ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression

Jesse M Engreitz[1,2], Rong Chen[1,3], Alexander A Morgan[1,3,4], Joel T Dudley[1,3,4], Rohan Mallelwar[5], and Atul J Butte[1,3,*]

[1]Division of Systems Medicine, Department of Pediatrics, [2]Department of Bioengineering, [3]Lucile Packard Children's Hospital, and [4]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA, USA, [5]Optra Systems Pvt. Ltd, 1, Dnyanesh, CTS No. 1179/3, Modern College Road, Shivajinagar, Pune, 411 005, India

SUPPLEMENTARY METHODS

ProfileChaser indexes and searches GEO DataSets using a combination of previously developed techniques for dimension reduction, data representation, and similarity measure (Engreitz et al. 2010a, Engreitz et al. 2010b). The following section describes our analytical pipeline for processing microarray experiments from GEO.

Data processing. For GEO DataSets, we matched probe identifiers to NCBI identifiers using AILUN (Chen et al. 2007). For species other than *H. sapiens*, we mapped genes to their unique human homologs with Homologene, discarding genes with multiple matches. To normalize expression values across datasets and platforms, we examined GEO annotations and value ranges for each dataset, and converted to log space as needed. We aggregated probes to genes using the fixed-effects meta-estimate, weighting the contribution of each probe by its variance.

Dimension reduction. Previously we applied independent component analysis to a compendium of 10,000 microarrays to identify *fundamental components* of human gene expression (Engreitz et al. 2010a). These 423 fundamental components represented coherent, functionally-relevant transcriptional programs that together spanned the space of human gene expression sampled in our compendium. To improve the speed and robustness of ProfileChaser, we projected each GEO microarray into this reduced feature-space using:

$$A = S^T X ,$$

where A is the reduced representation of the microarray experiment (423 features × thousands of profiles), S is the component matrix (thousands of genes × 423 features), and X is the original data in gene-space (thousands of genes × thousands of profiles). We found that this method, resulting in an approximately 50-fold reduction in the dimensionality of the data, yielded superior performance for comparing differential expression profiles, even across species and platforms (Engreitz et al. 2010b).

Data representation. ProfileChaser aims to index differential expression comparisons in GEO. To generate these profiles, we used the manually curated experimental variables defined in GEO DataSets to compare sets of microarrays. For each comparison, we created a *differential expression (DE) profile* by calculating the fold-change for each of the 423 fundamental components or features. In addition to fold-change, we calculated the probability that each fundamental component was differentially expressed using the empirical Bayes moderated *t*-statistic, implemented in the *limma* R package (Smyth 2004). *P*-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

Similarity measure. To compare DE profiles (vectors containing 423 elements), we use a weighted Pearson's correlation coefficient that considers the correct empirical Bayes *p*-value. Weights for the correlation are calculated by

$$w_i = \left[-\log(p_{i1}p_{i2})\right]^{1/2},$$

where p_{ij} is the corrected *p*-value for feature i in experiment j. Intuitively, features that are consistently differentially expressed in both DE profiles are given higher weights. When querying ProfileChaser, we calculate false discovery rate (FDR) for each retrieved result based on a null distribution of correlation coefficients between all 14,875 experimental comparisons. This FDR is likely an underestimate, since many of these experiments are in fact related to one another.

Identifying significant genes. To aid in identifying individual genes that contribute to this comparison, we also created DE profiles in gene-space for all GEO DataSet comparisons. We create scatterplots to show the global similarities and differences in expression between two DE profiles. The axes of these scatterplots represent the log_2 difference in expression between two conditions. The size of the point for gene i is directly proportional to the gene's contribution to the weighted correlation coefficient:

$$Area \propto w_i(x_{i1} - m_1)(x_{j2} - m_2),$$

where x_{ij} is the differential expression of gene i in profile j and m_j is the weighted mean of genes in profile j. Thus the largest points in the scatterplot represent genes that add positive contributions to the correlation coefficient (*i.e.*, genes that are differentially expressed in the same direction in both DE profiles).

