Accurate and Comprehensive Mapping of Multi-omic Data to Biological Pathways

Application Note

Integrated Biology

Authors

Anupama Rajan Bhat and Pramila Tata
Strand Life Sciences
Bangalore, India

Stephen Madden
Agilent Technologies, Inc.
Santa Clara, California

Abstract

This application note describes the use of Agilent-BridgeDB, an essential technology in Agilent's GeneSpring/Mass Profiler Professional (MPP) product to accurately map biological entities on pathways. It describes four case studies that demonstrate how Agilent-BridgeDB enables significantly more accurate mappings between experimentally identified biological entities (for example, genes, metabolites) and the corresponding entities in pathway databases. Common bioinformatics challenges like missing annotations, resolving enantiomers, and incomplete databases are overcome using the Agilent-BridgeDB technology.

Introduction

Pathway analysis provides a useful biological context for differentially expressed entities resulting from the analysis of high-throughput data in any 'omics' (for example genomics, transcriptomics, proteomics, or metabolomics) experiment. Pathways overrepresented or enriched in the entities of interest can provide mechanistic insights into the underlying biology of the conditions under study. Many popular pathway databases such as KEGG [1], BioCyc [2], and WikiPathways [3] provide detailed and well-annotated pathways. However, comparisons of pathway databases suggest that no single pathway database is comprehensive [4,5,6]. Further, it has been observed that these databases are partly complementary, and thus it is important for researchers to be able to access pathways from multiple sources simultaneously to gain a more complete picture and not miss possible biological interpretations. The Pathway Architect module in GeneSpring and MPP supports the import and analysis of pathways from these popular pathway databases. In addition, Pathway Architect also supports the import of pathways using standard formats such as BioPAX and GPML.

A lack of standardization in the names and identifiers of biological entities in pathways across multiple pathway databases results in the same entity being cited with different names or identifiers across databases and at times even across pathways within a single pathway database. In some cases, different entities of the same type (gene/protein/metabolite) within a pathway can cite identifiers from different databases as well. Furthermore, in the context of a GeneSpring/MPP experiment, the identifiers associated with entities of interest in the experiment may be different from the identifiers available with the pathway entities. This well-recognized identifier mapping problem poses a major challenge in pathway analysis and limits the matches between the entities from the experiment and their counterparts in pathways.

For example, the metabolite D-glucose (Figure 1) might be known alternatively as dextrose, meritose, or (3R,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol. It has 23 synonyms listed in the Human Metabolite Database (HMDB).

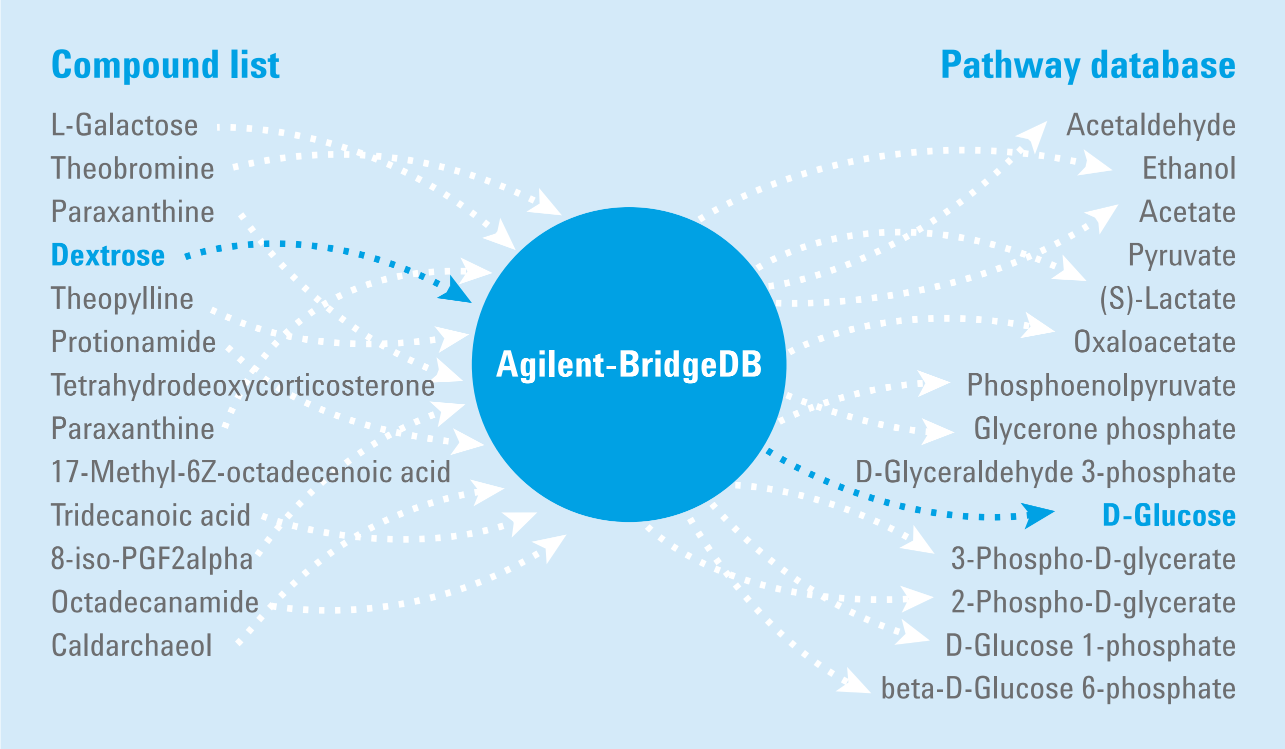


Figure 1. Mapping of metabolites using Agilent-BridgeDB.

To overcome this limitation, GeneSpring/MPP uses a modified version of the BridgeDB software framework [7] to ensure all possible matches between the experiment and the pathway are reported. Mapper files used by the framework provide the mapping between different entity databases to equate an entity in one dataset (pathway) with the same entity in another dataset (experiment). One way to visualize the mapping is in the form of a table where the rows connect all the synonyms and identifiers for an entity (Table 1).

