Database, 2022, 1–15
DOI: https://doi.org/10.1093/database/baac054
Original article

DATABASE
The Journal of Biological Databases and Curation

HumanMine: advanced data searching, analysis and cross-species comparison

Rachel Lyne, Adrián Bazaga, Daniela Butano, Sergio Contrino, Joshua Heimbach, Fengyuan Hu, Alexis Kalderimis, Mike Lyne, Kevin Reierskog, Radek Stepan, Julie Sullivan, Archie Wise, Yo Yehudi and Gos Micklem*

Department of Genetics, University of Cambridge, Downing Pl, Cambridge, CB2 3EH, UK

*Corresponding author: Tel: +44 (0)1223 333999; Fax: +44 (0)1223 333992; Email: gm263@cam.ac.uk

Citation details: Lyne, R., Bazaga, A., Butano, D. *et al.* HumanMine: advanced data searching, analysis and cross-species comparison. *Database* (2022) Vol. 2022: article ID baac054; DOI: https://doi.org/10.1093/database/baac054

Abstract

HumanMine (www.humanmine.org) is an integrated database of human genomics and proteomics data that provides a powerful interface to support sophisticated exploration and analysis of data compiled from experimental, computational and curated data sources. Built using the InterMine data integration platform, HumanMine includes genes, proteins, pathways, expression levels, Single nucleotide polymorphism (SNP), diseases and more, integrated into a single searchable database. HumanMine promotes integrative analysis, a powerful approach in modern biology that allows many sources of evidence to be analysed together. The data can be accessed through a user-friendly web interface as well as a powerful, scriptable web service Application programming interface (API) to allow programmatic access to data. The web interface includes a useful identifier resolution system, sophisticated query options and interactive results tables that enable powerful exploration of data, including data summaries, filtering, browsing and export. A set of graphical analysis tools provide a rich environment for data exploration including statistical enrichment of sets of genes or other biological entities. HumanMine can be used for integrative multistaged analysis that can lead to new insights and uncover previously unknown relationships.

Database URL: https://www.humanmine.org

Introduction

HumanMine (www.humanmine.org) is an integrated database of human genomics and proteomics data that provides a powerful interface to support sophisticated exploration and analysis of data compiled from experimental, computational and curated data sources. Built using the InterMine data integration platform, HumanMine includes genes, proteins, pathways, expression levels, single nucleotide polymorphisms, SNPs, diseases and more, integrated into a single searchable database. Therefore, HumanMine promotes integrative analysis, a powerful approach in modern biology that allows many sources of evidence to be analysed together. Data integration enables different pieces of information about a particular biological entity, or set of entities, to be surveyed together without the need to visit several databases. Such an approach has many advantages, including fewer false positives, compensation for missing or unreliable information and interrogation at different levels of genetic, genomic or proteomic regulation so helping in the understanding of complex biological systems. An important step in understanding such processes is also the interpretation of complementary data from model organisms. HumanMine interoperates with InterMine databases for a number of model organisms including mouse (MouseMine, (1)), rat (RatMine, (2)), zebrafish (ZebrafishMine, (3)), fly (FlyMine, (4)), worm (WormMine, http://intermine.wormbase.org/tools/wormmine), yeast (YeastMine, (4, 5)) and Arabidopsis (ThaleMine, (6)). Here we describe how HumanMine can be used for integrative multistaged analysis that can lead to new insights and uncover previously unknown relationships.

We will start by using the gene *PAX6* to illustrate the capabilities of HumanMine and then consider a more detailed use case in which HumanMine is used to support the exploration of disease comorbidities.

InterMine

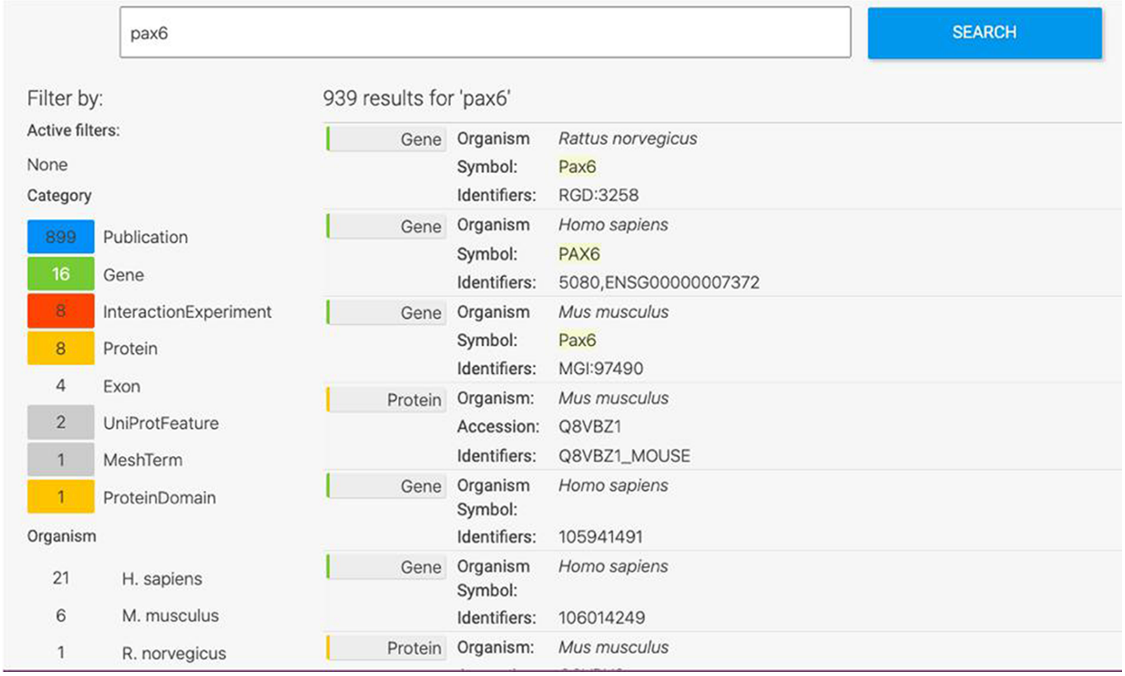
InterMine (7) is a data warehouse framework designed to facilitate the integration of diverse biological datasets and provide tools for bioinformatics analysis and visualization both through a web interface and through extensive web services generated automatically from the underlying data model. InterMine was originally developed as FlyMine to address these issues for the *Drosophila* community. It has since been adopted by other model organism databases (MODs) including mouse (1), yeast (5), worm (http://intermine.wormbase.org), fish (3) and rat (2) and a growing

Received 11 May 2022; Revised 7 June 2022; Accepted 21 June 2022

number of databases covering other animals, including Planaria (8), Hymenoptera (9), locusts (10) and armyworm (11). Plant species include Arabidopsis (6), Medicago (12), wheat (https://urgi.versailles.inra.fr/WheatMine), maize (11, 13), and several from the legume family (14) (see registry.intermine.org). InterMine is also used for drug development (TargetMine (15)) and by large-scale projects and consortia (PhytoMine (16), modMine (17), AllianceMine (https://www.alliancegenome.org/alliancemine)). HumanMine, described here, provides access to selected human data through the InterMine interface. A detailed description of the InterMine system is available elsewhere (7), but here we briefly review the data model and the relationships between different types of data. InterMine is able to integrate data from a wide variety of sources in various biological formats, including GFF3, FASTA, VCF, OBO, BioPAX (18), GAF, PSI (19) and Chado (20) and includes a powerful identifier resolution system such that any outdated identifiers from a dataset are mapped to the current ones. The underlying object-based core data model is based on the Sequence Ontology (SO (21)). This core is extended to include additional data types that do not fit the SO, such as protein–protein interactions. Each type in the model (e.g. Gene, Disease) is called a class and is described through a set of attributes (such as gene symbol and gene length). References between classes link related data types (e.g. Gene and Allele) and facilitate navigation and exploration of the data. InterMine allows users to work with single items or lists of items (for example, a list of genes) and provides a number of search and analysis interfaces, including a keyword search, a genomic regions search and an advanced query builder, which exposes the data model and relationships and so allows the user to construct complex searches over the integrated data. In addition, a comprehensive web services API makes it straightforward to access data supplied by InterMine databases within programs, with support for most commonly used scripting languages (22). As many aspects of HumanMine conform to the FAIR principles of data management (23), the features described help to make analyses reproducible and the data more findable, accessible, interoperable and reusable. HumanMine uses the recently released new interface to InterMine, BlueGenes (Manuscript in preparation).

Data

HumanMine currently includes around 40 different data sets (Supplementary Table S1). The latest human reference genome and annotation release are loaded from the NCBI (https://www.ncbi.nlm.nih.gov/genome/?term=human[organism]). Data are focussed on high-throughput sets covering the genome, transcriptome, interactome and proteome and include full genome and proteome annotation, gene-pathway and Gene Ontology annotations, expression data including tissue- and disease-specific expression, proteome-wide and genetic interaction data, disease and phenotype data including Genome-Wide Association Studies, clinically significant SNPs and cis-eQTLs. A full list of data sources, including dates and versions, is available via the HumanMine home page (https://www.humanmine.org/humanmine/results/All%20data%20sources). Data are updated quarterly and we hope to include new data sources as they arise. The results reported here were generated using HumanMine version 12, released February 2022.



Filter by:

Active filters:

None

Category

899	Publication
16	Gene
8	InteractionExperiment
8	Protein
4	Exon
2	UniProtFeature
1	MeshTerm
1	ProteinDomain

Organism

21	H. sapiens
6	M. musculus
1	R. norvegicus

939 results for 'pax6'

Gene	Organism	*Rattus norvegicus*
	Symbol:	Pax6
	Identifiers:	RGD:3258
Gene	Organism	*Homo sapiens*
	Symbol:	PAX6
	Identifiers:	5080,ENSG00000007372
Gene	Organism	*Mus musculus*
	Symbol:	Pax6
	Identifiers:	MGI:97490
Protein	Organism:	*Mus musculus*
	Accession:	Q8VBZ1
	Identifiers:	Q8VBZ1_MOUSE
Gene	Organism	*Homo sapiens*
	Symbol:	
	Identifiers:	105941491
Gene	Organism	*Homo sapiens*
	Symbol:	
	Identifiers:	106014249
Protein	Organism:	*Mus musculus*

Figure 1. Keyword search. A search for 'Pax6' returns a filterable menu of data classes and organisms with the number of entities found in each on the left and individual entities displayed on the right. Each individual entity provides a link to its report page.

