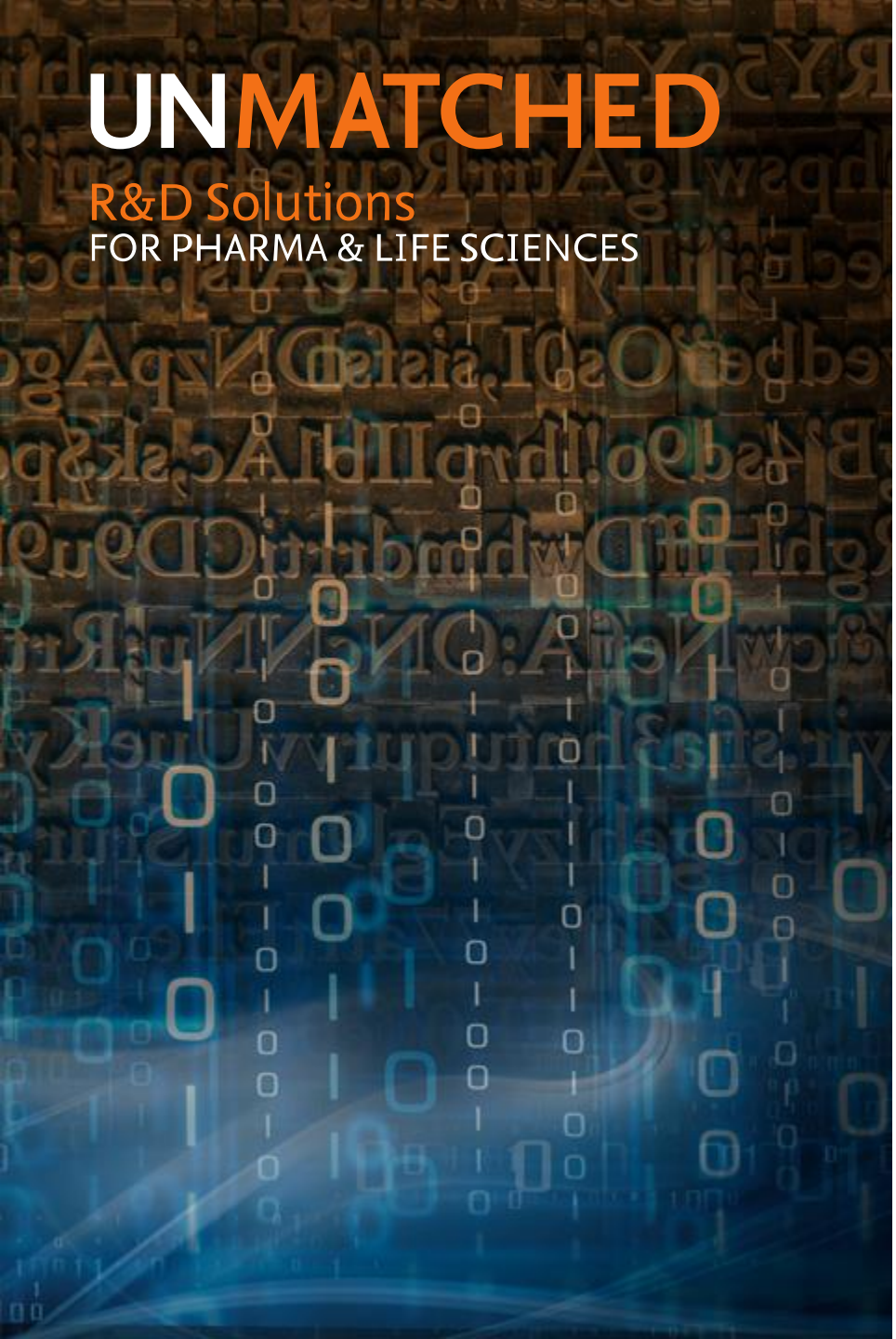


ELSEVIER

Pathway Studio-Assisted Biological Research

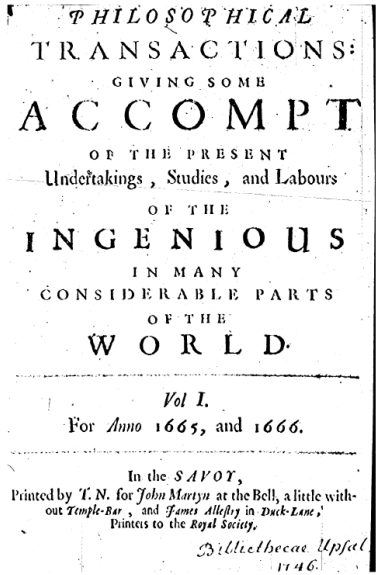
Integration of Tools, Data and Expertize to solve a customer problem

Stephen Sharp, Ph.D.

September 4, 2017

Scientific information and data exchange in practice for over 400 years

- Most scientific data is still published in unstructured format
  - 17th Century
  - 21st Century



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

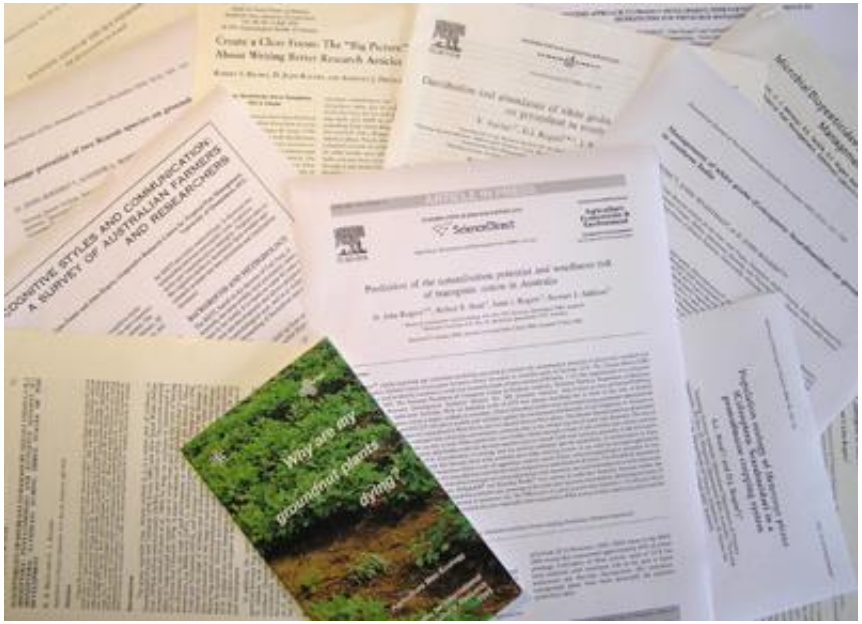
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

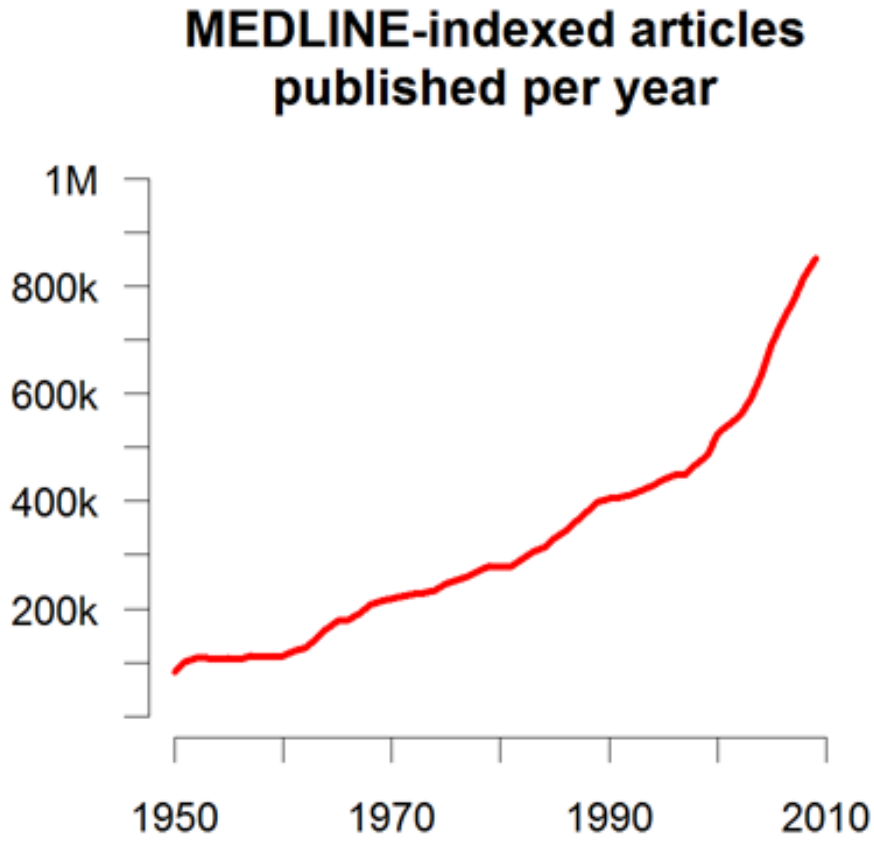
Bibliothecae Upsal
1746.

Royal Society of London
Oldest learned society (1660)
Oldest scientific journal (1665)

Scientific literature is exploding

- More than 1M new citations/year in Medline – HOW TO KEEP UP?

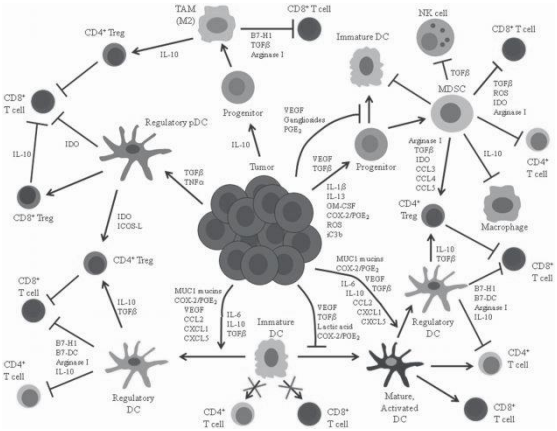




If only we knew what is known…

Text mining: analyzing text to extract information that is useful for particular purposes

Text mining



Amorphous information

- Hard to deal with
- Hard to deal with algorithmically

- Not scalable

Structured information

- Search
- Visualize
- Network analysis

- Scalable
- Compressed

Image Source: http://www.thesocialleader.com/wp-content/uploads/2011/03/paper-piles.jpg

History of MedScan Technology

- 2001- Ariadne Genomics developed **MedScan** – a tool to extract information for bio-molecular networks from literature

- 2012 - Elsevier acquired Ariadne Genomics and continued to develop **MedScan** and **PATHWAY STUDIO**

- **2017- MedScan** engine can be used within **PATHWAY STUDIO** and also independently

BIOINFORMATICS

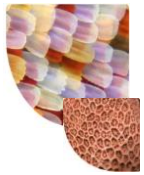
Vol. 19 no. 13 2003, pages 1699–1706
DOI: 10.1093/bioinformatics/btg207

MedScan, a natural language processing engine for MEDLINE abstracts

*Svetlana Novichkova, Sergei Egorov and Nikolai Daraselia**

Ariadne Genomics, Inc., 9100 Great Seneca HWY, Rockville, MD 20850, USA

Received on January 11, 2003; revised on February 11, 2003; accepted on March 17, 2003