Table 1. D-Glucose aligned with a synonym and some database identifiers.

Common name	Synonym	KEGG ID	Cas no.	HMDB ID	ChEBI ID
D-Glucose	Dextrose	C00031	50-99-7	001222	4167

There are two types of mapper files currently being used in GeneSpring/MPP-(a) Gene/Protein mapper file and (b) Metabolite mapper file. The Gene/Protein mapper file is organism specific, while the metabolite mapper is common for all organisms. The gene/protein mappers used in GeneSpring/MPP are from the Gladstone Institute and are primarily extracted from Ensembl [8]. The metabolite mapper is developed at Agilent Technologies.

Here we describe four case studies demonstrating the role of Agilent-BridgeDB and the mapper files in enhancing the pathway analysis capabilities of GeneSpring/MPP.

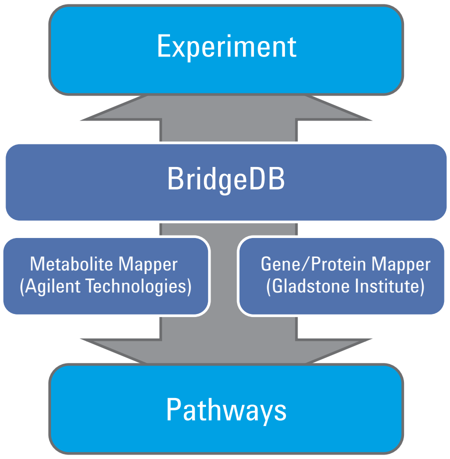
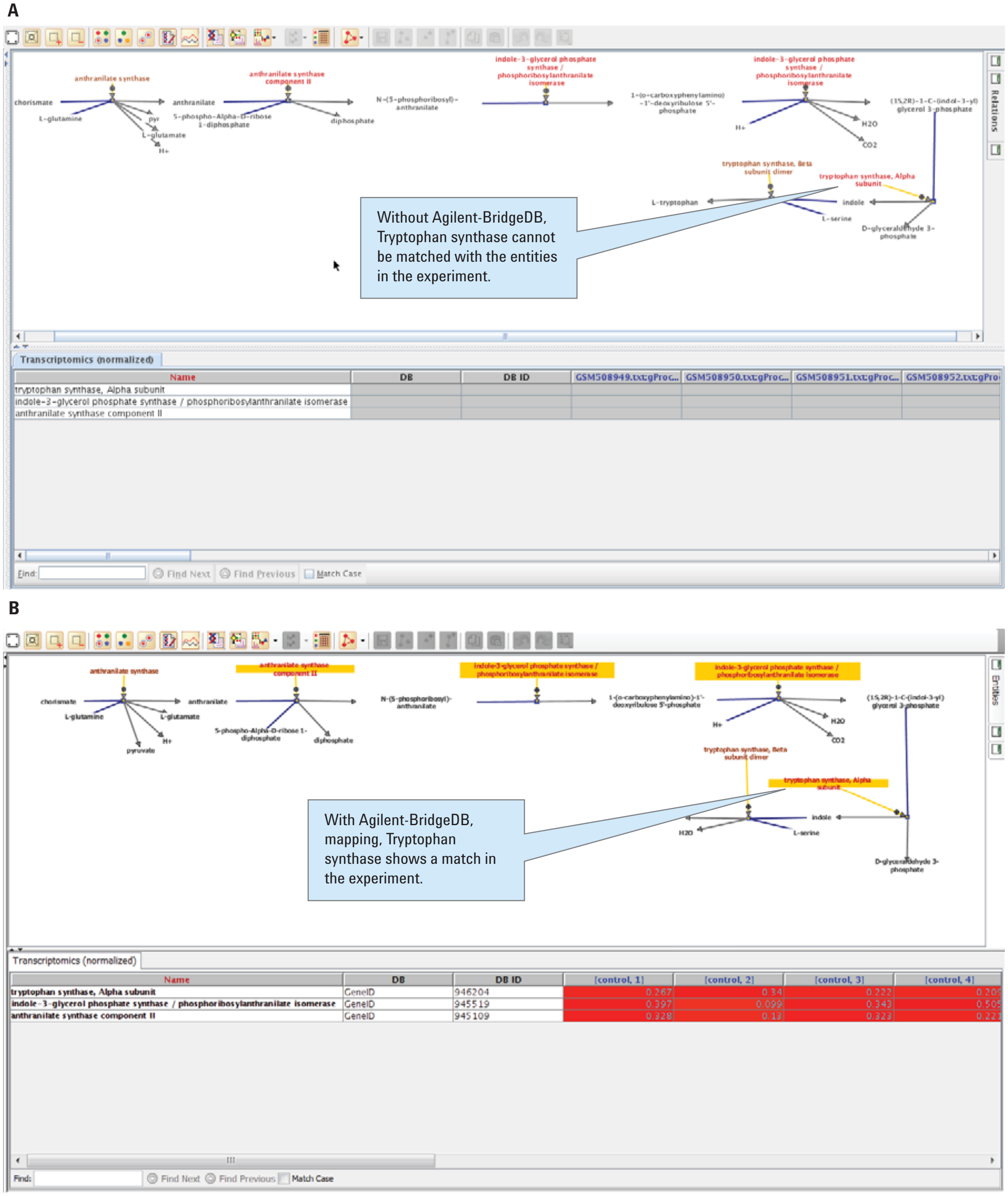


Figure 2. BridgeDB framework in GeneSpring using Agilent metabolite mapper Agilent-BridgeDB and the Gladstone Institute gene/protein mapper to map identifiers across pathways and experiment entities.

