GeneQuery:
Making non-trivial statements

Max Artyomov, Sep 23, 2019

Going beyond making obvious statements..

Systems Biology can actually be useful tool

Part I: getting comfortable "jumping" between the datasets

Going beyond making obvious statements..

So, you've got your RNA-seq….

Going beyond making obvious statements..

So, you've got your RNA-seq….

So, you've got your
differentially expressed genes…

Going beyond making obvious statements..

So, you've got your RNA-seq….

So, you've got your
differentially expressed genes…

So, you've got your
differentially regulated pathways…

Going beyond making obvious statements..

So, you've got your RNA-seq….

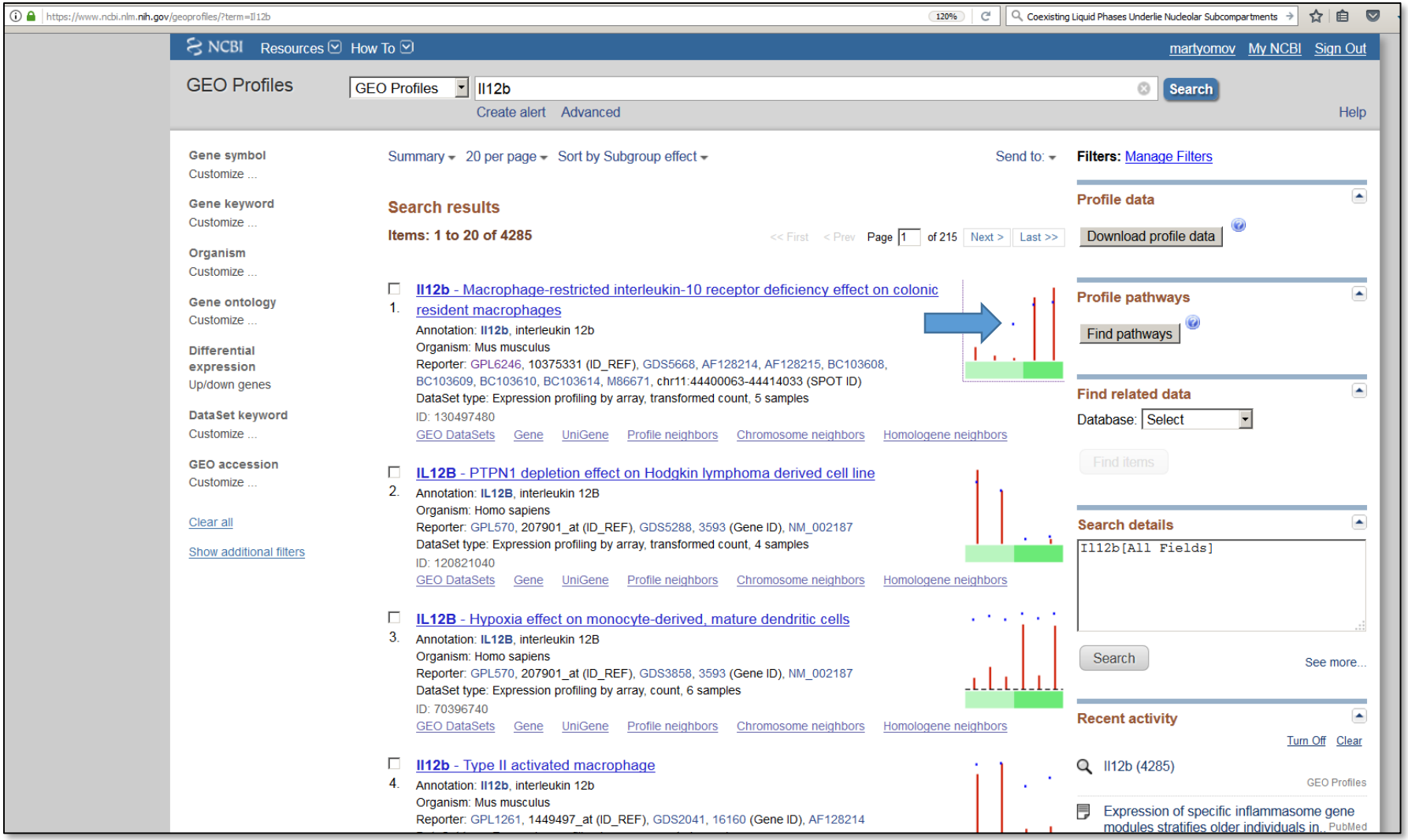
So, you've got your
differentially expressed genes…

So, you've got your
differentially regulated pathways…

What's next?

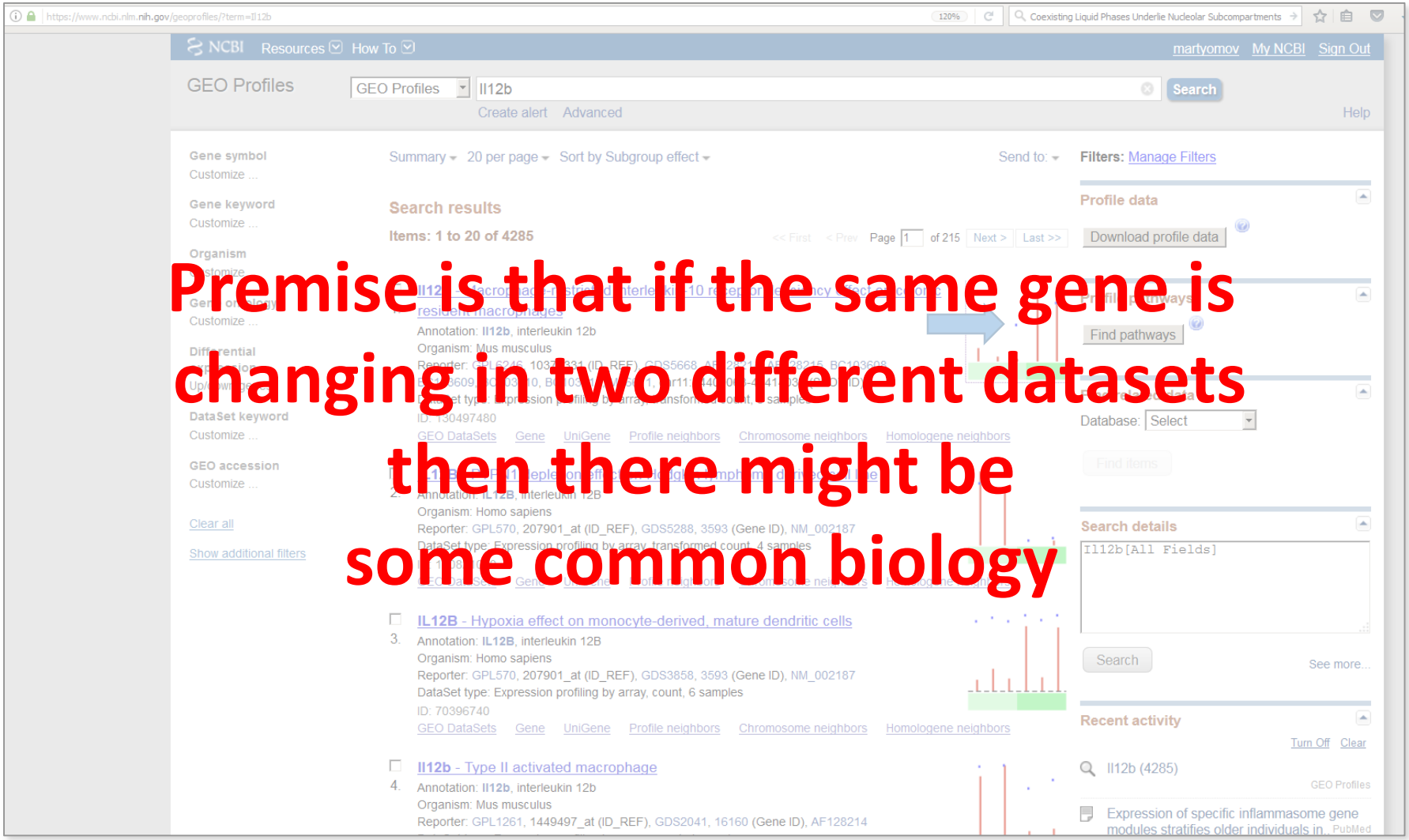
Single gene follow-up: GEO profiles tool

https://www.ncbi.nlm.nih.gov/geoprofiles/



Single gene follow-up: GEO profiles tool

https://www.ncbi.nlm.nih.gov/geoprofiles/



Premise is that if the same gene is changing in two different datasets then there might be some common biology

This is not good enough

- Single genes can have weird behavior with outliers
- Single genes do not necessarily define the phenotype

Working with gene sets is much more reliable!!

LETTER

https://doi.org/10.1038/s41586-018-0052-z

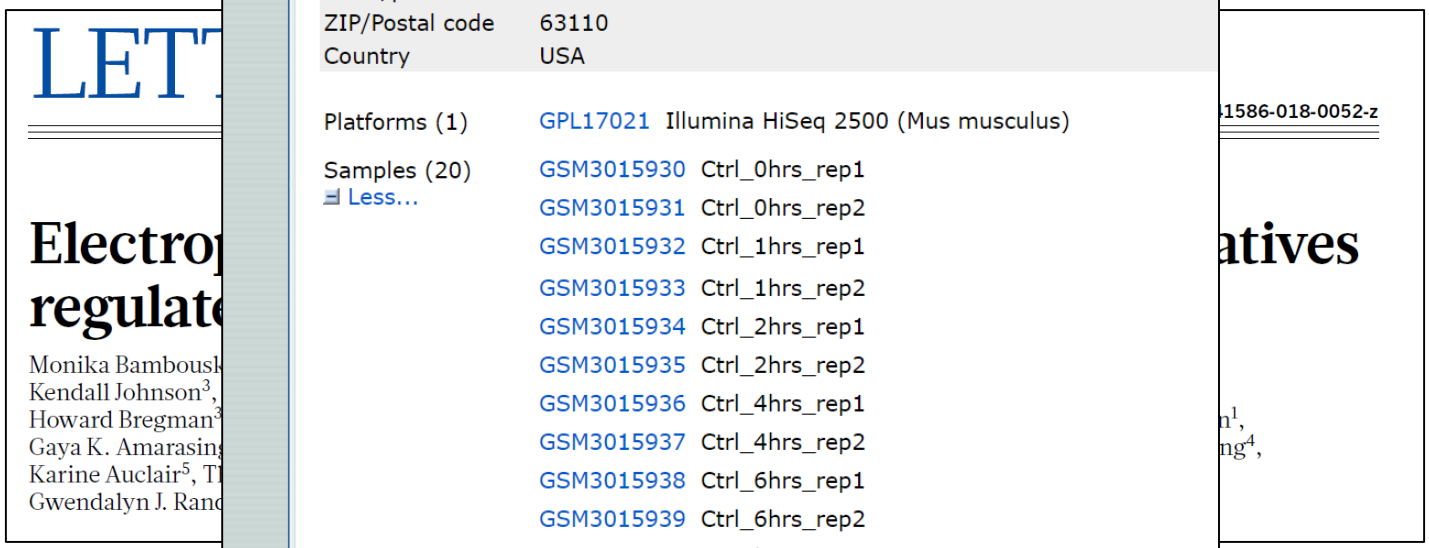
Electrophilic properties of itaconate and derivatives regulate the IκBζ–ATF3 inflammatory axis

Monika Bambouskova[1], Laurent Gorvel[1], Vicky Lampropoulou[1], Alexey Sergushichev[2], Ekaterina Loginicheva[1], Kendall Johnson[3], Daniel Korenfeld[1], Mary Elizabeth Mathyer[4], Hyeryun Kim[3], Li-Hao Huang[1], Dustin Duncan[5], Howard Bregman[3], Abdurrahman Keskin[6], Andrea Santeford[7], Rajendra S. Apte[7], Raghav Sehgal[8], Britney Johnson[1], Gaya K. Amarasinghe[1], Miguel P. Soares[9], Takashi Satoh[10], Shizuo Akira[10], Tsonwin Hai[11], Cristina de Guzman Strong[4], Karine Auclair[5], Thomas P. Roddy[3], Scott A. Biller[3], Marko Jovanovic[6], Eynav Klechevsky[1], Kelly M. Stewart[3], Gwendalyn J. Randolph[1] & Maxim N. Artyomov[1]*

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The raw and processed RNA-seq data have been deposited in the Gene Expression Omnibus with accession numbers GSE102190 and GSE110749. The original mass spectra may be downloaded from MassIVE (http://massive.ucsd.edu) using the identifier MSV000082101. Source data for the graphical representations found in all figures and Extended Data figures are provided. Source data for western blots (uncropped and unprocessed scans with size marker indications) are presented in Supplementary Fig. 1. All other data that support the findings of this study are available from the corresponding author upon reasonable request.

Bambouskova et al, Nature, 2018



acc.cgi?acc=GSE110749

Submission date	Feb 16, 2018
Last update date	Feb 01, 2019
Contact name	Maxim N. Artyomov
E-mail	martyomov@pathology.wustl.edu
Organization name	Washington University in St.Louis
Department	Immunology&Pathology
Street address	660 S. Euclid Avenue, Campus Box 8118
City	St.Louis
State/province	MO
ZIP/Postal code	63110
Country	USA

Platforms (1) GPL17021 Illumina HiSeq 2500 (Mus musculus)

Samples (20) Less...

