CompBio Version 2.5
User Manual

Updated July 24 th, 2023

Table of Contents

Section 1 - Account Management

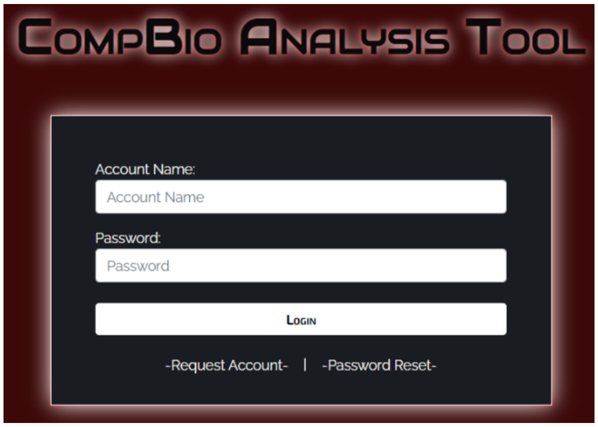
1.1 Requesting a CompBio Account

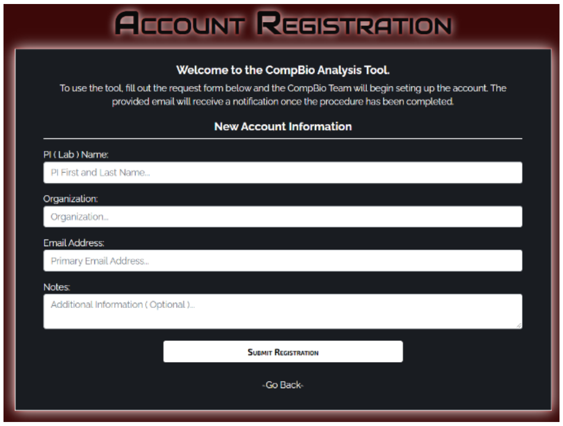
To begin using the CompBio platform, it is necessary to request an account. All Washington University investigators, researchers, and students may use the platform at no cost. For research laboratories, we do request that a single account be requested for the lab under the name of the PI. There is no limit to the number of users within the laboratory that may access the account and, as will be described later in the manual, each user can create their own project folder(s) for organizational purposes.

As the application is web-based, it can be run from any number of browsers and device types. However, as most testing is done with Google Chrome on PCs and FireFox on Macs we recommend one of these two browsers.

To access the CompBio interface and request and account use the URL - http://at-1.wucon.wustl.edu/

You will then be directed to the login page with the options below:





Account Registration

Welcome to the CompBio Analysis Tool.

To use the tool, fill out the request form below and the CompBio Team will begin seting up the account. The provided email will receive a notification once the procedure has been completed.

New Account Information

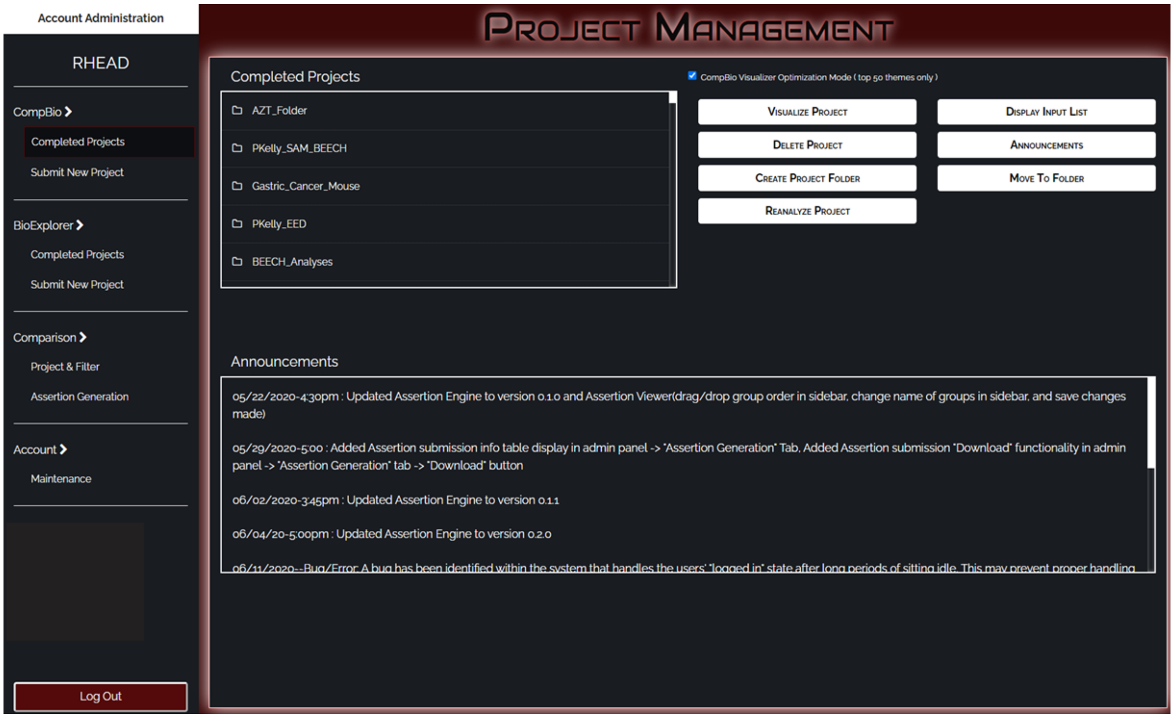
Click the **Request Account** link and you will be taken to the **Account Registration** page where you will be asked to fill out the pertinent information.

The **Primary Email Address** for the account will likely be the laboratories PI email address as it will be used for general account communications and maintenance in the future. Once the account request is received by the administration team, an account name with a temporary password will be emailed to that address.

1.2 Logging in to CompBio

Once a user account has been created for your lab, you will receive an account name and temporary password. To login into the system go to the login page using the URL - http://at-1.wucon.wustl.edu/ **Please note that the account id and password are both case sensitive.**

Once you have logged in, you will be taken to the **Project Management** page. This page contains two main panels. The first panel titled **Account Administration** to the left of the browser page contains links to all CompBio platform tools as well as your account information. The second panel contains two to three sub-sections that are specific to the tool you currently have chosen from the first panel. The upper-left sub-panel generally contains projects you have completed with that tool and the buttons to the upper-right allow enable the creation of new projects or actions on existing projects. In the case of CompBio, there will be a third lower sub-panel that contains a list of user announcements. These will typically be notifications of system downtime, updates, and so forth.



05/22/2020-4:30pm : Updated Assertion Engine to version 0.1.0 and Assertion Viewer(drag/drop group order in sidebar, change name of groups in sidebar, and save changes made)

05/29/2020-5:00 : Added Assertion submission info table display in admin panel -> "Assertion Generation" Tab, Added Assertion submission "Download" functionality in admin panel -> "Assertion Generation" tab -> "Download" button

06/02/2020-3:45pm : Updated Assertion Engine to version 0.1.1

06/04/20-5:00pm : Updated Assertion Engine to version 0.2.0

06/11/2020--Bug/Error: A bug has been identified within the system that handles the users' "logged in" state after long periods of sitting idle. This may prevent proper handling

1.3 Account Maintenance

Clicking the **Account Maintenance** link will take the user to a page that will allow for the **Primary Account Email** address and **Password** to be changed. This will change either for ALL account users so this should only be done when necessary for the account.

Section 2 - Project Management

The Project Management component of the CompBio platform is a relatively simple windows-like user interface. While the current functionality is fairly rudimentary, we continue to make improvements and welcome feedback to further enhance the user experience.

2.1 Submitting a Project

CompBio is a multi-omics analysis tool and can analyze one or multiple types of entities in a single input list. While CompBio is able to understand several different omics types, there are requirements for the input format on each and, in some cases, limitations on the synonyms that are understood by the platform. This is particularly true for metabolites given the wide range of names they can be identified by in the literature.

2.1.1 Entity definitions and limitations

There are five types of entities currently accepted as input for CompBio analysis- Genes, proteins, miRNAs, metabolites, and microbes. All entity lists must be newline delimited. Genes and proteins need to be input as official gene symbols ("ABCA1, IL34"). Microbes need to be input as "genus-species". Without a species name, the microbes will not be included in the analysis. Micro-RNAS need to be in the following format- "mir-1253"as seen in most publications. Metabolites need to be included using their complete names "Carbofuran, Dextromethorphan" etc. A unique feature of CompBio is that the input list can be a mixed list consisting of one or more types of entities.

Important Note: CompBio can currently analyze a list ranging from 5 to 2500 entities in a project. While it is possible to input and run a list larger than 2500, the randomization statistics are currently only computed to this number so larger projects will not have p-values or normalized enrichment scores on the themes.

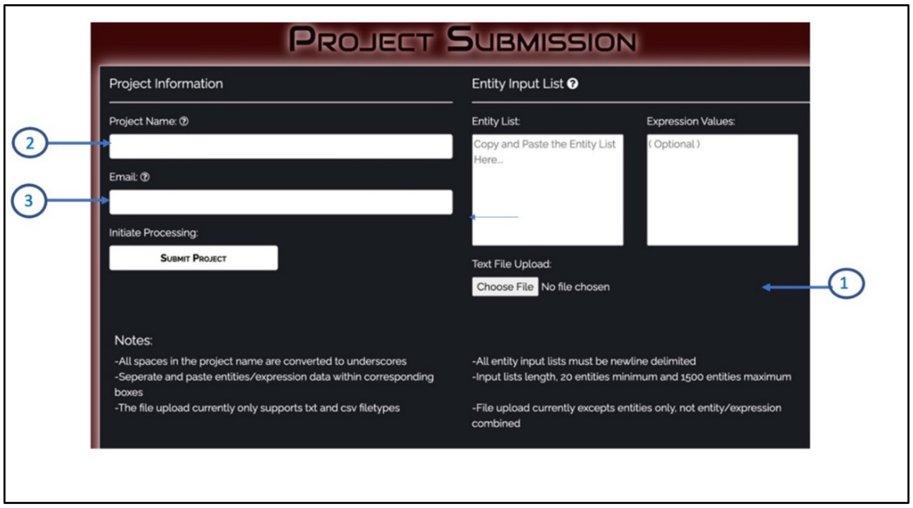
In the current version of CompBio, only mammalian gene symbols are considered; specifically human, mouse and rat. Other species are not included yet. However, if the gene symbol used for non-mammalian species in the same as that for mammals, those can get some hits. The metabolite table is extensive but is still in need of some more work because of the nature of those entities. Other entity types, like variants, will be added soon.

2.1.2 Expression values

CompBio offers the user an option to include expression values or any other numeric metric corresponding to the entity list. These can be fold changes, p-values or correlation coefficients. These are optional and not used in the creation of the map but are simply carried over into the "theme view". CompBio will generate an identical map even if the expression values are left blank. If the user is input a mixed list, this is a good way to distinguish between gene and protein entries with the same symbol. IL1B as a gene versus IL1B protein. The expression value for IL1B protein can have a "(p)" tag to help distinguish the two.

2.1.3 Text file upload

Instead of copy pasting the entity list into the predefined boxes, there is a "file upload" option. A text file (.txt or .csv) containing the entity list and expression values (optional) can be uploaded for a CompBio analysis using the **Choose File** button (1)



Project Information

Project Name:

Email:

Initiate Processing:

Entity Input List

Entity List:

Copy and Paste the Entity List Here...

Expression Values:

(Optional)

Text File Upload:

No file chosen

Notes:

-All spaces in the project name are converted to underscores
-Seperate and paste entities/expression data within corresponding boxes
-The file upload currently only supports txt and csv filetypes

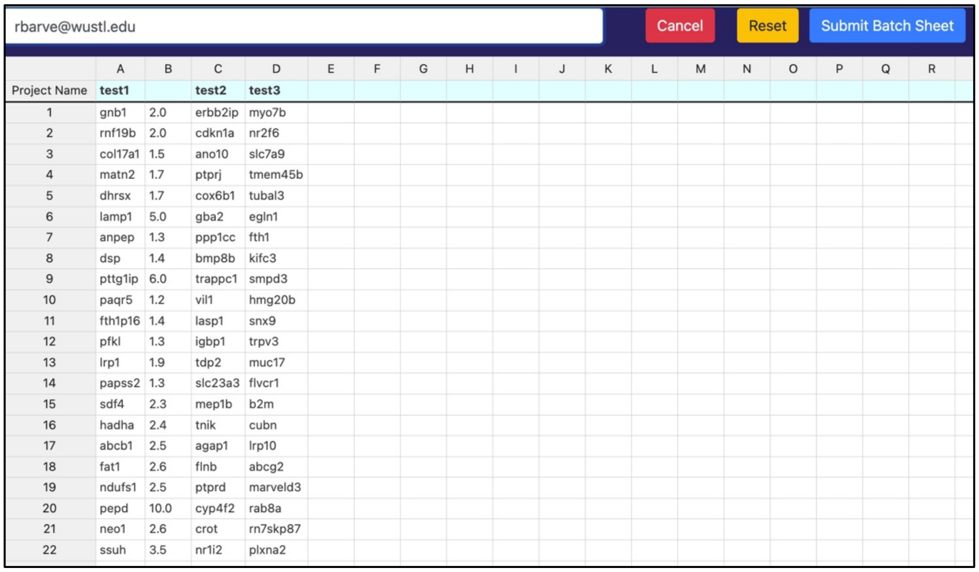
-All entity input lists must be newline delimited
-Input lists length, 20 entities minimum and 1500 entities maximum

-File upload currently excepts entities only, not entity/expression combined

Once the entity list is input, user needs to give a meaningful project name (2) and an email address (3) to submit the analysis by clicking **Submit Project**. An email will be sent notifying the user once the CompBio run is complete.

2.1.4 Batch submission

Multiple CompBio runs can be submitted at a time using the batch submission option.



rbarve@wustl.edu

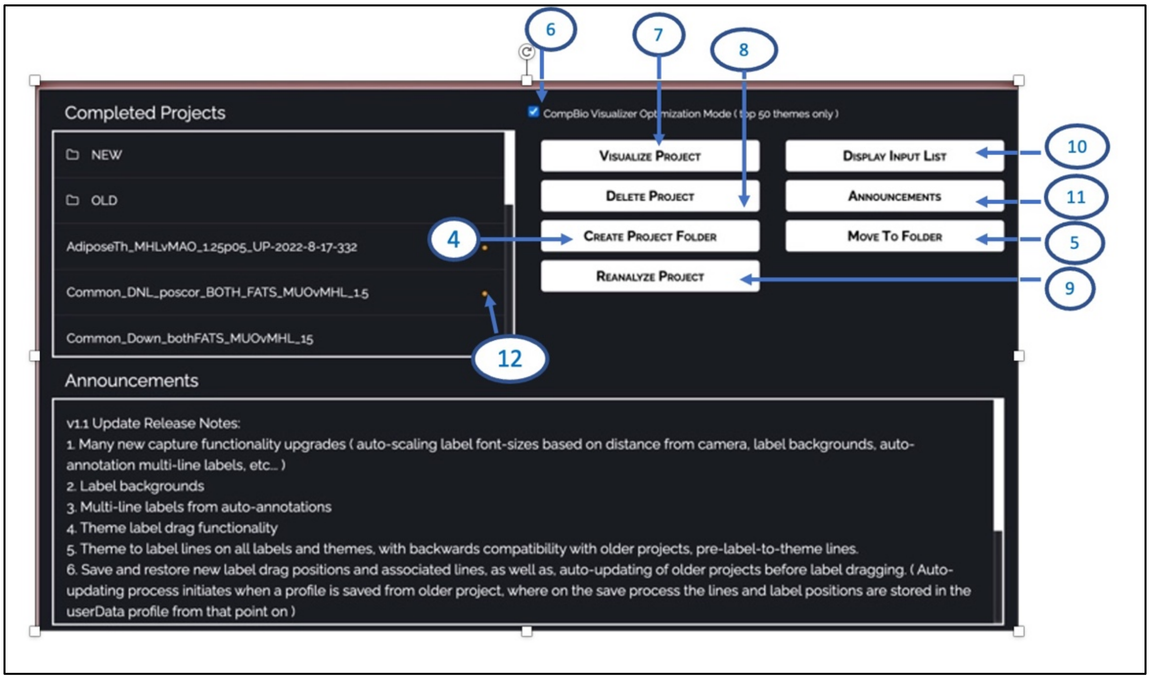
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Project Name	**test1**		**test2**	**test3**														
1	gnb1	2.0	erbb2ip	myo7b														
2	rnf19b	2.0	cdkn1a	nr2f6														
3	col17a1	1.5	ano10	slc7a9														
4	matn2	1.7	ptprj	tmem45b														
5	dhrsx	1.7	cox6b1	tubal3														
6	lamp1	5.0	gba2	egln1														
7	anpep	1.3	ppp1cc	fth1														
8	dsp	1.4	bmp8b	kifc3														
9	pttg1ip	6.0	trappc1	smpd3														
10	paqr5	1.2	vil1	hmg20b														
11	fth1p16	1.4	lasp1	snx9														
12	pfkl	1.3	igbp1	trpv3														
13	lrp1	1.9	tdp2	muc17														
14	papss2	1.3	slc23a3	flvcr1														
15	sdf4	2.3	mep1b	b2m														
16	hadha	2.4	tnik	cubn														
17	abcb1	2.5	agap1	lrp10														
18	fat1	2.6	flnb	abcg2														
19	ndufs1	2.5	ptprd	marveld3														
20	pepd	10.0	cyp4f2	rab8a														
21	neo1	2.6	crot	rn7skp87														
22	ssuh	3.5	nr1i2	plxna2														

Each project needs a name in the Project Name row and entities pasted below. Project test1 is submitted with entities and their expression values. Project test2 and test3 show examples where only entities are input. The progress of the jobs can be tracked using the tracker button on the main page.

2.2 Managing Completed Projects

2.2.1 Project directories

Once the project is completed and email notification will be sent to the user. A completed project can be moved into a directory/folder within the tool. To create a folder, use the **Create Project Folder** button (4). The new project folder will be visible on the left-hand side panel. To move the project simply select the project and use the **Move to Folder** button (5) to place it into the folder of choice.


6
7
8
Completed Projects
☑ CompBio Visualizer Optimization Mode (top 50 themes only)
NEW
OLD
AdiposeTh_MHLvMAO_1.25p05_UP-2022-8-17-332
Common_DNL_poscor_BOTH_FATS_MUOvMHL_1.5
Common_Down_bothFATS_MUOvMHL_15
4
12
Visualize Project
Delete Project
Create Project Folder
Reanalyze Project
Display Input List
Announcements
Move To Folder
10
11
5
9
Announcements

v1.1 Update Release Notes:
1. Many new capture functionality upgrades (auto-scaling label font-sizes based on distance from camera, label backgrounds, auto-annotation multi-line labels, etc...)
2. Label backgrounds
3. Multi-line labels from auto-annotations
4. Theme label drag functionality
5. Theme to label lines on all labels and themes, with backwards compatibility with older projects, pre-label-to-theme lines.
6. Save and restore new label drag positions and associated lines, as well as, auto-updating of older projects before label dragging. (Auto-updating process initiates when a profile is saved from older project, where on the save process the lines and label positions are stored in the userData profile from that point on)

2.2.2 *Annotation markers*

Each project map can be annotated by the user. The completed projects which have user defined annotation along with saved profiles will be marked with an orange dot (12) after the project name as shown above.

2.2.3 *Visualizer optimization*

On the project management page, there is a checkbox on the top right side (6). By default, that box is checked. This will limit the visualizer/display of the map to the top 50 themes. Most of the biology generated by the input list is covered within the top 35 or so themes. However, there may be certain cases where a user may want to look for more than 50 themes. This can be achieved by unchecking the box and all themes that can be generated from the input list will be displayed. There are a set of 6 buttons on the top right-hand side of the main panel, below the visualizer optimization check box. These buttons will help the user with basic functionality for the project.

2.2.4 *Visualize project*

Once the CompBio run is completed, the project map can be visualized using the **Visualize Project** button (7). A basic profile generated from the input list will appear in a separate window for the user to annotate and analyze.

2.2.5 *Delete project*

Using the **Delete Project** button (8) will permanently delete the project. Unless the project was present within the user's account long enough for it to appear in a backup, the project will not be retrievable. Thus, be certain this is acceptable before deleting a project.

2.2.6 Reanalyze project

The **Reanalyze Project** button (9) will fire a brand new CompBio run using the same input list with the current PubMed and latest version of the PCMM model. The project name will have the current date tagged on the end to distinguish it from the earlier/older run.

2.2.7 Display input list

The **Display Input List** button (10) will display the input list used to generate the CompBio project.

2.2.8 Announcements

The **Announcement Button** will display the announcements and updates from the CompBio team.

2.2.9 Move to folder

The completed CompBio project will be displayed in the main directory. To move to it to the desired folder, the **Move to Folder** button (5) can be used.

2.2.10 Project information

Once a project is selected, the bottom half of the main panel will display information about the project including the project name, submission date, email used for submission, number of entities.

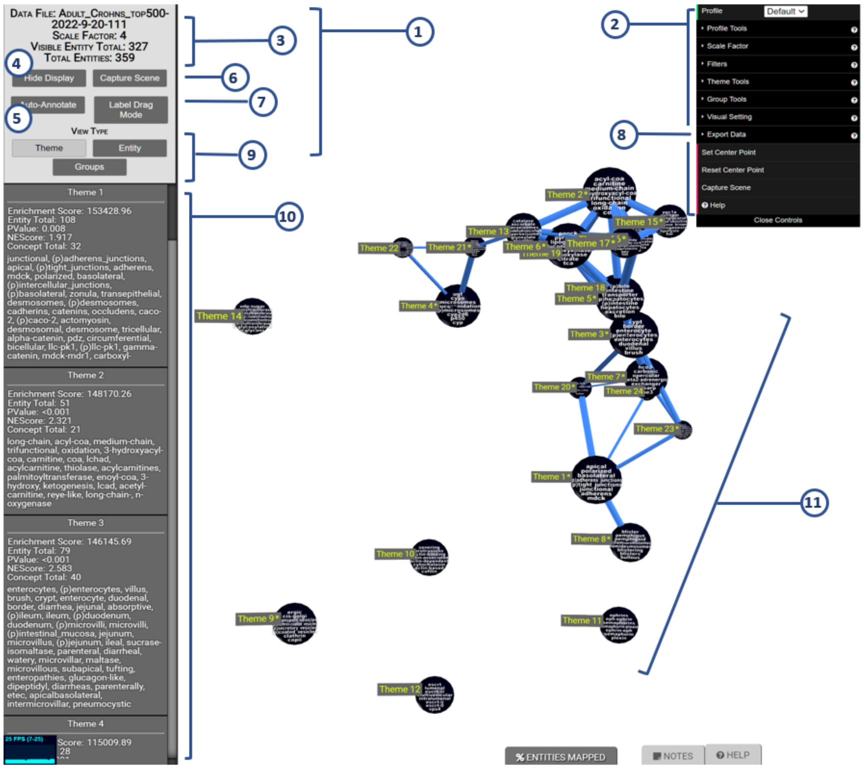
Section 3 - CompBio Analysis Projects

Section 3 of this manual will describe biological exploration and knowledge generation from completed CompBio knowledge maps, and the tools required to perform those analyses.

CompBio assembles contextually relevant information derived from an input list of entities (genes, proteins, miRNAs, metabolites, and/or bacterial species). This is accomplished by first identifying concepts enriched in literature that contains the provided entities. The subset of literature that contained the provided entities determines the context for the project. A conditional probability matrix is calculated for each concept versus every other identified concept describing the likelihood of encountering two concepts nearby each other *in the publications returned by the supplied entity list*. Since the probabilities are calculated using only records (publications) that contained the user supplied entities, the biological context is an important factor in the calculation of the distances between concepts. The CompBio algorithm then clusters the closely related concepts into themes.

Below is an overview of the CompBio Workspace. These elements will be discussed in the following sections.

Overview of CompBio Workspace Elements



Data File: Adult_Crohns_top500-2022-9-20-111
Scale Factor: 4
Visible Entity Total: 327
Total Entities: 359

Theme 1

Enrichment Score: 153428.96
Entity Total: 108
PValue: 0.008
NEScore: 1.917
Concept Total: 32

junctional, (p)adherens_junctions, apical, (p)tight_junctions, adherens, mdck, polarized, basolateral, (p)intercellular_junctions, (p)basolateral, zonula, transepithelial, desmosomes, (p)desmosomes, cadherins, catenins, occludens, caco-2, (p)caco-2, actomyosin, desmosomal, desmosome, tricellular, alpha-catenin, pdz, circumferential, bicellular, llc-pk1, (p)llc-pk1, gamma-catenin, mdck-mdr1, carboxyl-

Theme 2

Enrichment Score: 148170.26
Entity Total: 51
PValue: <0.001
NEScore: 2.321
Concept Total: 21

long-chain, acyl-coa, medium-chain, trifunctional, oxidation, 3-hydroxyacyl-coa, carnitine, coa, lchad, acylcarnitine, thiolase, acylcarnitines, palmitoyltransferase, enoyl-coa, 3-hydroxy, ketogenesis, lcad, acetyl-carnitine, reye-like, long-chain-, n-oxygenase

Theme 3

Enrichment Score: 146145.69
Entity Total: 79
PValue: <0.001
NEScore: 2.583
Concept Total: 40

enterocytes, (p)enterocytes, villus, brush, crypt, enterocyte, duodenal, border, diarrhea, jejunal, absorptive, (p)ileum, ileum, (p)duodenum, duodenum, (p)microvilli, microvilli, (p)intestinal_mucosa, jejunum, microvillus, (p)jejunum, ileal, sucrase-isomaltase, parenteral, diarrheal, watery, microvillar, maltase, microvillous, subapical, tufting, enteropathies, glucagon-like, dipeptidyl, diarrheas, parenterally, etec, apicalbasolateral, intermicrovillar, pneumocystic

Theme 4

Score: 115009.89
28

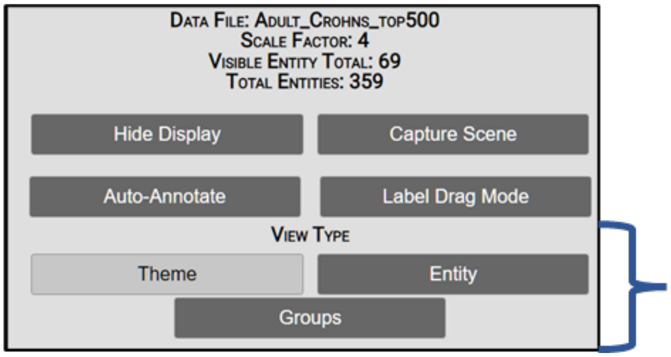
1. Project panel
2. Control panel
3. Project metrics
   a. Data file
   b. Scale factor
   c. Visible entities
   d. Total entities
4. Hide project panel
5. Auto-annotate project
6. Capture scene
7. Label drag mode on/off
8. Export data
9. Project view selection
10. Theme/Entity detail panel
11. Knowledge map

3.1 Biological Exploration and Knowledge Generation

CompBio has been designed to automate what has traditionally been a lengthy analysis process requiring biologists skilled in pathway analysis to wade through lists of differentially regulated entities, assembling groups of genes of similar function to discern pathways and processes of interest. While current tools can enable this effort to some degree, they tend to suffer from data inflation where a list of hundreds of genes can produce even greater numbers of statistically enriched pathways. CompBio, on the other hand, assembles enriched concepts into interconnected themes, allowing for rapid interpretation of the knowledge identified and contextually assembled from the input entity list.

3.1.1 View types – Theme, Entity, and Group

The CompBio project viewer has three different view types, **Theme**, **Entity**, and **Group**. Each view type will center the focus of the selection functions in the viewport on that particular type. The purpose is to allow the user to explore and visualize the data at different levels. Each of the types and their respective functions will be described in detail in the following sections. The view types are shown in the upper left section of the **Project Detail** panel and can be selected by clicking on the appropriate box. **Theme** view is the default view the user will be presented with upon loading a project into the visualizer and will be discussed first. **Entity** view and **Group** view will be discussed at the end of this section (3.1.3 and 3.1.4).



Data File: Adult_Crohns_top500
Scale Factor: 4
Visible Entity Total: 69
Total Entities: 359

3.1.2 Theme view

As described above, CompBio assembles enriched biological concepts into themes based on their associations through a method of natural language processing. As these themes are assembled without the use of predefined ontologies, the assembled themes generally represent one or more levels of biological process, function, or pathway. In some cases, a theme can contain information from a molecular function to the cell and/or tissue type that function is contextually associated with.

When initially visualizing a project, the Theme view will be displayed by default and the user will be presented with 3-dimensional ball-and-stick model in the central panel that contains the top themes based on Enrichment Score. Themes are placed close to other themes that they share contextual relationships with. While the absolute location of a theme is of no consequence, their relative location to each other is very informative. Proximal themes or theme groups will generally share many entities and describe closely related biology.