Case Study 1: Mapping different annotations in a pathway and an experiment

Figure 3 shows a pathway in a transcriptomics experiment in GeneSpring. Genes in the experiment are annotated with their Entrez Gene IDs. Table 2 shows an example of the properties available for one of the genes, 'trytophan synthase', in the BioCyc pathway in focus. Genes in this pathway do not cite an Entrez Gene ID, but are annotated with identifiers from other databases. Due to the absence of common identifiers or a bridging mechanism, the pathway does not show any enrichment and the entities do not show any matches with


A
anthranilate synthase
anthranilate synthase component II
indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase
chorismate
L-glutamine
pyr
L-glutamate
H+
anthranilate
5-phospho-Alpha-D-ribose 1-diphosphate
diphosphate
N-(5-phosphoribosyl)-anthranilate
1-(o-carboxyphenylamino)-1'-deoxyribulose 5'-phosphate
H+
H2O
CO2
(1S,2R)-1-C-(indol-3-yl) glycerol 3-phosphate
tryptophan synthase, Beta subunit dimer
tryptophan synthase, Alpha subunit
L-tryptophan
indole
L-serine
D-glyceraldehyde 3-phosphate
Relations
B
Entities
pyruvate

Without Agilent-BridgeDB, Tryptophan synthase cannot be matched with the entities in the experiment.

Transcriptomics (normalized)

Name	DB	DB ID	GSM508949.txt:gProc...	GSM508950.txt:gProc...	GSM508951.txt:gProc...	GSM508952.txt:gPro
tryptophan synthase, Alpha subunit						
indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase						
anthranilate synthase component II						

Find: ○ Find Next ○ Find Previous □ Match Case

With Agilent-BridgeDB, mapping, Tryptophan synthase shows a match in the experiment.

Transcriptomics (normalized)

Name	DB	DB ID	[control, 1]	[control, 2]	[control, 3]	[control, 4]
tryptophan synthase, Alpha subunit	GeneID	946204	0.267	0.34	0.222	0.209
indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase	GeneID	945519	0.397	0.099	0.343	0.505
anthranilate synthase component II	GeneID	945109	0.328	0.13	0.323	0.221

Find: ○ Find Next ○ Find Previous □ Match Case

Figure 3. Tryptophan biosynthesis pathway from BioCyc in a transcriptomics experiment. A) without Agilent-BridgeDB and B) with Agilent-BridgeDB. Yellow background color indicates matches with the experiment.

the experiment (Figure 3A). As a result, the pathway is ignored in the analysis.

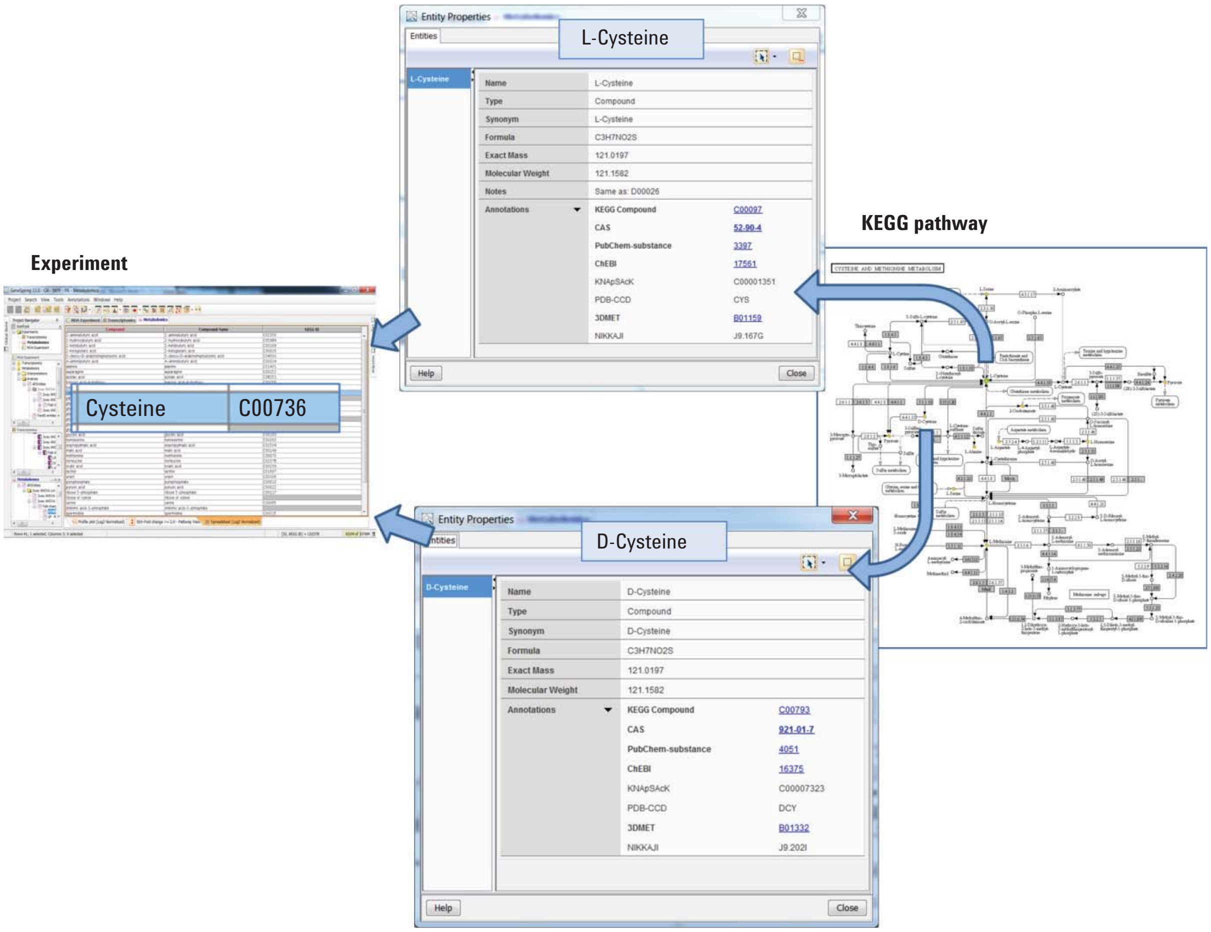
When the experiment is re-analyzed using Agilent-BridgeDB and the organism specific mapper files, mappings from pathway identifiers to experiment identifiers are retrieved and a match is identified. In the case of tryptophan synthase, the mapping from UniProt/TrEMBL identifier P0A877 (pathway) to Entrez ID 946204 (experiment) is available and is matched in pathway analysis (Figure 3B).