PATHWAY STUDIO®

Human
Mouse
Rat

From text to fact

text

24 M abstracts
3.5 M full texts

sentence

Our results suggest that a natural immune response mediated by γδ T lymphocytes may contribute to the immunosurveillance of melanoma

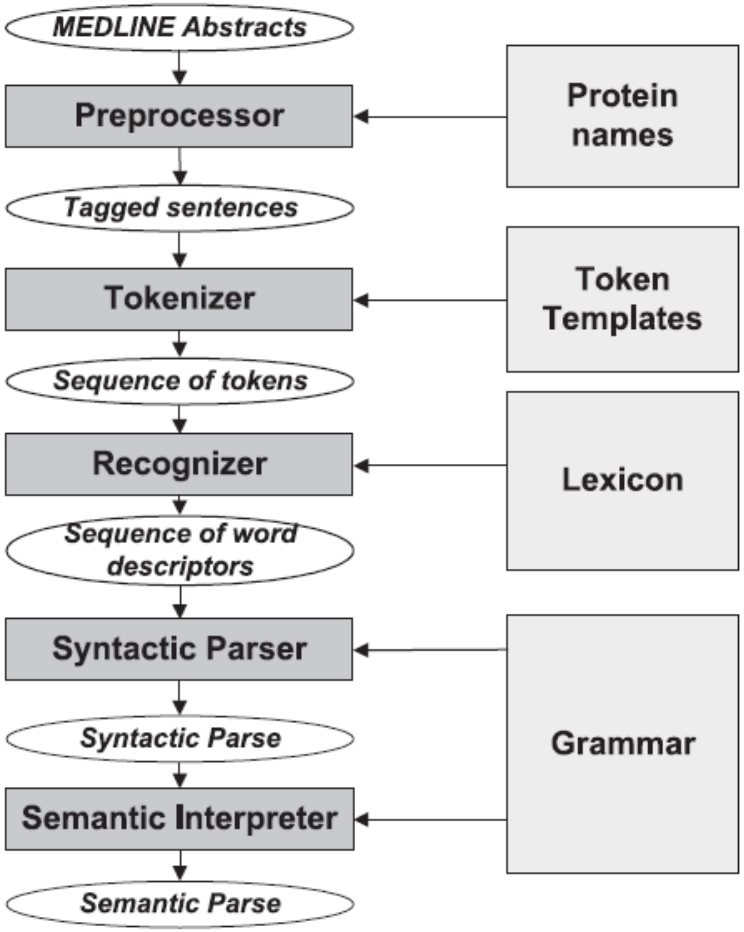
fact

standard name → standard link → standard name

5.6 M facts

Processing of Text using MedScan Technology

- Staged Processing:
  - Sentence tokenizing
  - Word recognition
  - Morphological analysis
  - Recognition of compound lexemes
  - Syntactic parsing
  - Semantic interpretation.



Text Mining with MedScan Technology

Natural Language Processing

The central idea of **MedScan´s** NLP algorithm is decomposing natural language sentences into semantic relationships (which we will also call semantic triplets). Each triplet is designed to represent a single semantic relationship between two singular noun phrases (NPs).

11940574:7 Because **Axin2** has been shown to associate with and inhibit **beta-catenin** abundance and function, we hypothesized that **Axin2**, which is affecting proliferation of MEF cells can work in a negative feedback pathway, regulating **Wnt** signaling and thus controlling apoptotic process.

Triplets identified:

Axin2 associate beta-catenin abundance
Axin2 inhibit beta-catenin function
Axin2 associate beta-catenin abundance
Axin2 inhibit beta-catenin function
Axin2 affect MEF cell line proliferation
Axin2 work negative feedback pathway
Axin2 regulate Wnt signaling
Axin2 control apoptotic process

Data Extraction using Elsevier NLP

How MedScan Identifies “Entities” and “Relationships”

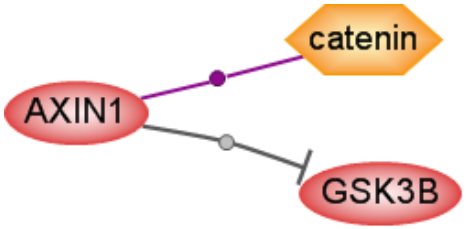
“Axin binds beta-catenin and inhibits GSK-3beta activity in hepatocytes.”

Text Mining recognizes “Semantic Triplets”

Entity	Relationship	Entity
Axin	binds	**beta-catenin**
Axin	inhibits	**GSK-3beta** activity

Stores this information in the database…

Pathway Studio visualizes the information in graphical format

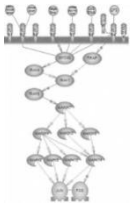


Axin - beta-catenin, relation: Binding celltype: hepatocyte
Axin -> GSK-3beta, relation: Regulation, effect: Negative celltype: hepatocyte

Solution overview

Pathway Studio®

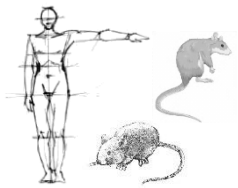
Knowledgebase

Biological relations extracted from literature	Manually curated pathways	Ontologies Annota-tions	Structural similarity for chemicals	Variation annotation from public sources
24M abstracts 3.5M full texts  6.2M relations	 >1800 pathways	 NCBI 		

Tools

Search Summarization Navigation Visualization	Experiment analysis: • Gene expression • Proteomics • Metabolomics • NGS	Text mining pipeline Easy to use text mining interface
	 	

Pathway Studio databases



Mammal — Protein centered

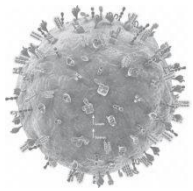
- Target interactions data
- Expression regulation
- Transport
- Synthesis
- Protein modification
- Functional characterization

Chem Effect — Chemical centered

- Chemical-target interactions
- Adverse events
- Drug repurposing

Disease FX — Disease centered

- Biomarkers
- Drug repurposing
- Clinical trials

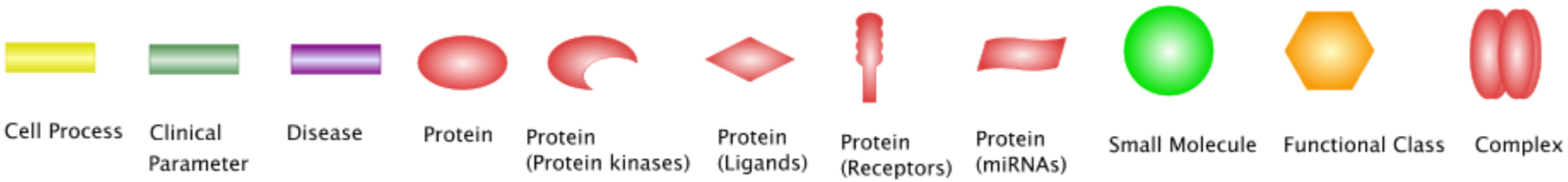


Cells — Cell centered

- Cell expression
- Cells role in the disease
- Cells as biomarkers

Types of objects (Entities) in the database

Entity	Example
Protein, Complex, Functional class	PARP1, miR133b, IL23, cytokine
Drug, Metabolite, Ion, Chemical	glucose, aspirin, estradiol, cetuximab
Disease	malaria
Clinical Parameter	heart rate, lesion size, Gleason score
Cell Process	apoptosis
Treatment	heat shock
Cell	regulatory T cell



Integrating text with other resources

All recognized terms have identifiers:

- Linking to other databases
- Mapping
- Integrating external data
- Use of ontologies

IUPAC Name:	4-[4-[[4-chloro-3-(trifluoromethyl)phen
Molecular Formula:	C21H16ClF3N4O3
PharmaPendium ID:	Sorafenib Tosylate
Rotatable Bond Count:	5
XLogP-AA:	4.1
Molecular Weight:	464.82495

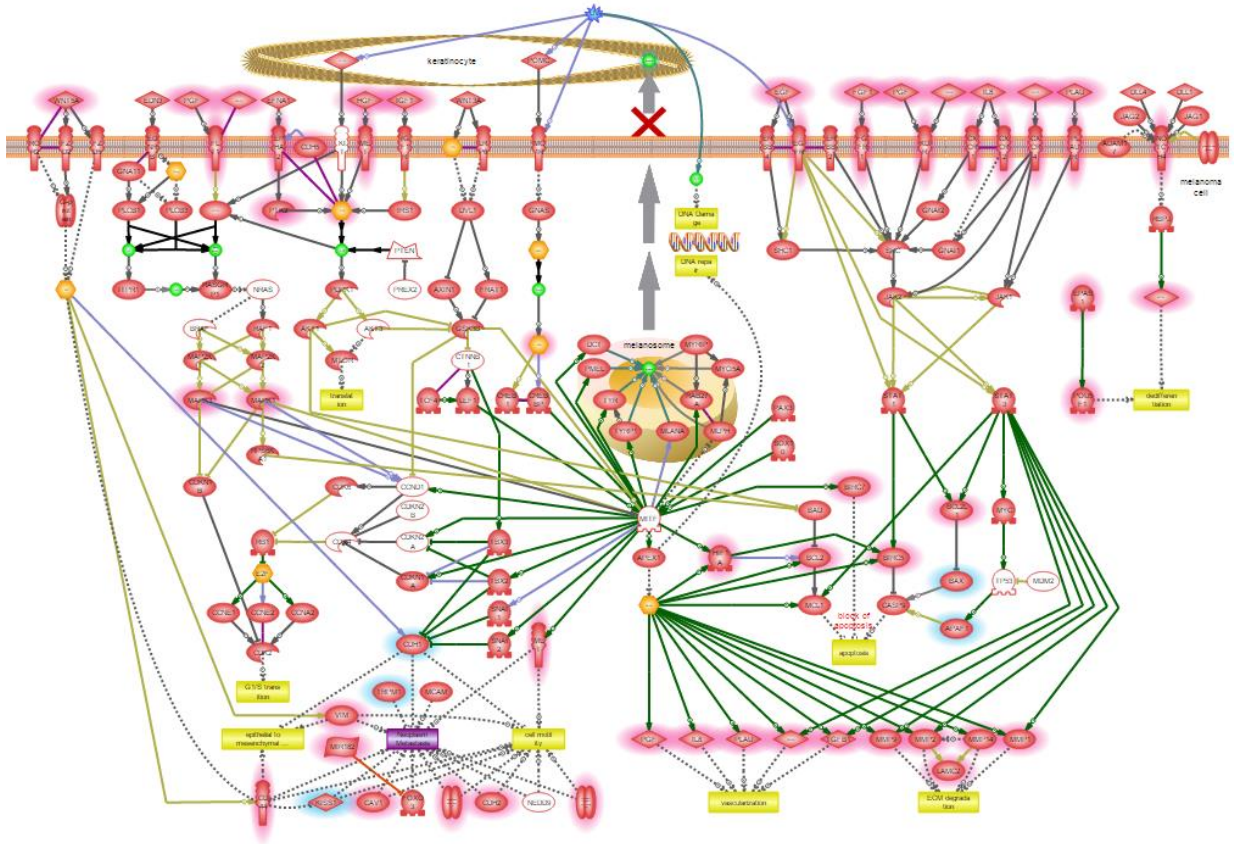
Notes:	This gene encodes a chromatin- proliferation, and tumor transfo [provided by RefSeq, Jul 2008]
Alias:	5830444G22; 5830444G22Rik; A polymerase); ADP-ribosyltransfe ADPRT 1; ADPRT I; ADPRT1; AI8 ADP-ribosyltransferase, nuclear; nuclear NAD ADP-ribosyltransfe polymerase; poly [ADP-ribose] p ribosyl)transferase; poly[ADP-rib
Connectivity:	5614
Primary Cell Localization:	Nucleus
Cell Localization:	Nucleus
Organism:	Homo sapiens
Human chromosome position:	1q41-q42
Rat chromosome position:	13q26
Mouse chromosome position:	1 84.44 cM
Owner:	public
URN:	urn:agi-llid:142
Date Created:	2015-12-18 06:22:20.026
Date Modified:	2015-12-18 06:22:20.027