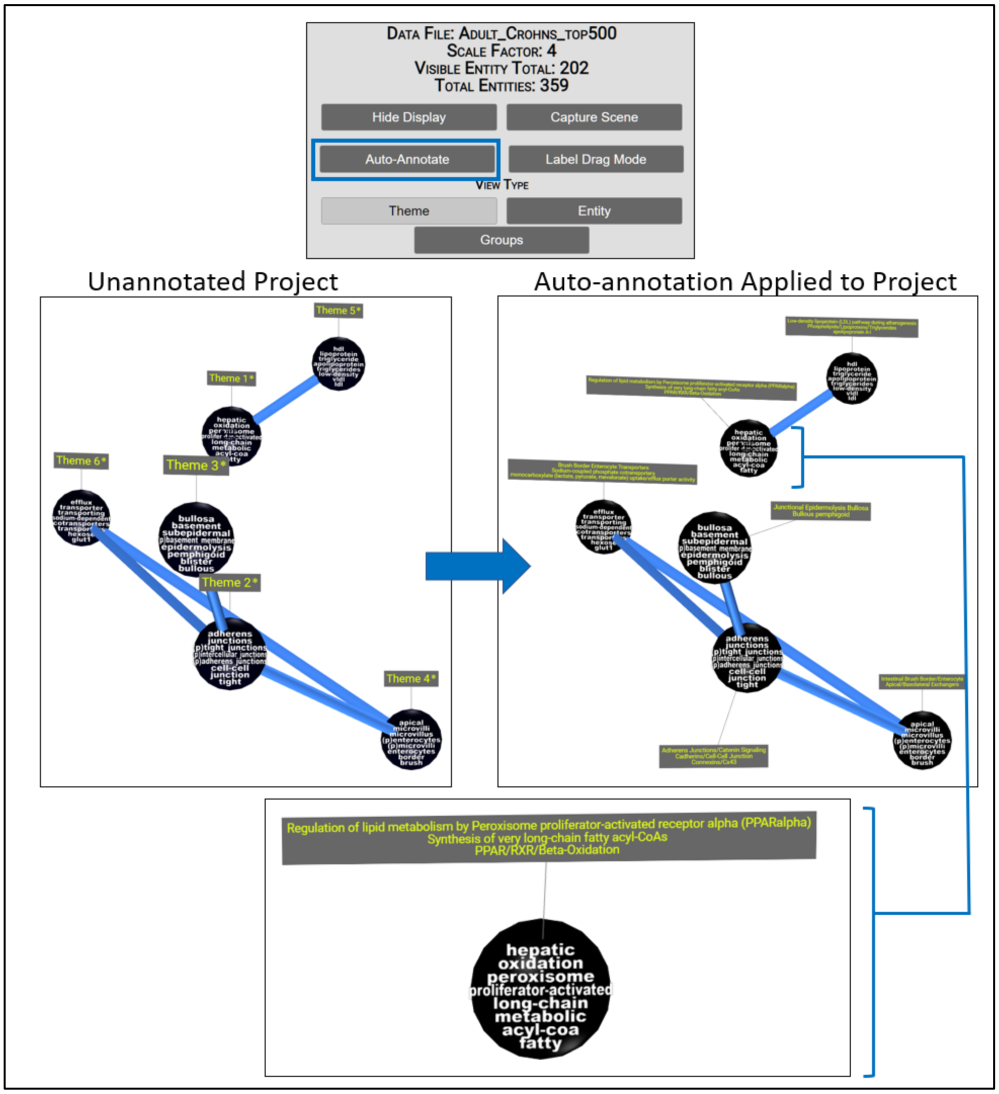
3.1.2.1 Theme annotation

Once the project is loaded into the visualizer the raw themes need to have annotation applied. Annotation is a summary description of the concepts contained in the theme. There are two ways to accomplish this. Annotation can be applied through manual annotation, where the user interprets the contents of each theme. Alternatively, the **Auto-Annotate** function can be implemented which will, through a natural language processing approach utilizing the theme contents and the project context, apply up to three titles to each theme. Potential titles are sourced from nearly 100,000 biological terms, processes, and pathways. Each title displays a the confidence score associated with the title autoannotation

Scores:

1.2 < Score < 1.5 – Modest confidence (90% accuracy with manual assessment)
1.5 < Score < 2.0 – Marked confidence (97% accuracy with manual assessment)
2.0< Score > 3.0 – High confidence (99% accuracy with manual assessment)
Score > 3.0 – Extremely high confidence (100% accuracy with manual assessment)

An overview of the workflow is shown below.

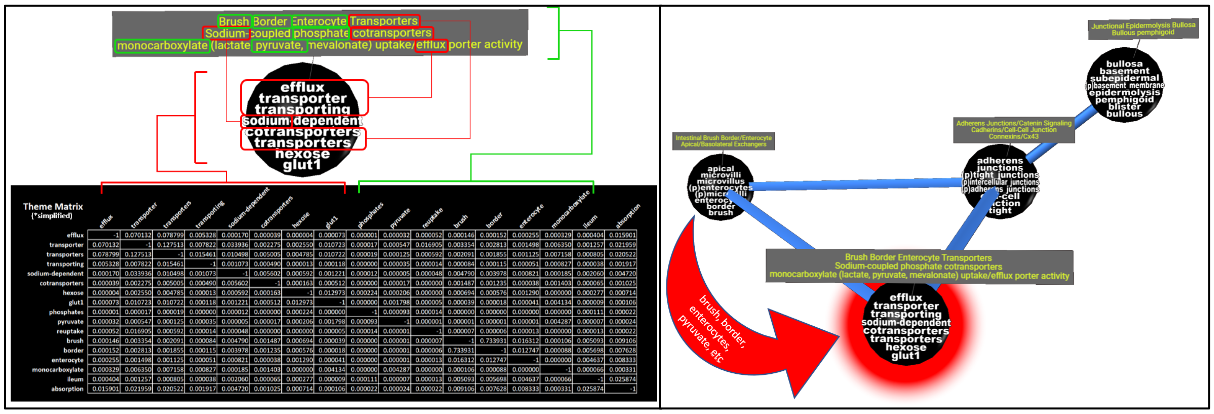


Data File: Adult_Crohns_top500
Scale Factor: 4
Visible Entity Total: 202
Total Entities: 359

Hide Display
Capture Scene
Auto-Annotate
Label Drag Mode
View Type
Theme
Entity
Groups
Unannotated Project
Auto-annotation Applied to Project
Theme 5*
Theme 1*
Theme 6*
Theme 3*
Theme 2*
Theme 4*
Low-density lipoprotein (LDL) pathway during atherogenesis
Phospholipids/Lipoproteins/Triglycerides
apolipoprotein A-I
Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)
Synthesis of very long-chain fatty acyl-CoAs
PPAR/RXR/Beta-Oxidation
Brush Border Enterocyte Transporters
Sodium-coupled phosphate cotransporters
monocarboxylate (lactate, pyruvate, mevalonate) uptake/efflux porter activity
Junctional Epidermolysis Bullosa
Bullous pemphigoid
Intestinal Brush Border/Enterocyte
Apical/Basolateral Exchangers
Adherens Junctions/Catenin Signaling
Cadherins/Cell-Cell Junction
Connexins/Cx43
hdl lipoprotein triglyceride apolipoprotein triglycerides low-density vldl ldl
hepatic oxidation peroxisome proliferator-activated long-chain metabolic acyl-coa fatty
efflux transporter transporting sodium-dependent cotransporters transporters hexose glut1
bullosa basement subepidermal (p)basement membrane epidermolysis pemphigoid blister bullous
adherens junctions (p)tight junctions (p)intercellular junctions (p)adherens junctions cell-cell junction tight
apical microvilli microvillus (p)enterocytes (p)microvilli enterocytes border brush

As can be seen above, three titles have automatically been applied to the theme. Please note that new projects have the auto-annotation precalculated when the job is first run, but older projects predating the implementation of this feature will take 1-2 minutes the first time auto-annotation is applied.

This process utilizes the conditional probability matrix described above to not only use theme members to inform annotation, but also closely related terms in other nearby themes. The images below represent graphically the auto-annotation process for a theme related to small-molecule transporters. If one were to only consider the concepts in the theme, a title of “Transporters” may be chosen as a best fit for this theme (shown below in red), but that ignores the contextual information in other nearby themes (shown in green). By allowing contextual inference additional descriptive concepts can be added to the theme title to supplement the understanding of what this theme represents in the context of this experiment. The **theme matrix** shown below represents the conditional probability matrix, showing the probability of encountering two concepts close by each other *in the publications that contained the user-supplied entities*. These associations are specific for the given experiment and are different than the co-occurrence frequency that would be observed in all scientific literature.


Brush Border Enterocyte Transporters
Sodium-coupled phosphate cotransporters
monocarboxylate (lactate, pyruvate, mevalonate) uptake/efflux porter activity
efflux
transporter
transporting
sodium-dependent
cotransporters
transporters
hexose
glut1
Junctional Epidermolysis Bullosa
Bullous pemphigoid
bullosa
basement
subepidermal
(p)basement_membrane
epidermolysis
pemphigoid
blister
bullous
Intestinal Brush Border/Enterocyte
Apical/Basolateral Exchangers
apical
microvilli
microvillus
(p)enterocytes
(p)microvilli
enterocy
border
brush
Adherens Junctions/Catenin Signaling
Cadherins/Cell-Cell Junction
Connexins/Cx43
adherens
junctions
(p)tight_junctions
(p)intercellular_junctions
(p)adherens_junctions
-cell
ction
tight
Brush Border Enterocyte Transporters
Sodium-coupled phosphate cotransporters
monocarboxylate (lactate, pyruvate, mevalonate) uptake/efflux porter activity
efflux
transporter
transporting
sodium-dependent
cotransporters
transporters
hexose
glut1
brush, border, enterocytes, pyruvate, etc

Theme Matrix (*simplified)	efflux	transporter	transporters	transporting	sodium-dependent	cotransporters	hexose	glut1	phosphates	pyruvate	reuptake	brush	border	enterocyte	monocarboxylate	ileum	absorption
efflux	-1	0.070132	0.078799	0.005328	0.000170	0.000039	0.000004	0.000073	0.000001	0.000032	0.000052	0.000146	0.000152	0.000255	0.000329	0.000404	0.015901
transporter	0.070132	-1	0.127513	0.007822	0.033936	0.002275	0.002550	0.010723	0.000017	0.000547	0.016905	0.003354	0.002813	0.001498	0.006350	0.001257	0.021959
transporters	0.078799	0.127513	-1	0.015461	0.010498	0.005005	0.004785	0.010722	0.000019	0.000125	0.000592	0.002091	0.001855	0.001125	0.007158	0.000805	0.020522
transporting	0.005328	0.007822	0.015461	-1	0.001073	0.000490	0.000013	0.000118	0.000000	0.000035	0.000014	0.000084	0.000115	0.000051	0.000827	0.000038	0.001917
sodium-dependent	0.000170	0.033936	0.010498	0.001073	-1	0.005602	0.000592	0.001221	0.000012	0.000005	0.000048	0.004790	0.003978	0.000821	0.000185	0.002060	0.004720
cotransporters	0.000039	0.002275	0.005005	0.000490	0.005602	-1	0.000163	0.000512	0.000000	0.000017	0.000000	0.001487	0.001235	0.000038	0.001403	0.000065	0.001025
hexose	0.000004	0.002550	0.004785	0.000013	0.000592	0.000163	-1	0.012973	0.000224	0.000206	0.000000	0.000694	0.000576	0.001290	0.000000	0.000277	0.000714
glut1	0.000073	0.010723	0.010722	0.000118	0.001221	0.000512	0.012973	-1	0.000000	0.001798	0.000005	0.000039	0.000018	0.000041	0.004134	0.000009	0.000106
phosphates	0.000001	0.000017	0.000019	0.000000	0.000012	0.000000	0.000224	0.000000	-1	0.000093	0.000014	0.000000	0.000000	0.000000	0.000000	0.000111	0.000022
pyruvate	0.000032	0.000547	0.000125	0.000035	0.000005	0.000017	0.000206	0.001798	0.000093	-1	0.000001	0.000001	0.000001	0.000001	0.004287	0.000007	0.000024
reuptake	0.000052	0.016905	0.000592	0.000014	0.000048	0.000000	0.000000	0.000005	0.000014	0.000001	-1	0.000007	0.000006	0.000013	0.000000	0.000013	0.000022
brush	0.000146	0.003354	0.002091	0.000084	0.004790	0.001487	0.000694	0.000039	0.000000	0.000001	0.000007	-1	0.733931	0.016312	0.000106	0.005093	0.009106
border	0.000152	0.002813	0.001855	0.000115	0.003978	0.001235	0.000576	0.000018	0.000000	0.000001	0.000006	0.733931	-1	0.012747	0.000088	0.005698	0.007628
enterocyte	0.000255	0.001498	0.001125	0.000051	0.000821	0.000038	0.001290	0.000041	0.000000	0.000001	0.000013	0.016312	0.012747	-1	0.000000	0.004637	0.008333
monocarboxylate	0.000329	0.006350	0.007158	0.000827	0.000185	0.001403	0.000000	0.004134	0.000000	0.004287	0.000000	0.000106	0.000088	0.000000	-1	0.000066	0.000331
ileum	0.000404	0.001257	0.000805	0.000038	0.002060	0.000065	0.000277	0.000009	0.000111	0.000007	0.000013	0.005093	0.005698	0.004637	0.000066	-1	0.025874
absorption	0.015901	0.021959	0.020522	0.001917	0.004720	0.001025	0.000714	0.000106	0.000022	0.000024	0.000022	0.009106	0.007628	0.008333	0.000331	0.025874	-1

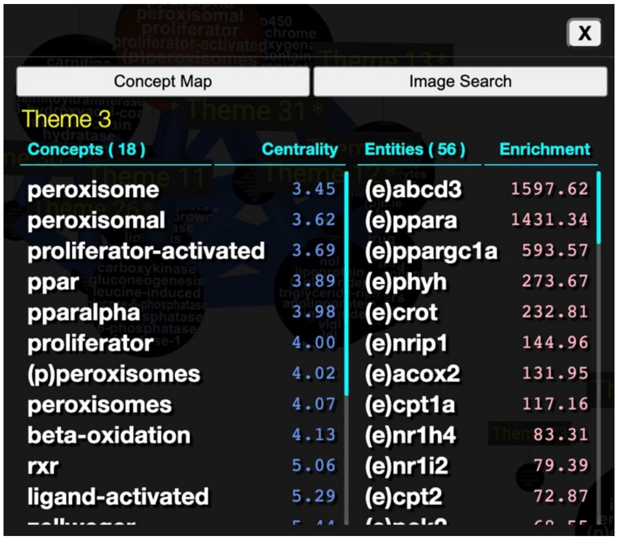
This approach allows the annotation process to take the contextual cues for the experiment into consideration while assigning titles to the theme. Given that there are typically several conceptual facets to each theme, the algorithm can assign up to three titles to each theme. In this manner, CompBio produces contextually relevant labels to themes in an automated fashion that can be accepted as-is or modified by the user for group creation.

Manual annotation can also be applied to the theme (theme tools are detailed in 5.4.4) allowing the user to specify a custom name for any or all themes.

3.1.2.2 *Concept-entity views*

The entities provided by the user when the CompBio project was created identify concepts that are clustered into themes. The concepts within a theme, and the entities that identified them, can be visualized by clicking on the theme of interest which brings up a **Theme Members Window**, shown below.

Theme Members Window



Theme 3

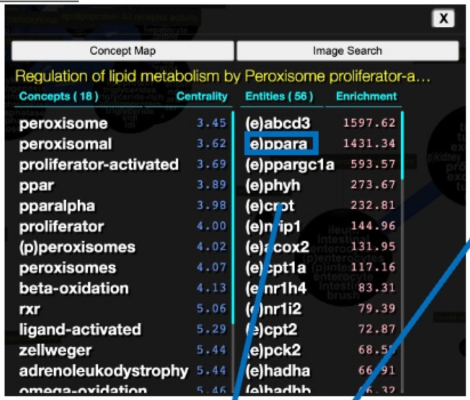
Concepts (18)	Centrality	Entities (56)	Enrichment
peroxisome	3.45	(e)abcd3	1597.62
peroxisomal	3.62	(e)ppara	1431.34
proliferator-activated	3.69	(e)ppargc1a	593.57
ppar	3.89	(e)phyh	273.67
pparalpha	3.98	(e)crot	232.81
proliferator	4.00	(e)nrip1	144.96
(p)peroxisomes	4.02	(e)acox2	131.95
peroxisomes	4.07	(e)cpt1a	117.16
beta-oxidation	4.13	(e)nr1h4	83.31
rxr	5.06	(e)nr1i2	79.39
ligand-activated	5.29	(e)cpt2	72.87

The theme concepts are listed on the left-hand side of the **Theme Members Window**, and the entities are listed on the right. The Centrality scores and Enrichment scores are discussed in sections 3.1.8 and 3.2.1, respectively.

3.1.2.3 Source literature

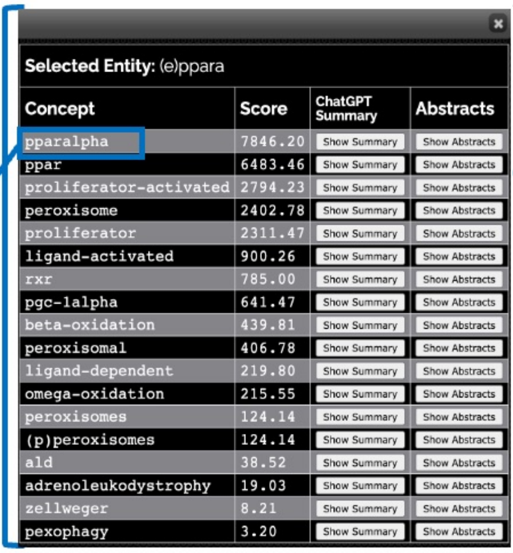
In CompBio, entities identify literature (abstracts and full text articles) and enriched concepts within those documents. The links between entities, concepts and the source literature is preserved in CompBio, allowing the user to drill down to the underlying literature that contains the entity (or one of its synonyms) and the concept by clicking on the concepts or entities within the **Theme Members Window**. This workflow is depicted below after clicking on a theme.

[**Section continues on next page with large image**]



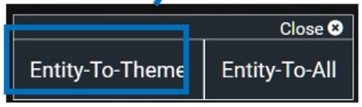
Regulation of lipid metabolism by Peroxisome proliferator-a...

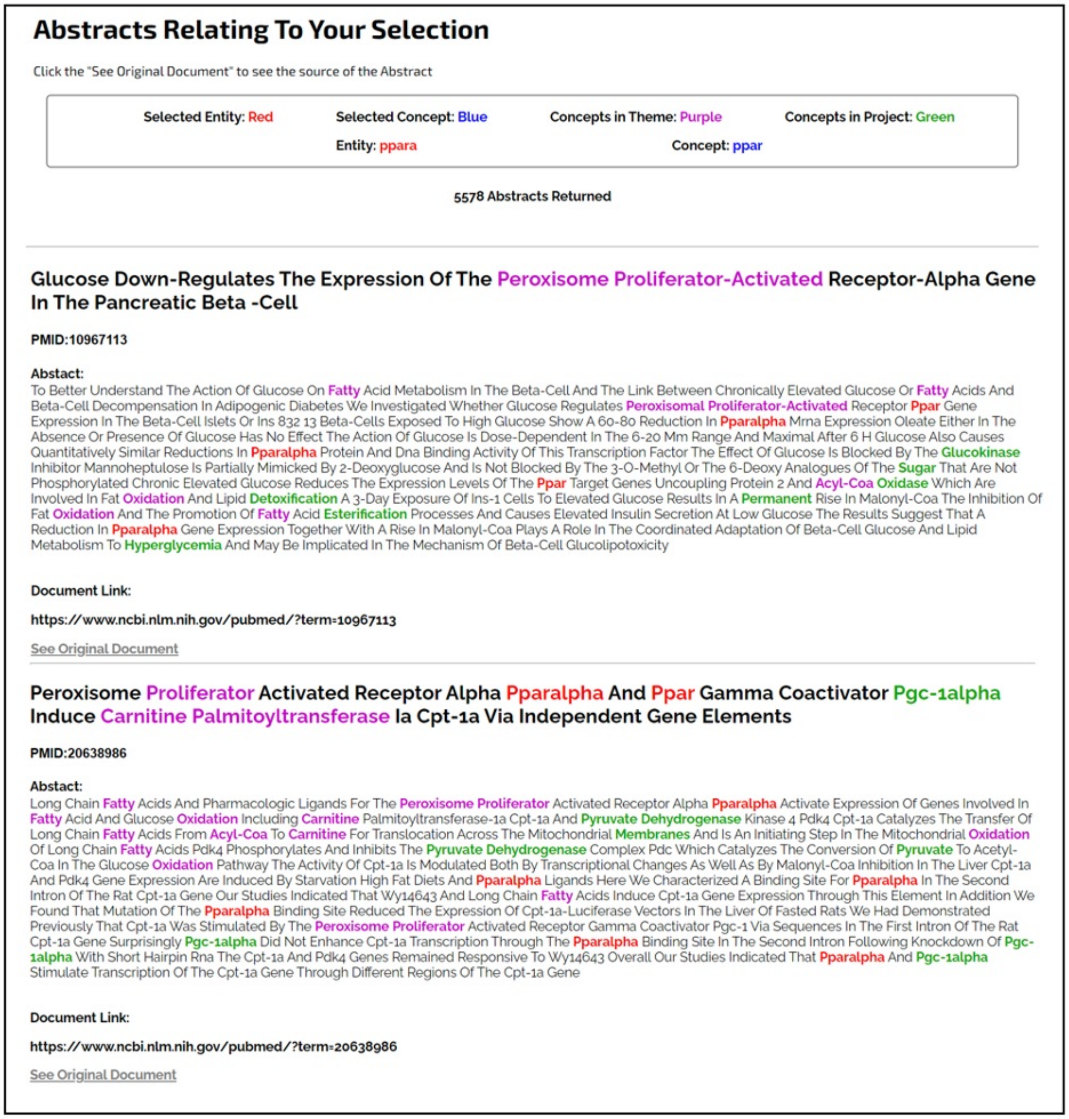
Concepts (18)	Centrality	Entities (56)	Enrichment
peroxisome	3.45	(e)abcd3	1597.62
peroxisomal	3.62	(e)ppara	1431.34
proliferator-activated	3.69	(e)ppargc1a	593.57
ppar	3.89	(e)phyh	273.67
pparalpha	3.98	(e)crot	232.81
proliferator	4.00	(e)nrip1	144.96
(p)peroxisomes	4.02	(e)acox2	131.95
peroxisomes	4.07	(e)cpt1a	117.16
beta-oxidation	4.13	(e)nr1h4	83.31
rxr	5.06	(e)nr1i2	79.39
ligand-activated	5.29	(e)cpt2	72.87
zellweger	5.44	(e)pck2	68.5
adrenoleukodystrophy	5.44	(e)hadha	66.91
omega-oxidation	5.46	(e)hadhb	66.32



Selected Entity: (e)ppara

Concept	Score	ChatGPT Summary	Abstracts
pparalpha	7846.20	Show Summary	Show Abstracts
ppar	6483.46	Show Summary	Show Abstracts
proliferator-activated	2794.23	Show Summary	Show Abstracts
peroxisome	2402.78	Show Summary	Show Abstracts
proliferator	2311.47	Show Summary	Show Abstracts
ligand-activated	900.26	Show Summary	Show Abstracts
rxr	785.00	Show Summary	Show Abstracts
pgc-1alpha	641.47	Show Summary	Show Abstracts
beta-oxidation	439.81	Show Summary	Show Abstracts
peroxisomal	406.78	Show Summary	Show Abstracts
ligand-dependent	219.80	Show Summary	Show Abstracts
omega-oxidation	215.55	Show Summary	Show Abstracts
peroxisomes	124.14	Show Summary	Show Abstracts
(p)peroxisomes	124.14	Show Summary	Show Abstracts
ald	38.52	Show Summary	Show Abstracts
adrenoleukodystrophy	19.03	Show Summary	Show Abstracts
zellweger	8.21	Show Summary	Show Abstracts
pexophagy	3.20	Show Summary	Show Abstracts





Abstracts Relating To Your Selection

Click the "See Original Document" to see the source of the Abstract

Selected Entity: Red **Selected Concept:** Blue **Concepts in Theme:** Purple **Concepts in Project:** Green

Entity: ppara **Concept:** ppar

5578 Abstracts Returned

Glucose Down-Regulates The Expression Of The Peroxisome Proliferator-Activated Receptor-Alpha Gene In The Pancreatic Beta -Cell

PMID:10967113

Abstact:
To Better Understand The Action Of Glucose On **Fatty** Acid Metabolism In The Beta-Cell And The Link Between Chronically Elevated Glucose Or **Fatty** Acids And Beta-Cell Decompensation In Adipogenic Diabetes We Investigated Whether Glucose Regulates **Peroxisomal Proliferator-Activated** Receptor **Ppar** Gene Expression In The Beta-Cell Islets Or Ins 832 13 Beta-Cells Exposed To High Glucose Show A 60-80 Reduction In **Pparalpha** Mrna Expression Oleate Either In The Absence Or Presence Of Glucose Has No Effect The Action Of Glucose Is Dose-Dependent In The 6-20 Mm Range And Maximal After 6 H Glucose Also Causes Quantitatively Similar Reductions In **Pparalpha** Protein And Dna Binding Activity Of This Transcription Factor The Effect Of Glucose Is Blocked By The **Glucokinase** Inhibitor Mannoheptulose Is Partially Mimicked By 2-Deoxyglucose And Is Not Blocked By The 3-O-Methyl Or The 6-Deoxy Analogues Of The **Sugar** That Are Not Phosphorylated Chronic Elevated Glucose Reduces The Expression Levels Of The **Ppar** Target Genes Uncoupling Protein 2 And **Acyl-Coa Oxidase** Which Are Involved In Fat **Oxidation** And Lipid **Detoxification** A 3-Day Exposure Of Ins-1 Cells To Elevated Glucose Results In A **Permanent** Rise In Malonyl-Coa The Inhibition Of Fat **Oxidation** And The Promotion Of **Fatty** Acid **Esterification** Processes And Causes Elevated Insulin Secretion At Low Glucose The Results Suggest That A Reduction In **Pparalpha** Gene Expression Together With A Rise In Malonyl-Coa Plays A Role In The Coordinated Adaptation Of Beta-Cell Glucose And Lipid Metabolism To **Hyperglycemia** And May Be Implicated In The Mechanism Of Beta-Cell Glucolipotoxicity

Document Link:

https://www.ncbi.nlm.nih.gov/pubmed/?term=10967113

See Original Document

Peroxisome Proliferator Activated Receptor Alpha Pparalpha And Ppar Gamma Coactivator Pgc-1alpha Induce Carnitine Palmitoyltransferase Ia Cpt-1a Via Independent Gene Elements

PMID:20638986

Abstact:
Long Chain **Fatty** Acids And Pharmacologic Ligands For The **Peroxisome Proliferator** Activated Receptor Alpha **Pparalpha** Activate Expression Of Genes Involved In **Fatty** Acid And Glucose **Oxidation** Including **Carnitine** Palmitoyltransferase-1a Cpt-1a And **Pyruvate Dehydrogenase** Kinase 4 Pdk4 Cpt-1a Catalyzes The Transfer Of Long Chain **Fatty** Acids From **Acyl-Coa** To **Carnitine** For Translocation Across The Mitochondrial **Membranes** And Is An Initiating Step In The Mitochondrial **Oxidation** Of Long Chain **Fatty** Acids Pdk4 Phosphorylates And Inhibits The **Pyruvate Dehydrogenase** Complex Pdc Which Catalyzes The Conversion Of **Pyruvate** To Acetyl-Coa In The Glucose **Oxidation** Pathway The Activity Of Cpt-1a Is Modulated Both By Transcriptional Changes As Well As By Malonyl-Coa Inhibition In The Liver Cpt-1a And Pdk4 Gene Expression Are Induced By Starvation High Fat Diets And **Pparalpha** Ligands Here We Characterized A Binding Site For **Pparalpha** In The Second Intron Of The Rat Cpt-1a Gene Our Studies Indicated That Wy14643 And Long Chain **Fatty** Acids Induce Cpt-1a Gene Expression Through This Element In Addition We Found That Mutation Of The **Pparalpha** Binding Site Reduced The Expression Of Cpt-1a-Luciferase Vectors In The Liver Of Fasted Rats We Had Demonstrated Previously That Cpt-1a Was Stimulated By The **Peroxisome Proliferator** Activated Receptor Gamma Coactivator Pgc-1 Via Sequences In The First Intron Of The Rat Cpt-1a Gene Surprisingly **Pgc-1alpha** Did Not Enhance Cpt-1a Transcription Through The **Pparalpha** Binding Site In The Second Intron Following Knockdown Of **Pgc-1alpha** With Short Hairpin Rna The Cpt-1a And Pdk4 Genes Remained Responsive To Wy14643 Overall Our Studies Indicated That **Pparalpha** And **Pgc-1alpha** Stimulate Transcription Of The Cpt-1a Gene Through Different Regions Of The Cpt-1a Gene

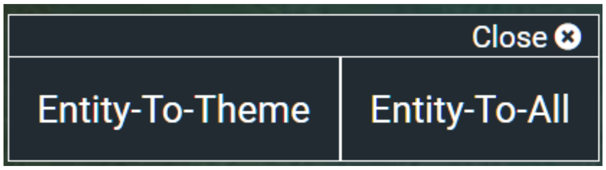
Document Link:

https://www.ncbi.nlm.nih.gov/pubmed/?term=20638986

See Original Document

Literature Scanning Page

Clicking on any entity listed on the right-hand side of the **Theme Members Window**, shown above, will spawn a **Scope Box** as shown below.



Scope Box

There are two options here, and this will determine the scope of the entity contribution scores returned.

- **Entity-To-Theme**: This option returns a list of concepts *from within the selected theme* and the score describing how much the selected entity contributed to the identification of that concept.
- **Entity-To-All**: This option returns a list of concepts *from the entire project* and the score describing how much the selected entity contributed to the identification of that concept.

