Supplementary Information

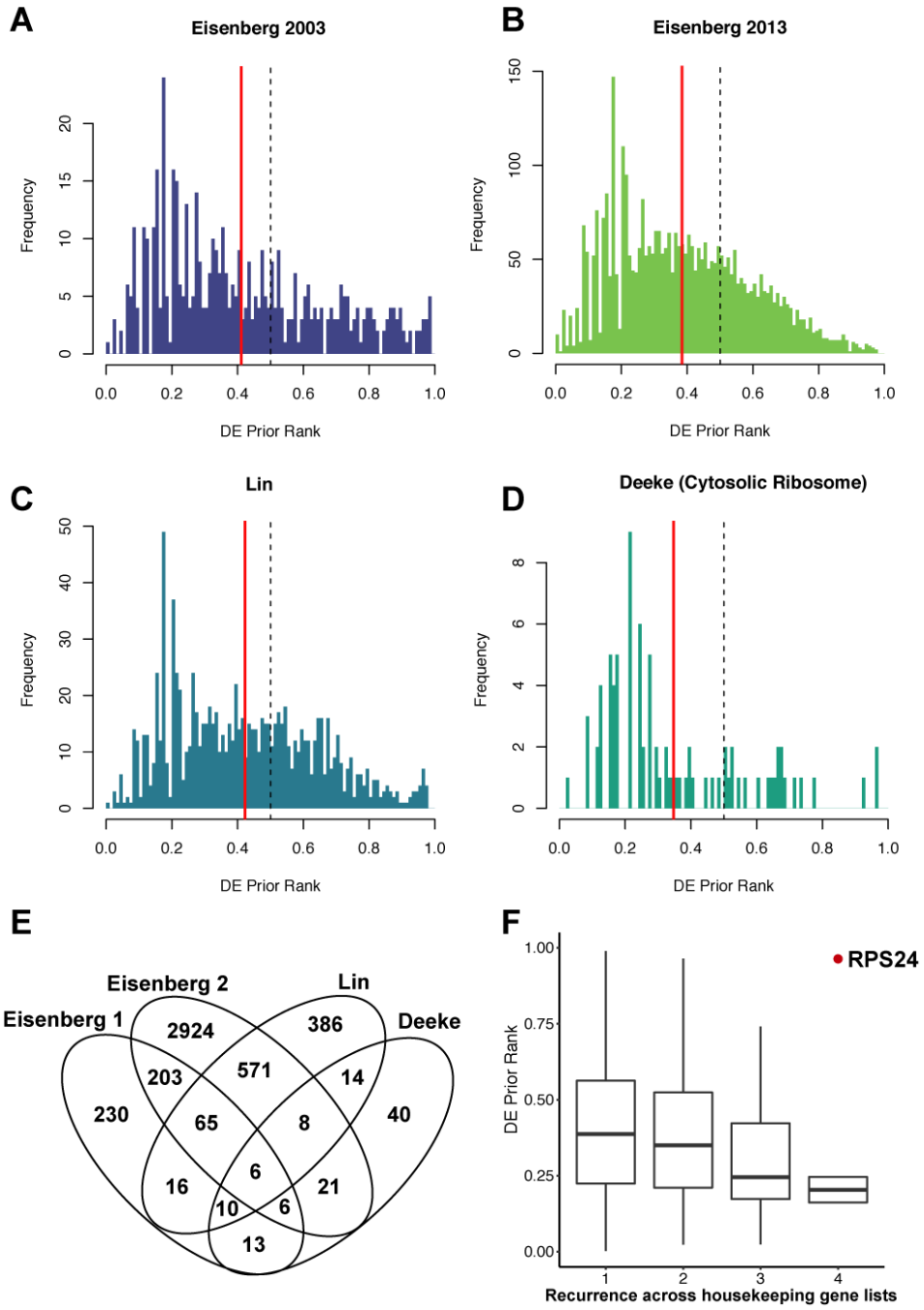
Predictability of Human Gene Expression Analysis

Megan Crow, Nathaniel Lim, Sara Ballouz, Paul Pavlidis, Jesse Gillis*
*Corresponding author: jgillis@cshl.edu

