-L 39-L 38-H 11-L 13-L 17-L 23-L 15-H 31-L 49-L 12-H 29-H 10-H 27-H 41-H 42-H 68-L 65-H 54-H 35-H 60-H 47-H 16-H 32-H 37-H
B
PHLDA2
SLC20A1
KRT10
SERPINF2
23-L 68-L 47-H 12-H 42-H 29-H 54-H 35-H 32-H 10-H 37-H 41-H 60-H 27-H 16-H 24-L 39-L 65-H 15-H 49-L 38-H 69-L 75-L 55-L 31-L 7-L 11-L 17-L 13-L

Figure 3


A
SLC20A1
Fraction survived
Survival (months)
B
KRT10
Proportion survived
Survival (months)
C
SIRT5
Proportion survived
Survival (months)
D
PHLDA2
Fraction survived
Survival (months)

Tables

Table 1. Differentially expressed probe sets from PDAC. Cox regression was performed to identify genes significantly associated with disease free survival time using gene expression profiles derived from 30 FFPE samples of PDAC. P-values were calculated using random permutations. Table continues on next page.

Table 1

Probe Set ID	Gene Symbol	P-value	Cox Score
1555847_a_at	LOC284454	0.002479	4.181493
205075_at	SERPINF2	0.002751	-3.95961
229112_at	SIRT5	0.009003	-3.53477
209069_s_at	H3F3B	0.002664	3.528287
210633_x_at	KRT10	0.006582	-3.51858
201055_s_at	HNRNPA0	0.003421	-3.30874
201920_at	SLC20A1	5.49E-05	3.221104
210183_x_at	PNN	0.008889	3.165033
204106_at	TESK1	0.008377	-3.02575
242644_at	TMC8	0.007954	-2.86055
203402_at	KCNAB2	0.004273	-2.75377
216190_x_at	ITGB1	0.006715	2.682047
203693_s_at	E2F3	0.00505	2.58083
220104_at	ZC3HAV1	0.002692	2.569461
209695_at	PTP4A3	0.004402	-2.4177
209803_s_at	PHLDA2	0.004642	2.266736
214953_s_at	APP	0.004736	2.171862
213428_s_at	COL6A1	0.006686	2.162502

202425_x_at	PPP3CA	0.003867	2.147291
229448_at	LASS1	0.007097	-2.01492
201829_at	NET1	0.003527	1.931678
214368_at	RASGRP2	0.002564	-1.86437
201286_at	SDC1	0.006043	1.731349
223617_x_at	ATAD3B	0.009448	-1.71537
238869_at	---	0.009118	1.682259
220430_at	GRRP1	0.006017	-1.56053
230494_at	---	0.007283	1.528856
237159_x_at	AP1S3	0.00682	1.41077
213503_x_at	ANXA2	0.007987	1.360014
1567219_at	---	0.009351	1.358106
225239_at	---	0.00966	1.207571
211509_s_at	RTN4	0.009551	1.170515
228617_at	XAF1	0.004465	1.070455
201889_at	FAM3C	0.005237	0.994979
209135_at	ASPH	0.005139	0.957955
203074_at	ANXA8 /// ANXA8L1 /// ANXA8L2	0.006878	0.840354
212097_at	CAV1	0.004835	0.782334

Chapter 4: Comparison of Molecular Signatures from Multiple Skin Diseases Identifies Mechanisms of Immunopathogenesis

Comparison of Molecular Signatures from Multiple Skin Diseases Identifies Mechanisms of Immunopathogenesis

Authors: Megan S. Inkeles[1], Philip O. Scumpia[2], William R. Swindell[3], David Lopez[1], Rosane M.B. Teles[2], Thomas G. Graeber[4], Stephan Meller[5], Bernhard Homey[5], James T. Elder[3,6], Michel Gilliet[7], Robert L. Modlin[2,8]*, Matteo Pellegrini[1]*

Affiliations:

[1]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095 USA.

[2]Division of Dermatology, University of California, Los Angeles, CA 90095 USA.

[3]Department of Dermatology, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA,

[4]Crump Institute for Molecular Imaging, Institute for Molecular Medicine, Johnson Comprehensive Cancer Center, California NanoSystems Institute, Department of Molecular and Medical Pharmacology, University of California, Los Angeles, CA 90095 USA.

[5]Department of Dermatology, University Hospital Duesseldorf, D-40225 Duesseldorf, Germany.

[6]Ann Arbor VA Hospital, Ann Arbor, MI 48105, USA

[7]Department of Dermatology, University Hospital Lausanne, CH-1011 Lausanne, Switzerland.

[8]Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, CA 90095 USA.

*To whom correspondence should be addressed: rmodlin@mednet.ucla.edu (RLM), matteop@mcdb.ucla.edu (MP).

Abstract

The ability to obtain gene expression profiles from human disease specimens provides an opportunity to identify common and distinct mechanistic pathways. However, multi-disease comparisons have been limited by the absence of data sets spanning a broad range of conditions. Our objective was to perform a concurrent study of multiple diseases in a single tissue in order to gain insight into disease pathogenesis. We performed an integrative analysis of publicly available microarray data from skin biopsy specimens comprising 16 conditions. Individual samples clustered by disease, from which disease-specific gene signatures were identified and validated using a random forest classifier that accurately predicted the diagnosis of publicly and prospectively collected samples. In one sample, the molecular classifier differed from the initial clinical diagnosis and correctly predicted the eventual diagnosis as the clinical presentation evolved. Unsupervised hierarchical clustering of assembled gene expression profiles yielded distinct disease groups according to common cellular and molecular pathways. Finally, when the expression of Type I versus Type II interferon (IFN) regulated gene programs was integrated with the skin database, a significant inverse correlation between IFN–β and IFN–γ programs was found across all conditions. Our study provides an integrative approach to the study of gene signatures from multiple skin conditions, providing insight into disease pathogenesis.

Introduction

Gene expression profiling technology, such as microarrays, provides the opportunity to identify disease specific genes and pathways. The NCBI Gene Expression Omnibus (GEO) is a community resource of publicly available data from multiple diseases, including less common diseases studied in specific laboratories (*27*). The ability to mine data from GEO provides a tremendous opportunity to compare gene expression profiles across multiple diseases.

Using GEO data, disease profiles have previously been compared across multiple data sets by normalizing to controls within each set; however, this practice limits the use of available data to those containing equivalent control profiles (*7, 52*). To overcome this limitation, data integration from multiple sources, particularly those from experiments containing a single disease, can be achieved with Frozen Robust Multi-array Average (fRMA), which normalizes samples to a standard reference set of microarrays, eliminating the need for control samples in each data set (*31*). Here we present a study that uses this approach on gene expression profiles derived from skin biopsy specimens. The skin represents an ideal organ to study molecular signatures at the site of disease due to the ease of access for lesional biopsies and its diverse manifestations of pathology. We assembled a database of publicly available skin microarray samples representing 16 inflammatory, infectious, and neoplastic conditions. This database was used to construct a classifier, perform functional analyses to identify representative pathways, and establish a spectrum of differentially expressed Type I vs. Type II interferon gene programs across these diseases.

Results

Data normalization with Frozen RMA

We searched the NCBI Gene Expression Omnibus (GEO) for microarray experiments performed on human skin samples associated with a dermatological disorder (www.ncbi.nlm.nih.gov/geo) (*27*). Data files for microarray analysis of 311 skin biopsy samples were downloaded, representing 16 conditions from 15 experiments and 14 laboratories (methods, Table S1). All samples used the Affymetrix HG U133 Plus 2.0 platform. In order to analyze this data as a single set, Frozen RMA (fRMA) was used to normalize samples (*31*). Normalized data from all series showed comparable probe set intensity distributions (Figure S1).

After unsupervised gene clustering, samples segregate by disease as well as groups of diseases with related pathogenesis

In order to determine whether the batch effects within a given disease were smaller than the differences between diseases, trees of filtered gene expression profiles were constructed for both samples and diseases (Figures 1, S2). Remarkably, we found that in diseases in which there were multiple batches of microarrays from different sources, including psoriasis, atopic dermatitis and leprosy, the batches from the same disease nearly always clustered together, despite coming from independent data sets (Figure 1). Furthermore, five batches of normal skin, each obtained from healthy control subjects from different laboratories, clustered together with little differentiation by batch. However, batch effects were not completely eliminated as samples from specific diseases often separated by lab or experiment. Furthermore, there were isolated cases of individual samples clustering with the incorrect disease (represented as a leaf with a different color than its neighbors in Figure 1). Squamous cell carcinoma (SCC) and basal cell carcinoma (BCC) samples from a single lab were split into two groups, with 17 samples clustering on the same branch as psoriasis and seven clustering in close proximity to irritant and

allergic conditions. Overall, these results suggest that the fRMA approach used in this context allows us to minimize the effect of batch and allow true disease signatures to predominate.

When the higher-level structure of the tree was examined, we found that branches of the tree could be annotated as disease groups with related pathogenesis. These distinct groups were categorized according to the following descriptions: i) keratinocyte proliferation and neoplastic growth (psoriasis, and approximately half the SCC and BCC samples), ii) wound (post-operative wound, burn), iii) normal, iv) allergic (allergic contact dermatitis, atopic dermatitis), v) malignant (mycosis fungoides, melanoma); and, vi) infectious (leprosy, chancroid). These relationships are consistent with those seen in the unrooted disease tree, which was built with one leaf per disease by averaging distances between all pairs of samples (Figure S2).

Proportional median metric for identifying disease specific gene signatures

Gene signatures were identified for each disease in order to build a disease classifier and perform downstream functional analysis. We developed the "proportional median" (PM) metric to identify highly expressed gene probe sets for each disease. The PM of a microarray probe set X in disease Y represents how highly expressed X was in Y, compared to all other diseases. Probe sets were ranked by PM for each disease, yielding 16 individual lists that corresponded to relative gene expression levels associated with each disease. Because lowly ranked genes often have low intensity and tend to be noisier than those with higher ranks, subsequent analysis utilized genes with high PM values (*53, 54*).

Random forest classifier accurately predicts disease diagnosis

We built a random forest multi-classifier using our disease expression profiles to predict disease status based on the expression of a limited number of signature genes (*37*). Briefly, the random

forest algorithm selects subsets of samples and genes to iteratively build multiple, parallel decision trees. The combination of random subset selection and iteration reduces the effect of noise and outliers on classifier training. In addition, cross-validation is built into the classifier training process by testing each decision tree with samples not used to build that tree. We used this classifier to assign each sample to one of the 16 conditions.

The classifier was trained on a list of PM-filtered probes, which we found improves classifier accuracy (*13*). Classifier training with 100 trees produced an error rate of 4.5% (Figure S3). Classifier performance for each disease was assessed by sensitivity, specificity, precision, and F1 score, a measure of accuracy equivalent to a weighted average of sensitivity and precision that has values between 0 (poor accuracy) and 1 (perfect accuracy). Sensitivity values ranged from 0.86 and 1.00 (mean 0.96) and specificity values ranged from 0.99 to 1.00, which corresponds to an average of 96% of disease samples being accurately classified and 99% of negative classifications being correct. The range of F1 scores was from 0.86 to 1.00 (mean 0.96) (Table S2). We also performed three-fold cross validation, which yielded aggregate F1 scores from 0.75 to 1.00 (mean 0.96) (Table S3).

To assess the classifier's robustness to overfitting and its ability to generalize, two-fold cross validation with separation by batch was carried out. We separated data into two groups of approximately equal size: in multi-batch conditions (leprosy, psoriasis, atopic dermatitis, and normal), data was separated according to batch; otherwise, data was randomly and equally partitioned. Two independent classifiers were built, each trained on one of the partitioned sets and tested on the other. PM values used in feature selection were computed only from one group of data to ensure that test data did not bias the classifier. Diseases which were separated by batch had F1 scores between 0.90 and 0.95 (mean 0.92), and diseases which

were randomly separated had F1 scores between 0.71 and 1.00 (mean 0.92), indicating little loss of accuracy with batch separated cross validation (Table 1).

To evaluate the classifier's performance on data not used for any of the previous analyses, we tested the 16 disease classifier trained on all data using gene expression profiles derived from additional skin biopsy specimens. These additional gene expression profiles were not included in the feature selection and training steps of this classifier. We generated 26 de-identified gene expression profiles derived from biopsy specimens of leprosy, psoriasis, atopic dermatitis, and normal skin. We also found public data for 168 gene expression profiles derived from biopsy specimens of psoriasis, atopic dermatitis, and melanoma (Table S4). The classifier was used to assess these 194 validation samples, yielding an overall sensitivity of 0.93, specificity of 0.99, and F1 statistics between 0.76 and 1.0 (Table 2).

Upon follow-up of the patients from whom we had collected samples, we discovered that one that was clinically diagnosed with atopic dermatitis, which we classified as psoriasis, had an unusual presentation. This patient had a history of an atopic diathesis including hay fever, increased total IgE levels as well as elevated levels of the eosinophilic cationic protein. The patient presented clinically with chronic dermatitis on the palms of the hands, as well as on the plantar side of the feet. Furthermore, inflammatory skin lesions on the arms and other locations were clinically diagnosed as atopic dermatitis, consistent with the atopic diathesis, and a sample was obtained for the present study. However, later the patient developed inflammatory plaques on the lower back, which were clinically diagnosed as psoriasis. Both atopic dermatitis and psoriasis were considered as a diagnosis for this patient at various stages; however, the co-occurrence of atopic dermatitis and psoriasis is rare, perhaps due to the opposing immunopathogenic mechanisms for the two diseases (*55-57*). Although we cannot be certain of

the initial diagnosis, our molecular classifier correctly predicted the diagnosis as the clinical course evolved.

Functional annotation of related disease signatures using cell type deconvolution and k-means clustering shows shared and unique mechanisms of disease

Using cell-type specific gene signatures developed in previous work by Swindell, et al., the relative enrichment of each cell type signature was assessed in each disease, and subjected to hierarchical clustering using Euclidian distance (Figure 2) (*58, 59*). We found that leprosy, sarcoidosis, chancroid, Stevens Johnson syndrome and mycosis fungoides were characterized by high levels of lymphocytes (T, NK, and B cells), macrophages and dendritic cells. However, in our original tree (Figure 1) mycosis fungoides did not cluster with these diseases, suggesting that despite the similarity of their cell types, mycosis fungoides must be distinct based on other factors. The absence of enrichment of T cells in psoriasis and atopic dermatitis only indicates that these diseases have a relatively lower T cell signature compared to the other conditions.

We used pathway analysis of gene signatures to further investigate common pathways within each disease group. For each group of diseases in Figure 1, combined gene signatures were constructed and evaluated for enriched functional terms. The P values associated with each term were then subjected to k-means clustering (Figure 3). Significantly enriched terms support previous findings in the literature: allergic diseases are enriched for “cell-cell adherens junction” (p-value 8.93×10^{-03}), and hyperproliferative/neoplastic diseases are enriched for keratinocyte and epithelial cell development (p-value 2.47×10^{-09}) (*60, 61*). Wound, malignant, and infectious groups share overlapping enriched GO terms, associated with response to wounding (Figure 3).

Functional analysis of PM signatures shows enrichment for genes and pathways corresponding to single diseases

In order to assess the relevance of individual disease PM signatures, additional pathway analysis was performed on the 250 probe sets with the highest PM values for each disease. DAVID and Ingenuity Pathways functional analyses of each individual disease signature often showed a correspondence to the disease of origin (data not shown). For example, Ingenuity Pathways Analysis of the melanoma PM signature revealed a significant enrichment of "biologic functions" relevant to melanocyte development and disorders ("differentiation of melanocytes," "Waardenburg's syndrome," and "albinism"; p-values of $4.4x10^{-09}$, $5.6x10^{-09}$, and $4.7x10^{-08}$, respectively) and cancer ("cancer" and "proliferation of cells"; p-values of $3.0x10^{-06}$ and $5.6x10^{-08}$, respectively), as well as a significant enrichment of the "canonical pathway" "Melanocyte Development and Pigmentation Signaling" pathway (p-value $9.2x10^{-05}$) (Figure 4C) (*25*). As a resource for the community, we have developed a web-based visualization tool (http://pathways-pellegrini.mcdb.ucla.edu/goTeles/dot_plot.html) that plots the expression of a gene in every microarray sample within our database; shown are two melanocyte development genes (Figure 4A,B).

Network analysis of the PM signatures was carried out using Ingenuity Pathways Analysis to visualize connected genes and pathways in each disease and further evaluate the functional significance of our signatures. Notably, a psoriasis network showed connections between *TCN1*, *OASL*, and *SPRR3* (Figure S4). *TCN1* and *OASL* are among the genes most consistently and strongly elevated in psoriasis lesions, and while both genes are expressed by keratinocytes, they are also expressed at appreciable levels in neutrophils. *OASL*, moreover, is potently induced by interferon (IFN)–γ and encodes a protein involved in anti-viral responses (*62, 63*). The psoriasis network identified here appears to provide a cellular nexus that connects

key elements of psoriasis pathogenesis, including differentiation-associated pathways, IFN-directed responses, and infiltrating inflammatory cells such as neutrophils. (*59, 64-67*).

Moreover, cell type specific deconvolution analysis of the psoriasis PM signature revealed gene expression patterns consistent with the presence of neutrophils in psoriasis lesions, with 80 of the top 250 probe sets by PM (32%, $p<10^{-10}$) significantly enriched in neutrophils (Figure S5). This supports recent work suggesting that IL17A-producing neutrophils may be as abundant in psoriasis lesions as IL17A-producing T cells, and a recent GWAS study highlighting the genetic contribution of innate immunity to this disease (*68, 69*). Furthermore, neutrophils are known to be a histological marker of psoriasis, specifically located in epidermal microabscesses in disease lesions (*70, 71*). The gene interactions identified here for psoriasis demonstrate the ability to identify cell type-specific pathways within a particular disease.

Type I vs. Type II interferon gene programs have a negative inverse correlation across a spectrum of skin diseases

Type I and Type II interferons (IFN) have opposing immunoregulatory roles in human disease, and previous work has shown different diseases or subtypes of disease to exhibit a range of IFN responses (*7, 72*). IFN–γ(Type II IFN) is involved in macrophage activation to fight bacterial infection, and is opposed by IFN–α/β (Type I IFN), which combats viral infection. Because the IFNs are weakly detected on microarrays, we used Type I and Type II-specific induced transcriptional profiles of human peripheral blood mononuclear cells to infer the expression of IFN signatures (*72, 73*). Since Type I and Type II IFN programs overlap in their downstream targets, care was taken to utilize genes that were exclusively regulated by either IFN–β or IFN–γ. Integration of the IFN gene expression profiles with our data set containing 16 different skin conditions demonstrated a significant, inverse correlation between IFN–β and IFN–γ regulated genes across all skin diseases studied, with high IFN–β scores corresponding to low

IFN–γ scores and vice versa (r=-0.66, p-value=0.006), underscoring the opposing roles of IFN–β and IFN–γ in skin disease (Figure 5). The Stevens Johnson syndrome samples, which were obtained from blister fluid rather than full thickness biopsies, had the most positive IFN–β profile and most negative IFN–γ profile. Nevertheless, even if these samples were omitted, the anti-correlation between IFN–β vs. IFN–γ profiles was still significant (r=-0.53, p-value=0.04).

Discussion

Insights into disease pathogenesis obtained by comparison of gene expression profiles are often limited because these comparisons are performed between either two different diseases or one disease versus healthy controls, and therefore cannot identify distinct and common mechanisms of pathogenesis. Here, we performed a cross-disease analysis of molecular profiles from multiple skin diseases. Using fRMA, it was possible to assemble a database of gene expression profiles from 311 samples spanning 16 conditions and visualize disease relationships on a hierarchical clustering tree. Remarkably we found that samples of a particular disease that were taken from different batches colocalized to the same branch. This was particularly striking in the case of normal skin, where five batches of samples taken from healthy control subjects not only clustered on the same branch, but were arranged with little differentiation by batch.