The difference here is the scope of the results returned. This choice about scope does not apply when clicking a concept in the Theme Members Window because concepts only occur uniquely in a single theme, whereas entities, as mentioned previously, can contribute to multiple themes. Typically, users are trying to understand how a given entity contributed to the selected theme, so the Entity-To-Theme option is more typically chosen. The Entity-To-All option informs the user on the role that entity plays in the entire project.

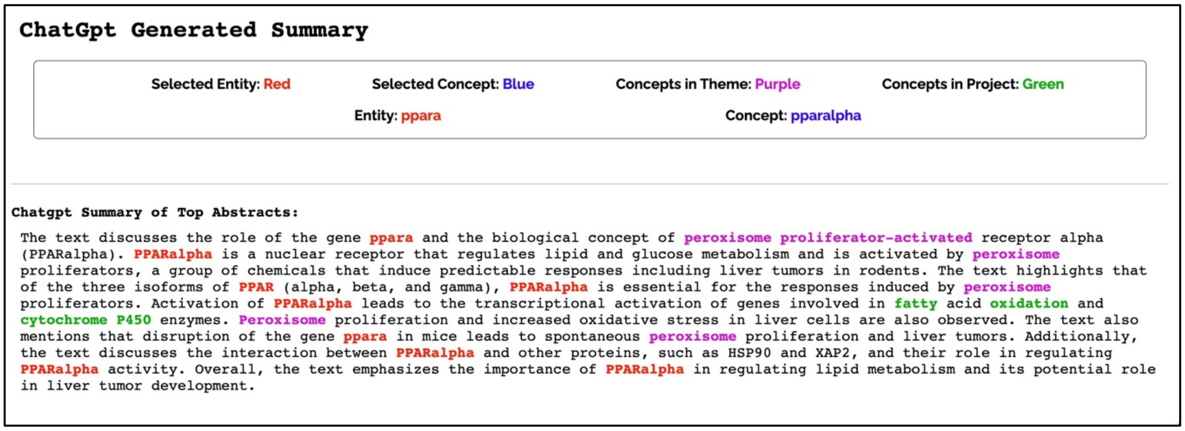
The literature results are returned in a **Literature Scanning Page** that contains condensed and formatted text. This condensed text lacks punctuation and is formatted to describe entity and concept occurrence in the knowledge map.

Below is a key describing the formatting contained in the **Literature Scanning Page:**

- **Red**: any occurrence of the entity (or any entity synonym) selected in the **Theme Members Window**
- **Blue**: any occurrence of the concept selected in the **Entity Window**.
- **Purple**: concepts appearing in the literature results that belong to the selected theme.
- **Green**: concepts contained in the current project that appear in the literature results, but do not belong to the selected theme.

The **Literature Scanning Page (**next page) summarizes the literature results from the selected entity-concept pair, but also serves as a gateway to the original underlying literature. The workflow to drill down into the original publications is shown below and simply requires clicking on the **See Original Document** link for the given abstract.

ChatGPT summaries: These summaries are generated by ChatGPT based on the information contained in the top abstracts identified by CompBio. They are meant to provide a short, rapid assessment of key information associated with the selected entity and concept in the context of the given theme. The number of abstracts can vary but typically falls into the range of the 8 to 10, if that many are available from the analysis. Please note, that the query restricts ChatGPT to simple summarization of the provided text to reduce the probability of ChatGPT introducing false statements. Within these constraints falsehoods have been rare, though not zero, so it is recommended to review the abstracts for confirmation if important information is being gleaned from the summary.



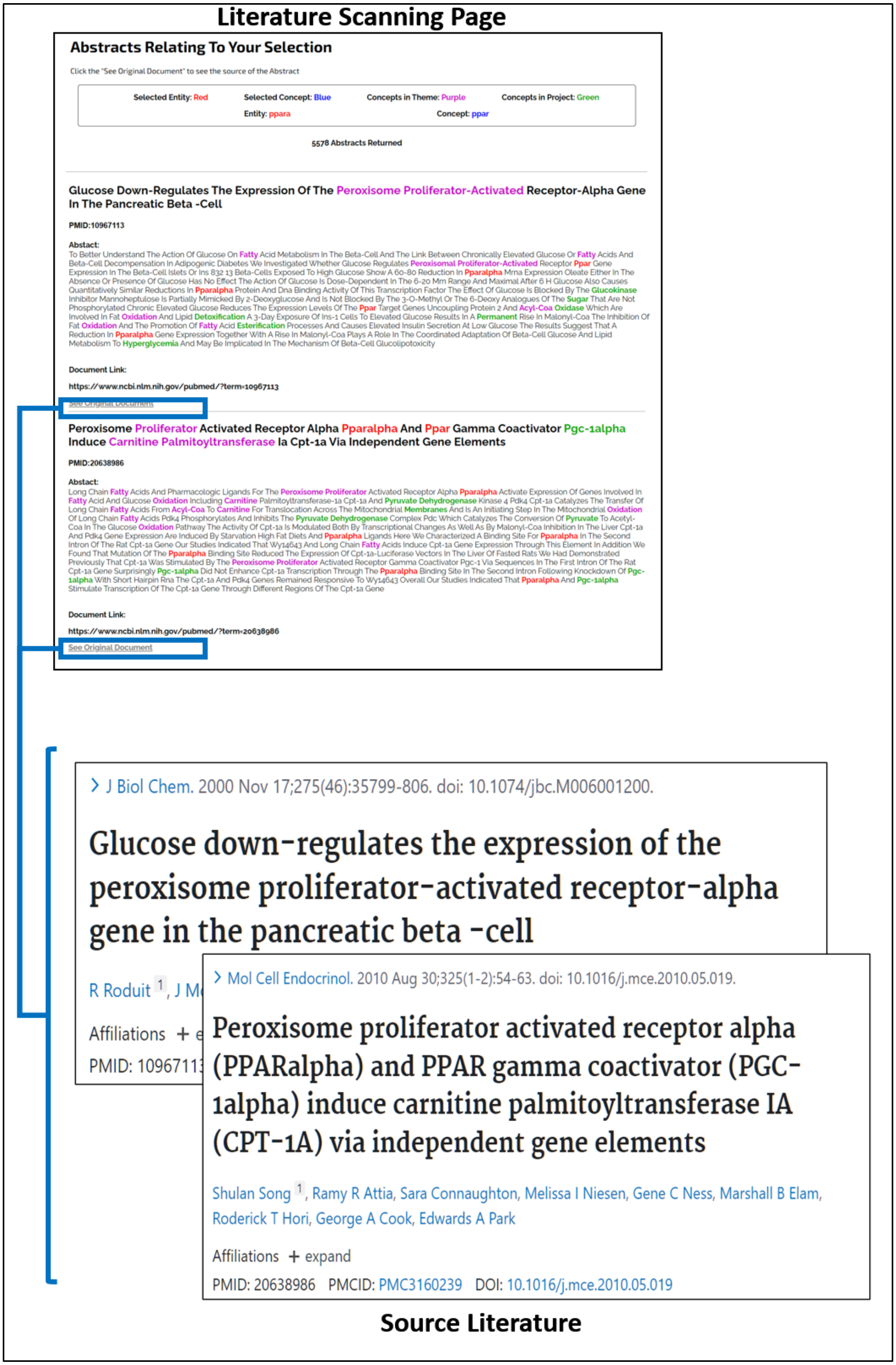
ChatGpt Generated Summary

Selected Entity: Red **Selected Concept:** Blue **Concepts in Theme:** Purple **Concepts in Project:** Green

Entity: ppara **Concept:** pparalpha

Chatgpt Summary of Top Abstracts:

The text discusses the role of the gene **ppara** and the biological concept of **peroxisome proliferator-activated** receptor alpha (PPARalpha). **PPARalpha** is a nuclear receptor that regulates lipid and glucose metabolism and is activated by **peroxisome** proliferators, a group of chemicals that induce predictable responses including liver tumors in rodents. The text highlights that of the three isoforms of **PPAR** (alpha, beta, and gamma), **PPARalpha** is essential for the responses induced by **peroxisome** proliferators. Activation of **PPARalpha** leads to the transcriptional activation of genes involved in **fatty** acid **oxidation** and **cytochrome P450** enzymes. **Peroxisome** proliferation and increased oxidative stress in liver cells are also observed. The text also mentions that disruption of the gene **ppara** in mice leads to spontaneous **peroxisome** proliferation and liver tumors. Additionally, the text discusses the interaction between **PPARalpha** and other proteins, such as HSP90 and XAP2, and their role in regulating **PPARalpha** activity. Overall, the text emphasizes the importance of **PPARalpha** in regulating lipid metabolism and its potential role in liver tumor development.



Literature Scanning Page

Abstracts Relating To Your Selection

Click the "See Original Document" to see the source of the Abstract

Selected Entity: Red Selected Concept: Blue Concepts in Theme: Purple Concepts in Project: Green

Entity: ppara Concept: ppar

5578 Abstracts Returned

Glucose Down-Regulates The Expression Of The Peroxisome Proliferator-Activated Receptor-Alpha Gene In The Pancreatic Beta -Cell

PMID:10967113

Abstact:
To Better Understand The Action Of Glucose On Fatty Acid Metabolism In The Beta-Cell And The Link Between Chronically Elevated Glucose Or Fatty Acids And Beta-Cell Decompensation In Adipogenic Diabetes We Investigated Whether Glucose Regulates Peroxisomal Proliferator-Activated Receptor Ppar Gene Expression In The Beta-Cell Islets Or Ins 832 13 Beta-Cells Exposed To High Glucose Show A 60-80 Reduction In Pparalpha Mrna Expression Oleate Either In The Absence Or Presence Of Glucose Has No Effect The Action Of Glucose Is Dose-Dependent In The 6-20 Mm Range And Maximal After 6 H Glucose Also Causes Quantitatively Similar Reductions In Pparalpha Protein And Dna Binding Activity Of This Transcription Factor The Effect Of Glucose Is Blocked By The Glucokinase Inhibitor Mannoheptulose Is Partially Mimicked By 2-Deoxyglucose And Is Not Blocked By The 3-O-Methyl Or The 6-Deoxy Analogues Of The Sugar That Are Not Phosphorylated Chronic Elevated Glucose Reduces The Expression Levels Of The Ppar Target Genes Uncoupling Protein 2 And Acyl-Coa Oxidase Which Are Involved In Fat Oxidation And Lipid Detoxification A 3-Day Exposure Of Ins-1 Cells To Elevated Glucose Results In A Permanent Rise In Malonyl-Coa The Inhibition Of Fat Oxidation And The Promotion Of Fatty Acid Esterification Processes And Causes Elevated Insulin Secretion At Low Glucose The Results Suggest That A Reduction In Pparalpha Gene Expression Together With A Rise In Malonyl-Coa Plays A Role In The Coordinated Adaptation Of Beta-Cell Glucose And Lipid Metabolism To Hyperglycemia And May Be Implicated In The Mechanism Of Beta-Cell Glucolipotoxicity

Document Link:

https://www.ncbi.nlm.nih.gov/pubmed/?term=10967113

See Original Document

Peroxisome Proliferator Activated Receptor Alpha Pparalpha And Ppar Gamma Coactivator Pgc-1alpha Induce Carnitine Palmitoyltransferase Ia Cpt-1a Via Independent Gene Elements

PMID:20638986

Abstact:
Long Chain Fatty Acids And Pharmacologic Ligands For The Peroxisome Proliferator Activated Receptor Alpha Pparalpha Activate Expression Of Genes Involved In Fatty Acid And Glucose Oxidation Including Carnitine Palmitoyltransferase-1a Cpt-1a And Pyruvate Dehydrogenase Kinase 4 Pdk4 Cpt-1a Catalyzes The Transfer Of Long Chain Fatty Acids From Acyl-Coa To Carnitine For Translocation Across The Mitochondrial Membranes And Is An Initiating Step In The Mitochondrial Oxidation Of Long Chain Fatty Acids Pdk4 Phosphorylates And Inhibits The Pyruvate Dehydrogenase Complex Pdc Which Catalyzes The Conversion Of Pyruvate To Acetyl-Coa In The Glucose Oxidation Pathway The Activity Of Cpt-1a Is Modulated Both By Transcriptional Changes As Well As By Malonyl-Coa Inhibition In The Liver Cpt-1a And Pdk4 Gene Expression Are Induced By Starvation High Fat Diets And Pparalpha Ligands Here We Characterized A Binding Site For Pparalpha In The Second Intron Of The Rat Cpt-1a Gene Our Studies Indicated That Wy14643 And Long Chain Fatty Acids Induce Cpt-1a Gene Expression Through This Element In Addition We Found That Mutation Of The Pparalpha Binding Site Reduced The Expression Of Cpt-1a-Luciferase Vectors In The Liver Of Fasted Rats We Had Demonstrated Previously That Cpt-1a Was Stimulated By The Peroxisome Proliferator Activated Receptor Gamma Coactivator Pgc-1 Via Sequences In The First Intron Of The Rat Cpt-1a Gene Surprisingly Pgc-1alpha Did Not Enhance Cpt-1a Transcription Through The Pparalpha Binding Site In The Second Intron Following Knockdown Of Pgc-1alpha With Short Hairpin Rna The Cpt-1a And Pdk4 Genes Remained Responsive To Wy14643 Overall Our Studies Indicated That Pparalpha And Pgc-1alpha Stimulate Transcription Of The Cpt-1a Gene Through Different Regions Of The Cpt-1a Gene

Document Link:

https://www.ncbi.nlm.nih.gov/pubmed/?term=20638986

See Original Document

> J Biol Chem. 2000 Nov 17;275(46):35799-806. doi: 10.1074/jbc.M006001200.

Glucose down-regulates the expression of the peroxisome proliferator-activated receptor-alpha gene in the pancreatic beta -cell

R Roduit [1], J M

Affiliations + e

PMID: 1096711

> Mol Cell Endocrinol. 2010 Aug 30;325(1-2):54-63. doi: 10.1016/j.mce.2010.05.019.

Peroxisome proliferator activated receptor alpha (PPARalpha) and PPAR gamma coactivator (PGC-1alpha) induce carnitine palmitoyltransferase IA (CPT-1A) via independent gene elements

Shulan Song [1], Ramy R Attia, Sara Connaughton, Melissa I Niesen, Gene C Ness, Marshall B Elam, Roderick T Hori, George A Cook, Edwards A Park

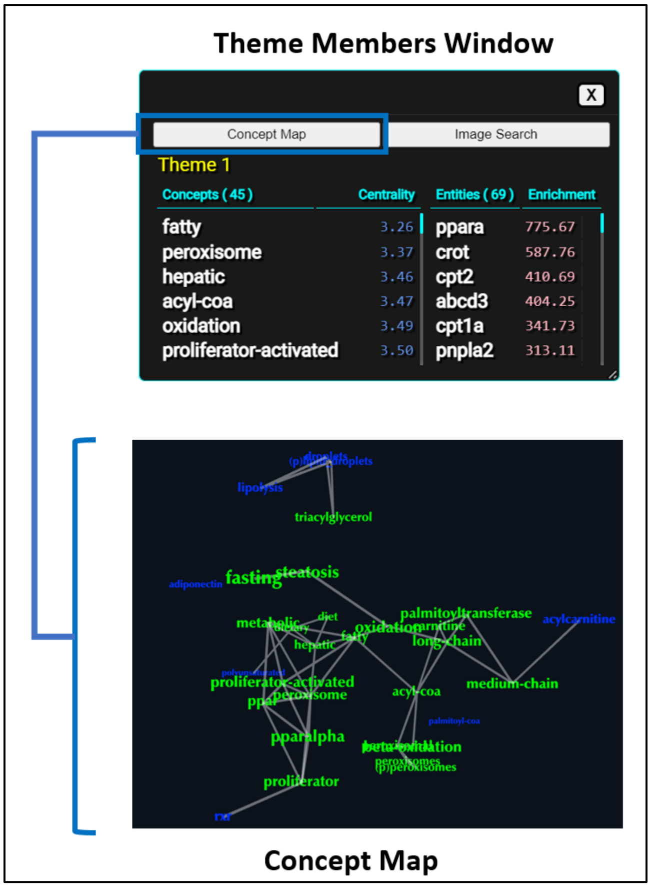
Affiliations + expand

PMID: 20638986 PMCID: PMC3160239 DOI: 10.1016/j.mce.2010.05.019

Source Literature

3.1.2.4 Concept map

A theme represents a clustering of the concepts that were positioned in three-dimensional space based on the conditional probability of those concepts co-occurring with each other in the literature identified by the entities. As such, the concepts within a theme therefore can be represented with a **Concept Map** describing their level of cooccurrence with each other. Concepts are ranked in the **Theme Members Window** according to how central they are to the theme (discussed in detail in 3.1.2.5.1). The **Concept Map** helps to understand the central concepts in the theme enabling the best manual theme annotation possible. The workflow for accessing the **Concept Map** is shown below. Note that the underlying literature is also accessible by clicking on edges in a workflow similar to that described in section 3.1.2.6 of this manual.

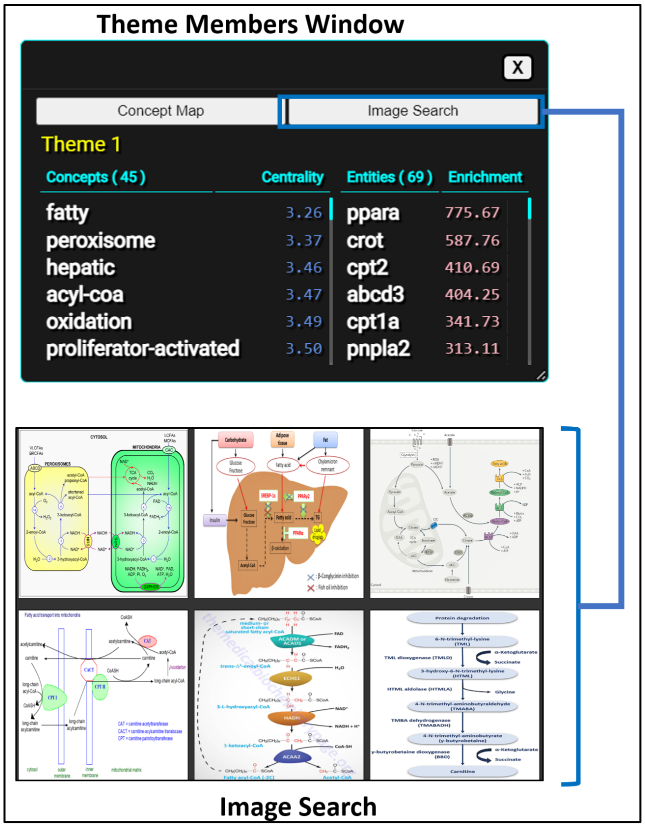


Theme 1

Concepts (45)	Centrality	Entities (69)	Enrichment
fatty	3.26	ppara	775.67
peroxisome	3.37	crot	587.76
hepatic	3.46	cpt2	410.69
acyl-coa	3.47	abcd3	404.25
oxidation	3.49	cpt1a	341.73
proliferator-activated	3.50	pnpla2	313.11

3.1.2.5 Image search

To inform the user about processes and pathways associated with the theme, a modified Google Image search has been implemented to return (primarily) scientific results related to the concepts contained in the theme. This can be useful in manual theme annotation as well as obtaining images for presentation purposes, or to obtain review articles related to theme concepts. This image searching utilizes all of the concepts within the theme, but customized searches are also possible (described in 3.1.2.5.2). The **Image Search** workflow is shown in the following image.



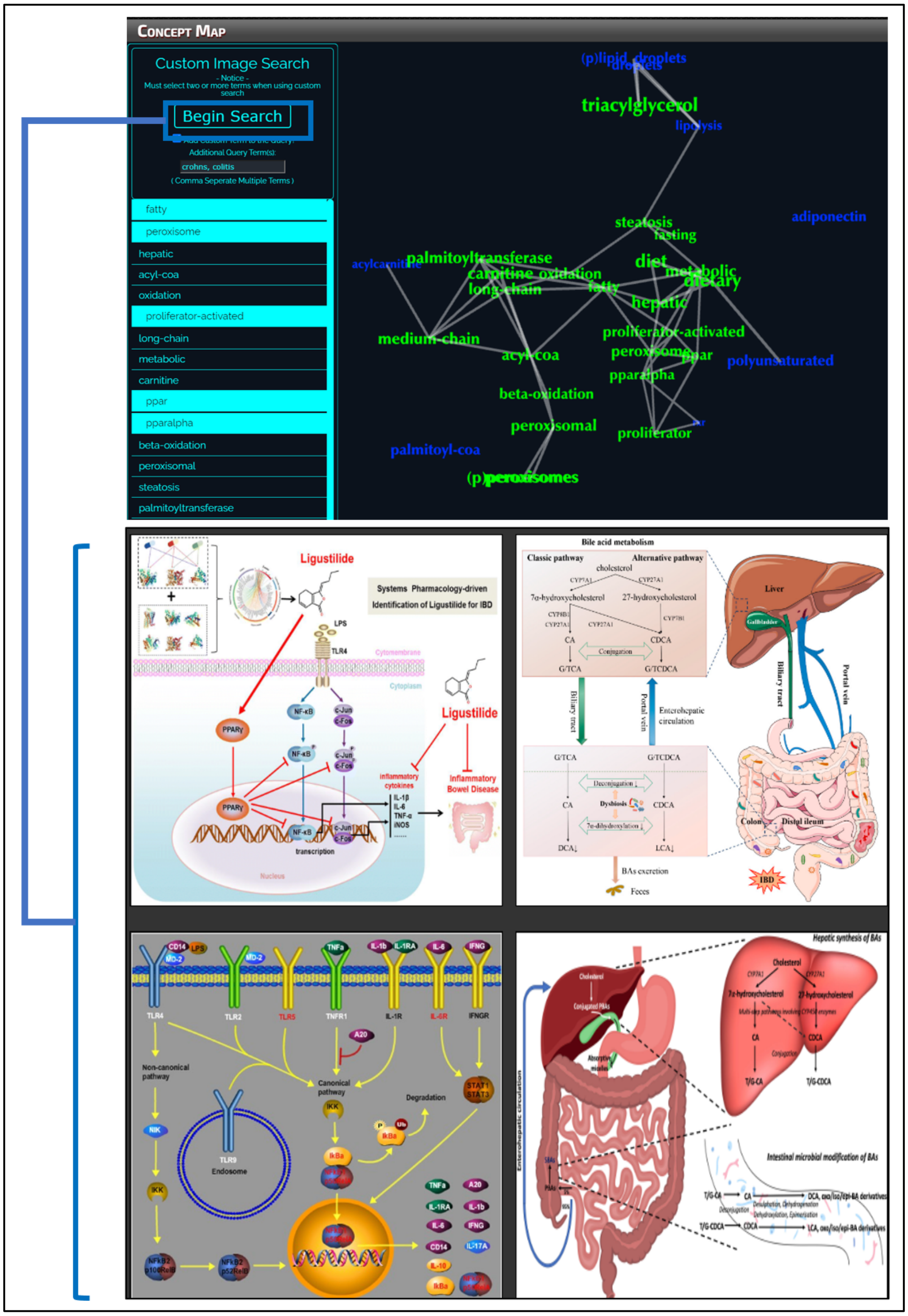
Concepts (45)	Centrality	Entities (69)	Enrichment
fatty	3.26	ppara	775.67
peroxisome	3.37	crot	587.76
hepatic	3.46	cpt2	410.69
acyl-coa	3.47	abcd3	404.25
oxidation	3.49	cpt1a	341.73
proliferator-activated	3.50	pnpla2	313.11

3.1.2.5.1 *Centrality*

The main CompBio knowledge map is created by arranging the project concepts in three-dimensional space based on their contextual cooccurrence, and then clustering them into themes. This approach means that each theme contains a subgraph of the concepts in 3D space. Accessing the Concept Map is described in 3.1.2.4. A **Centrality** score is calculated describing the distance of the concepts from the center of the theme, which is used to rank the concepts. Entities with a low distance from the center (low **Centrality** score) can imply high relevance or could describe a midpoint between two similar but distinct processes clustered into a single theme. In the **Concept Map** shown in 3.1.2.4, "fatty" is the most central concept, but the map shows "fatty" as a mid-point between two distinct but related concept branches, namely PPAR/Peroxisome concepts on the left and lipid metabolism on the right. It can be seen in 3.1.2.1 that the auto-annotation labels assigned to this theme correctly capture both of these biological facets of the same theme.

3.1.2.5.2 *Custom image search*

In addition to the **Image Search** described in 3.1.2.5, it is also possible to perform a custom search from within the **Concept Map** described in 3.1.2.4. This allows the image search to be performed on a subset of the concepts and can be helpful in themes that have more than one biological facet. Another powerful feature of this tool is that **Additional Query Term(s)** can be specified in a comma delimited list. This allows the user to understand the selected themes within the context of the additional terms. An example search is shown below; the search is initiated by clicking **Begin Search**.

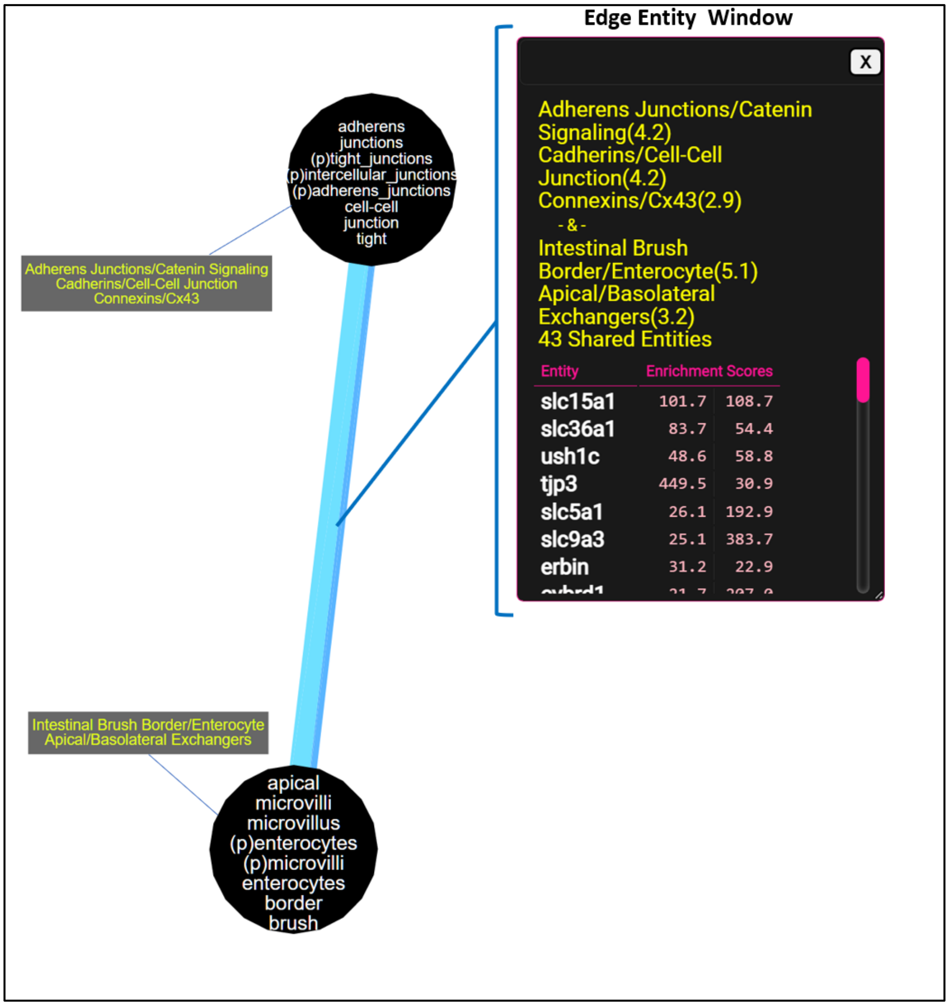


3.1.2.6 *Edges*

The knowledge map consists of entities and concepts clustered into themes represented by spheres. These themes can be, and often are, interconnected with each other by edges. An edge can potentially be formed so long as there is at least one gene that contributes to one or more concepts with an enrichment greater than 2.0 in both themes. Greater numbers of shared genes will cause the given edge to increase in thickness and low numbers of shared genes will similarly cause a thinner line to be displayed.

