Patterns, Volume *5*

Supplemental information

RummaGEO: Automatic mining of human and mouse gene sets from GEO

Giacomo B. Marino, Daniel J.B. Clarke, Alexander Lachmann, Eden Z. Deng, and Avi Ma'ayan

Supplemental Information

for

RummaGEO: Automatic Mining of Human and Mouse Gene Sets from GEO

Giacomo B. Marino[1], Daniel J. B. Clarke[1], Eden Z. Deng[1], Avi Ma'ayan[1,*]

[1]Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York 10029, NY USA

*To whom correspondence should be addressed:

Lead contact and corresponding author: avi.maayan@mssm.edu

Figure S1

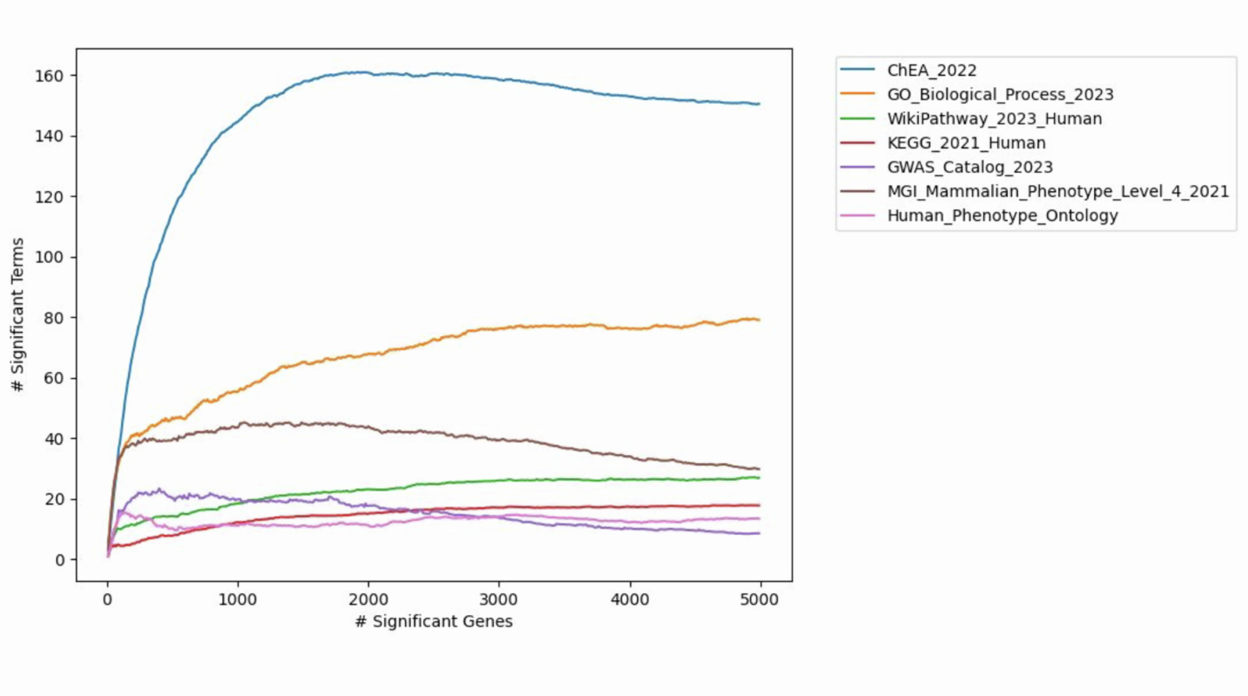


Fig. S1 Significantly enriched terms from selected Enrichr libraries varying based on the cutoff number of genes.