Our approach demonstrates that a multi-disease classifier can be built from disparate public data sources comprising over a dozen different conditions in a single tissue. We built this classifier from disease specific gene signatures, and found that it was accurate and robust to batch effect, maintaining mean sensitivity above 0.88 and mean specificity close to 1.00 for three classification schemes (aggregate, three-fold cross validation and two-fold cross validation separated by batch). The potential utility of molecular classification over the classic clinical criteria was demonstrated by the correct classification of an ambiguous case of psoriasis. Gene expression profiles have previously been used in classifier studies, but these typically involve only one or two groups of conditions (*18, 74-76*). Comparison of gene signatures in lesions from multiple diseases has been more limited, even in skin disease in which biopsy specimens are more readily accessible. A multi-disease classifier using epithelial cells from patients with psoriasis, atopic dermatitis, allergic contact dermatitis, and irritant contact dermatitis was

constructed, although this would be limited to diseases which had epidermal involvement (*77*). Our work expands this principle to a wide range of both inflammatory and neoplastic diseases, and demonstrates the potential value of this approach in comparing diverse conditions. This approach can be expanded to include a more diverse spectrum of diseases, as more data is publicly available, allowing for the comparative study of diseases for which skin biopsy specimens may not be widely available.

We also analyzed this data by three supervised approaches: analysis of cell type signatures, Gene Ontology pathways (GO term enrichment), and interferon response signatures. Together, these bioinformatic analyses provided insight into the distinct and related pathogenesis of the diseases. For example, hierarchical clustering of gene expression profiles revealed that leprosy and chancroid were located on the same branch which we termed “infectious” based upon their known etiologies. In the deconvolution analysis of cell type signatures, both diseases were characterized by the enrichment of similar lymphoid (CD3+, CD4+, CD8+, regulatory T cells, B cells) and myeloid (monocytes, macrophages, dendritic cells, neutrophils) expression profiles. Furthermore, Gene Ontology enrichment analysis identified the terms “lysosome”, “T cell differentiation” and “leukocyte adhesion”.

In studying the interferon responses across different pathologic processes, we focused on gene expression for genes regulated exclusively by IFN–β (Type I) or IFN–γ (Type II IFN). Type II IFN is necessary to fight intracellular bacteria and linked to Th1 mediated inflammatory conditions both in the skin and systemically (*78, 79*). It has become increasingly clear that Type I vs. Type II IFN responses are cross-regulatory (*78, 80-82*). An earlier multi-disease comparison found that the magnitude of an IFN gene signature distinguished different inflammatory skin diseases, but could not distinguish between the Type I vs. Type II IFN

patterns (*7*). Our own data demonstrated an inverse correlation between IFN–β vs. IFN–γ gene programs in leprosy lesions, which in the present study has been expanded to reveal an anti-correlation of Type I and Type II IFN responses across a wide range of skin diseases of different etiologies (*72*).

The present findings provide a rationale for further investigations to determine how these different IFN programs contribute to the pathogenesis of these diseases and identify treatment targets. The spectrum of IFN–β vs. IFN–γ gene program expression in the skin diseases studied here is consistent with current practices in treatment of skin disease. Typically, Type I IFN exhibits anti-proliferative effects, and is used to treat neoplasms, such as melanoma and BCC, which have a negative Type I IFN score. It should be noted that IFN–γ has been used to treat acute atopic dermatitis, even though chronic atopic dermatitis lesions, as studied here, express IFN–γ consistent with our findings (*83-87*). Anti-IFN–γ has also been shown in preliminary studies to have a positive effect on Th1-mediated autoimmune skin diseases, including psoriasis, which we found had a high Type II IFN score (*79*). The integrative analysis of interferon signatures in a diverse spectrum of skin conditions supports the value of the concurrent analysis of multiple disease gene expression profiles to gain insight into the pathogenesis of skin disease.

Methods

Microarray acquisition, normalization, and filtering

Data was obtained from the NCBI Gene Expression Omnibus (GEO) as described in supplemental methods and Table S1 (http://www.ncbi.nlm.nih.gov/geo/) (*27*). All data was normalized using the Frozen Robust Multiarray Average (fRMA) method and low signal probe sets were removed at a cutoff of a mean intensity of at least 15 in any disease. To reduce platform biases, only Affymetrix HG U133 Plus 2.0 arrays were used.

Additional validation samples were obtained by collecting biopsy specimens from patients with atopic dermatitis, normal skin, psoriasis, and leprosy. The samples were coded to remove patient identifying information before transport to the laboratory. Gene expression profiles were derived via mRNA microarrays, using the Affymetrix HG U133 Plus 2.0 platform as previously described and will be deposited to NCBI GEO (*33*).

Clustering

Unsupervised hierarchical clustering was used to group the normalized, filtered expression profiles using Matlab, Archaeopteryx and Hypertree programs (*88, 89*). Distances were calculated as one minus the Pearson correlation between two samples. A full rooted tree with one sample per leaf was built using UPGMA on the distances, and an unrooted tree with one leaf per disease using the average correlation distance between all samples of two diseases. .

Proportional median

Probe sets were ranked using the Proportional Median (*PM)*, which we define as the median intensity of a probe set within one disease divided by the median intensity of the same probe set across all samples. For each disease, probe sets were ranked in descending order by PM.

Random forest classifier

A random forest classifier was built using the Matlab TreeBagger class (see supplemental methods for full description). The classifier feature space was reduced by selecting the 25 probe sets with the highest PM values across the training set for each disease, and taking the unique combined set of these lists.

Three-fold cross validation was carried out on all diseases but Stevens Johnson Syndrome, which was excluded since its sample size was too small to be partitioned into three groups. Training data was randomly separated into three partitions of approximately equal size, and three classifiers were iteratively built on two thirds of the data and tested on the remaining third. PM values were recalculated within the training set for each classifier iteration, for the selection of the top 25 probe sets from each disease. Performance statistics were calculated by aggregating true positive, true negative, false positive, and false negative counts from all classifiers' predictions

Two-fold cross validation with batch separation was carried out by building two independent classifiers built using approximately half the data. Where applicable, data was partitioned by batch such that each classifier was trained on only one batch of data, otherwise, data was randomly divided equally. As in cross-validation, PM values were recalculated for the training data for each data partition. Testing was performed on the batch of data not used to build the classifier, and performance statistics were calculated in aggregate as in cross-validation.

Validation was carried out on independent microarray data sets of atopic dermatitis, normal skin, psoriasis, melanoma, and leprosy as described above. These microarray samples were

not used in classifier feature selection (i.e. PM calculation) or training. After discovery of a misdiagnosed patient, samples were reclassified and one sample was subtracted from atopic dermatitis and one was added to psoriasis. Performance was assessed as described for the batch classifier, by aggregating and counting using these modified values.

Pathway analysis

The 250 probe sets with the highest PM value for each disease were selected for pathway analysis using Ingenuity Pathways Analysis (http://www.ingenuity.com) and DAVID Functional Annotation Analysis (http://david.abcc.ncifcrf.gov/), using the top 250 probe sets by PM from each disease signature (*23-25*).

Cell-type specific signature enrichment

Cell-type specific expression profiles for 24 cell types were calculated as previously described and is described in more detail in the supplemental methods (*59*).

Group signatures

Probe set signatures from multiple diseases were integrated into five group signatures, based on our tree (Figure 1). The groups were hyperproliferative/neoplastic, wound, allergic, malignant, and infectious. Only individual samples that actually colocalized on the tree were used for group signature calculations; for example, the BCC and SCC samples adjacent to the psoriasis samples were used to calculate the hyperproliferative/neoplastic signature. For each group, PM was calculated for samples in each disease as described above. Probe sets were assigned a rank based on PM (with rank=1 assigned to the probe set with the highest PM value), and the mean rank and value for each probe set was calculated across all diseases in a

group. Group signatures were comprised of probe sets that had both a mean PM rank in the top 5% of all filtered probes, as well as an average PM value greater than or equal to 2.0.

Gene Ontology (GO) enrichment was performed via DAVID Functional Annotation Analysis of the group signatures. Enriched annotation terms were filtered for GO terms that passed a false discovery rate (FDR) of 5% to create "GO signatures". GO terms were clustered using k means clustering (k=8) of FDR values, which were displayed using a heat map.

IFN profile integration

Lists of probe sets either up or down regulated by IFN–β and IFN–γ were obtained as described previously in Teles, et al. (*72*). These lists were converted to gene names and filtered to eliminate any genes that were regulated by both IFN–β and IFN–γ. For genes corresponding to multiple probe sets, the probe set with the highest mean expression across all diseases was selected. Each list was filtered for genes showing the most differential expression across all skin samples by calculating the variance of gene intensity and selecting the 30 highest variance genes from each list. We chose to use 30 genes because this was the length of the smallest of the four lists.

IFN–β and IFN–γ scores were computed for each disease by first ranking all filtered microarray genes by intensity and computing the rank of each IFN-regulated gene in relation to all others in that same disease. The mean rank across all diseases was computed for each IFN-regulated gene, and a delta-rank was assigned to each IFN-regulated gene, in each disease, by taking the signed difference between the rank in that disease and the mean rank. The overall IFN–β and IFN–γ score was computed for each disease by subtracting the sum of down-regulated IFN

delta means from the sum of up-regulated IFN delta means, and a straight line was fitted to the plotted scores.

Acknowledgments:

We would like to thank Barry Bloom for his extensive and thoughtful input on the manuscript, and Steve Horvath for his helpful discussion regarding classifiers and clustering analyses.

Funding: JTE was supported by NIH grant R01 AR054966 to and by the Ann Arbor Veterans Affairs Hospital. MSI was supported in part by NIH/NIAMS grant P50 5P50AR063020. RLM is a recipient of the CHANEL-CERIES Research Award. **Author contributions**: MSI, POS, WRS, JTE, RLM, and MP wrote and edited the manuscript. DL built the website. MSI, WRS, JTE, RLM, RMBT, TGG, and MP designed the components of the analysis. MG, BH, and SM provided patient samples and microarray data. WRS performed deconvolution analysis. MSI performed all other analyses. **Competing interests**: No conflicts of interest to report.

Figure Legends

Figure 1. **Unsupervised hierarchical clustering tree of 311 skin samples.** Normalized, filtered microarray data was clustered using Pearson correlation distance and displayed in a tree using average distance. Each terminal leaf in the tree represents a biopsy sample and is colored according to disease, with colored bars to the right representing the majority disease diagnosis. Samples that clustered apart from other samples of the same diagnosis can be seen a leaf that differs in color from its neighbors. Numbers following disease name labels denote batches of the same disease, and lists of numbers following a disease name denote multiple batches clustering to the same tree branch with little or no differentiation by batch. Brackets to the far right delineate biological groups of neighboring diseases.

Figure 2. Cell-type specific signature enrichment. For each of 24 cell-type specific signatures, log fold changes were calculated per disease, with each fold change representing the enrichment for a particular cell type signature in that disease. Fold change vectors were clustered using Euclidean distance and displayed in a heatmap, where rows correspond to diseases and columns correspond to cell type. Note that enrichment scores are relative across each cell type. Black triangles denote FDR < 0.05 and directionality of fold change.

Figure 3. Functional annotation and k-means clustering of group signatures. Group signatures were annotated with enriched Gene Ontology (GO) terms, and the false discovery rate (FDR) for each GO term was clustered using k-means clustering (k=8) and visualized in a heatmap, where rows correspond to GO terms and columns correspond to disease groups. Each gray bar represents the log10 FDR for a particular GO term in a particular disease group. Colored bars to the right demarcate clusters of GO terms, and a summary of terms and p-values are provided for each color bar.

Figure 4. Visualization of melanoma proportional median (PM) signature. A,B. Visualization of raw intensities for MITF and TYR, two genes in the melanocyte development pathway. Each black circle represents an intensity value on a microarray for the specific probe set named. Red lines show median intensity values for each disease. Disease abbreviations are as follows: lepromatous leprosy (LLP), tuberculoid leprosy (TLP), reversal reaction leprosy (RR), erythema nodosum leprosum (ENL), chancroid (CH), mycosis fungoides (MF), sarcoid (SAR), Stevens Johnson Syndrome (SJS), psoriasis (PS), allergic contact dermatitis (ACD), irritant contact dermatitis (ICD), atopic dermatitis (ATD), burn (BU), acute wound (WA), post-operative wound (WPO), melanoma (MEL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), normal skin (NS). **C.** Visualization of melanocyte development pathway in Ingenuity Pathways analysis. Red genes indicate presence in a 250 probe set PM melanoma signature, with darker red denoting higher PM values. This signature was significantly enriched for the melanocyte pathway (p-value = 9.19×10^{-05}).

Figure 5. Type I and II interferon program cross-regulation. IFN–β and IFN–γ scores were calculated as the mean difference in delta ranks of genes specifically regulated by IFN–β or IFN–γ for each disease, relative to mean ranks across all diseases. Intuitively, high scores for each type of IFN represent high expression of IFN-stimulated genes, low expression of IFN-repressed genes, or both, such that placement on each axis shows the magnitude of expression of IFN–β or IFN–γ gene programs. The plot shows a significant negative inverse correlation (r=-0.66, p-value=0.006). Removing the outlier Stevens Johnson syndrome, the correlation remains significant (r=-0.53, p-value=0.04).

Figures

Figure 1

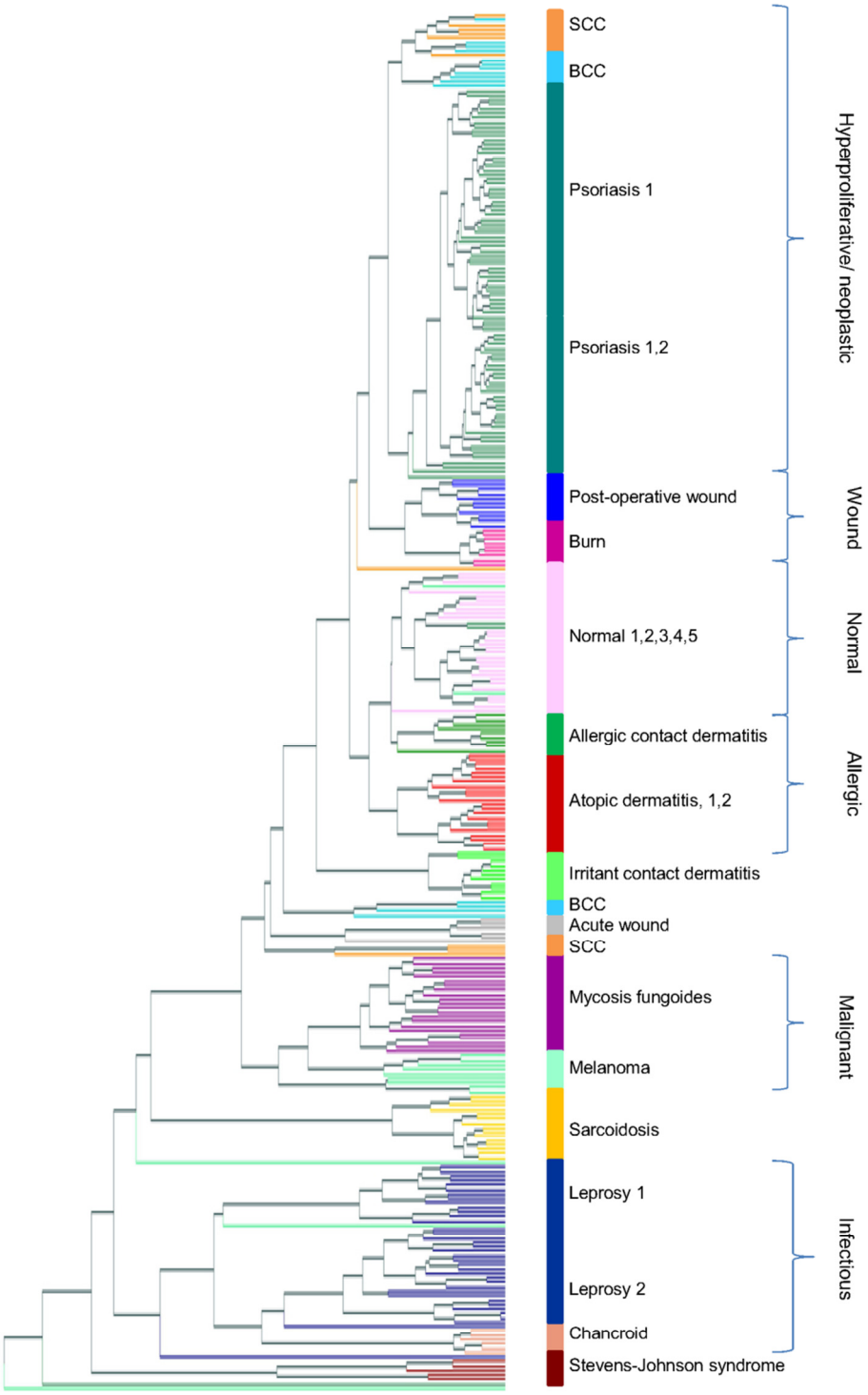


Figure 2

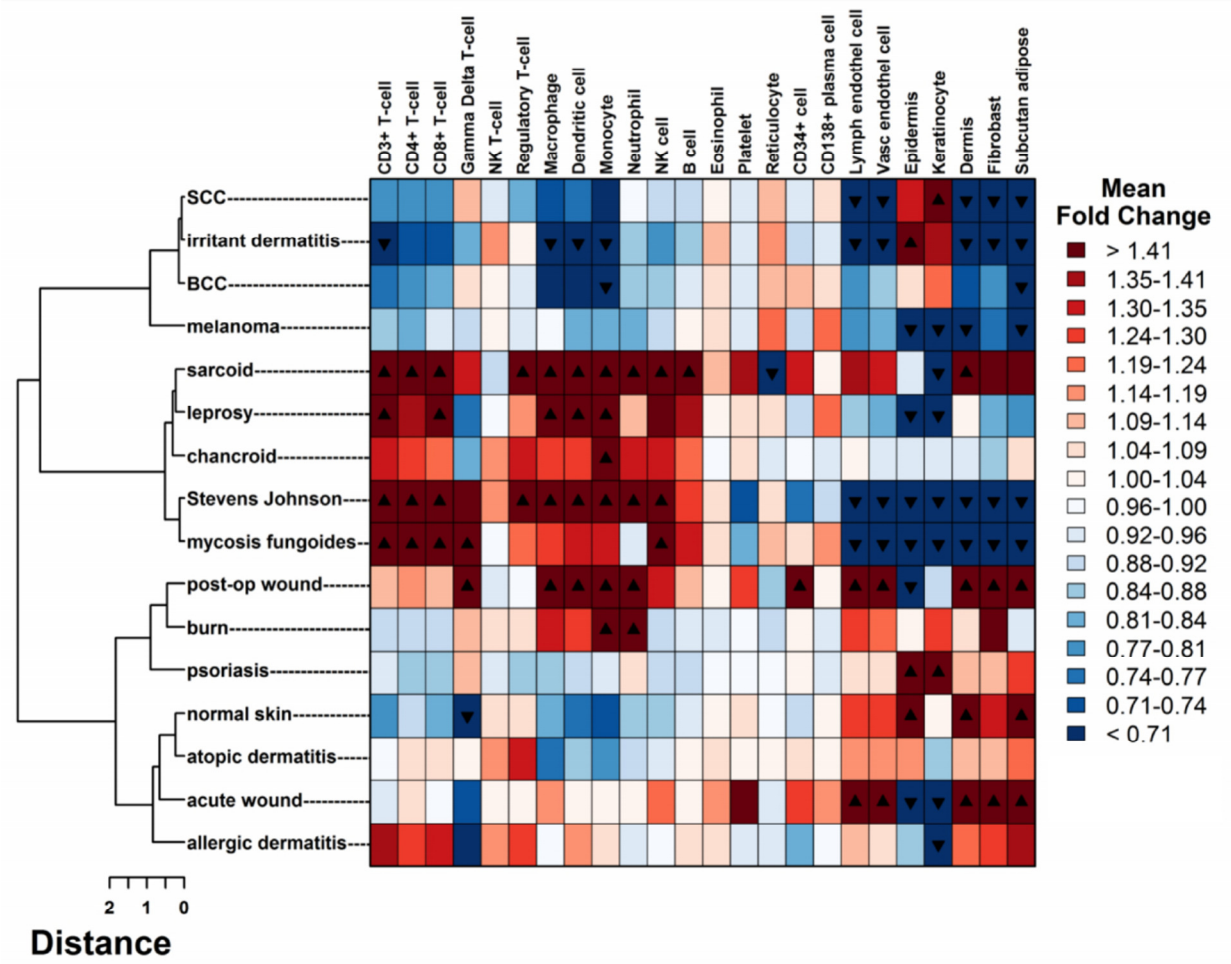

CD3+ T-cell
CD4+ T-cell
CD8+ T-cell
Gamma Delta T-cell
NK T-cell
Regulatory T-cell
Macrophage
Dendritic cell
Monocyte
Neutrophil
NK cell
B cell
Eosinophil
Platelet
Reticulocyte
CD34+ cell
CD138+ plasma cell
Lymph endothel cell
Vasc endothel cell
Epidermis
Keratinocyte
Dermis
Fibrobast
Subcutan adipose
SCC
irritant dermatitis
BCC
melanoma
sarcoid
leprosy
chancroid
Stevens Johnson
mycosis fungoides
post-op wound
burn
psoriasis
normal skin
atopic dermatitis
acute wound
allergic dermatitis
2 1 0
Distance
Mean Fold Change
> 1.41
1.35-1.41
1.30-1.35
1.24-1.30
1.19-1.24
1.14-1.19
1.09-1.14
1.04-1.09
1.00-1.04
0.96-1.00
0.92-0.96
0.88-0.92
0.84-0.88
0.81-0.84
0.77-0.81
0.74-0.77
0.71-0.74
< 0.71