3.1.2.6.1 *Shared entities*

Selection of an edge in the project window will spawn an **Edge Entity Window** that contains the theme labels and the entities shared between the two themes. As described in 3.1.2.6, edges can potentially be drawn if the same entity contributes to concepts in two themes, so long as the distance threshold criteria (described in 5.4.3) are met. These shared entities are displayed in the **Edge Entity Window** as shown below**.**

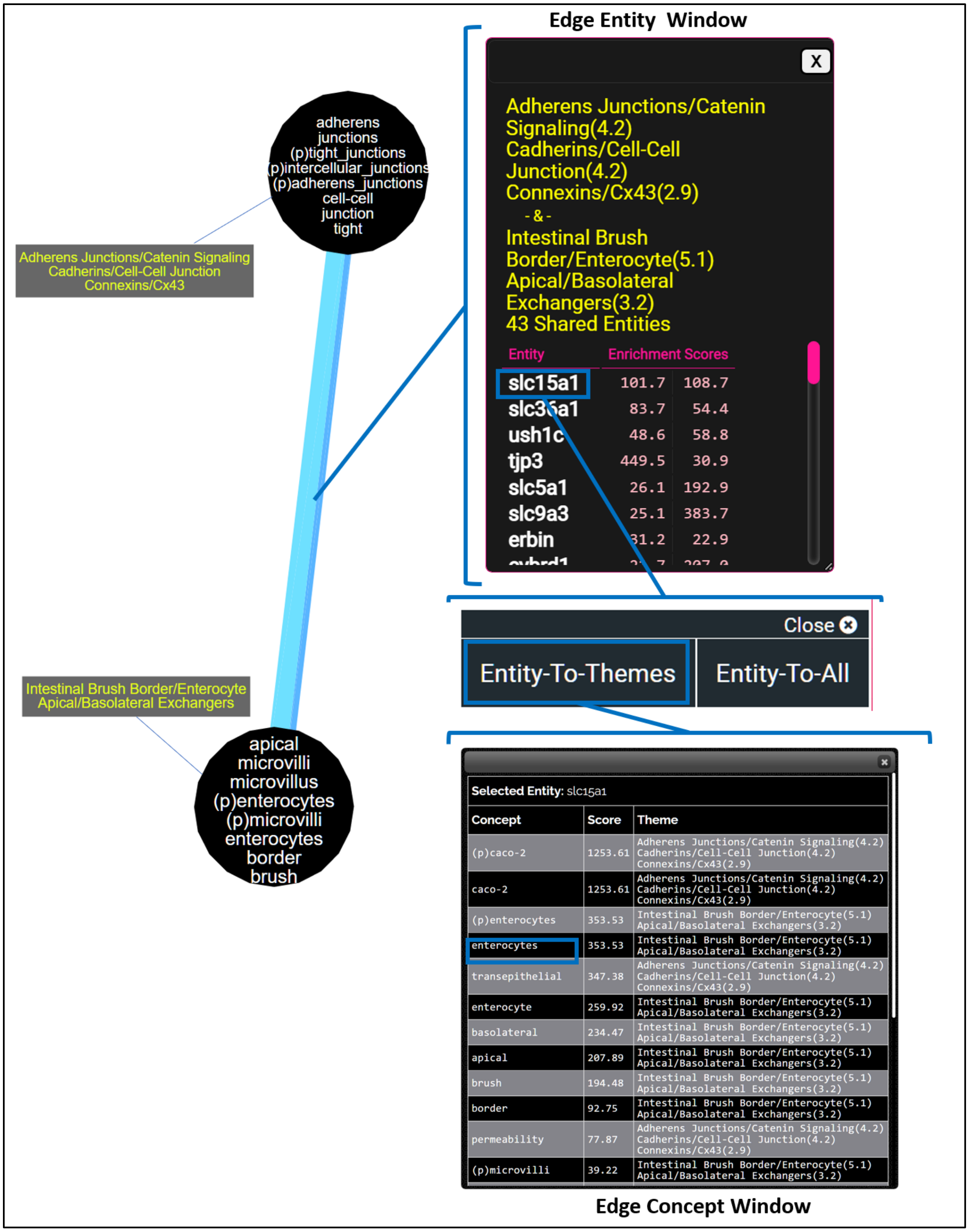


Adherens Junctions/Catenin Signaling(4.2)
Cadherins/Cell-Cell Junction(4.2)
Connexins/Cx43(2.9)
- & -
Intestinal Brush Border/Enterocyte(5.1)
Apical/Basolateral Exchangers(3.2)
43 Shared Entities

Entity	Enrichment Scores	
slc15a1	101.7	108.7
slc36a1	83.7	54.4
ush1c	48.6	58.8
tjp3	449.5	30.9
slc5a1	26.1	192.9
slc9a3	25.1	383.7
erbin	31.2	22.9

3.1.3.6.2 *Source literature*

The information that drove concept enrichment for this project and enabled edge formation is traceable back to the original abstracts. This enables the biologist to understand and drill down into the publications that identify a given concept being related to the selected gene. The figure below demonstrates clicking on an edge to display the shared contributing genes in the **Edge Entity Window** and drilling down to the abstracts that link that gene to the selected concept displayed in the **Edge Concept Window.** Clicking on any entry within the **Concept** will immediately take the user to the Abstract Scanning page with the ability to access the original abstracts or publications. See section 3.1.2.3 for more detail.



Edge Entity Window

Adherens Junctions/Catenin Signaling(4.2)
Cadherins/Cell-Cell Junction(4.2)
Connexins/Cx43(2.9)
- & -
Intestinal Brush Border/Enterocyte(5.1)
Apical/Basolateral Exchangers(3.2)
43 Shared Entities

Entity	Enrichment Scores	
slc15a1	101.7	108.7
slc36a1	83.7	54.4
ush1c	48.6	58.8
tjp3	449.5	30.9
slc5a1	26.1	192.9
slc9a3	25.1	383.7
erbin	31.2	22.9

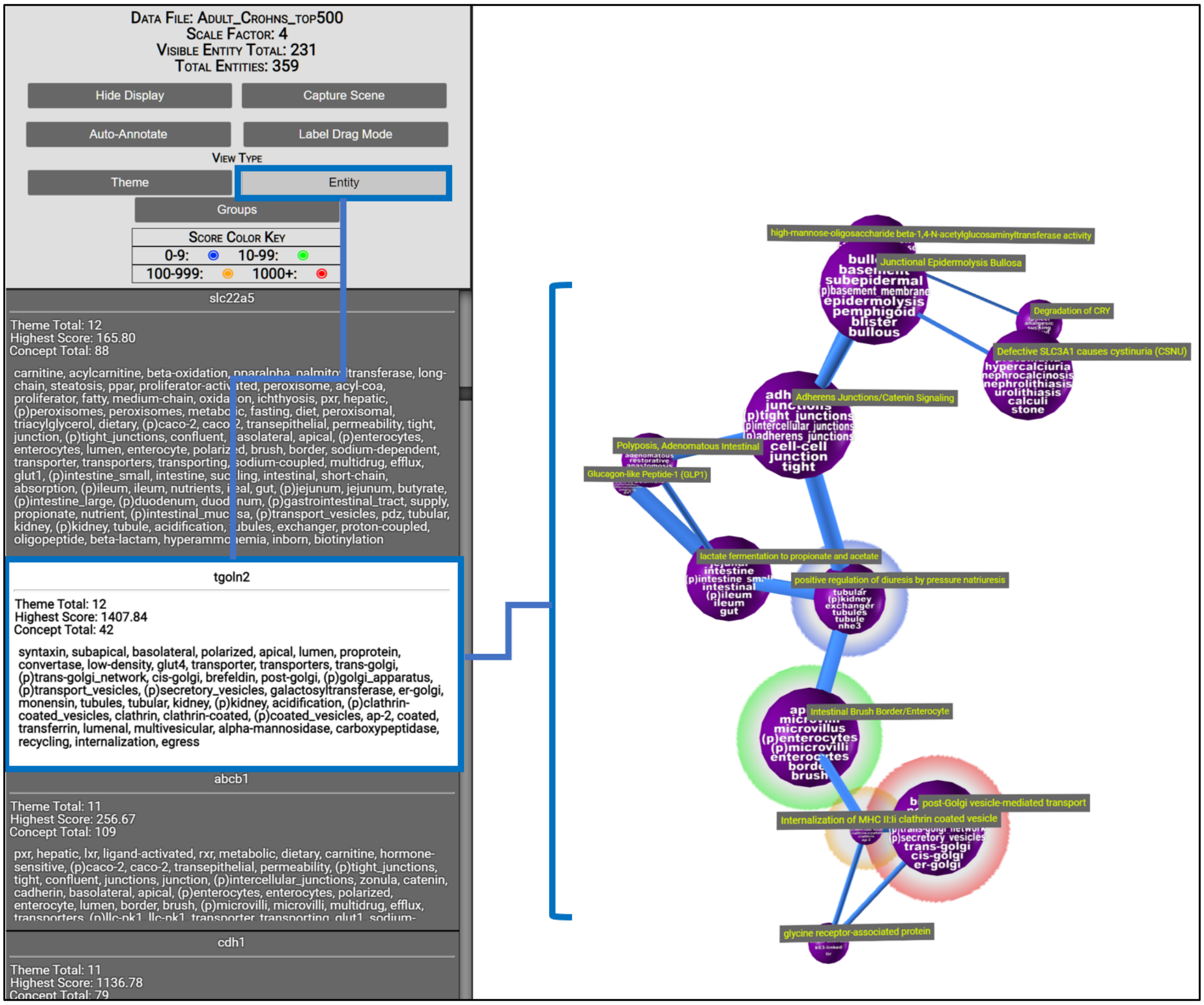
Selected Entity: slc15a1

Concept	Score	Theme
(p)caco-2	1253.61	Adherens Junctions/Catenin Signaling(4.2) Cadherins/Cell-Cell Junction(4.2) Connexins/Cx43(2.9)
caco-2	1253.61	Adherens Junctions/Catenin Signaling(4.2) Cadherins/Cell-Cell Junction(4.2) Connexins/Cx43(2.9)
(p)enterocytes	353.53	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
enterocytes	353.53	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
transepithelial	347.38	Adherens Junctions/Catenin Signaling(4.2) Cadherins/Cell-Cell Junction(4.2) Connexins/Cx43(2.9)
enterocyte	259.92	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
basolateral	234.47	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
apical	207.89	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
brush	194.48	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
border	92.75	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)
permeability	77.87	Adherens Junctions/Catenin Signaling(4.2) Cadherins/Cell-Cell Junction(4.2) Connexins/Cx43(2.9)
(p)microvilli	39.22	Intestinal Brush Border/Enterocyte(5.1) Apical/Basolateral Exchangers(3.2)

Edge Concept Window

3.1.3 *Entity view*

The **Entity View** allows the user to interact with the knowledge map from the standpoint of the entities, as opposed to the themes and concepts (**Theme View,** discussed in 3.1.2). The **Entity View** allows the user to see which entities contributed to concept identification across the themes, and to what degree. In the workflow shown below, after the **Entity** button is clicked, an entity can be selected from the **Theme/Entity Detail Panel**, and the visible themes will be haloed with colors representing the **Enrichment Score** of the selected entity for those themes. The color scale key is displayed in the **Project Panel** when the **Entity** button is clicked. This approach is useful to understand the selected entity contribution to all visible themes.



DATA FILE: ADULT_CROHNS_TOP500
SCALE FACTOR: 4
VISIBLE ENTITY TOTAL: 231
TOTAL ENTITIES: 359

Hide Display	Capture Scene
Auto-Annotate	Label Drag Mode

VIEW TYPE

Theme	Entity

Groups

SCORE COLOR KEY			
0-9:	🔵	10-99:	🟢
100-999:	🟠	1000+:	🔴

slc22a5

Theme Total: 12
Highest Score: 165.80
Concept Total: 88

carnitine, acylcarnitine, beta-oxidation, pparalpha, palmitoyltransferase, long-chain, steatosis, ppar, proliferator-activated, peroxisome, acyl-coa, proliferator, fatty, medium-chain, oxidation, ichthyosis, pxr, hepatic, (p)peroxisomes, peroxisomes, metabolic, fasting, diet, peroxisomal, triacylglycerol, dietary, (p)caco-2, caco-2, transepithelial, permeability, tight, junction, (p)tight_junctions, confluent, basolateral, apical, (p)enterocytes, enterocytes, lumen, enterocyte, polarized, brush, border, sodium-dependent, transporter, transporters, transporting, sodium-coupled, multidrug, efflux, glut1, (p)intestine_small, intestine, suckling, intestinal, short-chain, absorption, (p)ileum, ileum, nutrients, ileal, gut, (p)jejunum, jejunum, butyrate, (p)intestine_large, (p)duodenum, duodenum, (p)gastrointestinal_tract, supply, propionate, nutrient, (p)intestinal_mucosa, (p)transport_vesicles, pdz, tubular, kidney, (p)kidney, tubule, acidification, tubules, exchanger, proton-coupled, oligopeptide, beta-lactam, hyperammonemia, inborn, biotinylation

tgoln2

Theme Total: 12
Highest Score: 1407.84
Concept Total: 42

syntaxin, subapical, basolateral, polarized, apical, lumen, proprotein, convertase, low-density, glut4, transporter, transporters, trans-golgi, (p)trans-golgi_network, cis-golgi, brefeldin, post-golgi, (p)golgi_apparatus, (p)transport_vesicles, (p)secretory_vesicles, galactosyltransferase, er-golgi, monensin, tubules, tubular, kidney, (p)kidney, acidification, (p)clathrin-coated_vesicles, clathrin, clathrin-coated, (p)coated_vesicles, ap-2, coated, transferrin, lumenal, multivesicular, alpha-mannosidase, carboxypeptidase, recycling, internalization, egress

abcb1

Theme Total: 11
Highest Score: 256.67
Concept Total: 109

pxr, hepatic, lxr, ligand-activated, rxr, metabolic, dietary, carnitine, hormone-sensitive, (p)caco-2, caco-2, transepithelial, permeability, (p)tight_junctions, tight, confluent, junctions, junction, (p)intercellular_junctions, zonula, catenin, cadherin, basolateral, apical, (p)enterocytes, enterocytes, polarized, enterocyte, lumen, border, brush, (p)microvilli, microvilli, multidrug, efflux, transporters, (p)llc-pk1, llc-pk1, transporter, transporting, glut1, sodium-

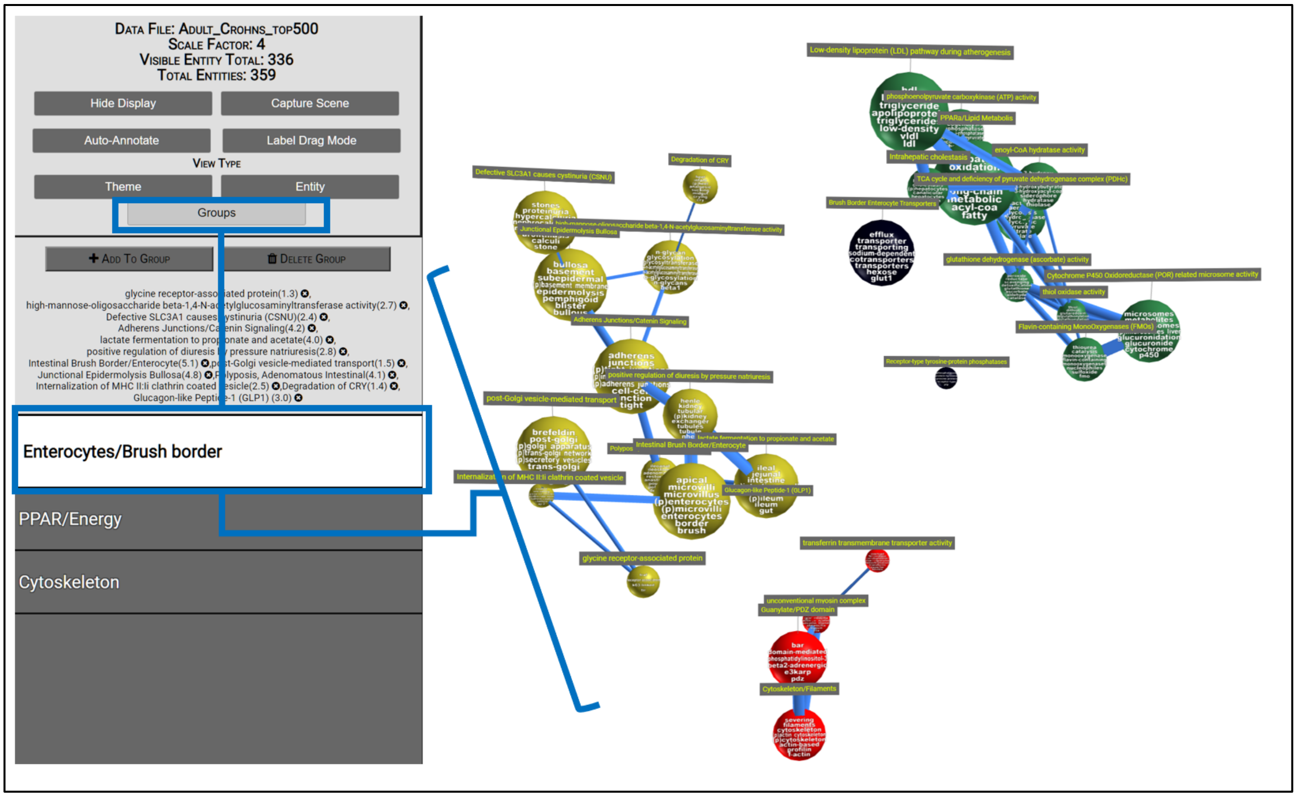
cdh1

Theme Total: 11
Highest Score: 1136.78
Concept Total: 79

high-mannose-oligosaccharide beta-1,4-N-acetylglucosaminyltransferase activity
Junctional Epidermolysis Bullosa
basement
subepidermal
(p)basement_membrane
epidermolysis
pemphigoid
blister
bullous
Degradation of CRY
Defective SLC3A1 causes cystinuria (CSNU)
hypercalciuria
nephrocalcinosis
nephrolithiasis
urolithiasis
calculi
stone
Adherens Junctions/Catenin Signaling
(p)tight_junctions
(p)intercellular_junctions
(p)adherens_junctions
cell-cell
junction
tight
Polyposis, Adenomatous Intestinal
Glucagon-like Peptide-1 (GLP1)
lactate fermentation to propionate and acetate
intestine
(p)intestine_small
intestinal
(p)ileum
ileum
gut
positive regulation of diuresis by pressure natriuresis
tubular
(p)kidney
exchanger
tubules
tubule
nhe3
Intestinal Brush Border/Enterocyte
microvillus
(p)enterocytes
(p)microvilli
enterocytes
brush
post-Golgi vesicle-mediated transport
Internalization of MHC II:Ii clathrin coated vesicle
(p)secretory_vesicles
trans-golgi
cis-golgi
er-golgi
glycine receptor-associated protein

3.1.4 *Group view*

Once the theme annotation process is complete, either by auto-annotation or manual annotation (5.1.4), themes can be bundled into groups of similar biological function (5.4.5). The **Groups** button then allows the user to interact with the knowledge map at the group level. The workflow shown below details entering **Group View** and selecting a group. In this mode, group-level coloring can be applied (5.4.5). Additionally, newly selected themes can be added to the group using the **+ Add to Group** button, removed from the theme by clicking the "x" bullet preceding the theme name, or the entire group can be deleted using the **Delete Group** button. Note that the **Delete Group** button does not delete the raw themes, but rather removes the group name from the **Group View** and discards any group-level formatting.



Data File: Adult_Crohns_top500
Scale Factor: 4
Visible Entity Total: 336
Total Entities: 359

glycine receptor-associated protein(1.3) ⊗,
high-mannose-oligosaccharide beta-1,4-N-acetylglucosaminyltransferase activity(2.7) ⊗,
Defective SLC3A1 causes cystinuria (CSNU)(2.4) ⊗,
Adherens Junctions/Catenin Signaling(4.2) ⊗,
lactate fermentation to propionate and acetate(4.0) ⊗,
positive regulation of diuresis by pressure natriuresis(2.8) ⊗,
Intestinal Brush Border/Enterocyte(5.1) ⊗,post-Golgi vesicle-mediated transport(1.5) ⊗,
Junctional Epidermolysis Bullosa(4.8) ⊗,Polyposis, Adenomatous Intestinal(4.1) ⊗,
Internalization of MHC II:Ii clathrin coated vesicle(2.5) ⊗,Degradation of CRY(1.4) ⊗,
Glucagon-like Peptide-1 (GLP1) (3.0) ⊗

Hide Display
Capture Scene
Auto-Annotate
Label Drag Mode
View Type
Theme
Entity
Groups
+ Add To Group
Delete Group
Enterocytes/Brush border
PPAR/Energy
Cytoskeleton
Low-density lipoprotein (LDL) pathway during atherogenesis
phosphoenolpyruvate carboxykinase (ATP) activity
PPARa/Lipid Metabolism
Intrahepatic cholestasis
enoyl-CoA hydratase activity
TCA cycle and deficiency of pyruvate dehydrogenase complex (PDHc)
Brush Border Enterocyte Transporters
glutathione dehydrogenase (ascorbate) activity
Cytochrome P450 Oxidoreductase (POR) related microsome activity
thiol oxidase activity
Flavin-containing MonoOxygenases (FMOs)
Receptor-type tyrosine-protein phosphatases
Degradation of CRY
Defective SLC3A1 causes cystinuria (CSNU)
Junctional Epidermolysis Bullosa
Adherens Junctions/Catenin Signaling
positive regulation of diuresis by pressure natriuresis
post-Golgi vesicle-mediated transport
lactate fermentation to propionate and acetate
Polyposis
Intestinal Brush Border/Enterocyte
Internalization of MHC II:Ii clathrin coated vesicle
Glucagon-like Peptide-1 (GLP1)
glycine receptor-associated protein
transferrin transmembrane transporter activity
unconventional myosin complex
Guanylate/PDZ domain
Cytoskeleton/Filaments

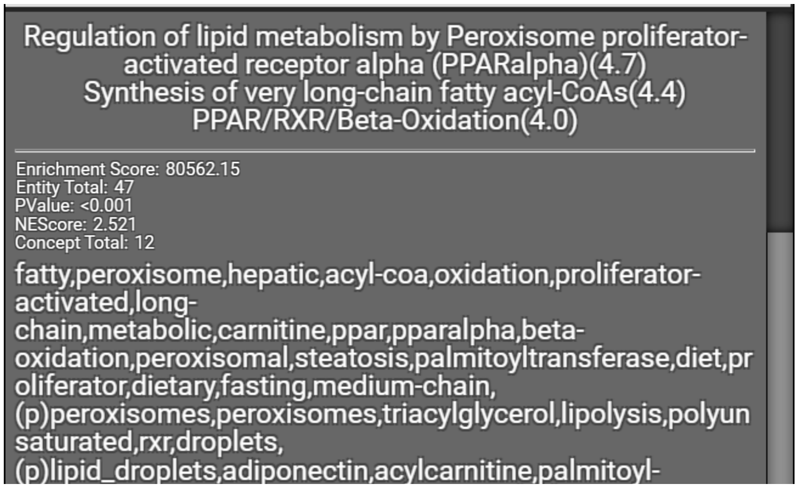
3.2 Understanding the Data

The focus of CompBio is to identify the enriched concepts given a list of input entities. To enable this objective, statistics and metrics are employed to allow the biologist tasked with data interpretation to assess the relative significance and enrichment levels of the identified concepts, as well as deciding objectively when to conclude the annotation process.

3.2.1 Significance

When a project is run in CompBio the project is compared against thousands of randomized comparator projects of similar entity list size. This allows the algorithm to calculate enrichment scores for concepts and p-values for the themes as compared to random noise. The level of precision reported in these metrics is determined by the number of randomized projects that have been run. In the current version of CompBio (v 2.5), p-values are reported down to the .001 level. Any p-value lower or equal to 0.001 are simply reported as "<0.001".

In the **Theme Detail Panel** on the lower left side of the workspace, significance and enrichment metrics are displayed as well as concepts contained within the themes and auto-annotation (if applied), as shown below.



Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)(4.7)
Synthesis of very long-chain fatty acyl-CoAs(4.4)
PPAR/RXR/Beta-Oxidation(4.0)

Enrichment Score: 80562.15
Entity Total: 47
PValue: <0.001
NEScore: 2.521
Concept Total: 12

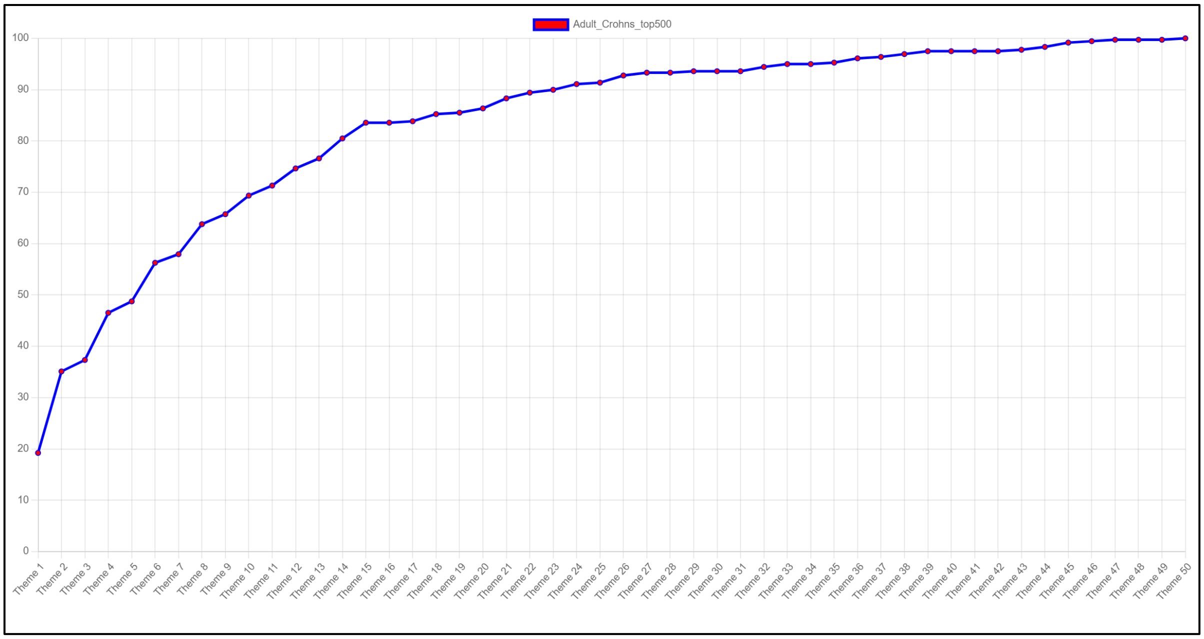
fatty,peroxisome,hepatic,acyl-coa,oxidation,proliferator-activated,long-chain,metabolic,carnitine,ppar,pparalpha,beta-oxidation,peroxisomal,steatosis,palmitoyltransferase,diet,proliferator,dietary,fasting,medium-chain,(p)peroxisomes,peroxisomes,triacylglycerol,lipolysis,polyunsaturated,rxr,droplets,(p)lipid_droplets,adiponectin,acylcarnitine,palmitoyl-

There are five key metrics displayed for each theme that enable the biologist to determine the robustness of the theme. Weaker scoring yet significant themes tend to provide a smaller degree of useful information towards the interpretation of the dataset as compared with the higher scoring themes. This will be discussed more thoroughly in the next section describing entity mapping efficiency.

- **Enrichment Score:** magnitude of enrichment of all of the concepts in the given theme over randomly generated themes of the same rank with similar numbers of input entities
- **Entity Total:** number of entities that contributed to identifying concepts in the theme
- **P-Value:** p-statistic calculated to a describe the confidence that the given theme is significantly enriched over random noise. So long as the **NEScore is >= 1.3**, p-values as high as 0.1 should be given consideration as a 0.05 cutoff tends to be overly strict with suitably enriched CompBio themes
- **NEScore:** the given theme's **Enrichment Score** normalized to the enrichment of the same theme rank in projects created with random entity lists within the same size range
- **Concept Total:** total number of concepts contained within the given theme

3.2.2 Entity mapping efficiency

One of the strengths of CompBio is the efficiency with which it is able to map the provided entities to biological concepts and themes. Other tools in widespread use tend to produce large amounts of statistically relevant pathway results that need to be reviewed and analyzed to determine what is meaningful. CompBio addresses these issues by producing a graph of the percentage of user-supplied entities mapped to concepts, which can be accessed through the **%Entities Mapped** button on the bottom right-hand side of the workspace (graph shown below).



As can be seen in the above graph, themes are listed across the X-axis and the cumulative number of entities mapped is shown on the Y-axis. By theme 23, over 90% of the provided entity list has been mapped to at least one biological concept. This provides a logical and meaningful cutoff when deciding what data to consider and what to set aside. This is important in the annotation workflow, as it is possible for a project to have over 50 significant themes and limiting the interpretation to just the primary mappings can save significant amounts of time in the annotation process.

3.2.3 *General quality control*

It is important to understand the level of biological signal in a dataset at the outset of analysis, and CompBio has several tools to enable this. The **Theme Detail Panel** contains metrics that describe the levels of enrichment of the biological signal as well as the level of significance. The **Enrichment Score** describes the degree of signal enrichment of the concepts in a given theme, over random noise. This can be thought of as a "fold change" over random noise. The theme order is defined by **Enrichment Score.**

The **NEScore** is based on the **Enrichment Score** but has been normalized to randomly generated themes of the same rank number. Therefore, the **NEScore** for theme 1 in a given project is calculated by normalizing the **Enrichment Score** for that theme against thousands of theme 1 **Enrichment Scores** from knowledge maps generated from randomly created lists of entities. This puts the themes on a common scale, regardless of rank, allowing for comparison of relative enrichment.

The **P-Value** is a significance calculated from the normally distributed enrichments of concepts identified in randomly created lists of entities.

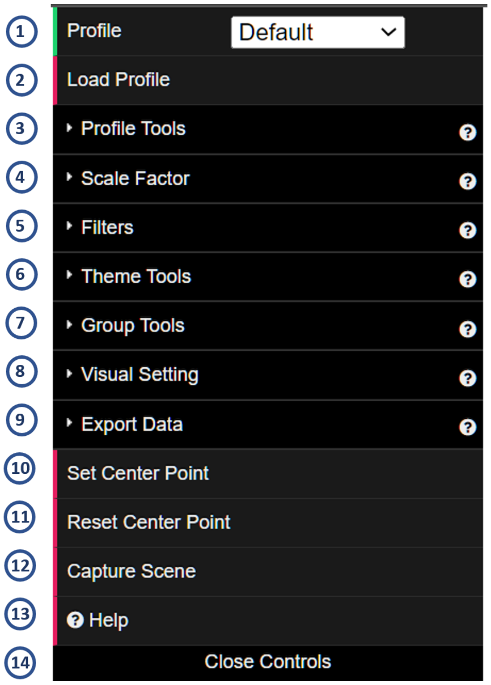
After testing with both random entity lists, positive control pathways, and well-characterized published datasets, the following guidelines are suggested for theme evaluation:

- **NEScore** >= 1.3
- **P-Value** < 0.1

If a theme fails to meet both of the above criteria, it is suggested to view the theme as failing to meet general quality control thresholds.

