Bioinformatics discovery of new disease drivers from genomic, proteomics, and metabolomics data

Andrew P. Hodges, Ph.D.
ahodges@sbpdiscovery.org
SBP
3 May 2016

SBP
SANFORD BURNHAM PREBYS

Viper-KR*

- **Viper-KR** applied to human tissue (obesity / adipose tissue) and cell culture transcriptome.
- Data from the Translational Research Institute (TRI) for Metabolism and Diabetes, SBP-Florida Hospital (PI Steven Smith)
- The results are promising: multiple computationally predicted gene targets were validated experimentally (7 out of 12)

Viper - Virtual Inference of Protein-activity by Enriched Regulon analysis (Mariano J. Alvarez, Federico Giorgi, Andrea Califano). KR – K-nearest neighbor and Random Forest.

Overview

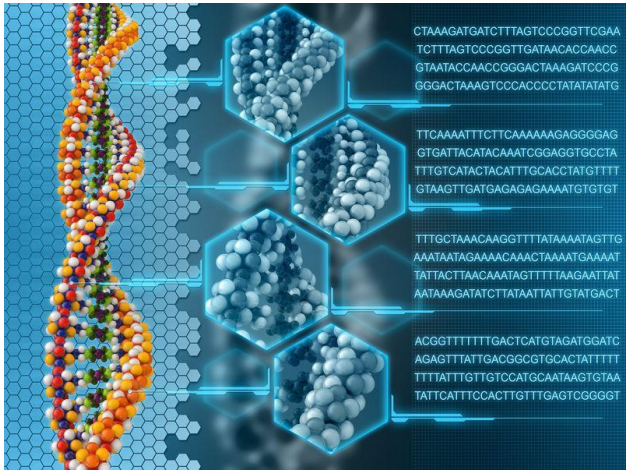
- Pt1 - Introduction/Background
- Pt2 - Resources
  - Public and proprietary data sources
  - Selected software and underlying algorithms
- Pt3 – Example Service: Master regulator/driver predictions
  - Example models from in vitro and clinical studies
  - New gene/protein drivers for follow-up confirmation
  - Integration against knowledge repositories/other resources

SBP
SANFORD BURNHAM PREBYS

Pt1 - Introduction

SBP
SANFORD BURNHAM PREBYS

Bioinformatics



- Major driver in current cancer research
- Oftentimes an afterthought
- *Should be the driver/guide for your studies!

- Quest for drivers/controllers of disease
  - Not cancer-specific… conceptually any disease
  - Molecular signatures guiding disease processes and drug responses
  - Limited by type(s) of data & available samples

NCI projects (2015)

SBP
SANFORD BURNHAM PREBYS

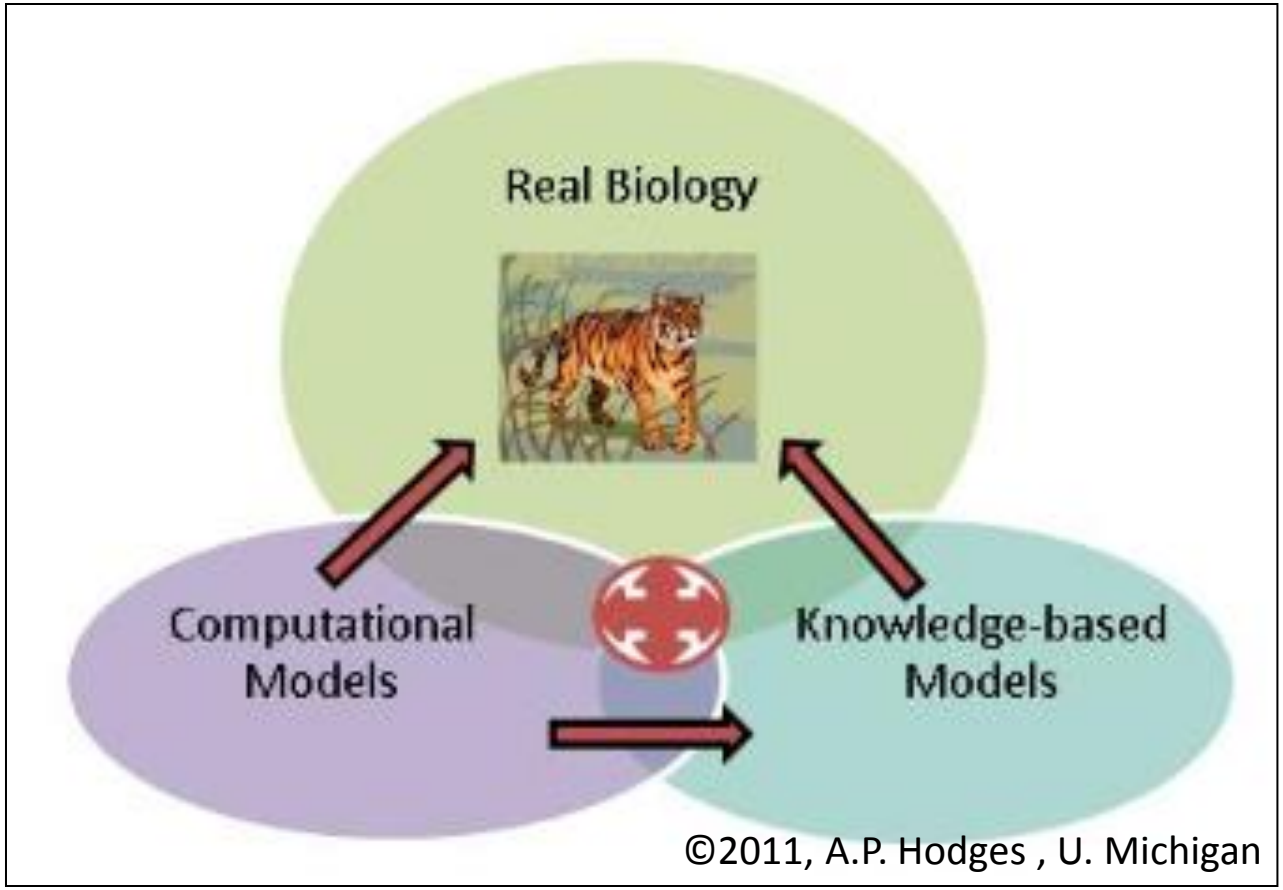
Major question today:

What are the drivers of pancreatic cancer progression in gemcitabine-treated patients?

SBP
SANFORD BURNHAM PREBYS

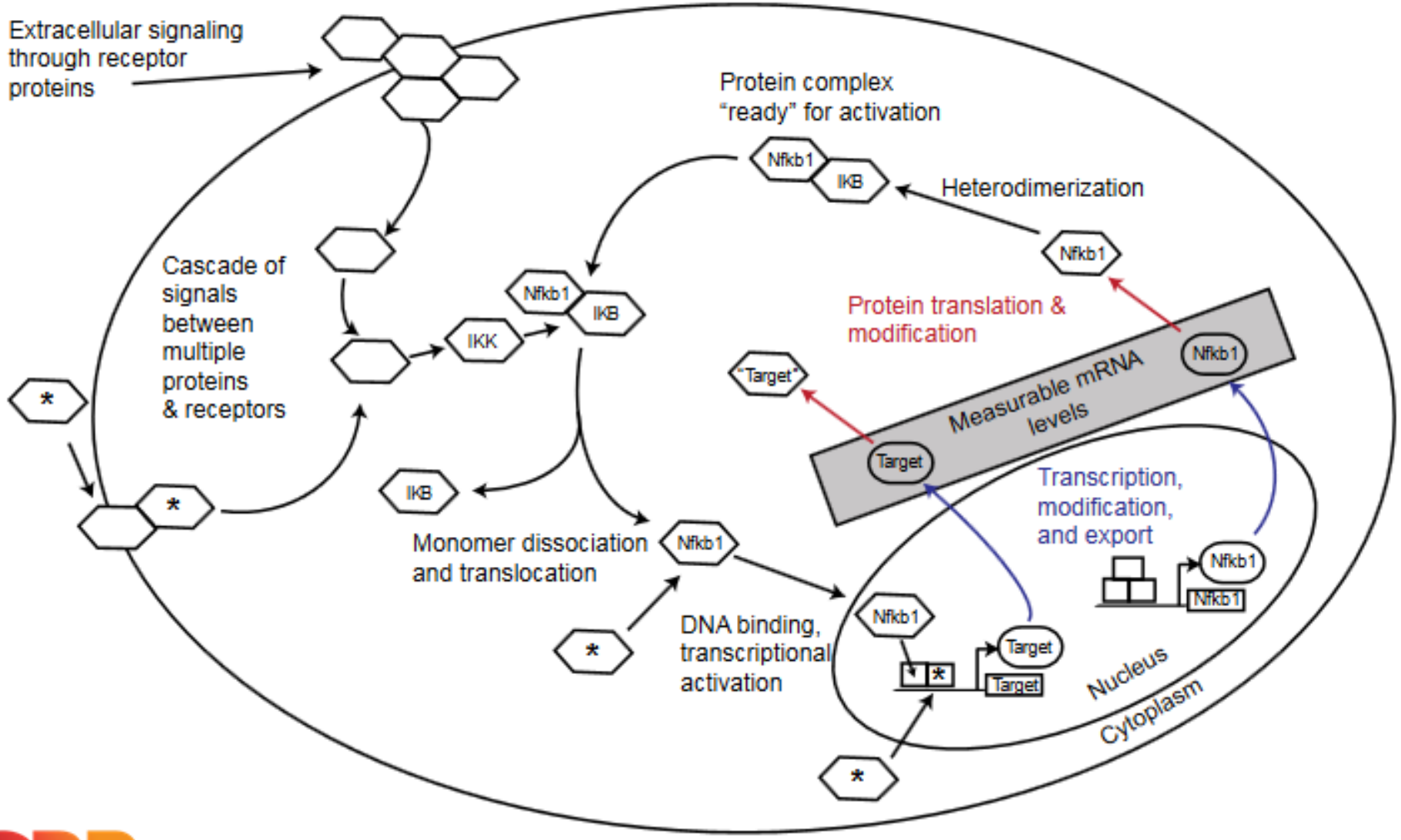
“Standing on two legs“

-A. Eroshkin



SBP
SANFORD BURNHAM PREBYS

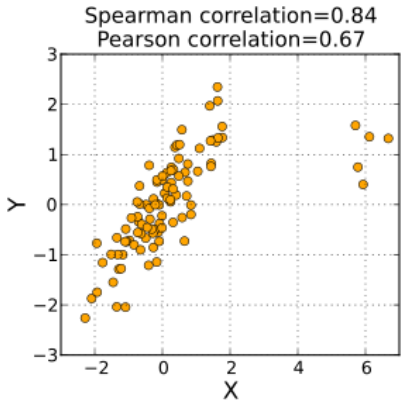
General framework for modeling (with opportunities & caveats): Transcriptional network inference



Machine learning: m*oving towards identifying drivers/controllers of processes & diseases*

- Naïve classic approach:
  - What are the best correlated interactors with my gene?

Pearson correlation (distance)
< Spearman correlation (rank)
< Mutual information (probability)



- Supervised & unsupervised learning
  - Is some prior knowledge used to guide/train predictions?
  - Identify factors influencing overall data behavior (e.g. PCA/dimensional redu.)
  - Many options for modeling:
    - SVMs, flux models, neural nets, ODE/PDE, fuzzy logic clustering, MI/BN models, ABMs, etc. --

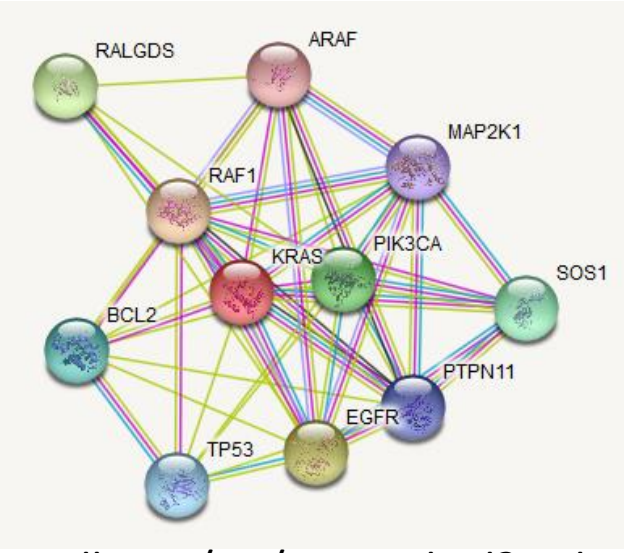
- Network models: can represent results from these approaches (e.g. Viper)

SBP
SANFORD BURNHAM PREBYS

Networks & centrality

- Assumption that ‘drivers’ are correlated with some disease phenotype
- Often assumed to be ‘highly connected’ to other pathway elements
- Hairball view… not interpretable!

- Network lingo:
  - Node/circle : represents gene/protein/molecular entity in some dataset/database
  - Edge/line/arrow: represents some interaction or inference



Knowledge-based and/or data-driven?

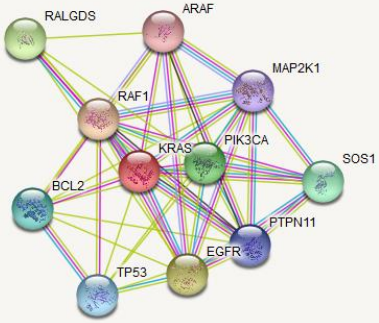
http://string-db.org/cgi/network.pl?taskId=2_rT3n5wjcWL

SBP
SANFORD BURNHAM PREBYS

- Edges can represent either known or putative interaction
  - Depends on knowledge database and/or algorithm
  - Type of interaction or inference can vary:

- **Know the interactions' provenance
  - Also, experimental/biological context of interactions? Instrumentation?

Neighborhood	Gene Fusion	Cooccurence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
				●	●	●		0.999
				●	●	●		0.999
				●	●	●		0.998
				●		●		0.997
				●	●	●		0.993



Node Size

small nodes:
protein of unknown 3D structure

large nodes:
some 3D structure is known or predicted

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

Others

textmining

co-expression

protein homology

SBP
SANFORD BURNHAM PREBYS

Pt2 – Some computational resources

SBP
SANFORD BURNHAM PREBYS

Recommended tools (DIY!):

https://www.bioconductor.org/

http://www.broadinstitute.org/cancer/software/genepattern/

http://www.cytoscape.org/

http://genomespace.org

D3.js (http://d3js.org)

http://basespace.illumina.com

http://www.metaboanalyst.ca/

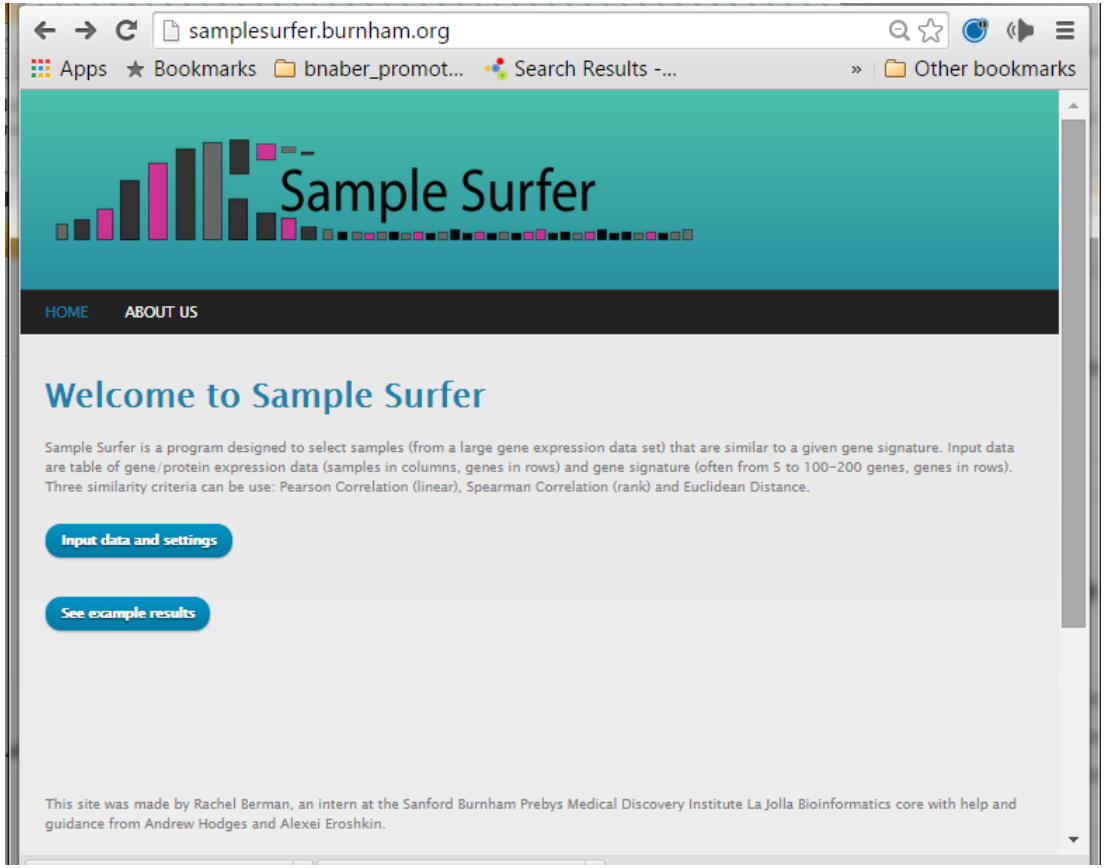
http://firebrowse.org/

***Google it! – thousands of resources available on the web (Github also)

New tools/services we offer

- Computational pipeline for **CRISPR/Cas9** knockout screens (our development) - from reads to “differentially affected sgRNA”
- **Viper-KR -** identifies master regulator genes
- **MetaboAnalyst** – integrative analysis of metabolic and genomic data
- **ChAMP** - Methylation data analysis pipeline
- **Our tool: Sample Surfer*** - selects samples (from a large gene expression data set) that are similar to a given gene/protein signature
- **Our tool: Regulattice*** - interactive analysis environment to analyze regulatory modules identified through machine learning, e.g., visualization of VIPER results
- **FireBrowse** - A simple and elegant way to explore cancer data.