Table 2. Annotations in BioCyc pathway for entity tryptophan synthase.

Property	**Valve**
Cellular location	Cytosol
DIP	DIP-35957N
DisProt	DP00252
EcoCyc	TRYPSYN-APROTEIN
EcoliWiki	b1260
InterPro	IPR013785
InterPro	IPR011060
InterPro	IPR018204
InterPro	IPR002028
Label	TrpA
ModBase	P0A877
Organism	Escherichia coli K-12 substr. MG1655
PDB	1V7Y
PDB	1WQ5
PDB	1XC4
PDB	1XCF
PR	PRO_000024117
PRIDE	P0A877
PROSITE	PS00167
Pfam	PF00290
Protein model portal	P0A877
RefSeq	NP_415776
SMR	P0A877
String	511145.b1260
Synonym	Try
Synonym	TrpA
Synonym	Alpha subunit
Synonym	TSase Alpha
Synonym	A protein
Uniprot/TrEMBL	P0A877

Case Study 2: Mapping of isomers between pathway and experiment

Figure 4 demonstrates a case in which Agilent-BridgeDB enables mapping of specific enantiomers to their D/L form. In this example, both the metabolomics experiment and the metabolites in the KEGG pathway have KEGG Compound identifiers. However, while the experiment contains the KEGG identifier for the D/L form of cysteine (C00736), the pathway cites the isomer specific identifiers: L-cysteine (C00097) and D-cysteine (C00793). Agilent-BridgeDB uses the mappings in the Agilent metabolite mapper to ensure the specific forms of the isomer get mapped to the generic form in the experiment.



Name	L-Cysteine	
Type	Compound	
Synonym	L-Cysteine	
Formula	C3H7NO2S	
Exact Mass	121.0197	
Molecular Weight	121.1582	
Notes	Same as: D00026	
Annotations	KEGG Compound	C00097
	CAS	52-90-4
	PubChem-substance	3397
	ChEBI	17561
	KNApSAcK	C00001351
	PDB-CCD	CYS
	3DMET	B01159
	NIKKAJI	J9.167G

Name	D-Cysteine	
Type	Compound	
Synonym	D-Cysteine	
Formula	C3H7NO2S	
Exact Mass	121.0197	
Molecular Weight	121.1582	
Annotations	KEGG Compound	C00793
	CAS	921-01-7
	PubChem-substance	4051
	ChEBI	16375
	KNApSAcK	C00007323
	PDB-CCD	DCY
	3DMET	B01332
	NIKKAJI	J9.202I

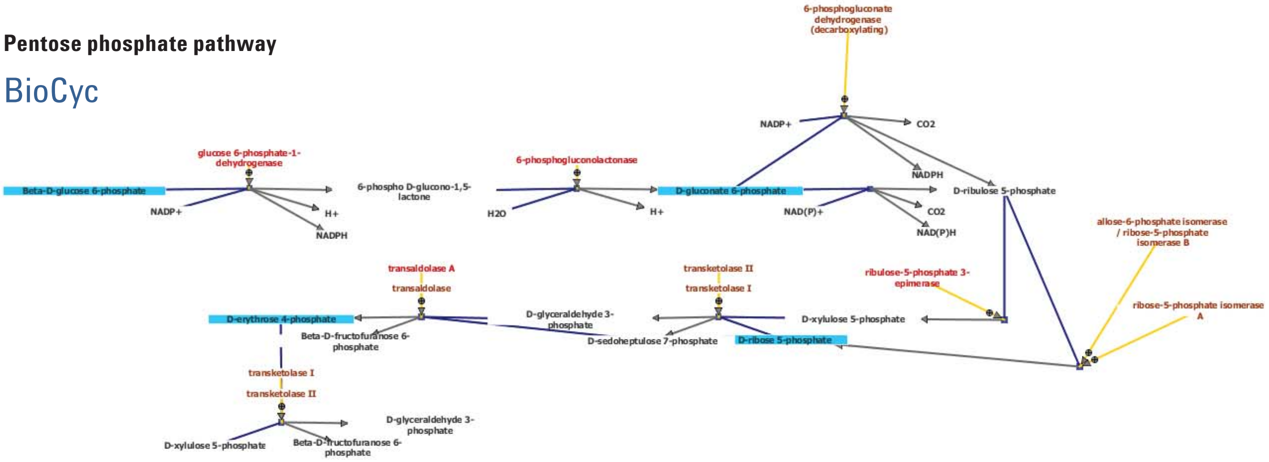
Figure 4. Specific isomers in a KEGG pathway are matched with the generic form in the experiment through Agilent-BridgeDB.

Case Study 3: Mapping multi-omic experiments to multiple pathway databases

Pathways from multiple sources contain complementary information and together are able to provide a more comprehensive picture of biological processes. The ability to map the same entity with different identifiers through Agilent-BridgeDB enables powerful analysis of pathways simultaneously from multiple sources in GeneSpring/MPP. This becomes useful for cases in which pathways from one source cannot be matched with the experiment due to missing annotations. For example, Figure 5 shows the pentose phosphate pathway from two sources, BioCyc and KEGG, enriched in a multi-omics experiment. Metabolites in both pathways could be matched with the experiment. However, proteins in the BioCyc pathway could not be matched with the transcriptomics experiment due to missing annotations, while proteins in the KEGG pathway could be. Thus in the absence of the metabolite mapper files, enrichment of the pentose phosphate experiment from BioCyc would have been overlooked.