3.2.4 Visuals and display control

The **Control Panel** located in the upper right hand of the workspace contains the tools for modifying the knowledge map. In the current version of CompBio, there are 14 primary sections in the **Control Panel**, shown below. Primary sections 2-9 are expandable, with subfunctions dropping down accordion-style after the user clicks on the primary section, which is denoted with a small triangle proceeding the primary section title.



A brief description of each of the tabs is provided below.

1. Profile: CompBio makes use of profiles to allow multiple views of the same knowledge map to be generated. This will be described in more detail in the **Profile Tools** section. The **Profile** pull-down box allows for the selection of a previously created profile or allows the user to revert to the original **Default** view.
2. Load Profile: Applies the profile selected.
3. Profile Tools: The profile tools primary section contains a drop-down with a single option. This option allows the user to either create a new profile or to delete the current profile. If a user profile is loaded, another primary section appears on the **Control Panel** allowing the user to **Save Progress**.
4. Scale Factor: This allows the user to modify the repulsive force that the themes apply on each other, modifying overall separation of the graph. This is useful if the knowledge map is tightly packed and greater readability is desired.
5. Filters: Controls for limiting the themes displayed in the knowledge map based on various criteria (described in detail in section 4.1.1.1)
6. Theme Tools: Controls for modifying the name and text in a theme as well as specifying theme-specific edge display (described in detail in section 5.4.4)
7. Group Tools: Controls for defining, naming, coloring, and display of groups of themes (described in detail in section 5.4.5)

8. Visual Setting: Text display tools, background, and various other visual control options (described in detail in section 5.4.6)

9. Export Data: Tools for compiling numerical data that describes the entire knowledge map, or a subset of the map (described in detail in section 5.4.7)

10. Set Center Point: Allows the user to modify the center point around which the knowledge map rotates

11. Reset Center Point: Reverts the center point back to the default location.

12. Capture Scene: (described in detail in section 5.4.7)

13. Help: Hyperlink to this document

14. Close Controls: Hides the **Control Panel**

Section 4 - CompBio Project Comparison

It is often desirable to compare two or more data sets to gain biological understanding of similarities and/or differences in them. Different human disease cohorts, animal models and human disease, different treatments within a model or disease, or time points within a treatment course are just a few examples of the conditions for which these comparisons are often sought. Within the CompBio platform, projects can be compared in multiple ways. Simple comparisons can be done at the concept or entity level of any two projects, and much more sophisticated contextual comparisons can be completed on libraries of projects with the Assertion Engine tool. This section will describe the utility and directions for each type of comparison.

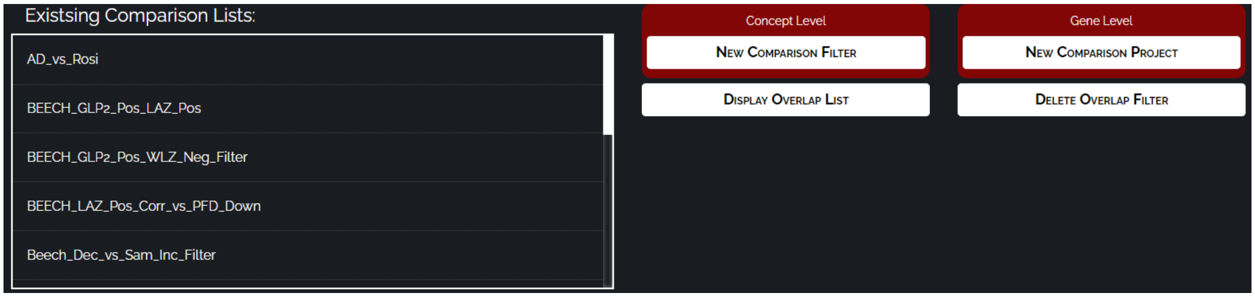
4.1 Basic Project Comparison

4.1.1 Concept comparison

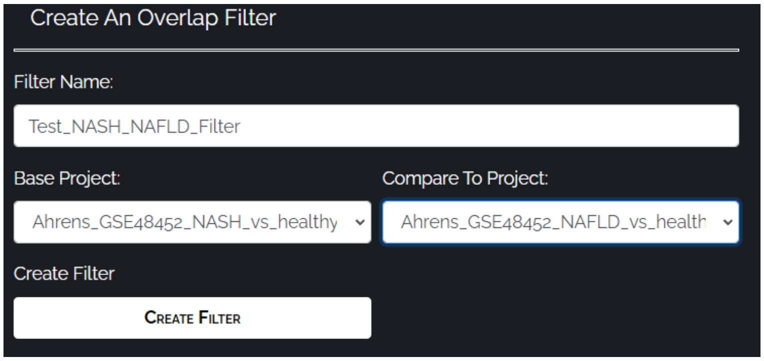
CompBio projects can be compared at the concept level by first creating a concept filter and then applying that filter to a given project. While these filters are often utilized with one or both projects from which they are created, they can be used with any CompBio project(s) within the users account. As these filters are created from a simple list comparison of concepts or entities from the comparator projects, the results should be viewed with care. Even though a given concept may exist within the top themes of two different projects, the context, or other concepts within those themes, may be different across the two projects. Thus, the user is primarily looking for themes in which numerous concepts are shared between the projects. Alternatively, entity level comparison will create a list of shared entities between the two projects and will create a new CompBio project from that shared list. The user can then observe the differences and similarities between the child project and the two parent projects.

4.1.1.1 Filter creation

The first step in project comparison is to create a filter that can be applied. To enable a project comparison, select the **Project & Filter** link under the **Comparison** tools section in the **Account Administration** panel. This will direct the user to the **Compare Projects** panel. To create the project level filter, click the **Concept Level (New Comparison Filter)** button.



The **Overlap Project** section will now be visible in the lower sub-panel. Enter the name of the filter you wish to create in the **Project Name**. You can then use the two drop down lists under **Base Project** and **Compare To Project** to select the two projects for comparison. Once selected, click the **Create Project** button. This will create an intersection of the concept lists that will be saved under the project name provided and can be used as a concept filter for any completed CompBio project.



Create An Overlap Filter

Filter Name: Test_NASH_NAFLD_Filter

Base Project: Ahrens_GSE48452_NASH_vs_healthy

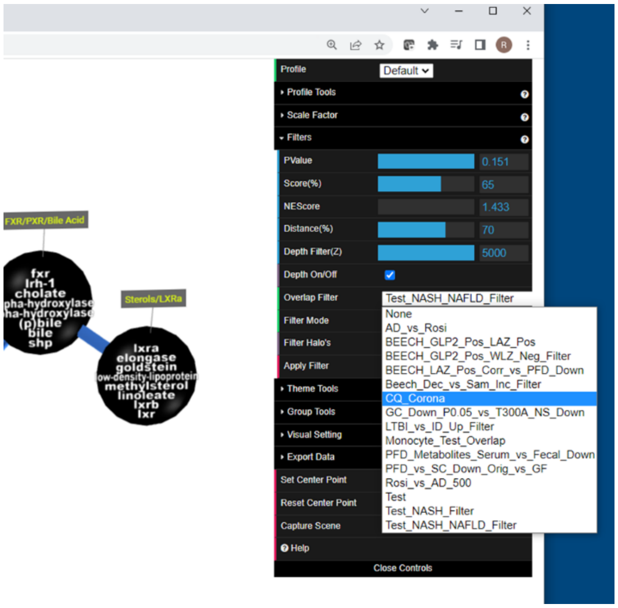
Compare To Project: Ahrens_GSE48452_NAFLD_vs_health

Create Filter

Clicking the **Display Overlap List** will show the complete list of intersecting concepts from the two projects. This is the list that will be applied as filter in the visualizer. Selecting a filter from the **Existing Comparison Lists** will display the basic information about that list including the creation date, the base and overlap project and total concepts in the overlap.

4.1.1.2 Filter application

Concept filters will now be visible in **Filters** tab of the **Controls** panel on the far right of the visualizer screen of any loaded CompBio project. To access the filter, open the Filters tab and click on the **Overlap Filter** white box. This will bring up a display of all filters available to apply to the project. Select the appropriate filter from that list.



Profile: Default

PValue: 0.151

Score(%): 65

NEScore: 1.433

Distance(%): 70

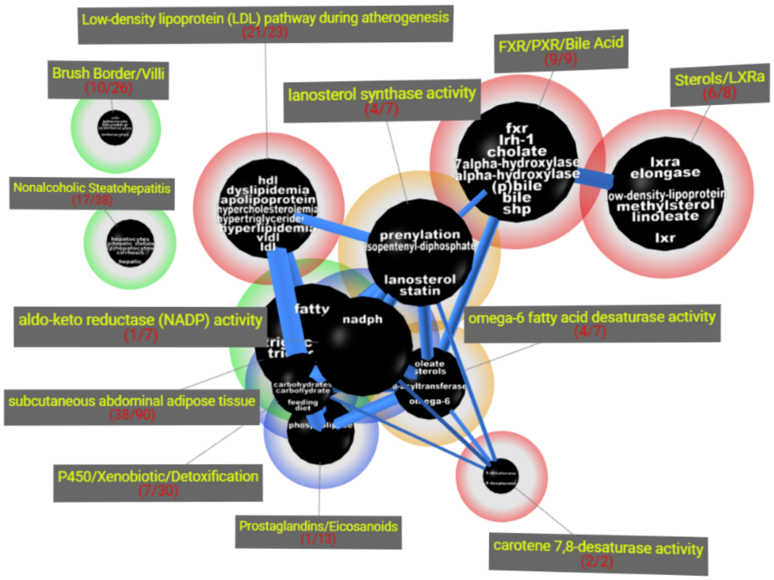
Depth Filter(Z): 5000

Depth On/Off: ☑

Overlap Filter: Test_NASH_NAFLD_Filter

None
AD_vs_Rosi
BEECH_GLP2_Pos_LAZ_Pos
BEECH_GLP2_Pos_WLZ_Neg_Filter
BEECH_LAZ_Pos_Corr_vs_PFD_Down
Beech_Dec_vs_Sam_Inc_Filter
CQ_Corona
GC_Down_P0.05_vs_T300A_NS_Down
LTBI_vs_ID_Up_Filter
Monocyte_Test_Overlap
PFD_Metabolites_Serum_vs_Fecal_Down
PFD_vs_SC_Down_Orig_vs_GF
Rosi_vs_AD_500
Test
Test_NASH_Filter
Test_NASH_NAFLD_Filter

Once the filter is selected, the user will then need to decide in which mode to utilize it. Just below the **Overlap Filter** option is the **Filter Mode** option. *Intersection* mode will display concepts that are shared with the filter list, and *Difference* mode will display concepts that are not shared with the filter list within the visualizer screen. **Note that these two modes are mutually exclusive.** Once the mode is selected, click the **Apply Filter** tab. Themes will now redisplay with the concepts that match the selected filter mode.

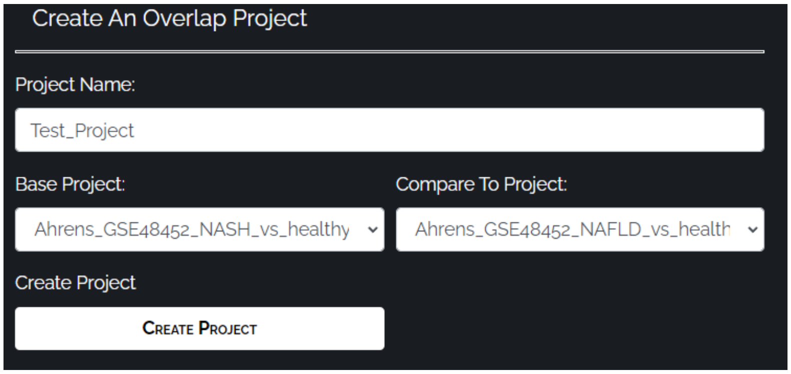


In the example above, Intersection mode was used with the filter list created from a NASH cohort and NAFLD cohort. The concepts still visible within the themes intersect with the filter list applied. The colored halos that are now present around the themes indicate how many of the concepts originally in the theme overlap with the filter list in heat-mapped fashion. Just below the theme annotation label are a pair of numbers. The first indicates the number of concepts in that theme that match with the filter lister and the second is total number of concepts in the theme. As the filter list was created by comparing NASH and NAFLD cohorts, this view indicates that themes associated with FXR/Bile Acid, Sterols/LXR, LDL/Atherogenesis, and Lanosterol Synthase activity are strongly preserved between NASH and NAFLD. Additionally, themes associated with Villi, NASH, and P450/Xenobiotics are modestly preserved, and themes related to Prostaglandins/Eicosanoids and NADP are weakly preserved. It should be noted that if a theme does not have a single preserved concept with the filter, the entire theme will disappear from the view. This was true in the case above as theme related to Flavin-containing MonoOxygenase activity was unique to NASH and no concepts overlapped with NAFLD. If the given filter had been applied in Difference mode, the view would have flipped. All currently displayed concepts would disappear, and the absent concepts would appear. Application of the filter in this manor would present the concepts that are unique to the NASH when compared to NAFLD project.

4.1.2 *Gene comparison and new project creation*

While the New Comparison Filter function will simply compare the two full lists of concepts from the completed projects and allow the user to view that filter applied as described above, the **Gene Level (New Comparison Project)** utility will create an intersection of the entities that were used to create those two projects and will generate a third, unique CompBio project using that intersected list. When comparing biology from different species, such as human and mouse, or from two different technology platforms where the entity lists may not be identical, it will generally be more enlightening to use the Comparison Filter. However, when different conditions are applied to common background, such as different drug treatments or a time course, it may be more desirable to compare at the entity level. Thus, the platform enables both types of comparisons.

As with the concept level comparison, the entity overlap project generation can be found under the **Project & Filter** link under the **Comparison** tools section in the **Account Administration** panel. After clicking the **Gene Level (New Comparison Project)** button, the **Create An Overlap Project** dialogue will appear in the lower sub-panel. Utilization is identical to that of the concept level comparison described in 4.1.1.1.



Create An Overlap Project

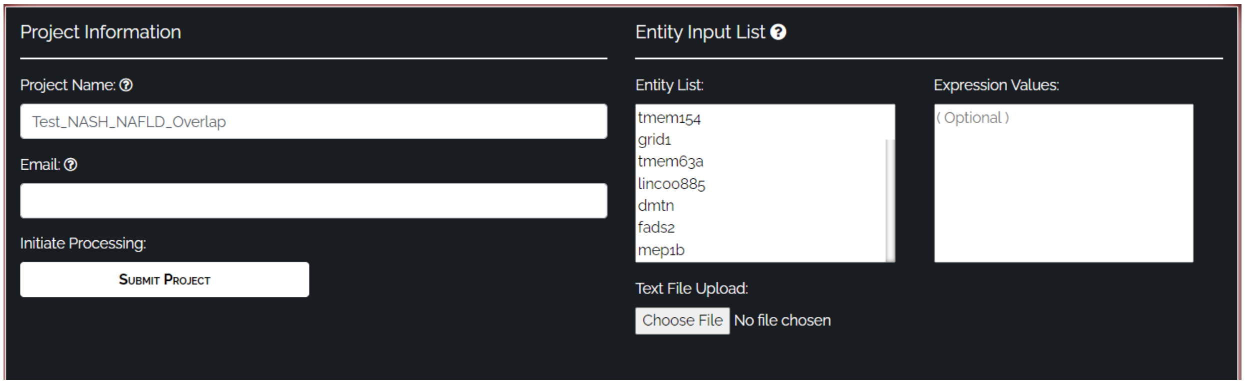
Project Name: Test_Project

Base Project: Ahrens_GSE48452_NASH_vs_healthy

Compare To Project: Ahrens_GSE48452_NAFLD_vs_health

Create Project

Clicking the **Create Project** button will take the user to the CompBio **Project Submission** page with the **Entity List** pre-loaded with the supplied **Project Name** and entity intersection list. To run the project, enter the Email notification address and click the Submit Project button. The user will be notified when the new project is complete.



Project Information

Project Name: Test_NASH_NAFLD_Overlap

Email:

Initiate Processing:

Entity Input List

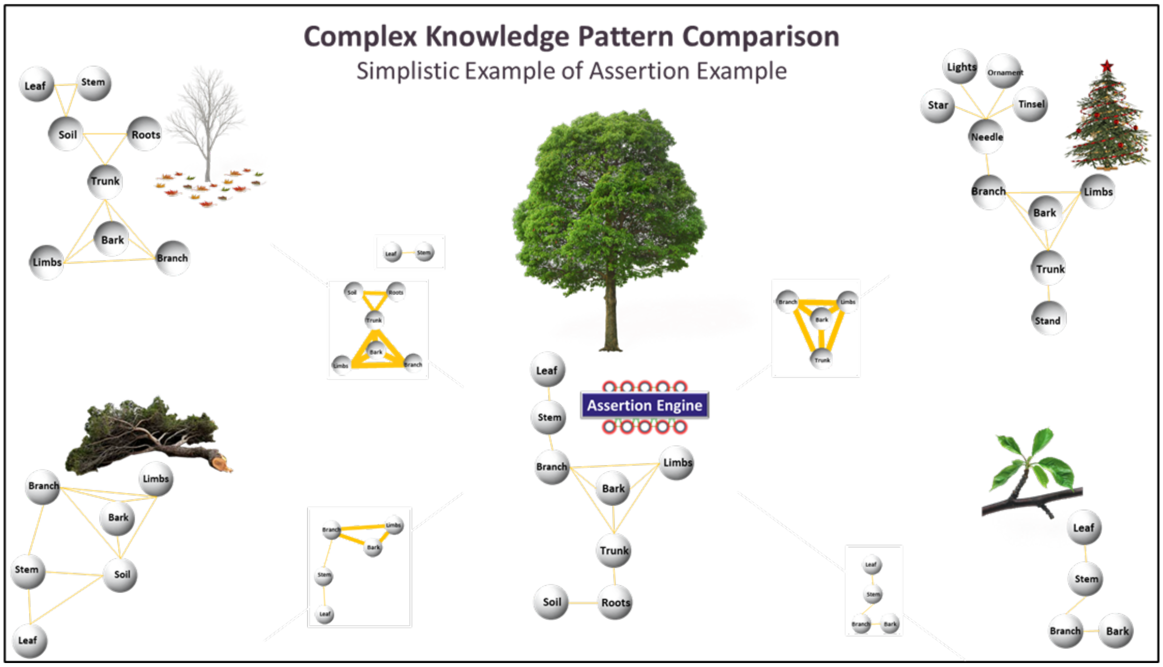
Entity List:
tmem154
grid1
tmem63a
linc00885
dmtn
fads2
mep1b

Expression Values: (Optional)

Text File Upload: No file chosen

4.2 Assertion Engine – ML Based Contextual Comparison

The Assertion Engine program within the CompBio platform enables the user to perform a comparison of projects, in pairs, to assess what biological concepts and processes are conserved between them. Unlike the **Project & Filter** functions that simply create an intersection of identical concepts or entities between two project that can then be applied as filter within the CompBio project visualizer, the Assertion Engine uses a machine learning-based, pattern similarity identification utility to detect patterns of contextually preserved biology. This means that the engine does not simply look to see if a concept exists within both projects, but pattern of connectivity between that concept and other concepts is preserved. This assessment is performed across the entirety of the concept space yielding a global map, with local regions, of preserved biological connectivity. Conceptually, this can be visualized in the example(s) of how similar or dissimilar different representations trees, or components of trees, are in the following figure. In this case, even though components are highly overlapping, the context of some components change. The Assertion Engine can recognize both the similarities and dissimilarities without ever training to understand what a tree or its components are. It merely identifies the level of preservation across components in any comparison and can, statistically, determine if the level preservation is significant or not.



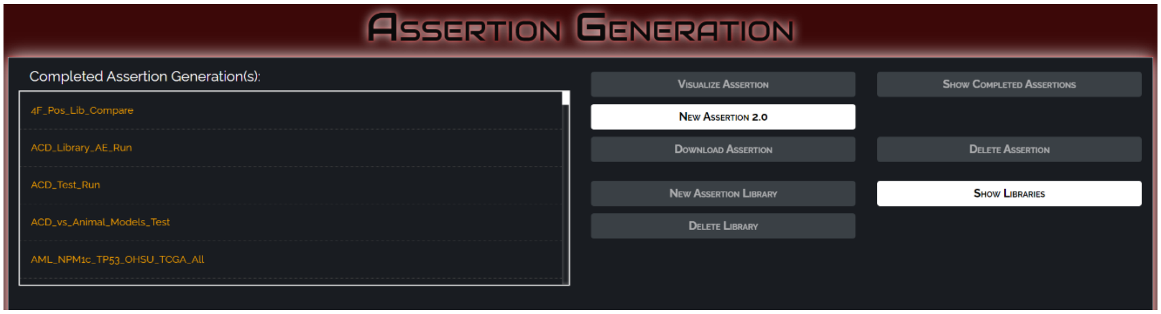
Complex Knowledge Pattern Comparison

Simplistic Example of Assertion Example

From a biological perspective, this ability allows the Assertion Engine to identify preserved biology in a unprecedented fashion across projects. Examples of how the engine has been used to date include comparison of different human disease cohorts with a common diagnosis, comparison of human disease to animal models, comparison of different intervention treatments in human or models, comparison of stages of progression in human disease or models, and time points within a time series. Though the engine can only compare two projects at a time, libraries of project can be created and passed to the engine to compute a full block of comparisons across many conditions or related projects of interest.

4.2.1 *Creating projects and libraries*

Accessing the Assertion Engine is done by clicking the **Assertion Generation** link within the **Comparison** section of the **Project Management** page. The inputs to the Assertion Engine are existing CompBio projects or project libraries. To begin an Assertion Engine run using existing completed projects click the **New Assertion 2.0** button in the upper right section of the **Assertion Generation** page.

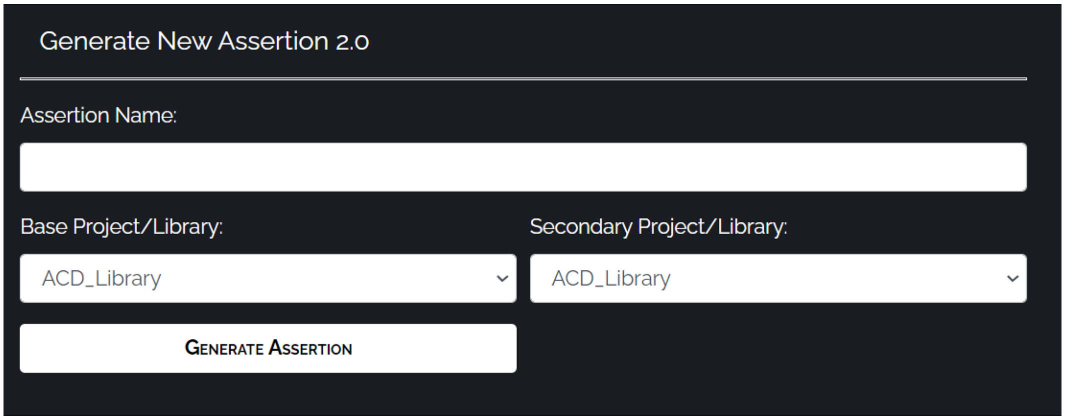


Assertion Generation

Completed Assertion Generation(s):

- 4F_Pos_Lib_Compare
- ACD_Library_AE_Run
- ACD_Test_Run
- ACD_vs_Animal_Models_Test
- AML_NPM1c_TP53_OHSU_TCGA_All

This will open the dialogue setting up the engine run and generating a comparison.



Generate New Assertion 2.0

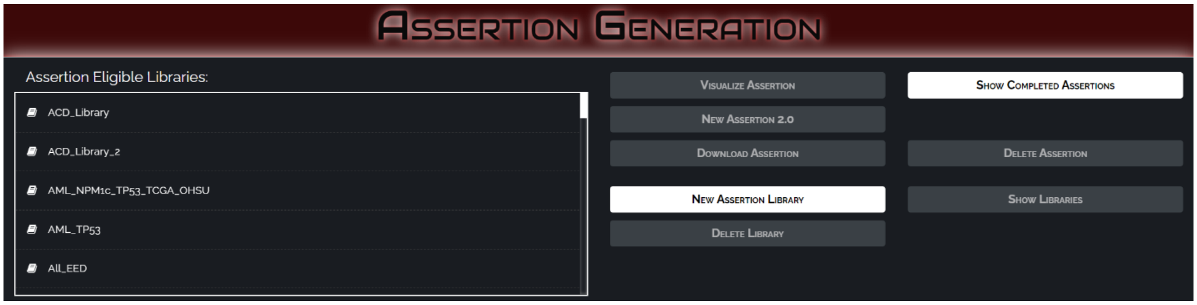
Assertion Name:

Base Project/Library: ACD_Library

Secondary Project/Library: ACD_Library

Select the base and secondary projects to compare using the two drop down lists. These lists will show the names of all completed projects available from within the user's account. Enter the name of the Assertion project. This can be any name that does not conflict with an existing Assertion Engine run name. Finally, click the **Generate Assertion** button. A typical comparison will take a minute or less.

If the user wishes to compare a large number of projects to a project or another group of projects, this can be done by first creating one or more libraries from existing CompBio projects. To create a library, first click the **Show Libraries** button in the upper right section of the **Assertion Generation** page. This will bring up a new list of available functions in the upper right section of the page while showing any existing libraries in the upper left section.

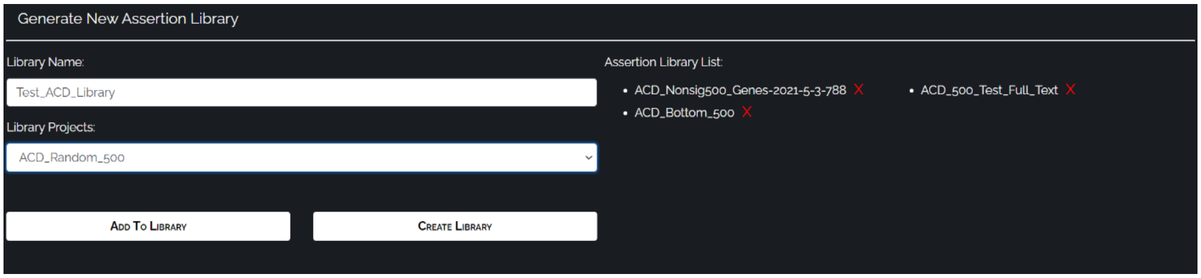


Assertion Generation

Assertion Eligible Libraries:

- ACD_Library
- ACD_Library_2
- AML_NPM1c_TP53_TCGA_OHSU
- AML_TP53
- All_EED

To create a new library, click on the **New Assertion Library** button. This will initiate new dialog in the lower portion of the page.



Generate New Assertion Library

Library Name: Test_ACD_Library

Library Projects: ACD_Random_500

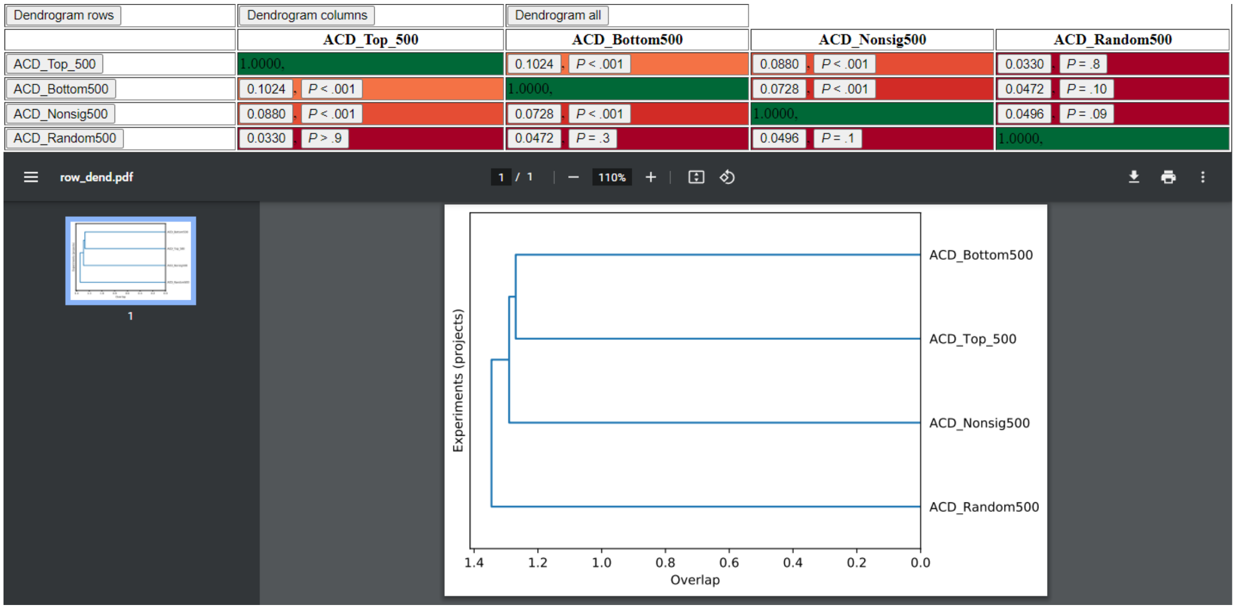
Assertion Library List:

- ACD_Nonsig500_Genes-2021-5-3-788 X
- ACD_Bottom_500 X
- ACD_500_Test_Full_Text X

First, enter the new libraries name in the **Library Name** dialogue box. Then, using the **Library Projects** drop down list, begin to select the project to be added to the library. It should be noted that while there is currently no limit on the number of projects that can be added to a library, once number of projects starts to exceed 8 to 10 projects, the resulting heatmap matrix can become a bit difficulty to read. Therefore, we recommend keeping the number of projects in a library to 10 or less. As the user adds project by selecting them from the drop-down list and then clicking the **Add To Library** button, they will appear in the lower right section under the Assertion Library List header. If the user decides to remove one or more projects, simply click the **red X** next to the projects name before clicking the **Create Library** button. Once the list is complete, click the **Create Library** button to generate the library. It will now appear in the **Assertion Eligible Libraries** list and can be selected in the **Generate New Assertion 2.0** dialogue. While individual project assertion comparisons generally take less than one minute to run, large library comparisons can take several minutes to an hour depending on the total number of assertions that result. Once the assertion run is completed the resulting library or project level assertion will display in the **Completed Assertion Generation(s):** list. Select the desired assertion within that list and click the **Visualize Assertion** button to view the results.