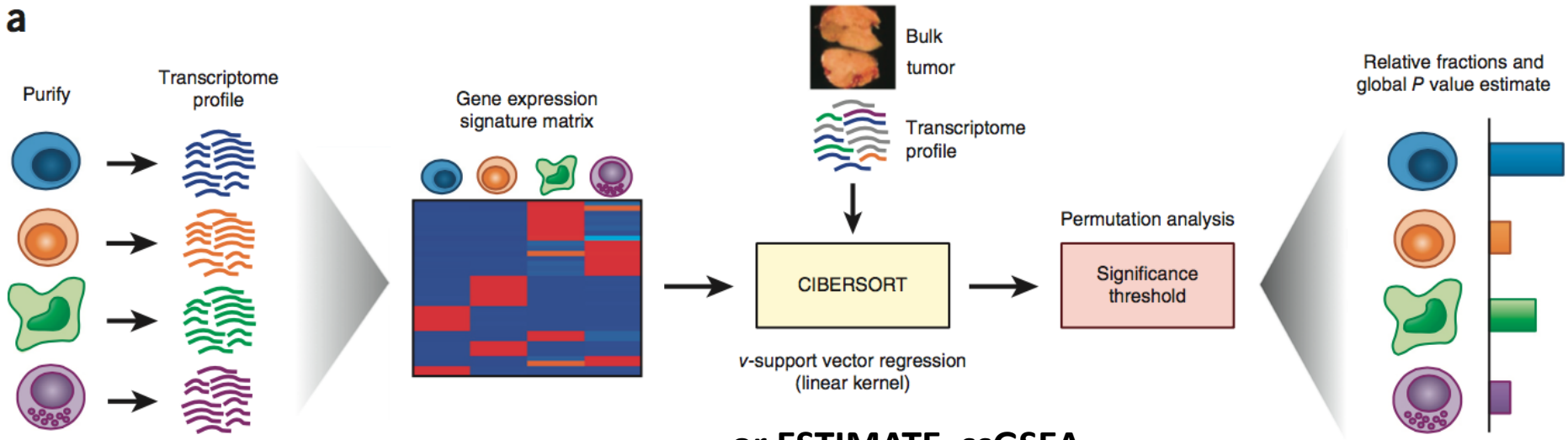
http://samplesurfer.burnham.org - intranet
http://samplesurfer.SBPdiscovery.org - outside of SBP



selects samples (from a large gene expression data set) that are similar to a given gene/protein signature

Godzik lab: How to measure host immune response from expression data?

Courtesy: A. Leblanc, Godzik Lab



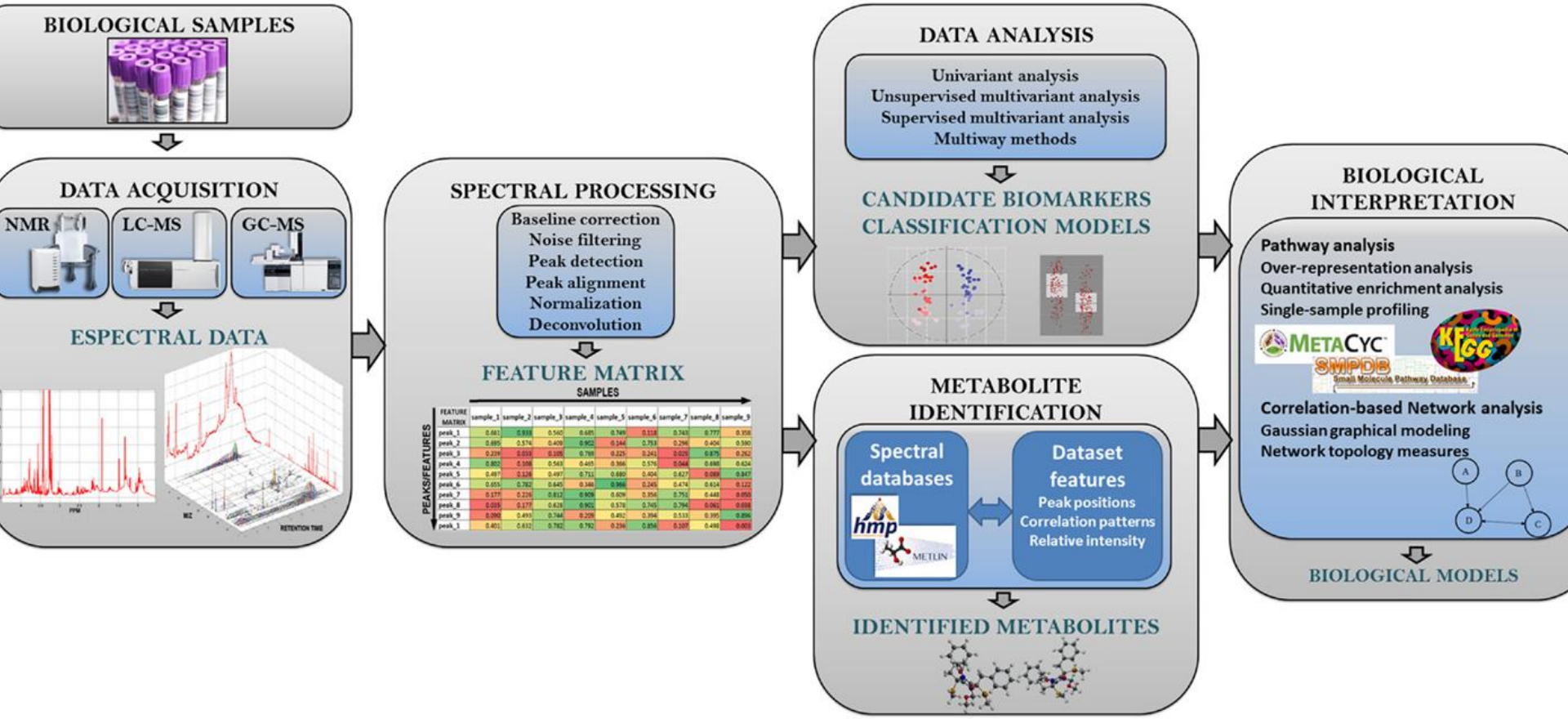
or ESTIMATE, ssGSEA, PERT, Cssam, CellMix

Modified figure from Newman et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 12:453-7.

SBP SANFORD BURNHAM PREBYS

**Also noteworthy: Cancer3D for identifying cancer driver mutations (structural basis) and drug targeting (http://cancer3d.org)

Metabolomics basic example:



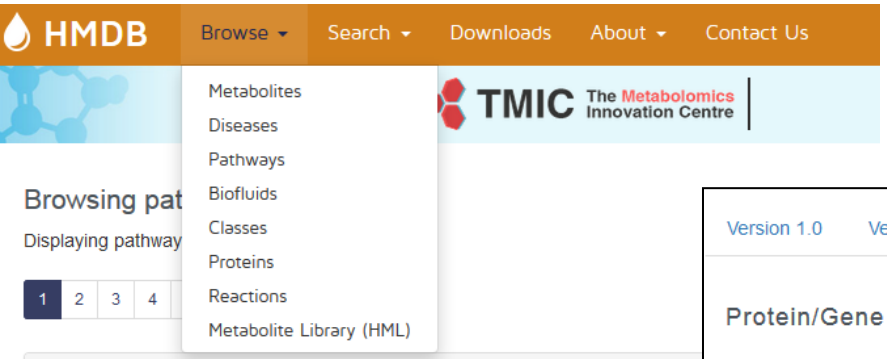
Analytical methods in untargeted metabolomics: state of the art in 2015
http://journal.frontiersin.org/article/10.3389/fbioe.2015.00023/full

SBP
SANFORD BURNHAM PREBYS

Tool	Type	Target	Features[a]	Website	Reference
MetaboAnalyst2	Web	MS and NMR	1–7	http://www.metaboanalyst.ca/	Xia et al. (2012)
XCMS	R	MS	1–3	http://metlin.scripps.edu/xcms/	Smith et al. (2006)
MetSign	MatLab	MS	1–3	http://metaopen.sourceforge.net/	Lommen and Kools (2012)
XCMS online	Web	LC-MS	1–4	https://xcmsonline.scripps.edu/	Tautenhahn et al. (2012b)
MAVEN	Application	LC-MS	1–7	http://genomics-pubs.princeton.edu/mzroll	Melamud et al. (2010)
mzMine2	Application	LC-MS	1–5	http://mzmine.sourceforge.net/	Pluskal et al. (2010)
MAIT	R	LC-MS	1–5	http://b2slab.upc.edu/software-and-downloads	Fernández-Albert et al. (2014)
OpenMS	Application	LC-MS	1–3	http://open-ms.sourceforge.net/	Sturm et al. (2008)
Metabolome express	Web	GC-MS	1–5	https://www.metabolome-express.org/	Carroll et al. (2010)
Metabolite detector	Application	GC-MS	1–4	http://md.tu-bs.de/	Hiller et al. (2009)
MetDAT	Web	MS	1–5	http://smbl.nus.edu.sg/METDAT2/	Biswas et al. (2010)
FOCUS	MatLab	NMR	1–4	http://www.urr.cat/FOCUS/	Alonso et al. (2013)
Automics	Application	NMR	1–2, 5	https://code.google.com/p/automics/	Wang et al. (2009)
Bayesil	Web	NMR	1–4	http://bayesil.ca/	Ravanbakhsh et al. (2014)
Speaq	Application	NMR	1–2, 5	https://code.google.com/p/speaq/	Vu et al. (2011)
MetaboLab	Application	NMR	1–2, 5	http://www.nmrlab.org.uk/	Ludwig and Gunther (2011)
rNMR	R	NMR	8	http://rnmr.nmrfam.wisc.edu/	Lewis et al. (2009)
MetaboMiner	Application	NMR	8	http://wishart.biology.ualberta.ca/metabominer/	Xia et al. (2008)
Muma	R	–	5	http://cran.r-project.org/web/packages/muma	Gaude et al. (2013)
MetaXCMS	R	MS and NMR	5	http://metlin.scripps.edu/metaxcms/	Tautenhahn et al. (2010)
BATMAN	R	NMR	3–4	http://batman.r-forge.r-project.org/	Hao et al. (2012)
AStream	R	LC-MS	4	http://www.urr.cat/AStream/AStream.html	Alonso et al. (2011)
Camera	R	LC-MS	4	http://metlin.scripps.edu/xcms/	Kuhl et al. (2011)
MetaboHunter	Web	NMR	4	http://www.nrcbioinformatics.ca/metabohunter/	Tulpan et al. (2011)
MetScape	Application	–	6–7	http://metscape.ncibi.org/	Gao et al. (2010)
IMPaLA	Web	–	6–7	http://impala.molgen.mpg.de/	Kamburov et al. (2011)
MetExplore	Web	–	6–7	http://metexplore.toulouse.inra.fr/	Cottret et al. (2010)
MetPA	Web	–	6–7	http://metpa.metabolomics.ca/	Xia and Wishart (2010a)
Cytoscape	Application	–	7	http://www.cytoscape.org/	Smoot et al. (2011)
Vanted	Application	–	7	http://vanted.ipk-gatersleben.de/	Rohn et al. (2012)
Paintomics	Web	–	7	http://www.paintomics.org/	García-Alcalde et al. (2011)

[a] This table provides a complete and updated list of the open-source software that is commonly used in the untargeted analysis of metabolomic data. This column refers to the features included in the tool: spectral pre-processing (1), spectral/peak alignment (2), peak detection (3), metabolite identification (4), data analysis (5), pathway analysis (6), pathway visualization (7), and 2D-NMR analysis (8).

Human Metabolite DB: HMDB (www.hmdb.ca/downloads)



Useful annotation source

Rival identifier to PubChem ID, ChEBI, compound name, etc.

Used in MetaboAnalyst

Version 1.0 | Version 2.5 | Current Version (3.6)

Protein/Gene Sequences (in FASTA Format)

Data Set	Released on	Protein Sequences	Gene Sequences
All Metabolite Metabolizing Enzymes	2016-04-24	Download	Download

Structures (in SDF Format)

Data Set	Released on	SDF File
Metabolite Structures	2016-04-24	Download

Metabolite and Protein Data (in XML format)

Data Set	Released on	XML File
All Metabolites	2016-04-24	Download
All Proteins	2016-04-23	Download

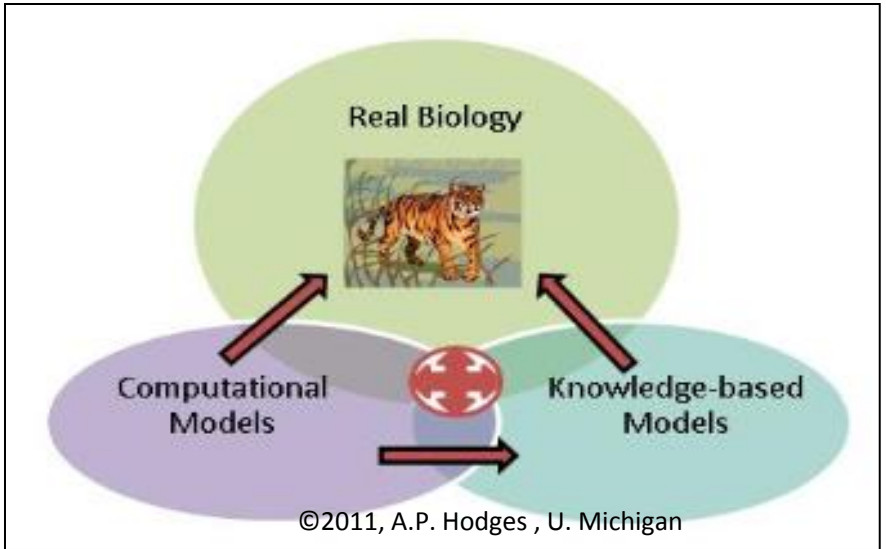
Spectra

Data Set	Download Link
Mass Spectra Image Files	Download
GC/MS Peak Lists	Download
NMR Spectra FIDS Files	Download
NMR Spectra Peaklist Files	Download
Spectra information (in XML format)	Download

SBP
SANFORD BURNHAM PREBYS

- MetaCore & MetaDrug ('the right leg')
  - Powerful tool for searching hubs, known/curated interactions across multiple biological scales/types, drug targeting, etc.
  - Great companion to predictions obtained from machine learning
  - Used to validate predictions from Viper/master regulator predictions

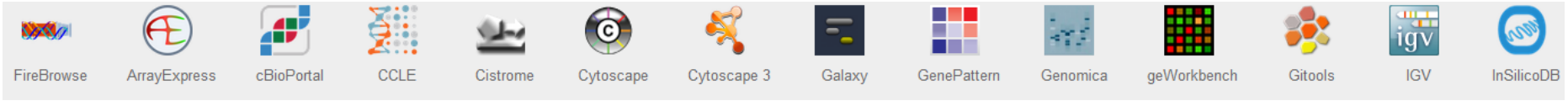
- NextBio
- IPA



SBP
SANFORD BURNHAM PREBYS

Genomespace (DIY sandbox)

http://genomespace.org/



- Many tools for next-gen data analyses (Broad Inst. & others)
- Public data championed by TCGA and other sites
- Personal workspace with 30+gb storage