Pentose phosphate pathway

BioCyc


6-phosphogluconate dehydrogenase (decarboxylating)
NADP+
CO2
glucose 6-phosphate-1-dehydrogenase
6-phosphogluconolactonase
NADPH
Beta-D-glucose 6-phosphate
6-phospho D-glucono-1,5-lactone
D-gluconate 6-phosphate
D-ribulose 5-phosphate
NADP+
H+
H2O
H+
NAD(P)+
CO2
NADPH
NAD(P)H
allose-6-phosphate isomerase / ribose-5-phosphate isomerase B
transaldolase A
transaldolase
transketolase II
transketolase I
ribulose-5-phosphate 3-epimerase
ribose-5-phosphate isomerase A
D-erythrose 4-phosphate
D-glyceraldehyde 3-phosphate
D-xylulose 5-phosphate
Beta-D-fructofuranose 6-phosphate
D-sedoheptulose 7-phosphate
D-ribose 5-phosphate
transketolase I
transketolase II
D-glyceraldehyde 3-phosphate
D-xylulose 5-phosphate
Beta-D-fructofuranose 6-phosphate

Pentose phosphate pathway

KEGG

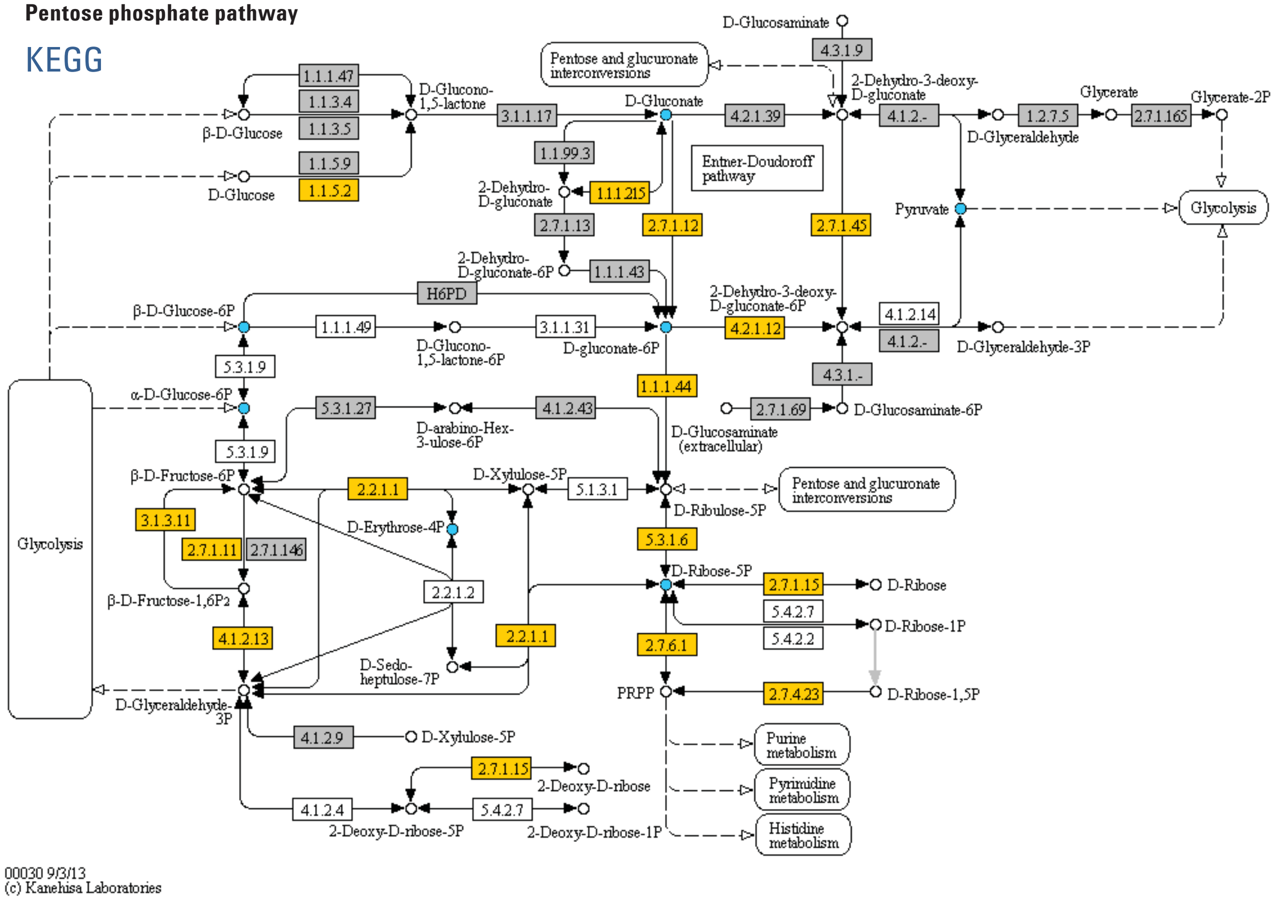
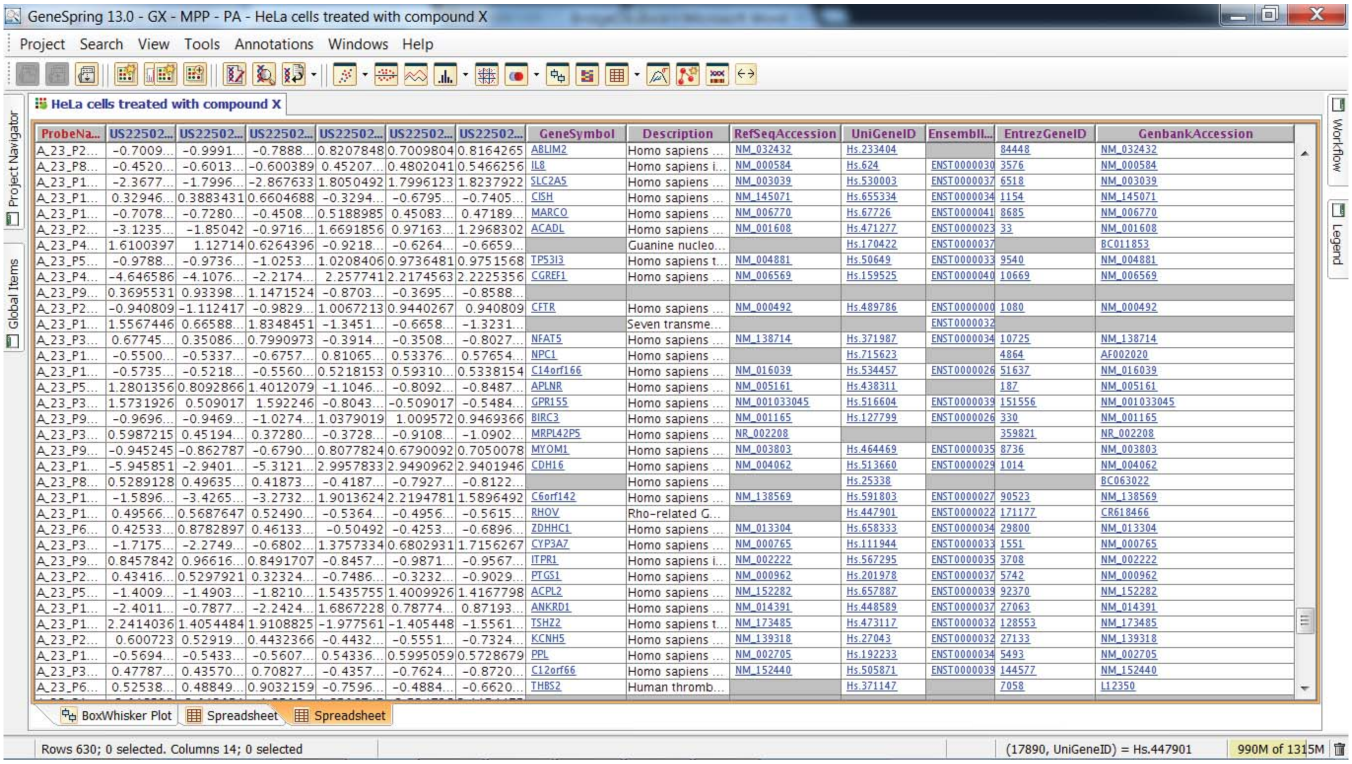