Figure 3

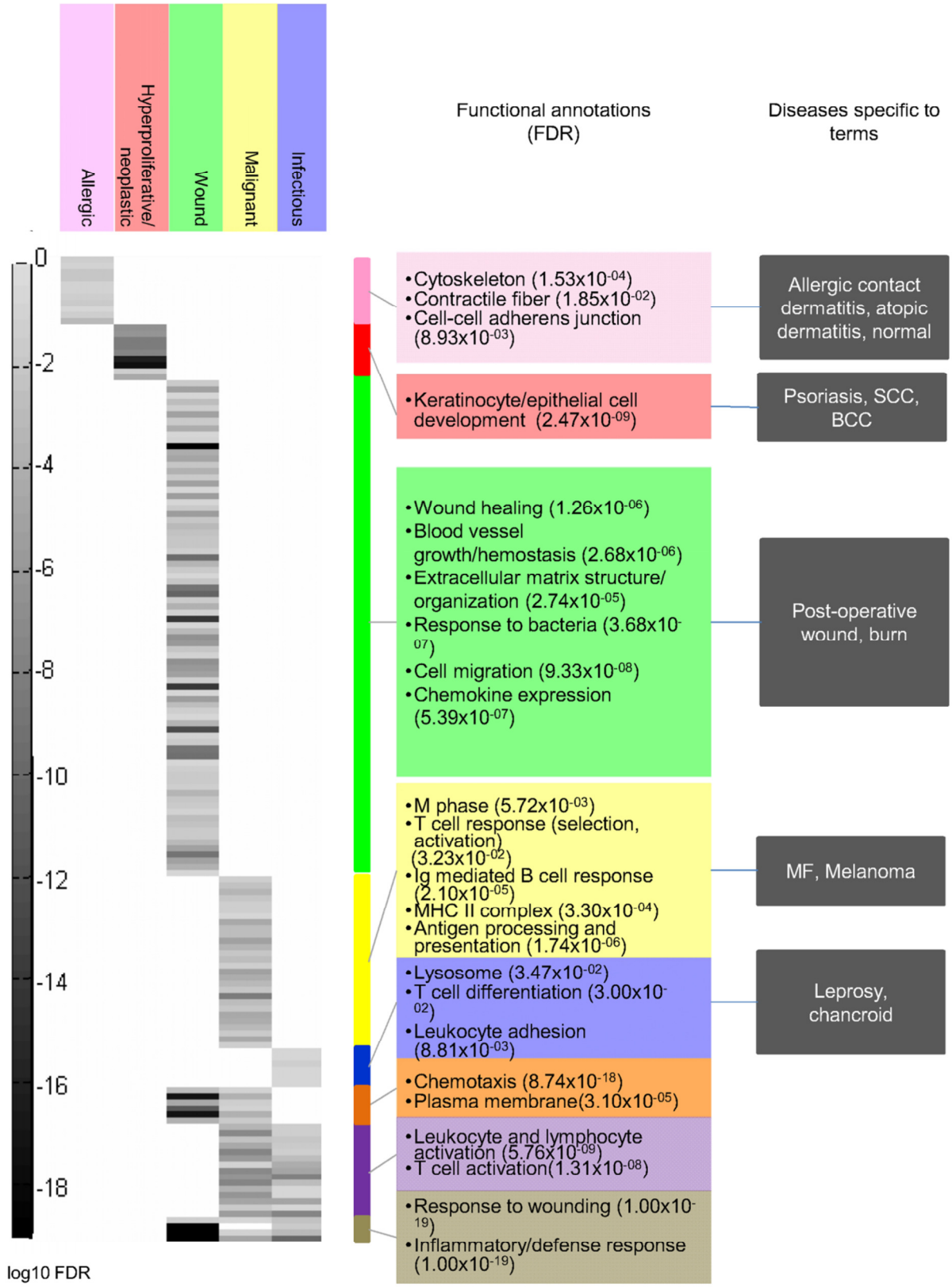

Allergic
Hyperproliferative/ neoplastic
Wound
Malignant
Infectious
Functional annotations (FDR)
Diseases specific to terms
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
log10 FDR
•Cytoskeleton (1.53x10-04)
•Contractile fiber (1.85x10-02)
•Cell-cell adherens junction (8.93x10-03)
Allergic contact dermatitis, atopic dermatitis, normal
•Keratinocyte/epithelial cell development (2.47x10-09)
Psoriasis, SCC, BCC
•Wound healing (1.26x10-06)
•Blood vessel growth/hemostasis (2.68x10-06)
•Extracellular matrix structure/ organization (2.74x10-05)
•Response to bacteria (3.68x10-07)
•Cell migration (9.33x10-08)
•Chemokine expression (5.39x10-07)
Post-operative wound, burn
•M phase (5.72x10-03)
•T cell response (selection, activation) (3.23x10-02)
•Ig mediated B cell response (2.10x10-05)
•MHC II complex (3.30x10-04)
•Antigen processing and presentation (1.74x10-06)
MF, Melanoma
•Lysosome (3.47x10-02)
•T cell differentiation (3.00x10-02)
•Leukocyte adhesion (8.81x10-03)
Leprosy, chancroid
•Chemotaxis (8.74x10-18)
•Plasma membrane(3.10x10-05)
•Leukocyte and lymphocyte activation (5.76x10-09)
•T cell activation(1.31x10-08)
•Response to wounding (1.00x10-19)
•Inflammatory/defense response (1.00x10-19)

Figure 4

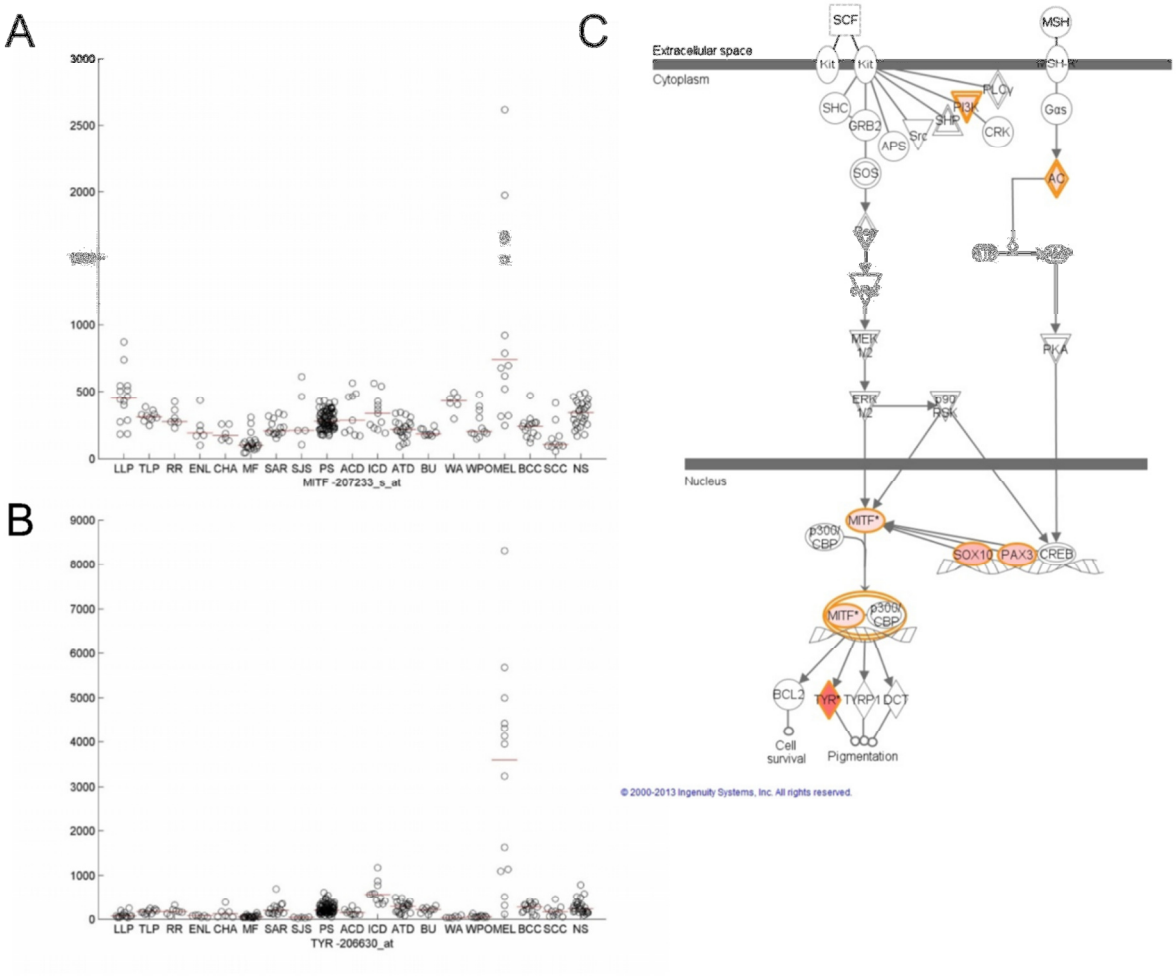
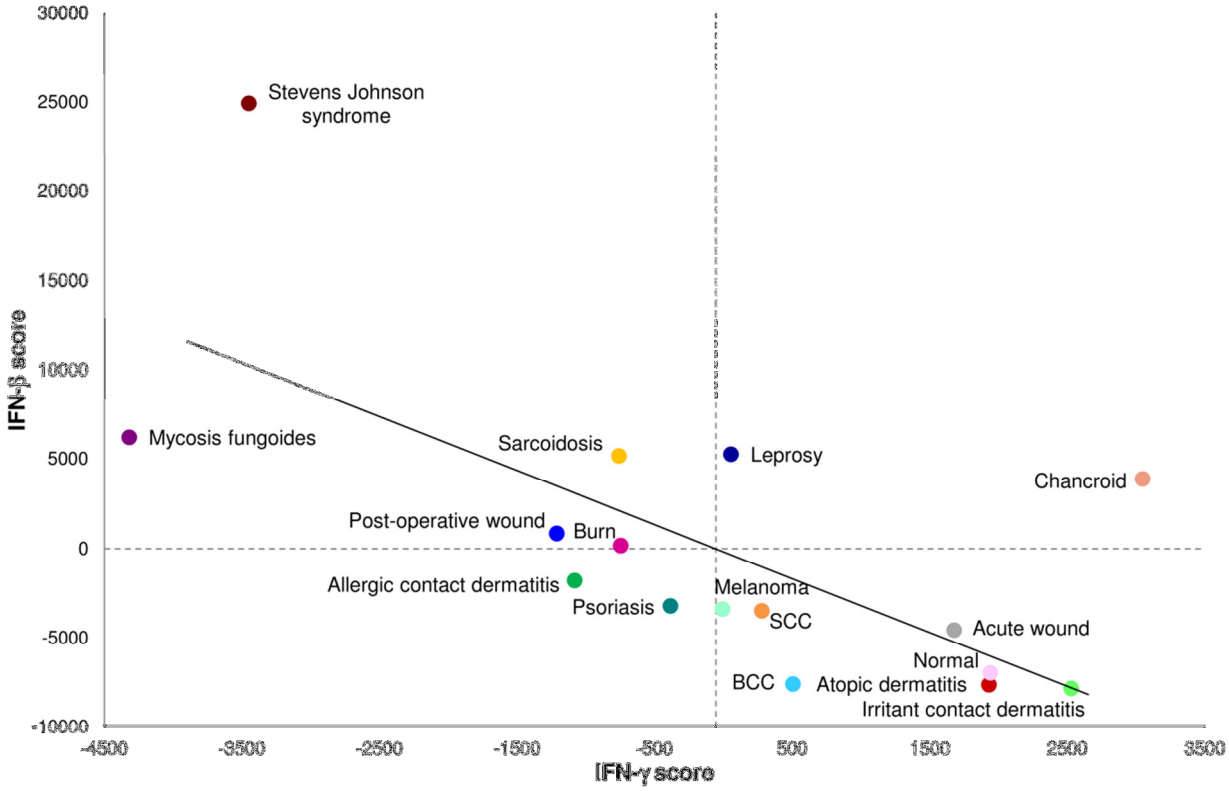

A
3000
2500
2000
1500
1000
500
0
LLP TLP RR ENL CHA MF SAR SJS PS ACD ICD ATD BU WA WPOMEL BCC SCC NS
MITF -207233_s_at
B
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
LLP TLP RR ENL CHA MF SAR SJS PS ACD ICD ATD BU WA WPOMEL BCC SCC NS
TYR -206630_at
C
Extracellular space
Cytoplasm
SCF
MSH
Kit
Kit
SHC
GRB2
APS
Src
PI3K
SHP
PLCγ
CRK
Gαs
SOS
AC
MEK 1/2
PKA
ERK 1/2
p90 RSK
Nucleus
MITF*
p300/ CBP
SOX10
PAX3
CREB
MITF*
p300/ CBP
BCL2
TYR*
TYRP1
DCT
Cell survival
Pigmentation
© 2000-2013 Ingenuity Systems, Inc. All rights reserved.

Figure 5


30000
25000
20000
15000
10000
5000
0
-5000
-10000
IFN-β score
-4500
-3500
-2500
-1500
-500
500
1500
2500
3500
IFN-γ score
Stevens Johnson syndrome
Mycosis fungoides
Sarcoidosis
Leprosy
Chancroid
Post-operative wound
Burn
Allergic contact dermatitis
Psoriasis
Melanoma
SCC
Acute wound
Normal
BCC
Atopic dermatitis
Irritant contact dermatitis

Tables

Table 1. Two-fold cross validation, with separation by batch where applicable.

Table 1

Condition	Sensitivity	Specificity	Precision	F1
Leprosy*	0.86	1.00	0.97	0.91
Psoriasis*	0.96	0.97	0.94	0.95
Chancroid	1.00	1.00	1.00	1.00
Allergic contact dermatitis	0.78	1.00	1.00	0.88
Irritant contact dermatitis	1.00	1.00	1.00	1.00
Atopic dermatitis*	0.95	0.99	0.88	0.91
Burn	1.00	1.00	1.00	1.00
Acute wound	0.83	1.00	1.00	0.91
Post-operative wound	1.00	1.00	1.00	1.00
Mycosis fungoides	1.00	1.00	1.00	1.00
BCC	0.87	1.00	0.93	0.90
Melanoma	0.86	0.98	0.67	0.75
SCC	0.60	1.00	0.86	0.71
Sarcoidosis	1.00	1.00	1.00	1.00
Stevens Johnson syndrome	0.80	1.00	1.00	0.89
Normal skin*	0.96	0.98	0.84	0.90

*Denotes diseases separated by batch

Table 2. Classifier performance on external validation data.

Table 2

Condition	Sensitivity	Specificity	Precision	F1
Leprosy 3	1.00	1.00	1.00	1.00
Psoriasis 3, 4, 5, 6, 7	0.97	0.97	0.99	0.98
Atopic Dermatitis 3, 4	0.80	0.98	0.73	0.76
Melanoma 2	0.90	1.00	1.00	0.95
Normal 6, 7	0.89	0.98	0.73	0.80

Supplementary Figures

Figure S1. Normalized distribution of fRMA normalized microarrays. Boxplots of the 311 skin microarrays used in this analysis. Each boxplot represents the intensity values for a single microarray. The line bisecting each boxplot represents the median microarray intensity value, and the bottom and top of each box denotes the 25th and 75th quartile, respectively.

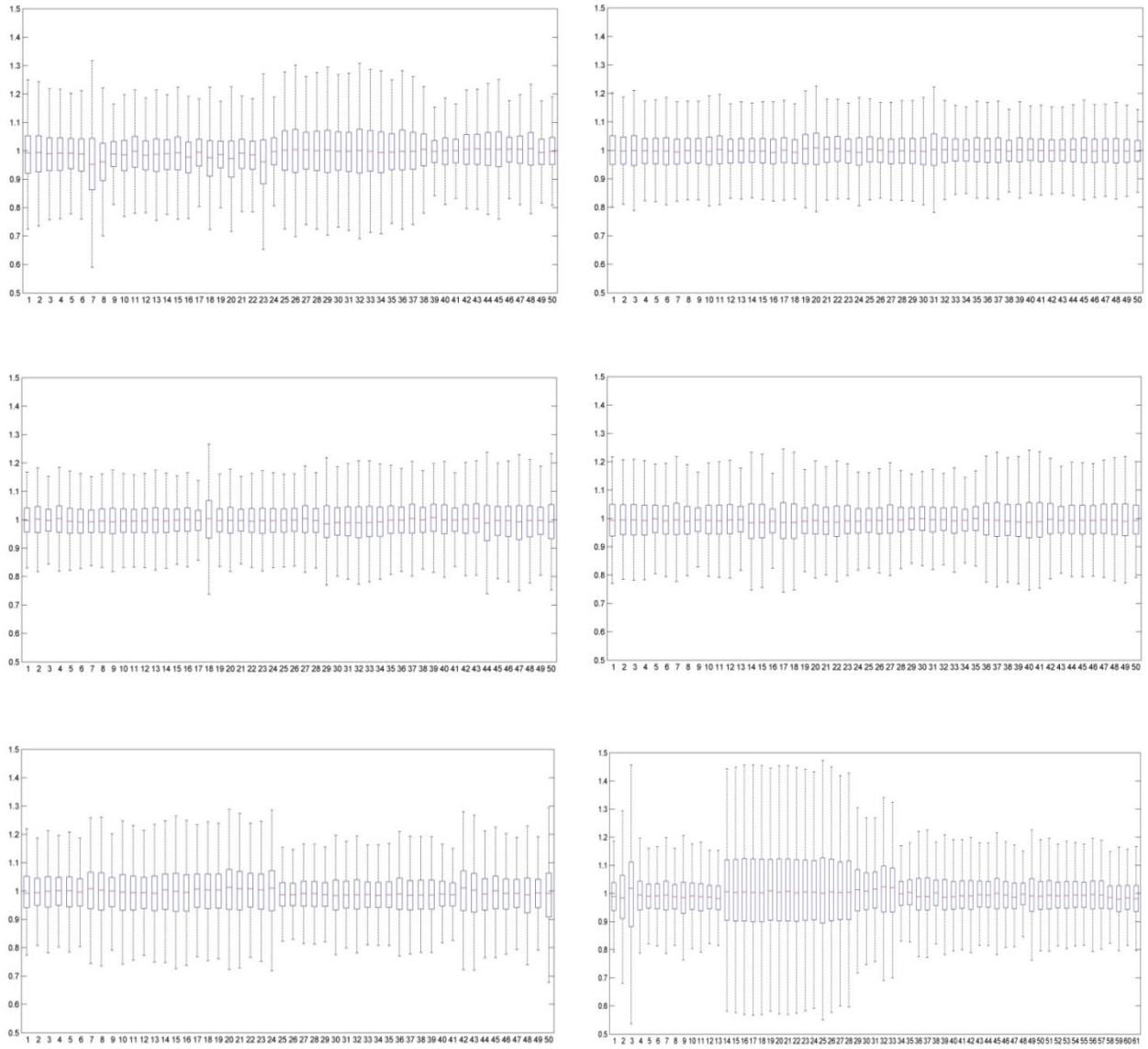
Figure S2. Unrooted unsupervised clustering tree of 16 skin conditions. Normalized, filtered microarray data was clustered using correlation as in Figure 1, and the average distance between all pairs of diseases was calculated. The tree was generated using the UPGMA method. Each leaf represents one disease, and distances between leaves are 1-correlation of disease expression profiles between all samples of two diseases.

Figure S3. Error rate for number of trees. Error rate was measured as a percentage for each number of trees by assessing out of bag predictions for accuracy. The plot shows the error rate at each value, and shows a stabilization of error rate at 5% once the number of trees used in this study is reached.