- Also, **geWorkbench** in GenomeSpace for AracNe
  - http://wiki.c2b2.columbia.edu/workbench/index.php/Home
  - Sometimes unstable/memory issues (Java-based)

SBP
SANFORD BURNHAM PREBYS

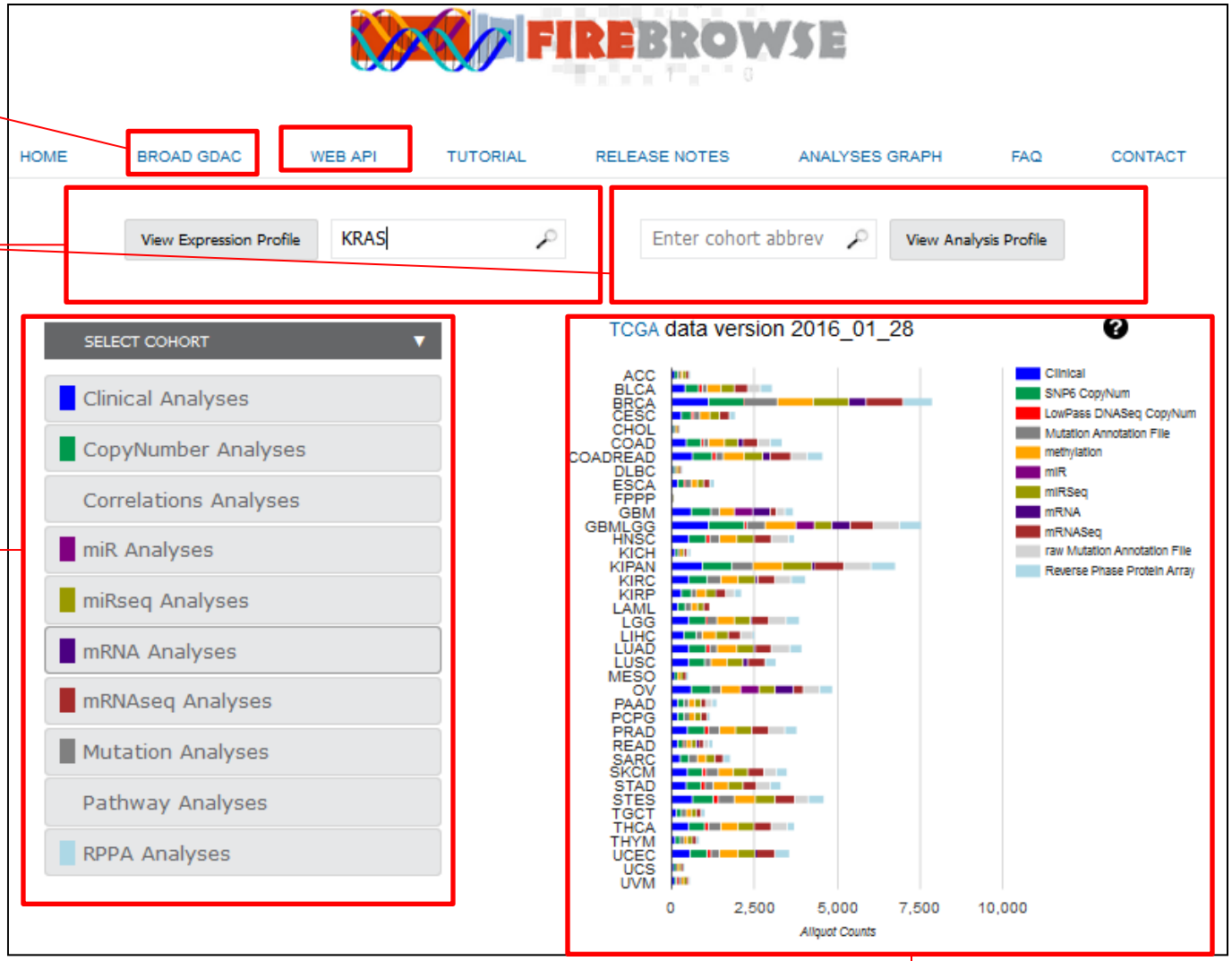
Summary table (TCGA)

Search by gene or cancer type

Compare 2 analyses (SNP vs Expression)

Clinical correlations

Basic pathway analysis



Download data directly (clickable image)

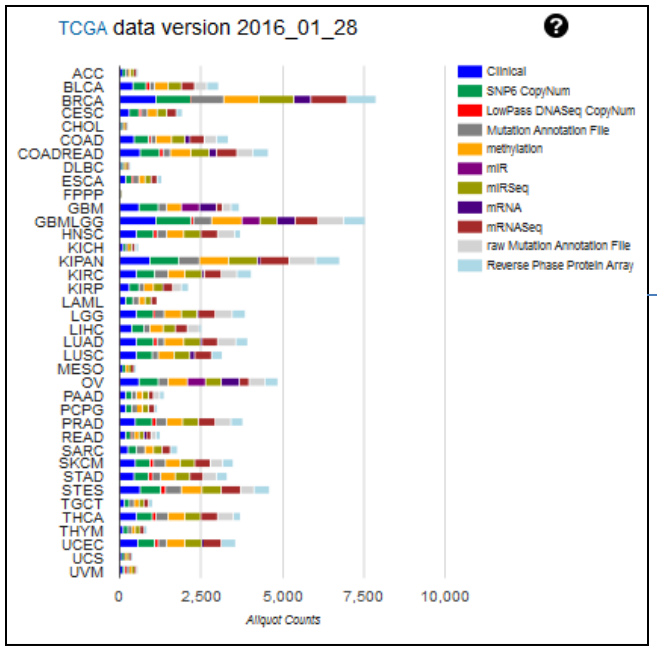
http://firebrowse.org/

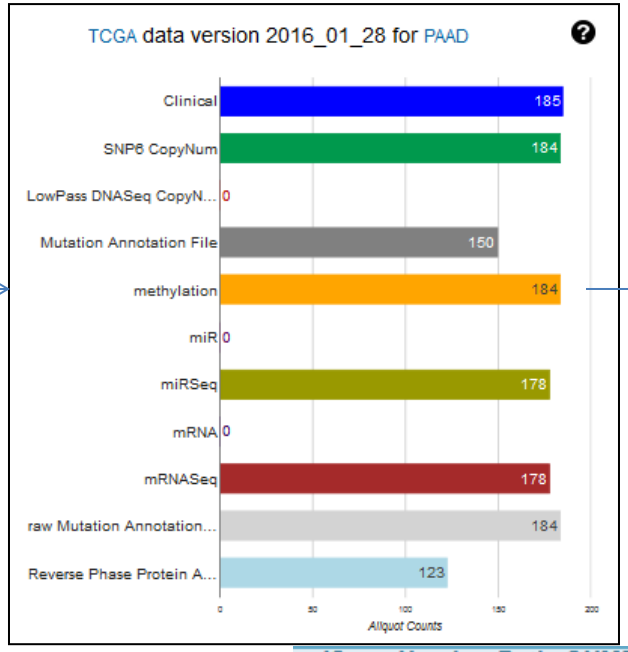
SBP
SANFORD BURNHAM PREBYS

Broad GDAC

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browse	Browse
Bladder urothelial carcinoma	BLCA	412	Browse	Browse
Breast invasive carcinoma	BRCA	1098	Browse	Browse
Cervical and endocervical cancers	CESC	307	Browse	Browse
Cholangiocarcinoma	CHOL	51	Browse	Browse
Colon adenocarcinoma	COAD	460	Browse	Browse
Colorectal adenocarcinoma	COADREAD	631	Browse	Browse
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	Browse	Browse
Esophageal carcinoma	ESCA	185	Browse	Browse
FFPE Pilot Phase II	FPPP	38	None	Browse
Glioblastoma multiforme	GBM	613	Browse	Browse
Glioma	GBMLGG	1129	Browse	Browse
Head and Neck squamous cell carcinoma	HNSC	528	Browse	Browse
Kidney Chromophobe	KICH	113	Browse	Browse
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	Browse	Browse
Kidney renal clear cell carcinoma	KIRC	537	Browse	Browse
Kidney renal papillary cell carcinoma	KIRP	323	Browse	Browse
Acute Myeloid Leukemia	LAML	200	Browse	Browse
Brain Lower Grade Glioma	LGG	516	Browse	Browse
Liver hepatocellular carcinoma	LIHC	377	Browse	Browse
Lung adenocarcinoma	LUAD	585	Browse	Browse
Lung squamous cell carcinoma	LUSC	504	Browse	Browse
Mesothelioma	MESO	87	Browse	Browse
Ovarian serous cystadenocarcinoma	OV	602	Browse	Browse
Pancreatic adenocarcinoma	PAAD	185	Browse	Browse
Pheochromocytoma and Paraganglioma	PCPG	179	Browse	Browse
Prostate adenocarcinoma	PRAD	499	Browse	Browse
Rectum adenocarcinoma	READ	171	Browse	Browse
Sarcoma	SARC	261	Browse	Browse
Skin Cutaneous Melanoma	SKCM	470	Browse	Browse
Stomach adenocarcinoma	STAD	443	Browse	Browse
Stomach and Esophageal carcinoma	STES	628	Browse	Browse
Testicular Germ Cell Tumors	TGCT	150	Browse	Browse
Thyroid carcinoma	THCA	503	Browse	Browse
Thymoma	THYM	124	Browse	Browse
Uterine Corpus Endometrial Carcinoma	UCEC	560	Browse	Browse
Uterine Carcinosarcoma	UCS	57	Browse	Browse
Uveal Melanoma	UVM	80	Browse	Browse

SBP
SANFORD BURNHAM PREBYS





Download

Correlation between copy number variations of arm-level result and molecular subtypes

Pancreatic Adenocarcinoma (Primary solid tumor)

28 January 2016 | analyses__2016_01_28 Maintainer Information Citation Information doi:10.7908/C1SQ8ZTT

- Overview
  - Introduction
  - Summary
- Results
  - Overview of the results

GET FULL TABLE

Table 1. Overview of the association between significant copy number variation of 80 arm-level events and 10 molecular subtypes. Shown in the table are P values (Q values). Thresholded by P value < 0.05 and Q value < 0.25, 268 significant findings detected.

		Clinical Features	*CN CNMF*	*METHLYATION CNMF*	*RPPA CNMF*	*RPPA CHIERARCHICAL*	*MRNASEQ CNMF*	*MRNASEQ CHIERARCHICAL*	*MIRSEQ CNMF*	*MIRSEQ CHIERARC*
	nCNV (%)	nWild-Type	Fisher's exact test	Fisher's exact test	Fisher's exact test	Fisher's exact test	Fisher's exact test	Fisher's exact test	Fisher's exact test	Fisher's ex
6q loss	82 (45%)	102	1e-05 (0.000571)	1e-05 (0.000571)	0.0225 (0.0922)	0.521 (0.679)	2e-05 (0.000941)	0.00089 (0.0132)	0.00127 (0.0164)	1e-05 (0.000571)
4q gain	9 (5%)	175	0.007 (0.0479)	0.571 (0.726)	0.359 (0.535)	0.0451 (0.142)	0.00798 (0.0519)	0.0089 (0.0535)	0.0399 (0.132)	0.0135 (0.0663)
7p gain	53 (29%)	131	0.00117 (0.0155)	0.0467 (0.144)	0.00733 (0.0489)	0.0474 (0.145)	0.0122 (0.0624)	0.205 (0.367)	0.144 (0.295)	0.0379 (0.128)
9p loss	74 (40%)	110	1e-05 (0.000571)	0.00027 (0.00502)	0.0261 (0.1)	0.0623 (0.171)	0.00172 (0.0183)	0.0429 (0.138)	0.0484 (0.147)	0.0226 (0.0922)
12p loss	19 (10%)	165	0.0234 (0.0936)	0.0349 (0.12)	7e-05 (0.00224)	0.00142 (0.0174)	0.00914 (0.0539)	0.0143 (0.0683)	0.595 (0.75)	0.00021 (0.00442)
17p loss	79 (43%)	105	1e-05 (0.000571)	1e-05 (0.000571)	0.0519 (0.153)	0.129 (0.276)	1e-05 (0.000571)	1e-05 (0.000571)	0.0001 (0.00276)	6e-05 (0.00209)

'Significant' results indicated with red dot

'Copy Number Ratio CNMF subtypes' versus *'RADIATION_THERAPY'*

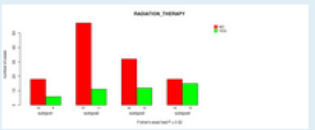
P value = *0.02* (Fisher's exact test), Q value = *0.15*

Table S9. Clustering Approach #1: *'Copy Number Ratio CNMF subtypes'* versus Clinical Feature #8: *'RADIATION_THERAPY'*

nPatients	NO	YES
ALL	125	44
subtype1	18	6
subtype2	57	11
subtype3	32	12
subtype4	18	15

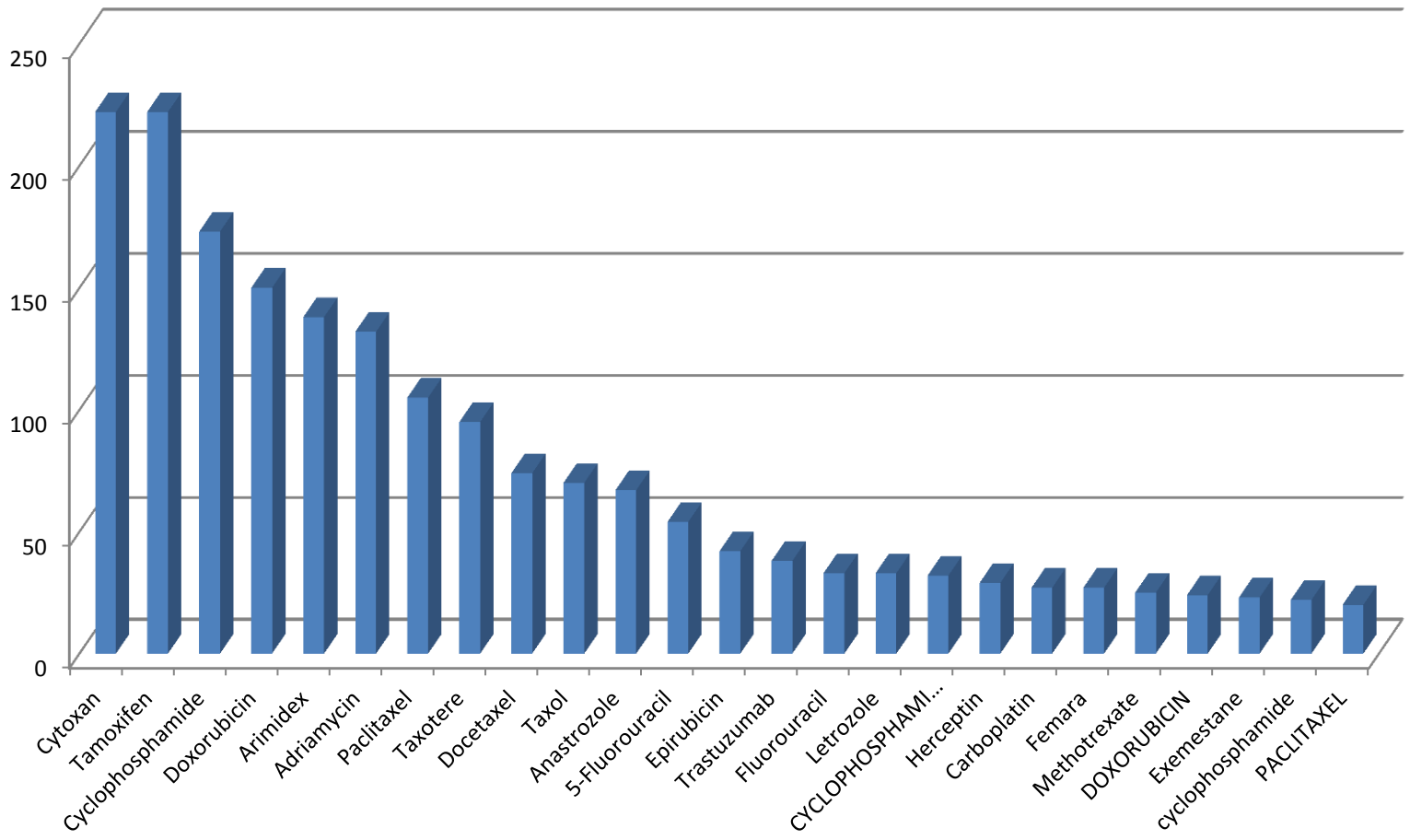
GET HIGH-RES IMAGE

Figure S8. Clustering Approach #1: *'Copy Number Ratio CNMF subtypes'* versus Clinical Feature #8: *'RADIATION_THERAPY'*



SANFORD BURNHAM PREBYS

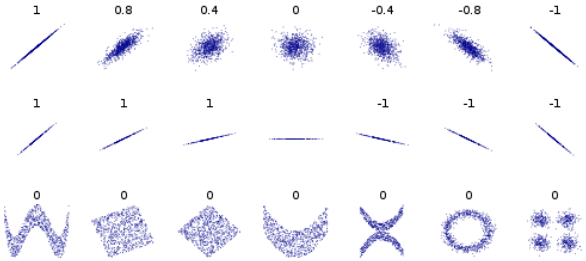
Drug information – BRCA dataset – not filtered by the FireBrowse tool!