4.2.2 *Global similarity assessment*

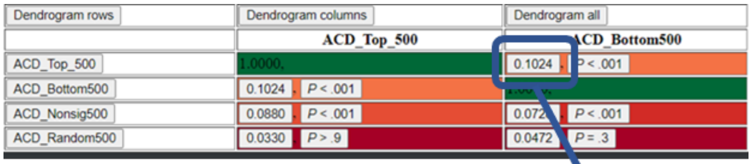
Upon clicking the **Visualize Assertion** button, the user will be taken to the **Global Similarity** comparison page showing project level comparison in a heatmap and dendrogram format. The pseudo heatmap section also displays the overlap value and a p-value that evaluates the significance of that overlap based on the size of the input projects.



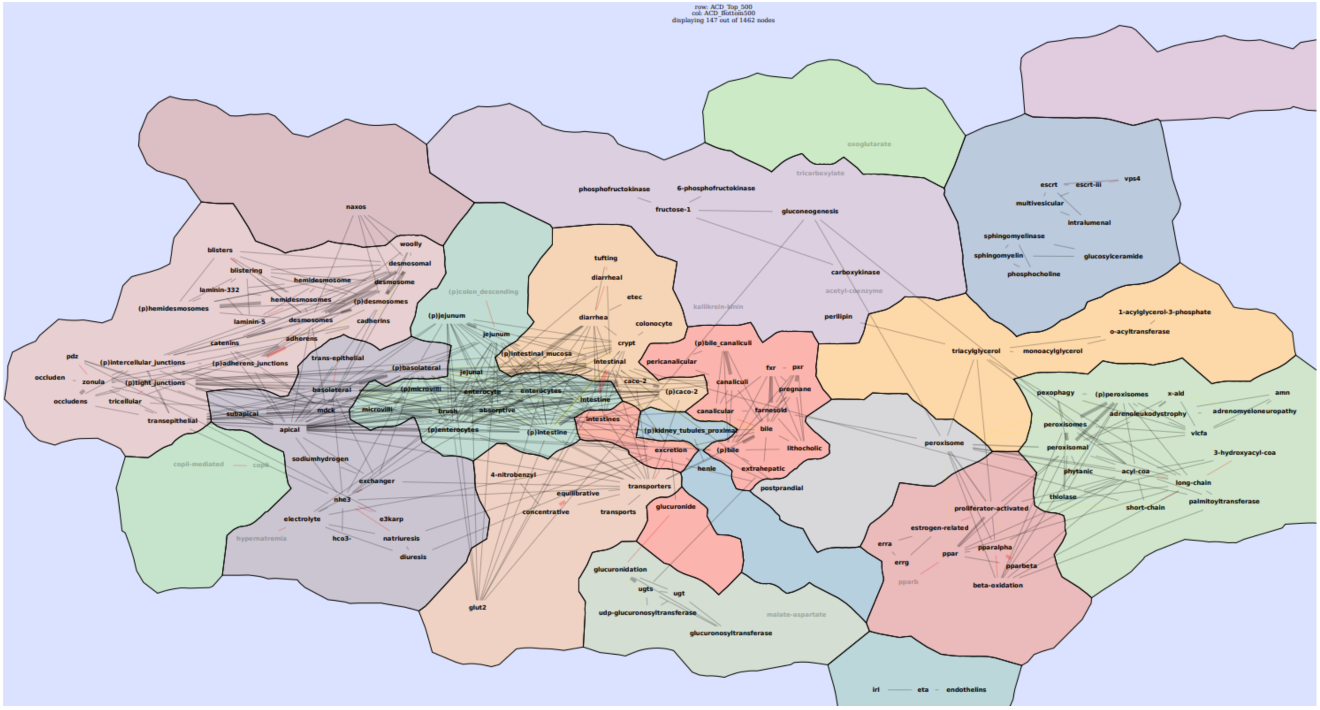
	ACD_Top_500	ACD_Bottom500	ACD_Nonsig500	ACD_Random500
ACD_Top_500	1.0000,	0.1024 , P < .001	0.0880 , P < .001	0.0330 P = .8
ACD_Bottom500	0.1024 , P < .001	1.0000,	0.0728 , P < .001	0.0472 P = .10
ACD_Nonsig500	0.0880 , P < .001	0.0728 , P < .001	1.0000,	0.0496 P = .09
ACD_Random500	0.0330 P > .9	0.0472 P = .3	0.0496 P = .1	1.0000,

The example in the above figure illustrates the comparison between 4 windows of genes selected from a list that was identified as differential (down regulated) between adult Crohn's Disease subjects and non-IBD controls. The Top 500 list contains the 500 genes with p-value < 0.01. As these genes were the most significantly different from the non-IBD controls, they are expected to represent base biological truth for the data set. The Bottom 500 list are genes with 0.01 < p-value < 0.05. While these genes are still nominally significant and would be expected to contain some relevant biology, the level of noise within the list is anticipated to be higher than the Top 500 gene list. The non-significant 500 (Nonsig500) gene list are the next 500 genes with 0.05 < p-value. These genes may contain some signal but likely contain more noise than signal. The Random 500 are genes randomly selected that are not contained with any of the first three lists. Based on these criteria, the first three lists would be expected to contain some degree of similar biology to one another but with increasing levels of noise. The Random 500 list would be expected to have no significant biological overlap with any of the first three. As can be seen in the figure, this is exactly what the Assertion Engine identified.

Each cell in the pseudo-heatmap contains a global overlap or similarity score for the project pair and a p-value indicating the level of significance for that overlap. Clicking on the **Overlap Score** button from the comparison of the top 500 ACD genes to the bottom 500 ACD genes will bring up the biological similarity map for those two projects. The thickness of the lines between concepts is an indicator of the level of contextual preservation between those concepts and what they, in turn, associate with. The colored regions are rough boundaries for sets of concepts that are tightly conserved. These will tend to be concepts associated with a particular process, pathway, cell type or other organized biological construct.



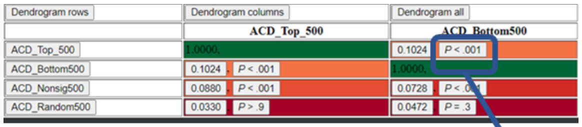
Dendrogram rows	Dendrogram columns	Dendrogram all
	ACD_Top_500	ACD_Bottom500
ACD_Top_500	1.0000,	0.1024 , P < .001
ACD_Bottom500	0.1024 , P < .001	1.0000,
ACD_Nonsig500	0.0880 , P < .001	0.072 , P < .001
ACD_Random500	0.0330 , P > .9	0.0472 , P = .3



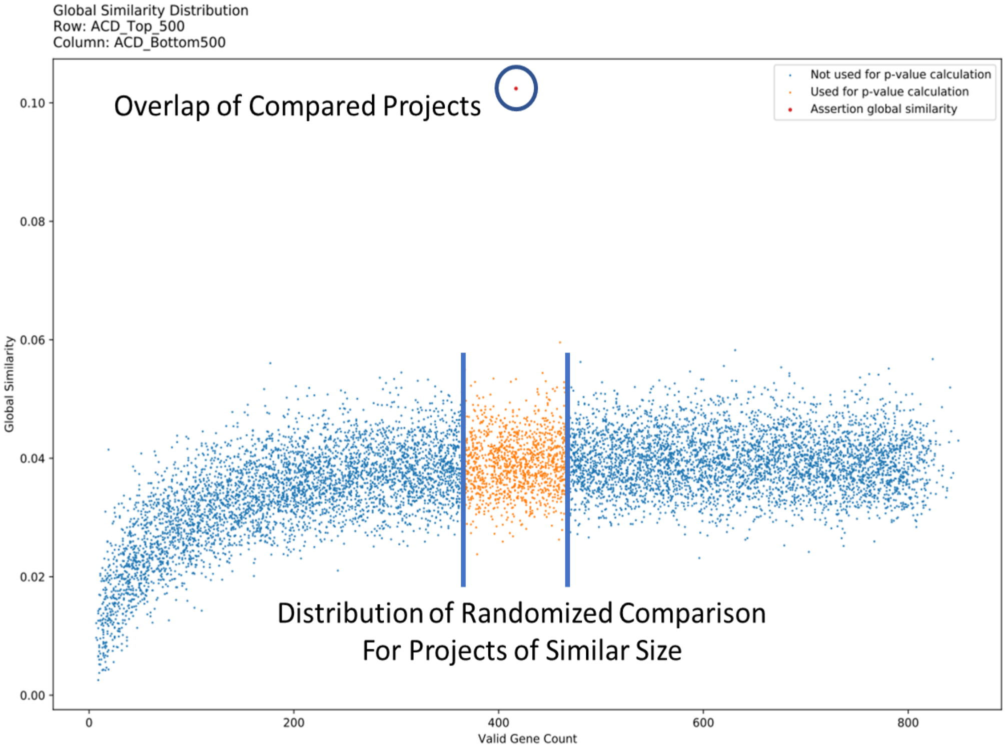
Each row of the pseudo-heatmap is based on the same comparator project. As such, the structure and regions of the map will remain the same as the user clicks the different **Overlap** button within the different cells in that row. This is not true for the columns as different cells in a column change the aspect to a different project.

As, in this case, the library was compared to itself, there is a visible green diagonal where each project was compared to itself. For these cases, the maximum possible overlap value of 1.0 is observed and, as such, no p-value is displayed as it is recognized as a self-comparison.

The **P-value** button also provides a graph to demonstrate just how far from the randomized range a given overlap is. The calculation of the p-value is done empirically by comparing randomized data sets of different sizes to real-world data sets of size of the base project in each case.



Dendrogram rows	Dendrogram columns	Dendrogram all
	ACD_Top_500	**ACD_Bottom500**
ACD_Top_500	1.0000,	0.1024 P < .001
ACD_Bottom500	0.1024 , P < .001	1.0000,
ACD_Nonsig500	0.0880 , P < .001	0.0728 , P < .001
ACD_Random500	0.0330 , P > .9	0.0472 , P = .3



Overlap of Compared Projects

Distribution of Randomized Comparison For Projects of Similar Size

Section 5 – Visualizer Controls and Navigation

5.1 Visualizer navigation

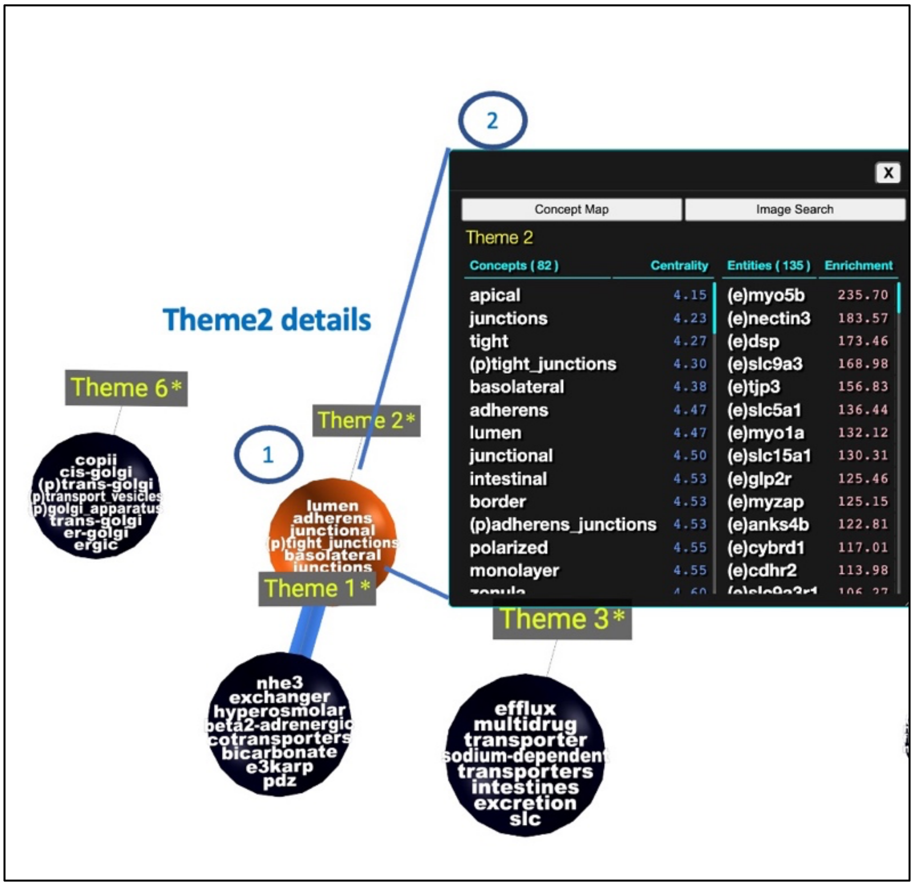
5.1.1 Workspace elements overview

Once the project is complete,the “visualize project” button (see section 2.2.3) will display the CompBio map. The visualization shows 3 panels- Left hand side panel gives the theme details with scores, p-values and concepts. The right-hand side shows the control panel for visual setting options and the middle panel is the actual 3D map generated from the input list. Refer section 3.1 for a detailed description.

5.1.2 Themes, concepts, and entities

The panel on the left-hand side shows the themes ordered by the total enrichment score from theme 1 to theme 50 (default settings is 50 themes. Can be opened up to greater than 50 using the visualizer checkbox (6)-see Visualizer Optimization section 2.2.3.)

Each theme will display the concepts that make up the theme. If the theme is clicked, (Theme2 in the figure below, it shows an orange hue), a separate box will appear in the main map (2) which shows details with concepts and genes that brought those concepts. If expression data or any numeric value is provided as input, those values will be displayed here next to the enrichment scores for each entity.

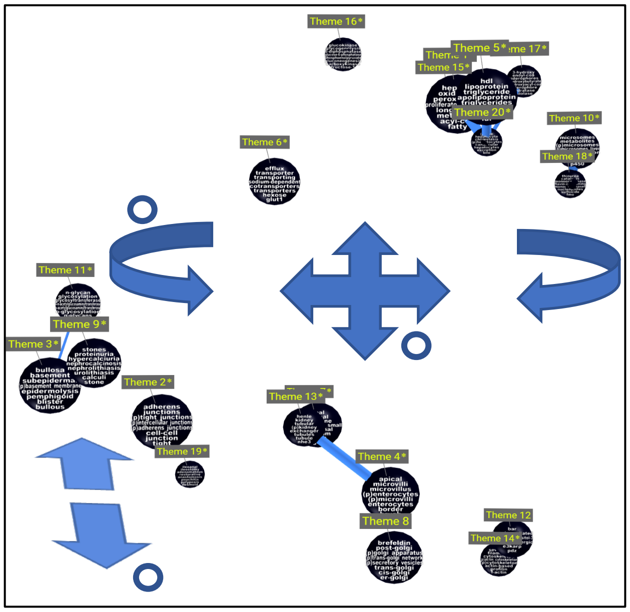


Theme 2

Concepts (82)	Centrality	Entities (135)	Enrichment
apical	4.15	(e)myo5b	235.70
junctions	4.23	(e)nectin3	183.57
tight	4.27	(e)dsp	173.46
(p)tight_junctions	4.30	(e)slc9a3	168.98
basolateral	4.38	(e)tjp3	156.83
adherens	4.47	(e)slc5a1	136.44
lumen	4.47	(e)myo1a	132.12
junctional	4.50	(e)slc15a1	130.31
intestinal	4.53	(e)glp2r	125.46
border	4.53	(e)myzap	125.15
(p)adherens_junctions	4.53	(e)anks4b	122.81
polarized	4.55	(e)cybrd1	117.01
monolayer	4.55	(e)cdhr2	113.98
zonula	4.60	(e)slc9a3r1	106.27

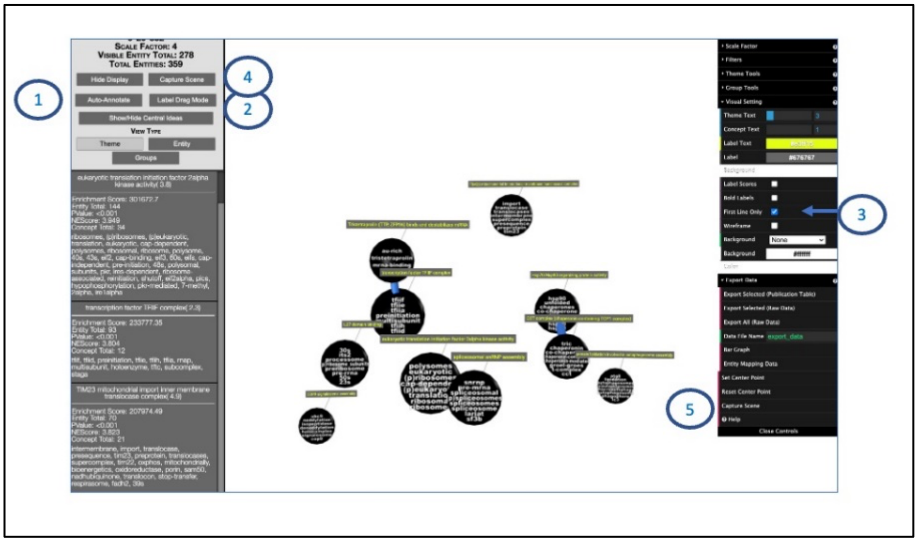
5.1.3 3D controls: pan, zoom, rotate

The central map can be maneuvered with the mouse- pan, zoom and rotation operations can be performed to get the perfect view for image capture. Left mouse click allows you to rotate; right mouse click for pan and the mouse wheel can be used to zoom in and out.



5.1.4 Auto-annotate

One of the recent features within CompBio is the ability to auto-annotate themes. Using the **Auto-Annotate** button (1) themes can be assigned a name based on the conditional probability matrix using the concepts within the theme as well as related themes. Details of the auto-annotation are described on section 3.1.6.



SCALE FACTOR: 4
VISIBLE ENTITY TOTAL: 278
TOTAL ENTITIES: 359

eukaryotic translation initiation factor 2alpha kinase activity(3.8)

Enrichment Score: 301672.7
Entity Total: 144
PValue: <0.001
NEScore: 3.949
Concept Total: 34

ribosomes, (p)ribosomes, (p)eukaryotic, translation, eukaryotic, cap-dependent, polysomes, ribosomal, ribosome, polysome, 40s, 43s, eif2, cap-binding, eif3, 60s, eifs, cap-independent, pre-initiation, 48s, polysomal, subunits, pic, ires-dependent, ribosome-associated, reinitiation, shutoff, eif2alpha, pics, hypophosphorylation, pkr-mediated, 7-methyl, 2alpha, ire1alpha

transcription factor TFIIF complex(2.3)

Enrichment Score: 233777.35
Entity Total: 93
PValue: <0.001
NEScore: 3.804
Concept Total: 12

tfiif, tfiid, preinitiation, tfiie, tfiih, tfiia, rnap, multisubunit, holoenzyme, tfiic, subcomplex, stage

TIM23 mitochondrial import inner membrane translocase complex(4.9)

Enrichment Score: 207974.49
Entity Total: 70
PValue: <0.001
NEScore: 3.823
Concept Total: 21

intermembrane, import, translocase, presequence, tim23, preprotein, translocases, supercomplex, tim22, oxphos, mitochondrially, bioenergetics, oxidoreductase, porin, sam50, nadhubiquinone, translocon, stop-transfer, respirasome, fadh2, 39s

Theme Text 3
Concept Text 1
Label Text
Label #676767
Background
Label Scores □
Bold Labels □
First Line Only ☑
Wireframe □
Background None
Background #ffffff
Color
Data File Name export_data

Many themes can have more than one label assigned. Using the **First Line Only** checkbox (3) only the top scoring label can be displayed. The labels can be moved and aligned for best display using the **Label Drag** button (2).

5.1.5 Capture scene

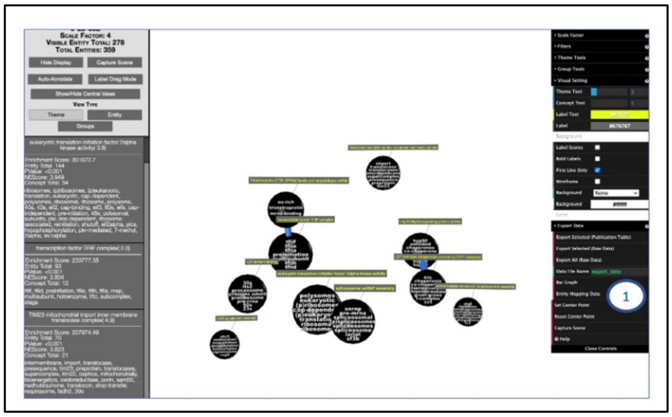
Once the themes are annotated and the final visual display of the map is created, the image can be saved using the **Capture Scene** button (4,5). The button is available within the control panel. Once the button is clicked a separate window with the map and a **Download Button** (1) will appear. A .svg image file will be saved on your desktop than can be edited using any graphic software such as Adobe illustrator.

5.1.6 Label drag mode

The labels assigned by auto-annotation can be moved around and aligned for best display using the **Label Drag** button (2).

5.1.7 Export unmapped entity data

Once the input list is submitted, the user can find out valid and mapped entities using the **Entity Mapping Data** button (1)



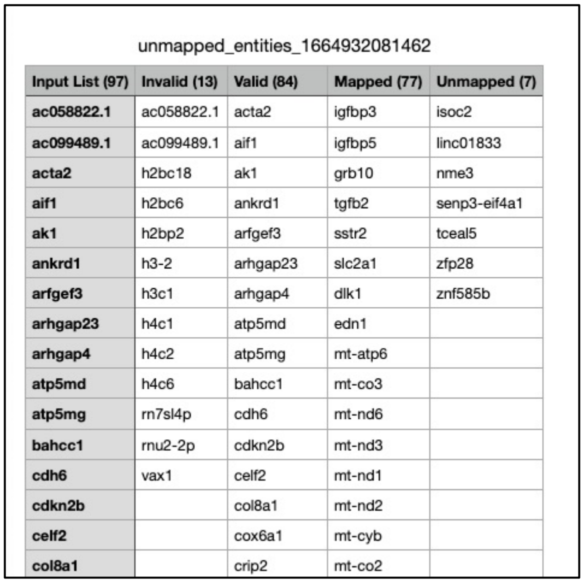
Scale Factor: 4
Visible Entity Total: 278
Total Entities: 359

eukaryotic translation initiation factor 2alpha kinase activity(3.8)

Enrichment Score: 301672.7
Entity Total: 144
PValue: <0.001
NEScore: 3.949
Concept Total: 34

Theme Text
Concept Text
Label Text
Label #676767
Background None
Background #ffffff
Data File Name export_data

Once this button is clicked, a.csv file will be downloaded on the computer showing the details of the input list as shown. The list will have 5 columns. First is the input list. Second column shows invalid entities- those entities that did not have a core memory in the model, i.e. those genes that are not recognized. The third column shows valid entities- those that the tool recognizes and can use in the analysis. The fourth column shows the mapped entities- those that are valid and meet a threshold minimum 3 abstracts. the last column shows the unmapped entities- those that are valid but do not meet the analysis threshold.



unmapped_entities_1664932081462

Input List (97)	Invalid (13)	Valid (84)	Mapped (77)	Unmapped (7)
ac058822.1	ac058822.1	acta2	igfbp3	isoc2
ac099489.1	ac099489.1	aif1	igfbp5	linc01833
acta2	h2bc18	ak1	grb10	nme3
aif1	h2bc6	ankrd1	tgfb2	senp3-eif4a1
ak1	h2bp2	arfgef3	sstr2	tceal5
ankrd1	h3-2	arhgap23	slc2a1	zfp28
arfgef3	h3c1	arhgap4	dlk1	znf585b
arhgap23	h4c1	atp5md	edn1	
arhgap4	h4c2	atp5mg	mt-atp6	
atp5md	h4c6	bahcc1	mt-co3	
atp5mg	rn7sl4p	cdh6	mt-nd6	
bahcc1	rnu2-2p	cdkn2b	mt-nd3	
cdh6	vax1	celf2	mt-nd1	
cdkn2b		col8a1	mt-nd2	
celf2		cox6a1	mt-cyb	
col8a1		crip2	mt-co2	

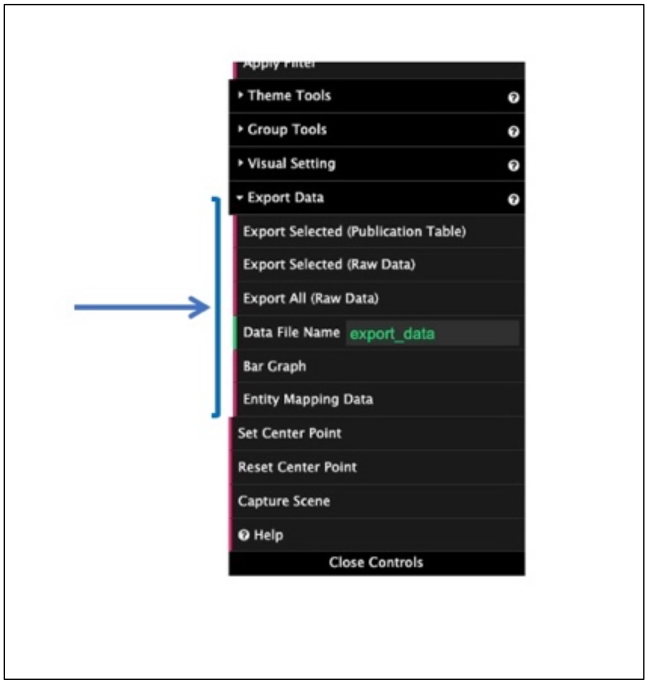
5.1.8 Filter creation - See section 4.1.1

5.1.9 Filter application - See section 4.1.2

5.2 Assertion Engine Figures - See section 4.2

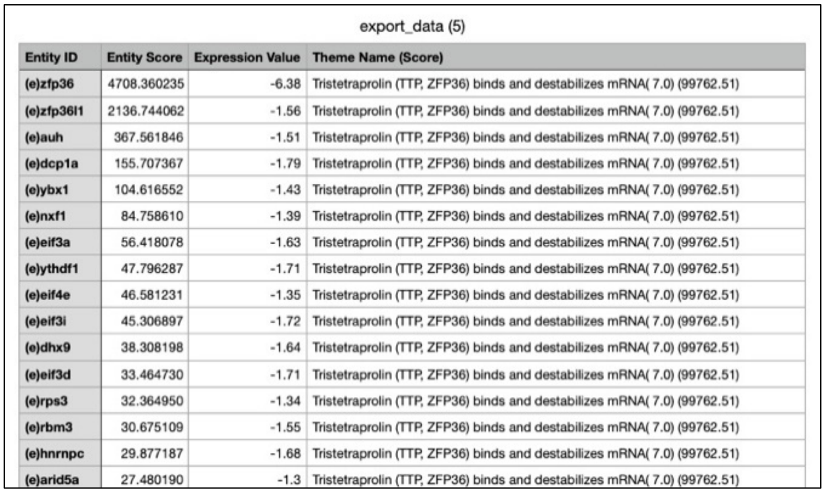
5.3 Data exports for Third Party Tools

The CompBio tool allows the user to export theme mapping data that be easily formatted into supporting material for manuscripts. These export tables can be downloaded from the export data pull down in the control panel.



5.3.1 Export selected (publication table)

Generate a csv file with the filename given in the **Data file name** box. This file will have entities and their scores for all the themes that are selected from the map.