HumanMine access and help

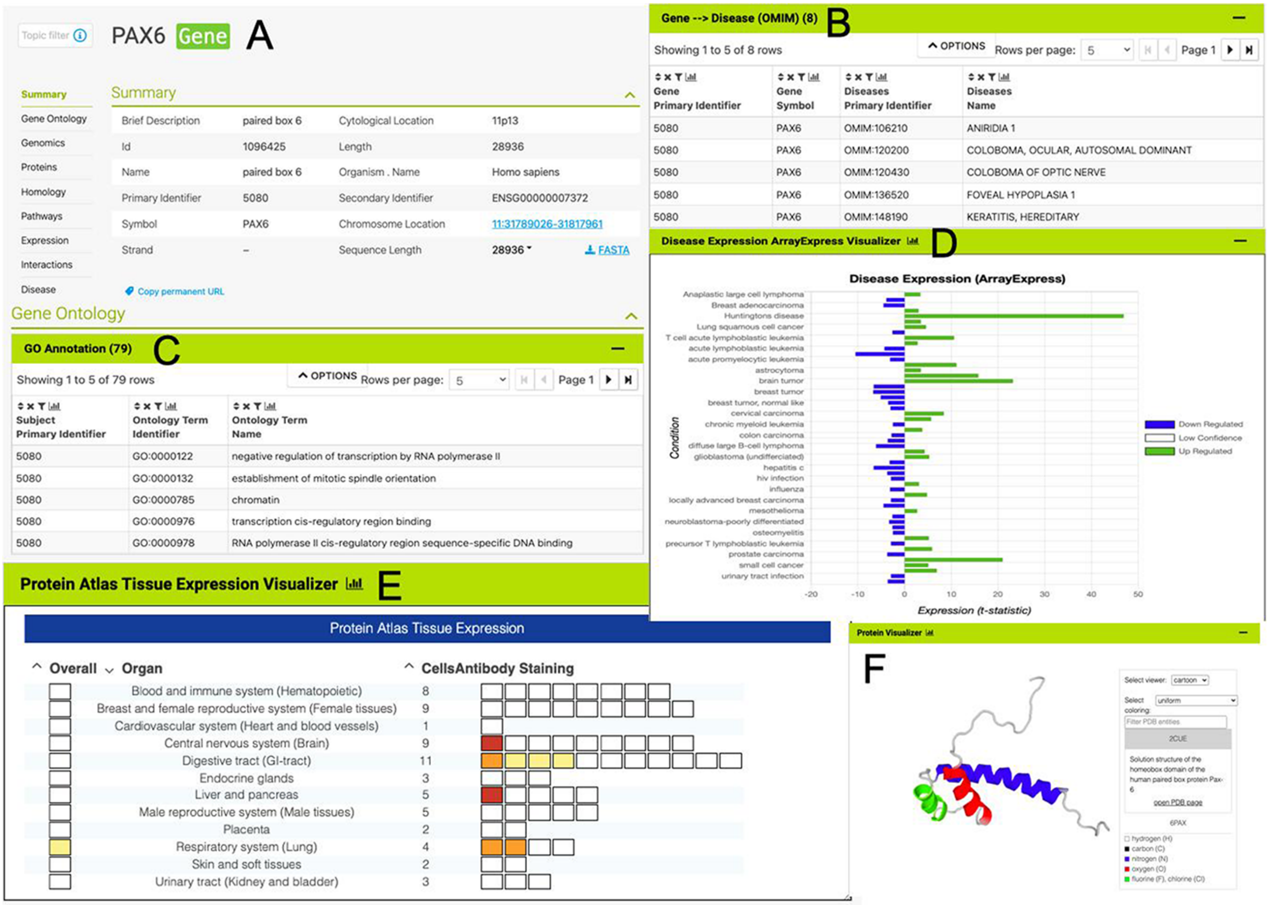
HumanMine is freely available for the public to use, but creating an account allows a user to save custom queries and lists permanently. If a user is not logged in then lists and queries are saved for the duration of the session.

Extensive help documentation is available to aid in the use of both the user interface and the web services together with a number of video tutorials (http://intermine.org/intermine-user-docs/; http://intermine.org/im-docs/docs/web-services/index/#api-and-client-libraries).

InterMine is an open source project and anyone with suitable computational expertise and computing resources can produce an in-house version of HumanMine using the relevant code repositories (humanmine (https://github.com/intermine/humanmine), humanmine-biosources (https://github.com/intermine/humanmine-bio-sources)). Recent work on building a software-as-a-service InterMine Cloud will make this type of process easier and available to non-programmers.

Exploratory data analysis: report pages and list analysis

The HumanMine interface provides a number of routes for exploring the integrated data, from keyword search and report pages to automated analysis of lists of data. For example, if a researcher wishes to explore data related to the *PAX6* gene they can use the simple keyword search. This returns links to genes, publications, interaction data, proteins, exons, UniProt protein features, mesh terms and a protein domain (Figure 1). The keyword search accepts any identifier type, such as a gene symbol or ontology term and keywords and allows use of operators like AND, OR and NOT between terms. All data types within the database are searched and the results returned are faceted by class, making it easier to find the data of interest, for instance, the *PAX6* related proteins from Homo sapiens. Following the *Homo sapiens PAX6* gene link presents a report page collating much information about its function including expression, phenotypes, sequence variants, disease and homology (Figure 2—the report page has



Brief Description	paired box 6	Cytological Location	11p13
Id	1096425	Length	28936
Name	paired box 6	Organism . Name	Homo sapiens
Primary Identifier	5080	Secondary Identifier	ENSG00000007372
Symbol	PAX6	Chromosome Location	11:31789026-31817961
Strand	–	Sequence Length	28936 FASTA

Gene Primary Identifier	Gene Symbol	Diseases Primary Identifier	Diseases Name
5080	PAX6	OMIM:106210	ANIRIDIA 1
5080	PAX6	OMIM:120200	COLOBOMA, OCULAR, AUTOSOMAL DOMINANT
5080	PAX6	OMIM:120430	COLOBOMA OF OPTIC NERVE
5080	PAX6	OMIM:136520	FOVEAL HYPOPLASIA 1
5080	PAX6	OMIM:148190	KERATITIS, HEREDITARY

Subject Primary Identifier	Ontology Term Identifier	Ontology Term Name
5080	GO:0000122	negative regulation of transcription by RNA polymerase II
5080	GO:0000132	establishment of mitotic spindle orientation
5080	GO:0000785	chromatin
5080	GO:0000976	transcription cis-regulatory region binding
5080	GO:0000978	RNA polymerase II cis-regulatory region sequence-specific DNA binding

Disease Expression ArrayExpress Visualizer
D
Disease Expression (ArrayExpress)
Anaplastic large cell lymphoma
Breast adenocarcinoma
Huntingtons disease
Lung squamous cell cancer
T cell acute lymphoblastic leukemia
acute lymphoblastic leukemia
acute promyelocytic leukemia
astrocytoma
brain tumor
breast tumor
breast tumor, normal like
cervical carcinoma
chronic myeloid leukemia
colon carcinoma
diffuse large B-cell lymphoma
glioblastoma (undifferciated)
hepatitis c
hiv infection
influenza
locally advanced breast carcinoma
mesothelioma
neuroblastoma-poorly differentiated
osteomyelitis
precursor T lymphoblastic leukemia
prostate carcinoma
small cell cancer
urinary tract infection
Condition
-20
-10
0
10
20
30
40
50
Expression (t-statistic)
Down Regulated
Low Confidence
Up Regulated
Protein Atlas Tissue Expression Visualizer
E
Protein Atlas Tissue Expression

^ Overall	⌄ Organ	^ CellsAntibody Staining	
□	Blood and immune system (Hematopoietic)	8	□□□□□□□□
□	Breast and female reproductive system (Female tissues)	9	□□□□□□□□□
□	Cardiovascular system (Heart and blood vessels)	1	□
□	Central nervous system (Brain)	9	🟥□□□□□□□□
□	Digestive tract (GI-tract)	11	🟧🟨🟨🟨□□□□□□□
□	Endocrine glands	3	□□□
□	Liver and pancreas	5	🟥□□□□
□	Male reproductive system (Male tissues)	5	□□□□□
□	Placenta	2	□□
🟨	Respiratory system (Lung)	4	🟧🟧□□
□	Skin and soft tissues	2	□□
□	Urinary tract (Kidney and bladder)	3	□□□

Solution structure of the homeobox domain of the human paired box protein Pax-6

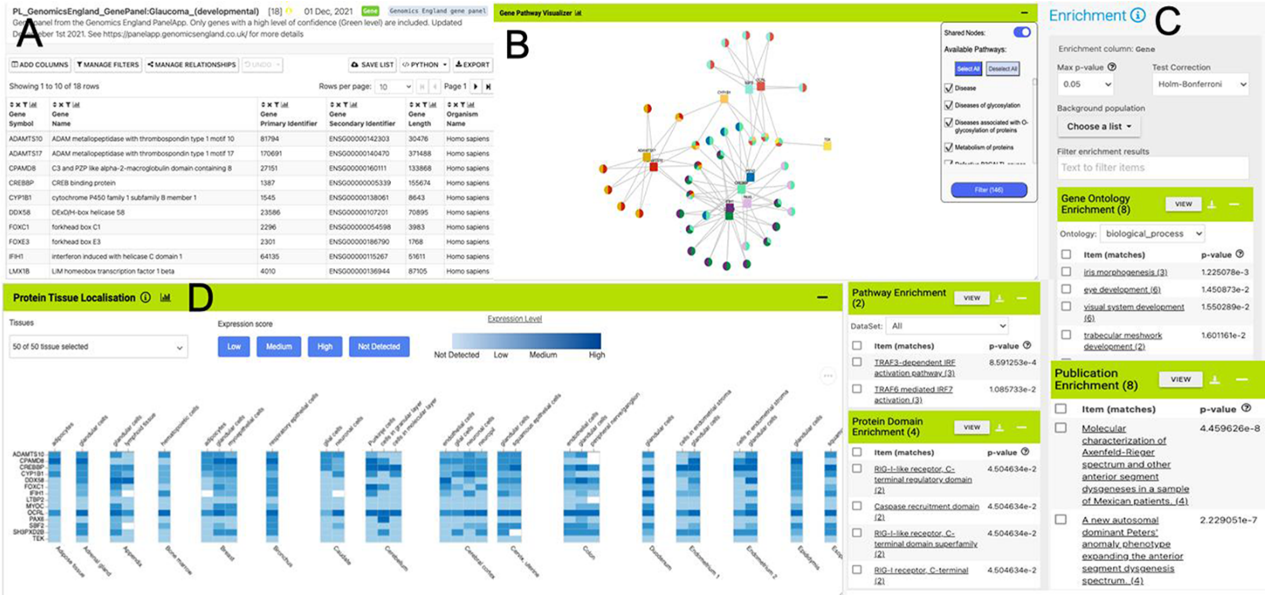
Figure 2. The PAX6 report page. Report pages present data through a range of interactive tables, graphs and visualizations depending on the data type. A selection of features from the report page for the human PAX6 gene are shown here. (A) A summary of the main identifiers and chromosomal location. (B). An interactive table of Gene Ontology annotations. Only the first five rows are shown. (C). A table of disease annotations (original data source: OMIM, https://www.omim.org). Only the first five rows are displayed. (D). A graph showing up- and downregulation of the PAX6 gene in various disease conditions (original data from ArrayExpress experiment E-MTAB-62, https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-62). (E). A graph showing protein localization data (from the Protein Atlas project, https://www.proteinatlas.org/humanproteome/tissue.). (F). A protein structure viewer pulling in data from the Protein Data Bank (https://www.rcsb.org).

the following permanent URL, formed using the NCBI identifier for the gene, 5080: https://www.humanmine.org/humanmine/gene:5080). Specifically, we learn that *PAX6* is expressed in the developing eye and brain, is involved in transcription and is associated with a number of eye-related disorders such as aniridia, keratitis and Peters Anomaly. Examination of gene expression and protein localization data shows us that *PAX6* is expressed not only in various brain structures but also in the pancreas and is upregulated in various cancers (Figure 2).

Within the report page for the *PAX6* gene, it is also possible to identify interesting gene sets ('lists') in which *PAX6* is found. For instance, *PAX6* is found in a number of Genomics England gene panels, highlighting its importance in many areas of development. Selecting the list 'PL_GenomicsEngland_GenePanel:Glaucoma_(developmental)', from the *PAX6* report page, allows us to explore aggregated data for this set of genes. Alongside a table detailing the individual members of the list, enrichment statistics show Gene Ontology terms for various visual and sensory organ development (such as iris morphogenesis (*P*-value: 1.20e-3) and eye development (*P*-value: 1.57e-2)) while enriched pathway terms include TRAF3-dependent IRF activation pathway (*P*-value: 7.83e-4). Widgets and tools then allow us to examine the chromosome distribution of the genes in the list, identify interacting partners through both a table and a network graph, examine heat maps of gene expression, visualize pathway interactions and Gene–GO term relationships (Figure 3).

The ability to create and use lists of entities is integral to data analysis within HumanMine. All lists can be viewed with such list analysis pages which, like the report pages, provide summarized data for the contents of the list. Lists of items with standardized annotations, such as Gene Ontology annotations, facilitate analysis by statistical enrichment methods. InterMine uses the hypergeometric distribution with a choice of multiple test corrections to provide enrichment statistics for lists of genes i.e. to discover whether an annotation is unexpectedly frequent within the group. Interactive tables of enrichment *P*-values, calculated for Gene Ontology terms, pathway annotations, protein domain annotations and publications are displayed on the gene list analysis pages. The tables can be adjusted for test correction, *P*-value threshold and background reference set (the default being all genes in the human genome with an annotation of the type being analysed, e.g. all human genes that have associated GO annotations) and subsets of the genes with a particular enriched term can easily be viewed and saved as further lists. In addition to the enrichment tables, a tools library allows for the easy addition of visualizations to the list pages by the database maintainer.