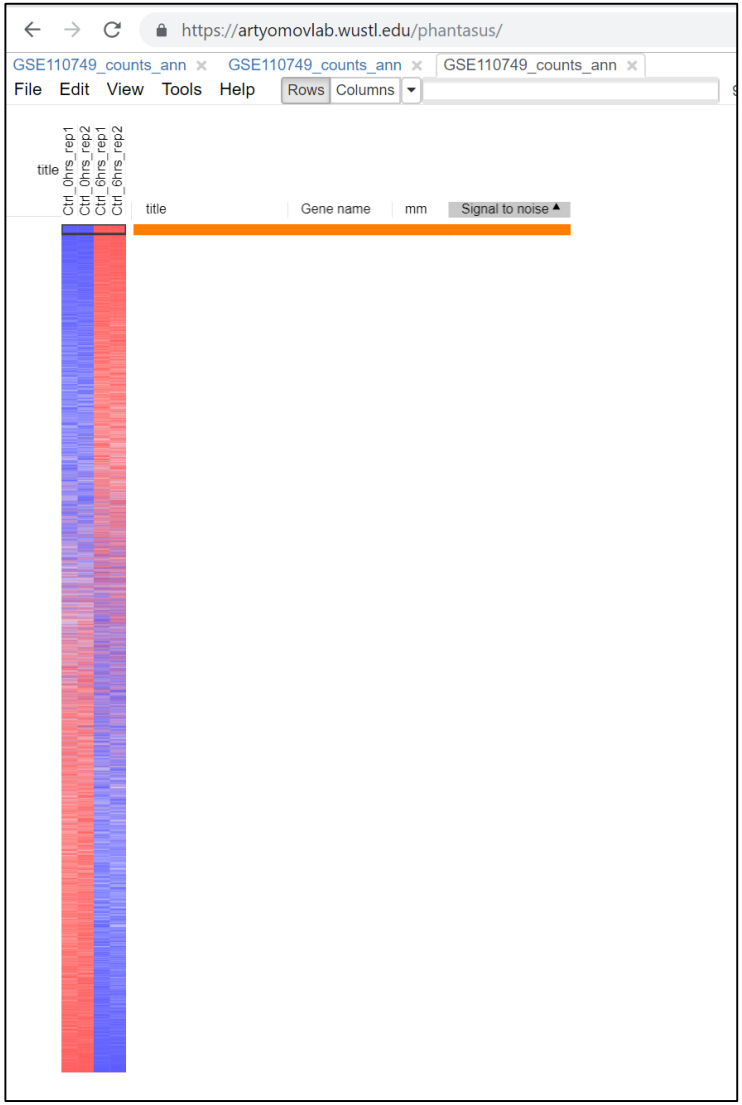
GSM3015930 Ctrl_0hrs_rep1
GSM3015931 Ctrl_0hrs_rep2
GSM3015932 Ctrl_1hrs_rep1
GSM3015933 Ctrl_1hrs_rep2
GSM3015934 Ctrl_2hrs_rep1
GSM3015935 Ctrl_2hrs_rep2
GSM3015936 Ctrl_4hrs_rep1
GSM3015937 Ctrl_4hrs_rep2
GSM3015938 Ctrl_6hrs_rep1
GSM3015939 Ctrl_6hrs_rep2
GSM3015940 DI_0hrs_rep1
GSM3015941 DI_0hrs_rep2
GSM3015942 DI_1hrs_rep1
GSM3015943 DI_1hrs_rep2
GSM3015944 DI_2hrs_rep1
GSM3015945 DI_2hrs_rep2
GSM3015946 DI_4hrs_rep1
GSM3015947 DI_4hrs_rep2
GSM3015948 DI_6hrs_rep1
GSM3015949 DI_6hrs_rep2

Bambouskova et al, Nature, 2018

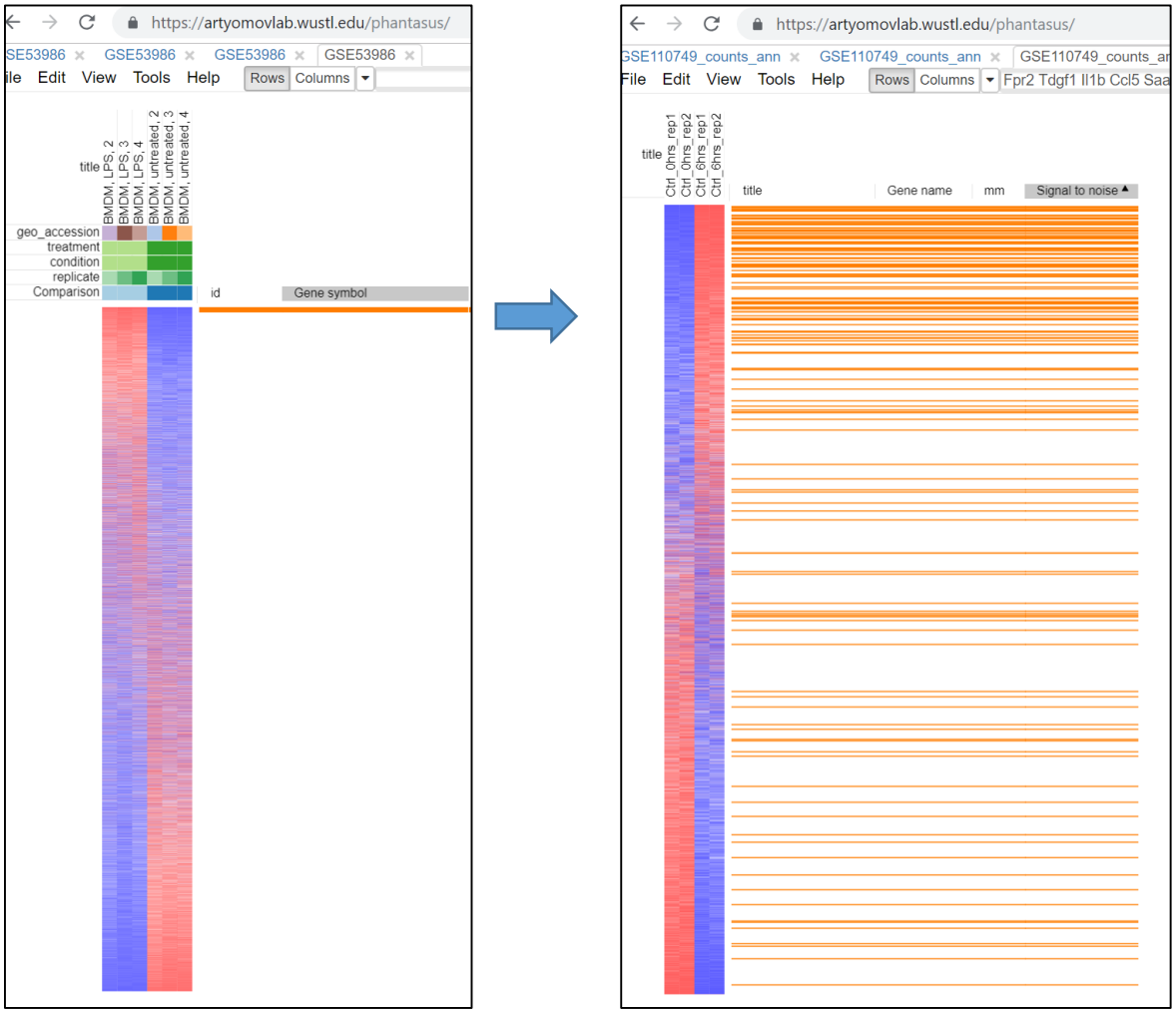
You know what to do!

Counts are available in folder

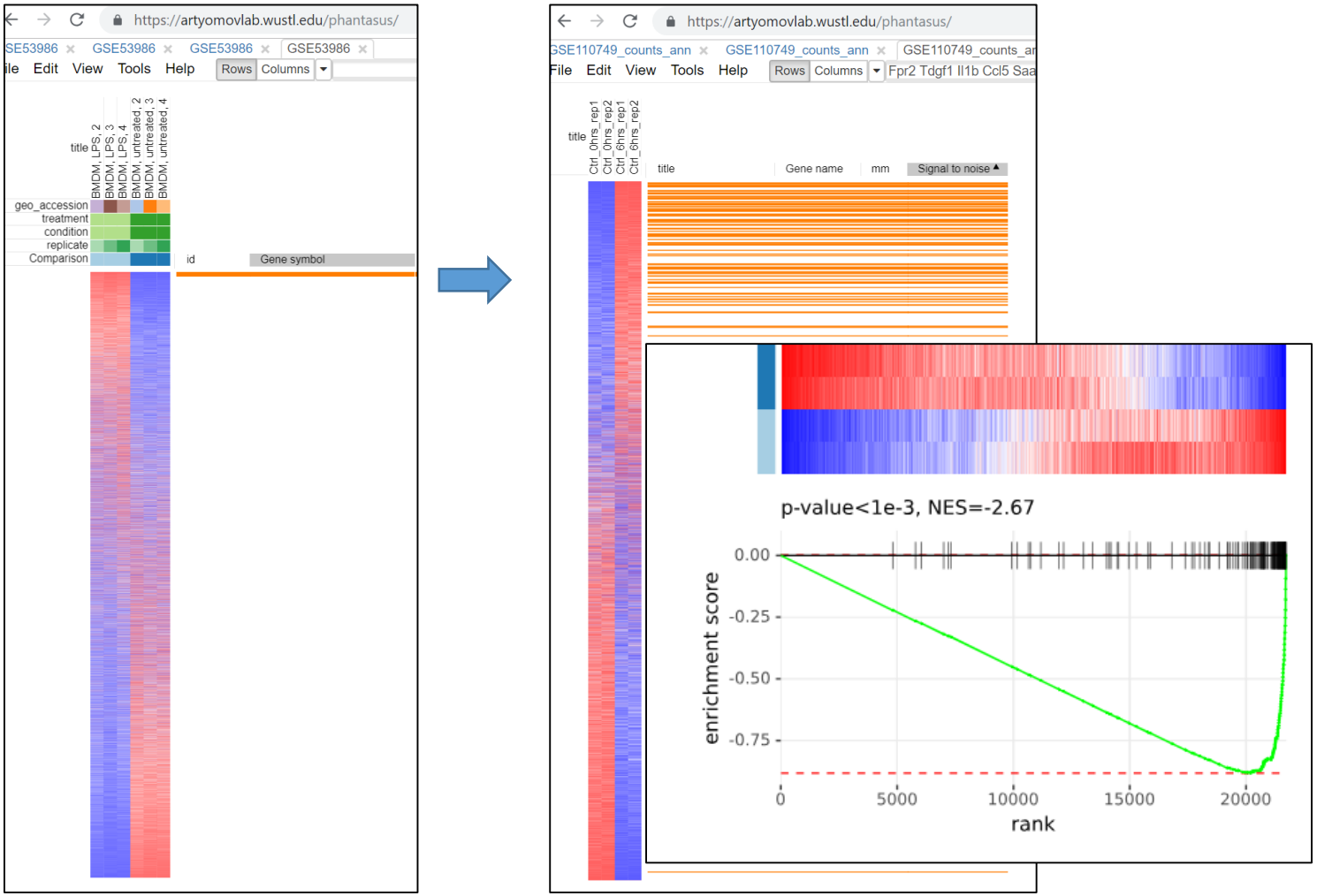
Log2->
qnorm->
Filter top 10k->
DE of 0h vs 6h LPS



The two datasets are connected!



The two datasets are connected!



But you kind of knew the answer!!!

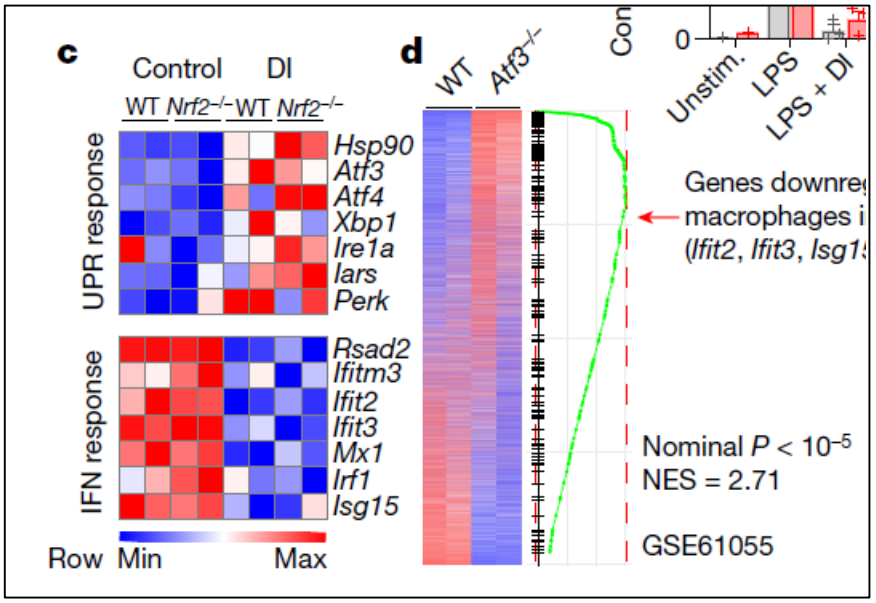
Do people do it in real science? – YES!!