Note that supplementary datasets S1 and S2 are provided separately

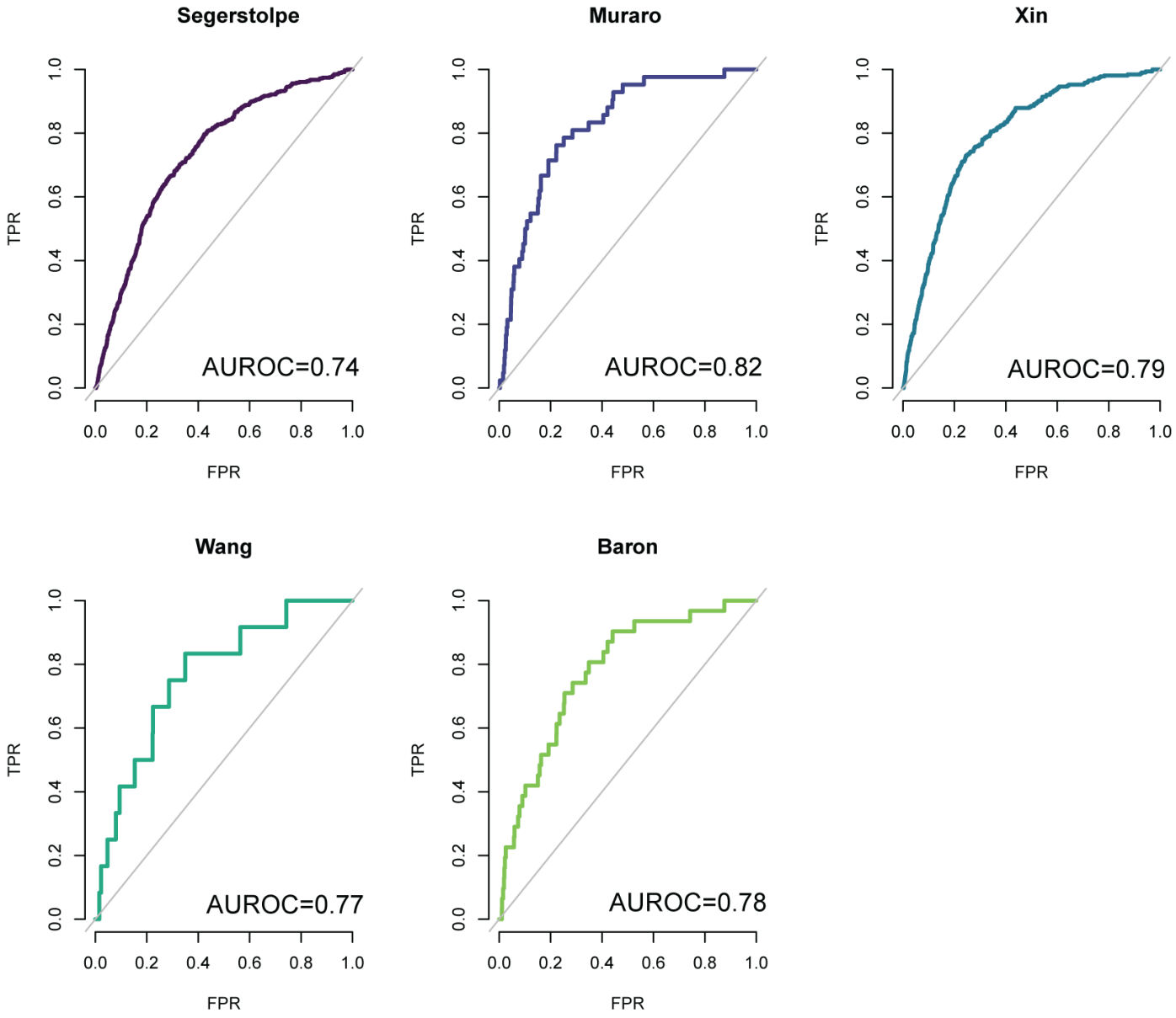
www.pnas.org/cgi/doi/10.1073/pnas.1802973116

Supplementary Figure 1 – Housekeeping gene sets are infrequently differentially expressed