Lists created within HumanMine or from an external source can be saved into the user's account for further use. In addition, HumanMine provides a number of public lists (sets of interesting genes from relevant publications or resources) such as the sets of rare-disease and cancer genes defined by the Genomics England Gene PanelApp project (https://panelapp.genomicsengland.co.uk, (24)) (only 'Green' level panels are included, which have a high level of evidence for the gene–disease association). Such public lists have a number of applications including identifying whether a gene of interest is present in any of the lists (shown on gene report pages), notifying users of significant overlaps between their lists and the public lists (currently available through the web services but soon to be developed for the user interface) and simply as a convenient list of genes known to be significantly involved in a particular disease.



PL_GenomicsEngland_GenePanel:Glaucoma_(developmental) [18] 01 Dec, 2021 Gene Genomics England gene panel

Gene panel from the Genomics England PanelApp. Only genes with a high level of confidence (Green level) are included. Updated December 1st 2021. See https://panelapp.genomicsengland.co.uk/ for more details

ADD COLUMNS | MANAGE FILTERS | MANAGE RELATIONSHIPS | UNDO | SAVE LIST | PYTHON | EXPORT

Showing 1 to 10 of 18 rows | Rows per page: 10 | Page 1

Gene Symbol	Gene Name	Gene Primary Identifier	Gene Secondary Identifier	Gene Length	Organism Name
ADAMTS10	ADAM metallopeptidase with thrombospondin type 1 motif 10	81794	ENSG00000142303	30476	Homo sapiens
ADAMTS17	ADAM metallopeptidase with thrombospondin type 1 motif 17	170691	ENSG00000140470	371488	Homo sapiens
CPAMD8	C3 and PZP like alpha-2-macroglobulin domain containing 8	27151	ENSG00000160111	133868	Homo sapiens
CREBBP	CREB binding protein	1387	ENSG00000005339	155674	Homo sapiens
CYP1B1	cytochrome P450 family 1 subfamily B member 1	1545	ENSG00000138061	8643	Homo sapiens
DDX58	DExD/H-box helicase 58	23586	ENSG00000107201	70895	Homo sapiens
FOXC1	forkhead box C1	2296	ENSG00000054598	3983	Homo sapiens
FOXE3	forkhead box E3	2301	ENSG00000186790	1768	Homo sapiens
IFIH1	interferon induced with helicase C domain 1	64135	ENSG00000115267	51611	Homo sapiens
LMX1B	LIM homeobox transcription factor 1 beta	4010	ENSG00000136944	87105	Homo sapiens

Enrichment

Enrichment column: Gene

Max p-value: 0.05

Test Correction: Holm-Bonferroni

Background population: Choose a list

Filter enrichment results: Text to filter items

Gene Ontology Enrichment (8)

Ontology: biological_process

Item (matches)	p-value
iris morphogenesis (3)	1.225078e-3
eye development (6)	1.450873e-2
visual system development (6)	1.550289e-2
trabecular meshwork development (2)	1.601161e-2

Pathway Enrichment (2)

DataSet: All

Item (matches)	p-value
TRAF3-dependent IRF activation pathway (3)	8.591253e-4
TRAF6 mediated IRF7 activation (3)	1.085733e-2

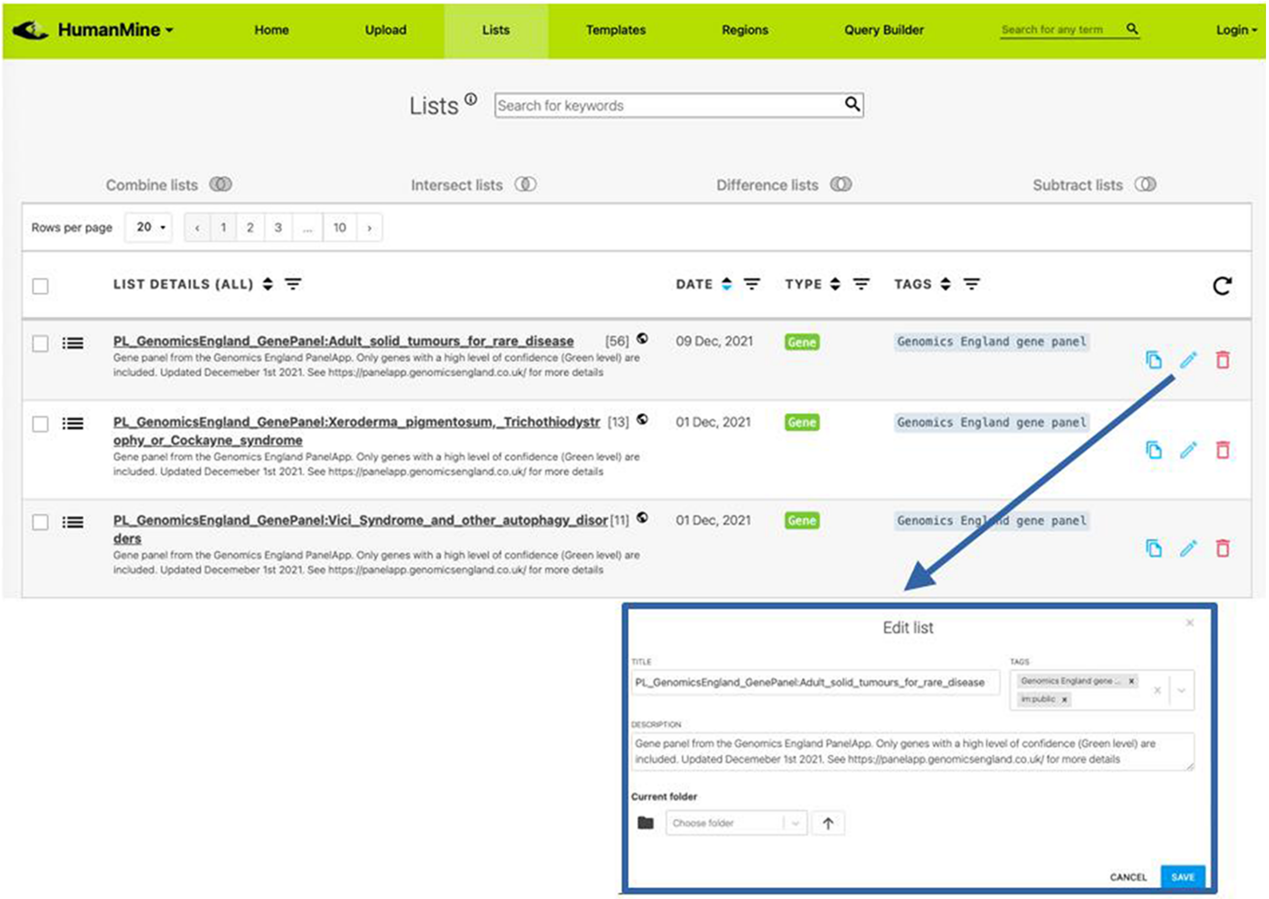
Protein Domain Enrichment (4)

Item (matches)	p-value
RIG-I-like receptor, C-terminal regulatory domain (2)	4.504634e-2
Caspase recruitment domain (2)	4.504634e-2
RIG-I-like receptor, C-terminal domain superfamily (2)	4.504634e-2
RIG-I receptor, C-terminal (2)	4.504634e-2

Publication Enrichment (8)

Item (matches)	p-value
Molecular characterization of Axenfeld-Rieger spectrum and other anterior segment dysgeneses in a sample of Mexican patients. (4)	4.459626e-8
A new autosomal dominant Peters' anomaly phenotype expanding the anterior segment dysgenesis spectrum. (4)	2.229051e-7

Figure 3. List analysis page for the public list: PL_GenomicsEngland_GenePanel:Glaucoma_(developmental). Like the report pages, list analysis pages provide a number of interactive tables and graphs. A selection is shown here. (A). Interactive table summarizing the contents of the list. (B) A network graph showing Gene–Pathway connections for genes in the list. Only genes that have two or more pathway connections are shown (this option can be toggled on the menu panel). The menu panel also allows filtering of the pathway annotations used in the graph. (C) Enrichment statistics for Gene Ontology terms, Publications, Protein domains and Pathway annotations. (D) A heat map showing protein localization for each gene in the list (original data from The Protein Atlas project https://www.proteinatlas.org/humanproteome/tissue).



□	LIST DETAILS (ALL)	DATE	TYPE	TAGS
□	PL_GenomicsEngland_GenePanel:Adult_solid_tumours_for_rare_disease [56] Gene panel from the Genomics England PanelApp. Only genes with a high level of confidence (Green level) are included. Updated Decemeber 1st 2021. See https://panelapp.genomicsengland.co.uk/ for more details	09 Dec, 2021	Gene	Genomics England gene panel
□	PL_GenomicsEngland_GenePanel:Xeroderma_pigmentosum,_Trichothiodystrophy_or_Cockayne_syndrome [13] Gene panel from the Genomics England PanelApp. Only genes with a high level of confidence (Green level) are included. Updated Decemeber 1st 2021. See https://panelapp.genomicsengland.co.uk/ for more details	01 Dec, 2021	Gene	Genomics England gene panel
□	PL_GenomicsEngland_GenePanel:Vici_Syndrome_and_other_autophagy_disorders [11] Gene panel from the Genomics England PanelApp. Only genes with a high level of confidence (Green level) are included. Updated Decemeber 1st 2021. See https://panelapp.genomicsengland.co.uk/ for more details	01 Dec, 2021	Gene	Genomics England gene panel

Edit list

TITLE: PL_GenomicsEngland_GenePanel:Adult_solid_tumours_for_rare_disease

TAGS: Genomics England gene ... ×, im:public ×

DESCRIPTION: Gene panel from the Genomics England PanelApp. Only genes with a high level of confidence (Green level) are included. Updated Decemeber 1st 2021. See https://panelapp.genomicsengland.co.uk/ for more details

Current folder: Choose folder

Figure 4. Lists. All lists, both public and private for a user account, can be viewed under the Lists tab.

All lists, both public and belonging to a user, can be viewed on the 'Lists' page, which has a number of functions for the organization and exploration of the lists available, including the addition of free-text descriptive annotation as 'tags', creation of folders, as well as search and filtering functions (Figure 4). Two or more lists can be combined using set operations: union, intersection, difference and subtraction.