LETTER

https://doi.org/10.1038/s41586-018-0052-z

Electrophilic properties of itaconate and derivatives regulate the IκBζ–ATF3 inflammatory axis

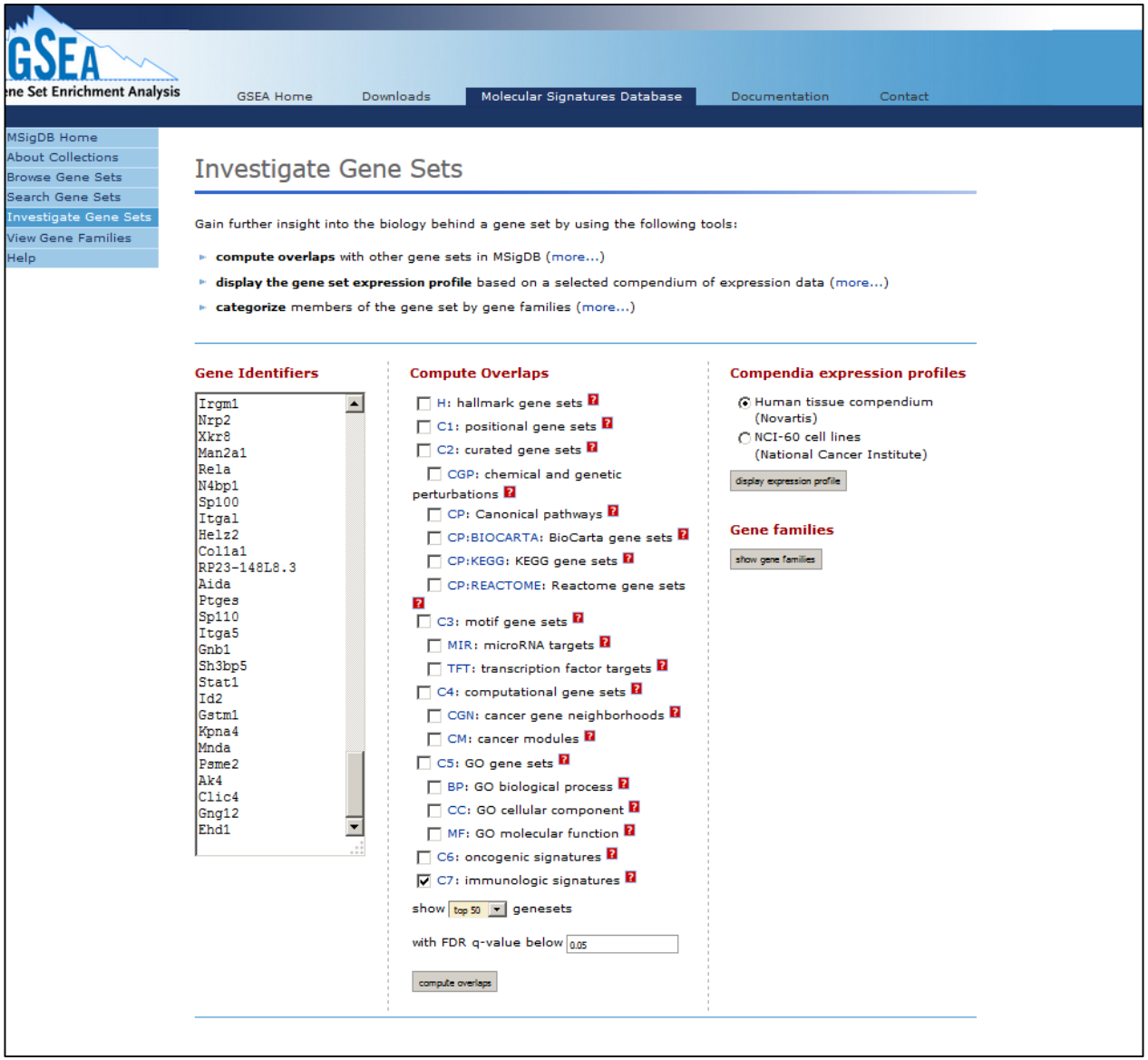
Monika Bambouskova[1], Laurent Gorvel[1], Vicky Lampropoulou[1], Alexey Sergushichev[2], Ekaterina Loginicheva[1], Kendall Johnson[3], Daniel Korenfeld[1], Mary Elizabeth Mathyer[4], Hyeryun Kim[3], Li-Hao Huang[1], Dustin Duncan[5], Howard Bregman[3], Abdurrahman Keskin[6], Andrea Santeford[7], Rajendra S. Apte[7], Raghav Sehgal[8], Britney Johnson[1], Gaya K. Amarasinghe[1], Miguel P. Soares[9], Takashi Satoh[10], Shizuo Akira[10], Tsonwin Hai[11], Cristina de Guzman Strong[4], Karine Auclair[5], Thomas P. Roddy[3], Scott A. Biller[3], Marko Jovanovic[6], Eynav Klechevsky[1], Kelly M. Stewart[3], Gwendalyn J. Randolph[1] & Maxim N. Artyomov[1]*



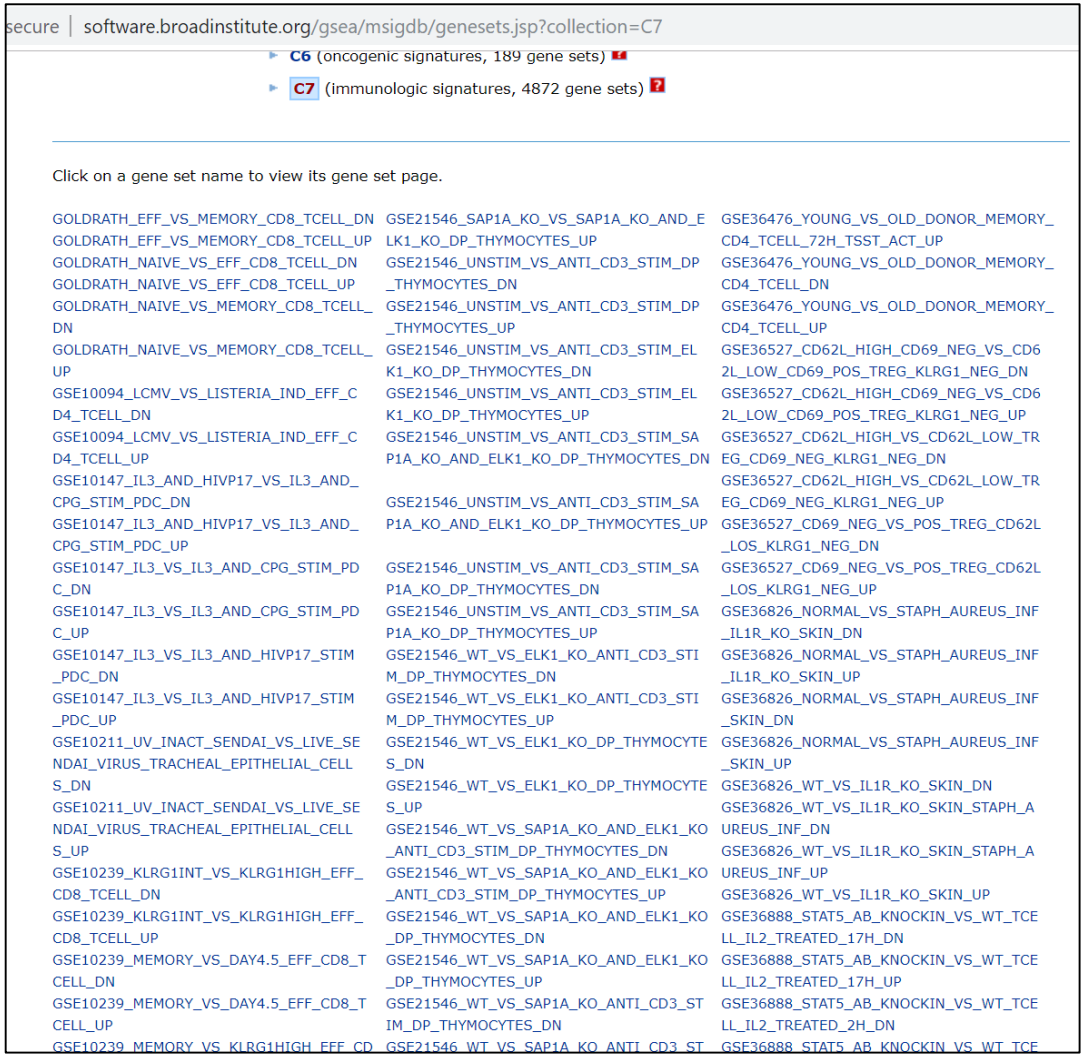
What if we make database?

Try to do diff expression between all conditions in all published datasets…

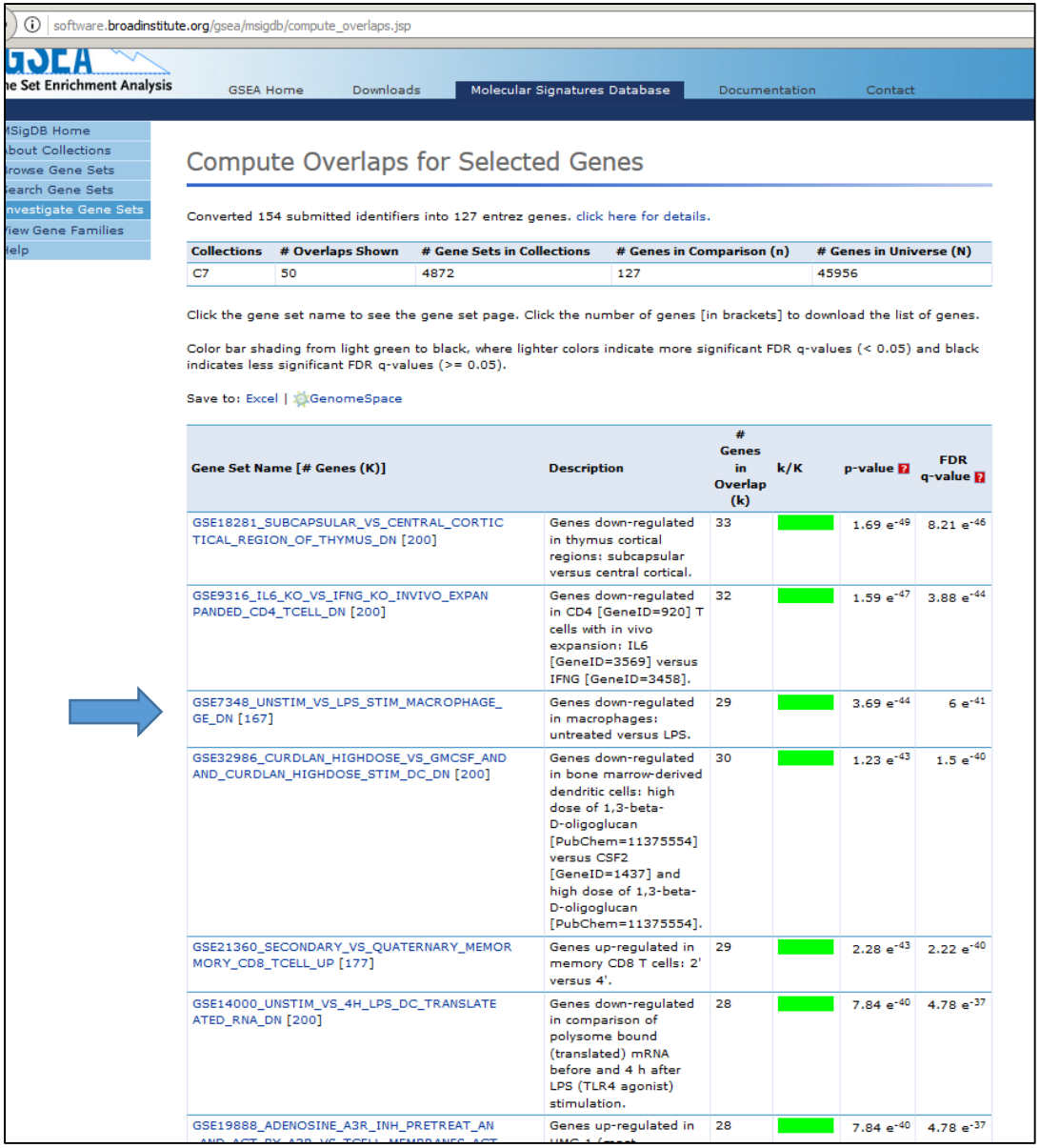
MsigDB c7


GSEA
ne Set Enrichment Analysis
GSEA Home
Downloads
Molecular Signatures Database
Documentation
Contact
MSigDB Home
About Collections
Browse Gene Sets
Search Gene Sets
Investigate Gene Sets
View Gene Families
Help
Investigate Gene Sets
Gain further insight into the biology behind a gene set by using the following tools:
compute overlaps with other gene sets in MSigDB (more...)
display the gene set expression profile based on a selected compendium of expression data (more...)
categorize members of the gene set by gene families (more...)
Gene Identifiers
Irgm1
Nrp2
Xkr8
Man2a1
Rela
N4bp1
Sp100
Itga1
Helz2
Col1a1
RP23-148L8.3
Aida
Ptges
Sp110
Itga5
Gnb1
Sh3bp5
Stat1
Id2
Gstm1
Kpna4
Mnda
Psme2
Ak4
Clic4
Gng12
Ehd1
Compute Overlaps
H: hallmark gene sets
C1: positional gene sets
C2: curated gene sets
CGP: chemical and genetic perturbations
CP: Canonical pathways
CP:BIOCARTA: BioCarta gene sets
CP:KEGG: KEGG gene sets
CP:REACTOME: Reactome gene sets
C3: motif gene sets
MIR: microRNA targets
TFT: transcription factor targets
C4: computational gene sets
CGN: cancer gene neighborhoods
CM: cancer modules
C5: GO gene sets
BP: GO biological process
CC: GO cellular component
MF: GO molecular function
C6: oncogenic signatures
C7: immunologic signatures
show top 50 genesets
with FDR q-value below 0.05
compute overlaps
Compendia expression profiles
Human tissue compendium (Novartis)
NCI-60 cell lines (National Cancer Institute)
display expression profile
Gene families
show gene families

MSigDB.C7 – attempt to make such database



Macrophage activation is found!



software.broadinstitute.org/gsea/msigdb/compute_overlaps.jsp

GSEA — Gene Set Enrichment Analysis

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

MSigDB Home
About Collections
Browse Gene Sets
Search Gene Sets
Investigate Gene Sets
View Gene Families
Help

Compute Overlaps for Selected Genes

Converted 154 submitted identifiers into 127 entrez genes. click here for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
C7	50	4872	127	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Excel | GenomeSpace

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
GSE18281_SUBCAPSULAR_VS_CENTRAL_CORTICTICAL_REGION_OF_THYMUS_DN [200]	Genes down-regulated in thymus cortical regions: subcapsular versus central cortical.	33		1.69 e-49	8.21 e-46
GSE9316_IL6_KO_VS_IFNG_KO_INVIVO_EXPANPANDED_CD4_TCELL_DN [200]	Genes down-regulated in CD4 [GeneID=920] T cells with in vivo expansion: IL6 [GeneID=3569] versus IFNG [GeneID=3458].	32		1.59 e-47	3.88 e-44
GSE7348_UNSTIM_VS_LPS_STIM_MACROPHAGE_GE_DN [167]	Genes down-regulated in macrophages: untreated versus LPS.	29		3.69 e-44	6 e-41
GSE32986_CURDLAN_HIGHDOSE_VS_GMCSF_ANDAND_CURDLAN_HIGHDOSE_STIM_DC_DN [200]	Genes down-regulated in bone marrow-derived dendritic cells: high dose of 1,3-beta-D-oligoglucan [PubChem=11375554] versus CSF2 [GeneID=1437] and high dose of 1,3-beta-D-oligoglucan [PubChem=11375554].	30		1.23 e-43	1.5 e-40
GSE21360_SECONDARY_VS_QUATERNARY_MEMORMORY_CD8_TCELL_UP [177]	Genes up-regulated in memory CD8 T cells: 2' versus 4'.	29		2.28 e-43	2.22 e-40
GSE14000_UNSTIM_VS_4H_LPS_DC_TRANSLATEATED_RNA_DN [200]	Genes down-regulated in comparison of polysome bound (translated) mRNA before and 4 h after LPS (TLR4 agonist) stimulation.	28		7.84 e-40	4.78 e-37
GSE19888_ADENOSINE_A3R_INH_PRETREAT_AN	Genes up-regulated in	28		7.84 e-40	4.78 e-37

You know what to do!