REFERENCES

Chen,R. et al. (2007) AILUN: reannotating gene expression data automatically. *BMC Nat Methods*, **4**(11), 548.

Engreitz,J.M. et al. (2010a) Independent component analysis: Mining microarray data for fundamental human gene modules. *J Biomed Inform*, **43**, 932-944.

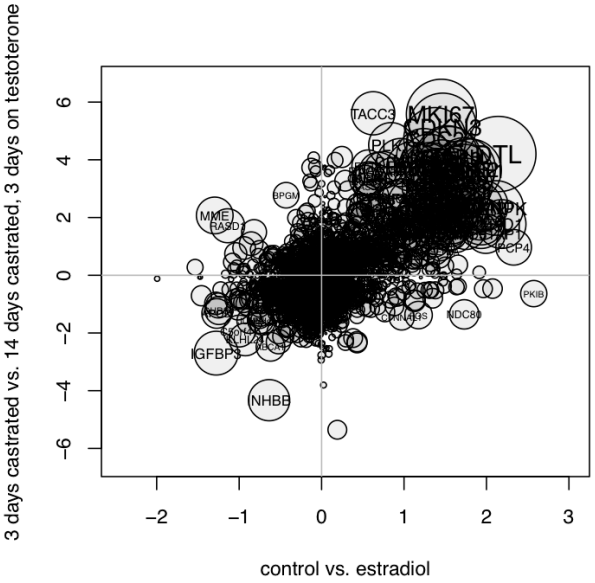
Engreitz,J.M. et al. (2010b) Content-based microarray search using differential expression profiles. *BMC Bioinformatics*, **11**(1), 603.

Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3:**Article 3.