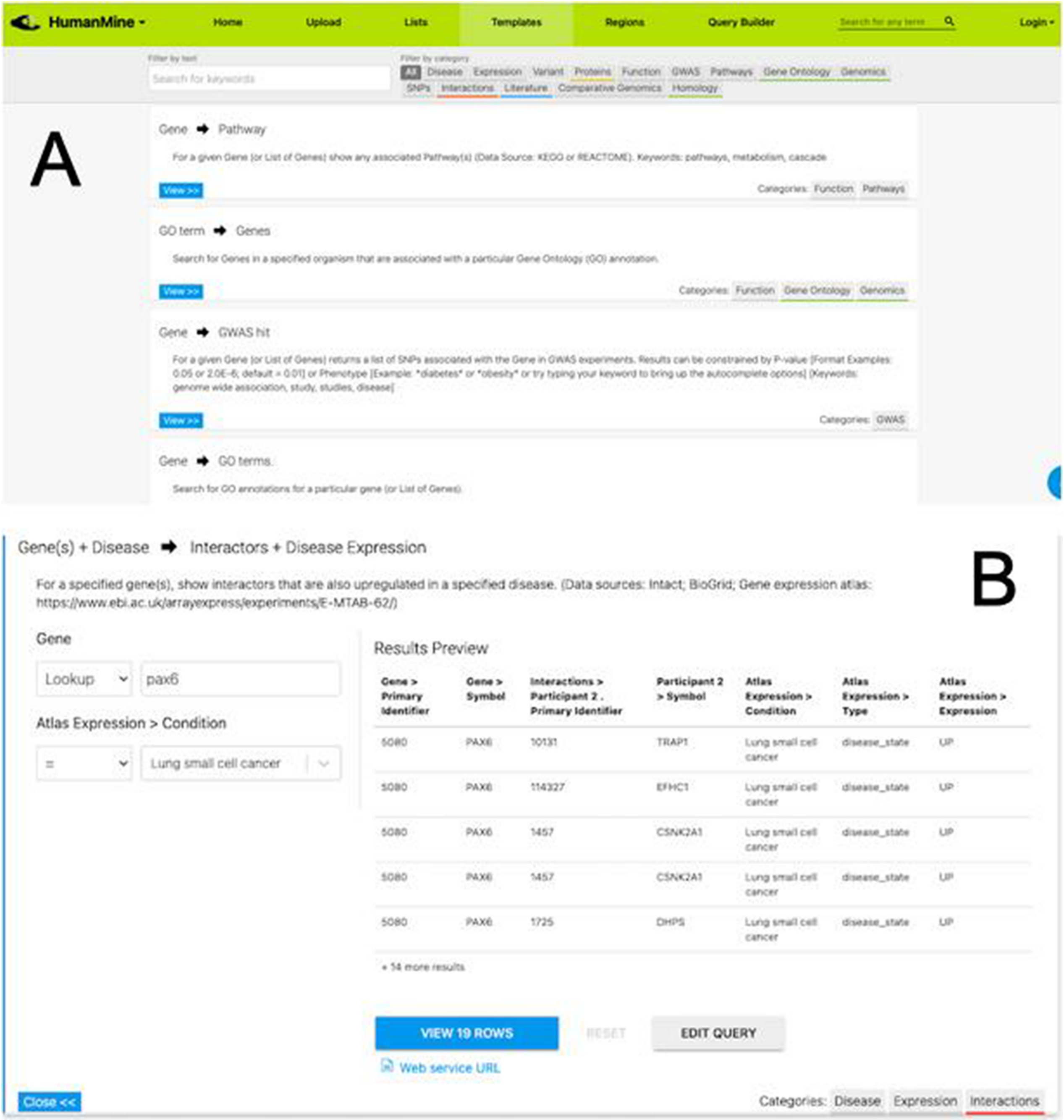
Structured data analysis: template searches and the query builder

After exploring the report pages for a particular entity, a researcher may wish to ask more structured or focussed questions about the data. Continuing with our *PAX6* example, we noted earlier that *PAX6* is upregulated in a number of cancers, including small cell lung cancer. To investigate this further a researcher may follow a number of lines of investigation. For example, to (i) identify other genes upregulated in small cell lung cancer and (ii) identify whether any of these genes interact with *PAX6*. A library of 'canned' or pre-formed searches, called template searches, allow the user to easily search across any integrated data set (Figure 5).

For example, the above two questions about *PAX6* could be investigated with a single template search, 'Gene(s) + Disease –> Interactors + Disease Expression'. This template combines disease expression data (ArrayExpress experiment E-MTAB-62 (25)) and interaction data (IntAct (26) and BioGrid (27)) and allows us to search for genes that interact with *PAX6* that are, according to the data, also upregulated in a defined disease state, here lung cancer.

A template search typically has one or more 'start' data types as input and one or more 'end' data types that are provided as output and allows the user to add constraints (or filters) to these to control the subset of data returned. Templates cover both simple searches and more complex searches combining two or more datasets, and in most cases can be executed with either a single item (e.g. a gene) or a list of items (such as a pre-saved gene list).

More advanced data mining is available through the query builder, which can be used to generate *de novo* queries or to modify template searches. For instance, the template 'Gene(s) + Disease –> Interactors + Disease Expression' can be displayed in the query builder and can be modified to return genes that interact with *PAX6* that are expressed in lung squamous cell cancer as well as small cell lung cancer (Figure 6). The query builder interface displays a hierarchical data model browser through which classes of data can be selected for combining into a search. Constraints that filter the data and the results that are provided as output can be defined through the interface together with more advanced features such as constraint logic (data types are combined with an 'AND' by default but can be configured for OR, see Figure 6) and data join type, i.e. whether the intersection of two datasets should be returned or the union of the datasets (the latter being returned by default). For instance, in our example, it is possible to return either all interacting genes regardless of whether they also show expression in the defined conditions or only return those interactors for which the expression conditions are satisfied. Template searches and the Query Builder present results as InterMine Results Tables. From here, we can, for example, use a column summary function to find that (i) 18 interacting genes are returned and (ii) that, using the column summary to filter the table, 14 of the genes are associated with small cell lung cancer and 12 with lung squamous cell cancer (Figure 7). The 'Save as list' function of the results table allows us to save the set of 14 upregulated small cell lung cancer genes that interact with *PAX6* enabling this set to be further analysed through the list analysis page, further template searches and set operations on lists. The Gene Pathway Visualizer on the list analysis page, for example, highlights a number of potentially important cancer genes including SOX2, CSNK2A1, LMX2, SMAD5, APP and TP73 (Figure 8). The intersection of these genes with a set known to be involved in small cell lung cancer from the DisGenNet (28) dataset



Gene ➡ Pathway

For a given Gene (or List of Genes) show any associated Pathway(s) (Data Source: KEGG or REACTOME). Keywords: pathways, metabolism, cascade

GO term ➡ Genes

Search for Genes in a specified organism that are associated with a particular Gene Ontology (GO) annotation.

Gene ➡ GWAS hit

For a given Gene (or List of Genes) returns a list of SNPs associated with the Gene in GWAS experiments. Results can be constrained by P-value (Format Examples: 0.05 or 2.0E-6; default = 0.01) or Phenotype [Example: *diabetes* or *obesity* or try typing your keyword to bring up the autocomplete options] (Keywords: genome wide association, study, studies, disease)

Gene ➡ GO terms.

Search for GO annotations for a particular gene (or List of Genes).

Gene(s) + Disease ➡ Interactors + Disease Expression

For a specified gene(s), show interactors that are also upregulated in a specified disease. (Data sources: Intact; BioGrid; Gene expression atlas: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-62/)

Gene: Lookup pax6

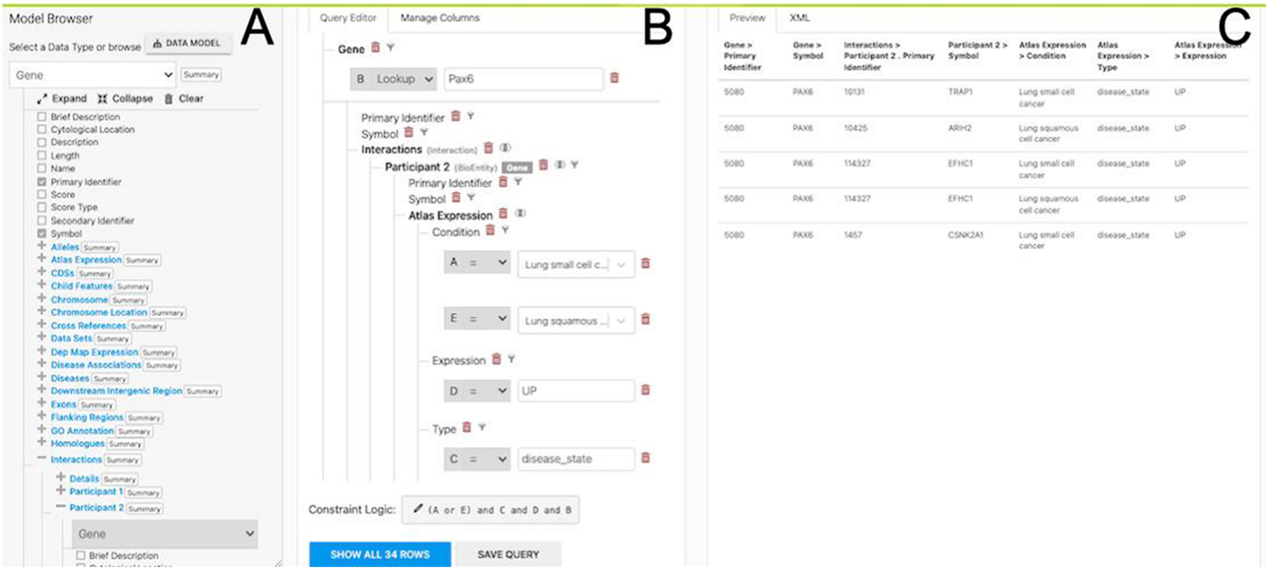
Atlas Expression > Condition: = Lung small cell cancer

Results Preview

Gene > Primary Identifier	Gene > Symbol	Interactions > Participant 2 . Primary Identifier	Participant 2 > Symbol	Atlas Expression > Condition	Atlas Expression > Type	Atlas Expression > Expression
5080	PAX6	10131	TRAP1	Lung small cell cancer	disease_state	UP
5080	PAX6	114327	EFHC1	Lung small cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP
5080	PAX6	1725	DHPS	Lung small cell cancer	disease_state	UP

+ 14 more results

Figure 5. (A) A library of 'template' searches is available from the 'Templates' tab. The template library can be searched using keywords or filtered using the various data category tags. (B) The 'Gene(s) + Disease Interactors + Disease Expression' template expanded. Each template provides one or more constraints that can be modified according to the search the user wishes to run. A preview of the template result is shown with options to view the full results or edit the query in the Query Builder.



Gene: B Lookup Pax6

Condition: A = Lung small cell c...

E = Lung squamous ...

Expression: D = UP

Type: C = disease_state

Constraint Logic: (A or E) and C and D and B

Gene > Primary Identifier	Gene > Symbol	Interactions > Participant 2 . Primary Identifier	Participant 2 > Symbol	Atlas Expression > Condition	Atlas Expression > Type	Atlas Expression > Expression
5080	PAX6	10131	TRAP1	Lung small cell cancer	disease_state	UP
5080	PAX6	10425	ARIH2	Lung squamous cell cancer	disease_state	UP
5080	PAX6	114327	EFHC1	Lung small cell cancer	disease_state	UP
5080	PAX6	114327	EFHC1	Lung squamous cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP

Figure 6. The Query Builder. The query builder consists of three main panels—the model Browser (A), the query editor (B) and the query preview (C). The 'Gene(s) + Disease Interactors + Disease Expression' search is shown with an extra constraint for 'condition = Lung squamous cell cancer' added. Each constraint in the query editor is labelled with a letter enabling the constraint logic to be edited here to give ((A or E) and C and D and B).

confirms that two, SOX2 and TP73 are known to be directly implicated in small cell lung cancers, suggesting that the rest of the list may also be worthy of further investigation in the context of lung cancer.

Thus the Results Table is a powerful query analysis tool in itself, providing many features for the manipulation of the results, including the column summaries (simple statistics describing the data in a column including the number of distinct attributes for categorical data, or, for numerical data, the mean, median, minimum and maximum), data filtering, sorting, list generation and the ability to add additional columns of data (Figure 7). The results tables can be exported in common formats including TAB/CSV and as a frictionless data package (https://frictionlessdata.io). Nucleotide and protein sequences can be exported in FASTA format. An additional feature of the results tables in the new BlueGenes interface is the automatic generation of enrichment statistics and the inclusion of analysis tools as seen on the list analysis pages (Figure 3). The enrichment is based on the first column of genes or proteins, however, if more than one column of genes or proteins is present in the table, another column can be selected for the enrichment calculations from the settings bar.