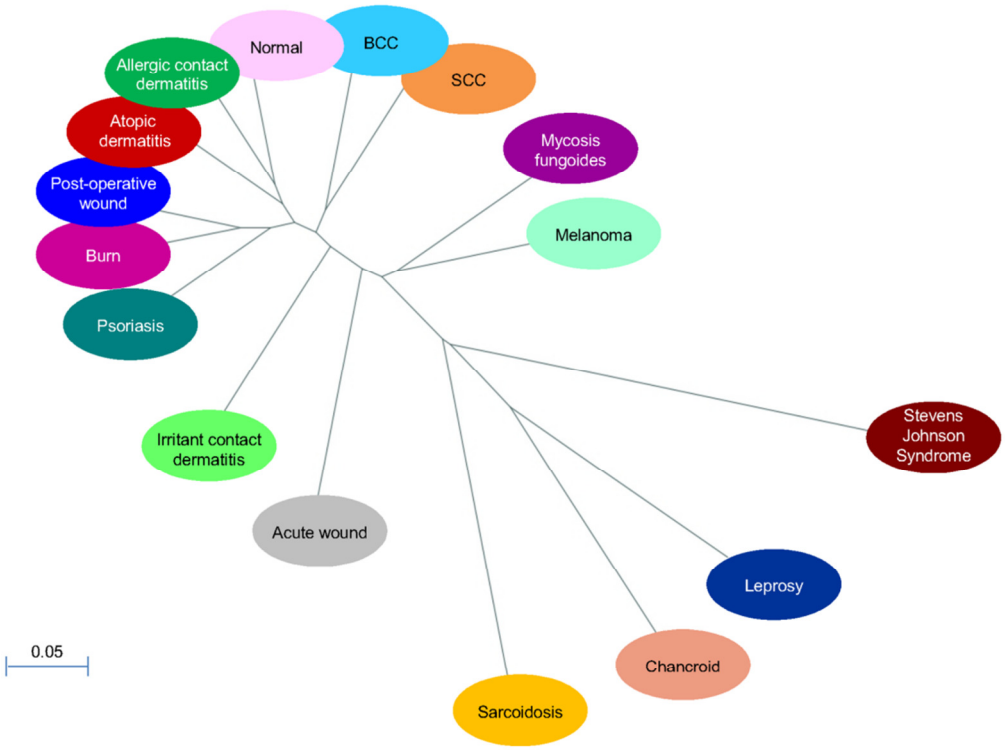
Figure S4. Psoriasis signature network. Pathways were constructed using the top 250 probe sets by PM in Ingenuity Pathways Analysis. Genes in the top 250 PM list for psoriasis are colored shades of red, with darker colors denoting higher PM values. Solid lines are direct connections and dotted lines are indirect connections.

Figure S5. Cell-type specific deconvolution of psoriasis signature. Cell type specific signatures were calculated for 24 different types of immune cells, and the relative expression of each probe set in the psoriasis PM signature was calculated for each cell type relative to all others. Each row in the heatmap represents a probe set in the top 250 probe sets by PM for psoriasis. Each column represents a cell type. Each box in the heatmap indicates fold change of that probe set in cell types relative to a reference set of 23 other cell types. Black arrows indicate FDR < 0.05, with orientation indicating directionality.

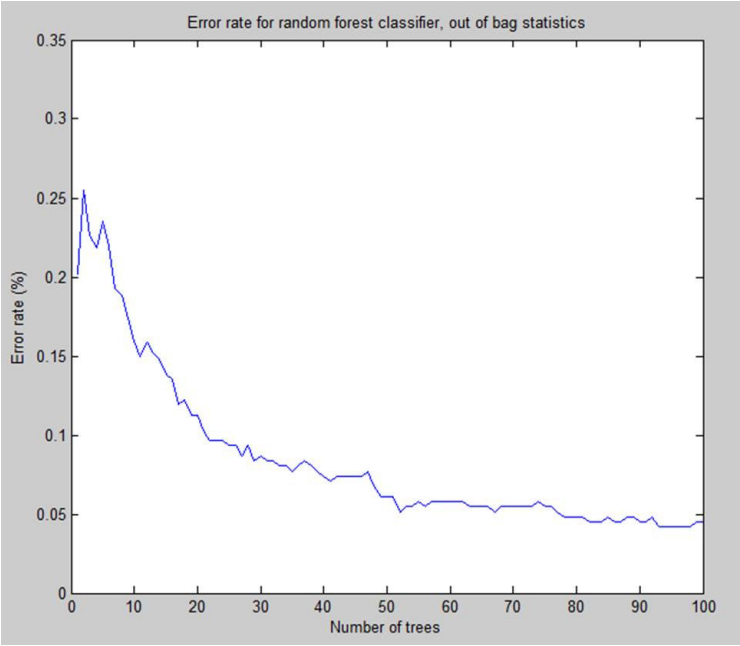
Supplementary Figure 1



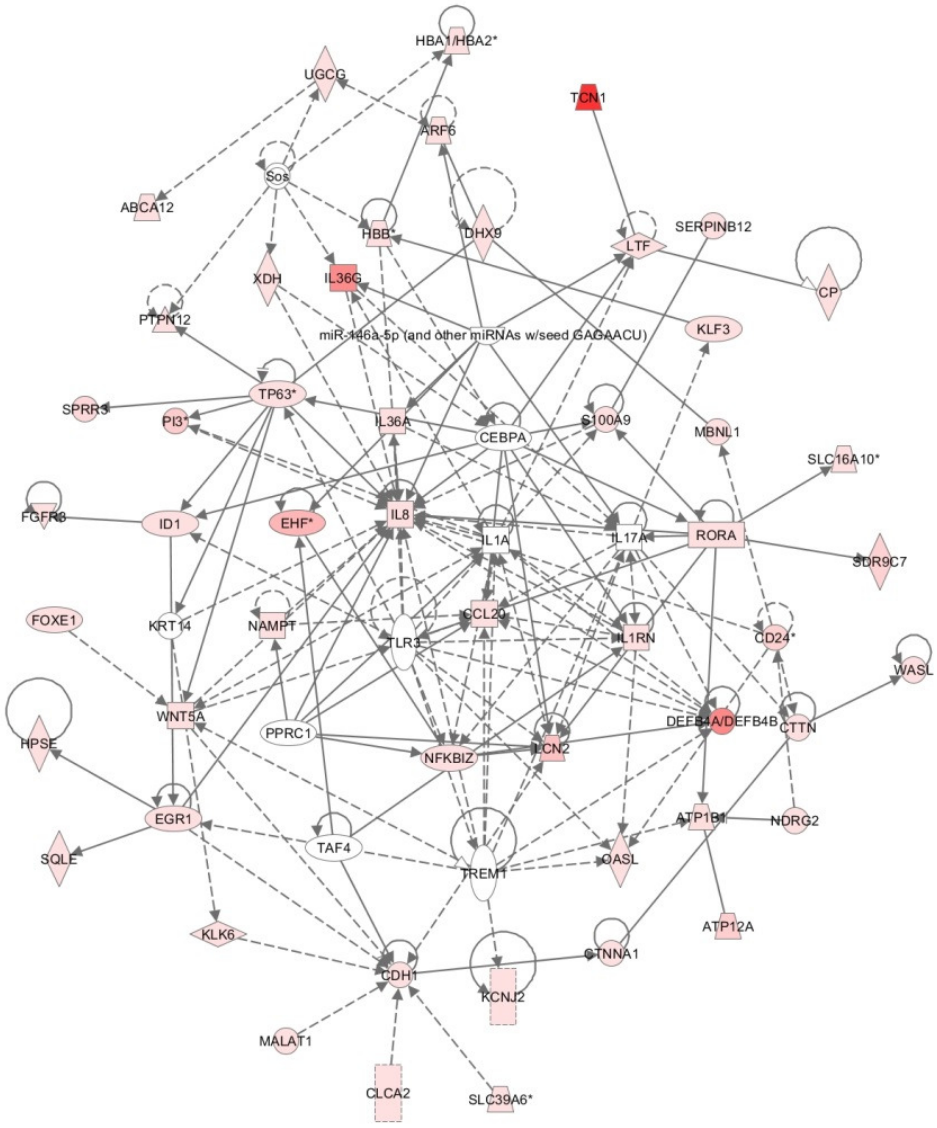
Supplementary Figure 2



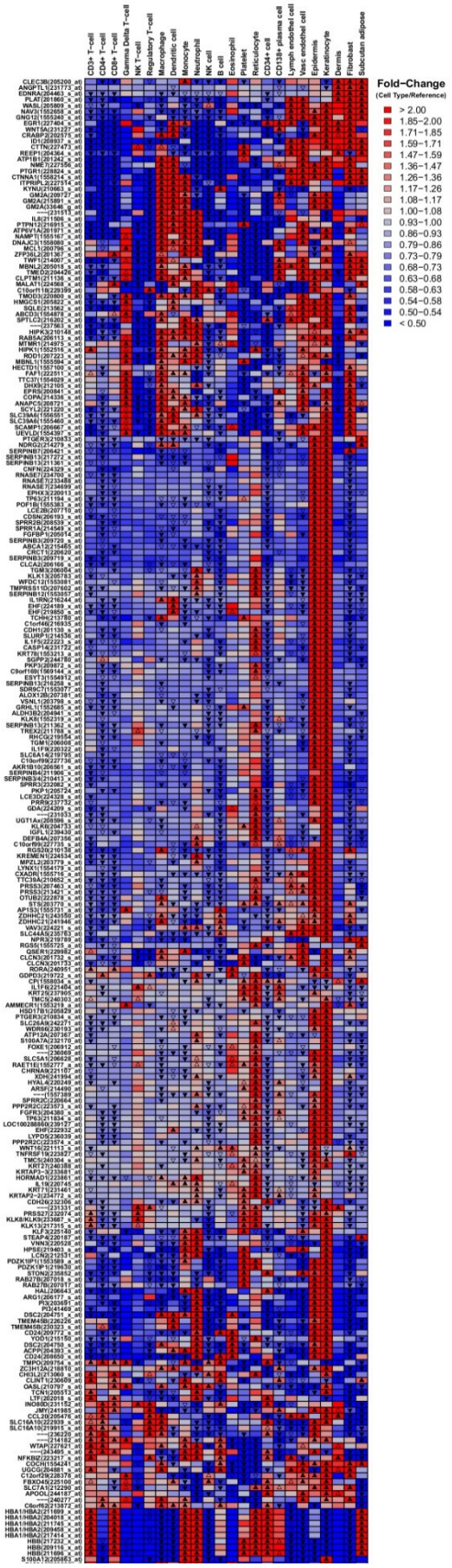
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5



Supplementary Tables

Table S1. Description of the skin samples used in this study to build the classifier and perform functional analysis

Supplementary Table 1

Disease	Number lesional samples	GEO accession #	Reference(s)
Psoriasis 1[1]	58	GSE13355	Nair et al., Nat Genet 2009 and Swindell et al., PLoS One 2011
Psoriasis 2[2]	33	GSE14905	Yao et al., PLoS One 2008
Leprosy 1*	24	GSE17763	Montoya et al., Cell Host Microbe 2009
Leprosy 2*	13	GSE16844	Lee et al., J Infect Dis 2010
Chancroid	6	GSE5547	Humphreys et al., Infect Immun 2007
Allergic contact dermatitis	9	GSE6281	Pederson et al., J Invest Dermatol 2007
Irritant contact dermatitis	11	GSE18206	Clemmensen et al., J Invest Dermatol 2010
Atopic dermatitis 1[3]§	9	GSE16161	Guttman-Yassky et al., J Allergy Clin Immunol 2009
Atopic dermatitis 2§	13	GSE32924	Suarez-Farinas et al., J Allergy Clin Immunol 2011
Burn[4]	9	GSE8506	N/A
Acute wound[5]	6	GSE28914	Nuutila et al., Wound Repair Regen 2012
Post-operative wound[5]	11	GSE28914	Nuutila et al., Wound Repair Regen 2012
Mycosis fungoides	22	GSE12902	van Doorn et al., Blood 2009
Basal cell carcinoma[6]	15	GSE7553	Riker et al., BMC Med Genomics 2008
Melanoma[6]	14	GSE7553	Riker et al., BMC Med Genomics 2008
Squamous cell carcinoma[6]	10	GSE7553	Riker et al., BMC Med Genomics 2008
Sarcoidosis	15	GSE32887	Judson et al., J Am Acad Dermatol 2012
Stevens-Johnson syndrome	5	GSE13726	Chung et al., Nat Med 2008
Normal skin 1[1]	6	GSE13355	Nair et al., Nat Genet 2009 and Swindell et al., PLoS One 2011
Normal skin 2[2]	6	GSE14905	Yao et al., PLoS One 2008
Normal skin 3[4]	3	GSE8506	N/A
Normal skin 4[3]§	9	GSE16161	Guttman-Yassky et al., J Allergy Clin Immunol 2009
Normal skin 5	4	GSE7553	Riker et al., BMC Med Genomics 2008

[1] [2] [3] [4] [5] [6] [7]: each from the same experiment and lab
*, §, †, ‡: each from the same lab

Table S2. Classifier internal validation performance. Sensitivity, specificity, and precision were calculated using "out-of-bag" predictions obtained during classifier training.

Supplementary Table 2

Condition	Sensitivity	Specificity	Precision	F1
Leprosy	0.97	1.00	0.97	0.97
Psoriasis	0.97	0.99	0.97	0.97
Chancroid	1.00	1.00	1.00	1.00
Allergic contact dermatitis	1.00	1.00	1.00	1.00
Irritant contact dermatitis	1.00	1.00	1.00	1.00
Atopic dermatitis	0.95	1.00	0.95	0.95
Burn	1.00	1.00	1.00	1.00
Acute wound	1.00	1.00	1.00	1.00
Post operative wound	1.00	1.00	1.00	1.00
Mycosis fungoides	0.96	1.00	0.96	0.96
BCC	0.93	1.00	0.93	0.93
Melanoma	0.86	0.99	0.86	0.86
SCC	0.88	1.00	0.88	0.88
Sarcoidosis	1.00	1.00	1.00	1.00
Stevens Johnson syndrome	1.00	1.00	1.00	1.00
Normal skin	0.87	0.99	0.87	0.87

Table S3. Classifier three-fold cross-validation performance. Three-fold, leave-one-out cross validation was performed on the data using the original dataset randomly partitioned into three sets of approximately equal size.

Supplementary Table 3

Condition	Sensitivity	Specificity	Precision	F1
Leprosy	1.00	1.00	1.00	1.00
Psoriasis	0.98	0.98	0.95	0.96
Chancroid	1.00	1.00	1.00	1.00
Allergic contact dermatitis	0.89	1.00	1.00	0.94
Irritant contact dermatitis	1.00	1.00	1.00	1.00
Atopic dermatitis	1.00	1.00	0.96	0.98
Burn	1.00	1.00	1.00	1.00
Acute wound	1.00	1.00	1.00	1.00
Post operative wound	1.00	1.00	1.00	1.00
Mycosis fungoides	1.00	1.00	1.00	1.00
BCC	0.93	1.00	0.93	0.93
Melanoma	0.86	1.00	1.00	0.92
SCC	0.60	1.00	1.00	0.75
Sarcoidosis	1.00	1.00	1.00	1.00
Stevens Johnson syndrome	1.00	0.99	0.90	0.95
Normal skin	1.00	1.00	1.00	1.00

Table S4. Description of the skin samples used in this study to perform classifier validation. Note that numbers and symbols from S1 are the same as in S4.

Supplementary Table 4

Disease	Number lesional samples	GEO accession #	Reference(s)
Leprosy 3[7]*	9	N/A	N/A
Psoriasis 3†	5	N/A	N/A
Psoriasis 4§	85	GSE30999	Suarez-Farinas et al., J Invest Dermatol 2012
Psoriasis 5‡	14	GSE34248	Bigler et al., PLoS One 2013
Psoriasis 6‡	24	GSE41662	Bigler et al., PLoS One 2013
Psoriasis 7‡	15	GSE41663	Bigler et al., PLoS One 2013
Atopic Dermatitis 3†	4	N/A	N/A
Atopic Dermatitis 4§	16	GSE36842	Gitler et al., J Allergy Clin Immunol 2012
Melanoma 2	10	GSE31879	N/A
Normal 6†	4	N/A	N/A
Normal 7[7]*	5	N/A	N/A

[1] [2] [3] [4] [5] [6] [7]: each from the same experiment and lab

*, §, †, ‡: each from the same lab

Supplementary methods

Data acquisition

A text search for "skin" was performed, and experimental series that contained microarrays from skin disease states were downloaded (*64, 65, 90-102*). All microarray data was derived from lesional skin biopsies, with the exception of SJS, which comprised lesional blister fluid. The series were additionally filtered for those utilizing the Affymetrix HG U133 Plus 2.0 platform in which raw CEL files were provided. We identified and downloaded 311 relevant skin samples spanning 15 diseases as well as normal skin. In some instances, multiple batches of the same disease were obtained, (designated as n_x where x denotes the batch). The samples included psoriasis, n = 91 (n_1 = 58, n_2 = 33); leprosy, n = 37 (n_1 = 24, n_2 = 13); *H ducreyi* infection (chancroid), n = 6; allergic contact dermatitis, n = 9; irritant contact dermatitis, n = 11; atopic dermatitis, n = 22 (n_1 = 9, n_2 = 13); burn, n = 9; acute wound, n = 6; post-operative wound, n = 11; mycosis fungoides, n = 22; basal cell carcinoma, n = 15; squamous cell carcinoma, n = 10; melanoma, n = 14; cutaneous sarcoidosis, n = 15; Stevens-Johnson syndrome, n = 5; normal skin, n = 28 (n_1 = 6, n_2 = 6, n_3 = 3, n_4 = 9, n_{5-4}) (Table S1). Normal skin was defined as skin taken from a healthy patient, and all normal skin samples came from a series that also contained lesional samples. 102 normal skin samples were available for download (n_1 = 64, n_2 = 22, n_3 = 3, n_4 = 9, n_{5-4}); to avoid bias due to disparate sample sizes, only six samples from n_1 and n_2 were used in this analysis. Validation samples were obtained by collecting biopsy specimens from patients with atopic dermatitis (n_3=3), normal skin (n_6=4, n_7=5), psoriasis (n_3=5), and leprosy (n_3=9); in addition, publicly available sets of psoriasis (n_4=85, n_5=14, n_6=24, n_7=15), atopic dermatitis (n_4=8), and melanoma (n_2=10) were downloaded (Table S5) (*65, 66, 103*)

Normalizing and filtering

Data was normalized using the Frozen Robust Multiarray Average (fRMA) normalization method (*31*). While frozen RMA is not a method for removing batch effects, it allowed the normalization of all samples to similar distributions and in this case, rendered the intra disease batch effects generally smaller than the inter disease differences. Traditional batch effect removal software was not appropriate for this analysis since such methods typically require all experimental conditions to be represented in each batch (*30*). Additionally, low intensity data was filtered by removing any probe sets that did not have a mean intensity of at least 15 in any disease. Filtering cutoffs were determined by comparing the intensity distribution from fRMA to that from standard RMA normalization using the same batch of data.

Online visualization tool

A Matlab standalone executable was built using Matlab Compiler Runtime (MCR). This executable is hosted by an Apache server and coordinates user input and output via HTML and Perl scripts. The tool is available at http://pathways-pellegrini.mcdb.ucla.edu/goTeles/dot_plot.html

Random Forest Classifier

The random forest algorithm builds a user-defined number of decision trees, each using a randomly selected subset of the training samples and a randomly selected subset of genes. A subset of samples is used to train the tree, with the remaining samples (termed “out of bag”) used to assess tree performance. Samples are evaluated by each decision tree independently and majority voting determines the final classification. Cross validation is built into the classifier training process, so that the internal tree performance assessment may be determined. Matlab’s TreeBagger class was used to build the classifier, with the oobError() and oobPredict() functions to characterize overall classifier performance and individual disease performance

statistics, respectively. Performance statistics for specific diseases were calculated by counting true positives, true negatives, false positives, and false negatives for out-of-bag predictions separately for each disease (i.e., each true positive was also counted as a true negative for all other diseases, and each false positive was also counted as a false negative for the true diagnosis disease). One instance of our classifier was built and used for all subsequent tests.

Cell-type specific signature enrichment

Briefly, 687 publicly available microarray samples on the Affymetrix HG U133 Plus 2.0 platform were selected as being representative of specific cell types, with the number of samples per cell type roughly equal. A moderated t-test and fold change criteria were employed to identify cell type specific signatures, by finding the 250 genes that were most significantly enriched in samples of one cell type, as compared to samples of the other 23 cell types.