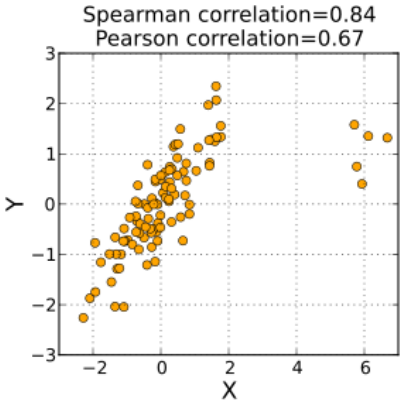
Firebrowser is not context-specific

SBP
SANFORD BURNHAM PREBYS

Evolving beyond Pearson corr. (‘the left leg’)



- **Pearson correlation** $\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$
  - Linear up/down relationships, sensitive to outliers

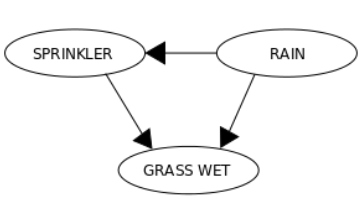


- **Spearman correlation** $r_s = \rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\operatorname{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}$
  - Up/down relationships, less sensitive to outliers
  - Idea of monotonically increasing/decreasing

- **Mutual information** $I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}$
  - Deals with linear & nonlinear relationships
  - Depends on (joint) probability distributions of data
  - Pairwise MI (e.g. in AracNe) almost always used for inference

- **Bayesian network/belief-based model**
  - Joint probabilities: e.g. P(A) = P(A|B)*P(B)
  - Similar to pairwise MI, but multiple controllers possible
  - More complicated: lots of ways to draw putative network

SPRINKLER		
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

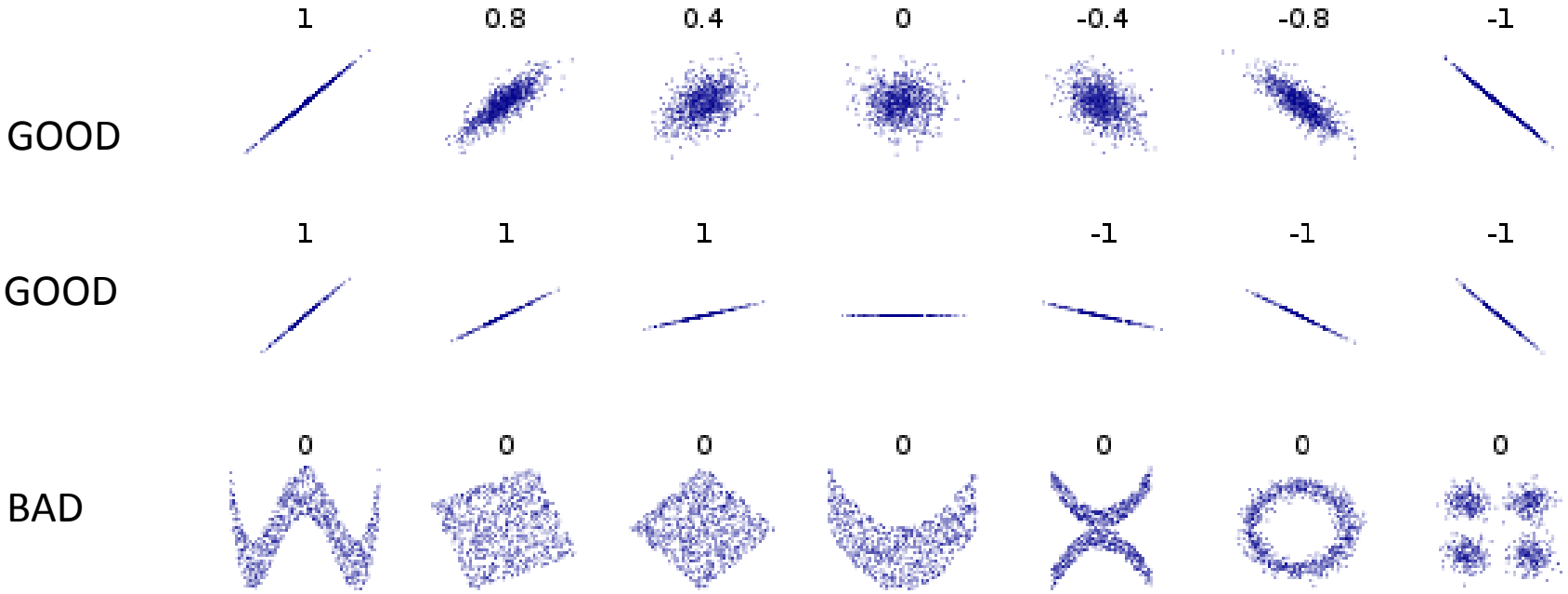


RAIN	
T	F
0.2	0.8

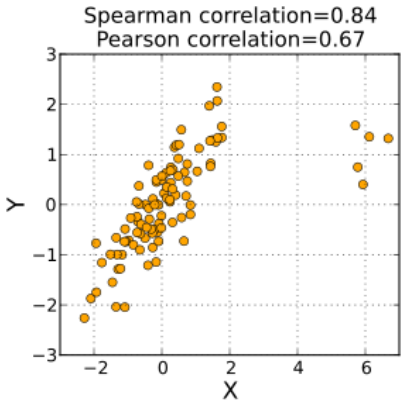
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

SBP
SANFORD BURNHAM PREBYS

When Pearson/Spearman metrics work & when they fail:

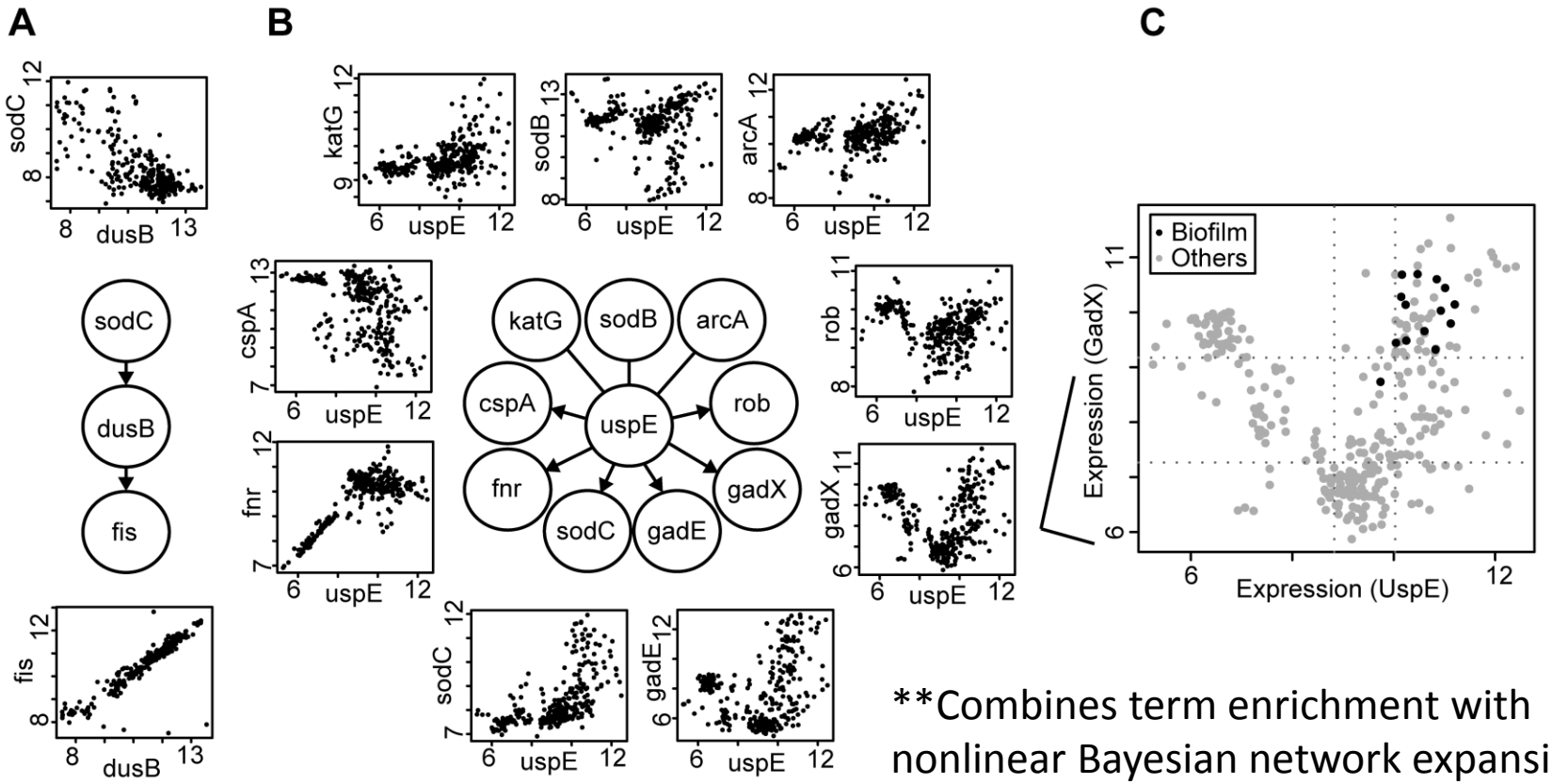


1. Don’t use Pearson correlations!
2. Stratify/select data based on PCA/clustering/etc.
3. Maintain sufficient # of samples



SBP
SANFORD BURNHAM PREBYS

BN/MI approaches identify nonlinear interactions and driver genes totally missed by Pearson/Spearman!

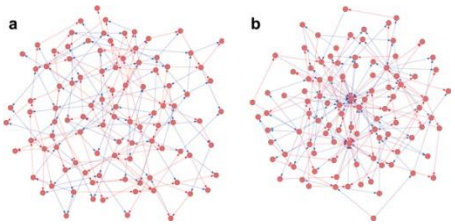


**Combines term enrichment with nonlinear Bayesian network expansion (BN+1) → novel functional prediction

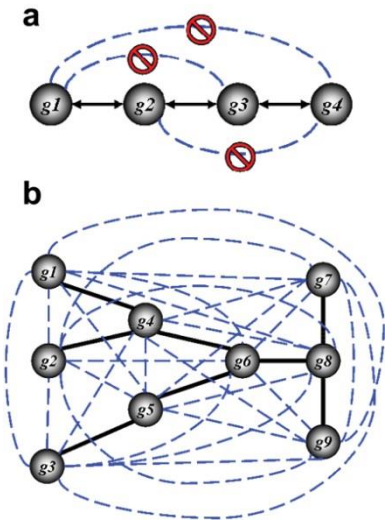
Hodges et al. 2010. Bayesian network expansion identifies new ROS and biofilm regulators. PLoS ONE. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009513

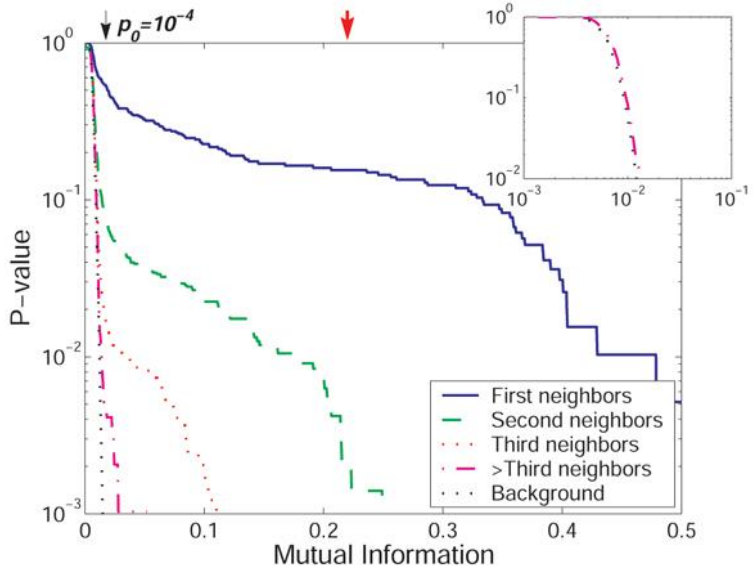
SBP
SANFORD BURNHAM PREBYS

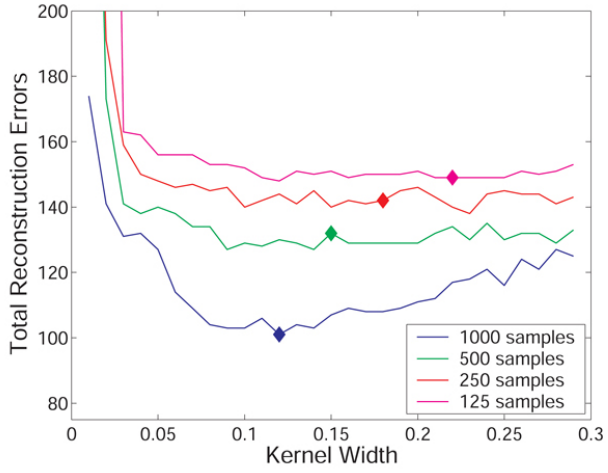
AraCNe MI



- Andrea Califano (Columbia U.)
- Mutual information
  - MI = 0 ➔ no information, poor edge/int.
  - MI = 1 ➔ ~perfect correlation, excellent edge/int.
- P-val approaches Inf. as MI → 1
  - P-value used in filtering interactions.
  - Filtering needed for N^2-N possible pairwise interactions!







http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810318/

Viper

SBP
SANFORD BURNHAM PREBYS

Viper:

- Identifies driver variables from a biological dataset
- Input: Your data or publicly-available
  - Transcriptomics, Proteomics, Metabolomics, Drug Assay, etc.
  - Matrix-style format (tab-delimited, row & column identifiers)
- Output: novel biological hypotheses
  - Master regulators
  - Implicated 'regulatory module' members
- Requires two phenotypic groups
  (e.g. cancer vs normal samples)
  - Can slice data from TCGA, CCLE, published drug assays, etc.
- Recommended minimum 10 samples/group
  - Possible to do ~3 samples, but likely no discriminatory power

SBP
SANFORD BURNHAM PREBYS

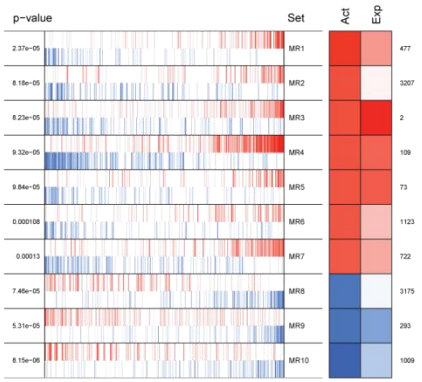
Viper Pipeline Overview:

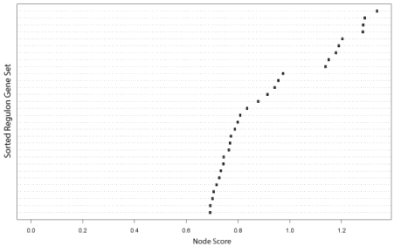
1. Data acquisition & preprocessing
2. Prediction of interactions (AracNe)
3. Compute gene signatures (paired T test)
4. Generate regulons (Viper)
5. Prioritize master regulators (random forest/KNN)
6. Validation/support (Metacore, NextBio, other provenance)
7. Advanced visualization & reporting (Regulattice, Metacore, Cytoscape, D3.js, etc.)