Using HumanMine to explore external data

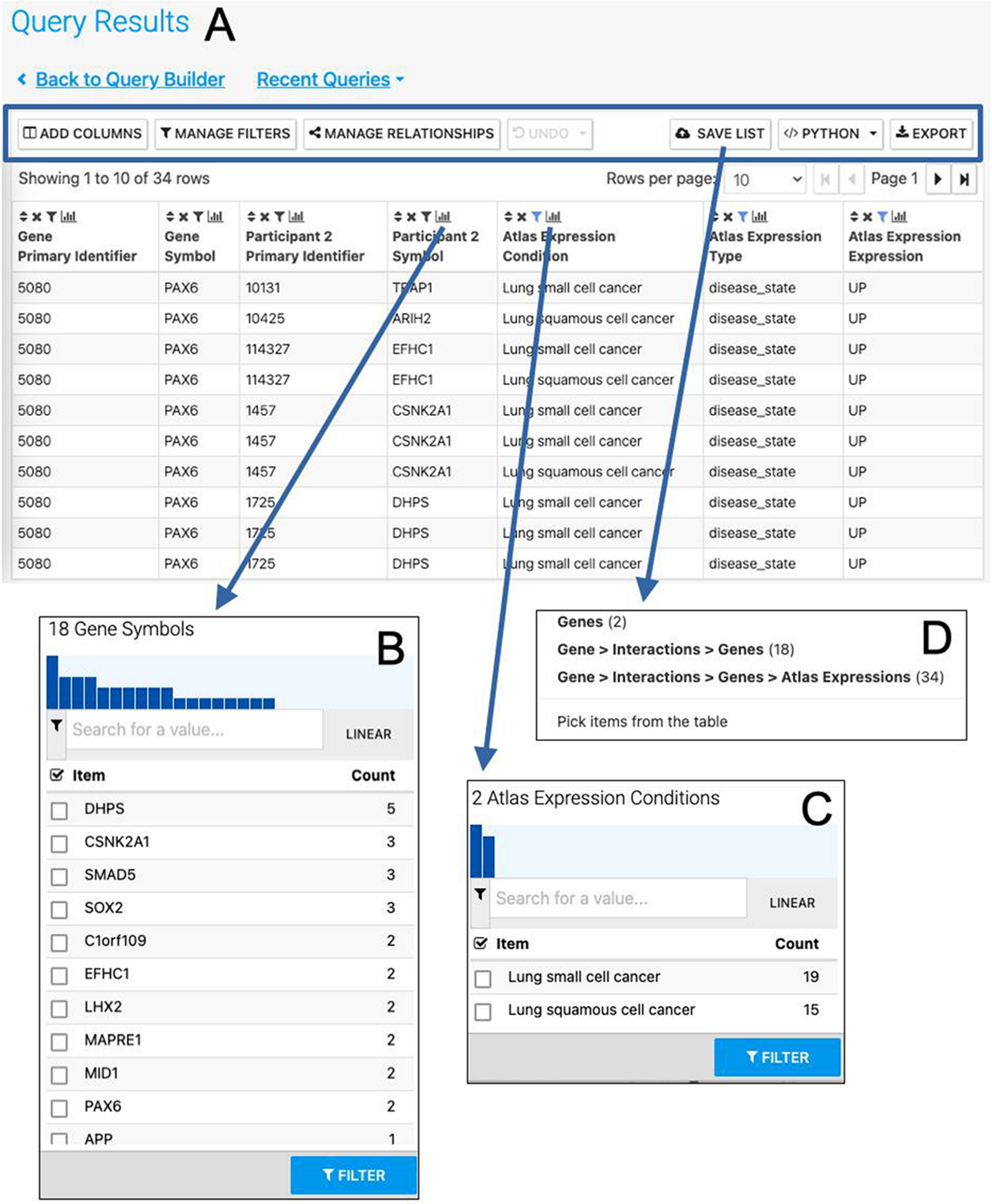
We learnt from the GO annotations on the report page that *PAX6* is a transcription factor involved in many cellular functions. It follows, therefore, that a researcher may wish to analyse genes that are regulated by *PAX6*. HumanMine provides two interfaces for the upload of external data, a list upload and a region search. For instance, a set of human genes already identified as *PAX6* target genes (29) can be uploaded to HumanMine (HumanMine public list PL_Pax6_Targets) using the list upload function enabling them to be analysed alongside the integrated data using any of the functionality discussed above. The template 'Gene → Gene Expression (Tissue, Disease; Array Express, E-MTAB-62)' can be run using this list as input and with 'Atlas Expression > type' set to disease_state and 'Condition' set to 'Contains' Lung. The resulting results table can be further filtered using the Atlas Expression Condition column summary to show data for just 'small cell lung cancer' and 'Lung small cell cancer' so identifying 51 upregulated genes, five of which (ASCL1, EPHA7, ID2, RUNXITI and SOX2) already have a known involvement in small cell lung cancer according to the DisGenNet data, with the others, therefore, providing a set of candidate genes for further research.

HumanMine allows users to upload any lists of objects for which data classes exist (for instance genes, proteins, transcripts). The list upload system includes a powerful identifier resolution system such that any outdated identifiers or synonyms can be updated to the current human annotation. For genes, HumanMine supports NCBI Entrez (30) gene identifiers, Ensembl (31) gene identifiers and HGNC (32) identifiers, with the main genome annotation being from NCBI. Thus, once a list has been uploaded it can be analysed in the same way as a list generated from within the database and will benefit from an automatically generated list analysis page.

The Region Search tool makes it possible to identify, for specified genomic spans, genomic sequence features such as genes and regulatory regions that overlap the spans. Thus the list of *PAX6* target genes could equally have been created by using a published set of ChIP-seq binding intervals (29) and searching for overlapping genes. The region search results interface allows such a set of genes to be easily saved as a list. The genomic regions can be extended by specified amounts beyond the 5′ and 3′ coordinates given, and a strand-specific search can be carried out if required. Features identified can be saved as lists or exported, either individually for each region specified or for all regions searched.

Web services

In addition to the user interface, HumanMine can be accessed through a comprehensive web services layer (12) through



Showing 1 to 10 of 34 rows

Gene Primary Identifier	Gene Symbol	Participant 2 Primary Identifier	Participant 2 Symbol	Atlas Expression Condition	Atlas Expression Type	Atlas Expression Expression
5080	PAX6	10131	TRAP1	Lung small cell cancer	disease_state	UP
5080	PAX6	10425	ARIH2	Lung squamous cell cancer	disease_state	UP
5080	PAX6	114327	EFHC1	Lung small cell cancer	disease_state	UP
5080	PAX6	114327	EFHC1	Lung squamous cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung squamous cell cancer	disease_state	UP
5080	PAX6	1725	DHPS	Lung small cell cancer	disease_state	UP
5080	PAX6	1725	DHPS	Lung small cell cancer	disease_state	UP
5080	PAX6	1725	DHPS	Lung small cell cancer	disease_state	UP

18 Gene Symbols

☑ Item	Count
□ DHPS	5
□ CSNK2A1	3
□ SMAD5	3
□ SOX2	3
□ C1orf109	2
□ EFHC1	2
□ LHX2	2
□ MAPRE1	2
□ MID1	2
□ PAX6	2
□ APP	1

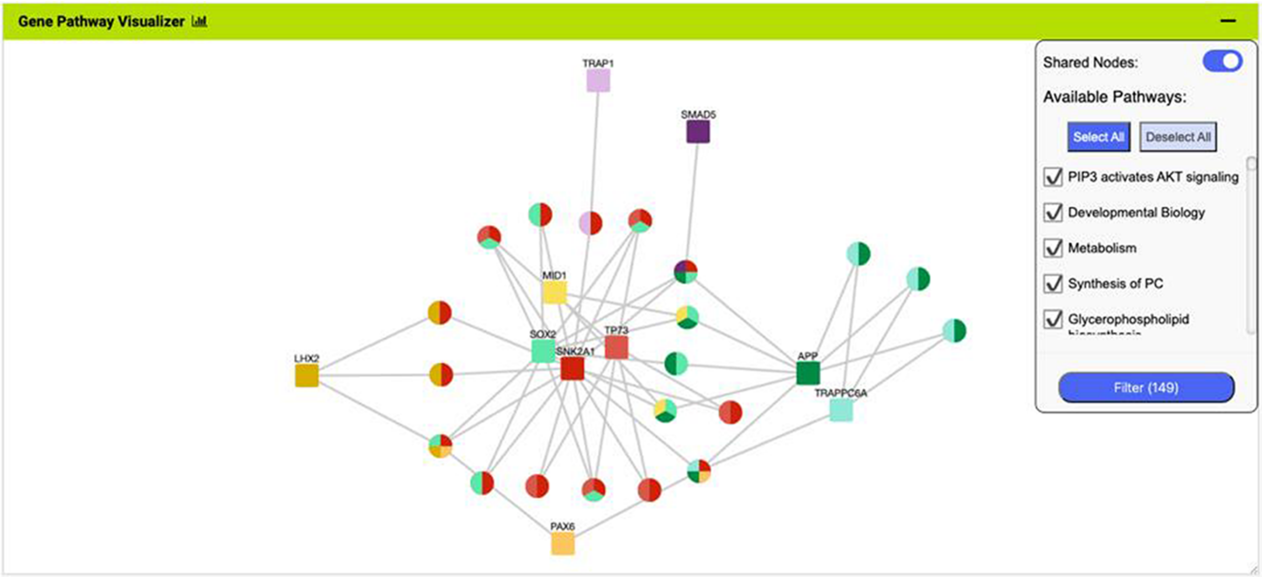
Genes (2)
Gene > Interactions > Genes (18)
Gene > Interactions > Genes > Atlas Expressions (34)
Pick items from the table

2 Atlas Expression Conditions

☑ Item	Count
□ Lung small cell cancer	19
□ Lung squamous cell cancer	15

Figure 7. The Results Table showing results from the query 'Gene(s) + Disease Interactors + Disease Expression' for the PAX6 gene with constraints on the disease name for small cell lung cancer or lung squamous cell cancer. The results tables provide many additional functions including 'Add columns' allowing additional data to be added, 'Manage filters' allowing filters on any column to be defined, 'Manage relationships' enabling either the union or intersect of classes of related data in the table to be defined, 'Save list' enabling subsets of items in the table to be saved as lists, 'Python', automatic code generation, available as a drop-down list of available languages and 'Export' (A). The column summary on the Participant 2 > Symbol column (the genes with which PAX6 interacts) allows the number of unique interacting genes (18) to be found (B). Using the column summary on the Atlas Expression > Condition column it is possible to see the number of rows for each disease condition. This could be used to filter the table to show just one of the disease conditions (C). The 'Save as list' function can be used to save any set of items from the table. Here it may be useful to save the set of interacting genes (Gene > Interactions > Genes (18)) (D). To save the set of interacting genes specific to one of the cancers, the table could be filtered first using the column summary function as described above.

which all the functionality of the user interface can be accessed. Client libraries in several languages are available (http://intermine.readthedocs.io/en/latest/web-services/#api-and-client-libraries) or RESTful endpoints may be accessed directly (http://iodocs.apps.intermine.org/). Automatic code generation through the user interface provides an easy starting point for those wishing to use the web services. For example, code for running the template mentioned earlier,



Shared Nodes: ●

Available Pathways:

Select All | Deselect All

☑ PIP3 activates AKT signaling

☑ Developmental Biology

☑ Metabolism

☑ Synthesis of PC

☑ Glycerophospholipid

Filter (149)

Figure 8. The Gene Pathway Visualizer on the list analysis page, showing Gene–Pathway associations for the 14 upregulated small cell lung cancer genes that interact with PAX6. A number of potentially important cancer genes including SOX2, CSNK2A1, LMX2, SMAD5, APP and TP73 can be seen. Only genes that have two or more pathway connections are shown (this option can be toggled on the menu panel).

'Gene → Gene Expression (Tissue, Disease; Array Express, E-MTAB-62)', can be generated directly from the template form (Figure 9). This code can then be added to a script that has imported the corresponding client library, so facilitating script-based analysis in which data from InterMine is dynamically retrieved. In addition, an R library, InteMineR (33) available from BioConductor (https://www.bioconductor.org/packages/release/bioc/html/InterMineR.html), allows queries against HumanMine to be run using the R programming language.