Figure 5. Multi-omics analysis results for the pentose phosphate pathway from BioCyc and KEGG. Matches with the experiment are indicated by the background color of the entity. Yellow indicates gene/protein matches with the transcriptomics experiment. Blue indicates metabolite matches with the metabolomics experiment.

Case Study 4: Mapping experiment entities with missing annotations

In some cases, the experiment may have more than one annotation column. It is possible that an entity with a missing identifier in one annotation has been assigned an identifier from another database. For example, Figure 6 shows a genomics experiment with multiple annotation columns: RefSeq Accession, UniGene ID, Ensembl ID, Entrez Gene ID, and Genbank Accession. Note that not all of the database identifiers are present for all entities. Mapping using any single database identifier will invariably lead to loss of matches due to missing annotations. However, pathway analysis in GeneSpring/MPP considers all available annotations for a specific entity in a predetermined order.



ProbeNa...	US22502...	US22502...	US22502...	US22502...	US22502...	US22502...	GeneSymbol	Description	RefSeqAccession	UniGeneID	Ensembll...	EntrezGeneID	GenbankAccession
A_23_P2...	-0.7009...	-0.9991...	-0.7888...	0.8207848	0.7009804	0.8164265	ABLIM2	Homo sapiens ...	NM_032432	Hs.233404		84448	NM_032432
A_23_P8...	-0.4520...	-0.6013...	-0.600389	0.45207...	0.4802041	0.5466256	IL8	Homo sapiens i...	NM_000584	Hs.624	ENST0000030	3576	NM_000584
A_23_P1...	-2.3677...	-1.7996...	-2.867633	1.8050492	1.7996123	1.8237922	SLC2A5	Homo sapiens ...	NM_003039	Hs.530003	ENST0000037	6518	NM_003039
A_23_P1...	0.32946...	0.3883431	0.6604688	-0.3294...	-0.6795...	-0.7405...	CISH	Homo sapiens ...	NM_145071	Hs.655334	ENST0000034	1154	NM_145071
A_23_P1...	-0.7078...	-0.7280...	-0.4508...	0.5188985	0.45083...	0.47189...	MARCO	Homo sapiens ...	NM_006770	Hs.67726	ENST0000041	8685	NM_006770
A_23_P2...	-3.1235...	-1.85042	-0.9716...	1.6691856	0.97163...	1.2968302	ACADL	Homo sapiens ...	NM_001608	Hs.471277	ENST0000023	33	NM_001608
A_23_P4...	1.6100397	1.12714	0.6264396	-0.9218...	-0.6264...	-0.6659...		Guanine nucleo...		Hs.170422	ENST0000037		BC011853
A_23_P5...	-0.9788...	-0.9736...	-1.0253...	1.0208406	0.9736481	0.9751568	TP53I3	Homo sapiens t...	NM_004881	Hs.50649	ENST0000033	9540	NM_004881
A_23_P4...	-4.646586	-4.1076...	-2.2174...	2.257741	2.2174563	2.2225356	CGREF1	Homo sapiens ...	NM_006569	Hs.159525	ENST0000040	10669	NM_006569
A_23_P9...	0.3695531	0.93398...	1.1471524	-0.8703...	-0.3695...	-0.8588...							
A_23_P2...	-0.940809	-1.112417	-0.9829...	1.0067213	0.9440267	0.940809	CFTR	Homo sapiens ...	NM_000492	Hs.489786	ENST0000000	1080	NM_000492
A_23_P1...	1.5567446	0.66588...	1.8348451	-1.3451...	-0.6658...	-1.3231...		Seven transme...			ENST0000032		
A_23_P3...	0.67745...	0.35086...	0.7990973	-0.3914...	-0.3508...	-0.8027...	NFAT5	Homo sapiens ...	NM_138714	Hs.371987	ENST0000034	10725	NM_138714