Figure S2

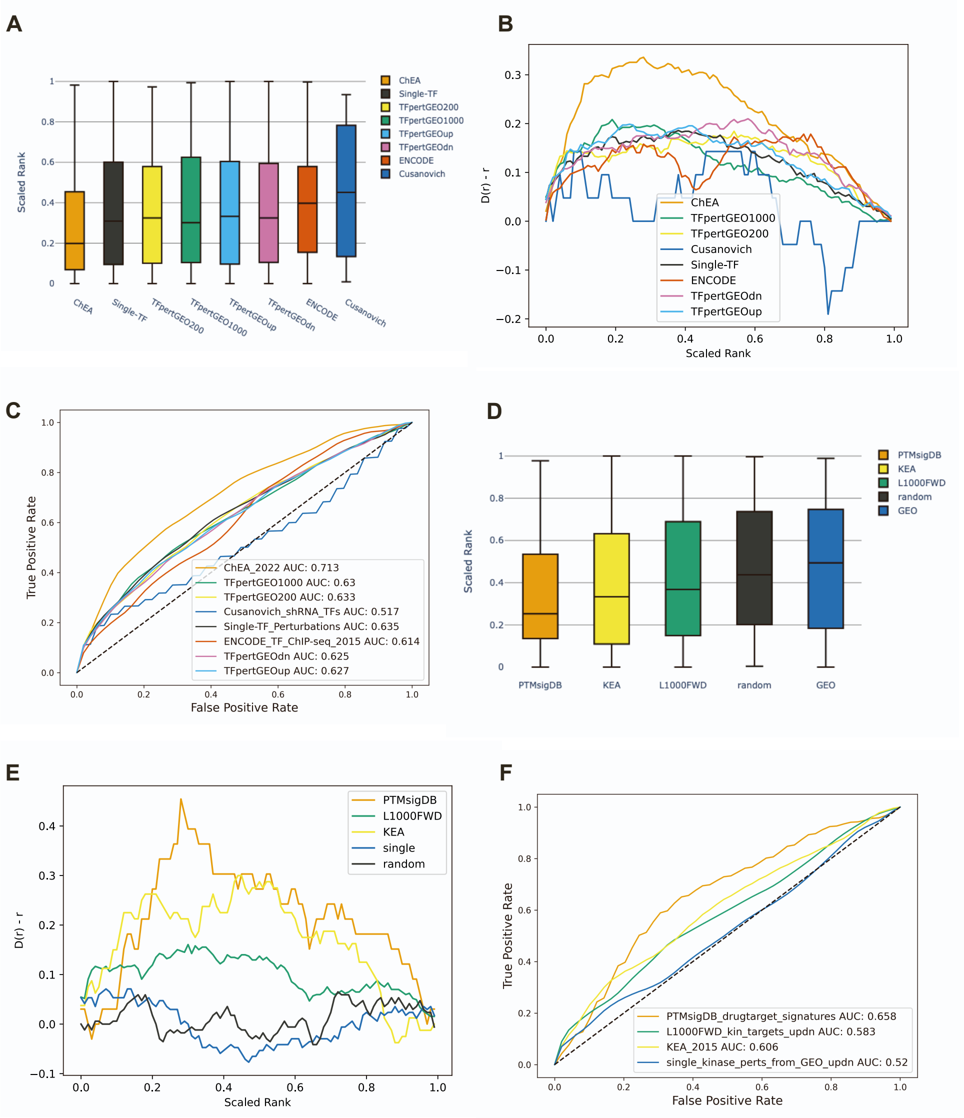
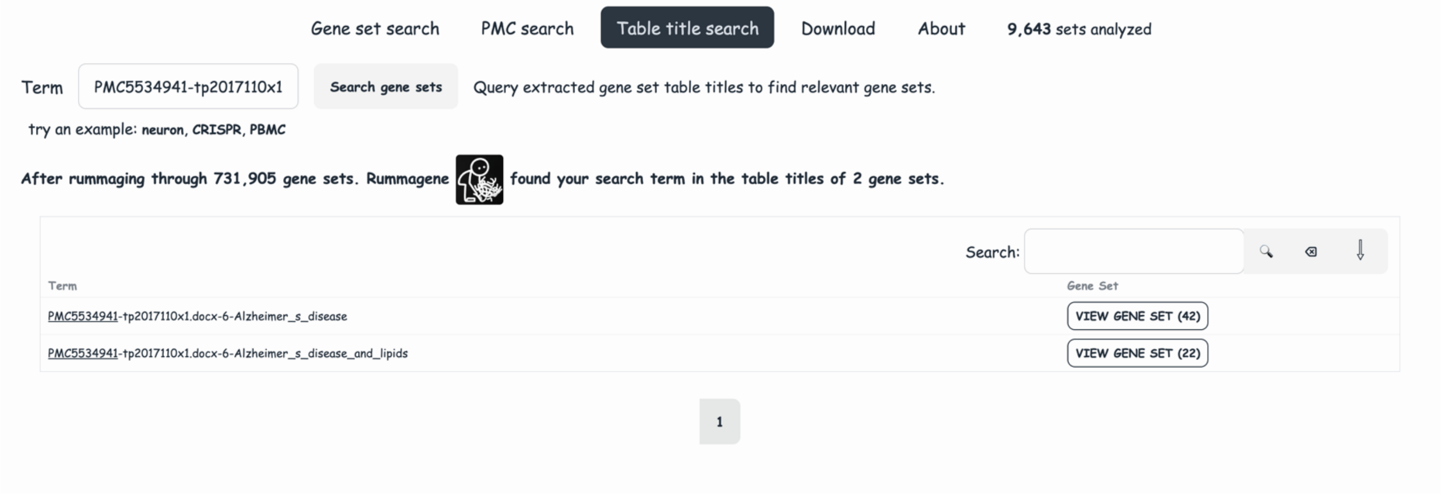