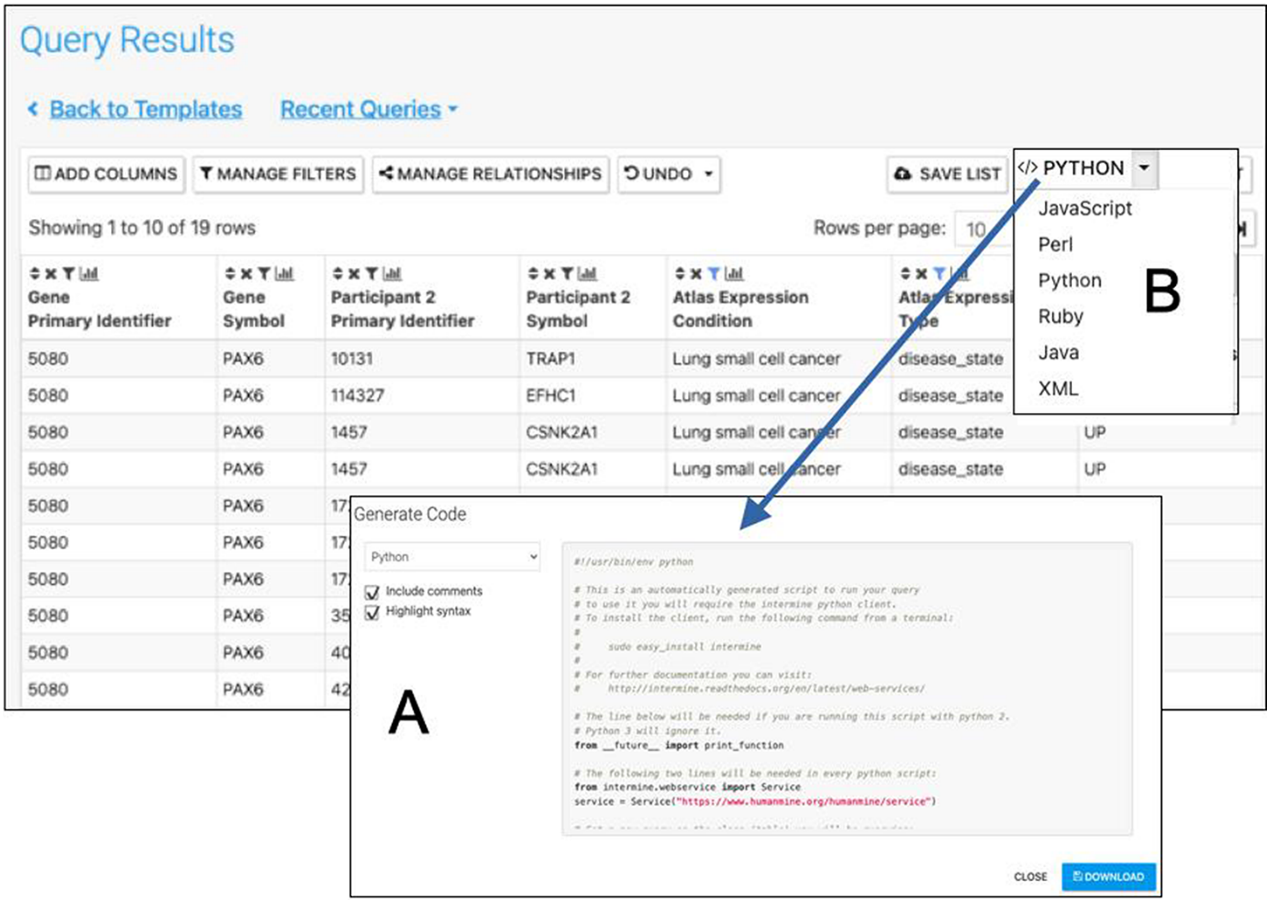
Use-case: using HumanMine to explore shared pathways in disease comorbidities

The combination of integrated data and analysis tools in HumanMine allows for the exploration of shared genes and pathways between two or more conditions. Asthma and type 2 diabetes are two of the most prevalent chronic diseases, presenting significant challenges to public health. While patients frequently only have one of either asthma or type 2 diabetes, the two diseases also have a well-documented but poorly understood comorbidity, in which they co-occur more frequently than would be expected by chance (34). Studying disease comorbidities allows both a better understanding of the pathology of the respective diseases, as well as identification of the underlying mechanisms of the comorbidity, which may provide opportunities for more targeted interventions and therapies.

For this example, data from multiple disease sources (OMIM (35), ClinVar (36), GWAS Catalogue (37), DisGenNet (28) and the Human Phenotype Ontology, HPO (38)) were searched for genes associated with either Type 2 diabetes or Asthma. The resulting lists of genes for each disease were combined by using the list intersection function to identify genes shared between the two diseases. Analysis of the resulting shared set of 912 genes (from now on called the shared set, Supplementary Table S2) provides a number of insights into the underlying causes of the comorbidity, some of which are discussed below (Figure 10). The enriched Gene Ontology terms and Pathway annotations for the shared set identify a number of underlying shared biological functions between the two conditions, in particular in the immune system, signal transduction and haemostasis (Supplementary Table S3A and B).

There is evidence to suggest that asthma may increase the risk of developing type 2 diabetes (39). The mechanisms by which this increased risk may occur, however, are not fully understood. It has been suggested that increased levels of some inflammatory cytokines known to be present in asthma may contribute to the development of type 2 diabetes (39). This proposal is backed up here by the finding that 249 genes from the 'shared set' are annotated with the high-level GO term 'Inflammatory response' and the enrichment of a number of pathways involved in inflammation, particularly the interleukin signalling pathways. The proinflammatory biomarkers interleukin-6 (IL6), C-reactive protein (CRP) and TNF-alpha (TNF), present in the 'shared set' and known to be elevated in low-grade inflammation, are thought to be major contributors to the development of type 2 diabetes (40, 41). Thus, it is possible that increased circulating levels of some inflammatory cytokines known to be present in asthma may also contribute to the development of type 2 diabetes.

Another important immune gene, TGFB1, present in the shared list, encodes the cytokine TGF-beta 1. Its roles in airway inflammation and remodelling in asthma are well-defined (42), and asthma severity is associated with different TGFB1 polymorphisms in humans (43). The role of TGF-beta 1 in type 2 diabetes is less established, however, an interesting hypothesis is that it drives dedifferentiation of beta cells (44), a process implicated in type 2 diabetes pathogenesis (45); dedifferentiation is also part of asthmatic airway remodelling (46). Targeting TGF-beta 1 remains challenging (47) but could offer new therapeutic potential for both diseases.



Query Results

Showing 1 to 10 of 19 rows

Rows per page: 10

Gene Primary Identifier	Gene Symbol	Participant 2 Primary Identifier	Participant 2 Symbol	Atlas Expression Condition	Atlas Expressi Type	
5080	PAX6	10131	TRAP1	Lung small cell cancer	disease_state	
5080	PAX6	114327	EFHC1	Lung small cell cancer	disease_state	
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP
5080	PAX6	1457	CSNK2A1	Lung small cell cancer	disease_state	UP
5080	PAX6	17				
5080	PAX6	17				
5080	PAX6	17				
5080	PAX6	35				
5080	PAX6	40				
5080	PAX6	42				

Generate Code

Python

☑ Include comments

☑ Highlight syntax

```
#!/usr/bin/env python

# This is an automatically generated script to run your query
# to use it you will require the intermine python client.
# To install the client, run the following command from a terminal:
#
#     sudo easy_install intermine
#
# For further documentation you can visit:
#     http://intermine.readthedocs.org/en/latest/web-services/

# The line below will be needed if you are running this script with python 2.
# Python 3 will ignore it.
from __future__ import print_function

# The following two lines will be needed in every python script:
from intermine.webservice import Service
service = Service("https://www.humanmine.org/humanmine/service")
```

Figure 9. Automatic code generation. From any results table, it is possible to view and copy code for the underlying query in various programming languages. Here the python code for the Gene(s) + Disease Interactors + Disease Expression result is shown (A). Code for Python, Perl, Ruby, Javascript and Java is available (B).

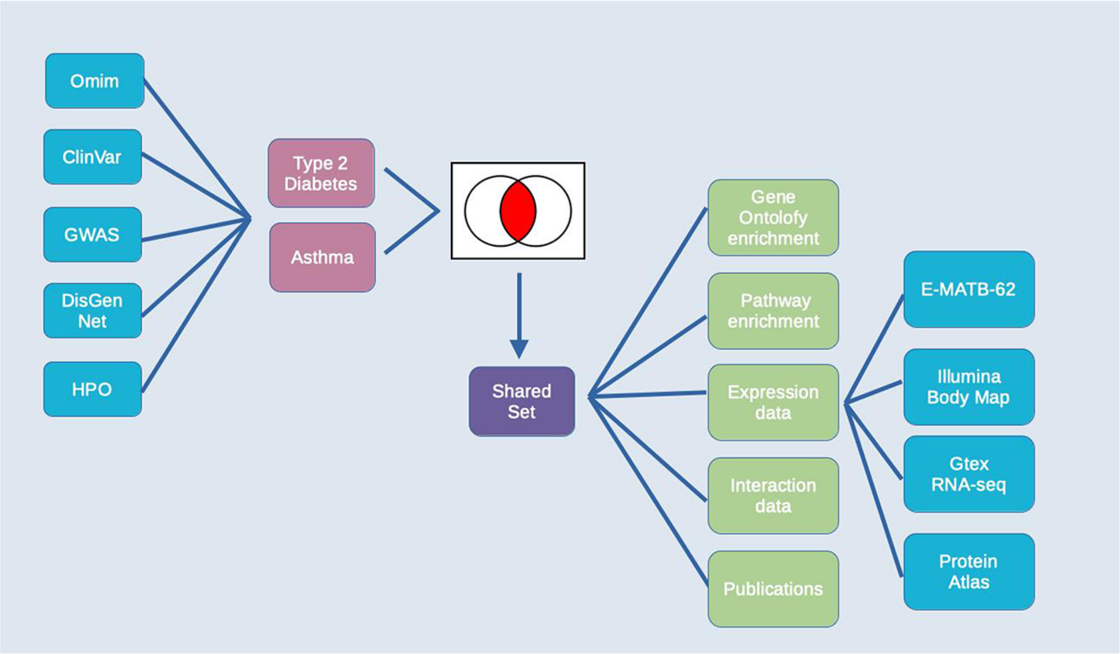


Figure 10. Exploring comorbidities using HumanMine. A schematic representation of the steps involved in the use-case 'Using HumanMine to explore shared pathways in disease comorbidities'. See text for details.