A_23_P1...	-0.5500...	-0.5337...	-0.6757...	0.81065...	0.53376...	0.57654...	NPC1	Homo sapiens ...		Hs.715623		4864	AF002020
A_23_P1...	-0.5735...	-0.5218...	-0.5560...	0.5218153	0.59310...	0.5338154	C14orf166	Homo sapiens ...	NM_016039	Hs.534457	ENST0000026	51637	NM_016039
A_23_P5...	1.2801356	0.8092866	1.4012079	-1.1046...	-0.8092...	-0.8487...	APLNR	Homo sapiens ...	NM_005161	Hs.438311		187	NM_005161
A_23_P3...	1.5731926	0.509017	1.592246	-0.8043...	-0.509017	-0.5484...	GPR155	Homo sapiens ...	NM_001033045	Hs.516604	ENST0000039	151556	NM_001033045
A_23_P9...	-0.9696...	-0.9469...	-1.0274...	1.0379019	1.009572	0.9469366	BIRC3	Homo sapiens ...	NM_001165	Hs.127799	ENST0000026	330	NM_001165
A_23_P3...	0.5987215	0.45194...	0.37280...	-0.3728...	-0.9108...	-1.0902...	MRPL42P5	Homo sapiens ...	NR_002208			359821	NR_002208
A_23_P9...	-0.945245	-0.862787	-0.6790...	0.8077824	0.6790092	0.7050078	MYOM1	Homo sapiens ...	NM_003803	Hs.464469	ENST0000035	8736	NM_003803
A_23_P1...	-5.945851	-2.9401...	-5.3121...	2.9957833	2.9490962	2.9401946	CDH16	Homo sapiens ...	NM_004062	Hs.513660	ENST0000029	1014	NM_004062
A_23_P8...	0.5289128	0.49635...	0.41873...	-0.4187...	-0.7927...	-0.8122...		Homo sapiens ...		Hs.25338			BC063022
A_23_P1...	-1.5896...	-3.4265...	-3.2732...	1.9013624	2.2194781	1.5896492	C6orf142	Homo sapiens ...	NM_138569	Hs.591803	ENST0000027	90523	NM_138569
A_23_P1...	0.49566...	0.5687647	0.52490...	-0.5364...	-0.4956...	-0.5615...	RHOV	Rho-related G...		Hs.447901	ENST0000022	171177	CR618466
A_23_P6...	0.42533...	0.8782897	0.46133...	-0.50492	-0.4253...	-0.6896...	ZDHHC1	Homo sapiens ...	NM_013304	Hs.658333	ENST0000034	29800	NM_013304
A_23_P3...	-1.7175...	-2.2749...	-0.6802...	1.3757334	0.6802931	1.7156267	CYP3A7	Homo sapiens ...	NM_000765	Hs.111944	ENST0000033	1551	NM_000765
A_23_P9...	0.8457842	0.96616...	0.8491707	-0.8457...	-0.9871...	-0.9567...	ITPR1	Homo sapiens i...	NM_002222	Hs.567295	ENST0000035	3708	NM_002222
A_23_P2...	0.43416...	0.5297921	0.32324...	-0.7486...	-0.3232...	-0.9029...	PTGS1	Homo sapiens ...	NM_000962	Hs.201978	ENST0000037	5742	NM_000962
A_23_P5...	-1.4009...	-1.4903...	-1.8210...	1.5435755	1.4009926	1.4167798	ACPL2	Homo sapiens ...	NM_152282	Hs.657887	ENST0000039	92370	NM_152282
A_23_P1...	-2.4011...	-0.7877...	-2.2424...	1.6867228	0.78774...	0.87193...	ANKRD1	Homo sapiens ...	NM_014391	Hs.448589	ENST0000037	27063	NM_014391
A_23_P1...	2.2414036	1.4054484	1.9108825	-1.977561	-1.405448	-1.5561...	TSHZ2	Homo sapiens t...	NM_173485	Hs.473117	ENST0000032	128553	NM_173485
A_23_P2...	0.600723	0.52919...	0.4432366	-0.4432...	-0.5551...	-0.7324...	KCNH5	Homo sapiens ...	NM_139318	Hs.27043	ENST0000032	27133	NM_139318
A_23_P1...	-0.5694...	-0.5433...	-0.5607...	0.54336...	0.5995059	0.5728679	PPL	Homo sapiens ...	NM_002705	Hs.192233	ENST0000034	5493	NM_002705
A_23_P3...	0.47787...	0.43570...	0.70827...	-0.4357...	-0.7624...	-0.8720...	C12orf66	Homo sapiens ...	NM_152440	Hs.505871	ENST0000039	144577	NM_152440
A_23_P6...	0.52538...	0.48849...	0.9032159	-0.7596...	-0.4884...	-0.6620...	THBS2	Human thromb...		Hs.371147		7058	L12350