Fig. S2 RummaGEO kinase and transcription factor libraries partial intersection benchmarking. A. Scaled rank (0 highest rank, 1 lowest rank) for transcription factor benchmarking libraries as computed by the Fisher's exact test; **B.** Deviation of the cumulative distribution for scaled ranks of each transcription factor from uniform distribution (Kolmogorov-Smirnov test for goodness of fit compared to uniform distribution: ChEA 2022 p = 3.83E-35; TFpertGEO1000 p = 1.07E-24; TFpertGEO200 p = 2.56E-21; Cusanovich shRNA TFs p = 5.47E-01; Single-TF Perturbations p = 2.01E-25; ENCODE TF ChIP-seq 2015 p = 1.01E-10; TFpertGEOdn p = 4.33E-19; TFpertGEOup p = 6.22E-21; **C.** 5,000 bootstrapped curves with downsampled negative class were generated to compute mean receiver operating characteristic (ROC) curves and mean area under the ROC curves (AUC) for transcription factors. **D.** Scaled rank (0 highest rank, 1 lowest rank) for kinase benchmarking libraries as computed by Fisher's exact test; **E.** Deviation of the cumulative distribution for scaled ranks of each kinase from uniform distribution (Kolmogorov-Smirnov test for goodness of fit compared to uniform distribution: PTMsigDB drugtarget signatures p = 1.05E-03; L1000FWD kin targets updn p = 1.06E-08; KEA 2015 p = 1.14E-03; single kinase perts from GEO updn p = 1.50E-01;
random P = 5.59E-01; **F.** 5,000 bootstrapped curves with downsampled negative class generated to compute mean receiver operating characteristic (ROC) curves and mean area under the ROC curves (AUC) for kinases.

Figure S3



Term: PMC5534941-tp2017110x1

Query extracted gene set table titles to find relevant gene sets.