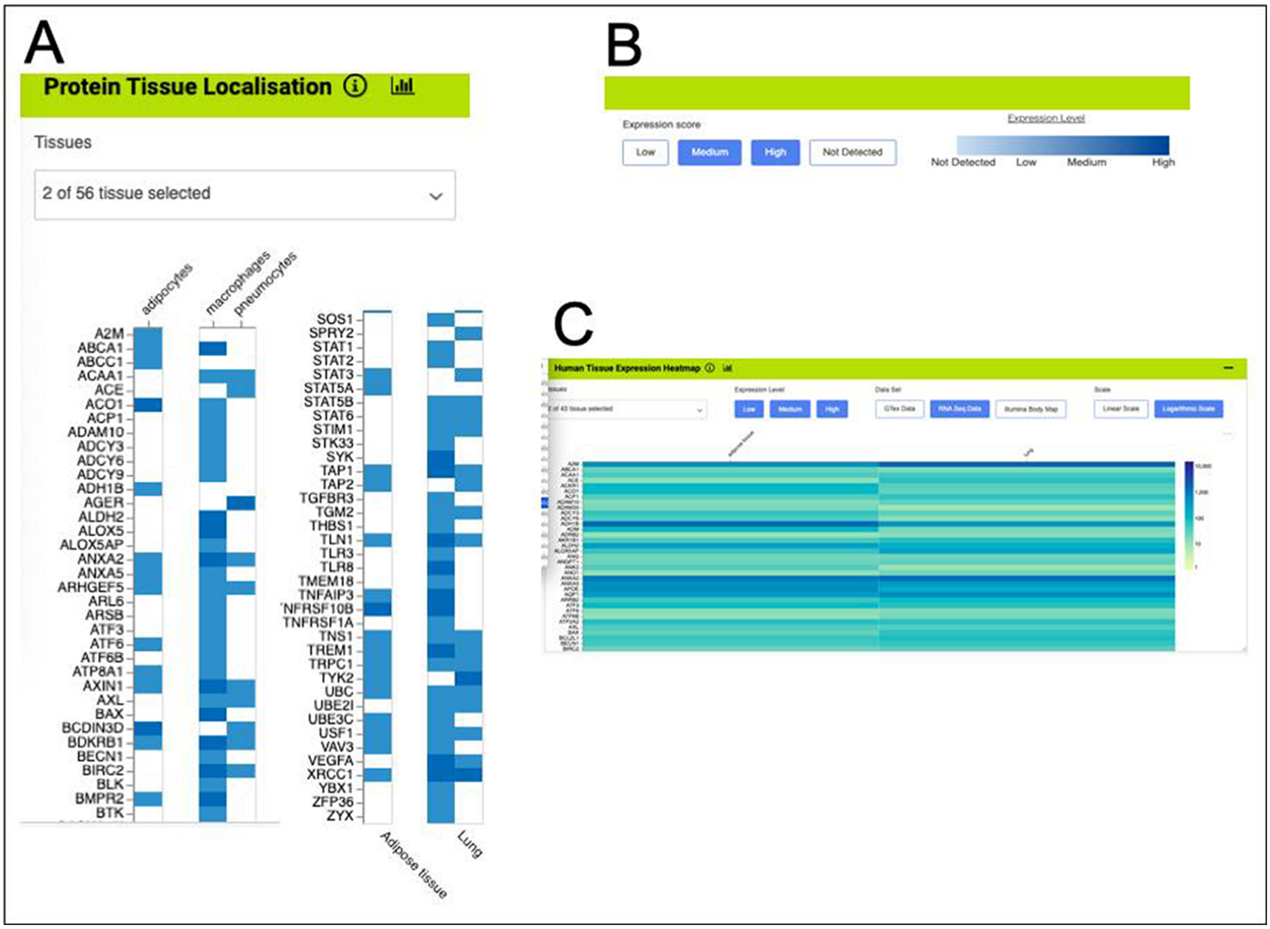
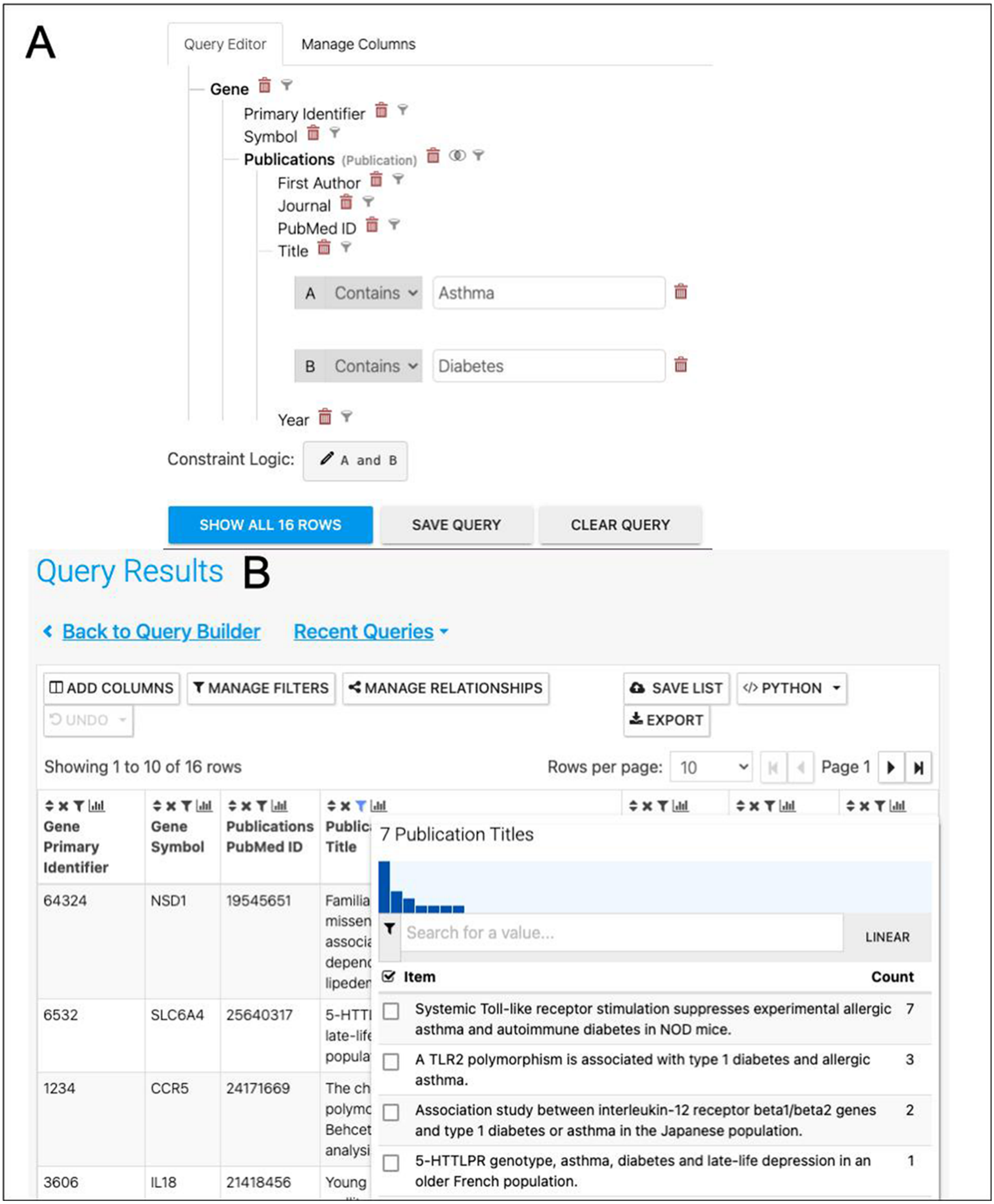


Figure 11. Gene expression heat maps for a selection of the genes from the shared set. (A). Protein tissue localization (original data from The Protein Atlas project, https://www.proteinatlas.org/humanproteome/tissue). The viewer has been filtered to show data for adipose tissue (adipocytes) and lung (macrophages and pneumocytes) only. (B). It is possible to toggle the expression score 'bins' on or off and a colour scale representing expression level is shown. (C). Heat map filtered to show RNA-seq data for adipose and lung (original data from The Protein Atlas project, https://www.proteinatlas.org/humanproteome/tissue). The viewer allows toggling between other expression data sets, showing different binned levels of expression and provides a scale for expression level.

Toll-like receptors (TLRs) are expressed on the membranes of leucocytes and activate immune cell responses in response to microbes. Such activation of TLRs in the lung can induce inflammation, inflammatory cell recruitment and cytokine release. In particular, TLR2 and TLR4 have been shown to sustain the inflammatory response in asthma (48). There is also evidence that TLR4 is directly involved in the pathophysiology of type 2 diabetes (49, 50) and it has been suggested that drugs targeting the TLR4 signal pathway could be candidates for the treatment of inflammatory comorbidities (51). The 'Pathway → Genes' template was used to retrieve all known genes of the TLR4 signalling pathway. List intersection shows that 45 genes from the shared set function in this pathway (out of a total of 128 pathway genes). Interestingly, 34 of these 45 proteins have a drug entry in DrugBank (52) [Source: TargetMine (15); https://targetmine.mizuguchilab.org], reinforcing the therapeutic potential of this pathway.

A number of other genes in the shared set warrant further analysis. For example, the enriched pathways include a number related to GPCR signalling with 128 genes being annotated with 'signalling by GPCR'. Included in these are BDKRB1 and BDKRB2, receptors for pro-inflammatory bradykinins, which have an established role in asthma (53) but less is known about their roles in type 2 diabetes. BDKRB1 has previously been identified as upregulated in allergic airway inflammation (54) and a mouse model of diabetes (55) and maybe worth further investigation.

The HumanMine integrated data allows many other datasets to be examined alongside the disease data. For instance, five datasets allow investigation of gene expression at the tissue level (E-MTAB-62 and E-MTAB-513 (from https://www.ebi.ac.uk/arrayexpress), GTex (from https://gtexportal.org) and Human Protein Atlas (both protein localization data and RNA-seq data, (56)). Such data allow other factors to be considered in the above analysis, for instance, genes from the shared set that are expressed in both the lung and adipose tissue. An initial analysis using the Human Protein Atlas (56) RNA-seq data indicates that 406 of the genes from the shared set are expressed in both tissues. These are enriched for the immune system and signal transduction pathways as discussed above and include IL6, TLR2 and TLR4. The expression viewers on the list analysis page allow for a visual overview of the genes in the shared list that are expressed in the lung and adipose tissue (Figure 11).



A Contains Asthma

B Contains Diabetes

Constraint Logic: A and B

Showing 1 to 10 of 16 rows

Rows per page: 10 Page 1

Gene Primary Identifier	Gene Symbol	Publications PubMed ID	Public Title
64324	NSD1	19545651	Familia missen associa depenc lipeder
6532	SLC6A4	25640317	5-HTTI late-life popula
1234	CCR5	24171669	The ch polymc Behcet analysi
3606	IL18	21418456	Young

7 Publication Titles

Search for a value...

Item	Count
□ Systemic Toll-like receptor stimulation suppresses experimental allergic asthma and autoimmune diabetes in NOD mice.	7
□ A TLR2 polymorphism is associated with type 1 diabetes and allergic asthma.	3
□ Association study between interleukin-12 receptor beta1/beta2 genes and type 1 diabetes or asthma in the Japanese population.	2
□ 5-HTTLPR genotype, asthma, diabetes and late-life depression in an older French population.	1

Figure 12. (A). A query to find publications in which the title includes both Asthma and Diabetes. (B). The results return seven publications as shown by the column summary.

Each expression set provides a slightly different result, emphasizing the benefit of being able to examine multiple datasets alongside each other.

Expanding the analysis further, the network of interacting proteins for the shared set can be examined, the interacting partners of each gene being found using the 'gene → interactions' template. Enrichment analysis of the resulting 'expanded' set highlights signalling by Rho GTPases. The RHO GTPases are emerging regulators of glucose homeostasis and have been implicated in beta cell dysfunction (57).

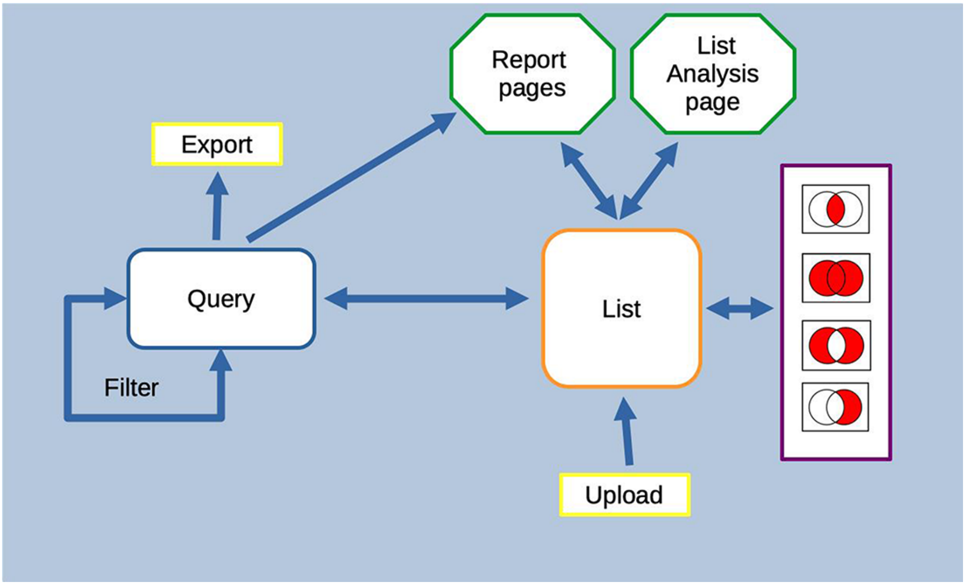


Figure 13. Features of the InterMine interface can be combined to create iterative workflows. For instance, the entities from the results of a query generated either using the query builder or through a template search can be saved as a list. The list can be fed into further queries or combined with other lists using set operations to create further lists. At any stage, individual entities or lists can be examined in more detail through the list analysis and report pages.