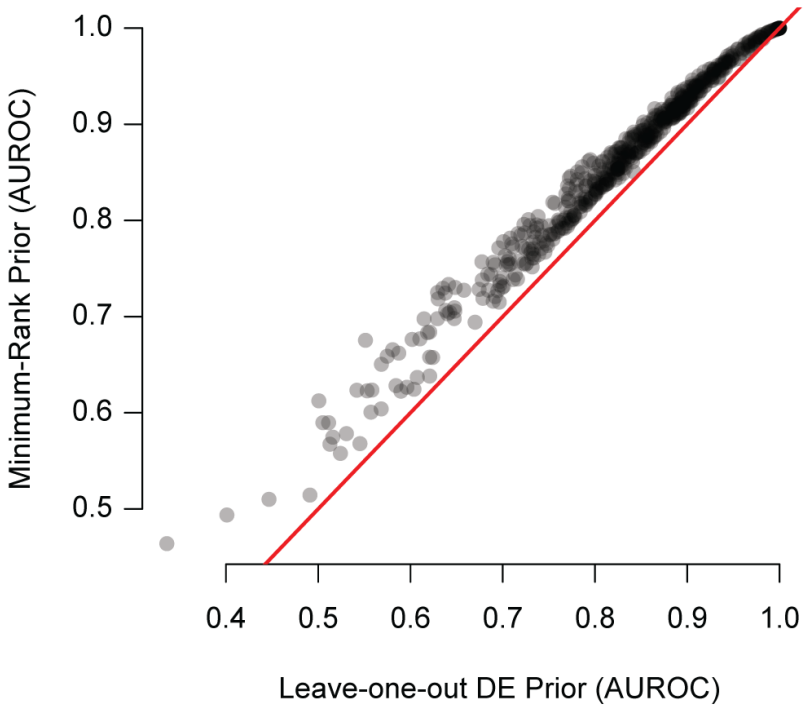
A-D: Distribution of DE prior ranks for housekeeping gene sets from bulk microarray (**A**), bulk RNA-seq (**B**), and single cell RNA-seq (**C&D**). In all cases mean ranks are low, confirming that they are infrequently differentially expressed. **E** - Venn diagram of gene overlaps in the four housekeeping sets. Few genes overlap between all four sets. **F** - Boxplot of DE prior rank compared to gene recurrence across housekeeping gene sets. Boxes represent quartiles, and the line is the median, with whiskers extending to 1.5 times the interquartile range. Genes that recur across a greater number of sets tend to have lower DE prior ranks, indicating that they are less commonly DE. RPS24 is a clear outlier to this trend.

Supplementary Figure 2 – The global DE prior predicts cell type markers from the pancreas



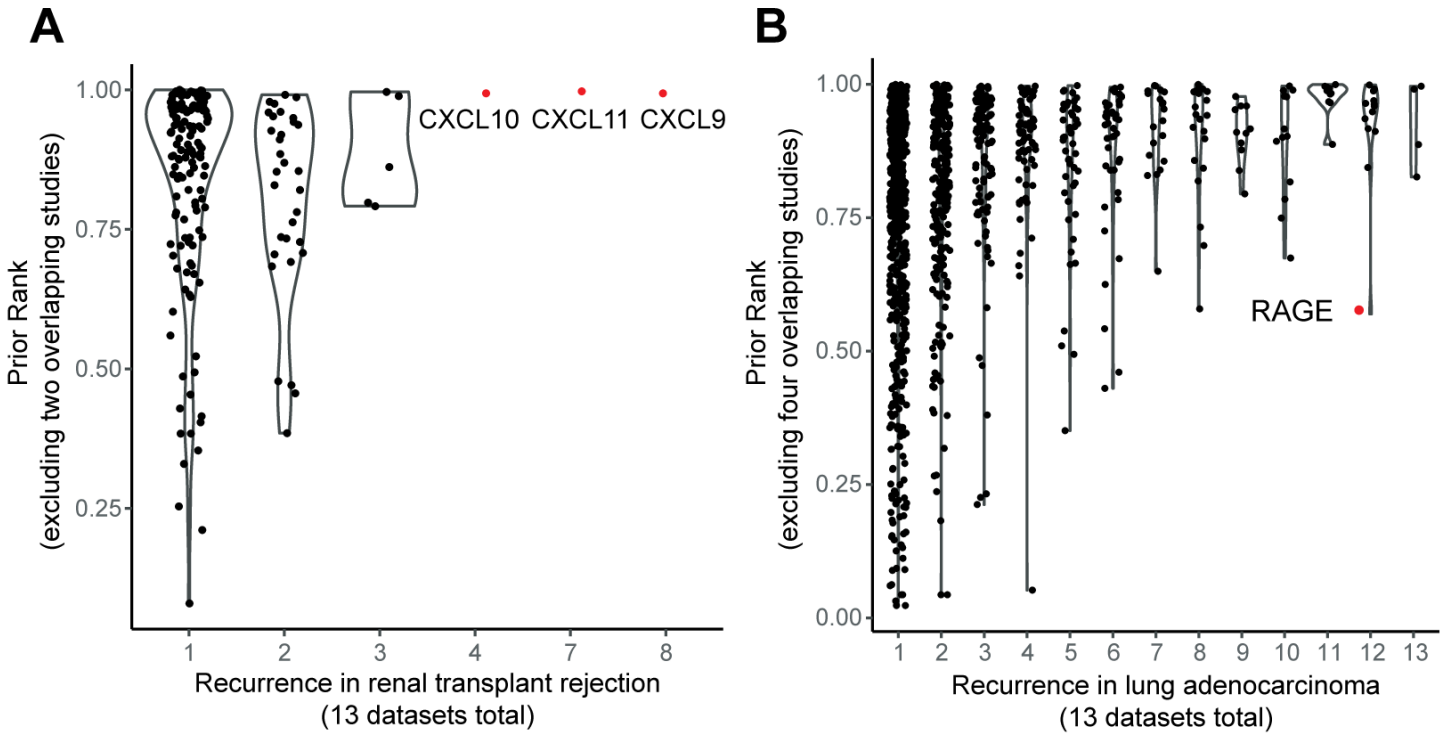
Panels depict DE prior predictions for alpha and beta cell markers from five independent single cell RNA-seq experiments, labeled by the first author of each study. In all cases, the prior has good performance (>0.7) suggesting that it is not overly biased toward studies of disease, cell lines, or other perturbations.