Cell-type specific enrichment was calculated for each skin condition using an adapted methodology from Swindell, et al (*59*). For each skin condition, a signature score was calculated based upon the 250 genes identified for a given cell type. For each gene, the fold-change between the gene's expression in one condition relative to all others was calculated, yielding a set of 250 log-transformed fold-change estimates, i.e., FC_1, FC_2, …, FC_{250}. To calculate the signature score, we then obtained the weighted arithmetic mean of the 250 fold-change estimates. Weights were equal to the square root of the genes rank, such that greater weight was assigned to genes most specifically expressed by a given cell type (i.e., lowest p-value; moderated t-test), with less weight assigned to genes less specifically expressed by a given cell type. This procedure was repeated with respect to each disease and each of the 24 cell types. Within each cell type, enrichment scores for each disease were calculated relative to

all others. Mean fold change vectors for each disease were then clustered based on Euclidian distance.

Chapter 5. Characterization and classification of leprosy subtypes

Introduction

Leprosy, a disease caused by infection with the intracellular bacteria *Mycobacterium leprae*, has affected humans for millennia: evidence of M leprae infection has been found in archaeological samples from the 1st century AD (*104*). While leprosy remains a rare diagnosis in countries like the United States, the disease is endemic to populations in underdeveloped countries—particularly South America, Asia and Africa (*105*). In all its forms, leprosy can cause permanent, debilitating nerve damage (*106*). Furthermore, leprosy carries an enduring social stigma, putting patients at risk for reduced access to treatment and social marginalization (*107*).

Beyond its historical, clinical, and social context, leprosy is a fascinating model of the human immune response to pathogens, since patient responses to *M leprae* fall on a spectrum. On one end, the tuberculoid form of the disease (T-lep) has one skin lesion from which bacterial organisms cannot be cultured. T-lep lesions are characterized histologically as granulomatous and usually self-resolve, although permanent nerve damage can still result from these lesions. Lepromatous leprosy (L-lep) is the other end of the spectrum, where patients have multiple, bacteria-rich lesions which can get progressively worse without treatment. The mechanism of these divergent responses can be traced back to the initial immune response to M leprae: T-lep corresponds to a Th1, cell-mediated immune response, whereas patients with L-lep develop a Th2, or humoral, immune response (*108*).

This response spectrum is continuous and fluid, such that patients can present as “borderline” tuberculoid or lepromatous (usually corresponding to the bacillary index, or the amount of bacteria present in lesions). Additionally, patients can spontaneously transition from L-lep to a T-lep state, which is termed reversal reaction (RR). Finally, patients (usually with L-lep) can develop a concurrent inflammatory reaction called erythema nodosum leprosum (ENL), which is characterized by systemic inflammation and subcutaneous nodules (*93, 109*). These borderline

and overlying conditions represent additional opportunities to characterize human immune states.

Since these divergent immune programs arise in reaction to the same pathogen, host gene expression must play some role in determining the clinical course of leprosy. Previous studies used microarrays to compare gene expression profiles derived from leprosy lesions in order to identify genes and pathways associated with each subtype. Most notably, the presence of an IL-10 induced phagocytic pathway in L-lep and a vitamin D-induced antimicrobial response in T-lep indicate mechanisms of differing immune programs (*92*). More recent work has indicated a pattern of negative inverse correlation in Type I versus Type II interferon (IFN) regulated gene programs that corresponds to leprosy subtypes (*72*). Lesions from patients with L-lep, T-lep, and RR were scored based on the expression of genes regulated by either IFN-γ or IFN-β. L-lep had high IFN-γ and low IFN-β scores, whereas T-lep and RR lesions had low IFN-γ and high IFN-β scores. However, these microarray studies were unable to capture certain key cytokines and molecules that had previously been verified in lesions using PCR. In particular, cathelicidin (CAMP) was transcribed at levels too low for detection, and IFN-β was either also transcribed at too low a level or its presence was too transitory to be measured by microarray (*72, 73*).

The various forms of leprosy also enable the study of disease subtypes. In Chapter 2, we built a disease classifier that diagnosed individual diseases. The analysis of a diverse range of diseases can be relatively straightforward when the diseases have significantly different pathophysiological bases. Disease subtypes also have variations in pathophysiology, but on a more subtle level, since they may share a common pathogen or disease mechanism. Developing a disease subtype classifier could be a first step in building a useful clinical

diagnostic tool, as well as a means to identify genes associated with each subtype and provide insight into disease pathogenesis.

In this analysis, we studied gene expression profiles from lesional biopsy samples of leprosy that were derived using microarray and mRNA sequencing technology. We used RNA-seq expression profiles to identify genes and pathways associated with two leprosy subtypes, in particular those genes that were not detectable on microarray platforms. We also used multiple batches of microarray gene expression profiles to perform a concurrent analysis of four leprosy subtypes in which we constructed a leprosy subtype classifier and used the proportional median ranking metric to build subtype specific gene signatures.

Results

Characterization of gene expression profiles in leprosy subtypes using RNA seq

We obtained gene expression profiles from messenger RNA-seq data from four leprosy skin biopsy specimens: one L-lep, two T-lep, and one RR. Samples were mapped to a reference human genome using TopHat and gene expression was quantified using Cufflinks. Differential expression analysis corresponded well with previous microarray studies of gene expression differences in L-lep versus T-lep. However, these gene lists did not offer insight beyond previous microarray studies of differential expression.

Since an initial differential expression analysis did not provide additional information into divergent leprosy subtype programs, we focused on the analysis of genes that could not be measured using microarray technology. Of particular interest was cathelicidin (CAMP), an antimicrobial peptide upregulated in T-lep that could not be quantified on a microarray yet was confirmed to be present in T-lep lesions by other methods (*92*). We were able to detect CAMP in T-lep lesions using RNA seq, with a representative T-lep lesion showing an RPKM of 1381.9. CAMP was present in RR lesions at a lower rate (RPKM=659.9). Surprisingly, CAMP was also present in L-lep lesions, although at a lower RPKM of 268.9 (Figure 1). However, IFNB1 and IFNA, two cytokines that were undetectable on microarray gene expression profiles, were also not detected on RNA-seq of L-lep lesions, where they are known to play a role in pathology (data not shown). Thus, while RNA seq can pick up the expression of some low count genes not seen on microarray, others remained elusive.

Characterization of gene expression profiles in leprosy subtypes using microarrays

Since microarrays have proven to be a reliable measure of leprosy gene expression profiles (with a few noted exceptions), and since we had access to a large, pre-existing data set of

lesional leprosy gene expression profiles derived via microarray, the rest of this chapter therefore focuses on the analysis of microarray data. Lesional biopsy samples of the following leprosy subtypes were obtained: lepromatous leprosy (LL, n=6), tuberculoid leprosy (TL, n=10), reversal reaction (RR, n=7), and erythema nodosum leprosum (ENL, n=7). Gene expression profiles were derived via Affymetrix HG U133 Plus 2.0 microarrays and normalized using frozen RMA (fRMA).

Similar to previous results, hierarchical clustering of gene expression profiles demonstrated separation of the leprosy samples into three groups: LL, ENL, and TL/RR (Figure 2). The tuberculoid and reversal reaction forms of the disease clustered homogenously and we were unable to separate the two forms by clustering. ENL formed a separate group that was equally dissimilar to both LL and TL/RR. There were a few exceptions to this clustering pattern. Notably, one ENL sample clustered with the TL/RR group, and one TL sample clustered with the ENL group. Additionally, three samples (1 TL, 1 RR, and 1 ENL) did not cluster with any group.

Proportional median signatures identify subtype specific genes for downstream analysis and random forest classification

In order to identify genes that were highly expressed in one subtype relative to all others, we calculated proportional median (PM) values for all filtered probe sets in every subtype. Briefly, PM is a metric developed previously that acts as a fold change for comparing three or more conditions. The PM is calculated for each probe set in each disease by dividing the median expression of that probe in that disease by the median expression of that same probe across all diseases. Thus, PM calculation allowed the ranking of probe sets according to relative expression in one subtype compared to all others.

To detect pathways associated with each subtype, we performed functional annotation analysis of the top 250 genes of each PM signature using the gene annotation tools DAVID and Ingenuity Pathways Analysis (*24, 25*). Notably, the probe set with the highest PM in TL corresponded to the gene MMP12. Matrix metalloproteinases (MMPs) such as MMP-2 and MMP-9 have an established role in the formulation of granulomas, which are a hallmark of the tuberculoid form of leprosy (*110*).

Random forest classifier predicts leprosy subtypes

A random forest classifier was trained on the leprosy subtype gene expression profiles. To perform feature selection, we used PM signatures to filter for the most informative features for each subtype. Using the unique set of the 25 probe sets with highest PM for each subtype, we built a classifier and measured performance by calculating sensitivity, specificity, precision, and F1 score, which is a composite statistic that combines sensitivity and precision and ranges from 0 (completely inaccurate) to 1 (perfect accuracy). The PM-random forest classifier had an overall accuracy of 87%, with F1 scores ranging from 0.77 to 1.0 (Table 1). Our classifier was able to effectively differentiate between TL and RR within the training set, although there was still evidence of overlap, with one sample each of TL and RR being diagnosed as the other. However, there was also an equivalent amount of overlap between ENL and TL/RR subtype classification (Table 2).

We validated our classifier using 22 independent samples that were not used in the PM calculation, feature selection, or classifier training steps. We obtained these validation samples from publicly available microarray data (6 ENL and 7 LL) and by obtaining de-identified leprosy skin biopsy specimens and deriving gene expression profiles via microarrays (3 LL, 3 TL, and 3 RR) (*93*). Our classifier identified 68% of these samples correctly; however, nearly a third of the

mis-classifications were TL samples classified as RR. When TL and RR are counted as a combined group (as they were shown to be indistinguishable according to hierarchical clustering), the classifier validation accuracy increases to 77% (Table 3).

Functional analysis reveals signatures and cell types associated with subtypes

In order to identify genes and pathways associated with leprosy subtypes, as well as discover novel connections between genes, we performed WGCNA on the filtered gene expression profiles in our leprosy subtype training set. WGCNA uses correlations to place genes into modules – similarly to a traditional clustering analysis – but raises each correlation to a power, thus lending more weight to strong, more reliable correlations while still factoring in weaker correlations (*26*). WGCNA analysis identified 21 modules of associated genes. To determine which modules were associated with each subtype, we performed module eigengene correlation to our samples by coding traits as a sparse binary matrix of zeroes and ones: each sample had a value of '1' for its corresponding subtype and '0' for all other subtypes (Figure 3).

Only one module, 'magenta', was significantly correlated to T-lep (correlation=0.39, p-value=0.03). This module contained 387 genes and notably contained the probe set corresponding to MMP12, which was the top gene for T-lep by PM. Analysis of the genes contained in the 'magenta' module by Ingenuity Pathways Analysis showed connections between MMP12 and genes such as SMAD3, SMAD7, VEGF, and PLAUR, which are active in tissue remodeling and angiogenesis, which is consistent with the granuloma formation that is characteristic of the tuberculoid form of leprosy (Figure 4) (*109, 111-114*). Angiogenesis in leprosy has mostly been studied in the context of L-lep, and has been correlated to lesional bacillary index (*115*). However, more recent literature supports the formation of new blood and lymphatic vessels equally across the leprosy subtypes, especially in reactional states such as

RR and ENL. Furthermore, lymphangiogenesis, for which VEGF is also a marker, is present at a higher level in the tuberculoid states of the disease (*116, 117*).

The WGCNA analysis identified six modules significantly associated with ENL. Using DAVID functional analysis, we found that two of these modules, brown and tan, were significantly enriched for the Gene Ontology (GO) term "cell adhesion" (FDR 4.25×10^{-24} and 0.058, respectively). Additionally, three other modules (green, yellow, and turquoise) were significantly enriched for the InterPro protein cadherin, which functions in cell-cell adhesion (FDR 1.03×10^{-5}, 2.92×10^{-5}, and 0.00114, respectively). The tan module was also enriched for blood vessel development (FDR 1.18×10^{-8}), and the brown module had an abundance of genes involved in epidermis development (FDR 1.03×10^{-15}). Histologically, ENL is characterized by a thickening of the epidermis, which is supported by these functional pathways (*118*). Additionally, the enrichment for cell adhesion pathways has previously been reported and may be associated with lymphocyte or neutrophil recruitment to ENL lesions (*93, 118*).

There were five WGCNA modules significantly associated with L-lep. The salmon and blue modules were enriched for terms consistent with the phagocytic program characteristic of L-lep, such as "lysosome" (salmon: FDR 0.0434, blue: FDR 1.60×10^{-39}) and "vacuole" (blue: FDR 5.91×10^{-34}) (*92*). The royal blue module contained a significant number of genes involved in immunoglobulin pathways (FDR 3.77×10^{-13}). This is consistent with reports that L-lep lesions often have a humoral, B cell component (*119*).

In order to determine whether expression of these pathways was due to enrichment of a particular cell type, we performed cell type specific deconvolution on our leprosy subtype gene expression profiles (Figure 5). Although findings for T-lep and RR were nonspecific, the cell

types enriched in ENL and L-Lep supported the findings from WGNCA. ENL was significantly enriched for a gene signature derived from lymphatic and blood vessel endothelial cells, as well as enriched for CD34+ cells, which are involved in cell adhesion (*120*). L-lep showed significant enrichment for the phagocytic macrophage, dendritic cells, and monocytes, which was consistent with WGCNA findings. Although the enrichment was not significant, L-lep also showed its third strongest enrichment for B cells, which was supported by pathways analysis of the WGCNA royal blue module showing an enrichment for immunoglobulin encoding genes.

Tissue immunostaining of leprosy skin lesions shows higher expression of MMP-12 in tuberculoid versus lepromatous forms

MMPs are known to be involved in granuloma formation, inflammation, and more specifically, the tuberculoid form of leprosy. Therefore, we investigated MMP12 further as a novel marker of tuberculoid leprosy via immunohistochemistry staining of fresh frozen leprosy tissue sections. Tissue was H&E stained to visualize cells and skin structure. We used antibodies for MMP-12 to probe for the presence of the protein in three samples each of L-lep, T-lep, and RR. We also probed for the presence of CD3 to ascertain binding of the antibodies, as well as IgG1 to probe for nonspecific binding. We selected representative examples of each subtype and antibody (Figure 6). Positive staining for MMP-12 was detected in T-lep and RR but not L-lep or ENL, confirming its ability to differentiate between the tuberculoid/reversal reaction morphologies and the lepromatous form.

Discussion

Leprosy has been well studied as a model of innate and adaptive immunity, making it an appealing model to train a disease subtype classifier. Furthermore, pathway and cell type enrichment has not yet been studied across four leprosy subtypes. We first performed a brief investigation of RNA-seq in two subtypes, L-lep versus T-lep, as a tool to detect genes that are expressed at low levels or are otherwise undetectable using microarrays. Then, we shifted our focus to the use of microarray gene expression profiles across all four subtypes. Using PM and a random forest classifier, we were able to successfully classify RR, L-lep, T-lep, and ENL in our original training set and an independent validation set. We supplemented analysis of the PM signatures for each subtype with WGCNA and cell type specific deconvolution. Using these analysis methods, we were able to identify a number of pathways associated with each subtype that support current knowledge of the disease, such as cell adhesion pathway in ENL and tissue remodeling in T-lep. Finally, we used a synthesis of our PM and functional analysis methods to identify a novel marker of T-lep that we confirmed in leprosy lesional tissue sections.

Methods

Leprosy RNA sequencing

These lesions were obtained from a leprosy clinic in Brazil, and shipped to the United States as flash-frozen, OCT-embedded blocks. mRNA was extracted directly from the lesions, and cDNA libraries were built using a slight variation on the standard Illumina protocol: to adjust for the lower cell density in tissue samples, an additional filtering step was implemented to prevent mRNA loss. The L-lep sample was sequenced on an Illumina GA IIx machine and the rest of the samples were sequenced on an Illumina HiSeq machine, resulting in approximately four times more reads in the latter samples (Table S1). Additionally, the T-lep and RR cDNA libraries were built from mRNA that was approximately ten times more concentrated than the mRNA used to build the L-lep library.

The analysis pipeline was identical for each of the four samples. First, sequenced reads were parsed and assessed for quality. Quality was determined according to a number of criteria including average Phred quality scores at each position in the read, read duplication, and presence of Illumina adaptor sequences. Reads with drop-offs in Phred quality were trimmed to remove positions with particularly low quality scores. Next, the trimmed reads were mapped to the Hg18 build of the human genome using TopHat, a short read mapper that performs gapped alignments to a reference genome (*121*). Mapping statistics were compared across samples to confirm that they were roughly equivalent (Table S1). Additionally, meta-gene plots were generated to confirm the absence of 3' bias.

The mapped reads were assembled into transcripts using Cufflinks (*122*). Rather than simply counting reads that map to genes, Cufflinks adjusts for reads that map to multiple locations, as well as for gene length. Thus, while gene expression in mRNA-seq experiments is usually represented as Reads Per Kilobase of exon model per Million mapped reads (RPKM), Cufflinks

uses the value Fragments Per Kilobase of exon model per Million mapped reads (FPKM) (*123*). FPKM can be thought of as a statistical estimate of RPKM, given that some reads map ambiguously. Using Cufflinks, FPKM was calculated for every gene in the Hg18 annotation. Additionally, RPKM counts were calculated using the HTSeq package in Python (*124*). Cufflinks tends to filter out transcripts that it deems "low-count" (*125*). Since a major focus in this study is low-count genes, we also visualized the mapped reads on the UCSC genome browser, and used this to follow up with a list of low-count genes that were relevant to either L-lep or T-lep, according to a review of innate immunity in leprosy (*108, 126*).

Microarray, normalization, and clustering

Skin lesional biopsy specimens were obtained from patients with the following leprosy subtypes ("training set"): lepromatous leprosy (n=6), tuberculoid leprosy (n=10), reversal reaction (n=7), and erythema nodosum leprosum (n=7). mRNA was extracted from lesions and gene expression profiles were derived via Affymetrix HG U133 Plus 2.0 microarrays as previously described in Bleharski, et al (*33*). Additional data for validation ("validation set") was obtained from a second batch of skin lesional biopsy specimens (lepromatous leprosy: n=3; tuberculoid leprosy: n=3; reversal reaction: n=3) and publicly available data on NCBI GEO (GSE16844; lepromatous leprosy: n=7; and erythema nodosum leprosum: n=6) (*93*).

Data were normalized using frozen RMA and filtered at a mean intensity of at least 150 in any one subtype. Filtered data were used in the proportional median signature, classifier, clustering and WGCNA analyses. Hierarchical clustering was performed using correlation distances and the "amap" package in R.

Proportional median signatures

The proportional median (PM) metric was defined as the median intensity of a probe set within one leprosy subtype divided by the median intensity of the same probe set across all samples. PM values were calculated for each subtype using the training set and ranked according to highest PM.

Random forest classifier

A random forest classifier was built from the training set using the Matlab TreeBagger class. PM values were used to select the most informative features for each subtype by using the top 25 probe sets by PM to build the classifier. True positives, false positives, and false negatives were calculated directly from out-of-bag predictions separately for each subtype. True negatives for a particular subtype were calculated from true positives of all other subtypes (i.e., a true positive for subtype A counted as a true negative for subtype B).

Cell type specific signature enrichment

Cell specific enrichment was calculated on the training set using an adapted methodology from Swindell, et al (*59*). Using 250 gene signatures for each cell type, signature scores were calculated for each subtype based on the weighted arithmetic mean of the fold-change between a particular gene's expression in one subtype relative to all others. Weights were assigned based on how specific each gene was to the particular cell type. Within each cell type, enrichment scores for each disease were calculated relative to all others. Mean fold change vectors for each disease were then clustered based on Euclidian distance.

Weighted Gene Correlation Network Analysis

Weighted Gene Correlation Network Analysis (WGCNA) was performed on filtered gene expression profiles of the training set ('wgcna' package in R) (*26*). Automatic network

construction was carried out with a power of 14 and a minimum module size of 50. For each module, networks were constructed using the topological overlap matrix. The top 50 probes from each network were selected by filtering by kME (intramodular connectivity) and converted to gene names before displaying. Networks were built using VisANT. Module correlation to leprosy subtypes was calculated by computing the correlation of each module eigengene to a binary matrix of traits which corresponded to individual subtypes. Correlation and significance calculations, as well as heatmap display, were calculated using built-in functions from the 'wgcna' R package.