SUPPLEMENTARY FIGURES

Rank	GEO	Title	Organism	Subset 1 vs. Subset 2	Factor	Score	q-value
1	GDS2324	Low concentrations of 17beta-estradiol effect on breast cancer cell line	*Homo sapiens*	0 pM vs. 100 pM	dose	0.8995	0.0007
2	GDS3285	Estrogen effect on breast cancer cell line: time course	*Homo sapiens*	6 h vs. 12 h	time	0.8778	0.0008
3	GDS1821	Muscle cell survival mediated by transcriptional coactivator p300	*Mus musculus*	24 h vs. 0 h	time	0.8429	0.0009
4	GDS2323	Estrogen-starved breast cancer cell line: time course	*Homo sapiens*	2 d vs. 0 d	time	0.8670	0.0009
5	GDS2324	Low concentrations of 17beta-estradiol effect on breast cancer cell line	*Homo sapiens*	0 pM vs. 30 pM	dose	0.8634	0.0009
6	GDS2324	Low concentrations of 17beta-estradiol effect on breast cancer cell line	*Homo sapiens*	0 pM vs. 60 pM	dose	0.8687	0.0009
7	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	4 h vs. 16 h	time	0.8434	0.0009
8	GDS1409	cAMP/protein kinaseA effect on cell-cycle regulation: timecourse	*Mus musculus*	24 h vs. 6 h	time	0.8267	0.0010
9	GDS2324	Low concentrations of 17beta-estradiol effect on breast cancer cell line	*Homo sapiens*	10 pM vs. 100 pM	dose	0.8352	0.0010
10	GDS2562	Prostate response to castration and subsequent hormone replacement	*Mus musculus*	3 days castrated vs. 14 days castrated, 3 days on testoterone	protocol	0.8341	0.0010
11	GDS2854	Myogenic transcription factor MyoD mutant expression effect on embryonic fibroblast: time course	*Mus musculus*	24 h vs. 12 h	time	0.8371	0.0010
12	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	1 h vs. 24 h	time	0.8354	0.0010
13	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	2 h vs. 16 h	time	0.8303	0.0010
14	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	2 h vs. 24 h	time	0.8356	0.0010
15	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	4 h vs. 10 h	time	0.8259	0.0010
16	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	4 h vs. 12 h	time	0.8308	0.0010
17	GDS2323	Estrogen-starved breast cancer cell line: time course	*Homo sapiens*	1 d vs. 0 d	time	0.8013	0.0011
18	GDS2367	Tamoxifen effect on breast cancer cell line expressing estrogen receptor alpha and beta	*Homo sapiens*	vehicle vs. tamoxifen	agent	0.8021	0.0011
19	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	0 h vs. 24 h	time	0.8092	0.0011
20	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	8 h vs. 24 h	time	0.8011	0.0011
21	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	*Mus musculus*	1 h vs. 16 h	time	0.8061	0.0011
22	GDS3482	X-linked inhibitor of apoptosis XIAP depletion effect on a colorectal cancer cell line	*Homo sapiens*	early passage vs. late passage	other	0.8178	0.0011
23	GDS568	Erythroid differentiation: G1E model	*Mus musculus*	30 h post-estradiol vs. 14 h post-estradiol	time	0.8182	0.0011
24	GDS911	Growth-arrested cell: serum deprivation and contact inhibition growth-arrest comparison	*Homo sapiens*	serum deprivation vs. asynchronous	growth protocol	0.8163	0.0011
25	GDS1549	Estrogen effect on estrogen receptor alpha positive breast cancer cell lines	*Homo sapiens*	control vs. estradiol	agent	0.7843	0.0012
26	GDS1873	Antiestrogen and aromatase inhibitor effect on breast cancer cells	*Homo sapiens*	control vs. hormone treatment	other	0.7764	0.0012
27	GDS2024	Lung immune response to Nippostrongylus brasiliensis infection: time course	*Mus musculus*	8 dpi vs. 3 dpi	time	0.7808	0.0012
28	GDS2324	Low concentrations of 17beta-estradiol effect on breast cancer cell line	*Homo sapiens*	30 pM vs. 100 pM	dose	0.7815	0.0012
29	GDS2367	Tamoxifen effect on breast cancer cell line expressing estrogen receptor alpha and beta	*Homo sapiens*	vehicle vs. estradiol	agent	0.7967	0.0012
30	GDS2965	Embryonic heart response to retinoic acid and dioxin: time course	*Danio rerio*	TCDD vs. control for TCDD	agent	0.7725	0.0012

Supplementary Figure S1. Top thirty search results for GDS3315 (control *vs*. estradiol).



Symbol	Description	Profile 1		Profile 2		Weight	Contribution
		Fold-change	*P*-value	Fold-change	*P*-value		
DTL	denticleless homolog (Drosophila)	2.1434	0.0000	4.2003	0.0000	2.5928	0.008518
MKI67	antigen identified by monoclonal antibody Ki-67	1.4525	0.0000	5.5799	0.0000	2.5252	0.007354
CDKN3	cyclin-dependent kinase inhibitor 3	1.4690	0.0000	5.1452	0.0000	2.4949	0.006787
TCF19	transcription factor 19	1.5498	0.0000	4.0422	0.0000	2.7139	0.006158
NCAPH	non-SMC condensin I complex	1.5243	0.0000	4.0983	0.0000	2.6024	0.005884
CDC2	cell division cycle 2	1.3751	0.0000	4.7299	0.0000	2.4992	0.005847
UHRF1	ubiquitin-like with PHD and ring finger domains 1	1.8028	0.0000	3.6458	0.0001	2.4209	0.005801
POLE2	polymerase (DNA directed)	1.7328	0.0000	3.6927	0.0000	2.4792	0.005776
ANLN	anillin	1.7215	0.0000	3.4996	0.0000	2.4587	0.005400
TTK	TTK protein kinase	1.6758	0.0000	3.6596	0.0001	2.2192	0.004952

Supplementary Figure S2. Comparison of differential expression profiles from GDS3315 (Profile 1: control vs. estradiol) and GDS2562 (Profile 2: 3 days castrated vs. 14 days castradated, 3 days on testosterone). Scatterplot displays the log_2 fold-change for genes in each comparison. The area of each point is proportional to each gene's contribution to the final correlation coefficient (see Supplementary Methods). Top genes include many proliferation markers, including MKI67, the locus that codes for Ki-67.