Supplementary Figure 3 – The minimum-rank prior is a robust approximation of the DE prior



Performance of the minimum-rank prior is plotted with respect to performance of the leave-one-out cross-validated prior, where each point represents a study in the compendium. Mean AUROCs are slightly higher for the minimum-rank prior (0.87 vs 0.83), as expected, but the linearity of the relationship suggests that it appropriately recapitulates the individual DE priors.

Supplementary Figure 4 – Re-interpreting DE genes from meta-analysis



A - DE genes from meta-analysis of 13 renal transplant rejection studies are plotted with respect to their prior rank and their recurrence. The prior was generated from the compendium, excluding two studies included in the transplant rejection meta-analysis. Results are consistent with those from the global DE prior (Figure 5C). The only genes that recur among most studies are all very frequently DE. **B -** DE genes from meta-analysis of 13 lung adenocarcinoma studies are plotted with respect to their prior rank and their recurrence. The prior was generated from the compendium, excluding four studies included in the lung cancer meta-analysis. Results are consistent with those from the global DE prior (Figure 5D), with highly recurrent genes showing high prior ranks.

Supplementary Table 1 – Top gene clusters and GO enrichment

Cluster labels provide a description of the GO terms associated with the genes in the module.

Cluster Description	Genes	Number of Significant GO terms (FDR<0.05)	GO Terms
Extracellular Matrix	ADH1B, ALAS2, BCHE, CD36, CLU, COL1A1, COL1A2, COL3A1, COL11A1, VCAN, CYP1B1, AKR1C2, DPP6, , EBF1, FABP4, FABP7, FN1, GPM6A, HBB, HBD, HOXB3, HOXC6, HPGD, ITGB8, KCNJ2, LPP, LUM, MGP, MME, NDP, OGN, PDK4, RORA, SFRP2, TNS1, WNT5A, XIST, AKR1C3, TSIX, CXCL14, AKAP12, RASGRP1, DHRS2, ABCA8, GPNMB, POSTN, FGL2, SULF1, ABI3BP, EGFL6, AKR1B10, CTHRC1, SESN3, SYNPO2, BOD1L1, H19, MALAT1, ANKRD36BP2	100	extracellular matrix, collagen fibril organization, blood vessel development, integrin binding, tissue development, wound healing
Interferon	BIRC3, TNFRSF17, HLA-DQA1, IFI27, IFIT1, IGHG1, IL7R, CXCL10, CXCL9, MMP12, POU2AF1, RGS1, CCL5, CXCL11, UBD, IFI44, CXCL13, IFI44L, LAMP3, ADAMDEC1, MZB1, FAM46C, RSAD2, ANKRD22, CMPK2, FAM26F, IGLL5	130	innate immune response, type I interferon response, chemokine activity, cytokine activity, antigen binding