Entrez GeneID:	11545; 142; 25591
Unigene ID:	Hs.177766; Mm.277779; Rn.11327
Swiss-Prot Accession:	A0A024R3T8; B1ANJ4; O35937; P0
Swiss-Prot ID:	PARP1_HUMAN; PARP1_RAT
OMIM ID:	173870
MGI ID:	MGI:1340806
Hugo ID:	270; HGNC:270
RGD ID:	2053
Ensembl ID:	ENSG00000143799; ENSMUSG000
GenBank ID:	AAA51599; AAA51663; AAA60000; ABM85752; AC121810; AC_000023; AK312339; AL359704; AL359742; A CAA46478; CBX74363; CH466555; NM_001618; NM_007415; NM_013
GO ID:	0000122; 0000723; 0003677; 0003 0023019; 0030225; 0032869; 0040
EC Number:	2.4.2.30
KEGG ID:	hsa:142; rno:25591
Homologene ID:	1222
PIR ID:	A29725; S21163; S26057
MedScan ID:	142

Rich pathway collections for modeling and experiment analysis

Over 1,800 pathways manually built by PhD level scientists (ongoing project)

- Signaling
- Metabolic
- Cell processes
- Disease
- Immunological
- Expression targets
- Toxicity
- Nociception



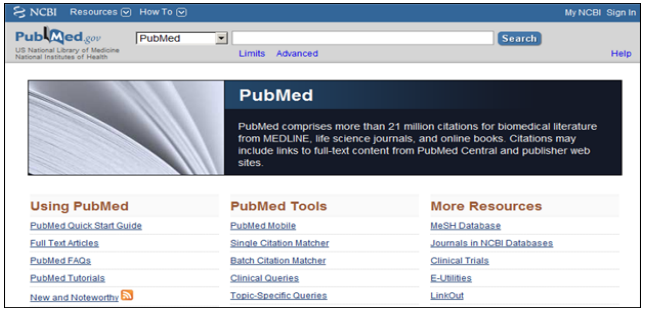
To use

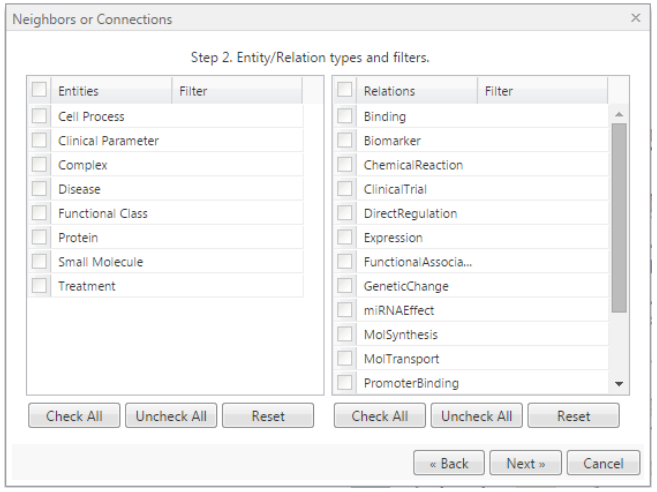
- As starting points to build pathways
- In experiment analysis
- In analysis of groups

Interface and tools to find answers to complex biological questions

- Visualization
- Summarization
- Intersect, subtract, and union facts and lists

- Filter by
  - Identifiers
  - Bibliography
  - Tissues, organs, cells, cell cultures, organisms
  - Drugability



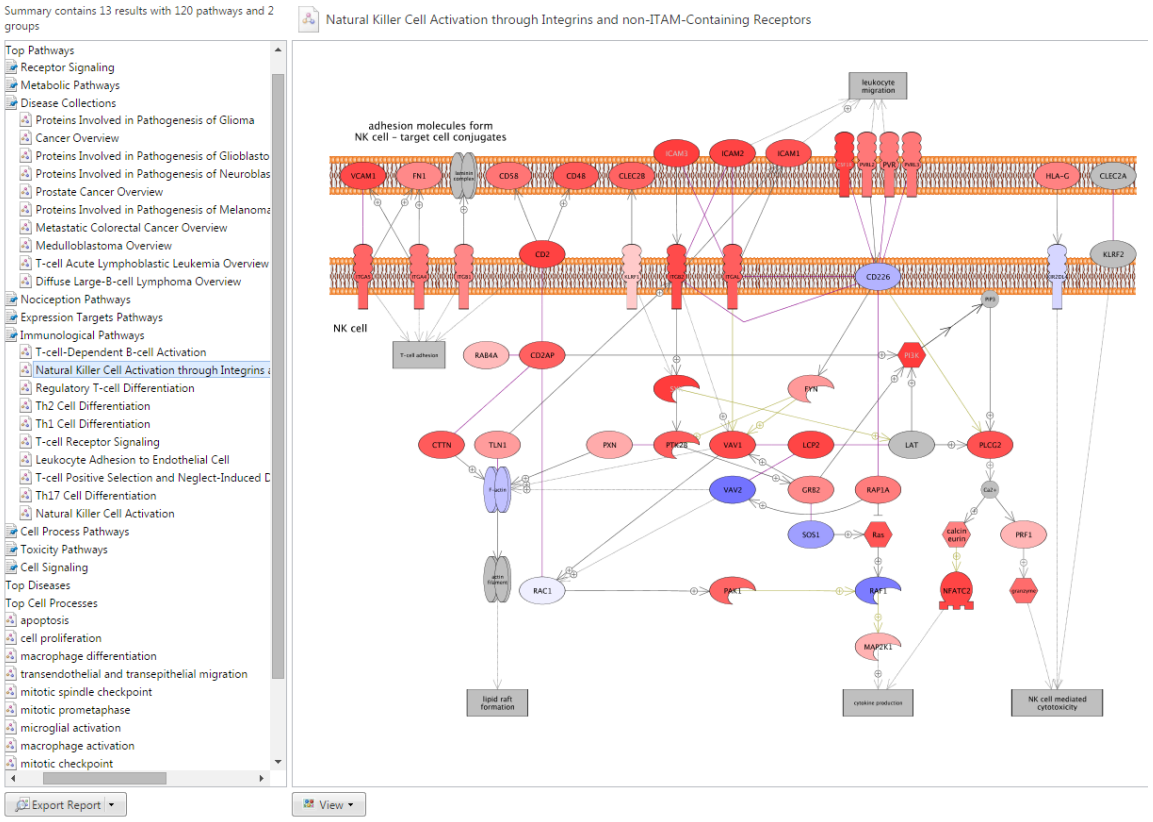


//CODiE//
2015 SIIA CODiE WINNER

Interface and tools to upload and analyze experimental data

- Upload experimental data or a list of concepts
- Calculate differential values

- Gene-Set Enrichment Analysis: interpret the data using Pathway Studio ontologies and pathway collections
- Sub-Network Enrichment Analysis (causal reasoning): interpret the data using 6.2 M biological relations extracted from the literature



//CODiE//
2015 SIIA CODiE WINNER

Elsevier R&D Solutions: data, software, capabilities

Content & Data

Elsevier and non-Elsevier textual information

Public and proprietary databases

Capabilities

- *Data extraction*
- *Data normalization*
- *Data integration*

Research & Discovery | **Lead ID & Valid** | **Pre-clinical** | **Clinical** | **Post-launch**

Databases Software

Research & Discovery	Lead ID & Valid	Pre-clinical	Post-launch
• *Disease modeling* • *Target identification* • *Biomarker discovery* • *Drug repositioning*	• *Lead identification and characterization* • *Synthesis optimization*	• *Lead prioritization for safety, delivery and efficacy* • *Translational medicine*	• *Monitoring adverse events*
Pathway Studio	Reaxys® Reaxys Medicinal Chemistry	PharmaPendium®	Embase®

Use-case centered integration and customization

Scopus

Professional Services

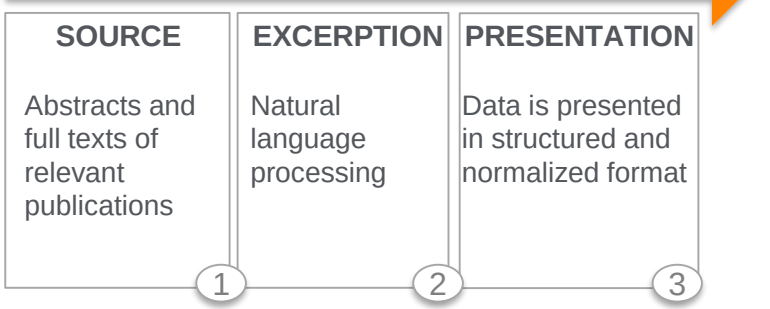
Elsevier Text Mining

ELSEVIER

Literature-extracted biology data: from disease mechanisms to targets

- What causes the disease of interest?
- What is the disease mechanism?
- What is it similar to?

Information published by researchers



Pathway Studio database contains:

>285,000 Biological concepts	**>7 MILLION** Relations between biological concepts	**>1,800** Manually built pathways, including disease models

The unique collection of data is sourced from:

>10,000 Journals, including **>1,700** Full text journals	**>4 MILLION** Full text articles	**>38 MILLION** Supporting sentences

Pathway Studio®

Rare diseases – when every piece matters

Nick Sireau at TEDx ImperialCollege
https://www.youtube.com/watch?v=B4UnVlU5hAY

- **Patients community**
- **Collaboration with medical researchers**
- **Drug repurposing candidate**
- **Fundraising**
- **Clinical Trial**

findacure

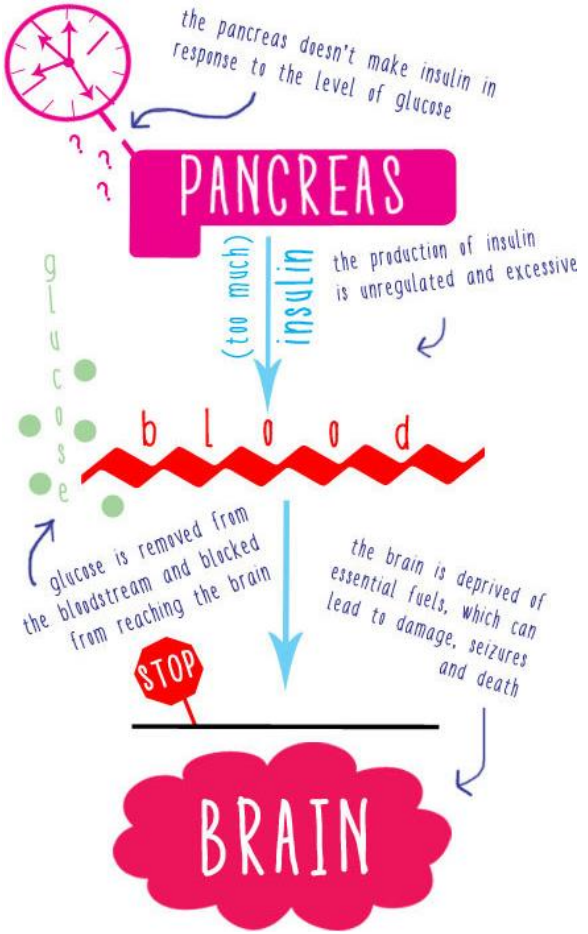
is a UK charity that is building the rare disease community to raise awareness, drive research and develop treatments.