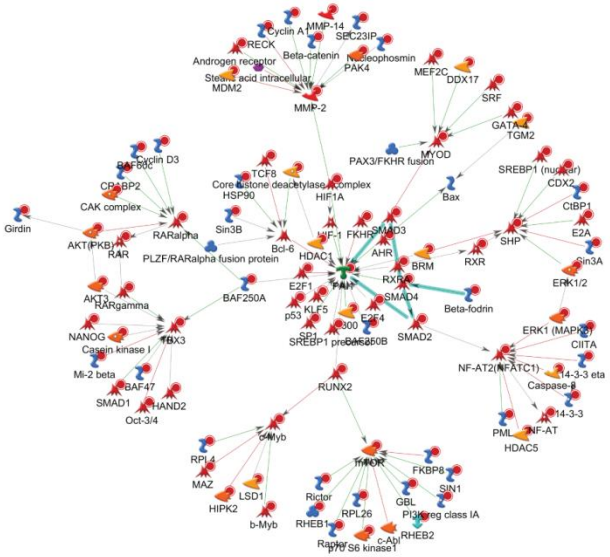
Optional:

Protein expression prediction (viper)

PEx signature matching to samples (Sample Surfer)







SBP
SANFORD BURNHAM PREBYS

“Suggestions”

1. Let your data guide the hypothesis.
    1. 2-group comparison for Viper
    2. Regression recommended for >2 groups

2. Small & homogenous datasets are often detrimental to biological inference.
    1. Variability = helpful for machine learning
    2. Recommended min. 5-10 samples per phenotypic group

SBP
SANFORD BURNHAM PREBYS

Like GSEA, but repeated for every candidate MR (e.g. all genes)

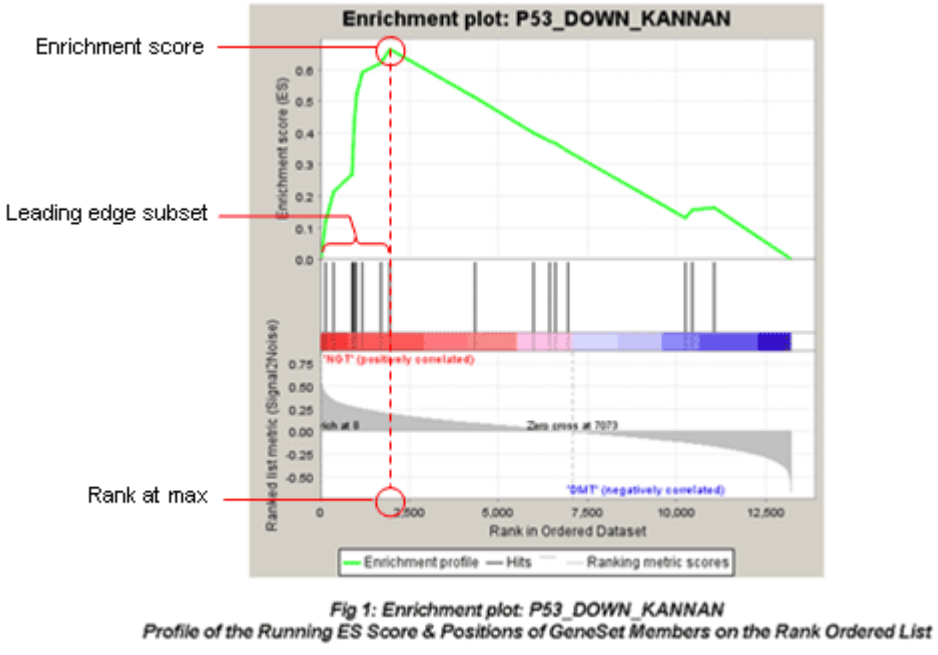
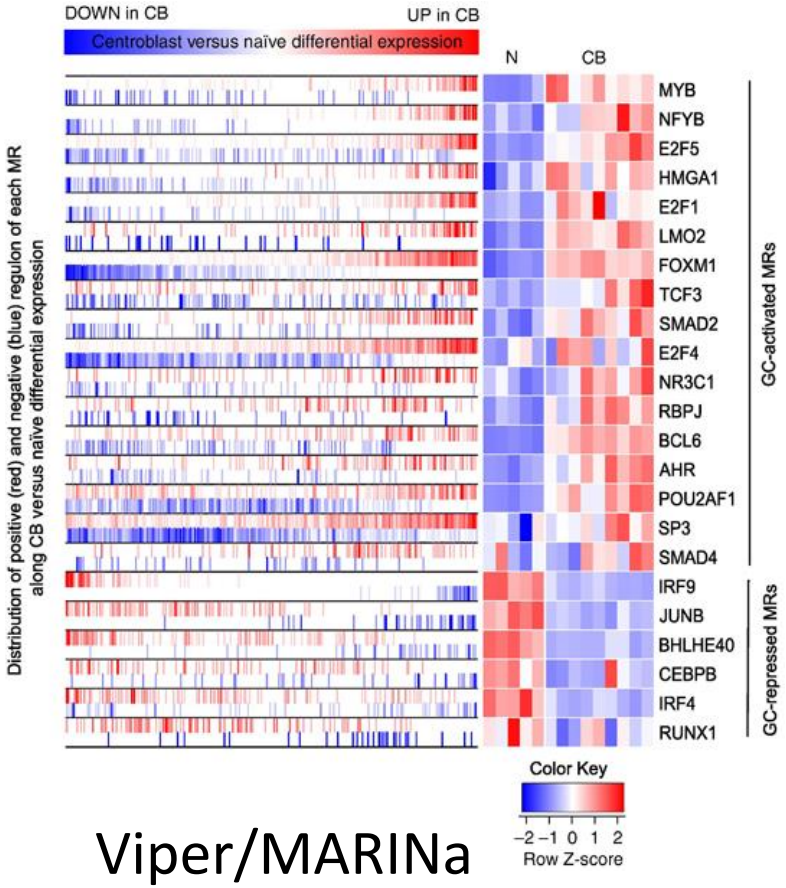


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

GSEA



Viper/MARINa

<email: ahodges@sbpdiscovery.org>

Part 3 - Examples

SBP
SANFORD BURNHAM PREBYS

Comparisons (so far):

- Pancreatic cancer
  - RNAseq analysis
    - RSEM processed (*different!)
  - Comparison of responders vs progressive disease
    - Both were treated with gemcitabine
- Lung cancer
  - Microarray & RNAseq
  - Comparing early vs late stage
- Breast cancer analysis
  - **chip details
  - RNAseq analysis
    - RPKM processed (standard format)
  - Comparison of 2 drugs & patient responses
    - Tamoxifen
    - Cytoxan
- Glioblastoma (Petrus/Vuori lab)
- Skeletal muscle (TRI)
  - Good vs poor ATP utilizers
  - Caloric restriction study
  - 5-drug analysis

SBP
SANFORD BURNHAM PREBYS

Typical scenario:

1. Pre-treat/normalize/batch correct/log transform as needed
2. Filter genes based on coefficient of variation, low abundance
3. Fold-change filtering & ANOVA
4. Run AracNe mutual information on the data
   - Compute interactome/set of possible molecular interactions
5. Run Viper (including student's T test for signature)
6. Assess results
   - Filter regulons based on NES, p-value, &/- FDR
   - Check enrichment of Master Regulators
   - Check regulon behavior

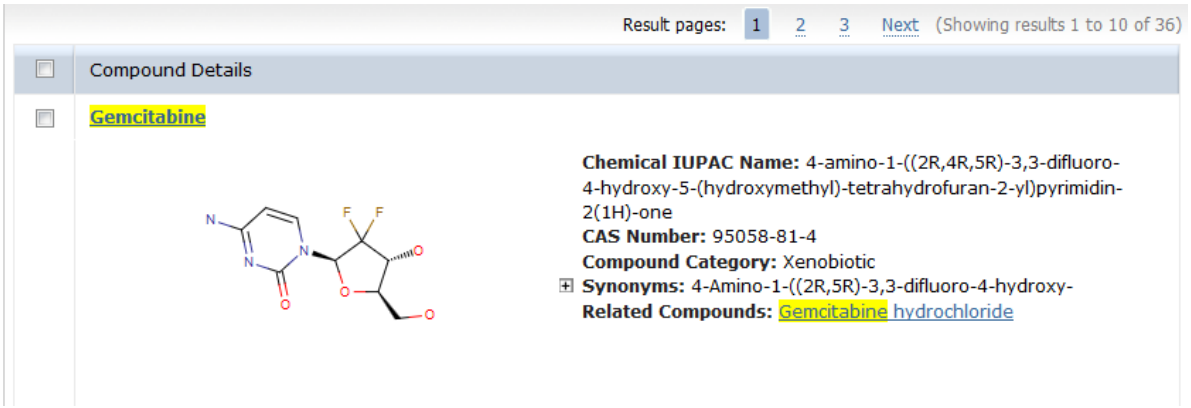
Pancreatic cancer analysis

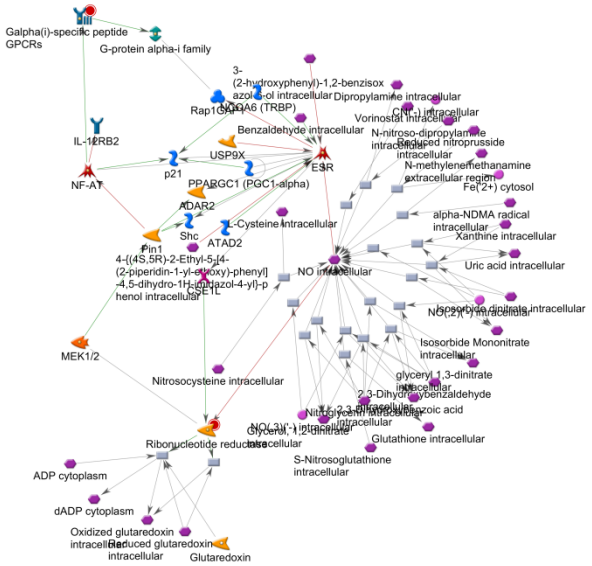
Method:

- 183 Illumina Hiseq V2 samples from TCGA
- RSEM – processed (non-RPKM)
- Used normalized version of gene data
- Selected all samples of patients treated with gemcitabine
- Compare responders and non-responders in Viper:
  - "Complete response" (16) vs "Clinical progressive disease" (20)
  - **another group, partial responders, show distinct differences vs both

SBP
SANFORD BURNHAM PREBYS

Gemcitabine





Description	An anticancer drug that belongs to the family of drugs called antimetabolites.
Drug Status	Clinical trial, FDA approved, Clinical trial - Phase II
IUPAC Name	4-amino-1-((2R,4R,5R)-3,3-difluoro-4-hydroxy-5-(hydroxymethyl)-tetrahydrofuran-2-yl)pyrimidin-2(1H)-one
Chemical Formula	C9H11F2N3O4
SMILES	Nc1ccn([C@@H]2O[C@H](CO)[C@@H](O)C2(F)F)c(=O)n1 |r|
InChI	InChI=1S/C9H11F2N3O4/c10-9(11)6(16)4(3-15)18-7(9)1 4-2-1-5(12)13-8(14)17/h1-2,4,6-7,15-16H,3H2,(H2,12,13,17)/t4-,6-,7-/m1/s1
Category	Xenobiotic
Melting Point	168.64°C
Molecular Weight	263.2

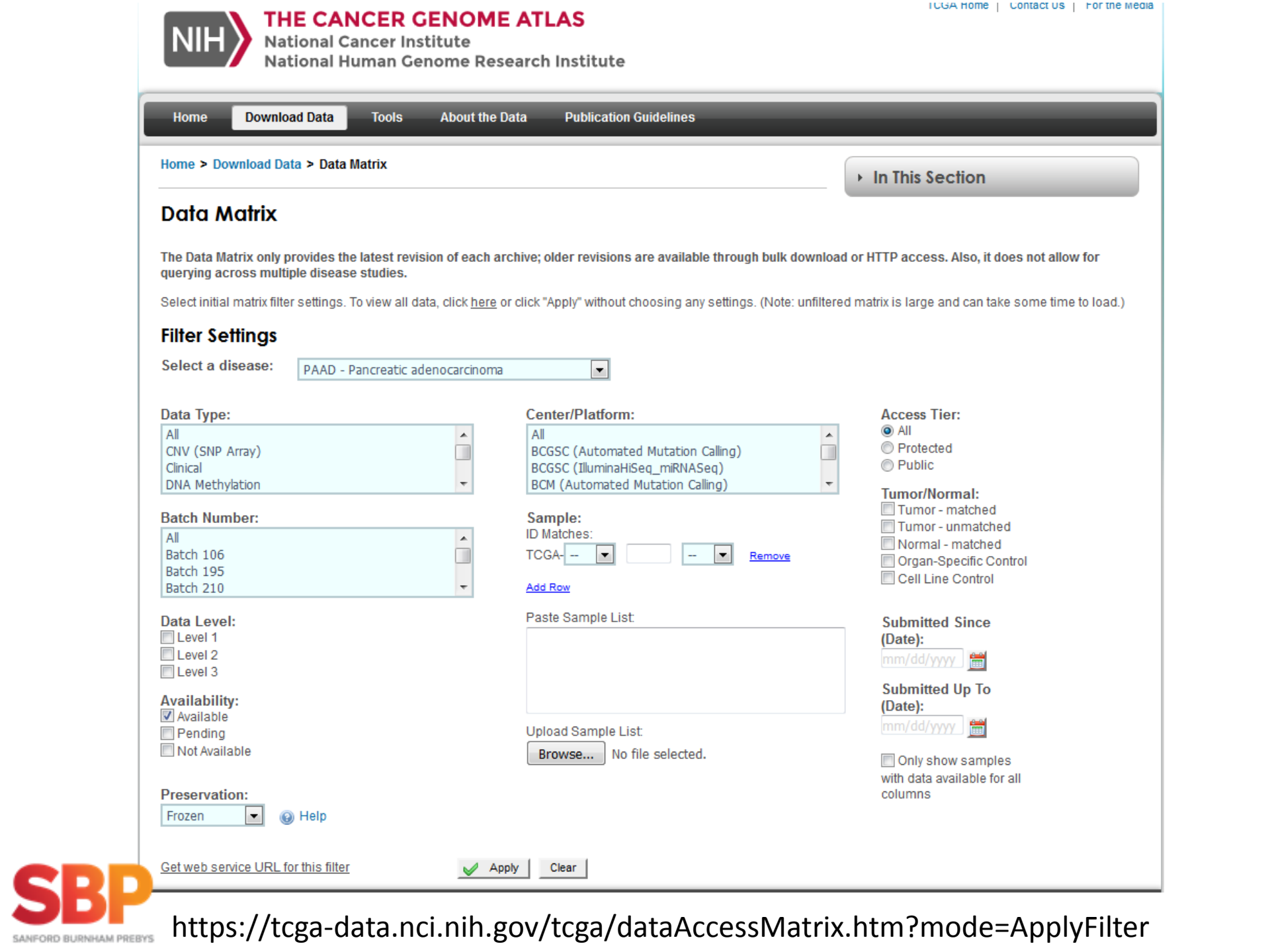
Target: RRM2B
Anti-neoplastic drug

MetaCore compound/structure search
http://portal.genego.com

SBP
SANFORD BURNHAM PREBYS

Indications

#	Regimen	Disease	Drug	Drug Status	Dosage
1	Single Agent	Carcinoma, Non-Small-Cell ...	Gemcitabine	Clinical trial	
2	Single Agent	Sarcoma	Gemcitabine	Clinical trial - Phase II	
3	Single Agent	Colorectal Neoplasms	Gemcitabine	Clinical trial - Phase II	
4	Single Agent	Endometrial Neoplasms	Gemcitabine	Clinical trial - Phase II	
5	Single Agent	Stomach Neoplasms	Gemcitabine	Clinical trial - Phase II	800 mg/m2, 1 times/week
6	Single Agent	Neoplasms	Gemcitabine	Clinical trial	2, 5, 10 mg, 1 times/14 days
7	Single Agent	Cholangiocarcinoma	Gemcitabine	Clinical trial - Phase II	1000 mg/m2
8	Single Agent	Pancreatic Neoplasms	Gemcitabine	FDA approved	1000 mg/m2
9	Combination	Pancreatic Neoplasms	Gemcitabine	Clinical trial - Phase II	
			Virulizin		
10	Combination	Carcinoma, Non-Small-Cell ...	Gemcitabine	Clinical trial - Phase II	
			Paclitaxel		
11	Combination	Carcinoma, Transitional Cell	Gemcitabine	Clinical trial - Phase II	
			Lonafarnib		
12	Combination	Carcinoma, Hepatocellular	Gemcitabine	Clinical trial - Phase II	
			Lonafarnib		
13	Combination	Carcinoma, Transitional Cell	Cisplatin	Clinical trial - Phase III	
			Gemcitabine		
14	Combination	Pancreatic Neoplasms	Gemcitabine	Clinical trial - Phase I	
			Ipilimumab		
15	Combination	Pancreatic Neoplasms	Gemcitabine	Clinical trial - Phase II	1000 mg/m2
			Infliximab		5 mg/kg
16	Combination	Cachexia	Gemcitabine	Clinical trial - Phase II	1000 mg/m2
			Infliximab		5 mg/kg
17	Combination	Urinary Bladder Neoplasms	Everolimus	Clinical trial - Phase I/II	5 mg
			Gemcitabine		2000 mg, 2 times/week



SBP SANFORD BURNHAM PREBYS

https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter

Drug & phenotype information

RNAseq data (lvl3, RSEM)

Data Matrix Datasets

The Data Matrix only provides the lat t revision of each archive; older revisions are available through bulk download or TP access. Also, it does not allow for querying across multiple disease udies.