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7348

Status	Public on Apr 22, 2007
Title	Gene Expression in Naive and Tolerant Macrophages stimulated with LPS
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	The inflammatory response initiated by microbial products signaling through Toll-like receptors (TLRs) of the innate immune system is essential for host defense against infection. Because inflammation can be harmful to host tissues, the innate response is highly regulated. Negative regulation of TLR4, the receptor for bacterial lipopolysaccharide (LPS), results in LPS tolerance, defined as hyporesponsiveness to repeated stimulation with LPS. LPS tolerance is thought to protect the host from excessive inflammation by turning off TLR4 signal, which then shuts down TLR-induced genes. However, TLR signaling induces hundreds of genes with very different functions. We reasoned that genes with different functions should have different requirements for regulation. Specifically, genes encoding proinflammatory mediators should be transiently inactivated to limit tissue damage, while genes encoding antimicrobial effectors, which directly target pathogens, should remain inducible in tolerant cells to protect the host from infection. Using an in vitro system of LPS tolerance in macrophages, here we show that TLR-induced genes may indeed be divided into two distinct categories based on their functions and regulatory requirements. Further, we show these distinct groups are regulated by gene-specific, and not signal-specific mechanisms. Keywords: Treatment Comparison
Overall design	We examined gene expression using affymetrix genechips in 3 groups of murine bone-marrow derived macrophages: Naive (untreated), Naive stimulated with LPS, and Tolerant stimulated with LPS. Two biological replicates were performed for each group.
Contributor(s)	Foster SL, Hargreaves DC, Medzhitov R
Citation(s)	Foster SL, Hargreaves DC, Medzhitov R. Gene-specific control of inflammation by TLR-induced chromatin modifications. *Nature* 2007 Jun 21;447(7147):972-8. PMID: 17538624
Submission date	Mar 22, 2007
Last update date	Jan 08, 2019
Contact name	Simmie L Foster
E-mail	simmie.foster@yale.edu
Phone	(203)785-6090
Fax	(203)785-4461
Organization name	Yale University School of Medicine
Department	Immunobiology
Lab	Medzhitov
Street address	300 Cedar St
City	New Haven
State/province	CT

nature Vol 447|21 June 2007|doi:10.1038/nature05836

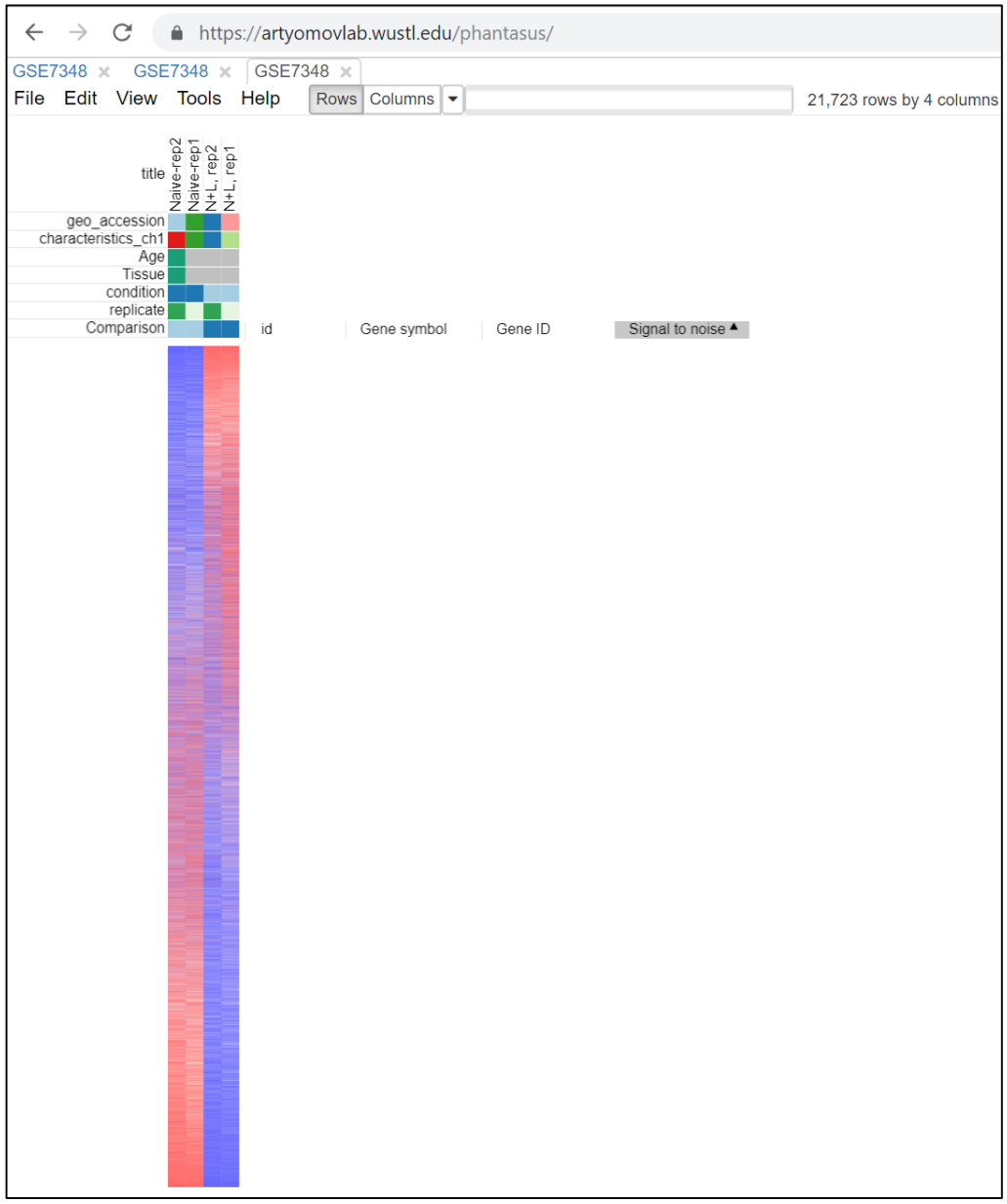
ARTICLES

Gene-specific control of inflammation by TLR-induced chromatin modifications

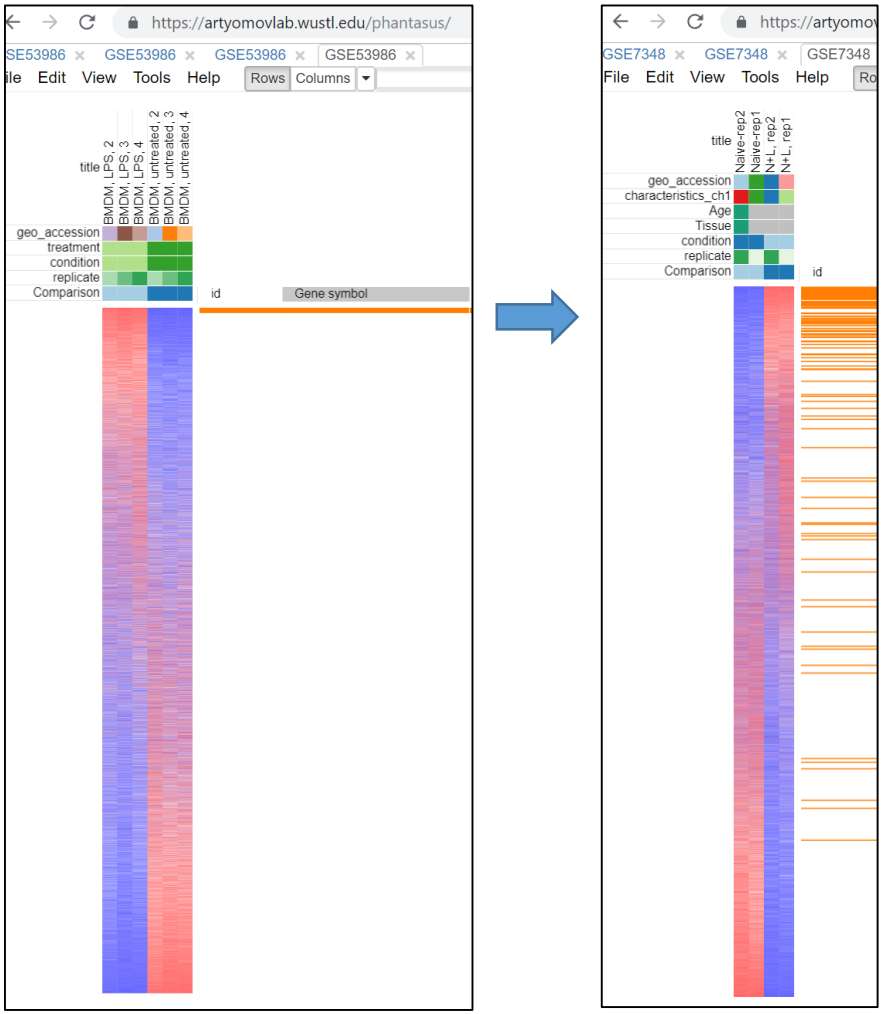
Simmie L. Foster[1]*, Diana C. Hargreaves[1]* & Ruslan Medzhitov[1]

Toll-like receptors (TLRs) induce a multi-component inflammatory response that must be tightly regulated to avoid tissue damage. Most known regulatory mechanisms target TLR signalling pathways and thus broadly inhibit multiple aspects of the inflammatory response. Given the functional diversity of TLR-induced genes, we proposed that additional, gene-specific regulatory mechanisms exist to allow individual aspects of the TLR-induced response to be differentially regulated. Using an *in vitro* system of lipopolysaccharide tolerance in murine macrophages, we show that TLR-induced genes fall into two categories on the basis of their functions and regulatory requirements. We demonstrate that representatives from the two classes acquire distinct patterns of TLR-induced chromatin modifications. These gene-specific chromatin modifications are associated with transient silencing of one class of genes, which includes pro-inflammatory mediators, and priming of the second class, which includes antimicrobial effectors. These findings illustrate an adaptive response in macrophages and reveal component-specific regulation of inflammation.

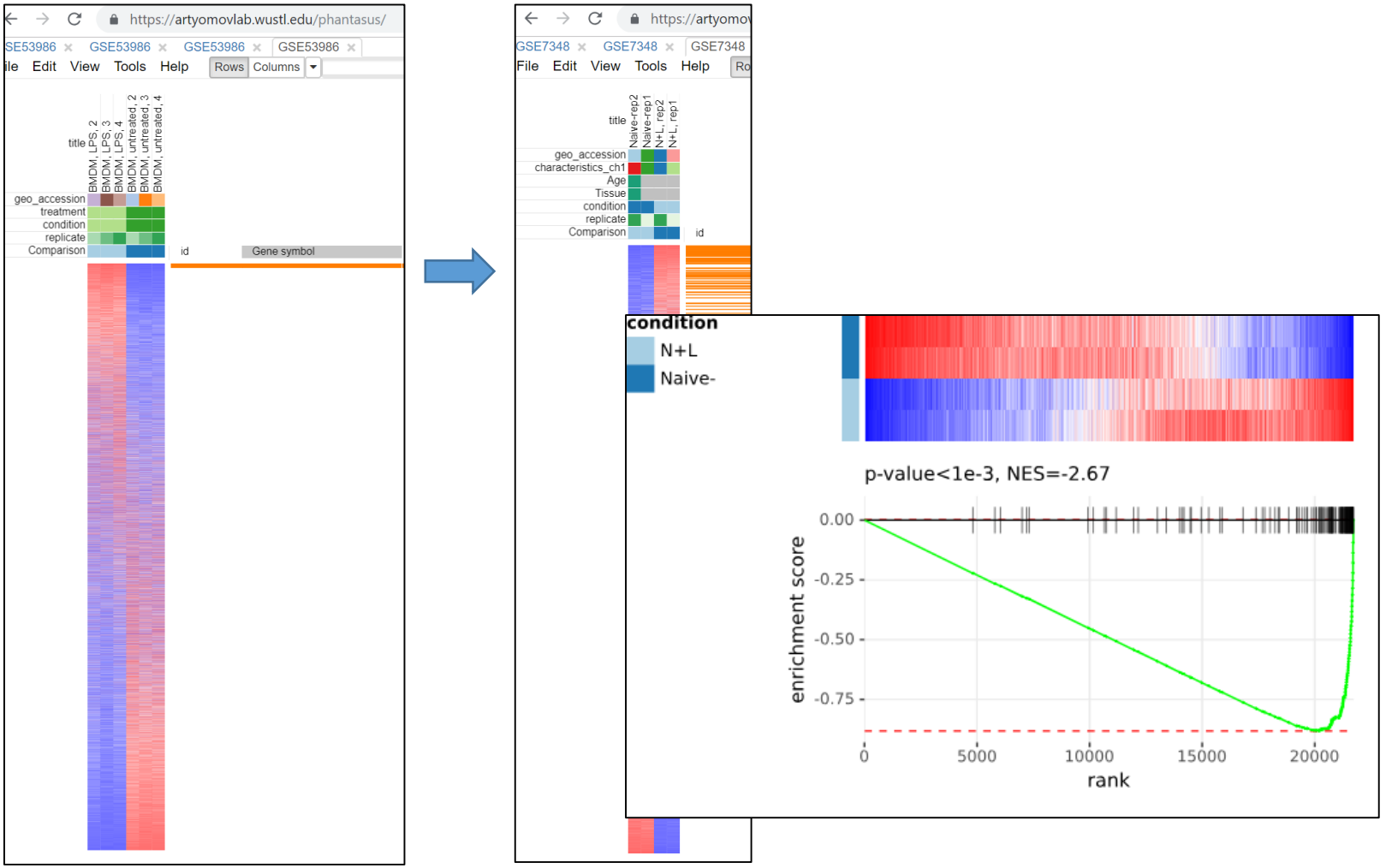
cited 860 times


https://artyomovlab.wustl.edu/phantasus/
GSE7348
GSE7348
GSE7348
File
Edit
View
Tools
Help
Rows
Columns
21,723 rows by 4 columns
title
Naive−rep2
Naive−rep1
N+L, rep2
N+L, rep1
geo_accession
characteristics_ch1
Age
Tissue
condition
replicate
Comparison
id
Gene symbol
Gene ID
Signal to noise