β

Rank	GEO	Title	Organism	Subset 1 vs. Subset 2	Factor	Score	q-value
1	GDS2617	Tumorigenic breast cancer cells (HG-U133A)	*Homo sapiens*	non-tumorigenic cancer cell vs. tumorigenic cancer cell	disease state	0.7438	0.0015
2	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 2030	individual	0.5435	0.0095
3	GDS1329	Molecular apocrine breast tumors	*Homo sapiens*	luminal tumor vs. basal tumor	disease state	0.5418	0.0096
4	GDS1925	Estrogen receptor alpha positive breast cancer cells response to hyperactivation of MAPK pathway	*Homo sapiens*	long-term E2 independent growth vs. EGFR	cell line	0.5312	0.0102
5	GDS1925	Estrogen receptor alpha positive breast cancer cells response to hyperactivation of MAPK pathway	*Homo sapiens*	long-term E2 independent growth vs. Raf-1	cell line	0.5299	0.0103
6	GDS2250	Basal-like breast cancer tumors	*Homo sapiens*	non-basal-like cancer vs. basal-like cancer	disease state	0.5107	0.0120
7	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1243	individual	0.5068	0.0126
8	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1688 vs. patient 2030	individual	0.5019	0.0128
9	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 2512	individual	0.4912	0.0146
10	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1677	individual	0.4821	0.0159
11	GDS2516	Interferons effect on endothelial cells	*Homo sapiens*	control vs. interferon alpha	agent	0.4743	0.0163
12	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1353	individual	0.4773	0.0163
13	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1687	individual	0.4732	0.0166
14	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 783	individual	0.4723	0.0167
15	GDS1329	Molecular apocrine breast tumors	*Homo sapiens*	apocrine tumor vs. basal tumor	disease state	0.4710	0.0169
16	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 275	individual	0.4660	0.0176
17	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1993	individual	0.4580	0.0186
18	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1688 vs. patient 1687	individual	0.4590	0.0186
19	GDS3210	Airway epithelial cells response to Sendai virus infection in vitro	*Mus musculus*	control vs. Sendai virus	infection	0.4590	0.0186
20	GDS2958	Tumor suppressor PTEN depletion effect on various cell lines	*Homo sapiens*	SKBR-3 vs. HCC827	cell line	0.4566	0.0190
21	GDS2958	Tumor suppressor PTEN depletion effect on various cell lines	*Homo sapiens*	mammary adenocarcinoma vs. non-small cell lung carcinoma	cell type	0.4566	0.0190
22	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1688 vs. patient 1353	individual	0.4569	0.0190
23	GDS2341	Type I and Type II interferons effect on lung epithelial cell line: time course	*Homo sapiens*	untreated vs. Type I IFN	agent	0.4548	0.0196
24	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 2791	individual	0.4515	0.0203
25	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1988	individual	0.4477	0.0210
26	GDS1381	Carboplatin sensitive and resistant ovarian carcinoma	*Homo sapiens*	patient 5 vs. patient 2	individual	0.4473	0.0211
27	GDS1381	Carboplatin sensitive and resistant ovarian carcinoma	*Homo sapiens*	patient 6 vs. patient 2	individual	0.4458	0.0214
28	GDS2414	Decidual stromal cell response to trophoblast conditioned medium: time course	*Homo sapiens*	control vs. trophoblast conditioned medium	agent	0.4441	0.0215
29	GDS3155	Dasatinib resistant and sensitive prostatic cancer cell lines	*Homo sapiens*	dasatinib resistant vs. dasatinib sensitive	other	0.4442	0.0215
30	GDS3017	Cervical cancer response to chemoradiotherapy	*Homo sapiens*	patient 1676 vs. patient 1690	individual	0.4439	0.0216