Tissue Immunostaining

Frozen tissue sections were blocked with normal horse serum before incubation with MMP-12 monoclonal antibody (mAb) and isotype control for 60 min, followed by incubation with biotinylated horse anti-mouse IgG for 30 min. Slides were counterstained with hematoxylin and mounted in crystal mounting medium (Biomeda, Foster City, CA) and were visualized using the ABC Elite system (Vector Laboratories, Burlingame, CA). Skin sections was examined using a Leica microscope (Leica, Heidelberg, Germany).

Figure Legends

Figure 1. UCSC Genome Browser view of cathelicidin expression (CAMP) in L-lep, T-lep, and reversal reaction (RR). RNA-seq gene expression profiling was performed on leprosy samples from L-lep, T-lep, and RR. Counts were normalized to counts per million mapped reads and visualized on the UCSC Genome Browser. Shown are the reads mapping to the gene CAMP.

Figure 2. **Hierarchical clustering of leprosy subtypes**. Microarray gene expression profiles derived from L-lep, T-lep, RR, and ENL skin lesional biopsy specimens were subjected to unsupervised hierarchical clustering. Clustering distance was calculated by correlation of filtered gene expression profiles. Trees were built using average linkage distance.

Figure 3. Weighted Gene Co-expression Network Analysis (WGCNA) of leprosy subtypes. WGCNA was performed on leprosy microarray gene expression profiles in order to find modules of significantly correlated genes. **A**. WGCNA dendrogram showing relationships of probe sets and modules. Each leaf on the dendrogram is a microarray probe set, and the module it belongs to is shown on the color bar below. Leaves further down on the dendrogram represent more closely connected modules. **B**. Correlation of module eigengenes (MEs) to leprosy subtypes. Modules are shown on the Y axis, labeled with the number of probe sets in each module. Correlation is shown for each square in the heatmap, with p-value below in parentheses.

Figure 4. Ingenuity Pathways Analysis network of MMP12 connections in WGCNA magenta module. The 387 probe sets in the WGCNA magenta module (associated with T-lep) were converted to gene names using Ingenuity Pathways Analysis and visualized in a network

using known connections. The genes with direct connections to MMP12 were isolated into a separate network.

Figure 5. Cell Type Specific Deconvolution. Cell type specific enrichment was calculated using 24 different cell types. Using signatures characteristic of each cell type, log fold changes were calculated per leprosy subtype, where each fold change represents the enrichment for a particular cell type signature in that subtype. Fold change vectors were clustered using Euclidean distance and displayed in a heatmap, with rows corresponding to leprosy subtypes and columns corresponding to cell types. Note that enrichment scores are relative across each cell type. Black triangles denote FDR < 0.05 and directionality of fold change.

Figure 6. MMP-12 immunostaining in T-lep, L-lep, RR, and ENL. Frozen sections of leprosy lesions were stained with H&E and then incubated with antibodies for MMP-12, CD3 (positive control) and IgG1 (negative control). Three sections of each subtype were stained; shown are representative pictures of each subtype at 10x magnification.

Figures

Figure 1

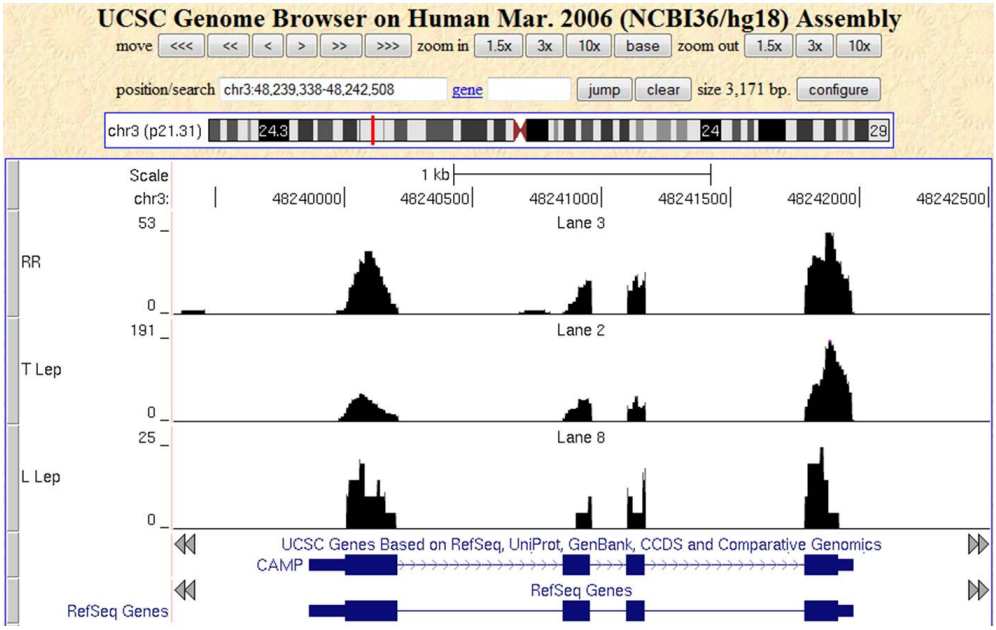


Figure 2

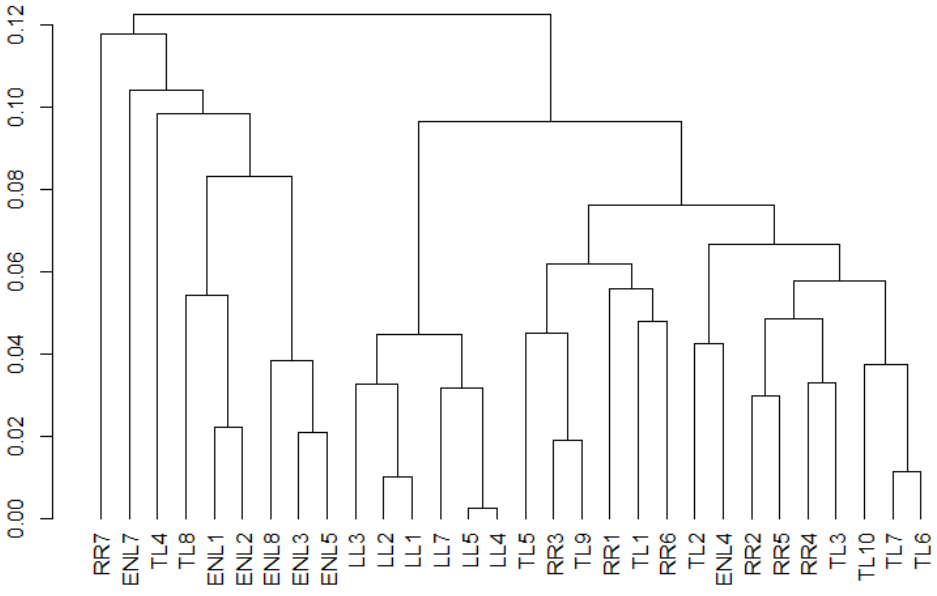


Figure 3

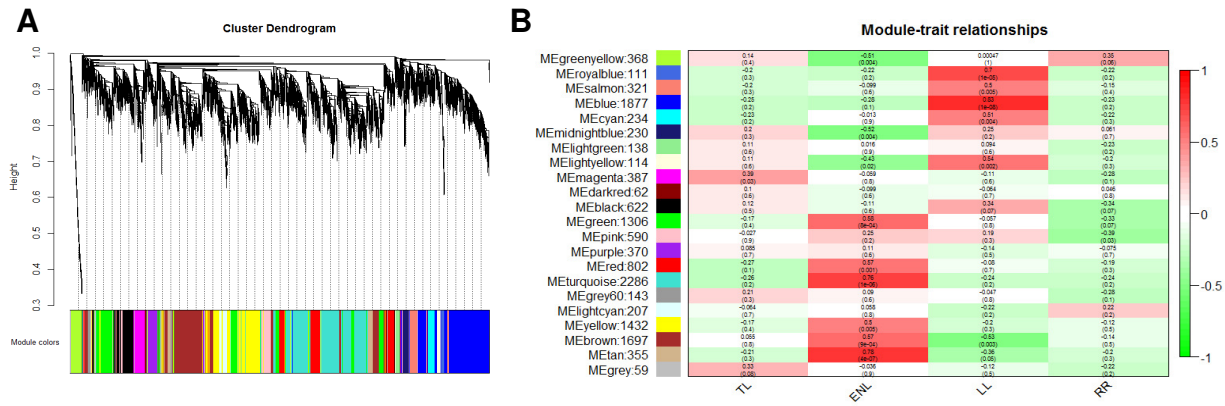


Figure 4

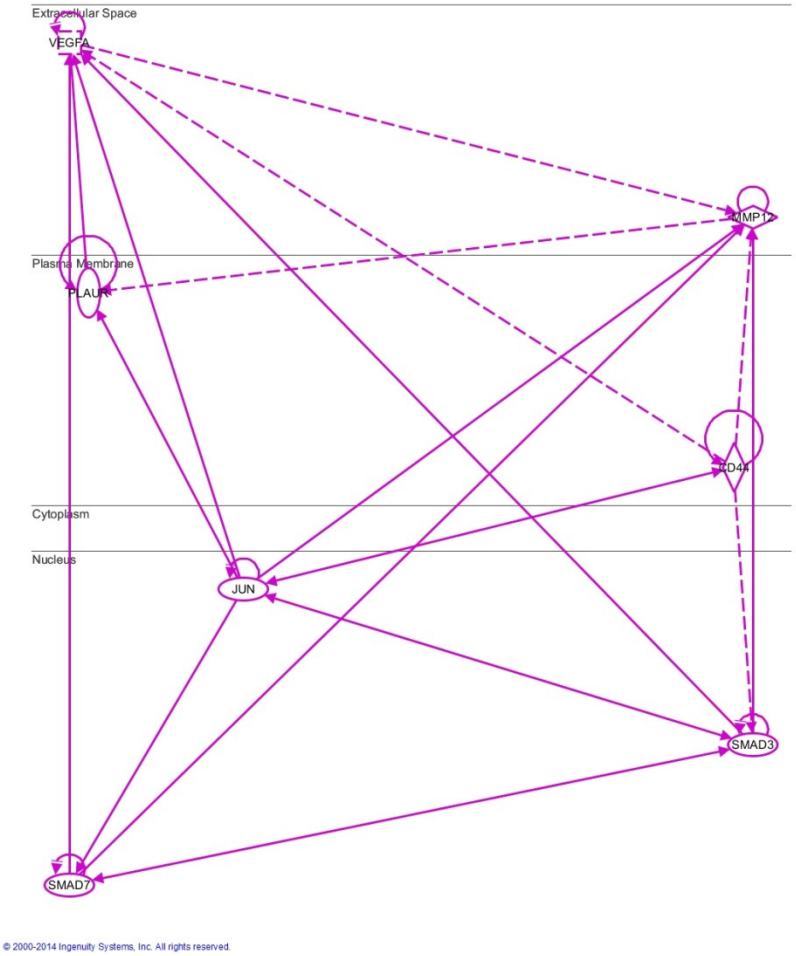


Figure 5

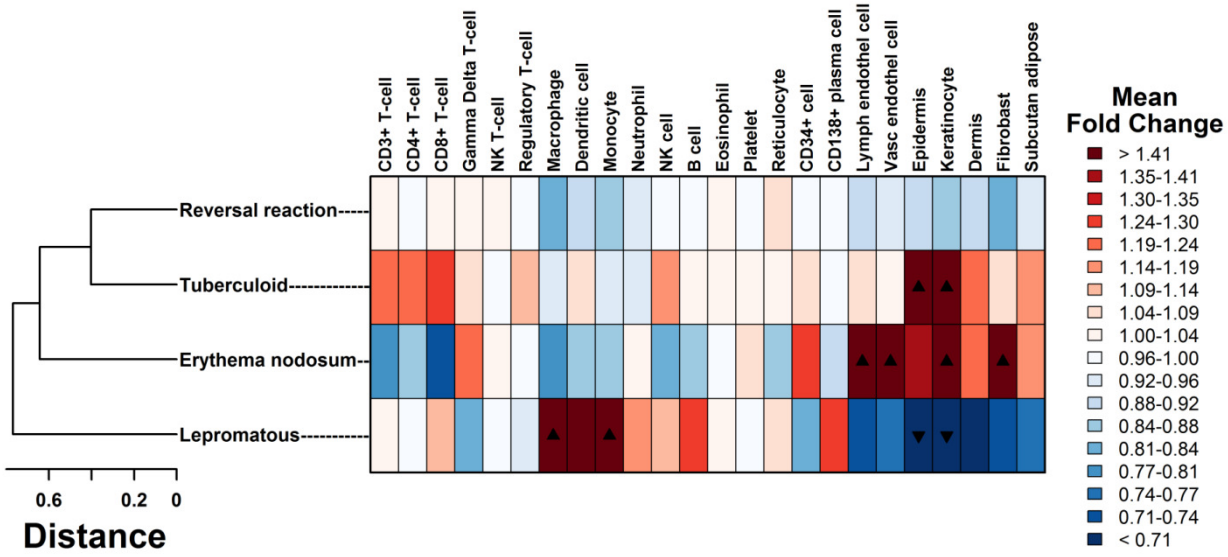
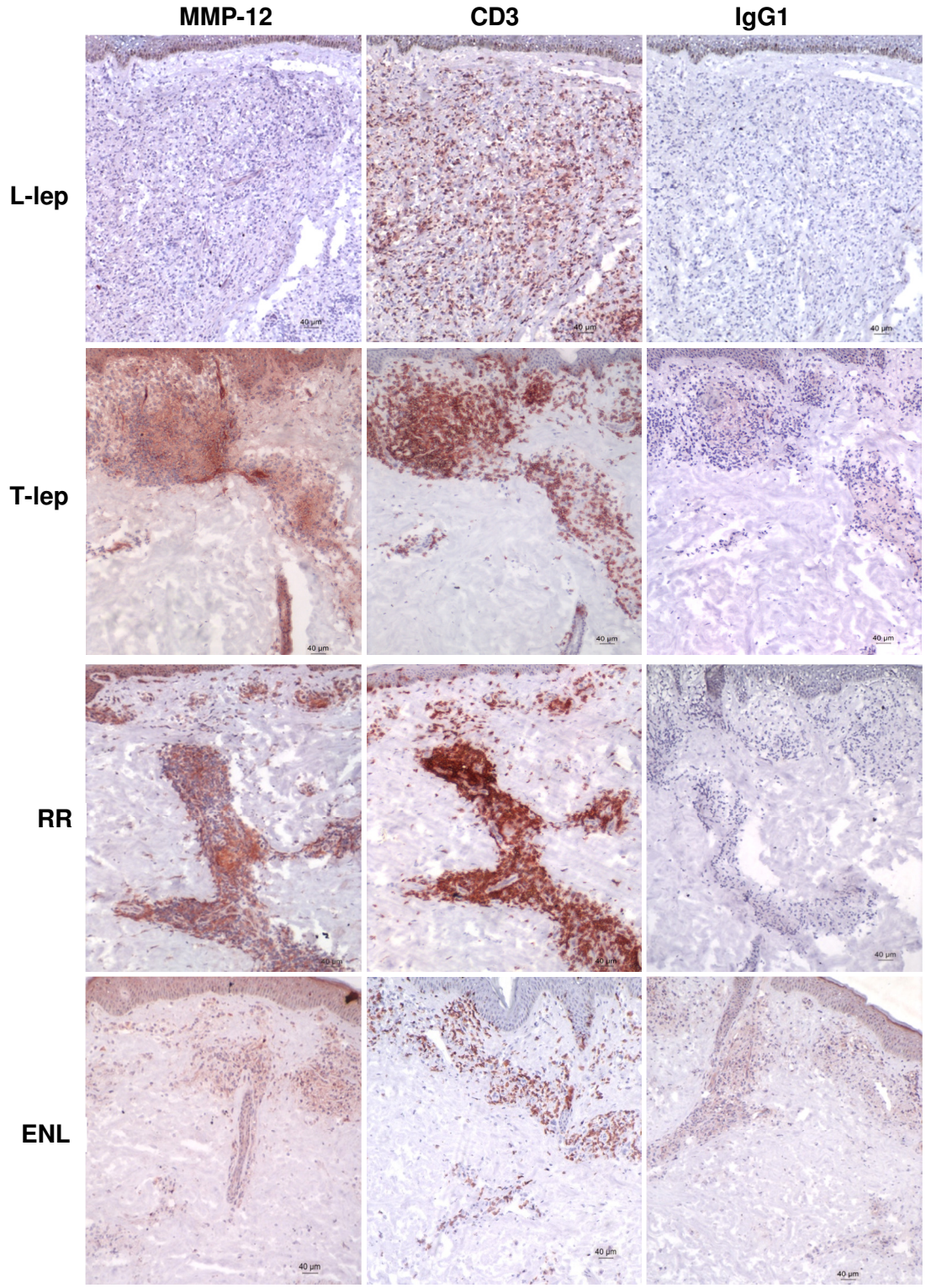

CD3+ T-cell
CD4+ T-cell
CD8+ T-cell
Gamma Delta T-cell
NK T-cell
Regulatory T-cell
Macrophage
Dendritic cell
Monocyte
Neutrophil
NK cell
B cell
Eosinophil
Platelet
Reticulocyte
CD34+ cell
CD138+ plasma cell
Lymph endothel cell
Vasc endothel cell
Epidermis
Keratinocyte
Dermis
Fibrobast
Subcutan adipose
Reversal reaction
Tuberculoid
Erythema nodosum
Lepromatous
0.6
0.2
0
Distance
Mean Fold Change
> 1.41
1.35-1.41
1.30-1.35
1.24-1.30
1.19-1.24
1.14-1.19
1.09-1.14
1.04-1.09
1.00-1.04
0.96-1.00
0.92-0.96
0.88-0.92
0.84-0.88
0.81-0.84
0.77-0.81
0.74-0.77
0.71-0.74
< 0.71

Figure 6



Tables

Table 1. Leprosy subtype classifier internal performance.

Table 1

	Sens	Spec	Prec	F1
L-Lep	1.00	1.00	1.00	1.00
T-Lep	0.90	0.90	0.82	0.86
RR	0.71	0.96	0.83	0.77
ENL	0.86	0.96	0.86	0.86

Table 2. Actual diagnosis versus predicted classification in leprosy subtype classifier training set.

Table 2

Training	Actual –ENL	Actual – L-Lep	Actual – RR	Actual – T-Lep
Predicted - ENL	6	0	0	1
Predicted - L-Lep	0	6	0	0
Predicted - RR	1	0	5	1
Predicted - T-Lep	0	0	1	9

Table 3. Actual subtype versus predicted subtype for independent validation leprosy samples. Samples which were not used at any step in the classifier training were run through the classifier.

Table 3

Validation	Predicted-ENL	Predicted-L-Lep	Predicted-RR	PredictedlT-Lep
Actual-ENL	6	0	0	0
Actual-L-lep	0	6	3	1
Actual-RR	0	0	3	0
Actual-T-lep	0	1	2	0

Supplementary Information

Table S1. Mapping statistics for leprosy RNA-seq

Supplementary Table 1

sample name	Illumina machine used	Read length	# of reads	% mapped	% mapped uniquely
L-Lep	Ga IIx	75	37568548	83.24	53.84
T-Lep 1	HiSeq	100	117904034	83.24	55.91
T-Lep 2	HiSeq	100	125313580	87.14	58.74
RR	HiSeq	100	115103849	84.36	59.46

Chapter 6: Conclusion

Gene expression profiling has become a common first step in the exploration of genes and pathways that contribute to human disease phenotypes. Their power lies in the ability to obtain a comprehensive picture of gene expression in any sample using methods that are fast and economically feasible. As a result of the widespread use of such technology, the body of private and public gene expression profile data has exploded, providing exciting opportunities to re-analyze the data and obtain new information in an even more efficient, cost-effective way.

The first section of this work was a brief study of microarray gene expression profiles in pancreatic ductal adenocarcinoma (PDAC). The study of PDAC remains problematic due to a lack of lesional biopsy specimens, especially those from patients with terminal disease. Currently, surgically removed tumors in patients without evidence of metastasis comprise the majority of tissue available for study. However, correlates of disease free survival found in these patients are still useful as potential targets for therapy or biomarkers for disease prognosis. Furthermore, the majority of surgical patients experience disease recurrence, implying that micro-metastases were present at the time of surgery and making these gene expression profiles directly relevant to more advanced cases (*127*).