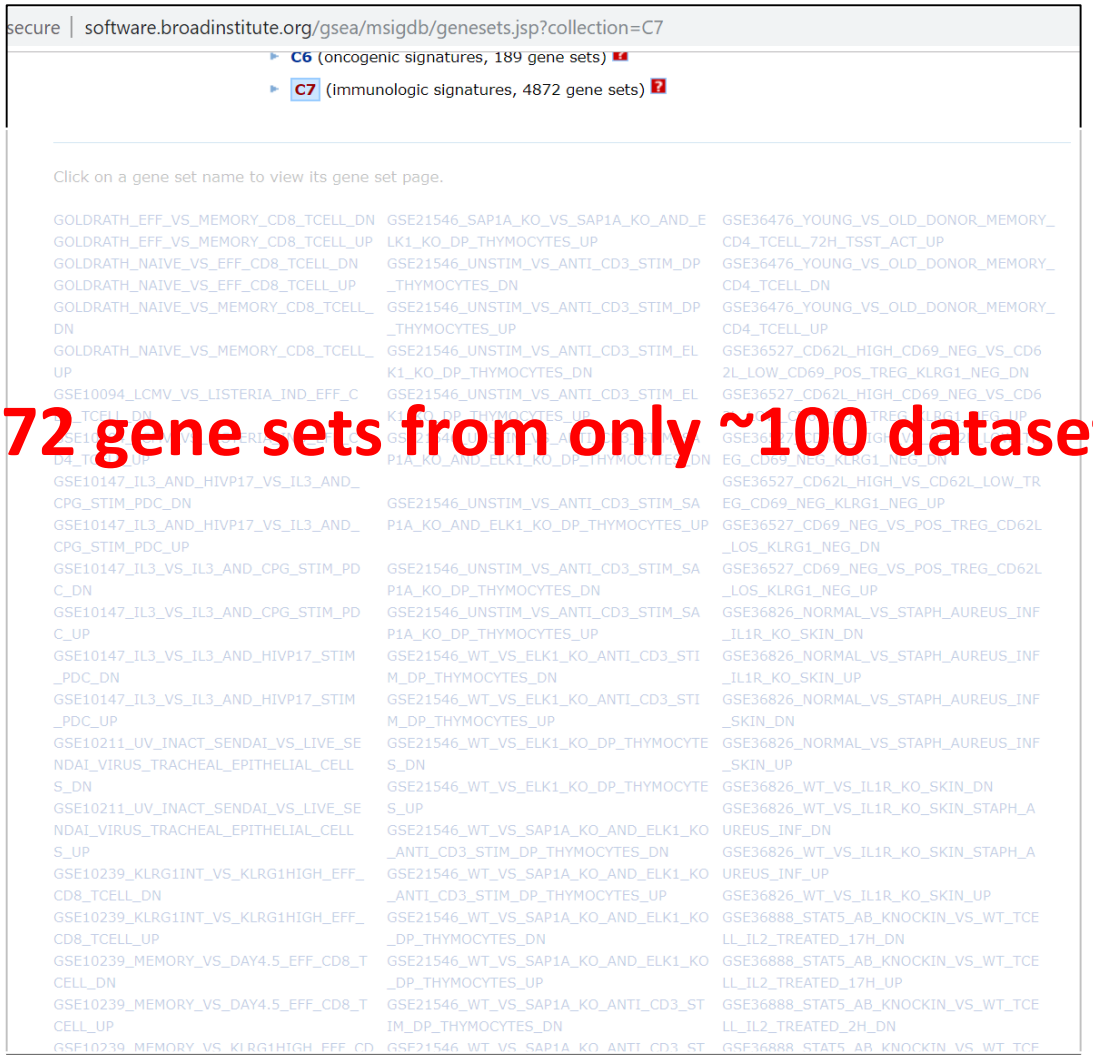
Differentially expressed genes behave very similar between two datasets!



Differentially expressed genes behave very similar between two datasets!

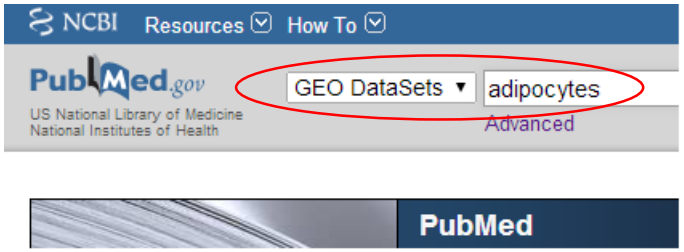


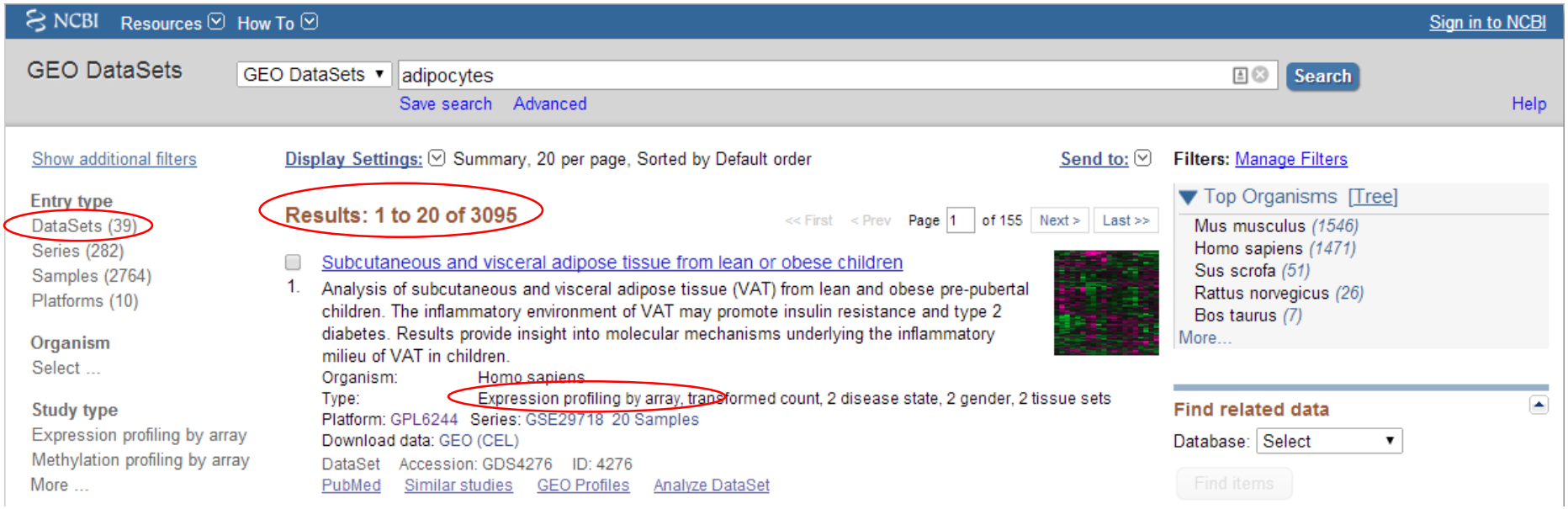
MSigDB.C7 – attempt to make such database



4872 gene sets from only ~100 datasets!!

GEO contains A LOT of datasets





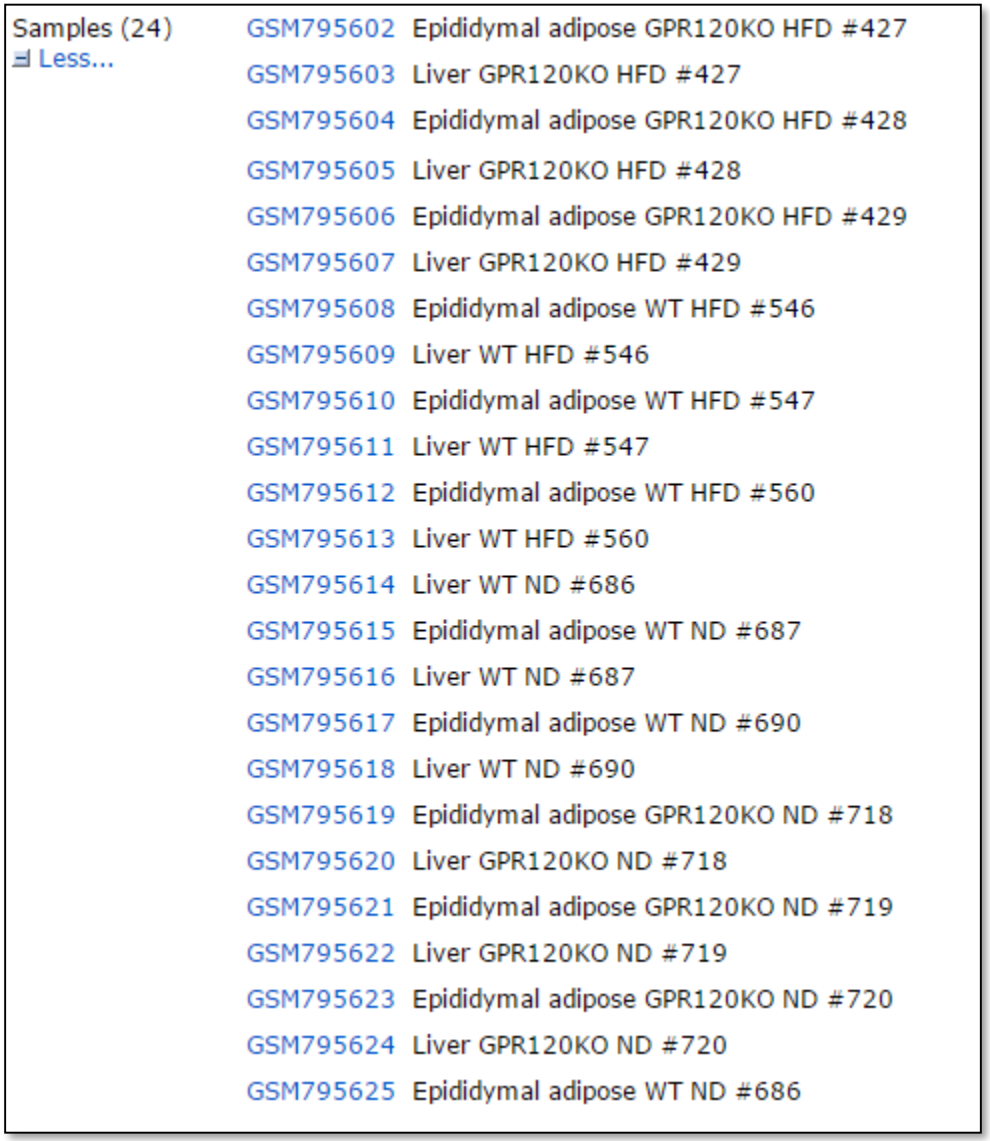
However, they are poorly annotated

Samples (24)
Less...

GSM795602	Epididymal adipose GPR120KO HFD #427
GSM795603	Liver GPR120KO HFD #427
GSM795604	Epididymal adipose GPR120KO HFD #428
GSM795605	Liver GPR120KO HFD #428
GSM795606	Epididymal adipose GPR120KO HFD #429
GSM795607	Liver GPR120KO HFD #429
GSM795608	Epididymal adipose WT HFD #546
GSM795609	Liver WT HFD #546
GSM795610	Epididymal adipose WT HFD #547
GSM795611	Liver WT HFD #547
GSM795612	Epididymal adipose WT HFD #560
GSM795613	Liver WT HFD #560
GSM795614	Liver WT ND #686
GSM795615	Epididymal adipose WT ND #687
GSM795616	Liver WT ND #687
GSM795617	Epididymal adipose WT ND #690
GSM795618	Liver WT ND #690
GSM795619	Epididymal adipose GPR120KO ND #718
GSM795620	Liver GPR120KO ND #718
GSM795621	Epididymal adipose GPR120KO ND #719
GSM795622	Liver GPR120KO ND #719
GSM795623	Epididymal adipose GPR120KO ND #720
GSM795624	Liver GPR120KO ND #720
GSM795625	Epididymal adipose WT ND #686