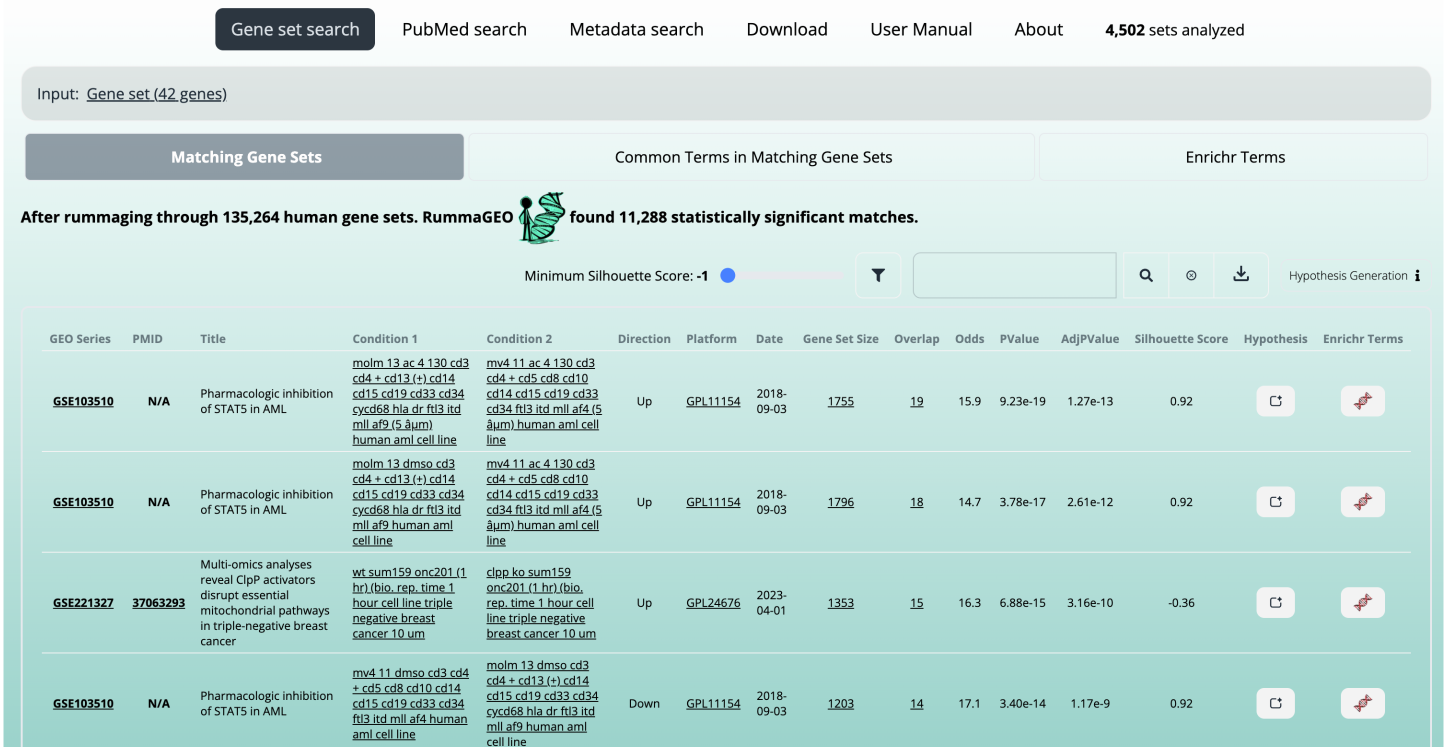
try an example: **neuron**, **CRISPR**, **PBMC**

After rummaging through 731,905 gene sets. Rummagene found your search term in the table titles of 2 gene sets.

Term	Gene Set
PMC5534941-tp2017110x1.docx-6-Alzheimer_s_disease	VIEW GENE SET (42)
PMC5534941-tp2017110x1.docx-6-Alzheimer_s_disease_and_lipids	VIEW GENE SET (22)

Fig. S3 Alzheimer's gene set on Rummagene (use case 1).