PAAD Data Matrix Options: Reset Matrix | Edit Filter | Remove Filter Help

Preservation: Frozen

Color Cells By: Availability **Scroll Size:** Standard

Build Archive

Legend:

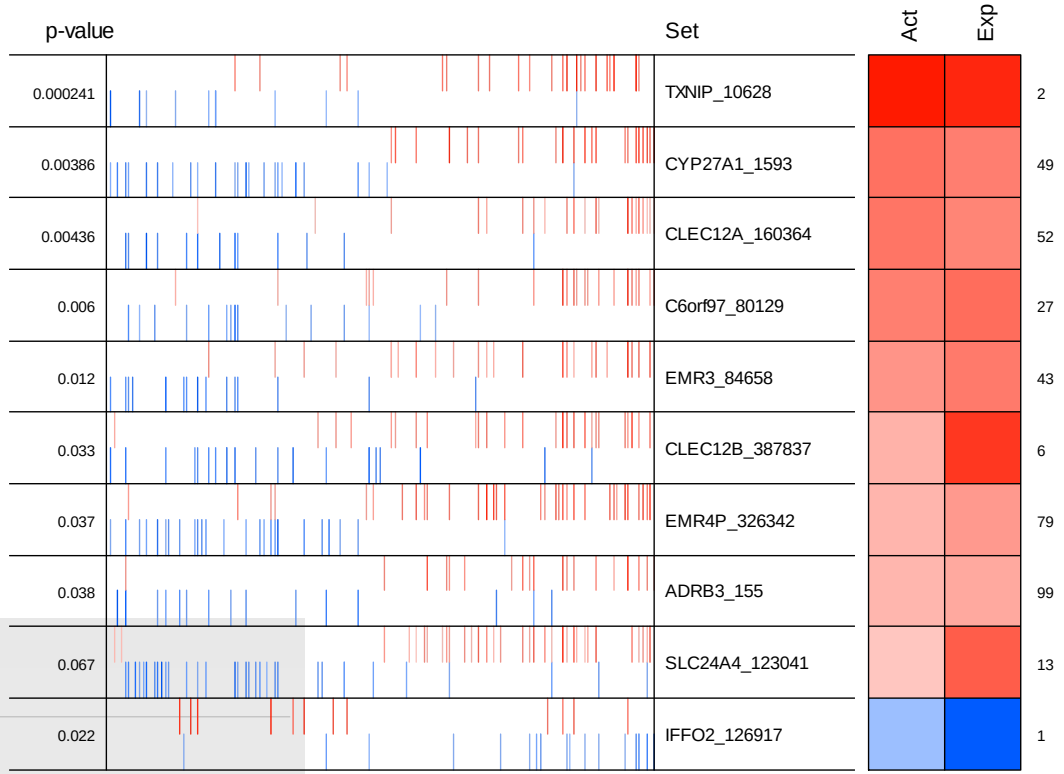
- A Available
- P Pending
- N Not Available
- Not Applicable

*Protected data

Batch/Sample	Level	Clinical		Methyl			CNV (SNP Array)			Somatic Mutations						miRNASeq	Exp-Protein			Fragment Analysis	RNASeqV2	Protected Mutations					
		XML	Biotab	JHU-USC HumanMethylation450			BI Genome_Wide_SNP_6			BI Mutation Calling	BI Automated Mutation Calling	BI Curated Mutation Calling	UCSC Automated Mutation Calling	BCGSC Automated Mutation Calling	BCM Automated Mutation Calling	BCGSC IlluminaHiSeq_miRNASeq	MDA MDA_RPPA_Core			NCH microsat_i	UNC IlluminaHiSeq_RNASeqV2	BI Mutation Calling	BI Automated Mutation Calling	BI Curated Mutation Calling	UCSC Automated Mutation Calling	BCGSC Automated Mutation Calling	BCM Automated Mutation Calling
				1	2	3	1*	2*	3	2	2	2	2	2	2	3	1	2	3	1*	3	2*	2*	2*	2*	2*	2*
Batch 106	TCGA-FZ-5919-01	A	A	A	A	A	A	A	A	A	A		A				A	A	A	A		A	A		A		
	TCGA-FZ-5920-01	A	A	A	A	A	A	A	A	A	A		A				A	A	A	A		A	A		A		
	TCGA-FZ-5921-01	A	A				A	A	A	A	A		A				A	A	A	A		A	A		A		
	TCGA-FZ-5922-01	A	A	A	A	A	A	A	A	A	A		A				A	A	A			A	A		A		
	TCGA-FZ-5923-01	A	A	A	A	A	A	A	A	A	A		A				A	A	A	A		A	A		A		
	TCGA-FZ-5924-01	A	A	A	A	A	A	A	A	A	A		A				A	A	A	A		A	A		A		
	TCGA-FZ-5926-01	A	A	A	A	A	A	A	A	A	A		A				A	A	A			A	A		A		
	TCGA-FZ-5919-11	A	A	A	A	A	A	A	A	A	A		A									A	A		A		

SBP SANFORD BURNHAM PREBYS

Initial Viper results:



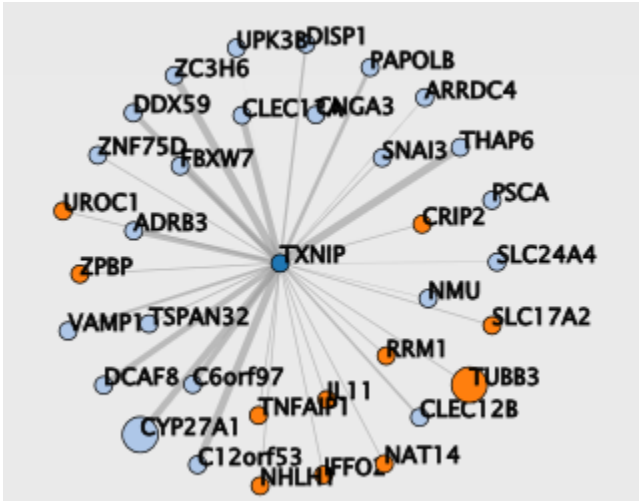
List-based regulon prioritization (All Mito list)

View | Edit | Track

Purpose: list master regulators and regulon information:

Toggle table view Curated human mitchondrial proteins from MitoCarta 2.0 (including synonyms):

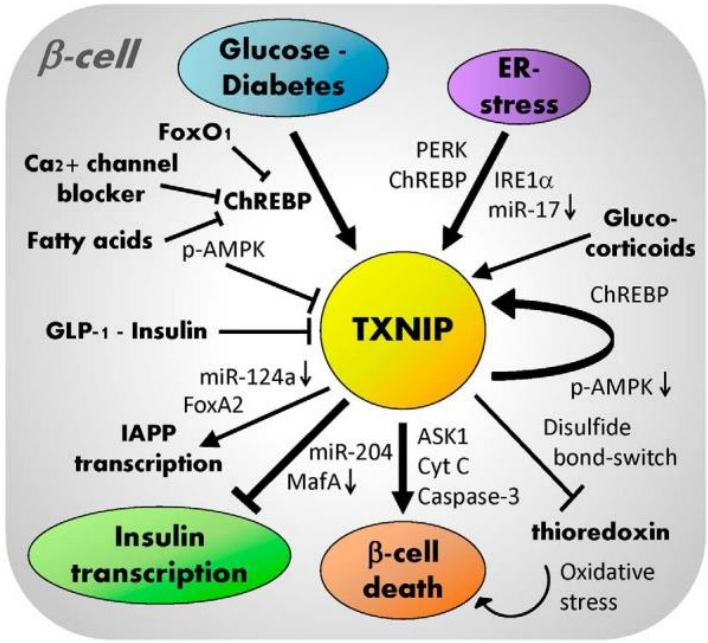
#	Probe	Master Regulator Gene	Regulatory Module Genes	Matched Genes	# Matched Genes	Total # in Regulon (inc. MR)	% Match
1	TXNIP_10628	TXNIP [Int.s]	ADRB3 ARRDC4 C12orf53 C6orf97 CLEC12A CLEC12B CNGA3 CRIP2 CYP27A1 DCAF8 DDX59 DISP1 FBXW7 IFFO2 IL11 NAT14 NHLH1 NMU PAPOLB PSCA RRM1 SLC17A2 SLC24A4 SNAI3 THAP6 TNFAIP1 TSPAN32 TUBB3 UPK3B UROC1 VAMP1 ZC3H6 ZNF75D ZPBP	CYP27A1 TUBB3	2	35	5.71
2	CYP27A1_1593	CYP27A1 [Int.s]	ACTBL2 ACTL6A ACYP1 ADRB3 C12orf53 C13orf29 C6orf201 C6orf97 CEP55 CLEC12A CLEC12B CNFN DCAF8 DISP1 EMR3 EMR4P FKBP5 GPC5 IFFO2 IGSF10 IL11 MBOAT2 NAA50 NAIP NAT14 NBPF14 NCRNA00081 NCRNA00152 NUDT15 PDCD10 PODXL RAB11FIP5 RRM1 RTN3 SEMA7A SLC17A2 SLC24A4 SNAI3 TAF1B TNNI3K TSPAN32 TXNIP UROC1 VAMP1 ZNF540 ZPBP	RAB11FIP5	2	47	4.26
3	CLEC12A_160364	CLEC12A [Int.s]	ACYP1 ADRB3 C12orf53 C6orf218 C6orf97 CLEC12B CNFN CRIP2 CYP27A1 DDX59 ELANE EMR3 EMR4P FKBP5 GFRA2 IGSF10 LOC100272216 MDFI NAIP NAT14 NUDT15 OVCH1 PKD1L3 RAB11FIP5 RABEP1 RRM1 RTN3 SLC17A2 SLC22A13 SLC24A4 SNAI3 TSPAN32 TXNIP UROC1	CYP27A1 RAB11FIP5	2	35	5.71
[illegible]	[illegible]	[illegible] [Int.s]	ACTG1 ACYP1 ADRB3 C12orf53 C6orf218 CLEC12A CLEC12B CNFN CRIP2 CSTF2T CYP27A1 DCAF8 EMR3 EMR4P GFRA2 GPR17 KRT17 KRT80 NAIP NAT14 NBPF14 NCRNA00152 NRM OVCH1 PKD1L3 PMEPA1 RTN3 RYR2 SLC17A2 SLC24A4 SMEK1 SNAI3 TFAP2A THAP6 TSPAN32 TXNIP UROC1 ZC3H6	CYP27A1	1	39	2.56



SANFORD BURNHAM PREBYS

Pancreatic Cancer
Top result: tumor suppressor perturbed in pancreatic cancer!

significantMRs	Tfmode(MR)	# genes in regulon	NES	p.value	FDR
TXNIP_10628	1.8265	34	3.67	0.000241	0.033
CYP27A1_1593	0.773951	46	2.89	0.00386	0.199
CLEC12A_160364	-0.415	34	2.85	0.00436	0.199
C6orf97_80129	-0.08728	38	2.75	0.006	0.206
EMR3_84658	-1.08802	39	2.51	0.012	0.329
IFFO2_126917	0.952298	31	-2.29	0.022	0.469
CLEC12B_387837	-1.23522	48	2.13	0.033	0.469
EMR4P_326342	-0.74868	59	2.09	0.037	0.469
ADRB3_155	-1.04984	34	2.07	0.038	0.469



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4116588/figure/F2/

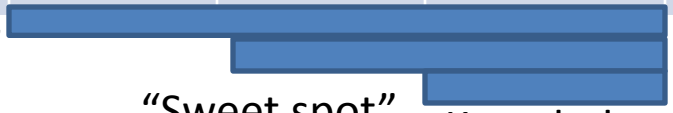
SBP
SANFORD BURNHAM PREBYS

Rank	Master Regulator	Name	Chr	Biological Role(s)
1	**TXNIP**	Thioredoxin interacting protein	1q21.2	Inhibits antioxidative function of thioredoxin; regulates cellular metabolism and ER stress; tumor suppressor
2	CYP27A1	Cytochrome P450 family 27 subfamily A member 1	2q35	Monooxygenase, catalyzing reactions in drug metabolism, cholesterol/steroid/lipid synthesis, mitochondrial protein
3	**CLEC12A**	C-type lectin domain family 12 member A	12p13.2	Negative regulator of granulocyte and monocyte function
4	C6orf97	Coiled-coil domain containing 170	6q25.1	**susceptibility locus in breast cancer (GWAS implicated)
5	**EMR3** (ADGRE3)	Adhesion G protein-coupled receptor E3	19p13.1	TM7 transmemberane protein; granulocyte marker; mediator: invasive variation in glioblastoma
6	IFFO2	Intermediate filament family orphan 2	1p36.13	Allergic diseases
7	**CLEC12B**	C-type lectin domain family 12 member B	12p13.2	Inhibitory receptor on myeloid cells
8	**EMR4P** (ADGRE4P)	Adhesion G protein-coupled receptor E4, pseudogene	19p13.3	Encoded protein not yet detected, thought to be soluble vs surface exprs'd
9	**ADRB3**	Adrenoceptor beta 3	8p11.23	Regulation of lipolysis and thermogenesis; mediates catecholamine-induced activation of adenylate cyclase via G proteins

SBP
SANFORD BURNHAM PREBYS

Implications of parameter selection: sample size, p-val filtering, data heterogeneity,

ANOVA p-value (feature filter)	0.05	0.05	0.05	0.01	0.01	0.01
AracNe p-value (interaction filter)	1E-2	1E-5	1E-7	1E-2	1E-5	1E-7
NES p-value (regulon filter)	0.05	0.05	0.05	0.05	0.05	0.05
# Interactions inferred (interactome)	915,585	190,330	66,770	4,932	810	210
# Regulons @ p<.05 cutoff (viper)	64	47	17	9	0	0
Median # features/regulon	532	92	59	38	0	0



"Too loose"

"Sweet spot"

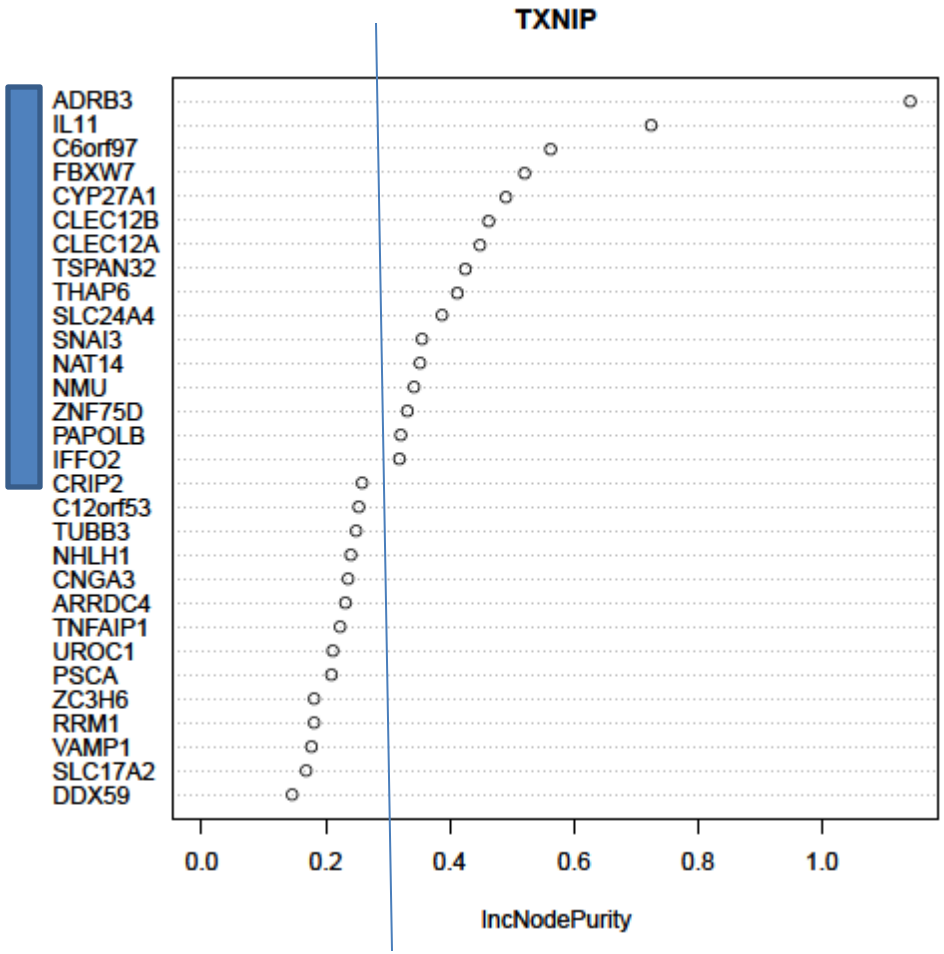
Knowledge-matched

"Too stringent"