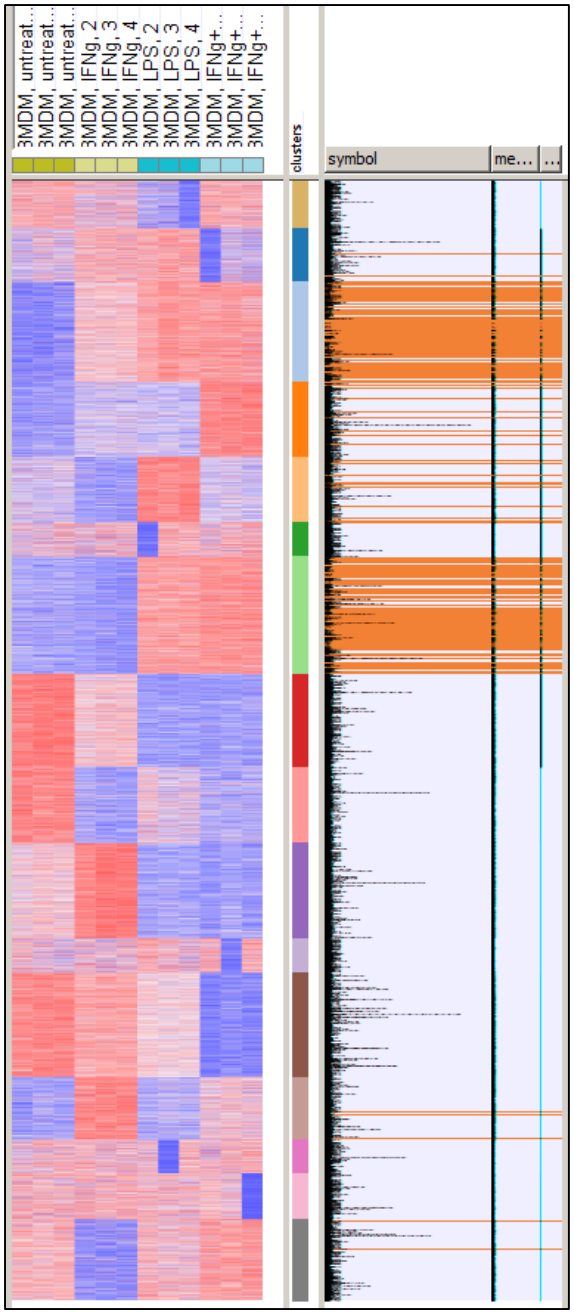
However, they are poorly annotated

It is virtually impossible to parse such entry without manual curation



Part II: how to get somewhere you haven’t been before

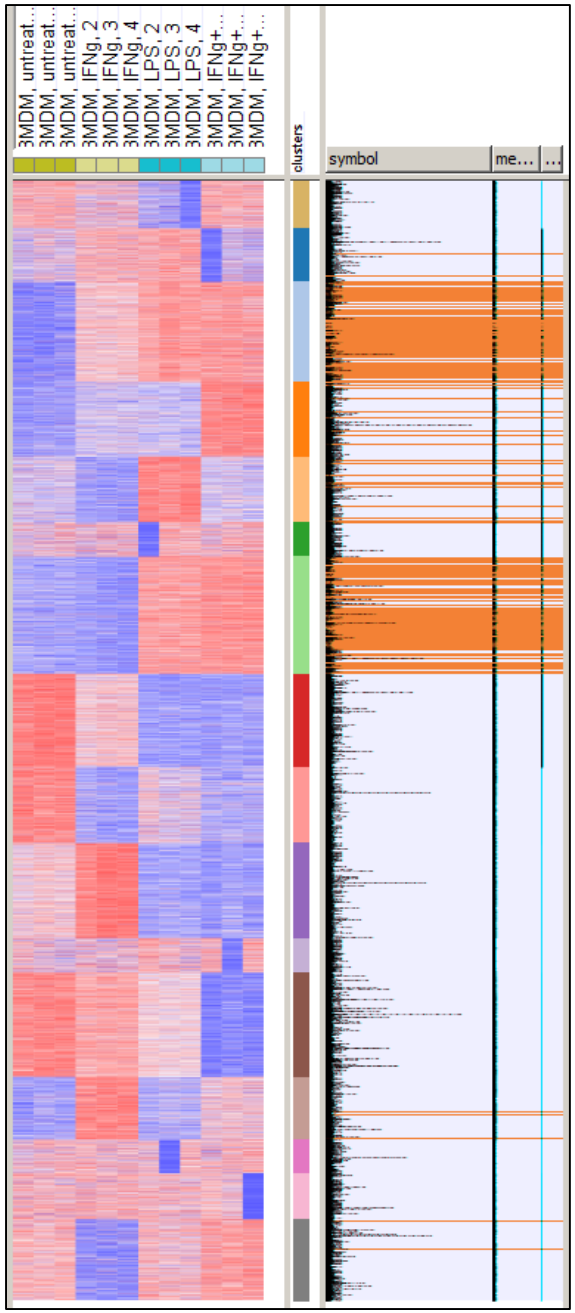
How it looks like in full dataset…



Genes upregulated by LPS in clustered heatmap

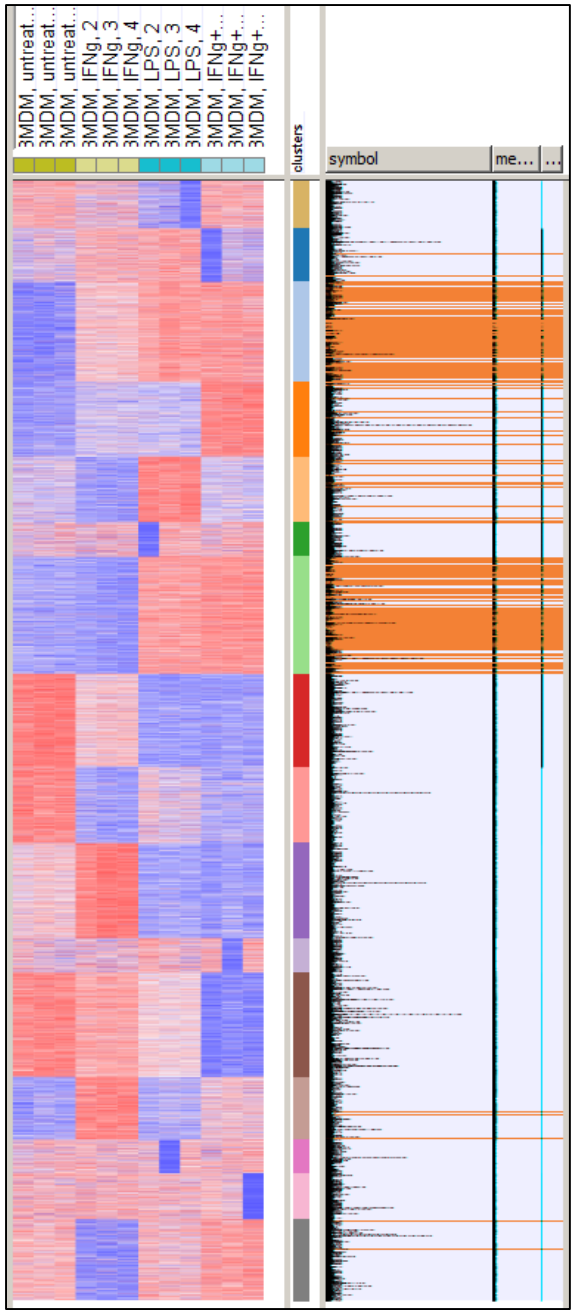
GSE53986
microarray data

Genes upregulated by LPS in your clustered heatmap

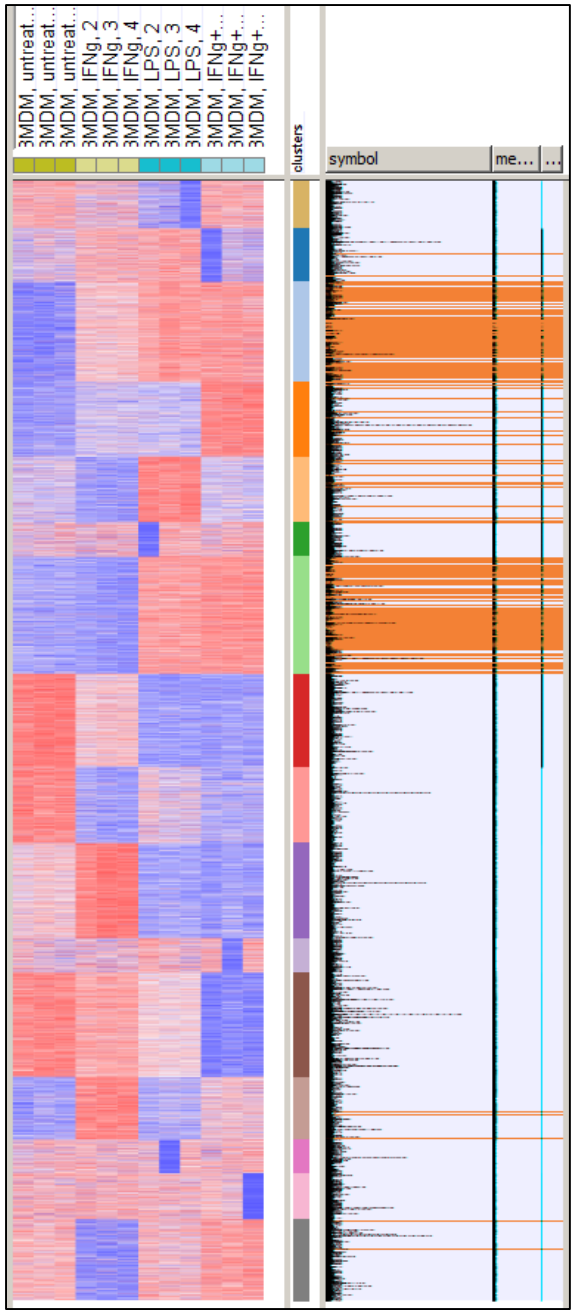


Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	100		$1.02\ e^{-120}$	$1.4\ e^{-117}$
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	51		$3.23\ e^{-63}$	$2.22\ e^{-60}$
REACTOME_IMMUNE_SYSTEM [933]	Genes involved in Immune System	102		$9.25\ e^{-50}$	$4.25\ e^{-47}$
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [270]	Genes involved in Cytokine Signaling in Immune system	41		$5.26\ e^{-26}$	$1.81\ e^{-23}$
REACTOME_INTERFERON_SIGNALING [159]	Genes involved in Interferon Signaling	33		$1.09\ e^{-25}$	$3.02\ e^{-23}$
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_EN_PROCESSING_PRESENTATION [251]	Genes involved in Class I MHC mediated antigen processing & presentation	39		$3.53\ e^{-25}$	$8.11\ e^{-23}$
REACTOME_INTERFERON_ALPHA_BETA_SIGNALIALING [64]	Genes involved in Interferon alpha/beta signaling	23		$2.3\ e^{-24}$	$4.53\ e^{-22}$
REACTOME_ADAPTIVE_IMMUNE_SYSTEM [539]	Genes involved in Adaptive Immune System	53		$5.85\ e^{-24}$	$1.01\ e^{-21}$
REACTOME_ANTIGEN_PROCESSING_UBIQUITINAINATION_PROTEASOME_DEGRADATION [212]	Genes involved in Antigen processing: Ubiquitination & Proteasome degradation	32		$1.66\ e^{-20}$	$2.55\ e^{-18}$
REACTOME_ER_PHAGOSOME_PATHWAY [61]	Genes involved	20		$1.96\ e^{-20}$	$2.7\ e^{-18}$

Genes upregulated by LPS in your clustered heatmap



Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]	Genes regulated by NF-kB in response to TNF [GeneID=7124].	45		$8.11 e^{-34}$	$1.12 e^{-30}$
HALLMARK_INFLAMMATORY_RESPONSE [200]	Genes defining inflammatory response.	44		$1.21 e^{-32}$	$8.31 e^{-30}$
REACTOME_IMMUNE_SYSTEM [933]	Genes involved in Immune System	86		$4.15 e^{-32}$	$1.91 e^{-29}$
HALLMARK_MTORC1_SIGNALING [200]	Genes up-regulated through activation of mTORC1 complex.	36		$9.36 e^{-24}$	$3.23 e^{-21}$
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [270]	Genes involved in Cytokine Signaling in Immune system	37		$3.79 e^{-20}$	$1.05 e^{-17}$
HALLMARK_HYPOXIA [200]	Genes up-regulated in response to low oxygen levels (hypoxia).	32		$1.16 e^{-19}$	$2.67 e^{-17}$
REACTOME_SIGNALING_BY_ILS [107]	Genes involved in Signaling by Interleukins	24		$1.24 e^{-18}$	$2.44 e^{-16}$
REACTOME_ADAPTIVE_IMMUNE_SYSTEM [539]	Genes involved in Adaptive Immune System	49		$1.73 e^{-18}$	$2.99 e^{-16}$
HALLMARK_GLYCOLYSIS [200]	Genes encoding proteins involved in glycolysis and gluconeogenesis.	27		$6.49 e^{-15}$	$9.94 e^{-13}$
HALLMARK_UV_RESPONSE_UP [158]	Genes up-regulated in response to ultraviolet (UV) radiation.	23		$1.27 e^{-13}$	$1.76 e^{-11}$
HALLMARK_HEME_METABOLISM [200]	Genes involved in metabolism of heme (a cofactor consisting of iron and porphyrin) and erythroblast differentiation.	25		$3.84 e^{-13}$	$4.81 e^{-11}$