Figure S4



Input: Gene set (42 genes)

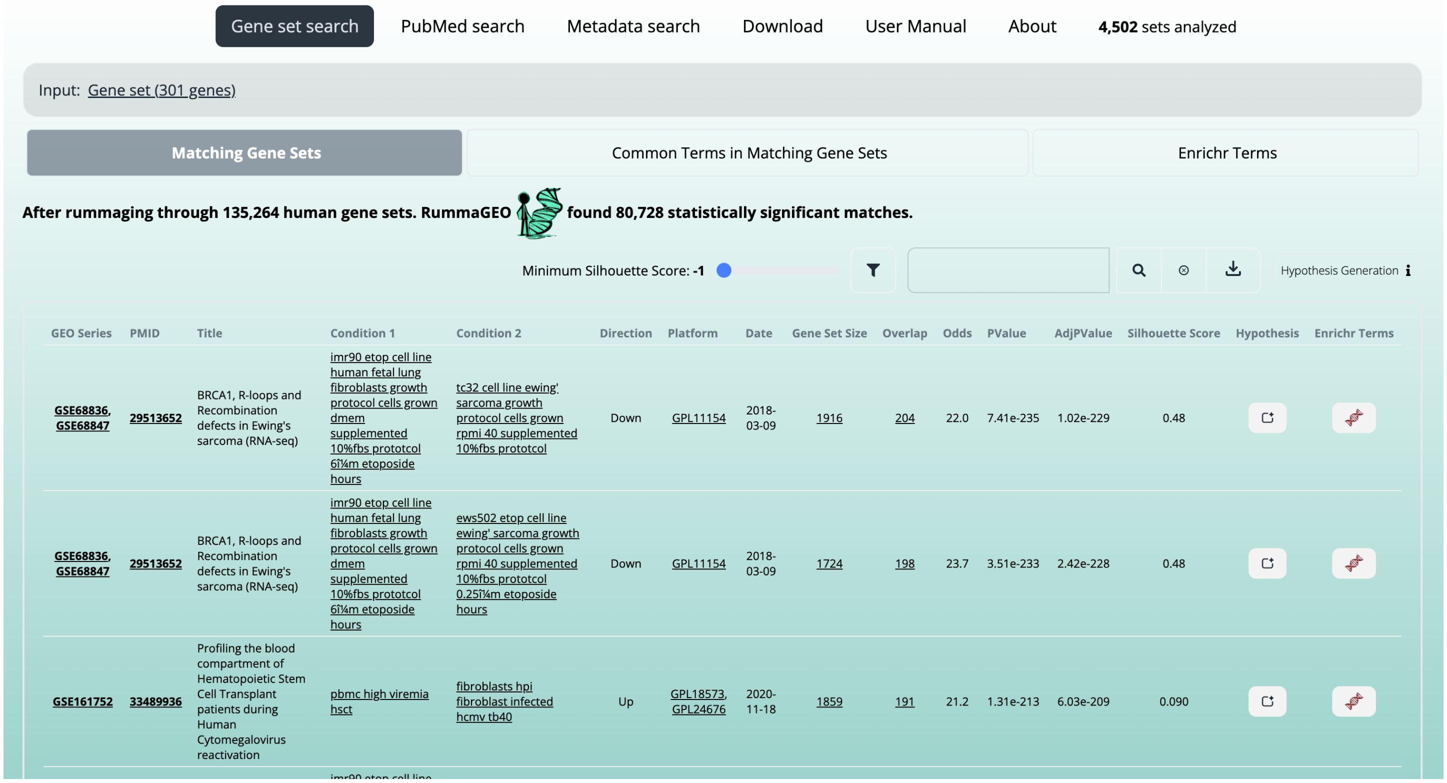
After rummaging through 135,264 human gene sets. RummaGEO found 11,288 statistically significant matches.