SBP
SANFORD BURNHAM PREBYS

Random Forest – impute most important regulon members for TXNIP

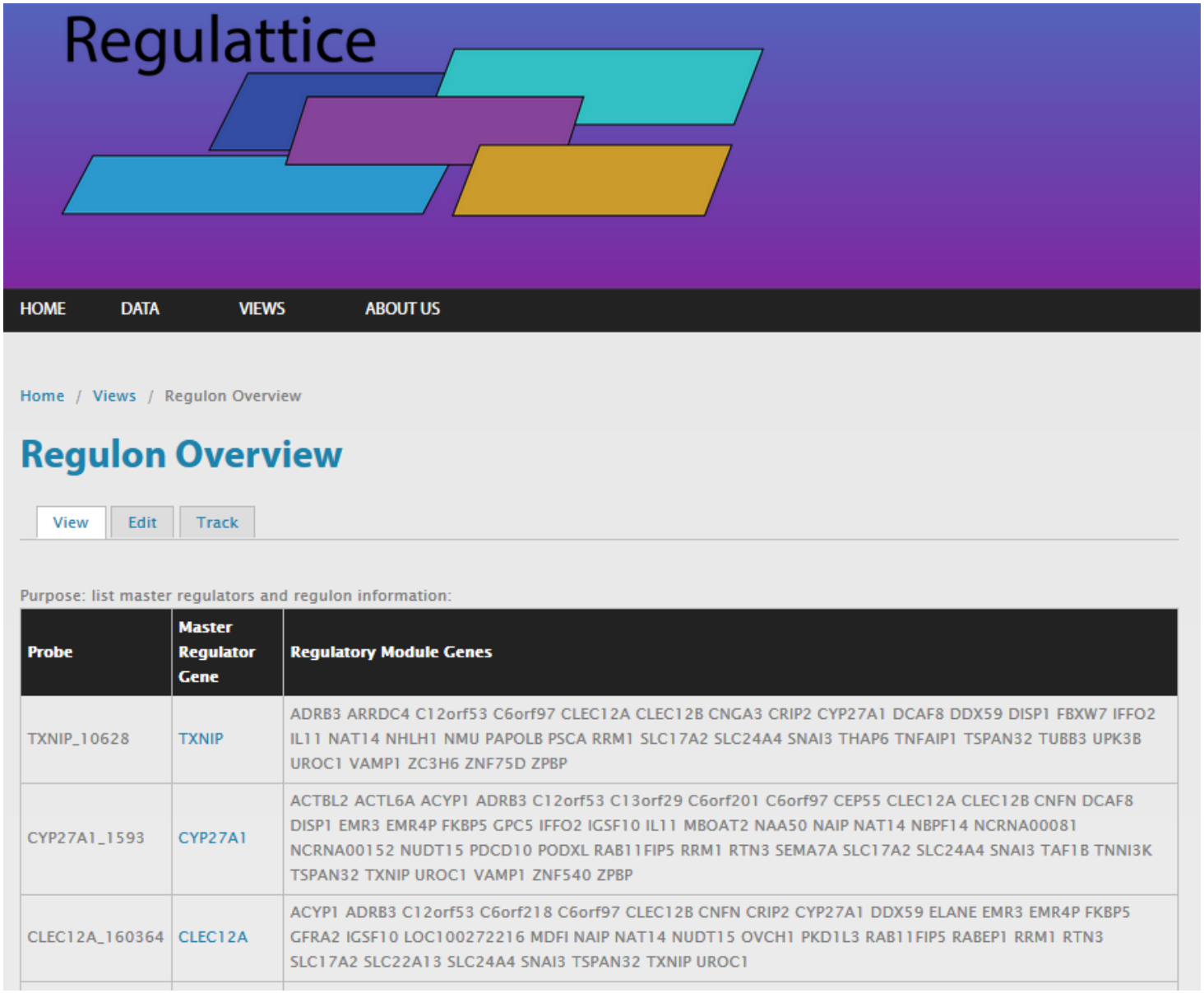
- VarImpPlot – also used in MetaboAnalyst
- Random forest imputation per regulon



Right: important driver candidates

Left: less supported regulon members

SBP
SANFORD BURNHAM PREBYS



http://regulattice.burnham.org
http://regulattice.sbpdiscovery.org (external)
Both require login/account access

SBP
SANFORD BURNHAM PREBYS

List-based regulon prioritization (All Mito list)

View | Edit | Track

Purpose: list master regulators and regulon information:

Toggle table view Curated human mitchondrial proteins from MitoCarta 2.0 (including synonyms):

#	Probe	Master Regulator Gene	Regulatory Module Genes	Matched Genes	# Matched Genes	Total # in Regulon (inc. MR)	% Match
1	TXNIP_10628	TXNIP [Int.s]	ADRB3 ARRDC4 C12orf53 C6orf97 CLEC12A CLEC12B CNGA3 CRIP2 CYP27A1 DCAF8 DDX59 DISP1 FBXW7 IFFO2 IL11 NAT14 NHLH1 NMU PAPOLB PSCA RRM1 SLC17A2 SLC24A4 SNAI3 THAP6 TNFAIP1 TSPAN32 TUBB3 UPK3B UROC1 VAMP1 ZC3H6 ZNF75D ZPBP	CYP27A1 TUBB3	2	35	5.71
2	CYP27A1_1593	CYP27A1 [Int.s]	ACTBL2 ACTL6A ACYP1 ADRB3 C12orf53 C13orf29 C6orf201 C6orf97 CEP55 CLEC12A CLEC12B CNFN DCAF8 DISP1 EMR3 EMR4P FKBP5 GPC5 IFFO2 IGSF10 IL11 MBOAT2 NAA50 NAIP NAT14 NBPF14 NCRNA00081 NCRNA00152 NUDT15 PDCD10 PODXL RAB11FIP5 RRM1 RTN3 SEMA7A SLC17A2 SLC24A4 SNAI3 TAF1B TNNI3K TSPAN32 TXNIP UROC1 VAMP1 ZNF540 ZPBP	RAB11FIP5	2	47	4.26
3	CLEC12A_160364	CLEC12A [Int.s]	ACYP1 ADRB3 C12orf53 C6orf218 C6orf97 CLEC12B CNFN CRIP2 CYP27A1 DDX59 ELANE EMR3 EMR4P FKBP5 GFRA2 IGSF10 LOC100272216 MDFI NAIP NAT14 NUDT15 OVCH1 PKD1L3 RAB11FIP5 RABEP1 RRM1 RTN3 SLC17A2 SLC22A13 SLC24A4 SNAI3 TSPAN32 TXNIP UROC1	CYP27A1 RAB11FIP5	2	35	5.71
4	C6orf97_80129	C6orf97 [Int.s]	ACTG1 ACYP1 ADRB3 C12orf53 C6orf218 CLEC12A CLEC12B CNFN CRIP2 CSTF2T CYP27A1 DCAF8 EMR3 EMR4P GFRA2 GPR17 KRT17 KRT80 NAIP NAT14 NBPF14 NCRNA00152 NRM OVCH1 PKD1L3 PMEPA1 RTN3 RYR2 SLC17A2 SLC24A4 SMEK1 SNAI3 TFAP2A THAP6 TSPAN32 TXNIP UROC1 ZC3H6	CYP27A1	1	39	2.56

SBP
SANFORD BURNHAM PREBYS

Master Regulator Enrichment in MetaCore

Viper with no ANOVA, AracNe p-value 1e-7

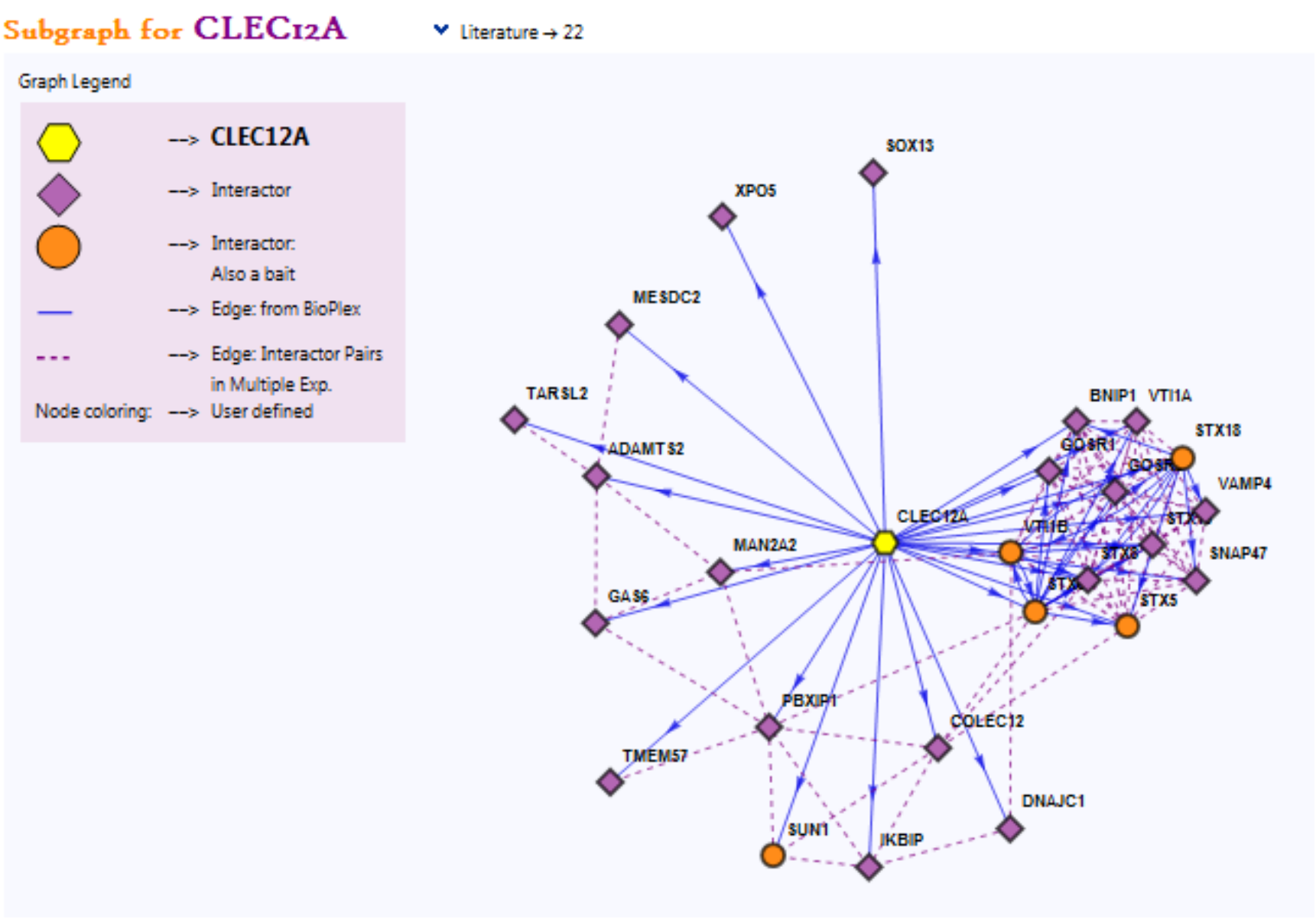
Diseases (by Biomarkers)

Export | Export to image

Total results: 10

#	Diseases	0 2 4 6 8 10 -log(pValue)	pValue	FDR	Ratio
1	Breast Neoplasms		3.553e-13	2.071e-10	222/9146
2	Breast Diseases		3.599e-13	2.071e-10	222/9147
3	Pancreatic Neoplasms		7.894e-13	2.680e-10	93/2630
4	Pancreatic Diseases		9.313e-13	2.680e-10	105/3166
5	Endocrine Gland Neoplasms		6.057e-12	1.394e-9	140/4938
6	Skin and Connective Tissue Diseases		4.266e-11	6.358e-9	267/12266
7	Carcinoma, Ductal		4.885e-11	6.358e-9	63/1566
8	Carcinoma, Squamous Cell		4.951e-11	6.358e-9	34/549
9	Neoplasms, Ductal, Lobular, and Medullary		4.971e-11	6.358e-9	64/1606
10	Skin Diseases		1.394e-10	1.558e-8	256/11689

SBP
SANFORD BURNHAM PREBYS



Bioplex interaction browser:

http://wren.hms.harvard.edu/bioplex/browseInteractions.php

SBP
SANFORD BURNHAM PREBYS

Summary

- Recommended multiple tools/resources for transcriptomics, proteomics and metabolomics analyses
- Showed information-based metrics (MI/BN) are more descriptive than Spearman/Pearson in complex datasets
- Presented viper analysis, one of several approaches to predict drivers/controllers guiding phenotype changes
- Applied Viper to understanding why pancreatic cancer progresses despite gemcitabine treatment
- Repeat the procedure for other phenotypes/cancers/samples?
  - Need your feedback/requests here!

SBP
SANFORD BURNHAM PREBYS

Acknowledgements

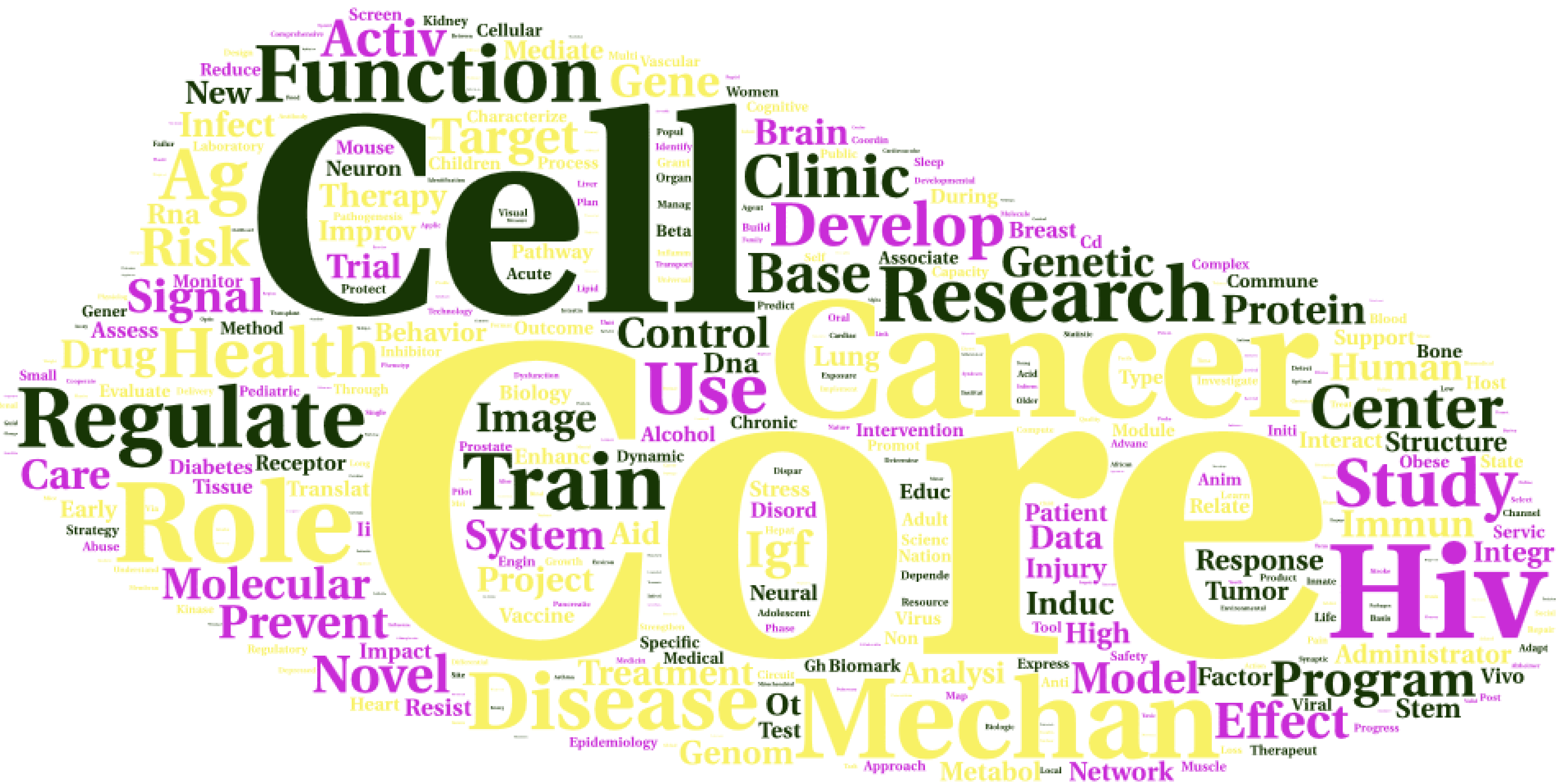
- Bioinformatics Core
  - Alexey Eroshkin
  - Vicky Guo
  - Stacy Huang
  - Craig Hauser
- Godzik Laboratory
  - Adam Godzik
  - Kai Post**
  - Andrew Leblanc**
  - Eduard Porto
  - Thomas Hrabe
- Cancer Center Support
- Powis Laboratory*
  - Garth Powis
  - Petrus De Jong
  - Claudia Miller
- Vuori Laboratory*
  - Kristiina Vuori
  - Darren (Ben) Finlay