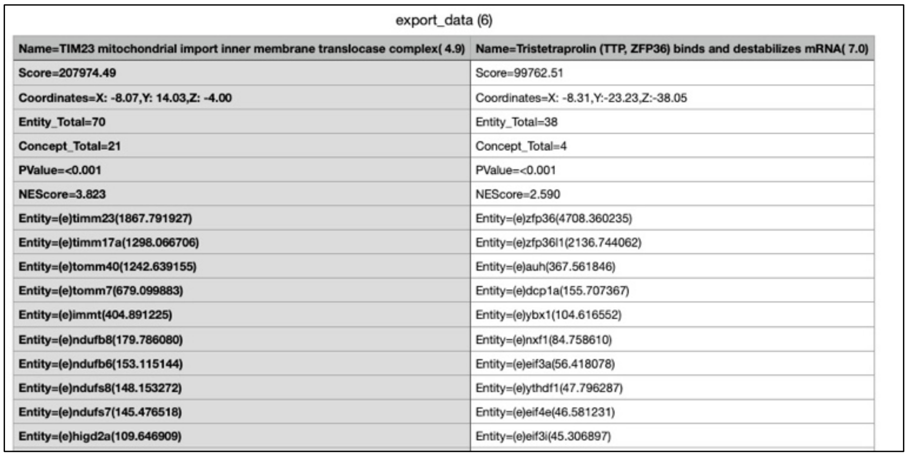


export_data (5)

Entity ID	Entity Score	Expression Value	Theme Name (Score)
(e)zfp36	4708.360235	-6.38	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)zfp36l1	2136.744062	-1.56	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)auh	367.561846	-1.51	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)dcp1a	155.707367	-1.79	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)ybx1	104.616552	-1.43	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)nxf1	84.758610	-1.39	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)eif3a	56.418078	-1.63	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)ythdf1	47.796287	-1.71	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)eif4e	46.581231	-1.35	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)eif3i	45.306897	-1.72	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)dhx9	38.308198	-1.64	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)eif3d	33.464730	-1.71	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)rps3	32.364950	-1.34	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)rbm3	30.675109	-1.55	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)hnrnpc	29.877187	-1.68	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)
(e)arid5a	27.480190	-1.3	Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0) (99762.51)

5.3.2 Export selected (raw data)

Exports theme details for all the themes that are selected within the map. Each column corresponds to a theme. For each theme the column will show all the metrics associated with that theme. The raw and normalized enrichment scores, p-value, total counts for entities and concepts and individual entity and concept that make up that theme.



export_data (6)

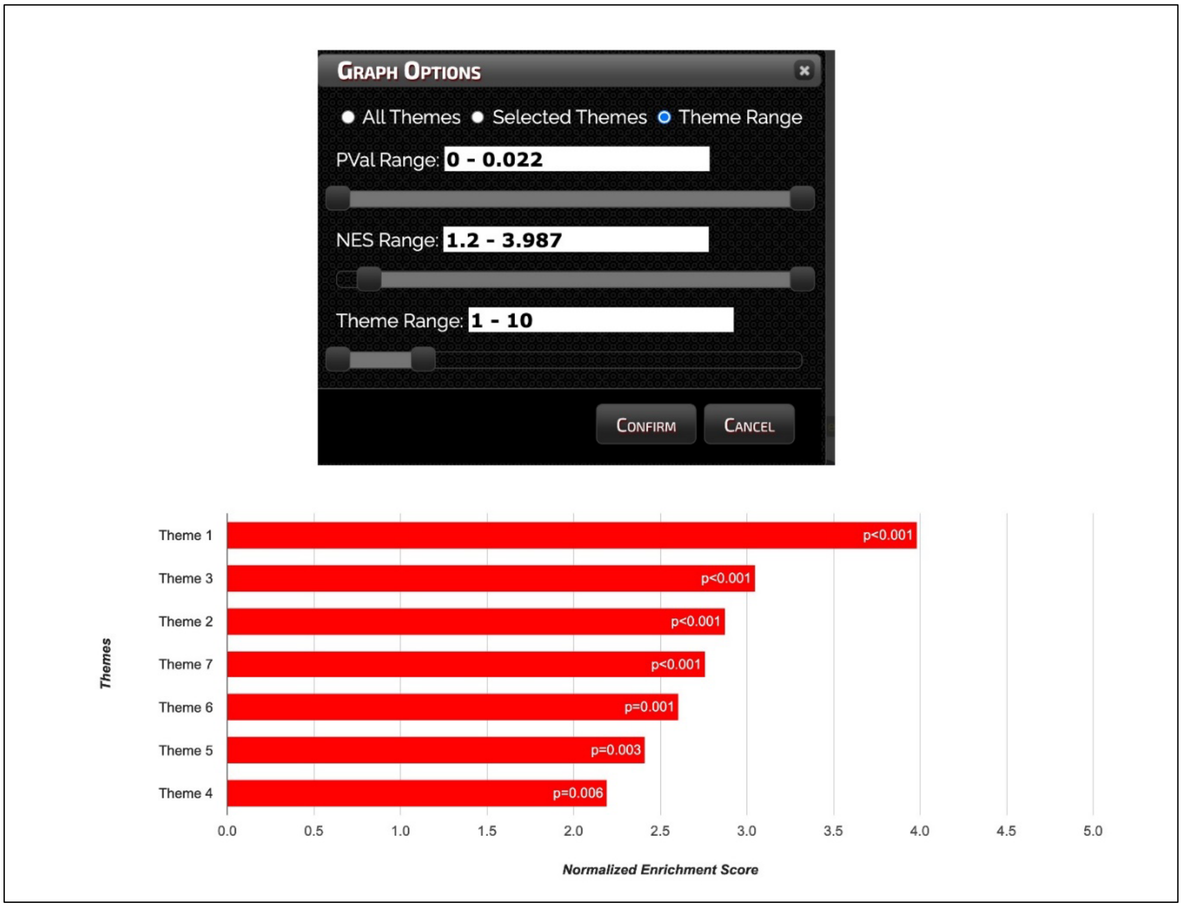
Name=TIM23 mitochondrial import inner membrane translocase complex(4.9)	Name=Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA(7.0)
Score=207974.49	Score=99762.51
Coordinates=X: -8.07,Y: 14.03,Z: -4.00	Coordinates=X: -8.31,Y:-23.23,Z:-38.05
Entity_Total=70	Entity_Total=38
Concept_Total=21	Concept_Total=4
PValue=<0.001	PValue=<0.001
NEScore=3.823	NEScore=2.590
Entity=(e)timm23(1867.791927)	Entity=(e)zfp36(4708.360235)
Entity=(e)timm17a(1298.066706)	Entity=(e)zfp36l1(2136.744062)
Entity=(e)tomm40(1242.639155)	Entity=(e)auh(367.561846)
Entity=(e)tomm7(679.099883)	Entity=(e)dcp1a(155.707367)
Entity=(e)immt(404.891225)	Entity=(e)ybx1(104.616552)
Entity=(e)ndufb8(179.786080)	Entity=(e)nxf1(84.758610)
Entity=(e)ndufb6(153.115144)	Entity=(e)eif3a(56.418078)
Entity=(e)ndufs8(148.153272)	Entity=(e)ythdf1(47.796287)
Entity=(e)ndufs7(145.476518)	Entity=(e)eif4e(46.581231)
Entity=(e)higd2a(109.646909)	Entity=(e)eif3i(45.306897)

Export all raw will have the same format except that it will have 50 columns, one for each of the 50 themes. All these tables can be used to import into R or other data processing tools and used as supplementary material for manuscripts.

5.3.3 Export bar graph

A bar graph can be created using the **Bar graph** option under the **Create Figure** selection from the control panel.

There are several options presented to the user. Either a bar graph of all themes can be created or a selected few. The themes can be selected from the map just by clicking the theme spheres. The other way is to give a range of p-values or NEScores to plot all these passing those cutoffs. An image as shown in the panel will appear with red bars showing the NEScores on the x-axis and the theme labels on the y axis. The p-values will also be displayed. The svg file containing the graph will be downloaded on the desktop. These can be incorporated into any manuscript figures.



Graph Options

○ All Themes ○ Selected Themes ● Theme Range

PVal Range: 0 - 0.022

NES Range: 1.2 - 3.987

Theme Range: 1 - 10

5.4 Visualizer Fine Controls

While many of the concepts being discussed in this section have been touched on previously, section 5.6 will delve into greater detail around use of the tools contained within the **Control Panel**. A figure detailing Control Panel components and a brief description of their functions is shown in section 3.2.4 of this manual.

5.4.1 Creating, naming, and saving a user project profile

When a project is initially visualized, the **Default Profile** is displayed that contains up to 50 themes, unless the Optimization Mode option is deselected on the **Project Management** page, at which point hundreds of themes could potentially be displayed. The **Default Profile** represents the starting point for your project, the raw, clustered output from CompBio. As the annotation process, described in section 3, proceeds, it is necessary to save the progress made. This is accomplished in CompBio through the use of profiles. Having multiple profiles allows multiple users to work on the same knowledge map and to create multiple views of the same project. To create a new profile, click the **Profile Tools** tab in the **Control Panel.** A drop-down tab opens with the **Create New** option. Three new drop-down tabs appear. In the **New Name** tab enter a suitable name for the new project profile and click **Set New Name**. A popup in the middle of the workspace will appear saying "Progress Saved", indicating that the profile has been successfully saved. This profile is now associated with the project and in the future can be selected from the

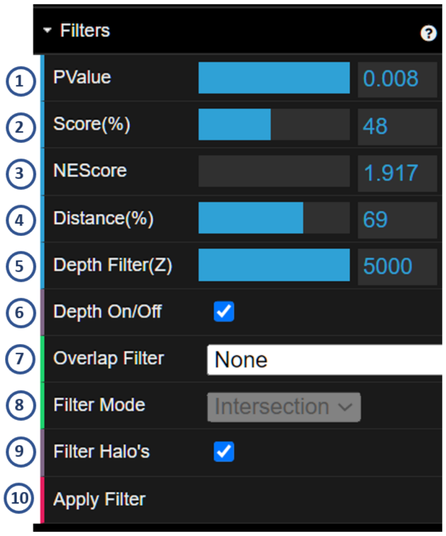
Profile pull-down and clicking **Load Profile**. Similarly, a profile can be deleted using the **Delete Profile**.

5.4.2 *Visualizer scaling*

The **Scale Factor** tab allows the user to modify the repulsive force that the themes apply on each other, modifying overall separation of the graph. This is useful if the knowledge map is tightly packed and greater readability is desired. To increase the separation of the graph, enter a value greater than the default value of 4, and click **Initiate Scaling**.

5.4.3 *Visualizer filters*

Filters that adjust the themes displayed are located in the **Filters** tab in the **Control Panel.** Clicking the **Filters** tab will open the filters as shown below. Tabs 7 – 10 are filters for project comparison and are described in detail in sections 5.1.1-2.



▾ Filters ?

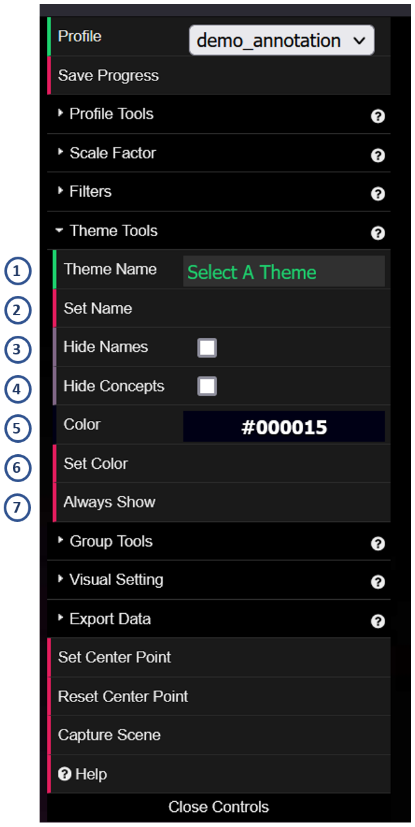
1 PValue 0.008
2 Score(%) 48
3 NEScore 1.917
4 Distance(%) 69
5 Depth Filter(Z) 5000
6 Depth On/Off ☑
7 Overlap Filter None
8 Filter Mode Intersection
9 Filter Halo's ☑
10 Apply Filter

1. The **P-Value** filter limits the themes displayed based on p-value, which was described in section 3.2.1. Dragging the slider to the left increases the stringency, requiring lower p-values for a theme to remain on the screen. Any changes to any of the filters is automatically applied to the project, but it is recommended to click **Save Progress** under frequently to ensure changes made to the knowledge map are save under the current profile.
2. The **Score(%)** filter limits the themes displayed based on their enrichment values. Dragging the slider to the left increases the stringency, requiring higher enrichment scores to remain on-screen.
3. The **NEScore** filter limits the themes displayed based on their normalized enrichment score. Dragging the slider to the left increases the stringency, requiring higher normalized enrichment scores to remain on-screen.
4. The **Distance (%)** filter limits the edges displayed based on their intra-theme distance in three-dimensional space. Dragging the slider to the left increases the stringency, requiring smaller distances between themes for an edge to be shown.
5. The **Depth (Z)** filter limits the themes displayed based on their position in three-dimensional space. Themes farther from the viewer will be removed from view as the slider is dragged to the left.
6. The **Depth On/Off** check box toggles the **Depth (Z)** filter on and off.

7. The **Overlap Filter** pull-down box allows for the selection of previously created comparison filters to limit the themes displayed based on concept overlap. See section 5.1.1-2 for details.
8. The **Filter Mode** pull-down box allows for the selection of **Intersection** or **Difference** options showing themes that overlap with concepts from the overlap filter or are unique form concepts in the filter, respectively.
9. The **Filter Halos** check box toggles theme halos that highlight the degree of overlapping concepts between the current project and the filter. The color scale is, from most relative overlap to least: red, orange, green, and blue.
10. The **Apply Filter** button must be pressed to apply overlap filter pull-down options.

5.4.4 Theme tools

During the annotation process, theme attributes are modified using controls in the **Theme Tools** tab in the **Control Panel.** Below is a figure marking the seven controls that reside within this tab.

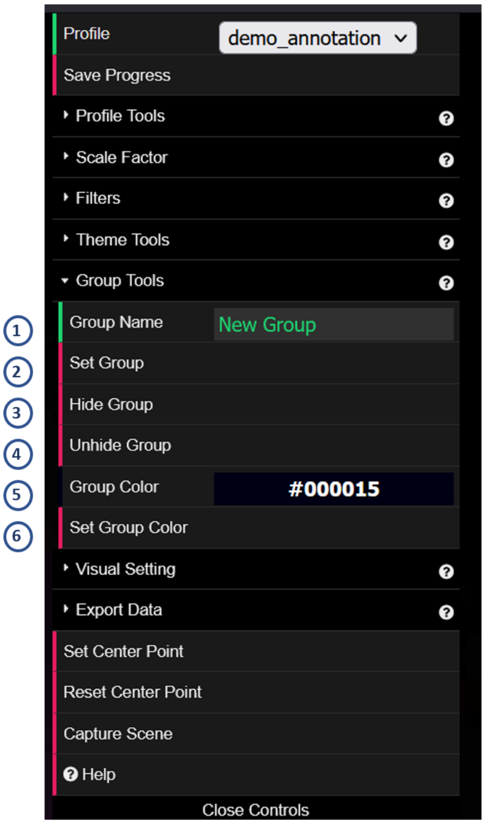


1. The **Theme Name** text box allows the user to specify a custom name for the selected theme.
2. The **Set Name** button applies the specified **Theme Name**. Any changes to the **Theme Name** will only be saved to the profile after clicking **Save Progress**.
3. **Hide Names** toggles the theme names displayed on the knowledge map on and off.
4. **Hide Concepts** toggles the theme concepts displayed on the knowledge map on and off.
5. The **Color** text box allows the user to either specify the theme color in hex color code or to pick the color using the mouse from a color palette that pop up on mouse hover.
6. The **Set Color** button applies the color chosen in the **Color** text box. Any changes to the theme color will only be saved to the profile after clicking **Save Progress**.

7. The **Always Show** button causes an edge to be drawn between two (and only two) selected themes, so long as they share at least one common gene, regardless of the maximum distance threshold specified with the **Distance (%)** filter (see section 5.6.3).

5.4.5 *Theme group creation and naming*

During the annotation process themes are bundled into similarly colored groups of themes with similar biological function using the controls in the **Group Tools** tab in the **Control Panel.** Below is a figure marking the six controls that reside within this tab.



Profile: demo_annotation

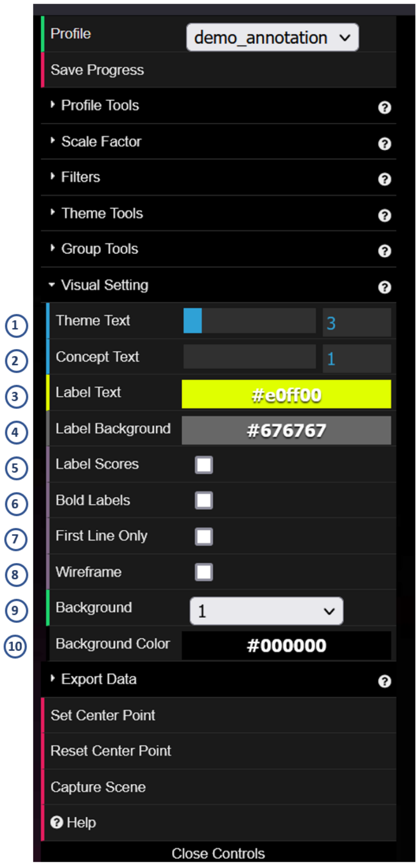
Group Name: New Group

Group Color: #000015

1. The **Group Name** text box allows the user to specify a custom group name for the selected themes.
2. The **Set Group** button applies the specified **Group Name**. Any changes to the **Group Name** will only be saved to the profile after clicking **Save Progress**.
3. **Hide Group** hides the currently selected group on the knowledge map.
4. **Hide Concepts** toggles the theme concepts displayed on the knowledge map on and off.
5. The **Color** text box allows the user to either specify the theme color in hex color code or to pick the color using the mouse from a color palette that pop up on mouse hover.
6. The **Set Color** button applies the color chosen in the **Color** text box. Any changes to the theme color will only be saved to the profile after clicking **Save Progress**.

5.4.6 *General visual settings*

The **Visual Settings** tab contains tools to modify various visual attributes of the knowledge map not contained in the themes or groups tabs. Below is a figure marking the ten controls that reside within this tab.



Profile demo_annotation

Theme Text 3

Concept Text 1

Label Text #e0ff00

Label Background #676767

Label Scores □

Bold Labels □

First Line Only □

Wireframe □

Background 1

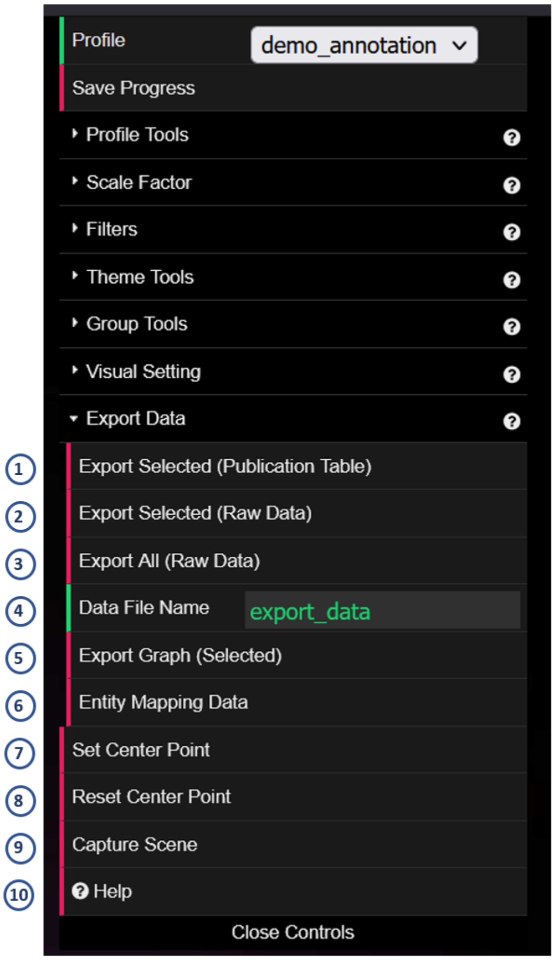
Background Color #000000

1. The **Theme Text** slider allows the user to specify a custom text size for the themes in the knowledge map.
2. The **Concept Text** slider allows the user to specify a custom text size for the concepts in the themes in the knowledge map.
3. **Label Text** allows the user to specify a custom label text color for the theme labels in the knowledge map.
4. **Label Background** allows the user to specify a custom label background color for the theme labels in the knowledge map.
5. The **Label Score** checkbox allows the user to toggle the numerical scores shown in the theme labels on and off.
6. The **Bold Labels** checkbox applies bold text formatting to the theme labels in the knowledge map.
7. The **First Line Only** checkbox restricts the auto-annotation theme labels, which can range from zero to three lines (section 3.1.6) to only a single line of annotation.
8. **Wireframe** checkbox converts the textured themes to wireframe rendering. This can be useful if the effect is desired for aesthetic reasons, or to improve performance on low-performing computer hardware.
9. The **Background** pulldown allows the user to select one of several preset backgrounds for the knowledge map.

10. The **Background Color** text box allows the user to either specify the background color in hex color code or to pick color the using the mouse from a color palette that pop up on mouse hover.

5.4.7 *Data exports*

The **Export Data** section of the **Control Panel** contains tools to export numerical and graphical data from the knowledge map for further analysis. Below is a figure marking the ten controls that reside within this tab.



Profile demo_annotation

Data File Name export_data

1. The **Export Selected (Publication Table)** button allows the user to export theme-level data detailing the entities and enrichment scores for selected themes in the knowledge map.
2. The **Export Selected (Raw Data)** button allows the user to export theme-level data detailing the entities, concepts, and enrichment scores in a less formatted configuration for selected themes in the knowledge map.
3. The **Export All (Raw Data)** button allows the user to export theme-level data detailing the entities and enrichment scores for all themes in the knowledge map.
4. The **Data File Name** text box allows the user to specify the exported file name.
5. The **Export Graph** button allows the user to export a bar graph of the selected theme enrichment scores.

6. The **Entity Mapping Data** button allows the user to export the original entity input list, as well as invalid genes (genes that had insufficient literature to contribute to concept identification in the knowledge map), mapped genes, and unmapped genes.
7. The **Set Center Point** sets the current visual center of the knowledge map as the point of rotation for the graph.
8. **Reset Center Point** button reverts the point of rotation of the knowledge map back to the original settings.
9. The **Capture Scene** button allows the user to capture the knowledge map in a new window as an SVG image with tools to move and edit the graph to suit the user's needs. Exportable as a publication-ready SVG image.
10. The **Help** button is a link to this document.

5.5 Publication Examples

Below are several examples of how published or submitted manuscripts have incorporated CompBio results to convey findings in publication-quality figures.

1. Simple table detailing the major CompBio findings.

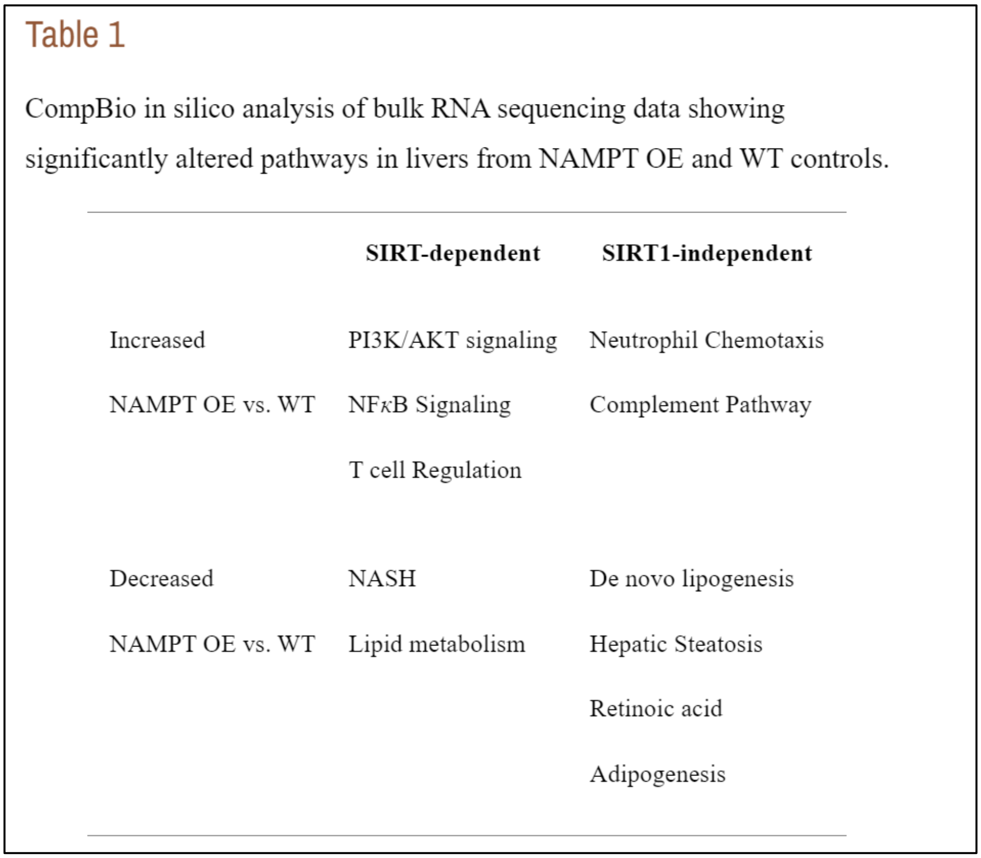


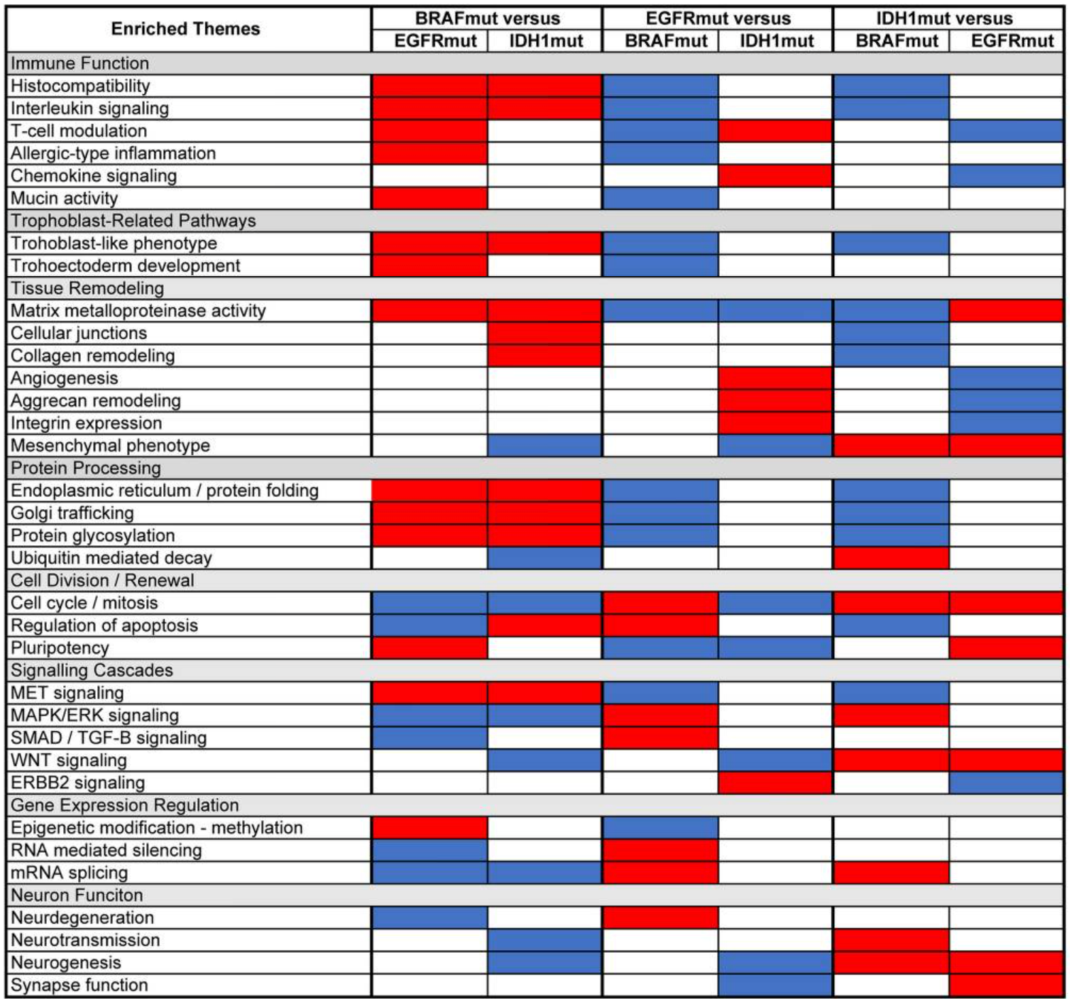
Table 1

CompBio in silico analysis of bulk RNA sequencing data showing significantly altered pathways in livers from NAMPT OE and WT controls.