Minimum Silhouette Score: **-1**

GEO Series	PMID	Title	Condition 1	Condition 2	Direction	Platform	Date	Gene Set Size	Overlap	Odds	PValue	AdjPValue	Silhouette Score	Hypothesis	Enrichr Terms
GSE103510	**N/A**	Pharmacologic inhibition of STAT5 in AML	molm 13 ac 4 130 cd3 cd4 + cd13 (+) cd14 cd15 cd19 cd33 cd34 cycd68 hla dr ftl3 itd mll af9 (5 âµm) human aml cell line	mv4 11 ac 4 130 cd3 cd4 + cd5 cd8 cd10 cd14 cd15 cd19 cd33 cd34 ftl3 itd mll af4 (5 âµm) human aml cell line	Up	GPL11154	2018-09-03	1755	19	15.9	9.23e-19	1.27e-13	0.92		
GSE103510	**N/A**	Pharmacologic inhibition of STAT5 in AML	molm 13 dmso cd3 cd4 + cd13 (+) cd14 cd15 cd19 cd33 cd34 cycd68 hla dr ftl3 itd mll af9 human aml cell line	mv4 11 ac 4 130 cd3 cd4 + cd5 cd8 cd10 cd14 cd15 cd19 cd33 cd34 ftl3 itd mll af4 (5 âµm) human aml cell line	Up	GPL11154	2018-09-03	1796	18	14.7	3.78e-17	2.61e-12	0.92		
GSE221327	**37063293**	Multi-omics analyses reveal ClpP activators disrupt essential mitochondrial pathways in triple-negative breast cancer	wt sum159 onc201 (1 hr) (bio. rep. time 1 hour cell line triple negative breast cancer 10 um	clpp ko sum159 onc201 (1 hr) (bio. rep. time 1 hour cell line triple negative breast cancer 10 um	Up	GPL24676	2023-04-01	1353	15	16.3	6.88e-15	3.16e-10	-0.36		
GSE103510	**N/A**	Pharmacologic inhibition of STAT5 in AML	mv4 11 dmso cd3 cd4 + cd5 cd8 cd10 cd14 cd15 cd19 cd33 cd34 ftl3 itd mll af4 human aml cell line	molm 13 dmso cd3 cd4 + cd13 (+) cd14 cd15 cd19 cd33 cd34 cycd68 hla dr ftl3 itd mll af9 human aml cell line	Down	GPL11154	2018-09-03	1203	14	17.1	3.40e-14	1.17e-9	0.92		

Fig. S4 Results of submitting the Alzheimer's gene set from Rummagene on RummaGEO (use case 1).

Figure S5



Input: Gene set (301 genes)

After rummaging through 135,264 human gene sets. RummaGEO found 80,728 statistically significant matches.

Minimum Silhouette Score: **-1**

GEO Series	PMID	Title	Condition 1	Condition 2	Direction	Platform	Date	Gene Set Size	Overlap	Odds	PValue	AdjPValue	Silhouette Score	Hypothesis	Enrichr Terms
GSE68836, GSE68847	29513652	BRCA1, R-loops and Recombination defects in Ewing's sarcoma (RNA-seq)	imr90 etop cell line human fetal lung fibroblasts growth protocol cells grown dmem supplemented 10%fbs prototcol 6î¼m etoposide hours	tc32 cell line ewing' sarcoma growth protocol cells grown rpmi 40 supplemented 10%fbs prototcol	Down	GPL11154	2018-03-09	1916	204	22.0	7.41e-235	1.02e-229	0.48		
GSE68836, GSE68847	29513652	BRCA1, R-loops and Recombination defects in Ewing's sarcoma (RNA-seq)	imr90 etop cell line human fetal lung fibroblasts growth protocol cells grown dmem supplemented 10%fbs prototcol 6î¼m etoposide hours	ews502 etop cell line ewing' sarcoma growth protocol cells grown rpmi 40 supplemented 10%fbs prototcol 0.25î¼m etoposide hours	Down	GPL11154	2018-03-09	1724	198	23.7	3.51e-233	2.42e-228	0.48		
GSE161752	33489936	Profiling the blood compartment of Hematopoietic Stem Cell Transplant patients during Human Cytomegalovirus reactivation	pbmc high viremia hsct	fibroblasts hpi fibroblast infected hcmv tb40	Up	GPL18573, GPL24676	2020-11-18	1859	191	21.2	1.31e-213	6.03e-209	0.090		
			imr90 etop cell line												

Fig. S5 Results from submitting SenoRanger gene set on RummaGEO (use case 2).