ELSEVIER

is partnering with Findacure scientists to help identify and evaluate treatments for congenital hypersinsulinism

Congenital hyperinsulinsm (CHI)

- **A rare genetic disease**
- **Permanently excessive level of insulin in the blood**
- Develops within the first few days of life
- Can lead to brain injury or even death
- In the most severe cases the only viable treatment is the removal of the pancreas, consigning the patient to a lifetime of diabetes



https://res.cloudinary.com/indiegogo-media-prod-cld/image/upload/c_limit,w_620/v1440424745/uzvnqzhvbpsrtthzxqpu.jpg

How can we help?

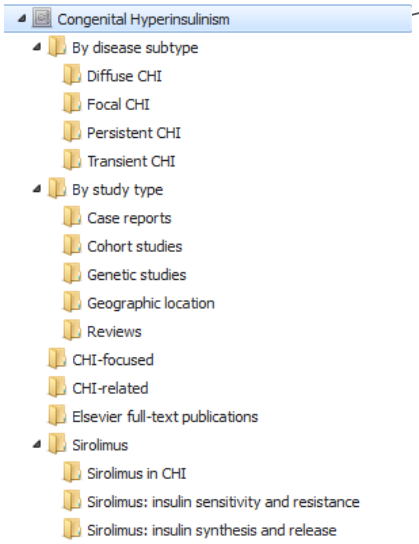
ELSEVIER

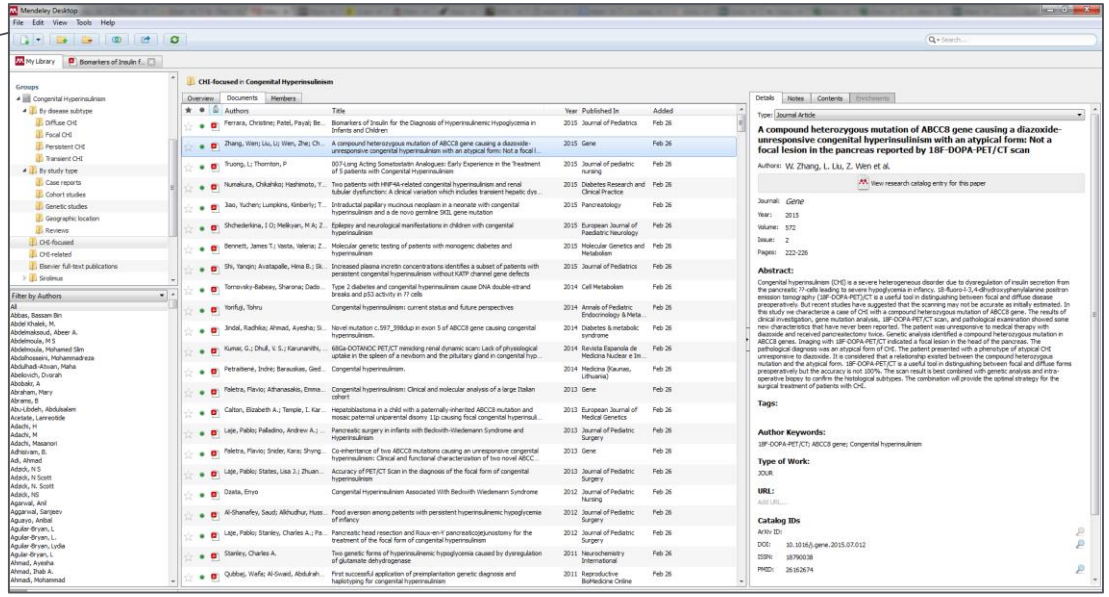
Congenital hyperinsulinism library

In support of Findacure's mission of education and knowledge sharing:

- Access to all Elsevier's ScienceDirect full-text publications covering CHI
- Collection of papers focused on different aspects of CHI
- Collection of papers focused on effects of sirolimus on CHI

ScienceDirect
Mendeley





Creating a comprehensive view of CHI with Elsevier R&D Solutions

- **CHI Library**
- **Disease, Target, Pathway, and Compound Analysis**
- **Research Landscape Analysis**

Information Assets Applied

- **Content**

  Elsevier's vast set of literature and patent data

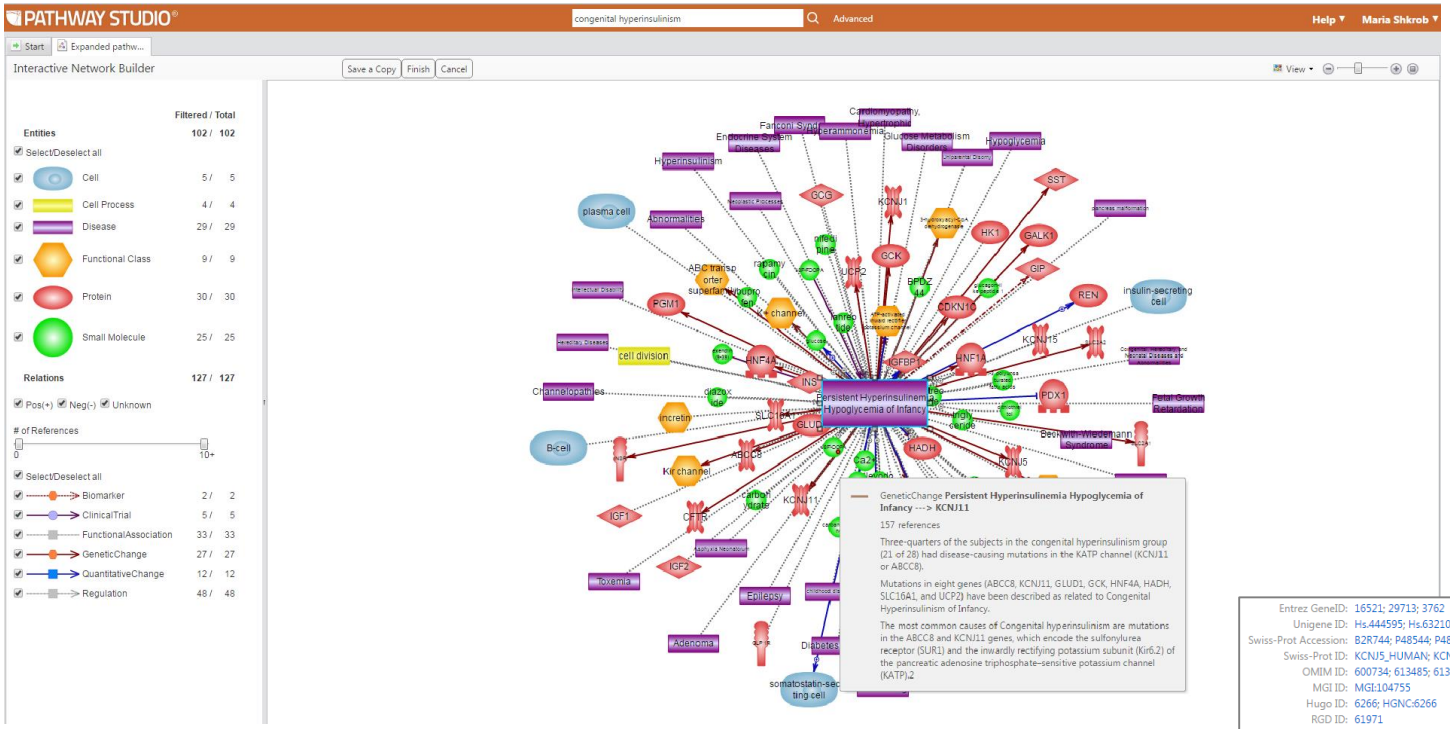
- **Data normalization**

  Taxonomies and dictionaries to normalize author names, institutions, drugs, targets, and other important terms

- **Information extraction**

  Finding semantic relationships, targets, pathways, drugs, and bioactivities

CHI: summarization and visualization of the findings



Pathway Studio®

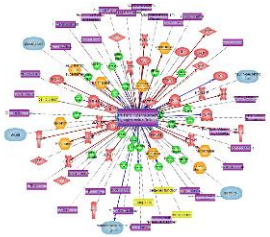
- Visualization and summarization of >7M literature findings
- Linking to non-literature sources

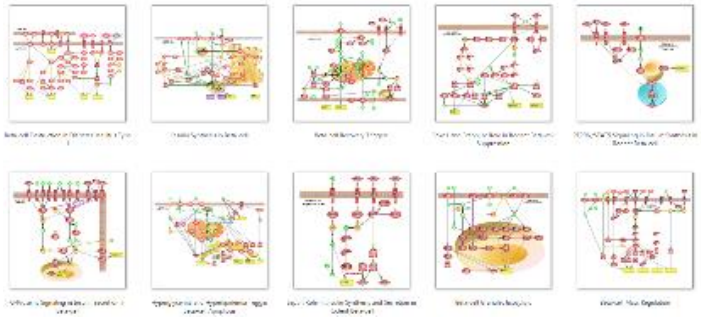
Primary Cell Localization:	Plasma membrane
Cell Localization:	Membrane
Organism:	Homo sapiens
Human chromosome position:	11q24
Rat chromosome position:	8q21
Mouse chromosome position:	9 17.65 cM

Entrez GeneID:	16521; 29713; 3762
Unigene ID:	Hs.444595; Hs.632109; Mm.69472; Rn.10047
Swiss-Prot Accession:	B2R744; P48544; P48545; P48548; P97508; Q3
Swiss-Prot ID:	KCNJ5_HUMAN; KCNJ5_MOUSE; KCNJ5_RAT
OMIM ID:	600734; 613485; 613677
MGI ID:	MGI:104755
Hugo ID:	6266; HGNC:6266
RGD ID:	61971
Ensembl ID:	ENSMUSG00000032034; ENSMUSP00000034
GenBank ID:	AAB01687; AAB07269; AAB53093; AAC42048 ABBA01023015; AC127149; AC132114; AC_0 BC069482; BC069499; BC069571; BC074838; L35771; L47208; NC_000011; NC_000075; NC XP_005271600; XP_006510102; XP_006510
GO ID:	0005242; 0005244; 0005886; 0005887; 0006
KEGG ID:	hsa:3762; mmu:16521; rno:29713
Homologene ID:	20248
PIR ID:	G02232; S48077

Building and refining the disease model

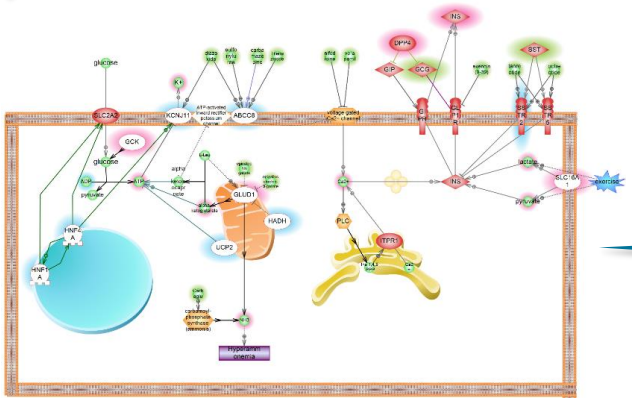
Literature-extracted data





Relevant pathways
(from a collection of 1800 models)