	SIRT-dependent	**SIRT1-independent**
Increased	PI3K/AKT signaling	Neutrophil Chemotaxis
NAMPT OE vs. WT	NFκB Signaling	Complement Pathway
	T cell Regulation	
Decreased	NASH	De novo lipogenesis
NAMPT OE vs. WT	Lipid metabolism	Hepatic Steatosis
		Retinoic acid
		Adipogenesis

Nat Commun. 2022; 13: 1074

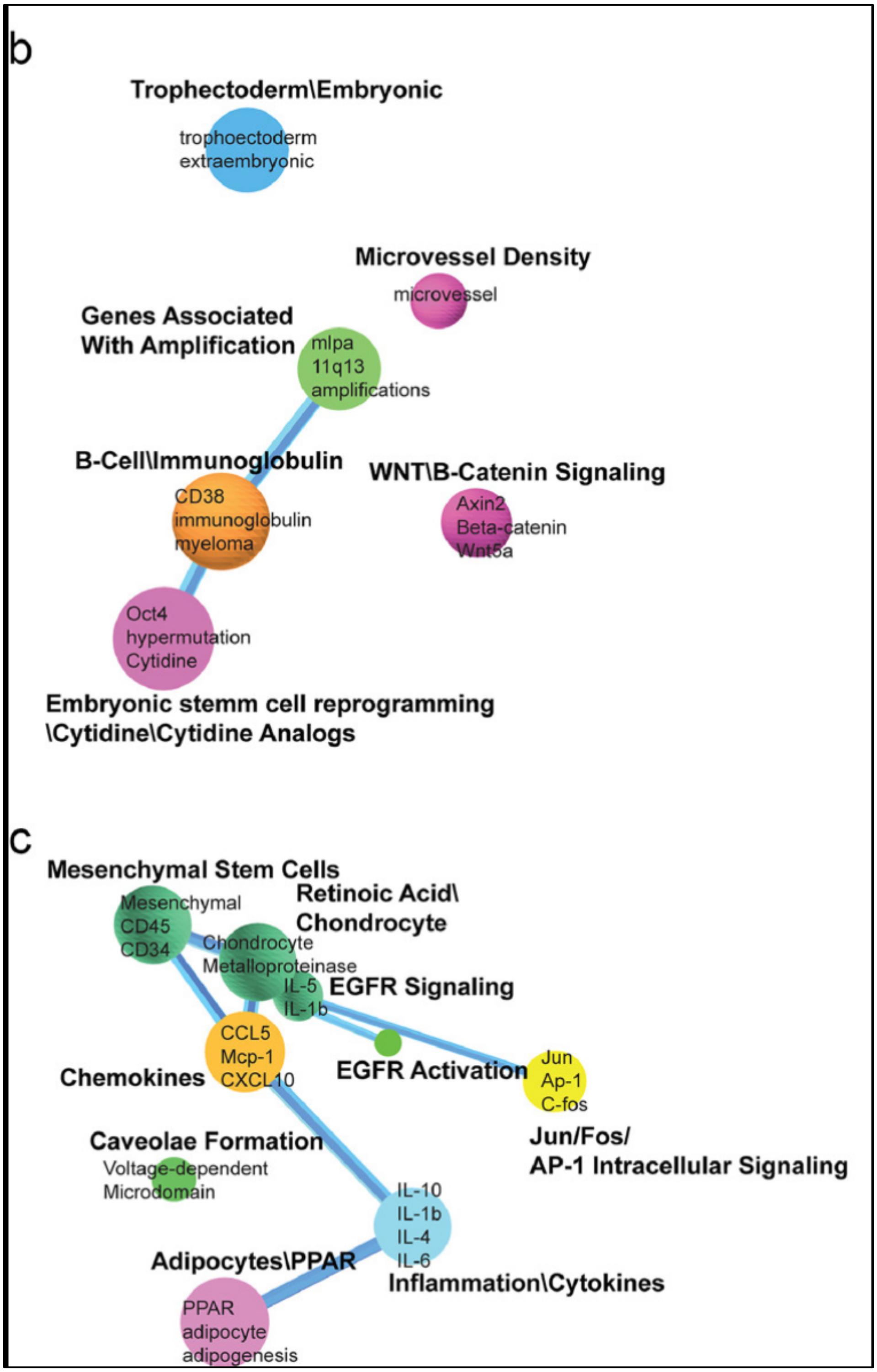
2. A heatmap-style summary of themes across experiments.



Enriched Themes	BRAFmut versus		EGFRmut versus		IDH1mut versus	
	EGFRmut	IDH1mut	BRAFmut	IDH1mut	BRAFmut	EGFRmut
Immune Function						
Histocompatibility	🟥	🟥	🟦		🟦	
Interleukin signaling	🟥	🟥	🟦		🟦	
T-cell modulation	🟥		🟦	🟥		🟦
Allergic-type inflammation	🟥		🟦			
Chemokine signaling				🟥		🟦
Mucin activity	🟥		🟦			
Trophoblast-Related Pathways						
Trohoblast-like phenotype	🟥	🟥	🟦		🟦	
Trohoectoderm development	🟥		🟦			
Tissue Remodeling						
Matrix metalloproteinase activity	🟥	🟥	🟦	🟦	🟦	🟥
Cellular junctions		🟥			🟦	
Collagen remodeling		🟥			🟦	
Angiogenesis				🟥		🟦
Aggrecan remodeling				🟥		🟦
Integrin expression				🟥		🟦
Mesenchymal phenotype		🟦		🟦	🟥	🟥
Protein Processing						
Endoplasmic reticulum / protein folding	🟥	🟥	🟦		🟦	
Golgi trafficking	🟥	🟥	🟦		🟦	
Protein glycosylation	🟥	🟥	🟦		🟦	
Ubiquitin mediated decay		🟦			🟥	
Cell Division / Renewal						
Cell cycle / mitosis	🟦	🟦	🟥	🟦	🟥	🟥
Regulation of apoptosis	🟦	🟥	🟥		🟦	
Pluripotency	🟥		🟦	🟦		🟥
Signalling Cascades						
MET signaling	🟥	🟥	🟦		🟦	
MAPK/ERK signaling	🟦	🟦	🟥		🟥	
SMAD / TGF-B signaling	🟦		🟥			
WNT signaling		🟦		🟦	🟥	🟥
ERBB2 signaling				🟥		🟦
Gene Expression Regulation						
Epigenetic modification - methylation	🟥		🟦			
RNA mediated silencing	🟦		🟥			
mRNA splicing	🟦	🟦	🟥		🟥	
Neuron Funciton						
Neurdegeneration	🟦		🟥			
Neurotransmission		🟦			🟥	
Neurogenesis		🟦		🟦	🟥	🟥
Synapse function				🟦		🟥

Sci Rep, 2021 Oct 8;11

3. Drilling down from themes (in wireframe) down to genes and concepts.



EBioMedicine, 2021 May;67:103347

4. Detailed pathway analysis and knowledge maps

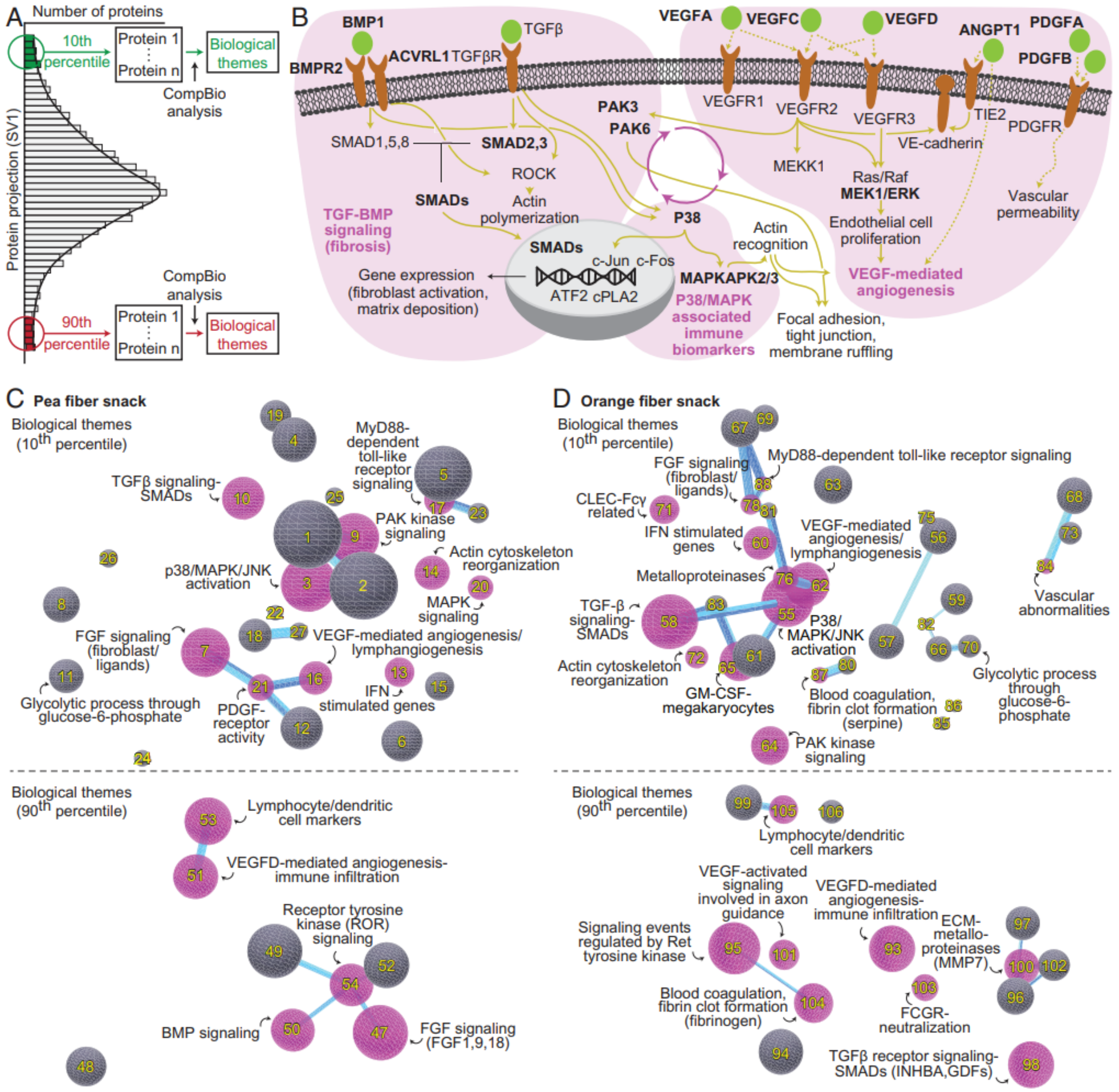
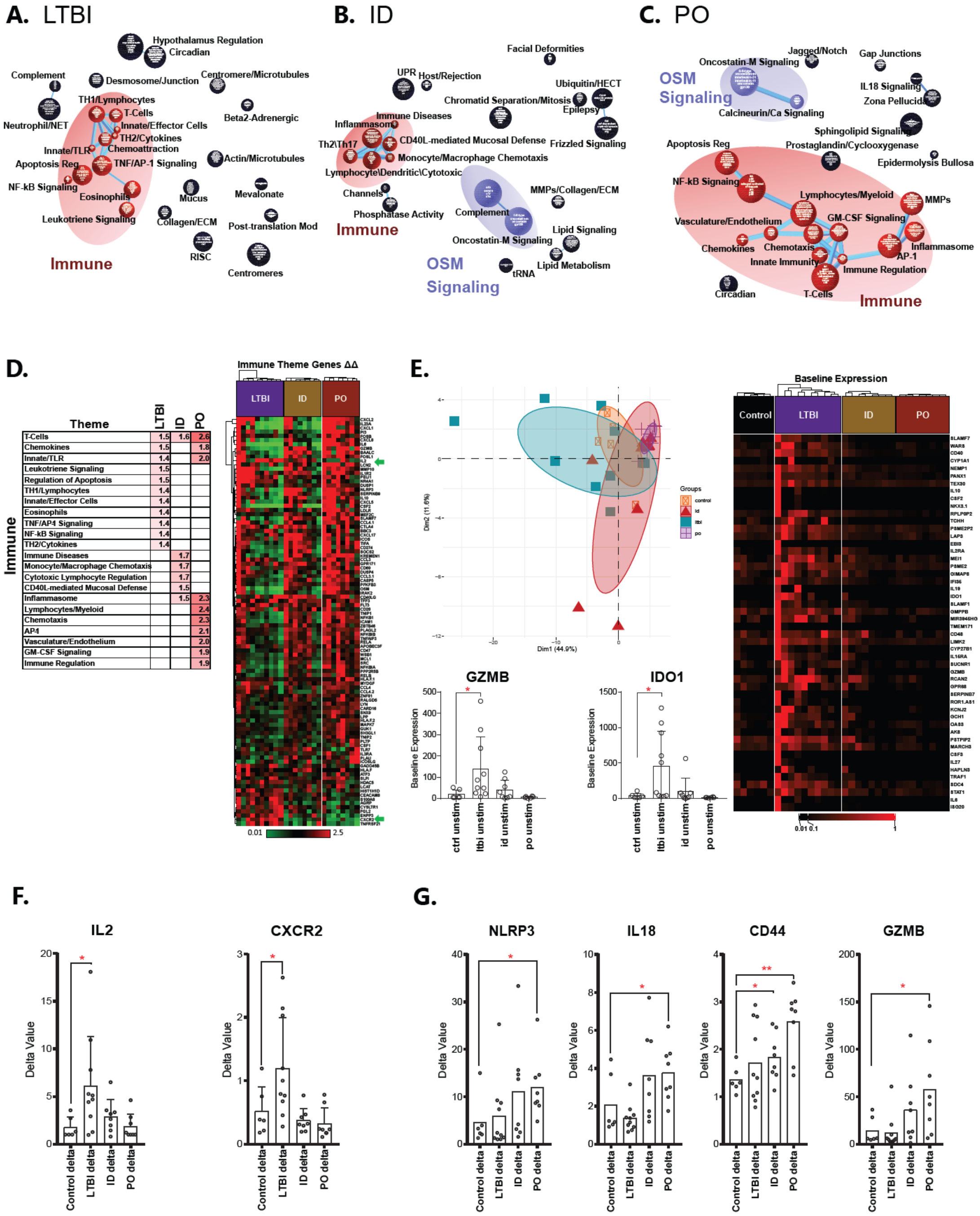

A
Number of proteins
10th percentile
Protein 1
Protein n
Biological themes
CompBio analysis
Protein projection (SV1)
CompBio analysis
90th percentile
Protein 1
Protein n
Biological themes
B
BMP1
BMPR2
ACVRL1
TGFβR
TGFβ
VEGFA
VEGFC
VEGFD
ANGPT1
PDGFA
PDGFB
VEGFR1
VEGFR2
VEGFR3
TIE2
PDGFR
VE-cadherin
PAK3
PAK6
SMAD1,5,8
SMAD2,3
ROCK
Actin polymerization
SMADs
TGF-BMP signaling (fibrosis)
MEKK1
Ras/Raf
MEK1/ERK
Endothelial cell proliferation
Vascular permeability
VEGF-mediated angiogenesis
P38
Actin recognition
SMADs
c-Jun c-Fos
ATF2 cPLA2
MAPKAPK2/3
P38/MAPK associated immune biomarkers
Gene expression (fibroblast activation, matrix deposition)
Focal adhesion, tight junction, membrane ruffling
C Pea fiber snack
Biological themes (10th percentile)
MyD88-dependent toll-like receptor signaling
TGFβ signaling-SMADs
PAK kinase signaling
Actin cytoskeleton reorganization
p38/MAPK/JNK activation
MAPK signaling
FGF signaling (fibroblast/ ligands)
VEGF-mediated angiogenesis/ lymphangiogenesis
IFN stimulated genes
Glycolytic process through glucose-6-phosphate
PDGF-receptor activity
19 4 5 25 17 23 10 1 9 2 3 14 20 26 8 22 18 27 7 21 16 13 15 11 12 6 24
Biological themes (90th percentile)
Lymphocyte/dendritic cell markers
VEGFD-mediated angiogenesis-immune infiltration
Receptor tyrosine kinase (ROR) signaling
BMP signaling
FGF signaling (FGF1,9,18)
53 51 49 54 52 50 47 48
D Orange fiber snack
Biological themes (10th percentile)
FGF signaling (fibroblast/ ligands)
MyD88-dependent toll-like receptor signaling
CLEC-Fcγ related
IFN stimulated genes
VEGF-mediated angiogenesis/ lymphangiogenesis
Metalloproteinases
TGF-β signaling-SMADs
P38/ MAPK/JNK activation
Actin cytoskeleton reorganization
GM-CSF-megakaryocytes
Blood coagulation, fibrin clot formation (serpine)
PAK kinase signaling
Vascular abnormalities
Glycolytic process through glucose-6-phosphate
67 69 88 63 78 81 71 60 75 56 68 73 84 76 62 83 58 55 59 82 57 66 70 72 65 61 87 80 86 85 64
Biological themes (90th percentile)
Lymphocyte/dendritic cell markers
VEGF-activated signaling involved in axon guidance
VEGFD-mediated angiogenesis-immune infiltration
Signaling events regulated by Ret tyrosine kinase
ECM-metallo-proteinases (MMP7)
Blood coagulation, fibrin clot formation (fibrinogen)
FCGR-neutralization
TGFβ receptor signaling-SMADs (INHBA,GDFs)
99 105 106 95 101 93 97 100 102 96 103 104 94 98

Fig. 3. Plasma proteomic biological themes correlated with changes in CAZyme gene abundances after consumption of the fiber-snack prototypes. (*A*) Schematic illustrating the distribution of plasma protein projections along singular vector 1 (SV1) from a CC-SVD analysis of the abundance of CAZyme genes versus changes in levels of plasma proteins after fiber snack consumption. Proteins with SV1 projections along the tails (α = 0.1) of the distribution are highlighted (green and red); these proteins, belonging to the 10th and 90th percentile groups, were analyzed using CompBio to identify biological themes enriched in each of the two percentile groups for each treatment. (*B*) Diagram illustrating pathways in which microbiome CAZyme-correlated plasma proteins (boldfaced) are involved in three interrelated biological themes (TGF-β/BMP-mediated fibrosis, P38/MAPK-associated immune biomarkers, and VEGF-mediated angiogenesis). (*C*, *D*) Network of biological themes identified from the CC-SVD and CompBio analysis. Themes are depicted as spheres; the number in each sphere corresponds to the theme enriched after pea or orange fiber treatment that is listed in *SI Appendix*, Dataset S6*D*. The size of a sphere is proportional to the CompBio enrichment score of its theme. The thickness of the lines connecting themes is proportional to the number of proteins shared between them. Purple spheres represent themes related to TGF-β-BMP signaling (fibrosis), p38/MAPK immune biomarkers, and VEGF-mediated angiogenesis that were enriched in the plasma proteomes of pea and orange fiber study participants (see Table 1).

Proc Natl Acad Sci, 2022 May 17;119(20)

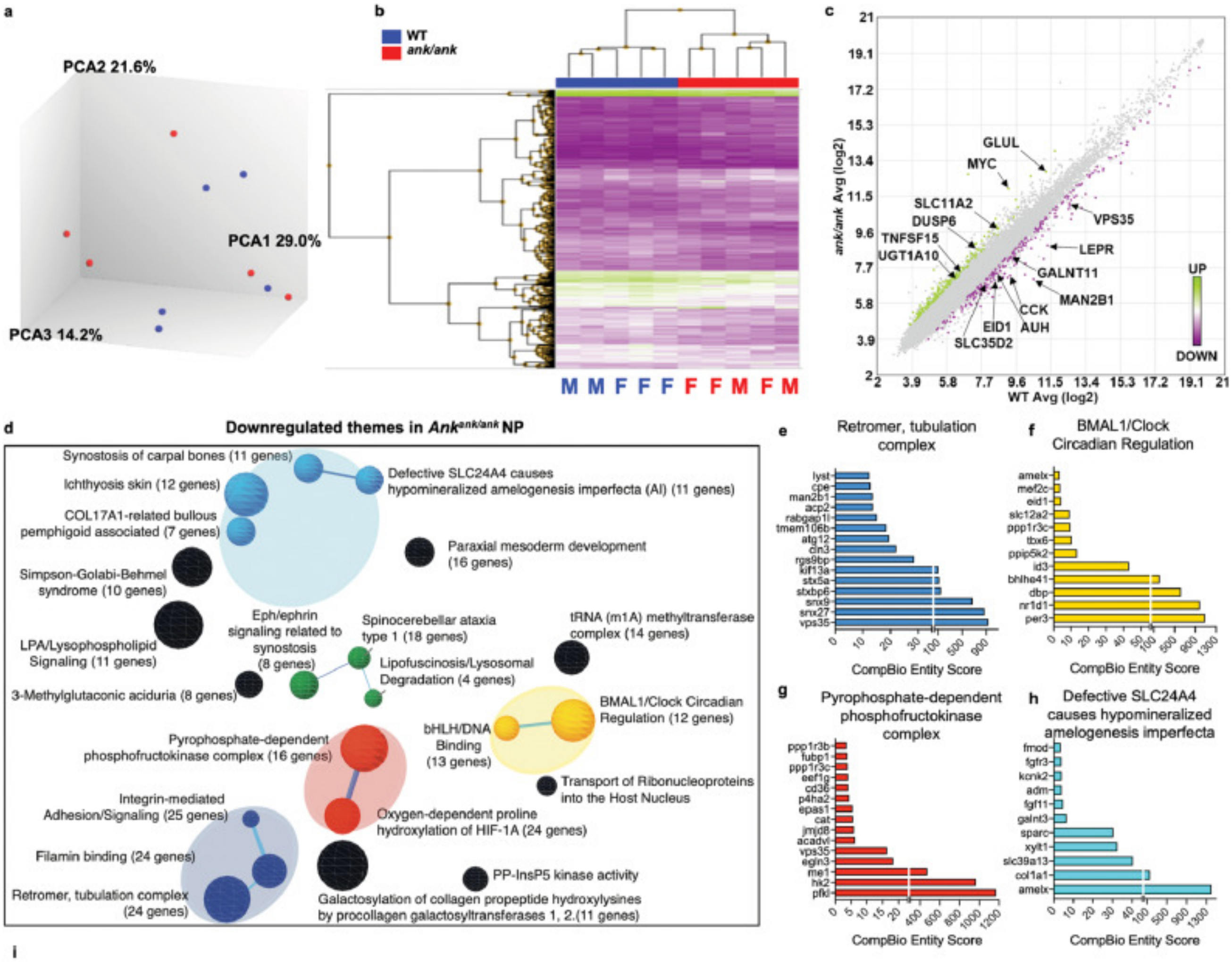
5. Drilling down from knowledge maps to themes across experiments, to expression of individual genes from themes.

Figure 3

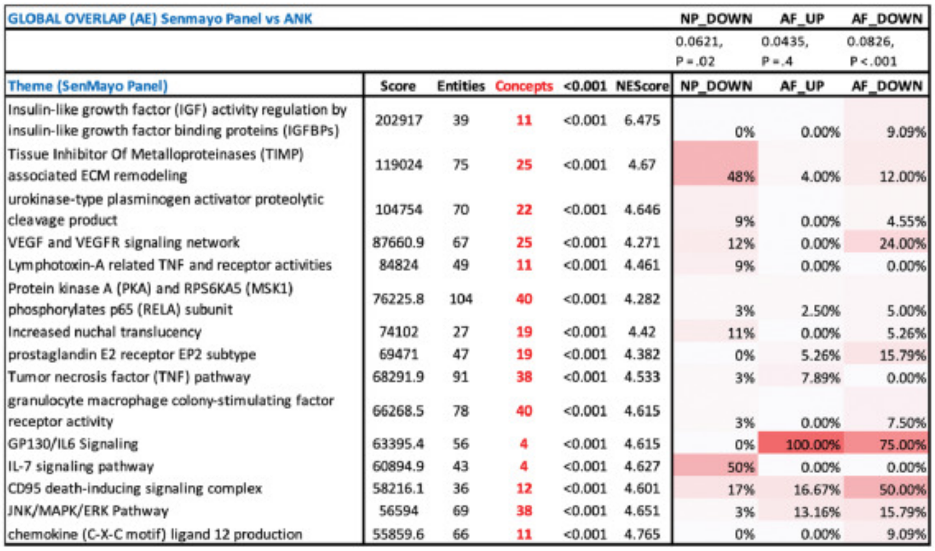


Theme	LTBI	ID	PO
T-Cells	1.5	1.6	2.6
Chemokines	1.5		1.8
Innate/TLR	1.4		2.0
Leukotriene Signaling	1.5		
Regulation of Apoptosis	1.5		
TH1/Lymphocytes	1.4		
Innate/Effector Cells	1.4		
Eosinophils	1.4		
TNF/AP4 Signaling	1.4		
NF-kB Signaling	1.4		
TH2/Cytokines	1.4		
Immune Diseases		1.7	
Monocyte/Macrophage Chemotaxis		1.7	
Cytotoxic Lymphocyte Regulation		1.7	
CD40L-mediated Mucosal Defense		1.5	
Inflammasome		1.5	2.3
Lymphocytes/Myeloid			2.4
Chemotaxis			2.3
AP4			2.1
Vasculature/Endothelium			2.0
GM-CSF Signaling			1.9
Immune Regulation			1.9

Silver et al., Manuscript in preparation



i



GLOBAL OVERLAP (AE) Senmayo Panel vs ANK						NP_DOWN	AF_UP	AF_DOWN
						0.0621, P = .02	0.0435, P = .4	0.0826, P < .001
Theme (SenMayo Panel)	**Score**	**Entities**	**Concepts**	**<0.001**	**NEScore**	**NP_DOWN**	**AF_UP**	**AF_DOWN**
Insulin-like growth factor (IGF) activity regulation by insulin-like growth factor binding proteins (IGFBPs)	202917	39	11	<0.001	6.475	0%	0.00%	9.09%
Tissue Inhibitor Of Metalloproteinases (TIMP) associated ECM remodeling	119024	75	25	<0.001	4.67	48%	4.00%	12.00%
urokinase-type plasminogen activator proteolytic cleavage product	104754	70	22	<0.001	4.646	9%	0.00%	4.55%
VEGF and VEGFR signaling network	87660.9	67	25	<0.001	4.271	12%	0.00%	24.00%
Lymphotoxin-A related TNF and receptor activities	84824	49	11	<0.001	4.461	9%	0.00%	0.00%
Protein kinase A (PKA) and RPS6KA5 (MSK1) phosphorylates p65 (RELA) subunit	76225.8	104	40	<0.001	4.282	3%	2.50%	5.00%
Increased nuchal translucency	74102	27	19	<0.001	4.42	11%	0.00%	5.26%
prostaglandin E2 receptor EP2 subtype	69471	47	19	<0.001	4.382	0%	5.26%	15.79%
Tumor necrosis factor (TNF) pathway	68291.9	91	38	<0.001	4.533	3%	7.89%	0.00%
granulocyte macrophage colony-stimulating factor receptor activity	66268.5	78	40	<0.001	4.615	3%	0.00%	7.50%
GP130/IL6 Signaling	63395.4	56	4	<0.001	4.615	0%	100.00%	75.00%
IL-7 signaling pathway	60894.9	43	4	<0.001	4.627	50%	0.00%	0.00%
CD95 death-inducing signaling complex	58216.1	36	12	<0.001	4.601	17%	16.67%	50.00%
JNK/MAPK/ERK Pathway	56594	69	38	<0.001	4.651	3%	13.16%	15.79%
chemokine (C-X-C motif) ligand 12 production	55859.6	66	11	<0.001	4.765	0%	0.00%	9.09%

Cell Death Dis. 2023 Jul 19;14(7):447 Ohnishi et al

5.6 Methods Description for Manuscripts

Platform Overview

CompBio (v2.5) is an artificial intelligence platform that enables the interpretation of omics and multi-omics data performed through a systematic analysis of literature pertaining to a list of biological entities (genes, proteins, miRNAs, or metabolites) provided the by user. Entities can be obtained from differential expression, correlation, or any number of relevant analysis methods. The first major component of the platform extracts and stores knowledge from PubMed abstracts and full-text articles using contextual language processing that is not restricted to fixed pathway and ontology knowledge bases. The extracted knowledge is maintained in a Persistent Contextual Memory Model (PCMM) for future use. The second major component is the generalized signal detection engine that can aggregate, normalize and compute across data from any stored dimension of the PCMM. Within the engine, biological concepts that are enriched with the input entities are identified and aggregated into higher-level biological themes (pathways/processes, cell types and structures, etc.) that are themselves inter-connected through their identified associations. The resulting knowledge maps are interactive with fully traceable information.

Scoring and Significance

Conditional probability analysis forms the foundation of gene, concept, and theme-level enrichment scores. The enrichment scores, in turn, are normalized and assessed for significance empirically, utilizing large, randomized comparator groups similar in construct to the query group. As such, the reported Normalized Enrichment Scores (NES) represent the magnitude to which the concepts/themes are enriched above random, and an empirically derived p-value identifies the likelihood of achieving that NES by chance.

Cross-Project Contextual Analysis – Assertion Engine v1.0

Fully contextualized analysis of multiple CompBio projects can be generated using the Assertion Engine (AE v1.0) component of the platform. The Assertion Engine is a machine-learning based engine capable of identifying preserved patterns between biological concepts of the compared projects and their inter-concept relationships. In other words, if a concept is present in both projects the AE determines how similar or different that concept's relationships are with the other concepts in the respective data sets. Strongly preserved groups of concept/inter-relationship associations will form high scoring sub-maps within the overall analysis provided to the user. A global score, representing the complete contextual biological similarity between the two is also computed. A score of 0.0 represents no similarity and a score of 1.0 represents complete similarity. P-values are also computed for the significance of the global similarity score.

Project and Theme Annotation

Within CompBio, biological themes and their relationships are computed in a fully automated fashion that is independent of existing ontologies. The biology described by the concepts in the themes can often be obvious to the user, however, it is generally desirable to have high-level annotation, at a process or pathway level, for the themes that is commonly used and recognized by a broad audience. To accommodate this, the auto-annotation system utilizes a language processing method that leverages the theme contents and over 100,000 biological terms, processes, and pathways to produce optimal annotations in a fully automated fashion. However, unlike traditional pathway/process tools, the CompBio annotation system not only uses the contents of a given theme when annotating it, but also considers the contents of neighboring themes as well. This provides overall annotation within the project that is, generally, more accurate and contextually specific. The same principles are utilized to generate project-level annotation as well.