Access to more PDAC samples and more sophisticated analysis techniques such mRNA sequencing are two strategies for the next steps in this analysis. A variety of PDAC microarray data sets are freely available on NCBI GEO, including primary tumors, metastatic tumors, xenografts of human PDAC tumors in mice, and stromal tissue (tissue surrounding primary tumors). These data could be integrated with new or existing data to yield large data sets that have an enhanced ability to explore genes and pathways associated with disease prognosis. On the high throughput sequencing front, alterations in PDAC genomic content were

characterized and RNA-seq gene expression profiles derived for a few tumors in a 2012 study (*128*). However, RNA sequencing has only been carried out on a small scale on tumors taken from actual surgical cases versus mouse xenografts and cell lines (*129, 130*). RNA-seq is particularly useful in the context of cancer gene expression profiles since it captures quantitative levels of gene expression as well as differential isoform expression and intronic regulatory elements.

The analysis strategy for the first chapter of this work is representative of the majority of current gene expression profile analyses: samples from two conditions are compared to find differentially expressed genes. Typical studies involving three or more conditions compare each condition in a pairwise fashion to a common control. However, many diseases—particularly those that manifest in the same tissue—have shared or interconnected pathophysiology, such that a multi-disease analysis would yield more information than a series of pairwise comparisons. The second chapter of this work focuses on techniques for simultaneously comparing gene expression profiles derived from lesional biopsy samples of a range of human diseases in a single tissue. Skin was chosen as the tissue of interest since a wide range of disease lesions manifest in skin and are readily accessible for biopsy excision. The backbone of these analyses is the successful integration of data from multiple batches, since the range and number of biopsy specimens studied far exceeds the resources of most investigators. Although batch effect is a concern in these data, we took advantage of data set features such as large sample size, independent, external validation, and robust algorithms like rank-based proportional median in order to identify strong biological signals.

This work has numerous opportunities for expansion. The most straightforward next step would be the integration of additional gene expression profiles—both increasing the sample size of

diseases currently in the data base and adding new diseases. For the problem of disease classification, careful consideration would be necessary for partitioning new samples into current training and test sets. While our multi-disease classifier had good performance both in the training set and for five externally validated conditions, it lacked independent validation sets for a majority of conditions. Therefore, although an increase in sample size should increase classifier accuracy, any new samples for diseases already in the data base would be prioritized to the validation set. The addition of new samples could be expedited by the inclusion of more microarray platforms, or even the integration of RNA-seq gene expression profiles with those derived via microarray. However, gene expression can be unstable between microarray platforms, making direct comparison of gene intensities a theoretical issue that has yet to be solved (*6, 131-133*). The integration of RNA-seq and microarray data is another area of active research that has been investigated with mixed results (*134, 135*).

The final section of this work examined the problem of classification and gene expression profile analysis in the context of multiple subtypes from one skin disease, leprosy, in which a range of patient immunological responses contribute to a disease with a spectrum of phenotypes. Many of the analysis techniques from the general skin disease analysis were applicable to the problem of leprosy subtypes. Indeed, using approaches including proportional medians, random forest classification, and cell type specific deconvolution, we were able to successfully classify leprosy subtypes as well as identify immunological and structural differences in gene expression that gave rise to the various disease phenotypes. This work provided a proof of concept that the subtler variations in disease subtypes can be identified using multi-disease comparison techniques, and future steps include the exploration of these methods in other skin diseases with variable subtypes such as psoriasis, or even infectious diseases like tuberculosis which has latent and active forms.

The development of high throughput gene expression profiling opened a window into the inner workings of tissues and cells. This view is, at times, bafflingly complex. Adding to this complexity are the myriad platforms and methods by which gene expression profiles may be obtained, resulting in a large collective body of transcriptome data that is too often underutilized. While simple, universally accepted methods will never exist for the integration of all types of genomic or transcriptomic data, one can leverage large sample size and robust biological signals in order to make the science of gene expression profile analysis more streamlined and efficient.

References

1. S. McGinn, I. G. Gut, DNA sequencing - spanning the generations. *N Biotechnol* **30**, 366-372 (2013); published online EpubMay (10.1016/j.nbt.2012.11.012).
2. J. Henson, G. Tischler, Z. Ning, Next-generation sequencing and large genome assemblies. *Pharmacogenomics* **13**, 901-915 (2012); published online EpubJun (10.2217/pgs.12.72).
3. D. Aird, M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, A. Gnirke, Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18 (2011)10.1186/gb-2011-12-2-r18).
4. S. C. Shin, d. H. Ahn, S. J. Kim, H. Lee, T. J. Oh, J. E. Lee, H. Park, Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS One* **8**, e68824 (2013)10.1371/journal.pone.0068824).
5. L. Hoopes, Genetic Diagnosis: DNA Microarrays and Cancer. *Nature Education* **1**, 3 (2008).
6. R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martínez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, W. Yu, Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**, 345-350 (2005); published online EpubMay (10.1038/nmeth756).
7. D. Wong, B. Kea, R. Pesich, B. W. Higgs, W. Zhu, P. Brown, Y. Yao, D. Fiorentino, Interferon and biologic signatures in dermatomyositis skin: specificity and heterogeneity across diseases. *PLoS One* **7**, e29161 (2012)10.1371/journal.pone.0029161).
8. J. A. DeWoody, K. C. Abts, A. L. Fahey, Y. Ji, S. J. Kimble, N. J. Marra, B. K. Wijayawardena, J. R. Willoughby, Of contigs and quagmires: next-generation sequencing pitfalls associated with transcriptomic studies. *Mol Ecol Resour* **13**, 551-558 (2013); published online EpubJul (10.1111/1755-0998.12107).
9. A. Oshlack, M. D. Robinson, M. D. Young, From RNA-seq reads to differential expression results. *Genome Biol* **11**, 220 (2010)10.1186/gb-2010-11-12-220).
10. E. Suárez, A. Burguete, G. J. Mclachlan, Microarray data analysis for differential expression: a tutorial. *P R Health Sci J* **28**, 89-104 (2009); published online EpubJun (
11. B. Duval, J. K. Hao, Advances in metaheuristics for gene selection and classification of microarray data. *Brief Bioinform* **11**, 127-141 (2010); published online EpubJan (10.1093/bib/bbp035).
12. R. Simon, Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol* **23**, 7332-7341 (2005); published online EpubOct (10.1200/JCO.2005.02.8712).
13. G. Janecek, W. Gansterer, M. Demel, G. Ecker, in *JMLR: Workshop and Conference Proceedings* (2008), vol. 4, pp. 90-105.
14. C. Ambroise, G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* **99**, 6562-6566 (2002); published online EpubMay (10.1073/pnas.102102699).
15. S. S. Koh, M. L. Opel, J. P. Wei, K. Yau, R. Shah, M. E. Gorre, E. Whitman, P. K. Shitabata, Y. Tao, A. J. Cochran, P. Abrishami, S. W. Binder, Molecular classification of melanomas and nevi using gene expression microarray signatures and formalin-fixed and paraffin-embedded tissue. *Mod Pathol* **22**, 538-546 (2009); published online EpubApr (10.1038/modpathol.2009.8).
16. L. Marchionni, R. F. Wilson, S. S. Marinopoulos, A. C. Wolff, G. Parmigiani, E. B. Bass, S. N. Goodman, Impact of gene expression profiling tests on breast cancer outcomes. *Evid Rep Technol Assess (Full Rep)*, 1-105 (2007); published online EpubDec (

17. R. Sanz-Pamplona, A. Berenguer, D. Cordero, S. Riccadonna, X. Solé, M. Crous-Bou, E. Guinó, X. Sanjuan, S. Biondo, A. Soriano, G. Jurman, G. Capella, C. Furlanello, V. Moreno, Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. *PLoS One* **7**, e48877 (2012)10.1371/journal.pone.0048877).
18. C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, D. Botstein, Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000); published online EpubAug (10.1038/35021093).
19. R. Sabatier, A. Gonçalves, F. Bertucci, Personalized medicine: Present and future of breast cancer management. *Crit Rev Oncol Hematol*, (2014); published online EpubMar (10.1016/j.critrevonc.2014.03.002).
20. M. C. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard, J. S. Parker, C. M. Perou, M. J. Ellis, T. O. Nielsen, Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst* **101**, 736-750 (2009); published online EpubMay (10.1093/jnci/djp082).
21. X. J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. T. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, D. C. Sgroi, A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5**, 607-616 (2004); published online EpubJun (10.1016/j.ccr.2004.05.015).
22. S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, B. R. Conklin, MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7 (2003).
23. d. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009); published online EpubJan (10.1093/nar/gkn923).
24. d. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009)10.1038/nprot.2008.211).
25. I. Systems.
26. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008)10.1186/1471-2105-9-559).
27. R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002); published online EpubJan (
28. S. Tian, J. G. Krueger, K. Li, A. Jabbari, C. Brodmerkel, M. A. Lowes, M. Suárez-Fariñas, Meta-analysis derived (MAD) transcriptome of psoriasis defines the "core" pathogenesis of disease. *PLoS One* **7**, e44274 (2012)10.1371/journal.pone.0044274).
29. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007); published online EpubJan (10.1093/biostatistics/kxj037).
30. C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, C. Liu, Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6**, e17238 (2011)10.1371/journal.pone.0017238).
31. M. N. McCall, B. M. Bolstad, R. A. Irizarry, Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242-253 (2010); published online EpubApr (10.1093/biostatistics/kxp059).

32. M. N. McCall, R. A. Irizarry, Thawing Frozen Robust Multi-array Analysis (fRMA). *BMC Bioinformatics* **12**, 369 (2011)10.1186/1471-2105-12-369).
33. J. R. Bleharski, H. Li, C. Meinken, T. G. Graeber, M. T. Ochoa, M. Yamamura, A. Burdick, E. N. Sarno, M. Wagner, M. Röllinghoff, T. H. Rea, M. Colonna, S. Stenger, B. R. Bloom, D. Eisenberg, R. L. Modlin, Use of genetic profiling in leprosy to discriminate clinical forms of the disease. *Science* **301**, 1527-1530 (2003); published online EpubSep (10.1126/science.1087785).
34. R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, T. P. Speed, Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003); published online EpubFeb (
35. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003); published online EpubApr (10.1093/biostatistics/4.2.249).
36. A. J. Hackstadt, A. M. Hess, Filtering for increased power for microarray data analysis. *BMC Bioinformatics* **10**, 11 (2009)10.1186/1471-2105-10-11).
37. L. Breiman, Random Forests. *Machine Learning* **45**, 5-32 (2001).
38. A. Prinzie, D. Van den Poel, Random Forests for multiclass classification: Random MultiNomial Logit. *Expert Systems with Applications* **34**, 1721-1732 (2008).
39. P. Langfelder, P. S. Mischel, S. Horvath, When is hub gene selection better than standard meta-analysis? *PLoS One* **8**, e61505 (2013)10.1371/journal.pone.0061505).
40. A. M. Yip, S. Horvath, Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* **8**, 22 (2007)10.1186/1471-2105-8-22).
41. A. L. Mihaljevic, C. W. Michalski, H. Friess, J. Kleeff, Molecular mechanism of pancreatic cancer--understanding proliferation, invasion, and metastasis. *Langenbecks Arch Surg* **395**, 295-308 (2010); published online EpubApr (10.1007/s00423-010-0622-5).
42. C. Guerra, M. Barbacid, Genetically engineered mouse models of pancreatic adenocarcinoma. *Mol Oncol* **7**, 232-247 (2013); published online EpubApr (10.1016/j.molonc.2013.02.002).
43. S. M. Hong, J. Y. Park, R. H. Hruban, M. Goggins, Molecular signatures of pancreatic cancer. *Arch Pathol Lab Med* **135**, 716-727 (2011); published online EpubJun (10.1043/2010-0566-RA.1).
44. V. M. Aris, M. J. Cody, J. Cheng, J. J. Dermody, P. Soteropoulos, M. Recce, P. P. Tolias, Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer. *BMC Bioinformatics* **5**, 185 (2004); published online EpubNov (10.1186/1471-2105-5-185).
45. U. Mahlknecht, A. D. Ho, S. Letzel, S. Voelter-Mahlknecht, Assignment of the NAD-dependent deacetylase sirtuin 5 gene (SIRT5) to human chromosome band 6p23 by in situ hybridization. *Cytogenet Genome Res* **112**, 208-212 (2006)10.1159/000089872).
46. C. Salaün, C. Leroy, A. Rousseau, V. Boitez, L. Beck, G. Friedlander, Identification of a novel transport-independent function of PiT1/SLC20A1 in the regulation of TNF-induced apoptosis. *J Biol Chem* **285**, 34408-34418 (2010); published online EpubNov (10.1074/jbc.M110.130989).
47. F. B. Müller, M. Huber, T. Kinaciyan, I. Hausser, C. Schaffrath, T. Krieg, D. Hohl, B. P. Korge, M. J. Arin, A human keratin 10 knockout causes recessive epidermolytic hyperkeratosis. *Hum Mol Genet* **15**, 1133-1141 (2006); published online EpubApr (10.1093/hmg/ddl028).
48. K. F. Tang, Y. Wang, P. Wang, M. Chen, Y. Chen, H. D. Hu, P. Hu, B. Wang, W. Yang, H. Ren, Upregulation of PHLDA2 in Dicer knockdown HEK293 cells. *Biochim Biophys Acta* **1770**, 820-825 (2007); published online EpubMay (10.1016/j.bbagen.2007.01.004).

49. R. A. Scelfo, C. Schwienbacher, A. Veronese, L. Gramantieri, L. Bolondi, P. Querzoli, I. Nenci, G. A. Calin, A. Angioni, G. Barbanti-Brodano, M. Negrini, Loss of methylation at chromosome 11p15.5 is common in human adult tumors. *Oncogene* **21**, 2564-2572 (2002); published online EpubApr (10.1038/sj.onc.1205336).
50. C. Partensky, Toward a better understanding of pancreatic ductal adenocarcinoma: glimmers of hope? *Pancreas* **42**, 729-739 (2013); published online EpubJul (10.1097/MPA.0b013e318288107a).
51. M. Herreros-Villanueva, E. Hijona, A. Cosme, L. Bujanda, Mouse models of pancreatic cancer. *World J Gastroenterol* **18**, 1286-1294 (2012); published online EpubMar (10.3748/wjg.v18.i12.1286).
52. D. Chaussabel, C. Quinn, J. Shen, P. Patel, C. Glaser, N. Baldwin, D. Stichweh, D. Blankenship, L. Li, I. Munagala, L. Bennett, F. Allantaz, A. Mejias, M. Ardura, E. Kaizer, L. Monnet, W. Allman, H. Randall, D. Johnson, A. Lanier, M. Punaro, K. M. Wittkowski, P. White, J. Fay, G. Klintmalm, O. Ramilo, A. K. Palucka, J. Banchereau, V. Pascual, A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150-164 (2008); published online EpubJul (10.1016/j.immuni.2008.05.012).
53. D. M. Mutch, A. Berger, R. Mansourian, A. Rytz, M. A. Roberts, The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics* **3**, 17 (2002); published online EpubJun (
54. Y. Tu, G. Stolovitzky, U. Klein, Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci U S A* **99**, 14031-14036 (2002); published online EpubOct (10.1073/pnas.222164199).
55. W. E. Beer, A. E. Smith, J. Y. Kassab, P. H. Smith, C. M. Rowland Payne, Concomitance of psoriasis and atopic dermatitis. *Dermatology* **184**, 265-270 (1992).
56. T. Henseler, E. Christophers, Disease concomitance in psoriasis. *J Am Acad Dermatol* **32**, 982-986 (1995); published online EpubJun (
57. S. Eyerich, A. T. Onken, S. Weidinger, A. Franke, F. Nasorri, D. Pennino, M. Grosber, F. Pfab, C. B. Schmidt-Weber, M. Mempel, R. Hein, J. Ring, A. Cavani, K. Eyerich, Mutual antagonism of T cells causing psoriasis and atopic eczema. *N Engl J Med* **365**, 231-238 (2011); published online EpubJul (10.1056/NEJMoa1104200).
58. W. R. Swindell, X. Xing, P. E. Stuart, C. S. Chen, A. Aphale, R. P. Nair, J. J. Voorhees, J. T. Elder, A. Johnston, J. E. Gudjonsson, Heterogeneity of inflammatory and cytokine networks in chronic plaque psoriasis. *PLoS One* **7**, e34594 (2012)10.1371/journal.pone.0034594).
59. W. R. Swindell, A. Johnston, J. J. Voorhees, J. T. Elder, J. E. Gudjonsson, Dissecting the psoriasis transcriptome: Inflammatory- and cytokine-driven gene expression in lesions from 163 patients. *BMC Genomics* **14**, 527 (2013).
60. G. M. O'Regan, A. Sandilands, W. H. McLean, A. D. Irvine, Filaggrin in atopic dermatitis. *J Allergy Clin Immunol* **122**, 689-693 (2008); published online EpubOct (10.1016/j.jaci.2008.08.002).
61. Y. Cai, C. Fleming, J. Yan, New insights of T cells in the pathogenesis of psoriasis. *Cell Mol Immunol* **9**, 302-309 (2012); published online EpubJul (10.1038/cmi.2012.15).
62. W. R. Swindell, A. Johnston, X. Xing, J. J. Voorhees, J. T. Elder, J. E. Gudjonsson, Modulation of Epidermal Transcription Circuits in Psoriasis: New Links between Inflammation and Hyperproliferation. *PLoS One* **8**, e79253 (2013)10.1371/journal.pone.0079253).
63. J. W. Schoggins, S. J. Wilson, M. Panis, M. Y. Murphy, C. T. Jones, P. Bieniasz, C. M. Rice, A diverse range of gene products are effectors of the type I interferon antiviral

response. *Nature* **472**, 481-485 (2011); published online EpubApr (10.1038/nature09907).
64. Y. Yao, L. Richman, C. Morehouse, M. de los Reyes, B. W. Higgs, A. Boutrin, B. White, A. Coyle, J. Krueger, P. A. Kiener, B. Jallal, Type I interferon: potential therapeutic target for psoriasis? *PLoS One* **3**, e2737 (2008)10.1371/journal.pone.0002737).
65. M. Suárez-Fariñas, S. J. Tintle, A. Shemer, A. Chiricozzi, K. Nograles, I. Cardinale, S. Duan, A. M. Bowcock, J. G. Krueger, E. Guttman-Yassky, Nonlesional atopic dermatitis skin is characterized by broad terminal differentiation defects and variable immune abnormalities. *J Allergy Clin Immunol* **127**, 954-964.e951-954 (2011); published online EpubApr (10.1016/j.jaci.2010.12.1124).
66. M. Suárez-Fariñas, K. Li, J. Fuentes-Duculan, K. Hayden, C. Brodmerkel, J. G. Krueger, Expanding the psoriasis disease profile: interrogation of the skin and serum of patients with moderate-to-severe psoriasis. *J Invest Dermatol* **132**, 2552-2564 (2012); published online EpubNov (10.1038/jid.2012.184).
67. W. Fujimoto, G. Nakanishi, J. Arata, A. M. Jetten, Differential expression of human cornifin alpha and beta in squamous differentiating epithelial tissues and several skin lesions. *J Invest Dermatol* **108**, 200-204 (1997); published online EpubFeb (
68. L. C. Tsoi, S. L. Spain, J. Knight, E. Ellinghaus, P. E. Stuart, F. Capon, J. Ding, Y. Li, T. Tejasvi, J. E. Gudjonsson, H. M. Kang, M. H. Allen, R. McManus, G. Novelli, L. Samuelsson, J. Schalkwijk, M. Ståhle, A. D. Burden, C. H. Smith, M. J. Cork, X. Estivill, A. M. Bowcock, G. G. Krueger, W. Weger, J. Worthington, R. Tazi-Ahnini, F. O. Nestle, A. Hayday, P. Hoffmann, J. Winkelmann, C. Wijmenga, C. Langford, S. Edkins, R. Andrews, H. Blackburn, A. Strange, G. Band, R. D. Pearson, D. Vukcevic, C. C. Spencer, P. Deloukas, U. Mrowietz, S. Schreiber, S. Weidinger, S. Koks, K. Kingo, T. Esko, A. Metspalu, H. W. Lim, J. J. Voorhees, M. Weichenthal, H. E. Wichmann, V. Chandran, C. F. Rosen, P. Rahman, D. D. Gladman, C. E. Griffiths, A. Reis, J. Kere, R. P. Nair, A. Franke, J. N. Barker, G. R. Abecasis, J. T. Elder, R. C. Trembath, C. A. S. o. P. (CASP), G. A. o. P. Consortium, P. A. G. Extension, W. T. C. C. C. 2, Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44**, 1341-1348 (2012); published online EpubDec (10.1038/ng.2467).
69. A. M. Lin, C. J. Rubin, R. Khandpur, J. Y. Wang, M. Riblett, S. Yalavarthi, E. C. Villanueva, P. Shah, M. J. Kaplan, A. T. Bruce, Mast cells and neutrophils release IL-17 through extracellular trap formation in psoriasis. *J Immunol* **187**, 490-500 (2011); published online EpubJul (10.4049/jimmunol.1100123).
70. J. G. Krueger, A. Bowcock, Psoriasis pathophysiology: current concepts of pathogenesis. *Ann Rheum Dis* **64 Suppl 2**, ii30-36 (2005); published online EpubMar (10.1136/ard.2004.031120).
71. M. Uribe-Herranz, L. H. Lian, K. M. Hooper, K. A. Milora, L. E. Jensen, IL-1R1 signaling facilitates Munro's microabscess formation in psoriasiform imiquimod-induced skin inflammation. *J Invest Dermatol* **133**, 1541-1549 (2013); published online EpubJun (10.1038/jid.2012.512).
72. R. M. Teles, T. G. Graeber, S. R. Krutzik, D. Montoya, M. Schenk, D. J. Lee, E. Komisopoulou, K. Kelly-Scumpia, R. Chun, S. S. Iyer, E. N. Sarno, T. H. Rea, M. Hewison, J. S. Adams, S. J. Popper, D. A. Relman, S. Stenger, B. R. Bloom, G. Cheng, R. L. Modlin, Type I Interferon Suppresses Type II Interferon-Triggered Human Anti-Mycobacterial Responses. *Science*, (2013); published online EpubFeb (10.1126/science.1233665).
73. S. J. Waddell, S. J. Popper, K. H. Rubins, M. J. Griffiths, P. O. Brown, M. Levin, D. A. Relman, Dissecting interferon-induced transcriptional programs in human peripheral blood cells. *PLoS One* **5**, e9753 (2010)10.1371/journal.pone.0009753).