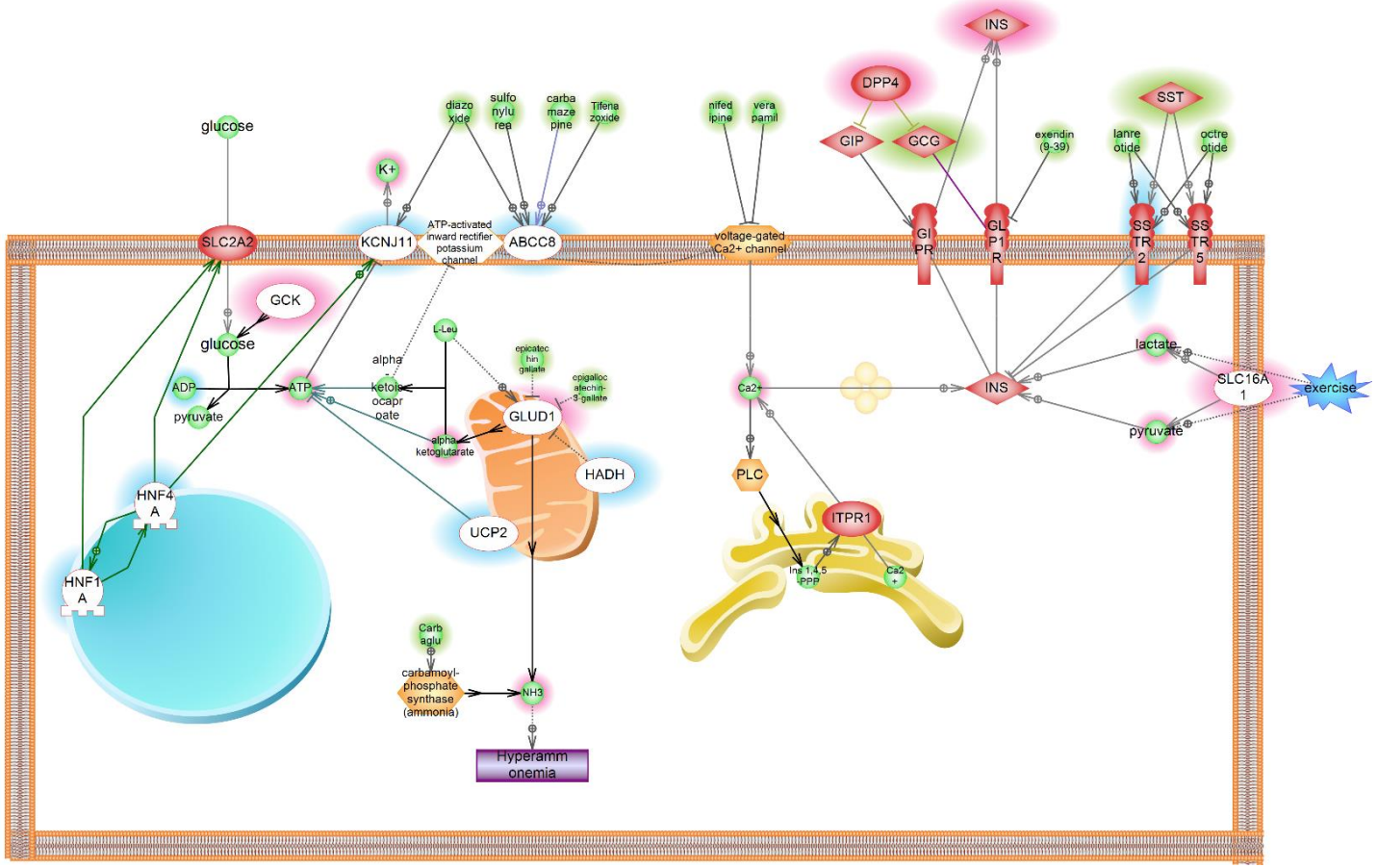
- Diseases
- Cell processes
- Signaling
- Metabolism
- Toxicity



Disease overview

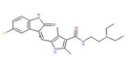
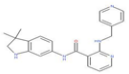
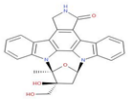
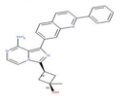
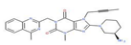
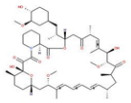
CHI: Building and refining the disease model

- Insulin secretion steps affected by CHI
- Role of mutated genes
- Drug targets
- Drugs



Approved compounds that may treat CHI

- Each binds to one or more targets related to the disease
- Can easily be obtained and tested in preclinical studies
- List includes compounds known to treat hyperinsulinism

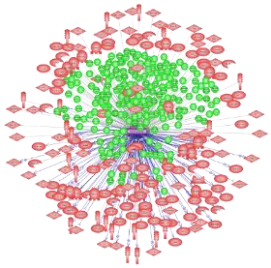
Molecule	N	NumberIn Group	Reaxys Registry Number (IDE.XRN)	Target	pX (DAT.PAUREUS)_Median	Chemical Name (IDE.CN)
	125	5	9364276	AKT1 ROCK1 IGF1R SGK1 FAK	7.1300 6.3400 6.2550 6.4750 6.3600	N-[2-(diethylamino)e... [3H]-Sunitinib Sunitinib 5-(5-Fluoro-2-oxo-1,... sunitinib sunitynib SU 11248
	136	4	11751576	AKT1 JNK1 ROCK1 SGK1	7.1100 7.8300 7.3800 7.4200	[14C]-Motesanib Motesanib motesanib N-(2,3-dihydro-3,3-d... N-(3,3-dimethyl-2,3-... N-(3,3-dimethylindol... AMG 706
	193	3	9305136	JNK1 ROCK1 IKKb	7.9600 7.2500 7.9200	Lestaurtinib (9S-(9α,10β,12α))-2,... CEP-701 CEP701 lestaurtinib
	231	3	18476426	INSR ROCK1 IGF1R	7.5100 6.2000 7.6200	Linsitinib OSI-906 cis-3-[8-amino-1-(2-... linsitinib cis-3-[8-amino-1-(2-... cis-3-[8-amino-1-(2-...
	366	2	11300181	Dipeptidyl peptidase 4 Dipeptidyl peptidase 4|Dipeptidyl peptidase 8|Dipeptidyl peptidase 2|Dipeptidyl peptidase 9	9 9	[14C]-Linagliptin [3H]-Linagliptin Linagliptin 1-[(4-methyl-quinazo... 8-[3(R)-aminopiperid... Tradjenta ® Tradjenta®
	1133	2	5848501	TNF FRAP	7.2900 8.1550	[14C]-Rapamycin Rapamycin Sirolimus RAPAMUNE® RAPAMUNE® temsirolimus NSC-226080

From pathways to treatments

Automated analysis to combine disease data with drug data

Step 1

Find all targets that could be used to affect the disease state



- 88 targets related to hyperinsulinism with ≥3 literature references
- Full relationship information

Name	Effect	PMID	Selected_Sentences	Types_of_Members	Number_of_References
Regulation: ABCC8 ---> Hyperinsulinism		16075046 16075046 15356046 15356046 15292329 14715863 12414839	Mutations in ID{6833=SUR1 (ABCC8)} are the most common cause of ID{9000000,9010758=hyperinsulinism} of infancy, accounting for almost 50% of cases (6, 34). For ID{6833=SUR1}, impaired binding/hydrolysis and/or transduction of Mg nucleotides will decrease ID{12115272=KATP channel} activity, producing ID{9000000,9010758=hyperinsulinism} of infancy	Protein -> Disease	47
Regulation: GCK ---> Hyperinsulinism		16075046 15919746 15356046 15356046 16574664 10488074 19717002	As described above, mutations in ID{0,2645,2746=GCK and GLUD1} cause ID{9000000,9010758=hyperinsulinism} of infancy, probably by enhancing metabolic ID{1190010=ATP} generation and decreasing ID{12115272=KATP channel} activity. CONTEXT{10003409} In addition, ID{2645=glucokinase} mutations that increase enzyme activity cause ID{9000000,9010389,9010758=hyperinsulinemic	Protein -> Disease	46
Regulation: INSR ---> Hyperinsulinism		19770178 19752219 17018838 16644916 16214940 16214940 10194465	However, ID{3643=insulin receptor} inactivation in the ID{10000000,8803116=liver} also resulted in ID{9000000,9010758=hyperinsulinemia}, ID{10000000,8803116=hepatic} ID{9000000,9012899=insulin resistance}, and peripheral ID{9000000,9012899=insulin resistance} (Michael et al. 2000, Baudry et al. 2002, Mauvais-Jarvis et al. 2002, Fisher & Kahn 2003)	Protein -> Disease	45

From pathways to treatments

Automated analysis to combine disease data with drug data

Step 1

Find all targets that could be used to affect the disease state

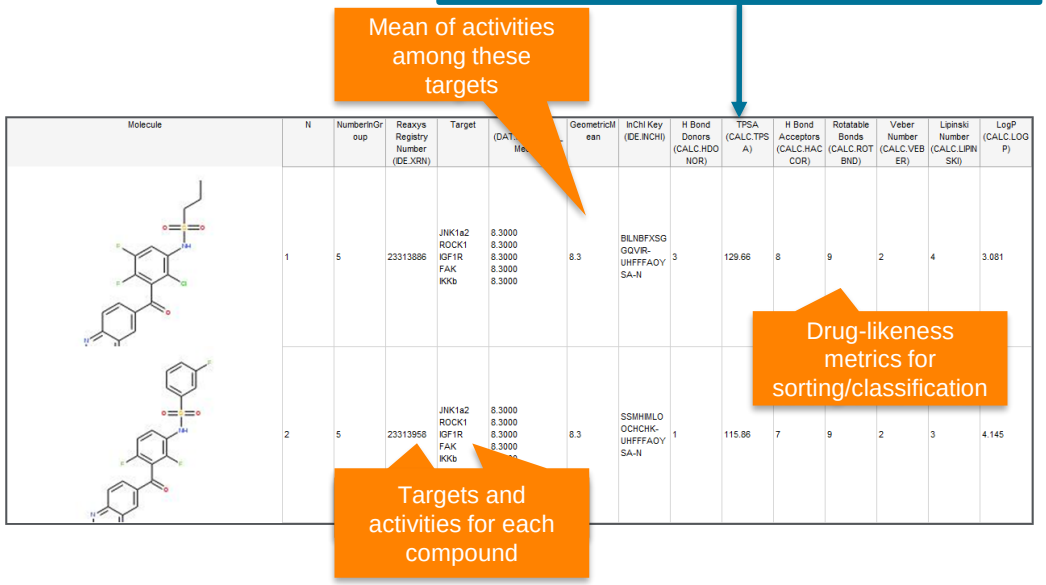
Step 2

Query for each protein to find compounds that target it (>6 log units)

Step 3

Collate data by compound to summarize the targets/activities related to disease that the compound hits
Compute geometric mean of activities for ranking
Rank by number of targets and geometric mean of activities against targets

- All compounds that were observed to bind to targets in pathway
- Sorted by number of active targets.
- Too many targets may suggest lack of specificity.