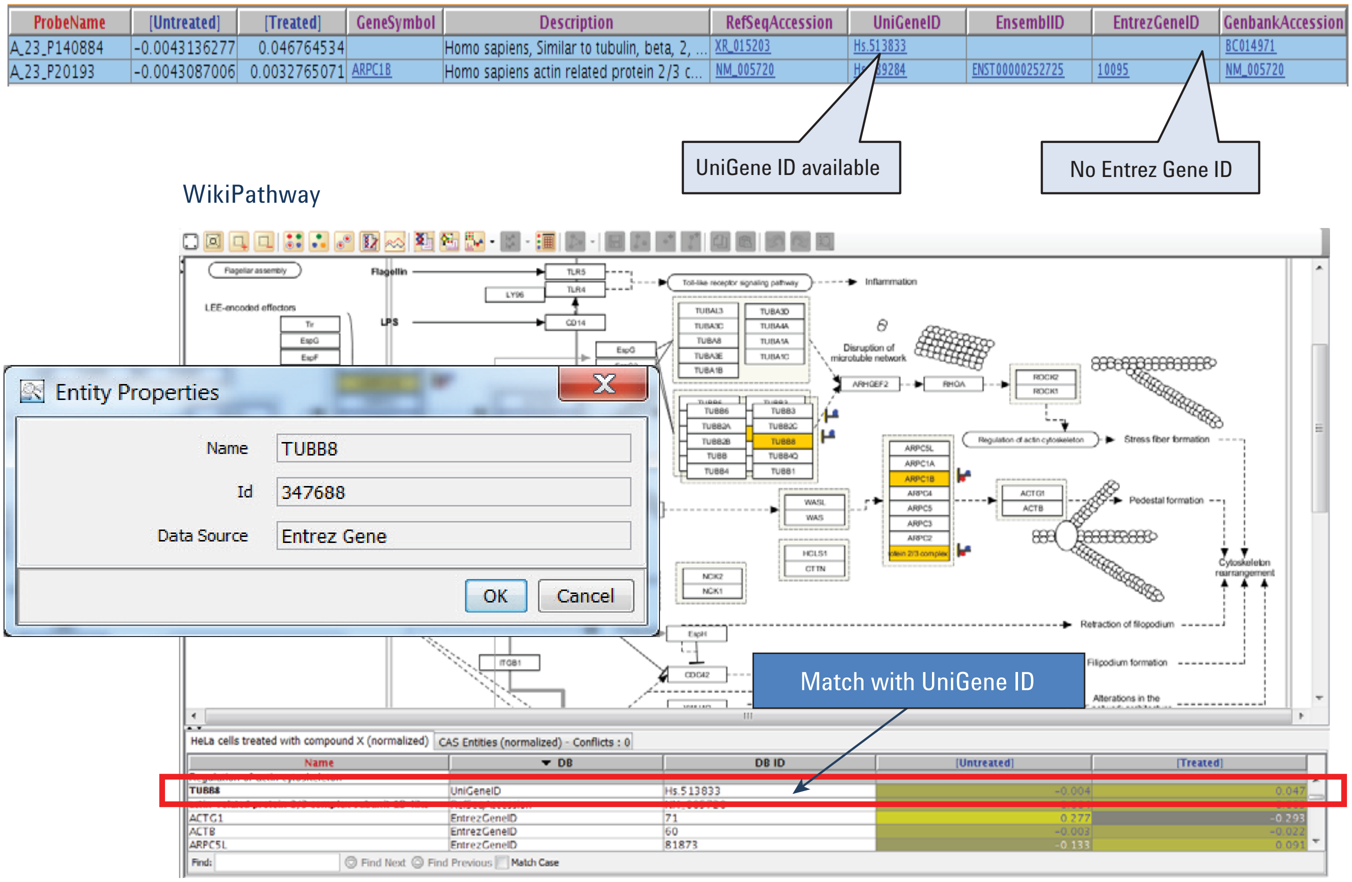
Rows 630; 0 selected. Columns 14; 0 selected

(17890, UniGeneID) = Hs.447901

990M of 1315M

Figure 6. Multiple complementary annotation columns in a GeneSpring experiment.

This ensures that entities with sparse annotations are also mapped. In Figure 7 the experiment has the Entrez Gene ID annotation column, but an identifier is not available specifically for the putative ‘tubulin’ gene. Therefore, Agilent-BridgeDB attempts to match a pathway entity with other available identifiers for this gene. In this case it retrieves a mapping to the UniGene ID and the pathway entity in WikiPathways is matched with the experiment.



ProbeName	[Untreated]	[Treated]	GeneSymbol	Description	RefSeqAccession	UniGeneID	EnsemblID	EntrezGeneID	GenbankAccession
A_23_P140884	-0.0043136277	0.046764534		Homo sapiens, Similar to tubulin, beta, 2, ...	XR_015203	Hs.513833			BC014971
A_23_P20193	-0.0043087006	0.0032765071	ARPC1B	Homo sapiens actin related protein 2/3 c...	NM_005720	H[illegible]89284	ENST00000252725	10095	NM_005720

UniGene ID available

No Entrez Gene ID

WikiPathway

Entity Properties

Name: TUBB8

Id: 347688

Data Source: Entrez Gene

Match with UniGene ID

HeLa cells treated with compound X (normalized) | CAS Entities (normalized) - Conflicts : 0

Name	DB	DB ID	[Untreated]	[Treated]
TUBB8	UniGeneID	Hs.513833	-0.004	0.047
ACTG1	EntrezGeneID	71	0.277	-0.293
ACTB	EntrezGeneID	60	-0.003	-0.022
ARPC5L	EntrezGeneID	81873	-0.133	0.091

Find: ○ Find Next ○ Find Previous □ Match Case

Figure 7. Pathway entity with Entrez Gene ID is matched to its counterpart in the experiment through its UniGene ID by Agilent-BridgeDB, since the Entrez ID is not available in the experiment.

Conclusions

Specific examples have been presented across different pathway databases (KEGG, BioCyc, and WikiPathways) and ‘omics techniques (genomics, transcriptomics, and metabolomics) available in GeneSpring/MPP. Each of them demonstrated that researchers can get more accurate and comprehensive mappings of their experimental data to pathway databases due to the Agilent-BridgeDB technology. Biological entities that are missing specific annotations in either the experiment or pathway can still be mapped, resulting in more useful information. Multi-omics experiments are more likely to indicate pathways enriched in multiple ‘omic technologies since GeneSpring/MPP has mappers for genes, proteins, and metabolites. Successful mapping helps drive research forward by highlighting important pathways and making planning for the next experiment significantly more effective.

References

1. Kanehisa, *et al*. *Nucleic Acids Research*, **42**:D199 (2014).
2. Caspi, *et al. Nucleic Acids Research*, **38**:D473 (2010).
3. Kelder, *et al*. *Nucleic Acids Research*, **40**:D1301 (2012).
4. Stobbe, *et al*. *BMC Systems Biology*, **5**:165 (2011).
5. Soh, *et al*. *BMC Bioinformatics*, **11**:449 (2010).
6. Altman, *et al. BMC Bioinformatics*, **14**:112 (2013).
7. van Iersel, *et al*. *BMC Bioinformatics*, **11**(1):5 (Jan 4, 2010).
8. Flicek, *et al. Nucleic Acids Research*, **42**:D749 (2014).

For More Information

These data represent typical results. For more information on our products and services, visit our Web site at www.agilent.com/chem.

www.agilent.com/chem

Agilent shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Information, descriptions, and specifications in this publication are subject to change without notice.

For Research Use Only. Not for use in diagnostic procedures.

PR7000-0382

Printed in the USA February 10, 2016
5991-5192EN

Agilent Technologies