Inflammation	ADM, ALOX5, ANXA3, AREG, BCL2A1, CD69, CHI3L1, CSTA, DUSP4, EREG, GJB2, CXCL1, CXCL2, CXCL3, HMOX1, HSPA6, IL1B, IL6, CXCL8, KRT19, LCN2, LTF, TACSTD2, MMP1, CEACAM6, NNMT, SERPINB2, SERPINA1, PTGS2, PTX3, RARRES1, S100A8, S100A9, S100A12, CCL2, CCL20, CXCL6, CXCL5, SPP1, TCN1, THBS1, TNFAIP3, TNFAIP6, CXCR4, TFPI2, VNN1, KYNU, SOCS3, GDF15, CRISP3, G0S2, MS4A4A, CHST15, MUCL1, FDCSP	223	defense response, neutrophil migration, cell motility, positive regulation of phosphorylation
Cell Cycle	BCAT1, GPX2, MYBL1, RRM2, TOP2A, NMU, SLC7A11, ANLN, PBK, BMS1P20, ASPM	4	cell cycle, mitotic cell cycle
Stress Response	ATF3, CA2, HBEGF, DUSP1, DUSP2, DUSP6, EGR1, EGR2, EGR3, FOS, FOSB, HSPA1B, ID1, INSIG1, JUN, MT1E, MT1X, NR4A2, NR4A3, BHLHE40, IER3, KLF4, MAFF, C8orf4, SIK1, C11orf96	75	transcription regulatory region DNA binding, positive regulation of transcription from RNA polymerase II promoter, response to growth factor
Y-chromosome	UTY, ZFY, KDM5D, USP9Y, DDX3Y, EIF1AY, NLGN4Y, TTTY14, TXLNGY	0	

Supplementary Table 2 – Cancer and transplant rejection datasets

Lung adenocarcinoma and renal transplant rejection datasets re-analyzed for Figure 5 and Supplementary Figure 4. Studies are listed first by phenotype, then in decreasing order based on the total number of samples. Studies that are overlap with the compendium are indicated by a star in the right-most column.

Phenotype	Accession	Platform	First Author	N Control	N Cases	
Lung Adenocarcinoma	GSE31210	GPL570	Okayama	20	226	*
Lung Adenocarcinoma	GSE32863	GPL6884	Selamat	58	58	
Lung Adenocarcinoma	GSE19188	GPL570	Hou	65	45	*
Lung Adenocarcinoma	GSE43458	GPL6244	Kabbout	30	80	
Lung Adenocarcinoma	GSE10072	GPL96	Landi	49	58	
Lung Adenocarcinoma	GSE30219	GPL570	Rousseaux	14	85	*
Lung Adenocarcinoma	GSE43767	GPL6480	Feng	15	69	
Lung Adenocarcinoma	GSE63459	GPL6883	Robles	32	33	
Lung Adenocarcinoma	GSE7670	GPL96	Su	27	27	
Lung Adenocarcinoma	GSE27262	GPL570	Wei	25	25	*
Lung Adenocarcinoma	GSE12236	GPL5188	Xi	20	20	
Lung Adenocarcinoma	GSE2514	GPL8300	Stearman	19	20	
Lung Adenocarcinoma	GSE21933	GPL6254	Lo	21	11	
Transplant Rejection	GSE36059	GPL570	Reeve	281	122	*
Transplant Rejection	GSE48581	GPL570	Halloran	222	78	
Transplant Rejection	GSE21374	GPL570	Einecke	206	76	
Transplant Rejection	GSE50058	GPL570	Khatri	58	43	*
Transplant Rejection	GSE22459	GPL570	Park	25	40	
Transplant Rejection	GSE50084	GPL6244	Ó Broin	33	28	
Transplant Rejection	GSE44131	GPL6244	Hayde	12	45	
Transplant Rejection	GSE34748	GPL570	Dean	36	20	
Transplant Rejection	GSE53605	GPL571	Maluf	32	23	
Transplant Rejection	GSE9489	GPL570	Saint-Mezard	32	15	
Transplant Rejection	GSE47097	GPL6883	Rekers	4	36	
Transplant Rejection	GSE65326	GPL10558	Toki	7	16	
Transplant Rejection	GSE1563	GPL8300	Flechner	15	7	