74. M. A. Care, S. Barrans, L. Worrillow, A. Jack, D. R. Westhead, R. M. Tooze, A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PLoS One* **8**, e55895 (2013)10.1371/journal.pone.0055895).
75. Z. Yi, Z. Li, S. Yu, C. Yuan, W. Hong, Z. Wang, J. Cui, T. Shi, Y. Fang, Blood-based gene expression profiles models for classification of subsyndromal symptomatic depression and major depressive disorder. *PLoS One* **7**, e31283 (2012)10.1371/journal.pone.0031283).
76. D. V. Nguyen, D. M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* **18**, 1216-1226 (2002); published online EpubSep (
77. M. Kamsteeg, P. A. Jansen, I. M. van Vlijmen-Willems, P. E. van Erp, D. Rodijk-Olthuis, P. G. van der Valk, T. Feuth, P. L. Zeeuwen, J. Schalkwijk, Molecular diagnostics of psoriasis, atopic dermatitis, allergic contact dermatitis and irritant contact dermatitis. *Br J Dermatol* **162**, 568-578 (2010); published online EpubMar (10.1111/j.1365-2133.2009.09547.x).
78. G. Trinchieri, Type I interferon: friend or foe? *J Exp Med* **207**, 2053-2063 (2010); published online EpubSep (10.1084/jem.20101664).
79. B. Skurkovich, S. Skurkovich, Inhibition of IFN-gamma as a method of treatment of various autoimmune diseases, including skin diseases. *Ernst Schering Res Found Workshop*, 1-27 (2006).
80. P. D. Ling, M. K. Warren, S. N. Vogel, Antagonistic effect of interferon-beta on the interferon-gamma-induced expression of Ia antigen in murine macrophages. *J Immunol* **135**, 1857-1863 (1985); published online EpubSep (
81. R. Yoshida, H. W. Murray, C. F. Nathan, Agonist and antagonist effects of interferon alpha and beta on activation of human macrophages. Two classes of interferon gamma receptors and blockade of the high-affinity sites by interferon alpha or beta. *J Exp Med* **167**, 1171-1185 (1988); published online EpubMar (
82. M. Rayamajhi, J. Humann, K. Penheiter, K. Andreasen, L. L. Lenz, Induction of IFN-alphabeta enables Listeria monocytogenes to suppress macrophage activation by IFN-gamma. *J Exp Med* **207**, 327-337 (2010); published online EpubFeb (10.1084/jem.20091746).
83. S. Alazemi, M. A. Campos, Interferon-induced sarcoidosis. *Int J Clin Pract* **60**, 201-211 (2006); published online EpubFeb (10.1111/j.1742-1241.2005.00651.x).
84. A. Grassegger, R. Höpfl, Significance of the cytokine interferon gamma in clinical dermatology. *Clin Exp Dermatol* **29**, 584-588 (2004); published online EpubNov (10.1111/j.1365-2230.2004.01652.x).
85. M. Grewe, K. Gyufko, E. Schöpf, J. Krutmann, Lesional expression of interferon-gamma in atopic eczema. *Lancet* **343**, 25-26 (1994); published online EpubJan (
86. M. Grewe, S. Walther, K. Gyufko, W. Czech, E. Schöpf, J. Krutmann, Analysis of the cytokine pattern expressed in situ in inhalant allergen patch test reactions of atopic dermatitis patients. *J Invest Dermatol* **105**, 407-410 (1995); published online EpubSep (
87. Q. Hamid, M. Boguniewicz, D. Y. Leung, Differential in situ cytokine gene expression in acute versus chronic atopic dermatitis. *J Clin Invest* **94**, 870-876 (1994); published online EpubAug (10.1172/JCI117408).
88. M. V. Han, C. M. Zmasek, phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**, 356 (2009)10.1186/1471-2105-10-356).
89. J. Bingham, S. Sudarsanam, Visualizing large hierarchical clusters in hyperbolic space. *Bioinformatics* **16**, 660-661 (2000); published online EpubJul (

90. R. P. Nair, K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, B. J. Feng, A. Ruether, S. Schreiber, M. Weichenthal, D. Gladman, P. Rahman, S. J. Schrodi, S. Prahalad, S. L. Guthery, J. Fischer, W. Liao, P. Y. Kwok, A. Menter, G. M. Lathrop, C. A. Wise, A. B. Begovich, J. J. Voorhees, J. T. Elder, G. G. Krueger, A. M. Bowcock, G. R. Abecasis, C. A. S. o. Psoriasis, Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat Genet* **41**, 199-204 (2009); published online EpubFeb (10.1038/ng.311).
91. W. R. Swindell, A. Johnston, S. Carbajal, G. Han, C. Wohn, J. Lu, X. Xing, R. P. Nair, J. J. Voorhees, J. T. Elder, X. J. Wang, S. Sano, E. P. Prens, J. DiGiovanni, M. R. Pittelkow, N. L. Ward, J. E. Gudjonsson, Genome-wide expression profiling of five mouse models identifies similarities and differences with human psoriasis. *PLoS One* **6**, e18266 (2011)10.1371/journal.pone.0018266).
92. D. Montoya, D. Cruz, R. M. Teles, D. J. Lee, M. T. Ochoa, S. R. Krutzik, R. Chun, M. Schenk, X. Zhang, B. G. Ferguson, A. E. Burdick, E. N. Sarno, T. H. Rea, M. Hewison, J. S. Adams, G. Cheng, R. L. Modlin, Divergence of macrophage phagocytic and antimicrobial programs in leprosy. *Cell Host Microbe* **6**, 343-353 (2009); published online EpubOct (10.1016/j.chom.2009.09.002).
93. D. J. Lee, H. Li, M. T. Ochoa, M. Tanaka, R. J. Carbone, R. Damoiseaux, A. Burdick, E. N. Sarno, T. H. Rea, R. L. Modlin, Integrated pathways for neutrophil recruitment and inflammation in leprosy. *J Infect Dis* **201**, 558-569 (2010); published online EpubFeb (10.1086/650318).
94. T. L. Humphreys, L. Li, X. Li, D. M. Janowicz, K. R. Fortney, Q. Zhao, W. Li, J. McClintick, B. P. Katz, D. S. Wilkes, H. J. Edenberg, S. M. Spinola, Dysregulated immune profiles for skin and dendritic cells are associated with increased host susceptibility to Haemophilus ducreyi infection in human volunteers. *Infect Immun* **75**, 5686-5697 (2007); published online EpubDec (10.1128/IAI.00777-07).
95. M. B. Pedersen, L. Skov, T. Menné, J. D. Johansen, J. Olsen, Gene expression time course in the human skin during elicitation of allergic contact dermatitis. *J Invest Dermatol* **127**, 2585-2595 (2007); published online EpubNov (10.1038/sj.jid.5700902).
96. A. Clemmensen, K. E. Andersen, O. Clemmensen, Q. Tan, T. K. Petersen, T. A. Kruse, M. Thomassen, Genome-wide expression analysis of human in vivo irritated epidermis: differential profiles induced by sodium lauryl sulfate and nonanoic acid. *J Invest Dermatol* **130**, 2201-2210 (2010); published online EpubSep (10.1038/jid.2010.102).
97. E. Guttman-Yassky, M. Suárez-Fariñas, A. Chiricozzi, K. E. Nograles, A. Shemer, J. Fuentes-Duculan, I. Cardinale, P. Lin, R. Bergman, A. M. Bowcock, J. G. Krueger, Broad defects in epidermal cornification in atopic dermatitis identified through genomic analysis. *J Allergy Clin Immunol* **124**, 1235-1244.e1258 (2009); published online EpubDec (10.1016/j.jaci.2009.09.031).
98. A. I. Riker, S. A. Enkemann, O. Fodstad, S. Liu, S. Ren, C. Morris, Y. Xi, P. Howell, B. Metge, R. S. Samant, L. A. Shevde, W. Li, S. Eschrich, A. Daud, J. Ju, J. Matta, The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Med Genomics* **1**, 13 (2008)10.1186/1755-8794-1-13).
99. R. van Doorn, M. S. van Kester, R. Dijkman, M. H. Vermeer, A. A. Mulder, K. Szuhai, J. Knijnenburg, J. M. Boer, R. Willemze, C. P. Tensen, Oncogenomic analysis of mycosis fungoides reveals major differences with Sezary syndrome. *Blood* **113**, 127-136 (2009); published online EpubJan (10.1182/blood-2008-04-153031).
100. K. Nuutila, A. Siltanen, M. Peura, J. Bizik, I. Kaartinen, H. Kuokkanen, T. Nieminen, A. Harjula, P. Aarnio, J. Vuola, E. Kankuri, Human skin transcriptome during superficial

cutaneous wound healing. *Wound Repair Regen* **20**, 830-839 (2012); published online Epub2012 Nov-Dec (10.1111/j.1524-475X.2012.00831.x).
101. M. A. Judson, R. M. Marchell, M. Mascelli, A. Piantone, E. S. Barnathan, K. J. Petty, D. Chen, H. Fan, H. Grund, K. Ma, F. Baribaud, C. Brodmerkel, Molecular profiling and gene expression analysis in cutaneous sarcoidosis: the role of interleukin-12, interleukin-23, and the T-helper 17 pathway. *J Am Acad Dermatol* **66**, 901-910, 910.e901-902 (2012); published online EpubJun (10.1016/j.jaad.2011.06.017).
102. W. H. Chung, S. I. Hung, J. Y. Yang, S. C. Su, S. P. Huang, C. Y. Wei, S. W. Chin, C. C. Chiou, S. C. Chu, H. C. Ho, C. H. Yang, C. F. Lu, J. Y. Wu, Y. D. Liao, Y. T. Chen, Granulysin is a key mediator for disseminated keratinocyte death in Stevens-Johnson syndrome and toxic epidermal necrolysis. *Nat Med* **14**, 1343-1350 (2008); published online EpubDec (10.1038/nm.1884).
103. J. Bigler, H. A. Rand, K. Kerkof, M. Timour, C. B. Russell, Cross-study homogeneity of psoriasis gene expression in skin across a large expression range. *PLoS One* **8**, e52242 (2013)10.1371/journal.pone.0052242).
104. C. D. Matheson, K. K. Vernon, A. Lahti, R. Fratpietro, M. Spigelman, S. Gibson, C. L. Greenblatt, H. D. Donoghue, B. Zissu, Molecular exploration of the first-century Tomb of the Shroud in Akeldama, Jerusalem. *PLoS One* **4**, e8319 (2009)10.1371/journal.pone.0008319).
105. W. H. Organization. (2014), vol. 2014.
106. L. C. Rodrigues, D. N. j. Lockwood, Leprosy now: epidemiology, progress, challenges, and research gaps. *Lancet Infect Dis* **11**, 464-470 (2011); published online EpubJun (10.1016/S1473-3099(11)70006-8).
107. J. Rafferty, Curing the stigma of leprosy. *Lepr Rev* **76**, 119-126 (2005); published online EpubJun (
108. R. L. Modlin, The innate immune response in leprosy. *Curr Opin Immunol* **22**, 48-54 (2010); published online EpubFeb (10.1016/j.coi.2009.12.001).
109. S. Kamath, S. A. Vaccaro, T. H. Rea, M. T. Ochoa, Recognizing and managing the immunologic reactions in leprosy. *J Am Acad Dermatol*, (2014); published online EpubApr (10.1016/j.jaad.2014.03.034).
110. R. M. Teles, R. B. Teles, T. P. Amadeu, D. F. Moura, L. Mendonça-Lima, H. Ferreira, I. M. Santos, J. A. Nery, E. N. Sarno, E. P. Sampaio, High matrix metalloproteinase production correlates with immune activation and leukocyte migration in leprosy reactional lesions. *Infect Immun* **78**, 1012-1021 (2010); published online EpubMar (10.1128/IAI.00896-09).
111. M. Shibuya, Vascular endothelial growth factor-dependent and -independent regulation of angiogenesis. *BMB Rep* **41**, 278-286 (2008); published online EpubApr (
112. M. Schiller, D. Javelaud, A. Mauviel, TGF-beta-induced SMAD signaling and gene regulation: consequences for extracellular matrix remodeling and wound healing. *J Dermatol Sci* **35**, 83-92 (2004); published online EpubAug (10.1016/j.jdermsci.2003.12.006).
113. P. A. Suwanabol, K. C. Kent, B. Liu, TGF-β and restenosis revisited: a Smad link. *J Surg Res* **167**, 287-297 (2011); published online EpubMay (10.1016/j.jss.2010.12.020).
114. M. P. Crippa, Urokinase-type plasminogen activator. *Int J Biochem Cell Biol* **39**, 690-694 (2007)10.1016/j.biocel.2006.10.008).
115. S. S. Bhandarkar, C. Cohen, M. Kuruvila, T. H. Rea, J. B. Mackelfresh, D. J. Lee, R. L. Modlin, J. L. Arbiser, Angiogenesis in cutaneous lesions of leprosy: implications for treatment. *Arch Dermatol* **143**, 1527-1529 (2007); published online EpubDec (10.1001/archderm.143.12.1527).

116. C. T. Soares, P. S. Rosa, A. P. Trombone, L. R. Fachin, C. C. Ghidella, S. Ura, J. A. Barreto, A. e. F. Belone, Angiogenesis and lymphangiogenesis in the spectrum of leprosy and its reactional forms. *PLoS One* **8**, e74651 (2013)10.1371/journal.pone.0074651).
117. C. Cursiefen, L. Chen, L. P. Borges, D. Jackson, J. Cao, C. Radziejewski, P. A. D'Amore, M. R. Dana, S. J. Wiegand, J. W. Streilein, VEGF-A stimulates lymphangiogenesis and hemangiogenesis in inflammatory neovascularization via macrophage recruitment. *J Clin Invest* **113**, 1040-1050 (2004); published online EpubApr (10.1172/JCI20465).
118. I. P. Kahawita, D. N. Lockwood, Towards understanding the pathology of erythema nodosum leprosum. *Trans R Soc Trop Med Hyg* **102**, 329-337 (2008); published online EpubApr (10.1016/j.trstmh.2008.01.004).
119. J. Touw, E. M. Langendijk, G. L. Stoner, A. Belehu, Humoral immunity in leprosy: immunoglobulin G and M antibody responses to Mycobacterium leprae in relation to various disease patterns. *Infect Immun* **36**, 885-892 (1982); published online EpubJun (
120. G. U. Gangenahalli, V. K. Singh, Y. K. Verma, P. Gupta, R. K. Sharma, R. Chandra, P. M. Luthra, Hematopoietic stem cell antigen CD34: role in adhesion or homing. *Stem Cells Dev* **15**, 305-313 (2006); published online EpubJun (10.1089/scd.2006.15.305).
121. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009); published online EpubMay (10.1093/bioinformatics/btp120).
122. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010); published online EpubMay (10.1038/nbt.1621).
123. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008); published online EpubJul (10.1038/nmeth.1226).
124. A. S, P. PT, H. W, HTSeq — A Python framework to work with high-throughput sequencing data. *Under review*, (2014).
125. J. S. Cumbie, J. A. Kimbrel, Y. Di, D. W. Schafer, L. J. Wilhelm, S. E. Fox, C. M. Sullivan, A. D. Curzon, J. C. Carrington, T. C. Mockler, J. H. Chang, GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. *PLoS One* **6**, e25279 (2011)10.1371/journal.pone.0025279).
126. D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, W. J. Kent, The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-770 (2014); published online EpubJan (10.1093/nar/gkt1168).
127. M. Kayahara, K. Funaki, H. Tajima, H. Takamura, I. Ninomiya, H. Kitagawa, T. Ohta, Surgical implication of micrometastasis for pancreatic cancer. *Pancreas* **39**, 884-888 (2010); published online EpubAug (10.1097/MPA.0b013e3181ce6daa).
128. W. S. Liang, D. W. Craig, J. Carpten, M. J. Borad, M. J. Demeure, G. J. Weiss, T. Izatt, S. Sinari, A. Christoforides, J. Aldrich, A. Kurdoglu, M. Barrett, L. Phillips, H. Benson, W. Tembe, E. Braggio, J. A. Kiefer, C. Legendre, R. Posner, G. H. Hostetter, A. Baker, J. B. Egan, H. Han, D. Lake, E. C. Stites, R. K. Ramanathan, R. Fonseca, A. K. Stewart, D. Von Hoff, Genome-wide characterization of pancreatic adenocarcinoma patients using next generation sequencing. *PLoS One* **7**, e43192 (2012)10.1371/journal.pone.0043192).

129. M. Yu, D. T. Ting, S. L. Stott, B. S. Wittner, F. Ozsolak, S. Paul, J. C. Ciciliano, M. E. Smas, D. Winokur, A. J. Gilman, M. J. Ulman, K. Xega, G. Contino, B. Alagesan, B. W. Brannigan, P. M. Milos, D. P. Ryan, L. V. Sequist, N. Bardeesy, S. Ramaswamy, M. Toner, S. Maheswaran, D. A. Haber, RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. *Nature* **487**, 510-513 (2012); published online EpubJul (10.1038/nature11217).
130. G. Young, K. Wang, J. He, G. Otto, M. Hawryluk, Z. Zwirco, T. Brennan, M. Nahas, A. Donahue, R. Yelensky, D. Lipson, C. E. Sheehan, A. B. Boguniewicz, P. J. Stephens, V. A. Miller, J. S. Ross, Clinical next-generation sequencing successfully applied to fine-needle aspirations of pulmonary and pancreatic neoplasms. *Cancer Cytopathol* **121**, 688-694 (2013); published online EpubDec (10.1002/cncy.21338).
131. M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, P. Pavlidis, Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res* **33**, 5914-5923 (2005)10.1093/nar/gki890).
132. F. Liu, W. P. Kuo, T. K. Jenssen, E. Hovig, Performance comparison of multiple microarray platforms for gene expression profiling. *Methods Mol Biol* **802**, 141-155 (2012)10.1007/978-1-61779-400-1_10).
133. W. P. Kuo, T. K. Jenssen, A. J. Butte, L. Ohno-Machado, I. S. Kohane, Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405-412 (2002); published online EpubMar (
134. S. S. Chavan, M. A. Bauer, E. A. Peterson, C. J. Heuck, D. J. Johann, Towards the integration, annotation and association of historical microarray experiments with RNA-seq. *BMC Bioinformatics* **14 Suppl 14**, S4 (2013)10.1186/1471-2105-14-S14-S4).
135. X. Xu, Y. Zhang, J. Williams, E. Antoniou, W. R. McCombie, S. Wu, W. Zhu, N. O. Davidson, P. Denoya, E. Li, Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC Bioinformatics* **14 Suppl 9**, S1 (2013)10.1186/1471-2105-14-S9-S1).