From pathways to treatments: PipelinePilot implementation combines data sources

Automated analysis combines bioassay data with text-mined data

Step 1

Find all targets that could be used to affect the disease state

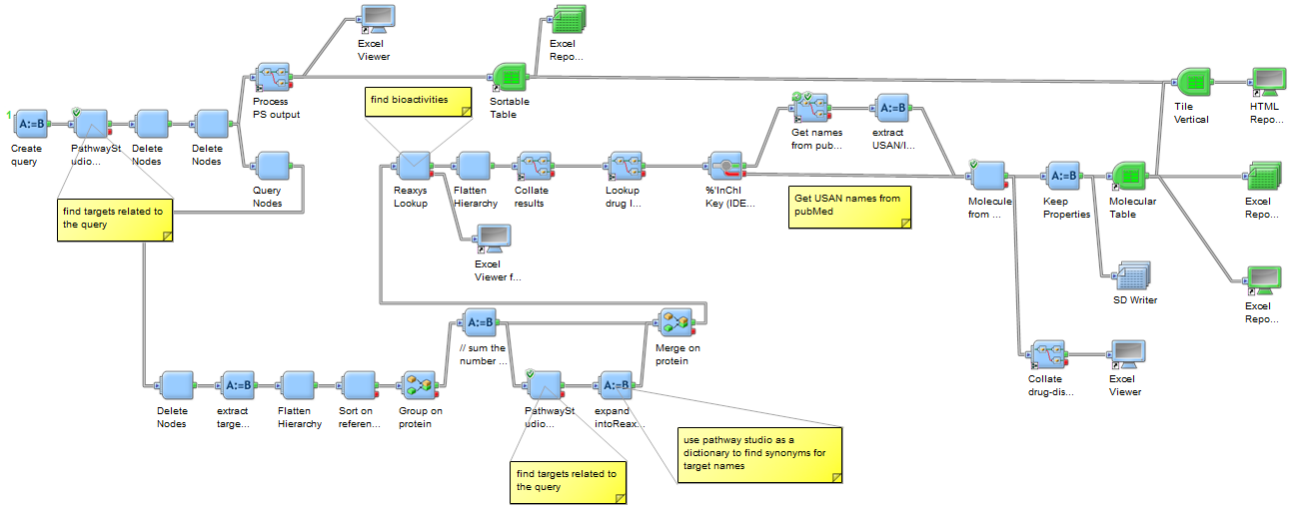
Step 2

Query for each protein to find compounds that target it (>6 log units)

Step 3

Collate data by compound to summarize the targets/activities related to disease that the compound hits

- Compute geometric mean of activities for ranking
- Rank by number of targets and geometric mean of activities against targets



From pathways to treatments: adjusting workflow

- Types of connections to a disease
- Place in a disease model
- Supporting evidence (good or bad)
- Adverse events
- Target class/localization
- Overall connectivity
- Biomarkers
- Signature-based
- Role in processes associated with a disease

- Drugs approved/passed safety
- Potential off-target activity
- Metabolism/transport
- Polypharmacology

- New drugs

Summary

- Used extensive Elsevier's content, tools and capabilities to provide information about a rare disease:
  - Literature-extracted biology data to find targets and summarize what is known about the disease mechanism
  - Bioactivity data to find drugs that target those targets
  - Normalized names of authors and institution to find collaborators/research centers

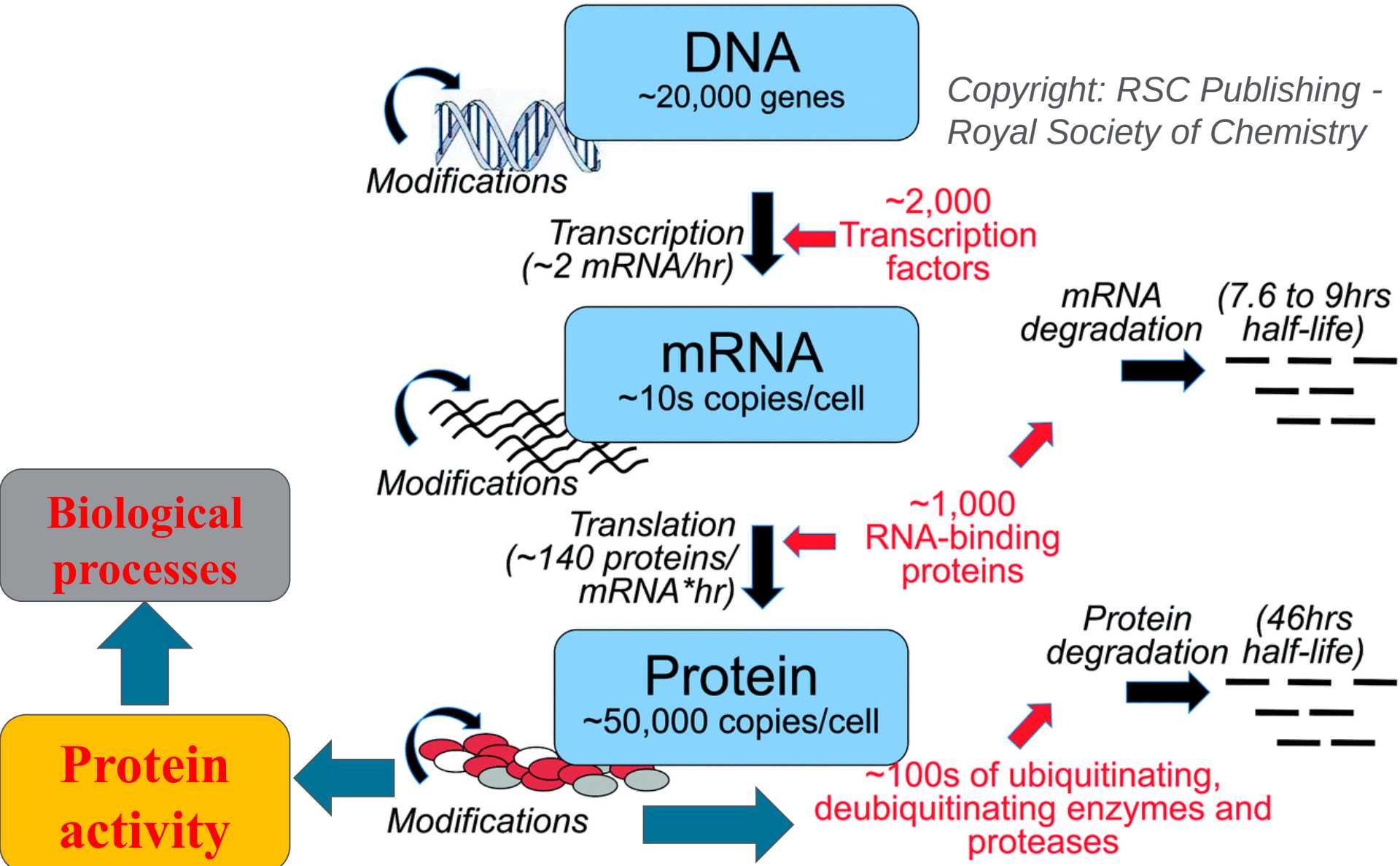
- Once the output of interest is decided, answer generation can be automated:

  Provide a disease name and get:
  - List of targets with supporting information
  - Sorted list of approved drugs with supporting information
  - KOLs and institutes

ELSEVIER

PATHWAY ANALYSIS FOR PERSONALIZED ONCOLOGY

Central dogma 2017: How to measure protein activity?



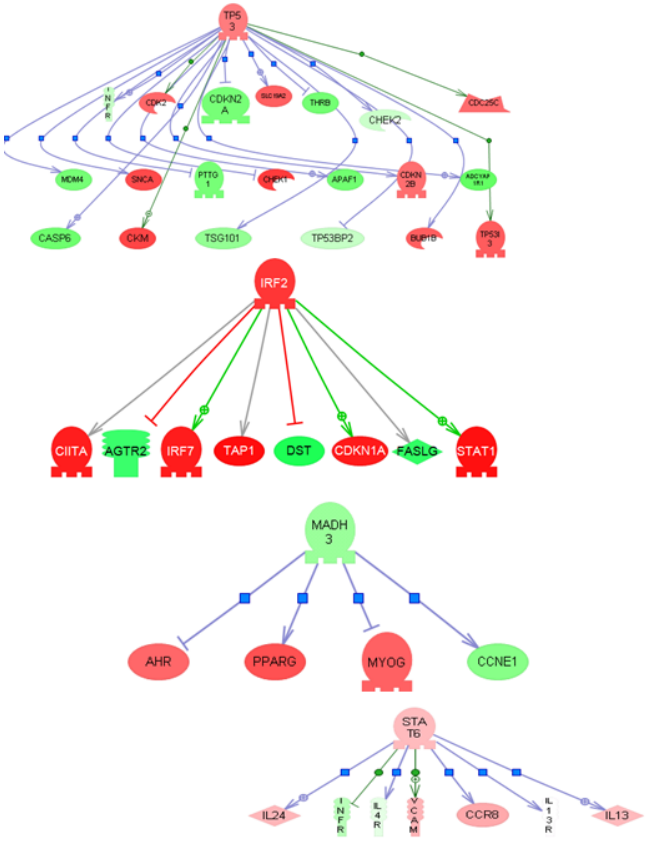
ELSEVIER

SNEA: sub-network enrichment analysis

- Calculates protein activity from the observed changes of its downstream targets

SNEA	Reverse Causal Reasoning
Mann-Whitney enrichment test	Fisher's overlap test

Lower p-value (more significant)

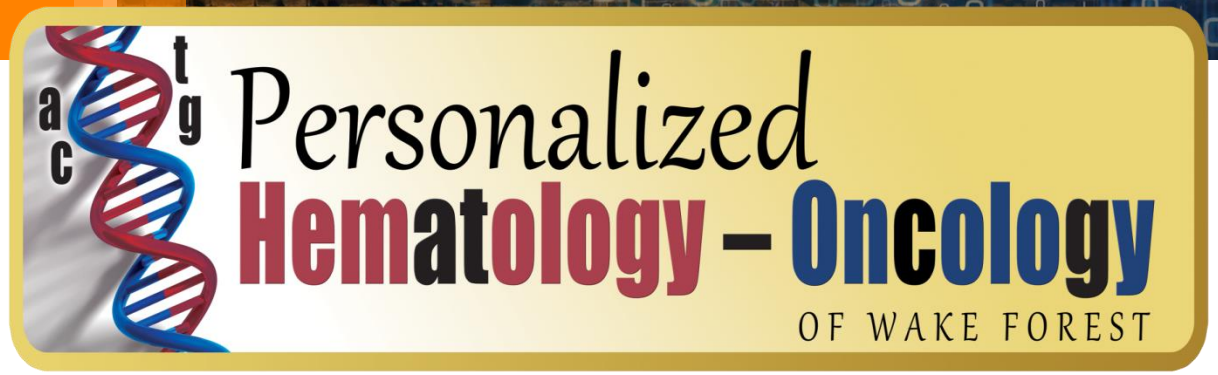


- **SNEA builds networks from all genes/proteins measured in the experiment using all relations in the database.**
- **SNEA can include indirect regulation i.e. expression regulatory cascades consisting of 2-3 steps**
- **Significant network centers may be found that are not measured in the primary dataset**
- **No prior curation of gene sets is required.**
- **Can work with partial information about TF targets. Does not require knowledge about all targets for TF**
- **P-value is sensitive to the size of the chip**

Higher p-value (less significant)

Molecular networks in microarray analysis.
Sivachenko A, Yuryev A, Daraselia N, Mazo I. J Bioinform Comp. Biol.