  - Stephen Smith
  - Lauren Sparks
  - Stephanie Parsons

Appendix

SBP
SANFORD BURNHAM PREBYS

NIH Projects funded in 2015: Wordcloud based on titles



SBP
SANFORD BURNHAM PREBYS

All NIH projects (2015)

NCI projects (2015)

SBP
SANFORD BURNHAM PREBYS

TCGA major variables – clinical info (available in the Biotab annotation set)

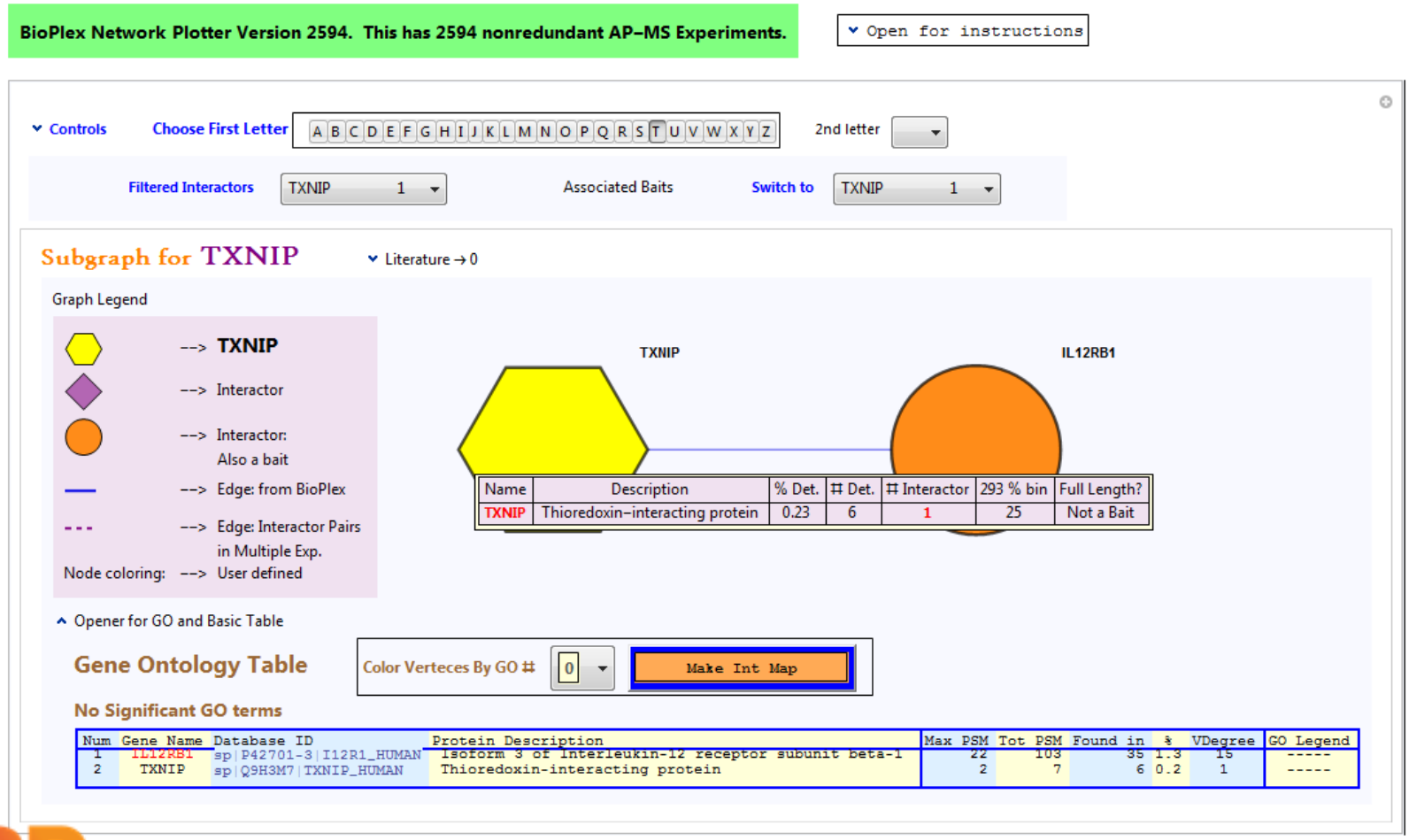
- bcr_patient_uuid – patient id: max = 23, min = 1 - match vs microarray
- bcr_patient_barcode – shortened form of patient id (TCGA id)
- bcr_drug_uuid – extended barcode (like *_patient_uuid)
- bcr_drug_bacrcode – drug barcode (also TCGA- id)
- form_completion_date
- pharmaceutical_therapy_drug_name* - common drug name
- clinical_trial_drug_classification – most n/a or not available
- pharmaceutical_therapy_type – most are chemo, followed by hormone or immunotherapy, etc.
- Pharmaceutical_tx_started_days_to
- ….

SBP
SANFORD BURNHAM PREBYS

Bioplex interaction browser:

http://wren.hms.harvard.edu/bioplex/browseInteractions.php

- Only 1 direct protein interaction identified for TXNIP (AP-MS)



SBP
SANFORD BURNHAM PREBYS

MetaCore enrichment of Master Regulators (2 cutoffs for p-values)

tabolic Networks

Export | Export to image — Total results: 10

#	Networks	0 0.25 0.5 0.75 1 1.25 1.5 1.75 -log(pValue)	pValue	FDR	Ratio
1	Lyso-Phosphatidylserine pathway		5.551e-3	6.046e-2	2/81
2	Phosphatidylinositol-4,5-diphosphate pathway		7.113e-3	6.046e-2	2/92
3	Vitamin, mediator and cofactor metabolism_Vitamin D3		6.121e-2	2.133e-1	1/44
4	Steroid metabolism_Cholesterol metabolism		1.036e-1	2.133e-1	1/76
5	GalNAcbeta1-3Gal pathway		1.216e-1	2.133e-1	1/90
6	Sphingomyelin pathway		1.292e-1	2.133e-1	1/96
7	N-acyl-sphingosine phosphate pathway		1.330e-1	2.133e-1	1/99
8	Acetyl-L-carnitine pathway		1.343e-1	2.133e-1	1/100
9	Lipid metabolism_Triacylglycerol metabolism		1.355e-1	2.133e-1	1/101
10	Glutamic acid pathway		1.380e-1	2.133e-1	1/103

back to to

Processes

Export | Export to image — Total results: 10

#	Processes	0 0.5 1 1.5 2 2.5 3 3.5 4 -log(pValue)	pValue	FDR	Ratio
1	fast, calcium ion-dependent exocytosis of neurotransmitter		6.413e-6	9.741e-3	2/7
2	detection of calcium ion		7.038e-6	9.741e-3	3/15
3	negative regulation of SREBP signaling pathway		1.921e-5	1.329e-2	2/3
4	regulation of SREBP signaling pathway		1.921e-5	1.329e-2	2/3
5	regulation of cell migration involved in sprouting angiogenesis		3.080e-5	1.705e-2	3/24
6	negative regulation of triglyceride biosynthetic process		3.835e-5	1.736e-2	2/4
7	regulation of sprouting angiogenesis		4.950e-5	1.736e-2	3/28
8	regulation of potassium ion transmembrane transport		5.501e-5	1.736e-2	4/81
9	dicarboxylic acid transport		6.053e-5	1.736e-2	4/83
[illegible]	[illegible] of oxidative stress-induced neuron intrinsic apoptotic signaling pathway		6.381e-5	1.736e-2	2/5

SBP
SANFORD BURNHAM PREBYS

Diseases (by Biomarkers)

Export | Export to image

Total results: 10

#	Diseases	0 0.5 1 1.5 2 2.5 3 3.5 -log(pValue)	pValue	FDR	Ratio
1	Neoplasms, Complex and Mixed		2.991e-5	2.778e-2	7/347
2	Thymoma		3.856e-4	1.134e-1	3/55
3	Thymus Neoplasms		6.888e-4	1.134e-1	3/67
4	Cholangiocarcinoma		7.529e-4	1.134e-1	4/159
5	Thromboembolism		9.568e-4	1.134e-1	3/75
6	Dysplastic Nevus Syndrome		1.080e-3	1.134e-1	2/19
7	Nervous System Diseases		1.122e-3	1.134e-1	26/5717
8	Bronchitis		1.196e-3	1.134e-1	3/81
9	Atrial Fibrillation		1.239e-3	1.134e-1	3/82
10	Beckwith-Wiedemann Syndrome		1.452e-3	1.134e-1	2/22

back to top

Toxicity Networks

Export | Export to image

Total results: 10

#	Networks	0 0.25 0.5 0.75 1 1.25 1.5 1.75 -log(pValue)	pValue	FDR	Ratio
1	Chemotaxis_NF-kB regulation		8.055e-3	2.426e-1	2/43
2	Transport_Vesicle-mediated transport_membrane biogenesis		1.732e-2	2.426e-1	2/64
3	Protein folding_ATFs regulation		2.052e-2	2.426e-1	2/70
4	Cell cycle_APC regulation of G1_S		2.762e-2	2.426e-1	2/82
5	Protein folding_Calcineurin, FKBPs		5.373e-2	2.426e-1	1/17
6	Protein folding_p53 regulation		6.902e-2	2.426e-1	1/22
7	Cell cycle_G1_S transition of mitotic cell cycle		9.005e-2	2.426e-1	1/29
8	Blood coagulation_Fibrinogen signaling		9.597e-2	2.426e-1	1/31
9	Signal transduction_Neuropeptide signaling pathway_Substance P, GRP, Neuromedulin B, Neuropeptide FF, Urotensin-2		1.048e-1	2.426e-1	1/34
10	Inflammation_Kallikrein signaling		1.106e-1	2.426e-1	1/36

Pathway Maps

Export | Export to image

Total results: 10

#	Maps	-log(pValue)	pValue	FDR	Ratio
1	NETosis in SLE		2.080e-3	1.504e-1	2/31
2	Development_Positive regulation of STK3/4 (Hippo) pathway and negative regulation of YAP/TAZ function		1.025e-2	1.504e-1	2/70
3	HBV regulation of DNA repair and apoptosis leading to HCC		3.045e-2	1.504e-1	1/14
4	Involvement of VEGF signaling in the progression of lung cancer		3.686e-2	1.504e-1	1/17
5	Development_Role of proteases in hematopoietic stem cell mobilization		3.899e-2	1.504e-1	1/18
6	Development_WNT signaling pathway. Part 1. Degradation of beta-catenin in the absence WNT signaling		4.111e-2	1.504e-1	1/19
7	Ethanol/Acetaldehyde-dependent stimulation of MMP-9 expression in HCC		4.111e-2	1.504e-1	1/19
8	PR action in breast cancer: stimulation of metastasis		4.323e-2	1.504e-1	1/20
9	Development_Role of G-CSF in hematopoietic stem cell mobilization		4.534e-2	1.504e-1	1/21
10	Role of ZNF202 in regulation of expression of genes involved in atherosclerosis		4.534e-2	1.504e-1	1/21

back to top

Process Networks

Export | Export to image

Total results: 10

#	Networks	-log(pValue)	pValue	FDR	Ratio
1	Transport_Synaptic vesicle exocytosis		2.143e-3	1.007e-1	4/175
2	Development_Neurogenesis_Synaptogenesis		1.925e-2	2.840e-1	3/180
3	Cell adhesion_Synaptic contact		2.040e-2	2.840e-1	3/184
4	Cell adhesion_Cell-matrix interactions		2.914e-2	2.840e-1	3/211
5	Reproduction_Progesterone signaling		3.022e-2	2.840e-1	3/214
6	Proteolysis_Proteolysis in cell cycle and apoptosis		6.055e-2	3.761e-1	2/125
7	Signal transduction_Androgen receptor nuclear signaling		6.140e-2	3.761e-1	2/126
8	Chemotaxis		7.112e-2	3.761e-1	2/137
9	Neurophysiological process_GABAergic neurotransmission		7.203e-2	3.761e-1	2/138
10	Development_Neuromuscular junction		8.036e-2	3.777e-1	2/147

SBP
SANFORD BURNHAM PREBYS

MitoCarta Provenance (Known mitochondrial proteins)

10	**Below are descriptions for the columns (in tables A,B):**	
11	TrainingDataset	Tmito, Tnot_mito, or T_possible_mito (indicating evidence based on NCBI GO mitochondrial annotation or MitoP2 database, bu
12	HumanGeneID	NCBI Entrez Gene ID for human ortholog of mouse gene (based on reciprocal BLASTP hit, expect<1e-3)
13	MouseOrthologGeneID	NCBI Entrez Gene ID for mouse gene
14	Symbol	Official Entrez gene symbol
15	Synonyms	NCBI Entrez gene synonyms
16	Description	NCBI Entrez gene description
17	EnsemblGeneID	Ensembl Gene Identifier (based on Ensembl mapper)
18	ProteinLength	Length of longest RefSeq protein isoform
19	TargetP_Score	TargetP confidence score (1-5, 1 is most confident) of mitochondrial targeting signal. Score 0 indicates no mitochondrial predic
20	MitoDomain_Score	MitoDomain indicates exclusively mito domain; SharedDomain indicates shared mito & non-mito domain; NonMitoDomain indi
21	CoexpressionGnfN50_Score	N50 score for coexpression across mouse GNFv3 tissue atlas (N50 = number of 50 nearest neighbors that are in Tmito)
22	PGC_Induction_Score	Foldchange in transcript induction following overexpression of PGC1a, known to induce mitochondrial biogenesis
23	YeastMitoHomolog_Score	OrthologMitoHighConf (HomologMitoHighConf) if yeast ortholog (homolog) is annotated mitochondrial in SGD with high confide
24	RickettsiaHomolog_Score	Ortholog if 1:1 ortholog between mouse and *Rickettsia prowazekii*; Homolog if homology between mouse and *Rickettsia prowa*
25	MSMS_Score	Category indicating MS/MS abundance (coverage) and enrichment in subtractive proteomics (either pure-enriched, crude-enrich
26	MitoCarta2_score	Naïve Bayes score (sum of log-likelihood ratios for each of 7 above features)
27	MitoCarta2_FDR	Estimated corrected false discovery rate for all predictions >= given score
28	MitoCarta2_List	1 if gene is on final MitoCarta2.0 list
29	MitoCarta2_Evidence	Type of experimental support
30	hg19_Chromosome	Chromosome (hg19 assembly)
31	hg19_Start	Start position (hg19)
32	hg19_Stop	Stop position (hg19)
33	MSMS_NUM_TISSUES	Number 0-14 tissues where gene products were found by MS/MS
34	MSMS_NUM_PEPTIDES_UNIQ	Number of unique peptides, based on pooling spectra from 14 tissues
35	MSMS_NUM_SPECTRA	Number of spectra corresponding to this gene, based on pooling spectra from 14 tissues
36	MSMS_TOTAL_INTENSITY	total peak intensity, based on pooling spectra from 14 tissues
37	MSMS_PERCENT_COVERAGE	Coverage (percent of amino acids covered by MS/MS spectra), based on pooling spectra from 14 tissues
38	Tissues	List of tissues with evidence of protein, based on 14 tissues

From Human.MitoCarta2.0.xls file, 1st tab. 2nd tab is filtered 1158 human genes