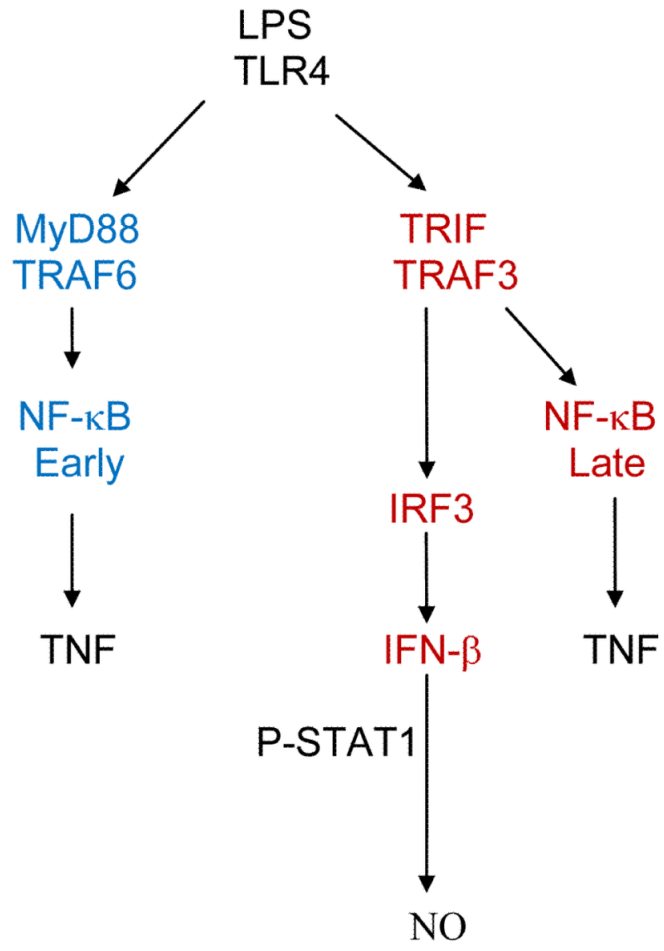


Genes upregulated by LPS in clustered heatmap

Ifng responsive inflammatory genes

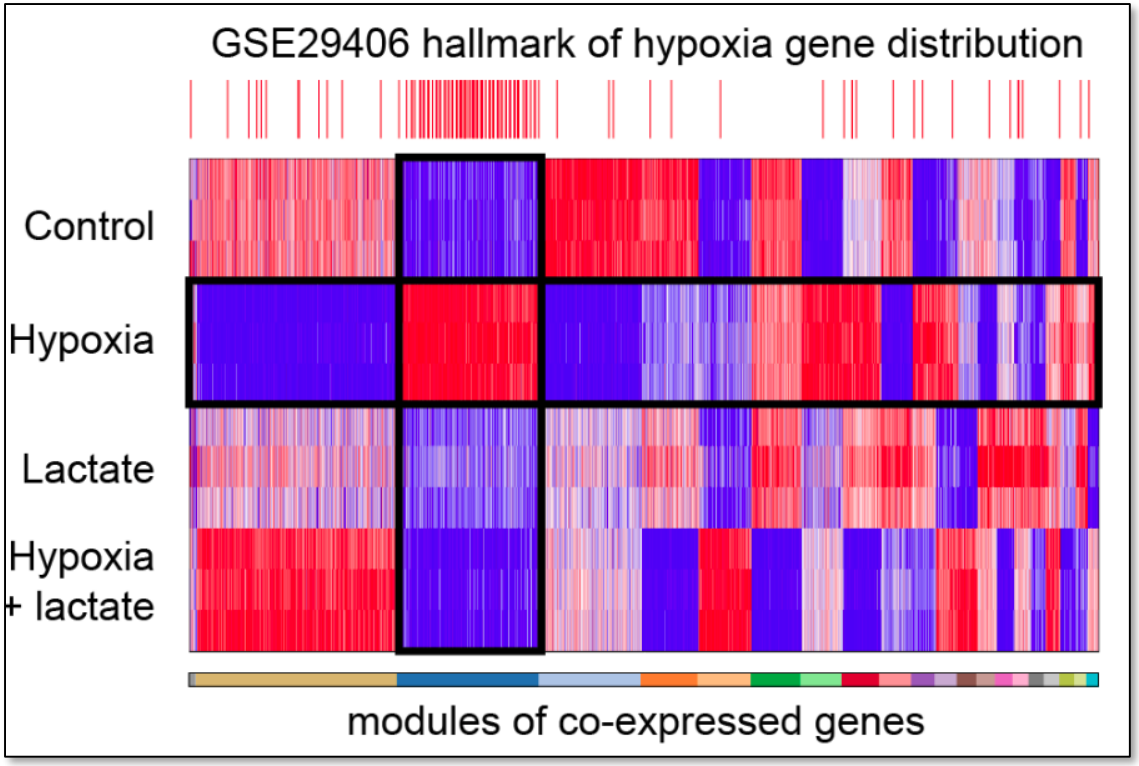
Ifng non-responsive inflammatory genes

These two transcriptional modules are due to two regulatory modes downstream TLR4



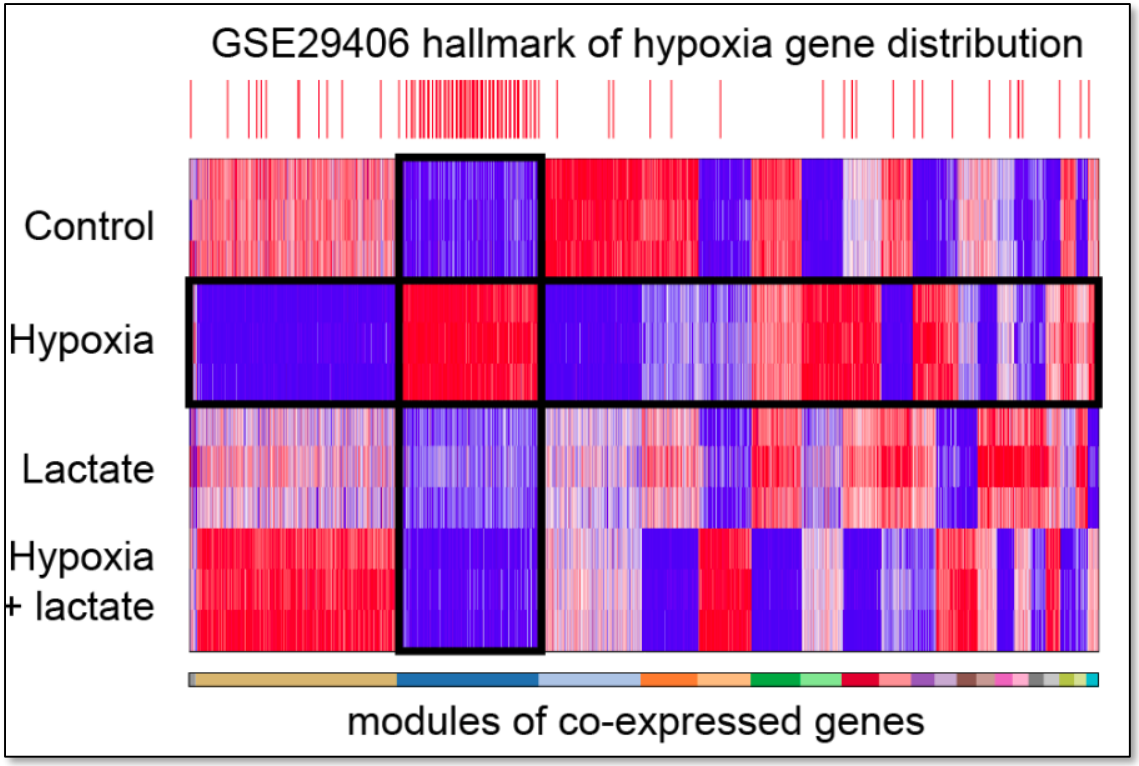
Observation:
biologically meaningful gene sets
are typically co-expressed in relevant datasets!!

Observation:
biologically meaningful gene sets
are typically co-expressed in relevant datasets!!



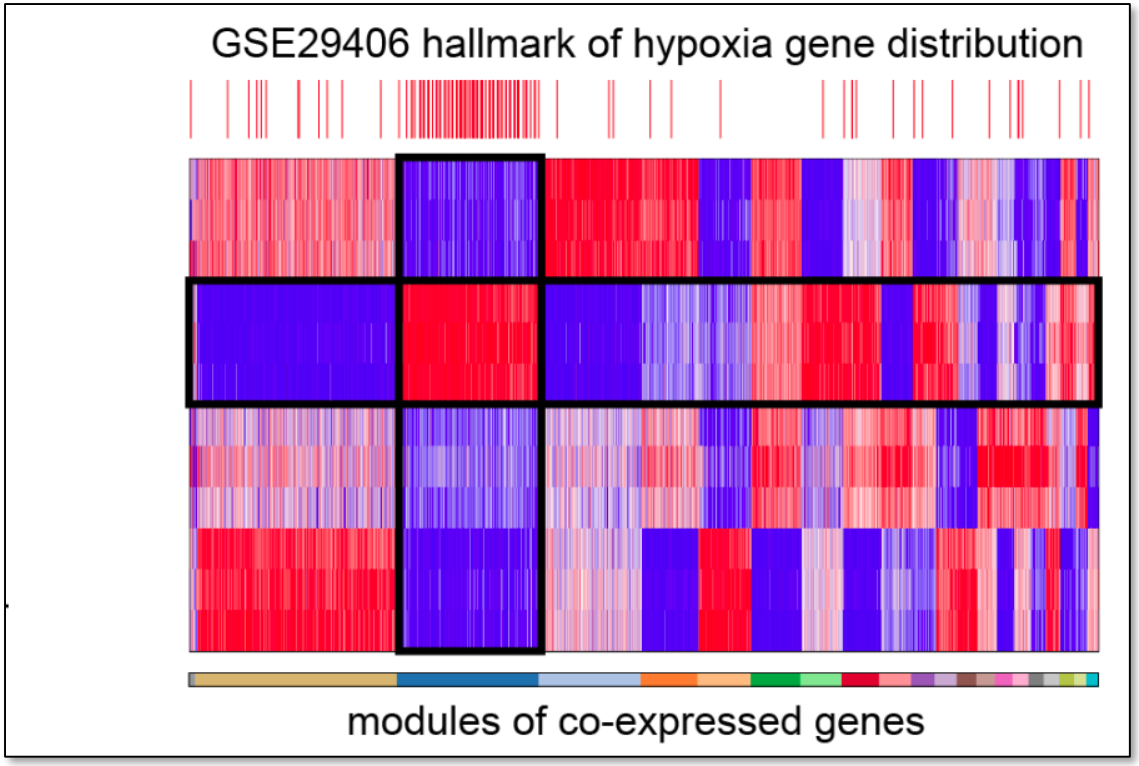
Predeus et al, submitted 2017

Why is it important:
Co-expression clustering does NOT require knowledge of conditions



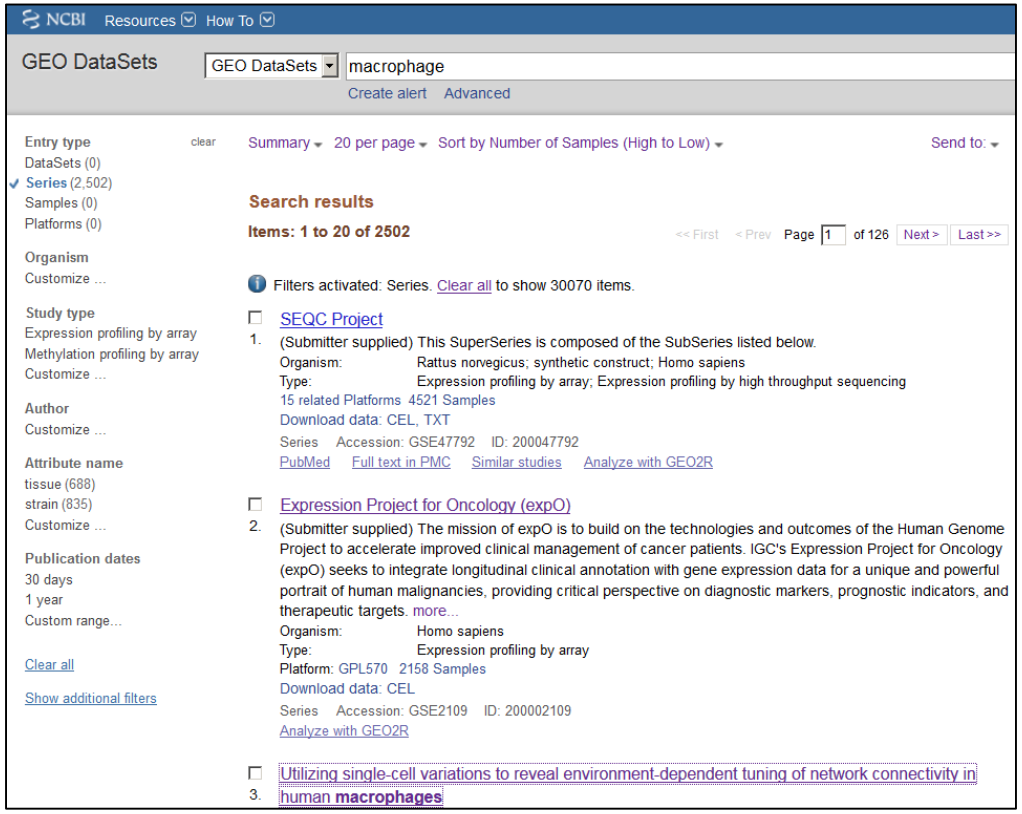
Predeus et al, submitted 2017

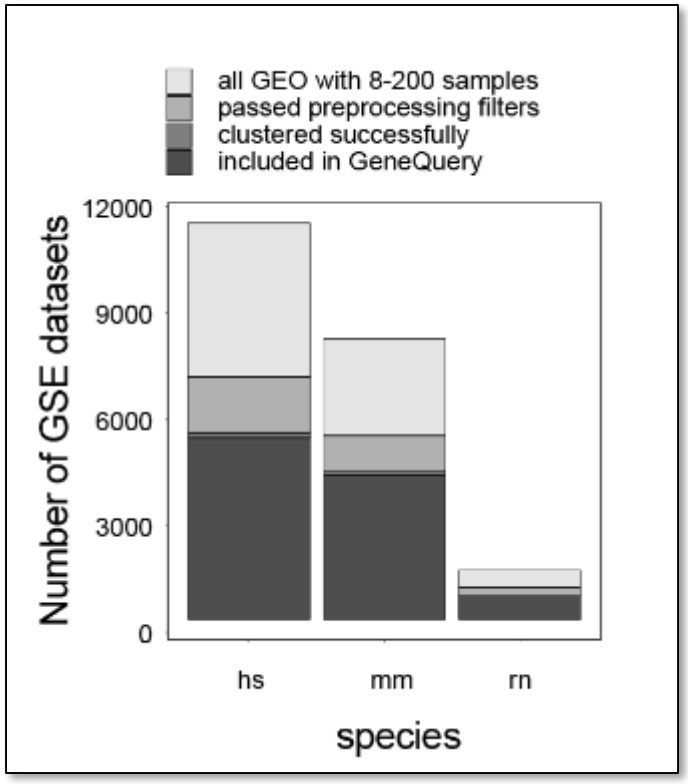
Why is it important:
Co-expression clustering does NOT require knowledge of conditions



Predeus et al, submitted 2017

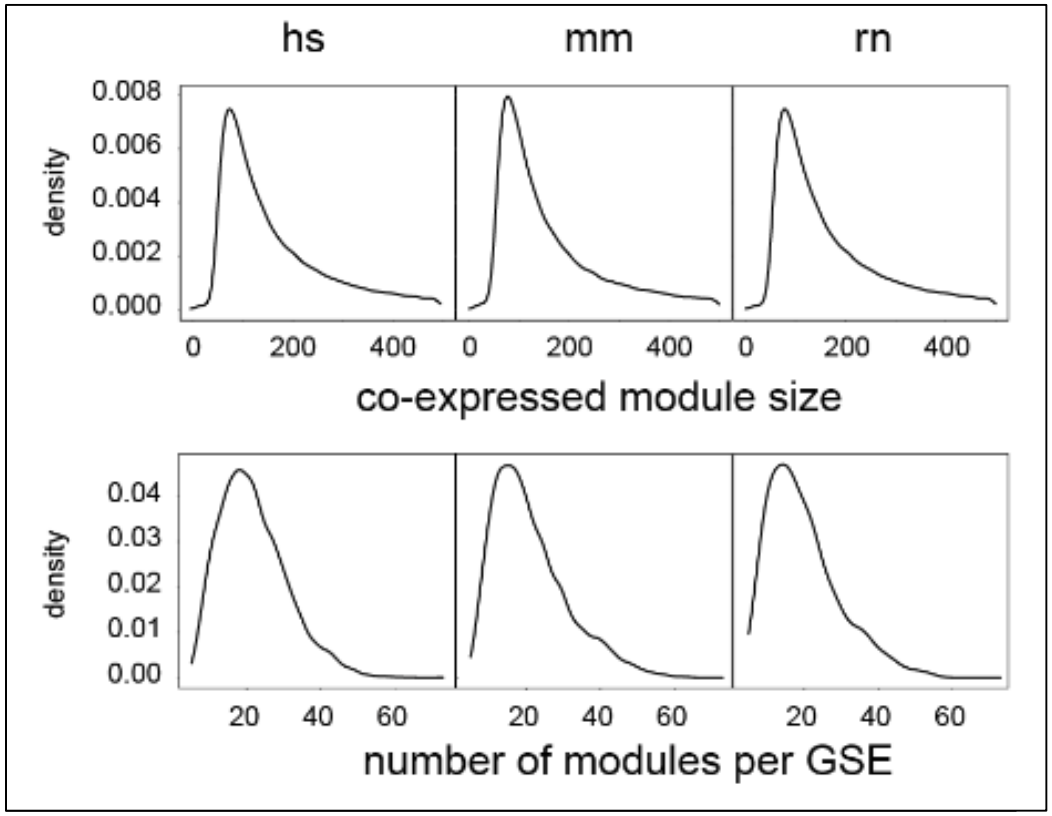
You can cluster whole GEO!!