They are also well known to play an important role in the pathophysiology of asthma and recent advances have suggested novel roles for RhoA in regulating allergic airway inflammation. The RHO GTPases are promising targets for therapeutic treatment of asthma and two major Rho-kinase inhibitors have been developed (58).

A further way to enhance the analysis would be to investigate the literature for publications exploring the asthma/diabetes comorbidity, for instance, by searching for papers that mention both in either their titles or abstracts. In addition to searching data, HumanMine makes it easy to begin a literature search for the interesting sets of genes identified, by means of the PubMed-to-gene mappings provided by PubMed. Editing the template 'Gene → Publications' in the query builder to remove the gene constraint and adding two constraints on the publication title (Figure 12) allows publications in which the title includes both Asthma and Diabetes to be easily identified. Seven publications are returned by the search, interestingly identifying a TLR2 polymorphism associated with type 1 diabetes and allergic asthma (Figure 12). In addition, the publication enrichment for the shared set can also be searched for interesting clues. Although no papers directly analysing diabetes and asthma appear to cite an unexpectedly high number of the shared set members, there are some potentially interesting leads—for instance, 37 genes from a paper entitled 'Analyses of shared genetic factors between asthma and obesity in children'.

While it is not possible to provide a full analysis of the data and its implications in the therapeutic treatment of asthma and diabetes here, the above use-case illustrates how HumanMine facilitates an in-depth exploration of the data available and highlights many areas where further research could be focussed. The method outlined above can be applied to any combination of disease comorbidity and can easily be adapted to other research areas.

Conclusion

HumanMine is an advanced biological data warehouse for human genomics and proteomics data, exposing the data through intuitive graphical user interfaces and as well as allowing programs to access data through extensive web services. The *PAX6* and disease comorbidity examples illustrate how HumanMine can be used for powerful explorative and iterative data analysis. The combination of list creation, list operations, querying and filtering allows a workflow of tasks to be built up, the results of which can be examined at various stages using enrichment tools, list tools and report pages (Figure 13). Furthermore, the combination of automatic code generation and web services makes it easy to replicate such steps in a script and so repeat the analysis at scale with different starting conditions and parameters.

Supplementary data

Supplementary data are available at *Database* Online.

Funding

This work was supported by the Wellcome Trust [099133, 208381].

Conflict of interest

None declared.

References

1. Motenko,H., Neuhauser,S.B., O'Keefe,M. *et al.* (2015) MouseMine: a new data warehouse for MGI. *Mamm. Genome*, **26**, 325–330.

2. Wang,S. J., Laulederkind,S.J.F., Hayman,G.T. *et al.* (2013) Analysis of disease-associated objects at the Rat Genome Database. *Database*, **2013**, bat046.
3. Ruzicka,L., Bradford,Y.M., Frazer,K. *et al.* (2015) ZFIN, The zebrafish model organism database: Updates and new directions. *Genesis*, **53**, 498–509.
4. Lyne,R., Smith,R., Rutherford,K. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, **8**, R129.
5. Balakrishnan,R., Park,J., Karra,K. *et al.* (2012) YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database*, **2012**, bar062.
6. Krishnakumar,V., Contrino,S., Cheng,C. Y. *et al.* (2016) ThaleMine: a warehouse for arabidopsis data integration and discovery. *Plant Cell Physiol*, **58**, e4. 10.1093/pcp/pcw200.
7. Smith,R.N., Aleksic,J., Butano,D. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, **28**, 3163–3165.
8. Rozanski,A., Moon,H., Brandl,H. *et al.* (2019) PlanMine 3.0-improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic Acids Res.*, **47**, D812–D820.
9. Elsik,CG., Tayal,A., Diesh,CM. *et al.* (2016) Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Research*, **44**(D1), D793–800. 10.1093/nar/gkv1208.
10. Yang,P., Hou,L., Wang,X. *et al.* (2019) Core transcriptional signatures of phase change in the migratory locust. *Protein Cell*, **10**, 883–901.
11. Yang,P., Wang,D., Guo,W. *et al.* (2021) FAWMine: an integrated database and analysis platform for fall armyworm genomics. *Insect Sci.*, **28**, 590–601.
12. Krishnakumar,V., Kim,M., Rosen,B.D. *et al.* (2015) MTGD: the Medicago truncatula genome database. *Plant Cell Physiol.*, **56**, e1.
13. Shamimuzzaman,M., Gardiner,J.M., Walsh,A.T. *et al.* (2020) MaizeMine: a data mining warehouse for the maize genetics and genomics database. *Front. Plant Sci.*, **11**, 592730.
14. Redsun,S., Hokin,S., Cameron,C.T. *et al.* (2022) Doing genetic and genomic biology using the legume information system and associated resources. *Methods Mol. Biol.*, **2443**, 81–100.
15. Chen,Y. A., Tripathi,L.P. and Mizuguchi,K. (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, **6**, e17844.
16. Goodstein,D.M., Shu,S., Howson,R. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
17. Contrino,S., Smith,R.N., Butano,D. *et al.* (2012) modMine: flexible access to modENCODE data. *Nucleic Acids Res.*, **40**, D1082–D1088.
18. Demir,E., Cary,M.P., Paley,S. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.
19. Sivade Dumousseau,M., Alonso-López,D., Ammari,M. *et al.* (2018) Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinform.*, **19**, 134.
20. Zhou,P., Emmert,D. and Zhang,P. (2006) Using Chado to store genome annotation data. *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9.6.
21. Eilbeck,K., Lewis,S.E., Mungall,C.J. *et al.* (2005) The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
22. Kalderimis,A., Lyne,R., Butano,D. *et al.* (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, **42**, W468–W472.
23. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J.J. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
24. Martin,A.R., Williams,E., Foulger,R.E. *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.
25. Uhlen,M., Oksvold,P., Fagerberg,L. *et al.* (2010) Towards a knowledge-based Human Protein Atlas. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
26. Orchard,S., Ammari,M., Aranda,B. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
27. Oughtred,R., Stark,C., Breitkreutz,B. J. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
28. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J. *et al.* (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
29. Coutinho,P., Pavlou,S., Bhatia,S. *et al.* (2011) Discovery and assessment of conserved Pax6 target genes and enhancers. *Genome Res.*, **21**, 1349–1359.
30. Sayers,E.W., Bolton,E.E., Brister,J.R. *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **50**, D20–D26.
31. Howe,K.L., Achuthan,P., Allen,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
32. Tweedie,S., Braschi,B., Gray,K. *et al.* (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.
33. Kyritsis,K.A., Wang,B., Sullivan,J. *et al.* (2019) InterMineR: an R package for InterMine databases. *Bioinformatics*, **35**, 3206–3207.
34. Ehrlich,S.F., Quesenberry,C.P., Jr, Van Den Eeden,S.K. *et al.* (2010) Patients diagnosed with diabetes are at increased risk for asthma, chronic obstructive pulmonary disease, pulmonary fibrosis, and pneumonia but not lung cancer. *Diabetes Care*, **33**, 55–60.
35. Amberger,J.S., Bocchini,C.A., Schiettecatte,F. *et al.* (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
36. Landrum,M.J., Lee,J.M., Benson,M. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
37. Buniello,A., MacArthur,J.A.L., Cerezo,M. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
38. Köhler,S., Gargano,M., Matentzoglu,N. *et al.* (2021) The Human Phenotype Ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
39. Song,Y., Klevak,A., Manson,J.E. *et al.* (2010) Asthma, chronic obstructive pulmonary disease, and type 2 diabetes in the Women's Health Study. *Diabetes Res. Clin. Pract.*, **90**, 365–371.
40. Liu,S., Tinker,L., Song,Y. *et al.* (2007) A prospective study of inflammatory cytokines and diabetes mellitus in a multiethnic cohort of postmenopausal women. *Arch. Intern. Med.*, **167**, 1676–1685.
41. Song,Y., Manson,J.E., Tinker,L. *et al.* (2007) Circulating levels of endothelial adhesion molecules and risk of diabetes in an ethnically diverse cohort of women. *Diabetes*, **56**, 1898–1904.
42. Yang,Y.C., Zhang,N., Van Crombruggen,K. *et al.* (2012) Transforming growth factor-beta1 in inflammatory airway disease: a key for understanding inflammation and remodeling. *Allergy*, **67**, 1193–1202.
43. Ierodiakonou,D., Postma,D.S., Koppelman,G.H. *et al.* (2013) TGF-β1 polymorphisms and asthma severity, airway inflammation, and remodeling. *J. Allergy Clin. Immunol.*, **131**, 582–585.
44. Toren-Haritan,G. and Efrat,S. (2015) TGFβ pathway inhibition redifferentiates human pancreatic Islet β cells expanded in vitro. *PLoS One*, **10**, e0139168.
45. White,M.G., Marshall,H.L., Rigby,R. *et al.* (2013) Expression of mesenchymal and α-cell phenotypic markers in islet β-cells in recently diagnosed diabetes. *Diabetes Care*, **36**, 3818–3820.

46. Vignola,A.M., Chanez,P., Bonsignore,G. *et al.* (2000) Structural consequences of airway inflammation in asthma. *J. Allergy Clin. Immunol.*, **105**, S514–7.
47. Al-Alawi,M., Hassan,T. and Chotirmall,S.H. (2014) Transforming growth factor β and severe asthma: a perfect storm. *Respir. Med.*, **108**, 1409–1423.
48. Lafferty,E.I., Qureshi,S.T. and Schnare,M. (2010) The role of toll-like receptors in acute and chronic lung inflammation. *J. Inflamm.*, **7**, 57.
49. Ladefoged,M., Buschard,K. and Hansen,A.M.K. (2013) Increased expression of toll-like receptor 4 and inflammatory cytokines, interleukin-6 in particular, in islets from a mouse model of obesity and type 2 diabetes. *Apmis*, **121**, 531–538.
50. Amyot,J., Semache,M., Ferdaoussi,M. *et al.* (2012) Lipopolysaccharides impair insulin gene expression in isolated islets of Langerhans via Toll-Like Receptor-4 and NF-κB signalling. *PLoS One*, **7**, e36200.
51. Lucas,K. and Maes,M. (2013) Role of the Toll Like receptor (TLR) radical cycle in chronic inflammation: possible treatments targeting the TLR4 pathway. *Mol. Neurobiol.*, **48**, 190–204.
52. Wishart,D.S., Feunang,Y.D., Guo,A.C. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
53. Ricciardolo,F.L.M., Folkerts,G., Folino,A. *et al.* (2018) Bradykinin in asthma: Modulation of airway inflammation and remodelling. *Eur. J. Pharmacol.*, **827**, 181–188.
54. Christiansen,S.C., Eddleston,J., Woessner,K.M. *et al.* (2002) Upregulation of functional kinin B1 receptors in allergic airway inflammation. *J. Immunol.*, **169**, 2054–2060.
55. Gabra,B.H. and Sirois,P. (2002) Role of bradykinin B(1) receptors in diabetes-induced hyperalgesia in streptozotocin-treated mice. *Eur. J. Pharmacol.*, **457**, 115–124.
56. Uhlén,M., Fagerberg,L., Hallström,B.M. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
57. Møller,L.L.V., Klip,A. and Sylow,L. (2019) Rho GTPases-emerging regulators of glucose homeostasis and metabolic health. *Cells*, **8**, 434. 10.3390/cells8050434.
58. Zhang,Y., Saradna,A., Ratan,R. *et al.* (2020) RhoA/Rho-kinases in asthma: from pathogenesis to therapeutic targets. *Clin. Transl. Immunol.*, **9**, e01134.