ELSEVIER



www.wakeforest-personalized-hemonc.com

11635 Northpark Drive, Suite 250, Wake Forest, NC 27587

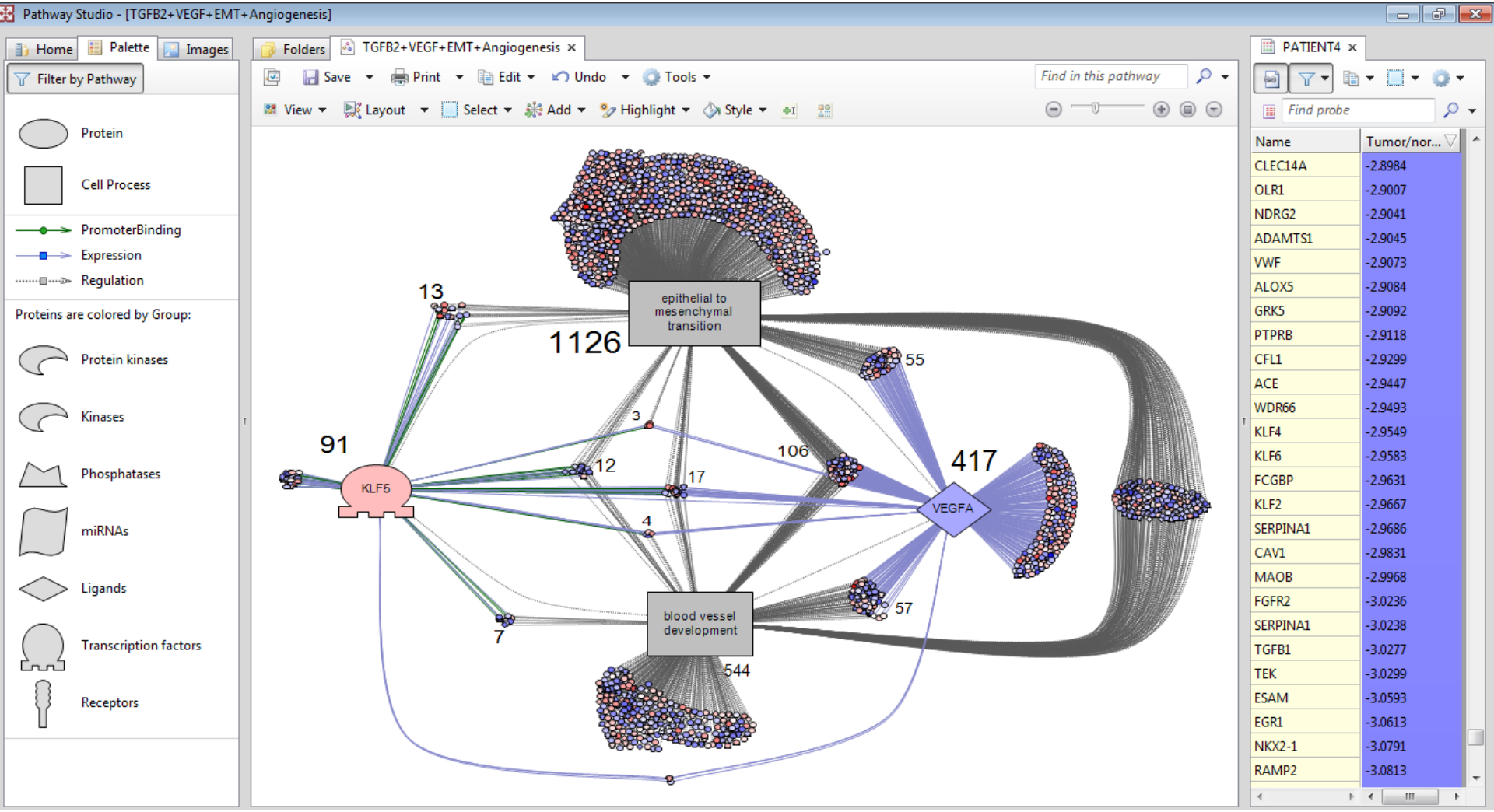
Gene expression profiling for targeted cancer treatment

Luminita Castillos[1], PhD, MBA, Francisco Castillos[1], III, MD and Anton Yuryev[2], PhD

[1]Personalized Hematology-Oncology of Wake Forest, PLLC, NC 27587, USA

[2]Elsevier, MD 20852, USA

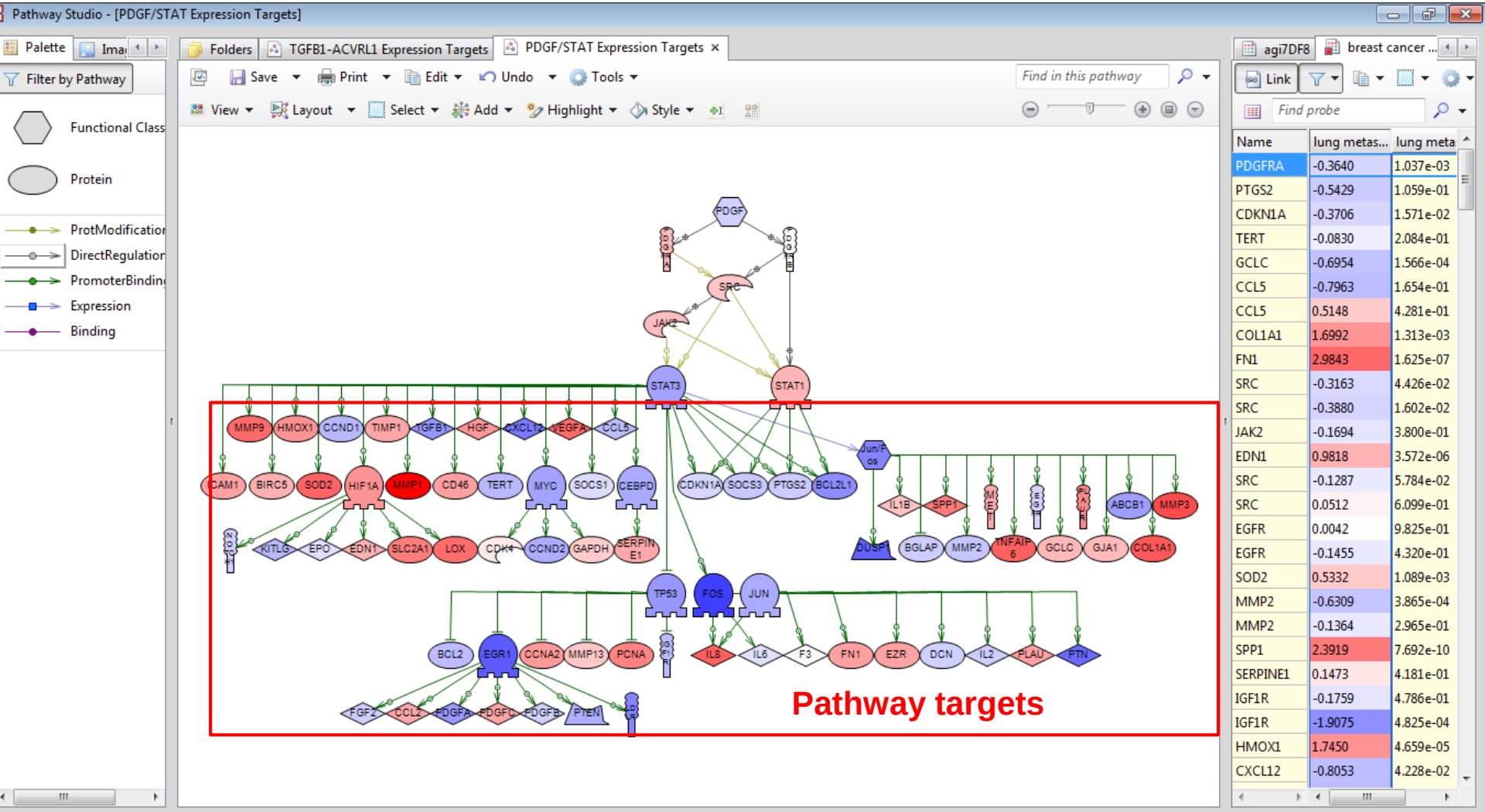
Example of expression regulators and Cell processes identified by SNEA in lung cancer patient



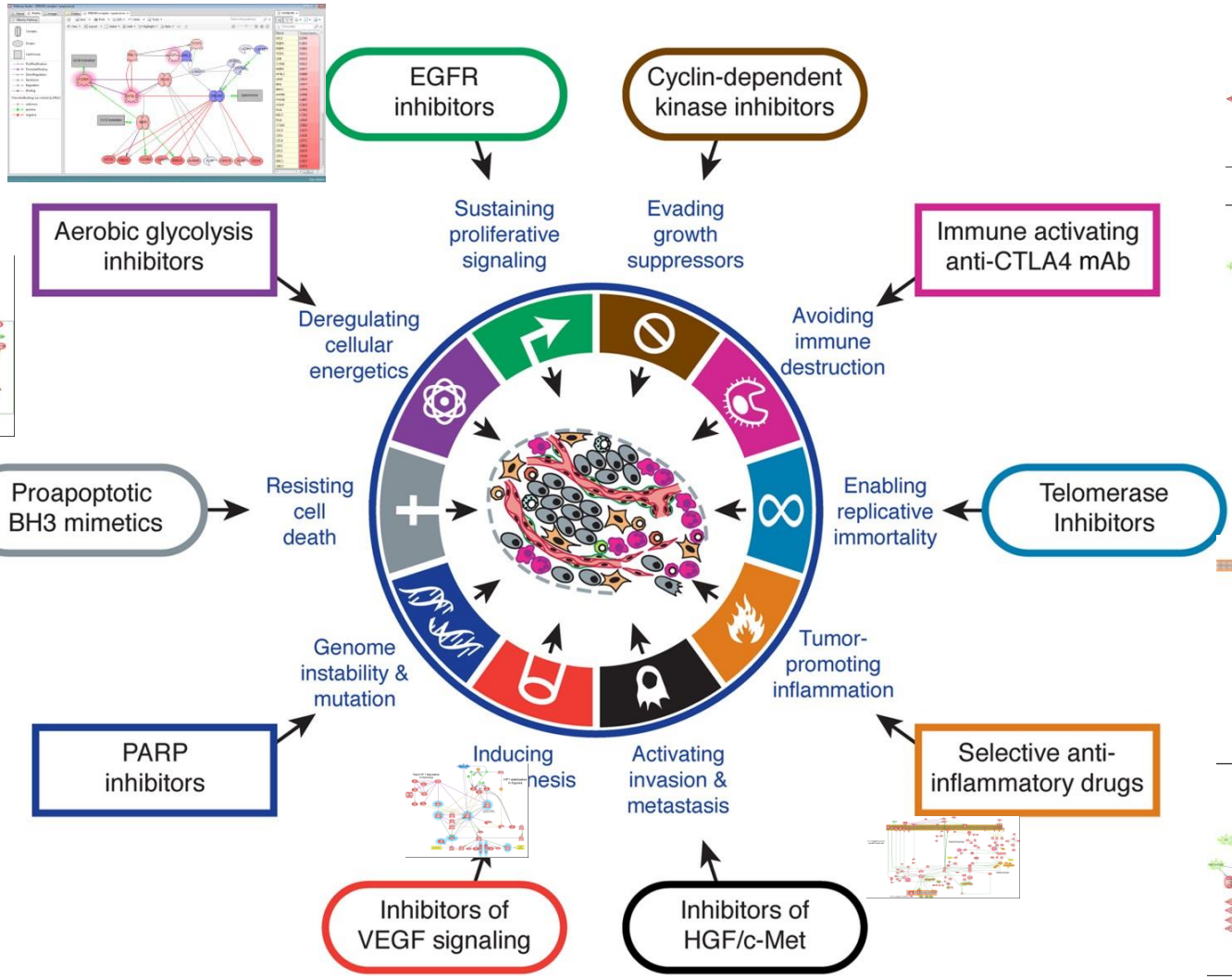
Common misconception:

Pathway activity ≠ **Differential Expression of its components**

Pathway activity = **Differential Expression of its expression targets**



Pathway Activity signatures identify targets for anti-cancer drugs



Hanahan & Weinberg. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74

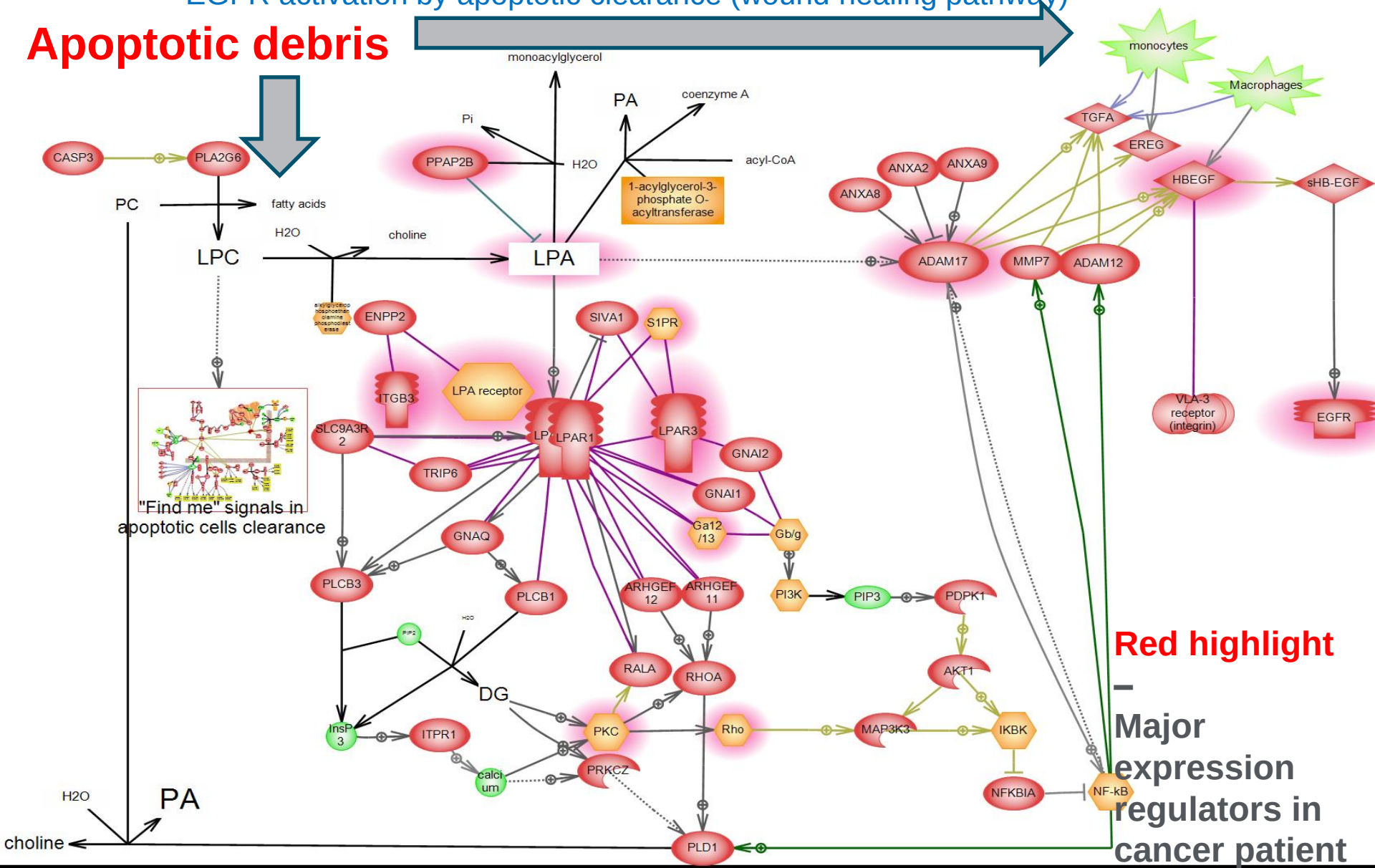
Major steps to calculate pathway activity signature

1. **Calculates major expression regulators from the expression of their targets**
2. **Maps major expression regulators on cancer pathway collection**
3. **Calculate pathway activity signature**
    1. Pathway activity signatures are short and therefore can classify patients better
    2. Pathway activity allow selection of drugs inhibiting the active pathway(s) instead of inhibiting single target

Cancer pathways: Insights to cancer biology

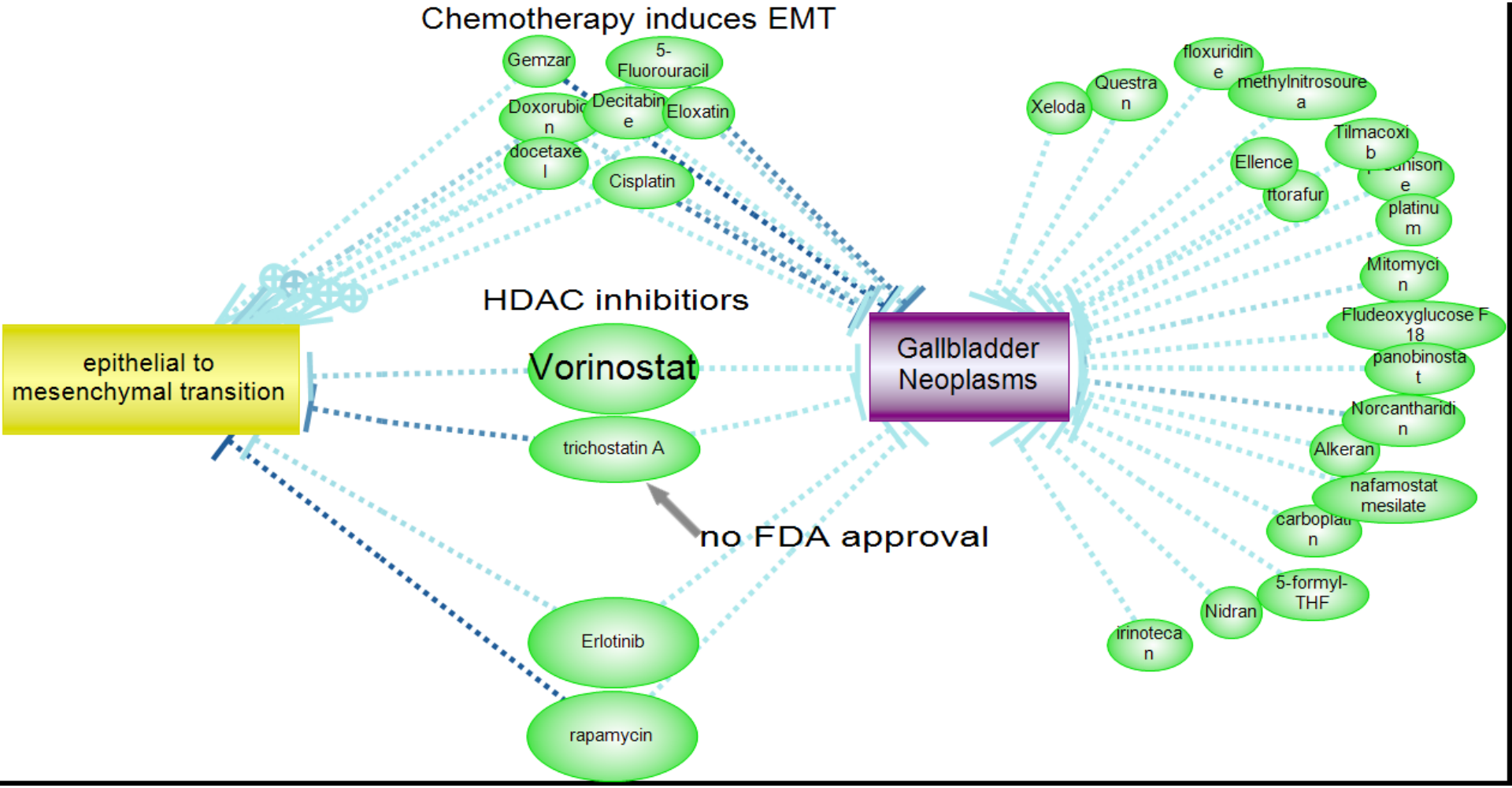
EGFR activation by apoptotic clearance (wound healing pathway)

Apoptotic debris



Red highlight – Major expression regulators in cancer patient

How to select anti-cancer drugs in Pathway Studio

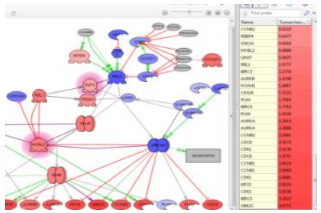


Precision Oncology

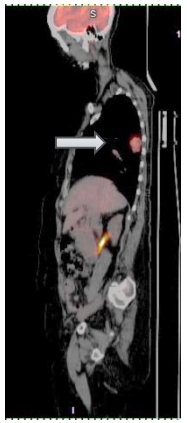
Personalized Hematology – Oncology OF WAKE FOREST

ELSEVIER

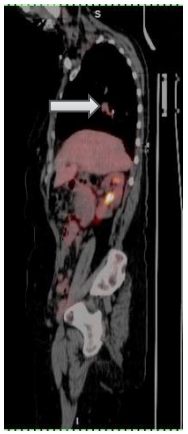
Biopsy | Transcript omics | Data analysis | Targets | Drugs



Lung met before treatment



No lung met after treatment

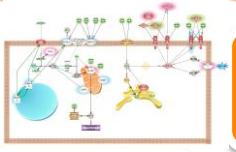


Rare diseases

ELSEVIER

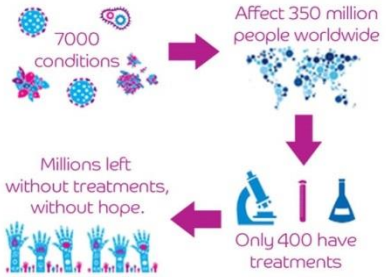
findacure

Congenital hyperinsulinism



Disease mechanism | Targets | Drugs | Publications | Experts Network

findacure



Patient Empowerment

R&D Solutions

Thank you