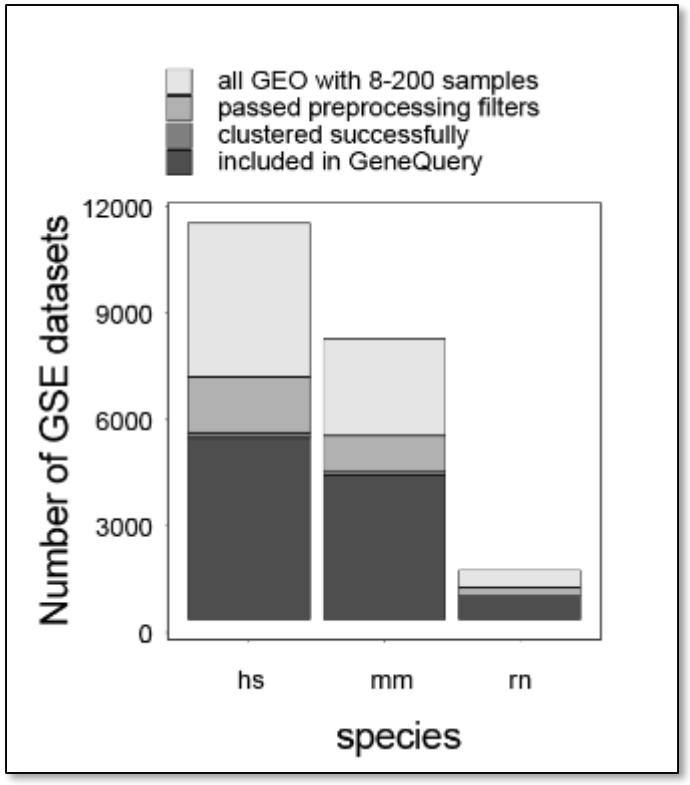




Predeus et al, submitted 2017

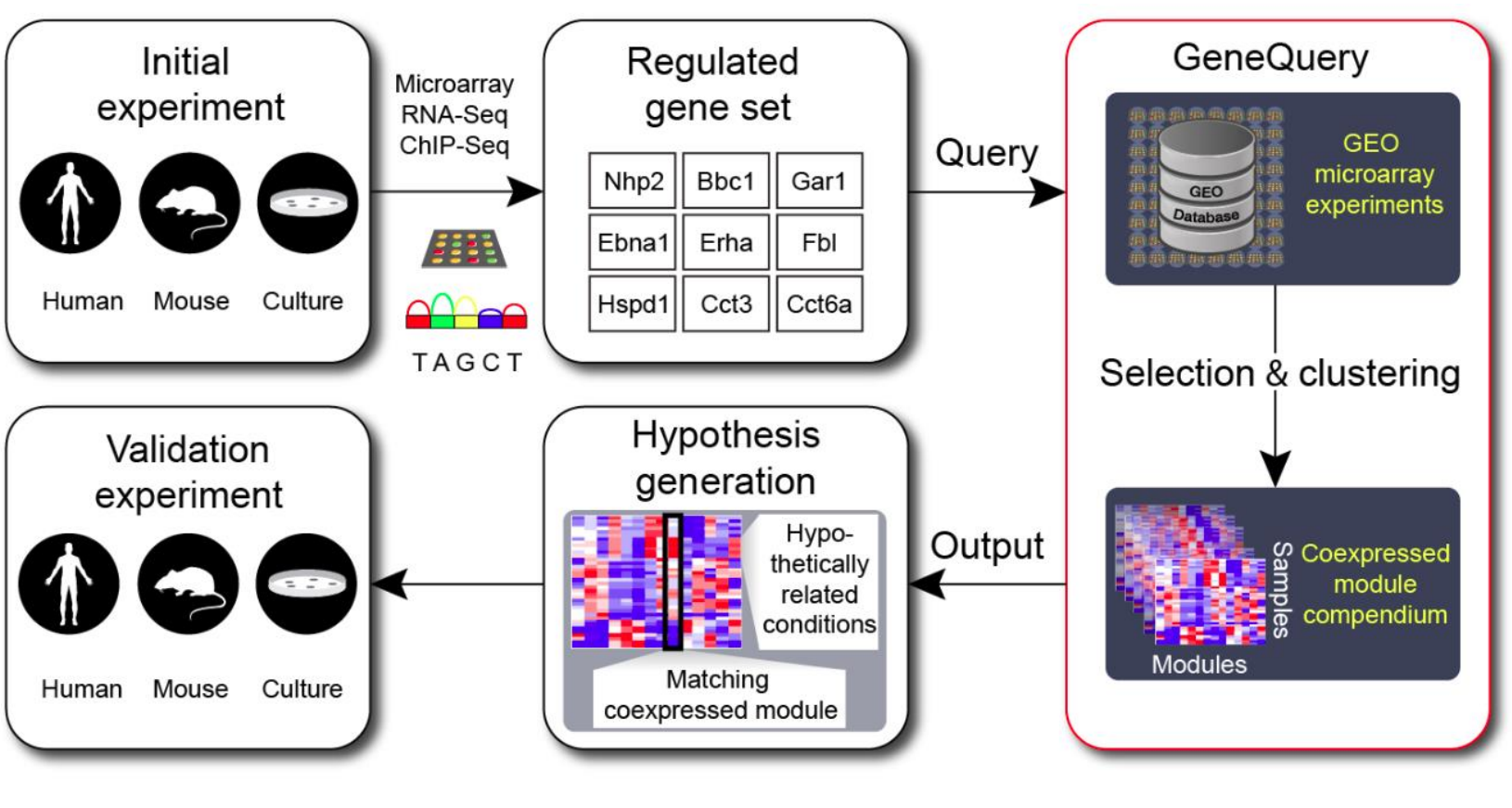
You can cluster whole GEO!!





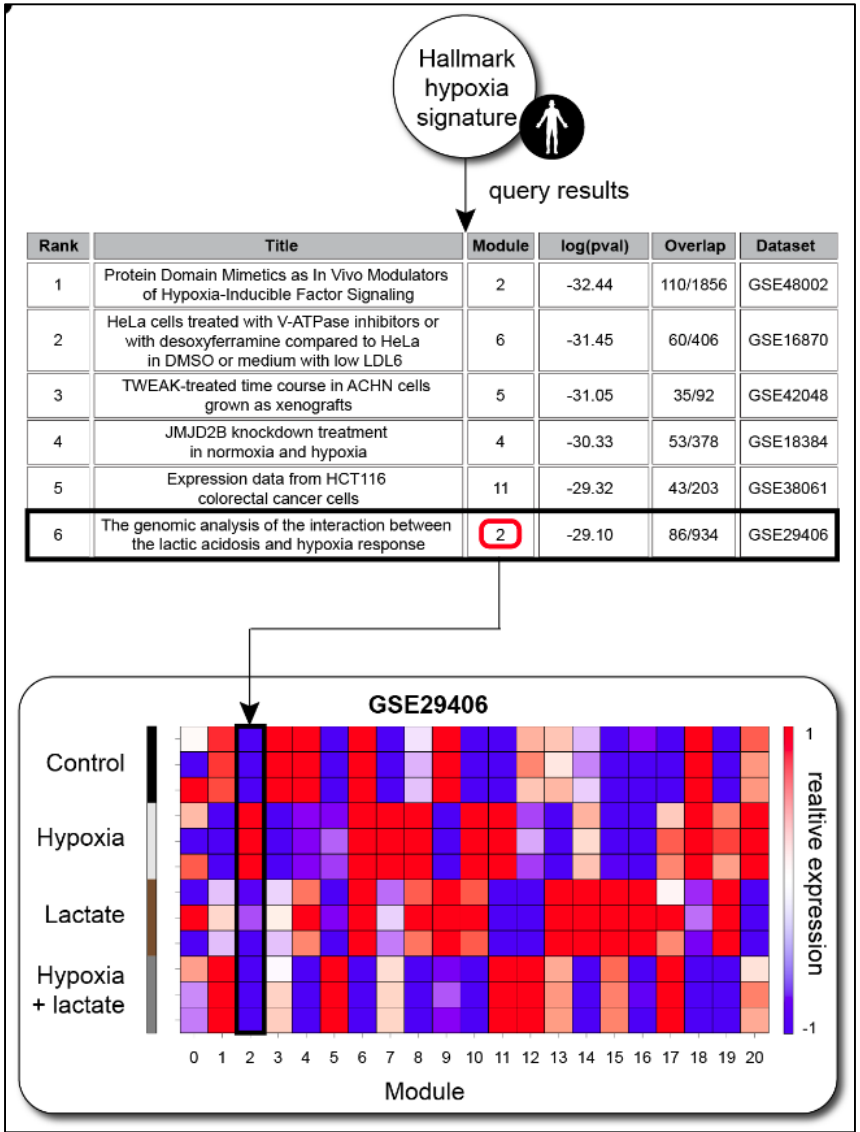
Predeus et al, submitted 2017

Get the idea for your next experiment!



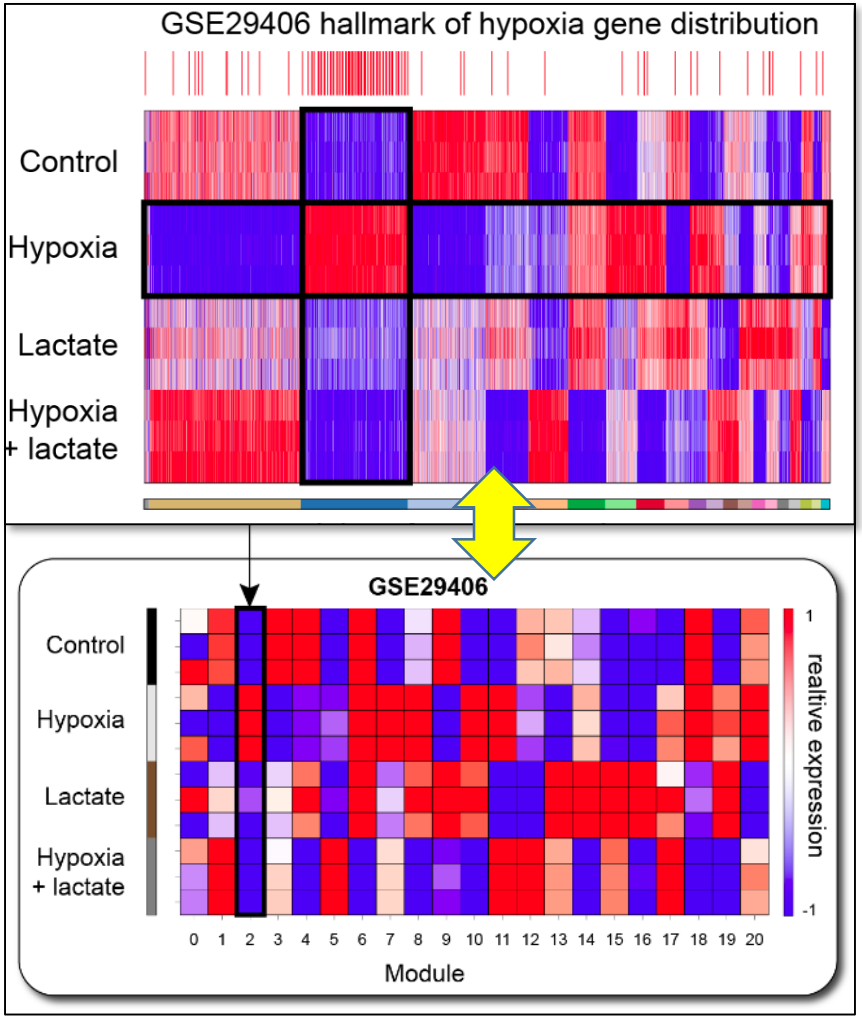
Get the idea for your next experiment!

http://artyomovlab.wustl.edu/genequery/searcher/
http://genome.ifmo.ru/genequery/searcher/



Get the idea for your next experiment!

http://artyomovlab.wustl.edu/genequery/searcher/
http://genome.ifmo.ru/genequery/searcher/



Eigengene is averaged representation of the cluster behavior

Let’s practice with our inflammatory macrophage datasets!

Explore gmt

Let’s practice with some hallmark signatures!