Supplementary Figure S3. Top thirty search results for GDS2618 (tumorigenic cancer cells vs. non-tumorigenic cancer cells). GDS2617, which represents the same samples run on a companion platform (HG-U133B), is identified as the top hit, despite the fact that HG-U133A and HG-U133B measure only 4431 of the same genes (out of 13,780 and 10,044 genes, respectively). This search identifies dasatinib as a potential inhibitor of breast cancer stem cells (see Result 29).

Sample Subsets				
Samples	Factors			Title
	time	agent	growth protocol	
GSM25978	0 h	control	no selection	A404 SMC Differentiation Control Replicate 1
GSM25979	0 h	control	no selection	A404 SMC Differentiation Control Replicate 2
GSM26006	0 h	control	no selection	A404 SMC Differentiation Control Replicate 3
GSM26007	0 h	control	no selection	A404 SMC Differentiation Control Replicate 4
GSM26008	0 h	control	no selection	A404 SMC Differentiation Control Replicate 5
GSM26009	0 h	control	no selection	A404 SMC Differentiation Control Replicate 6
GSM26010	48 h	retinoic acid	no selection	A404 SMC Differentiation RA48 Replicate 1
GSM26011	48 h	retinoic acid	no selection	A404 SMC Differentiation RA48 Replicate 2
GSM26012	48 h	retinoic acid	no selection	A404 SMC Differentiation RA48 Replicate 3
GSM26013	48 h	retinoic acid	no selection	A404 SMC Differentiation RA48 Replicate 4
GSM26014	48 h	retinoic acid	no selection	A404 SMC Differentiation RA48 Replicate 5
GSM26015	96 h	retinoic acid	no selection	A404 SMC Differentiation RA96 Replicate 1
GSM26016	96 h	retinoic acid	no selection	A404 SMC Differentiation RA96 Replicate 2
GSM26017	96 h	retinoic acid	no selection	A404 SMC Differentiation RA96 Replicate 3
GSM26018	96 h	retinoic acid	no selection	A404 SMC Differentiation RA96 Replicate 4
GSM26019	96 h	retinoic acid	no selection	A404 SMC Differentiation RA96 Replicate 5
GSM26020	96 h	retinoic acid	no selection	A404 SMC Differentiation RA96 Replicate 6
GSM26021	96 h	retinoic acid	puromycin resistance	A404 SMC Differentiation Puromycin Replicate 1
GSM26022	96 h	retinoic acid	puromycin resistance	A404 SMC Differentiation Puromycin Replicate 2
GSM26023	96 h	retinoic acid	puromycin resistance	A404 SMC Differentiation Puromycin Replicate 3
GSM26024	96 h	retinoic acid	puromycin resistance	A404 SMC Differentiation Puromycin Replicate 4
GSM26025	96 h	retinoic acid	puromycin resistance	A404 SMC Differentiation Puromycin Replicate 5
GSM26026	96 h	retinoic acid	puromycin resistance	A404 SMC Differentiation Puromycin Replicate 6

Supplementary Figure S4. Example of a GEO Dataset with a multifactorial experimental design (GDS799, http://http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS799). ProfileChaser compares all arrays annotated in each subset; for example, we generate a differential expression profile for all arrays labeled with "no selection" compared to all arrays labeled with "puromycin resistance." However, this comparison is partially confounded in that the "no selection" subset includes samples generated at multiple time points and with differing application of retinoic acid. The results page of the web server indicates the additional factors in each experimental design, but all results should be interpreted carefully through inspection of the experimental design defined by GEO and the original study references. For more information, see the tutorial on the ProfileChaser web site.